Investigation of analytical bioinformatic approaches to genomic and transcriptomic data

Karen Crawford

July 2022

School of Medicine

Cardiff University

Supervisors:

Prof. Valentina Escott-Price

Dr Dobril Ivanov

Prof. Michael O'Donovan

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Summary

Alzheimer's disease (AD) is a devastating form of neurodegeneration that is characterised by the formation of amyloid plaques and tau tangles in the brain. Genome-wide association studies (GWAS) have identified over 70 risk loci. How these functionally relate to AD is still yet to be fully explored. The work presented in this thesis aims to integrate and interrogate three publicly available genetic and RNA-sequencing datasets. This is with the aim to increase our understanding of the mechanisms underlying AD biology. This was achieved by utilising a variety of bioinformatic analyses.

Chapter 1 introduces the background of AD, an overview of the relevant literature and the aims of this thesis. Chapter 2 gives an overview of some of the bioinformatic methodology used throughout this thesis. Chapter 3 uses linear mixed-effect models in addition to principal component analysis to combine the ROSMAP, MSBB and MayoRNAseq bulk brain RNA-sequencing datasets into a single dataset. This dataset was then utilised in chapter 4 to perform a differential gene expression analysis followed by a gene ontology enrichment analysis. This identified that GWAS prioritised genes are not enriched in differential gene expression derived from case-control bulk RNA-sequencing data. This analysis also implicated pathways associated with mitochondrial processes and the endoplasmic reticulum in AD. Chapter 5 explores a cis- and trans- eQTL analysis of differentially expressed genes that were identified in chapter 4. This identified *SST*, *TAC1*, *MAF1* and *SCGN* as potential candidate risk genes for AD. Chapter 6 compares the results of the differential gene expression analysis (from chapter 4) to three published Transcriptome-Wide Association studies (TWAS) results. This identified that in AD, TWAS signals are not enriched in bulk brain DGE analysis. Chapter 7 is a discussion of the results of this thesis and directions for possible future study.

Acknowledgements

I am extremely grateful to my supervisor Valentina Escott-Price without whom I could not have completed this PhD. Her invaluable advice and insights throughout, along with her enthusiasm have encouraged me throughout the PhD process. I also wish to thank my other supervisors Dobril Ivanov and Michael O'Donovan for their advice and feedback throughout the creation of this thesis.

I would like to thank the MRC for funding this PhD and therefore for all of the experience and knowledge I have gained over the past four years. I would also like to thank Cardiff University for their support enabling me to complete this thesis.

I would like to acknowledge that the data available in the AD Knowledge Portal would not be possible without the participation of research volunteers and the contribution of data by collaborating researchers.

I would like to thank Ganna Leonenko, Alun Meggy, the AD field team, Peter Holmans and Ioanna Katzourou for their contributions to this thesis.

I would like to thank my family and also thank my closest friends Amy, Carly, Colette, Ella and Thea for their unwavering support, love, humour and perspective. You are all angels and the most fabulous friends anyone could ask for.

I also wish to thank my husband Julian for his love, support and encouragement throughout. Without you this thesis would not have been completed and I look forward to all the adventures and ventures to come now that this chapter of our lives has come to an end.

Finally, I would like to mention my son Leo for being the biggest source of joy these past two years and thank him for teaching me what is truly important in life.

Abbreviations

AD	Alzheimer's disease
AMP-AD	Accelerating Medicines Partnership – Alzheimer's Disease
APOE	Apolipoprotein E
APP	Amyloid precursor protein
BP	Base pair
BM10	Brodmann area 10 (frontal pole)
BM22	Brodmann area 22 (parahippocampal gyrus)
BM36	Brodmann area 36 (inferior frontal gyrus)
BM44	Brodmann area 44 (superior temporal gyrus)
CDR	Clinical dementia rating
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
CI	Cognitive impairment
CPM	Counts per million
CQN	Conditional quantile normalization
CSF	Cerebrospinal fluid
DNA-seq	DNA Sequencing
DEG	Differentially expressed gene
DGE	Differential gene expression
EOAD	Early-onset Alzheimer's disease
eGene	The gene of an eQTL SNP-gene pair
eQTL	Expression quantitative trait loci
FDA	Food and Drug Administration
FDR	False discovery rate

FUMA	Functional mapping and annotation
GC	Guanine-Cytosine
GO	Gene ontology
GTEx	Genotype-Tissue Expression (project)
GWAS	Genome-wide association study
GWAX	Genome-wide association study-by-proxy
IGAP	International Genomics of Alzheimer's Project
Kb	Kilobase
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LD	Linkage disequilibrium
LMEM	Linear mixed-effect model
LOAD	Late-onset Alzheimer's disease
MAGMA	Multi-marker Analysis of Genomic Annotation
MAP	Memory and Aging Project
Mb	Megabase
MCI	Mild cognitive impairment
MSBB	Mount Sinai Brain Bank
NCI	No cognitive impairment
NFT	Neurofibrillary tangle
PA	Pathological ageing
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PMI	Post-mortem interval
PSEN	Presenilin
PSP	Progressive supranuclear palsy

QC	Quality control
QQ	Quantile-quantile
RIN	RNA integrity number
ROS	Religious Orders Study
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
SD	Standard deviation
SNP	Single nucleotide polymorphism
TWAS	Transcriptome-wide association study
WGS	Whole genome sequencing

Contributions and statements of others work

Ganna Leonenko supported this work by advising and performing the ancestry analysis and genetic QC (chapter 3).

Alun Meggy and the AD field team advised on the definition of AD phenotype based on the data we had available to us and their own professional experience alongside Ganna Leonenko and Dobril Ivanov (chapter 3).

Peter Holmans produced the list of most up-to-date GO terms and this was provided to me by Ioanna Katzourou with Peter's permission (chapter 4).

Dobril Ivanov provided the software CATMAP and associated example scripts for running the gene ontology pathway analyses which I amended for performing my own analyses (chapter 4).

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

The purpose of this chapter is to introduce the wider context and history of the AD field in which the work presented in this thesis sits. This chapter will also provide a summary of the overall aims of the work included in this thesis.

1.1 Alzheimer's Disease

1.1.1 Overview

Alzheimer's disease (AD) was originally described by Alois Alzheimer in 1907. He described an individual suffering from memory loss, paranoia, disorientation to space and time and hallucinations. The individual declined over a period of four years until their eventual death. Upon autopsy, brain atrophy was evident and changes to neurons and glial cells were documented (Alzheimer 1907; Alzheimer et al. 1995). AD is the most common form of dementia, which is an umbrella term for a variety of neurodegenerative disorders which all typically feature cognitive impairment, and mental and physical deterioration. AD accounts for 50-70% of all dementia cases (Winblad et al. 2016).

There are a few forms of AD, and they can be divided based on age of visible symptom onset and genetic predisposition. AD is rare in younger individuals and over 95% of AD cases are what is known as sporadic or late-onset AD (LOAD). This occurs when the visible symptoms develop after 65 years of age although disease processes are likely to be earlier. Conversely, early-onset AD (EOAD) occurs when age of disease onset and visible symptoms are before 65 years of age (Bali et al. 2012). Rare autosomal dominant forms of AD also exist due to mutations in the amyloid precursor protein gene (*APP*) and presenilin genes (*PSEN1* and *PSEN2*). Knowledge of

these mutations has contributed to current understanding of both early- and late-onset AD disease processes (Van Cauwenberghe et al. 2016).

1.1.2 Pathology of AD

The pathology of the AD brain can be divided into microscopic and macroscopic features. The microscopic changes usually seen in AD include progressive loss of the synaptic connections between neurons and eventually the neurons themselves. The hippocampus (which is important for memory formation) is one of the first areas to be affected. As the disease progresses, macroscopic features are evident such as cortical atrophy. The brain appears to be shrunken and the frontal and temporal cortices have enlarged sulcal spaces (Drew 2018; Knopman et al. 2021).

Further microscopic changes include the accumulation of the amyloid-beta peptide. Amyloid-beta is produced by the cleavage of APP in the membrane of neurons. Amyloid-beta forms oligomers compromising neuronal membrane integrity which is thought to result in synaptic disfunction. Fibrils of amyloid-beta aggregate into what are known as plaques. These insoluble plaques form between neurons and interfere with their function (Drew 2018; Knopman et al. 2021).

Another microscopic change is seen with tau, which is a microtubule-associated protein that is present in the cytoplasm of axons. Misfolded tau protein aggregates into neurofibrillary tangles (NFT) inside of neurons, displacing intracellular organelles. Misfolded tau can pass through synapses to other neurons, where it catalyses further misfolding of tau (Drew 2018; Knopman et al. 2021).

1.1.3 Staging

The pathophysiology and symptomology of AD is understood to be on a continuum. Patients transition from normal cognition to a pre-dementia phase of AD referred to

as mild cognitive impairment (MCI). This is then followed by a progressive dementia advancing through mild, moderate and severe stages of AD (Sperling et al. 2011; Davis et al. 2018).

Symptoms of AD typically include both cognitive and neuropsychiatric symptoms and vary depending on disease stage (Atri 2019b). Symptoms initially start as memory problems, impaired judgement or behaviour but do not affect the independence of the individual (Davis et al. 2018). Early pathological changes can already be seen in the cortex and hippocampus (Breijyeh and Karaman 2020). As the disease progresses, cognition worsens and neuropsychiatric symptoms such as depression, apathy, hallucinations, and delusions can become exacerbated. Over time these symptoms can have an increasingly adverse effect on daily function, quality of life and they can have a huge impact on caregivers (Lyketsos et al. 2011).

Eventually the individual will decline to a severe AD stage. At this stage, pathological changes in the entire cortex area can be seen. Patients will have lost independence due to severe cognitive impairment and potentially due to difficulties with swallowing and urination. Eventually death occurs due to complications such as pneumonia, urinary tract infection, dehydration and sepsis (Breijyeh and Karaman 2020). The symptoms and rate of decline varies between individuals and are difficult to predict and are not well understood (Davis et al. 2018).

1.1.4 Diagnosis

Symptoms of AD are very heterogenous and in the early stages of disease are often misattributed to other conditions, dismissed or ignored leading to delays in diagnosis and missed opportunities for intervention. Diagnosis of AD is challenging as there are no biomarker tests available for clinical use. Diagnosis is mainly based on a clinical and exclusionary approach (Scheltens et al. 2021).

Diagnosis is usually decided based on a clinical interview with the patient and an informant in addition to cognitive and physical examinations. Blood tests are performed to exclude conditions that may cause cognitive symptoms. Brain imaging is used to identify structural changes in the brain. Genetic testing may be used when an autosomal dominant case of AD may be suspected but routine genetic testing is not currently recommended (Lane et al. 2018).

Research is currently focused on identifying a potential biomarker that would be usable in the clinical setting. Amyloid-beta, total tau, and phosphorylated tau levels in cerebrospinal fluid (CSF) have been considered for use as biomarkers and have been shown to be moderately successful in diagnosing not only AD dementia but prodromal AD. However, a blood based biomarker would be preferred as it is less invasive to collect than CSF (Blennow and Zetterberg 2018) . Recently a blood biomarker assay of phosphorylated tau has been developed. This assay was able to differentiate AD dementia from cognitively unimpaired older adults with an area under the curve (AUC) of 90.21-98.24% (Karikari et al. 2021).

Recently a research framework has been proposed to diagnose individuals living with AD using a range of biomarkers. The biomarkers are usually a mixture of imaging and biofluid markers (often CSF) and aim to capture the following underlying AD biology: amyloid-beta deposition, pathological tau, and neurodegeneration. This is referred to as Amyloid/Tau/Neurodegeneration (ATN) classification (Jack et al. 2018). The framework has the advantage in that it aims to provide individuals a biologically driven classification of AD. The framework considers AD as a continuum rather than a binary AD case-control status based solely on observed symptoms. The ATN classification has the advantage that additional biomarkers can be added as understanding of disease biology progresses, and this is an active area of research (Kasuga et al. 2022).

Tau protein phosphorylated at threonine 181 (p-tau) is a biomarker used for tau deposition (T) and CSF total tau is used as a biomarker for neurodegeneration (N) within the ATN framework. However, this has been criticised as they are so highly correlated that there may be limited value in using total tau when diagnosing AD using the ATN framework (Soldan et al. 2019; Cousins et al. 2021). CSF neurofilament light chain protein has been suggested as an alternative marker of N. Neurofilament light chain is an axonal protein that has been shown to be elevated in CSF in a range of neurodegenerative conditions including AD and dementia (de Jong et al. 2007). Neurofilament light chain has been suggested as a promising biomarker for identifying preclinical disease as it is dynamic and captures different biology to total tau (Mattsson-Carlgren et al. 2020; Kasuga et al. 2022)

1.1.5 Epidemiology and projections of AD

AD is not simply a product of normal ageing and the disease is not only distressing to the patient, but it is also onerous to their families, caregivers and the healthcare system (Atri 2019b). In 2015, it was estimated that nearly 47 million people were living with AD globally. In 2019, this had risen to 57.4 million people. This number is predicted to rise to more than 167 million by 2050 with some models predicting that 1 in 85 people would be living with the disease if no cure or prevention is found (Brookmeyer et al. 2007; Prince et al. 2015). AD is not only devastating at a personal level but has a large economic cost. The current worldwide annual cost of dementia is estimated to be one trillion US dollars and predicted to reach two trillion US dollars by 2030 if no incidence-altering treatment is found (Prince et al. 2015; Wittenberg et al. 2020; Nichols 2022).

Globally there is an increasingly ageing population due in part to individuals living longer. It is therefore anticipated that the global burden of AD and related dementias will increase. In 2019, it was estimated that almost 885,000 older people in the UK have dementia. If no new interventions are discovered, this is forecasted to be 1.6

million by 2040. Projections suggest that overall expenditure on care for older people with dementia will have risen by 249% in the period between 2015 to 2040. Total costs for dementia-related health and social care are predicted to rise from 0.8% of GDP to 1.9% of GDP in this same period. This is excluding unpaid care costs which usually impact the families of those affected by AD (Wittenberg et al. 2020). These figures are forecasted specifically with respect to the UK, but if a similar outcome is seen globally, it is clear that there is an urgent need for new treatments to reduce the global burden of disease. An intervention that could delay both disease onset and progression by one year could have a huge impact with an estimated reduction in 9.2 million fewer cases worldwide (Brookmeyer et al. 2007).

1.1.6 Treatments

There is currently no cure for AD and treatment strategies aim to retain quality of life and manage symptoms (Atri 2019a). Over the past 10 years, more than 100 AD drug candidates have either not made it past the drug development stage or through clinical trials (Dujardin et al. 2022). It is understood that the pathological processes underlying AD could begin decades before symptoms manifest (Atri 2019a). This could explain why clinical trials have been unsuccessful so far. Many trials included patients that are in the advanced stage of disease when widespread and likely irreversible damage to the brain has taken hold (Mehta et al. 2017). More modern trials are focussed on including individuals with milder disease, MCI and even utilising the ATN framework (Grill et al. 2019; Gregory et al. 2022; Kasuga et al. 2022).

Other criticisms of trials to date include problems with trial design, and potentially targeting the wrong pathological substrate. Amyloid-beta plaques and their elimination have been the primary target for many drugs, but success has been limited (Mehta et al. 2017; Dujardin et al. 2022). In 2021, the US Food and Drug Administration (FDA) approved Aducanumab, a monoclonal antibody which targets amyloid-beta. However only clearance of amyloid-beta has been demonstrated, and as such the efficacy of the drug on the impact on AD is unknown. Additionally, the

FDA advisory committee had advised that the drug should not be approved as there was a lack of evidence that Aducanumab would result in cognitive improvement. The decision to approve the drug by the FDA lead to three committee members resigning. In addition, there were further concerns over its safety profile. The European Medicines Agency did not approve the drug for use in the EU partially because brain scans of patients using the drug were found to show abnormalities such as swelling and bleeding. Additionally, they stated that no link between reducing amyloid in the brain and cognitive improvement was found and that the trial results were conflicting (Mahase 2021; Dujardin et al. 2022).

1.1.7 Risk factors for AD

The underlying cause of pathological changes in AD remain unknown. AD is considered a multifactorial disease with several risk factors (Zhang et al. 2020b). These can be separated into modifiable and non-modifiable risk factors, and these are summarised in Figure 1-1. It is estimated that around 35% of life-time risk of dementia is modifiable by factors such as diet, education, health care and socio-economic status which are often interlinked (Livingston et al. 2017). Other modifiable risk factors include exposure to air pollution, alcohol use and infections. Obesity, type 2 diabetes and cardiovascular diseases are also known risk factors for AD and dementia (Breijyeh and Karaman 2020). It is known that a wide variety of modifiable risk factors contribute to risk for AD but as to how they contribute to underlying disease mechanisms remains elusive. Health initiatives to prevent AD are largely aimed at encouraging individuals to lead active lifestyles with a healthy diet, especially at middle age due to a lack of pharmacological interventions (Silva et al. 2019). Targeting poverty and inequality would also be of benefit especially for individuals experiencing the most deprivation (Livingston et al. 2017).

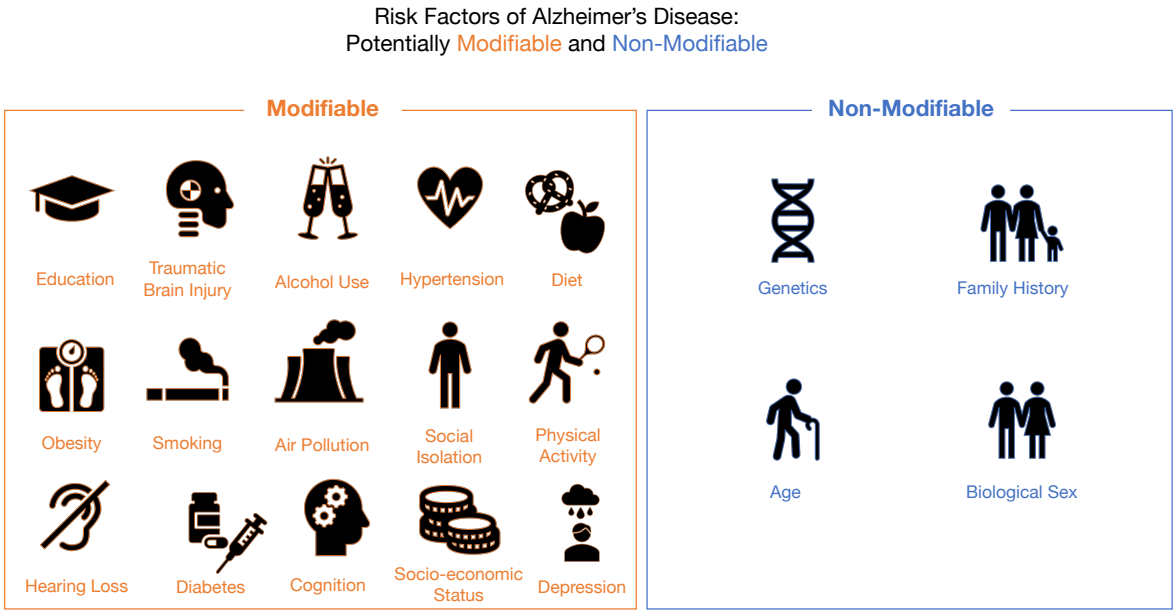


Figure 1-1 - A list of potentially modifiable and non-modifiable risk factors for AD

Risk factors that are non-modifiable include age, sex, genetics and family history. Ageing is considered to be one of the most important risk factors in AD for several reasons. Firstly, AD is rare in young people and most cases have a late onset which starts over the age of 65 years (Breijyeh and Karaman 2020). Furthermore, ageing is known to be a complex process that impacts on every cell in the body. It has been hypothesised that systems driving brain ageing such as amyloid-beta processing, inflammation, and mitochondria dysfunction contribute to AD risk. The female prevalence of AD is higher than that in males, which could be due in part to women having a greater lifespan than men on average (4.5 years) (Riedel et al. 2016).

As age is one of the most important risk factors for AD, recent research has given more attention to the concept and definition of age and ageing. Age can be defined using chronological age which is determined by the day a person is born. However, chronological age is not neccesssarily an accurate indicator of the biological process of ageing. Biological age has been proposed as a method to predict the ageing status of an individual or tissue (Milicic et al. 2022).

One measure of biological age is the epigenetic clock. Epigenetic clocks rely on DNA methylation, a process which involves the addition of a tag known as a methyl group to parts of the genome in which cytosine bases are bound to guanine bases through a phosphate group (CpG). When methylated, CpGs can act as binding sites for proteins that alter the DNA's structure. At numerous CpGs, methylation has been shown to decrease gene expression which could in turn affect biological function (Drew 2022). Horvath developed a ground-breaking epigenetic clock which applied an algorithm on the same 353 CpGs to predict biological age regardless of the DNA origin. This allowed the comparing of biological age from different tissues (Horvath 2013). More recently, another epigenetic clock has been developed specifically for human cortex that has the potential to identify phenotypes associated with biological ageing in the brain (Shireby et al. 2020). With time, these methods are likely to help increase our understanding of the impact biological ageing has on neurodegenerative diseases.

Genetics are also known to play an important role in AD and have contributed to our current understanding of AD disease processes. It is known that rare mutations account for familial forms of early-onset AD. Sporadic early-onset AD and late-onset AD share common pathological features and both are considered to be highly heritable (Li et al. 2021). The heritability of late-onset AD is estimated to be 58-79% (Gatz et al. 2006) and familial early-onset AD shows a heritability of over 90% (Wingo et al. 2012; Sims et al. 2020). Common polymorphisms in the apolipoprotein E (*APOE*) gene are known to be a major genetic risk factor for AD.

1.1.8 Early genetic studies of AD and the amyloid cascade hypothesis

Early genetic studies in AD have informed the main hypothesis of how AD occurs which has remained for over three decades now. In the early nineties, mutations were identified in *APP*, *PSEN1* and *PSEN2* which led to the development of the amyloid cascade hypothesis (Hardy and Allsop 1991; Hardy and Higgins 1992). The hypothesis is based on the idea that accumulation of amyloid-beta protein in the

brain is the main causal agent in the pathogenesis of AD. Neurofibrillary tangles, cell loss and eventually dementia follow (Hardy and Higgins 1992; Hardy and Selkoe 2002).

Amyloid-beta is produced by the enzymes beta-secretase and gamma-secretase by cutting APP in two places, which then creates a peptide fragment. Mutations in *PSEN1* and *PSEN2* were found to affect the location where gamma-secretase cuts APP resulting in a variant of amyloid-beta that clumps together more easily. This triggers the accumulation of amyloid-beta oligomers and then further accumulation produces insoluble fibrils which aggregate into plaques. This aggregation of amyloid-beta is thought to then trigger a cascade of disease (Makin 2018).

Further knowledge of disease has come from individuals with Down syndrome who have three copies of chromosome 21 (trisomy 21). The *APP* gene is found on chromosome 21, and amyloid plaques and tau neurofibrillary tangles are usually found in individuals with Down syndrome by the age of 40. The lifetime risk of AD in individuals with Down syndrome is more than 90% and Down syndrome is now considered a genetic cause of AD (Fortea et al. 2021).

1.1.9 The *APOE* gene

The strongest genetic risk factor for sporadic late-onset AD is the ε4 allele of the *APOE* gene. ApoE is a lipoprotein that is expressed in the brain. It is involved in cholesterol and lipid transportation and neuronal growth (Li et al. 2021). Three different alleles of *APOE* encode three different isoforms of ApoE. These are Apo E2, E3, and E4 and encoded by the ε2, ε3 and ε4 alleles of the *APOE* gene respectively (Liu et al. 2013). The three ApoE isoforms are determined by combinations of two SNPs (rs429358 and rs7412). There are three possible haplotypes which are presented in Table 1-1 and six possible genotypes (ε2/ε2, ε2/ε3, ε2/ε4, ε3/ε3, ε3/ε4, ε4/ε4) (Wu et al. 2020).

rs429358	rs7412	*APOE* allele
T	T	ε2
T	C	ε3
C	C	ε4

Table 1-1 The three APOE haplotypes (ε2/ε3/ε4) formed by two single nucleotide polymorphisms: rs429358 and rs7412

APOE ε4 carriers have an increased risk of AD with ε4/ε4 homozygotes carrying a 14.9-fold higher odds ratio (OR) of AD risk (Liu et al. 2013). Disease risk is also higher for ε4 heterozygotes (OR = 2.6 for ε2/ε4 and OR = 3.2 for ε3/ε4) (Liu et al. 2013; Li et al. 2021). In contrast the ε2 allele has been found to be protective against AD (OR=0.6) (Liu et al. 2013; Li et al. 2021).

1.1.10 GWAS and its applications in AD

The identification of susceptibility variants for common disease with high minor allele frequency, like *APOE*, led to the 'common disease common variant' hypothesis. This hypothesis is that common diseases like AD are influenced by genetic variation that is common within the population and that effect sizes for any one variant must be small relative to that found for rare disorders. Therefore, if common alleles with low penetrance contribute to the heritability of common disease, then multiple variants of low penetrance are contributing to disease susceptibility (Bush and Moore 2012). Some AD variants with varying frequency and risk effect size are summarised in Figure 1-2.

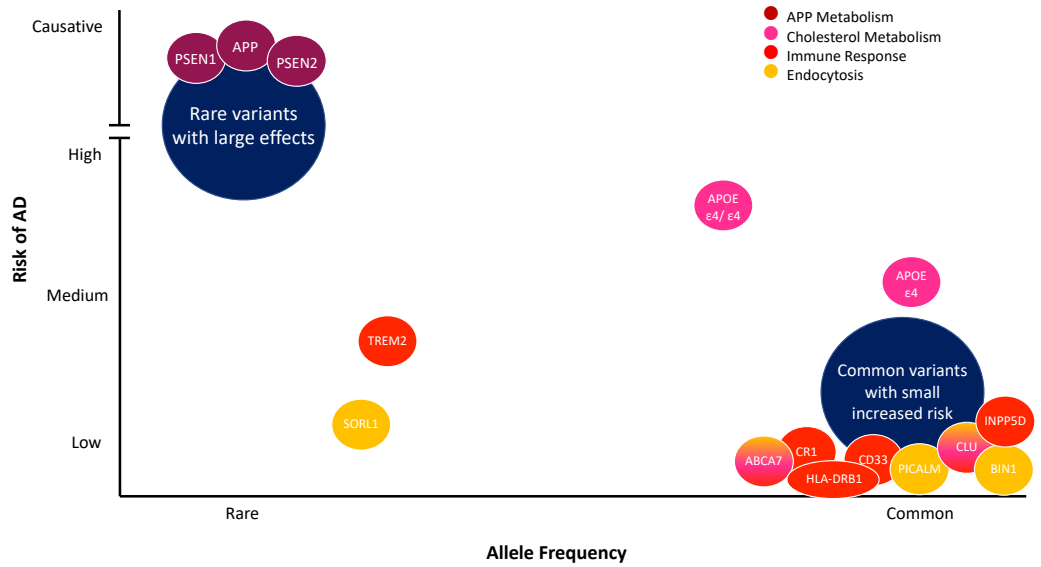


Figure 1-2 A selection of risk genes associated with AD presented by their risk allele frequency and the strength of their genetic effect. Colours in the legend indicate pathways in which the genes are involved. Adapted from (Lane et al. 2018) and (König and Stögmann 2021)

The arrival of high throughput genotyping has allowed genome-wide association studies (GWAS) to explore the genetic architecture of human disease and identify variants of small effect sizes that potentially contribute to the heritability of complex disorders. GWAS examines the association between millions of SNPs across the genome and a trait of interest with no underlying assumptions about disease biology (Uffelmann et al. 2021).

An important concept of GWAS to take into consideration is linkage disequilibrium (LD) (Li et al. 2021). LD is the degree to which an allele of a SNP or variant is inherited with an allele of another SNP within a population. The identification of an index SNP in GWAS means the index SNP can be used as a tag SNP for the surrounding region of LD (Bush and Moore 2012). This aspect of LD is important to take note of when considering the results from GWAS at it means that identified variants may not be causal but tagging the causal variant. Additionally, it means that findings from one population may not be applicable to another population due to differences in LD structure (Bush and Moore 2012; Li et al. 2021). Due to LD potentially obscuring

which is the causal variant driving a trait-association, any causal relationship cannot be ascertained from GWAS alone (Wainberg et al. 2019).

The first AD GWAS were started in 2007 where only SNPs within the *APOE* locus reached genome-wide significance and included around 2,000 cases and controls. As AD GWAS sample numbers increased, subsequent analyses identified more AD susceptibility loci including *CLU*, *PICALM*, *CR1*, *BIN1*, *ABCA7*, *MS4A*, *CD2AP*, *CD33* and *EPHA1* (Li et al. 2021).

In 2013, the International Genomics of Alzheimer's Project (IGAP) published a meta-analysis of previous AD GWAS including 17,008 cases and 37,154 controls (stage 1). This was followed by replication in an independent cohort of 8,572 disease cases and 11,312 controls (stage 2) (Lambert et al. 2013). These two stages of the meta-analysis confirmed previous GWAS findings and found a further 11 novel susceptibility loci: *HLA-DRB5*, *HLA-DRB1*, *PTK2B*, *SORL1*, *SLC24A4*, *INPP5D*, *MEF2C*, *NME8*, *ZCWPW1*, *CELF1*, and *FERMT2*.

In order to build on the previous findings of GWAS and increase statistical power, the first AD GWAS-by-proxy (GWAX) was performed in 2018. This utilised the UK Biobank by including 314,278 proxy AD cases and controls. The proxy cases and controls were based on self-reported parental history of dementia as opposed to an AD diagnosis of the individuals included in the study. These were then meta-analysed with the previous IGAP GWAS. This led to the identification of 27 AD susceptibility loci including three novel ones: *ADAM10*, *BCKDK/KAT8* and *CR1* (Marioni et al. 2018). The authors also performed expression quantitative trait loci (eQTL) analysis which is an approach used to identify variants associated with gene expression on the basis that a proportion of transcripts are under genetic control. A transcript that is correlated with a risk variant in a relevant tissue has often been considered to be a candidate susceptibility gene (Lawrenson et al. 2015). This led to the authors reporting *TOMM40*, *KAT8* and *CR1* as candidate causal genes in AD (Marioni et al. 2018).

In 2019, a larger GWAX was published consisting of 534,403 individuals. This was achieved by meta-analysing IGAP stage 1, a new sample from the Psychiatric Genetics Consortium (n=17,477), exome-wide data from the Alzheimer's Disease Sequencing Project and GWAX from the UK Biobank (71,880 proxy cases and 383,378 proxy controls). This analysis identified 29 susceptibility loci in which nine were novel: *ADAMTS4*, *HESX1*, *CLNK*, *CNTNAP2*, *ADAM10*, *APH1B*, *KAT8*, *ALPK2* and *AC074212.3* (Jansen et al. 2019). The authors used three gene-mapping strategies to link the AD-associated variants to candidate causal genes. This included positional gene-mapping, eQTL analysis and chromatin interaction mapping as implemented by the platform Functional Mapping and Annotation (FUMA) (Watanabe et al. 2017). This was followed by genome-wide gene-based association analysis using Multi-marker Analysis of Genomic Annotation (MAGMA) (de Leeuw et al. 2015). 215 potential causative genes were identified in total, but only 16 genes were implicated by all four approaches with seven of these genes being located outside of the *APOE* locus: *HLA-DRA*, *HLA-DRB1*, *PTK2B*, *CLU*, *MS4A3*, *SCIMP*, *RABEP1*.

Later in 2019 IGAP produced the largest GWAS of clinically diagnosed AD. IGAP Stage 1 was increased to 21,892 cases and 41,944 controls. The meta-analysis with stages 2 and 3 produced a final sample size of 35,274 cases and 59,163 controls with 24 susceptibility loci being discovered including three novel loci: *IQCK*, *ADAMTS1*, and *WWOX* (Kunkle et al. 2019). In order to determine candidate AD genes, Kunkle et al. determined a priority score for genes located within 500Kb of the LD region for the risk locus associated with each lead SNP. The priority score was determined by taking the sum of different lines of evidence including exonic functional annotation, eQTLs in AD relevant tissue then all tissues, correlation between expression and Braak (tau) pathology, differential expression in AD and evidence based on biological pathways. 400 candidate causative genes were identified across 24 loci including *ADAM10*, *ADAMTS1*, *ACE*, *IQCK*, *WWOX* and *MAF* (Kunkle et al. 2019). MAGMA pathway analyses were performed in the 2018 GWAX and both 2019 studies and all three implicated amyloid and tau processing, lipid metabolism, and the immune system.

In 2021, a meta-analysis of the UK Biobank GWAX dataset with the stage 1 data from Kunkle et al. (Schwartzentruber et al. 2021). This included 53,042 AD proxy cases, 21,982 AD cases and 397,844 controls (Li et al. 2021; Schwartzentruber et al. 2021). Schwartzentruber et al. identified 37 risk loci with novel associations near *CCDC6*, *TSPAN14*, *NCK2* and *SPRED2*. The authors performed colocalization between 36 of the discovered risk loci (excluding *APOE*) and 109 eQTL datasets representing a wide range of tissues, cell types and conditions. They also performed fine-mapping using FINEMAP, GCTA and PAINTOR and network analysis. Evidence from these analyses led the authors to suggest the following candidate causal genes: *CCDC6*, *TSPAN14*, *NCK2*, *SPRED2*, *BIN1*, *APH1B*, *PTK2B*, *PILRA*, *CASS4*, *ABCA7*, *SORL1*, *PICALM*, *SPI1*, *CR1*. Immune response, phagocytosis and the complement cascade pathways were found to be implicated in AD using gene enrichment analysis (Schwartzentruber et al. 2021).

In 2021, the 2019 GWAX was also expanded to include samples from additional cohorts including GR@CE and 23andMe. This resulted in the inclusion of 90,338 (44,613 proxy) cases and 1,036,225 (318,246 proxy) controls. This study identified 38 loci, with seven of them not being previously reported: *AGRN*, *TNIP1*, *HAVCR2*, *TMEM106B*, *GRN*, *NTN5*, and *LILRB2*. The authors' work utilised gene set analysis, chromatin enrichment analysis, eQTL enrichment analysis and functional consequence enrichment analysis which provided evidence for microglial pathways, amyloid and tau aggregation, and the immune system as AD- associated pathways (Wightman et al. 2021).

More recently in 2022, another study has been published including 111,326 cases (which are a mixture of clinically diagnosed and proxy cases) and 677,663 controls. Their samples came from the following consortia/datasets: EADB, GR@CE, EADI, GERAD/PERADES, DemGene, Bonn, the Rotterdam study, the CCHS study, NxC and UK Biobank. The samples included share a larger overlap with previously published GWAS and GWAX. The authors reported that they found 75 risk loci, of which 42

were new at the time of their analysis. Their gene prioritisation analysis identified 55 genes and the authors highlighted that six of them are expressed at a low level in microglia (*ICA1L*, *EGFR*, *RITA1*, *MYO15A*, *LIME1* and *APP*) which the authors concluded emphasised that AD and related dementias result from multiple cell types in the brain. A pathway enrichment analysis using their results confirmed the involvement of amyloid/tau pathways and implicated microglial endocytosis in AD (Bellenguez et al. 2022).

GWAS have shown that AD has a genetic underpinning that is highly polygenic and has allowed researchers to identify genetic variants associated with AD. These variants can be combined into a polygenic risk score (PRS) that captures part of an individual's susceptibility to disease (Lewis and Vassos 2020). PRSs have been shown to discriminate between pathologically confirmed AD cases and controls achieving an accuracy between 75-84% (Escott-Price et al. 2015; Escott-Price et al. 2017a). Prediction of disease status based on PRS alone is insufficient for precision medicine but can be used for other applications. One example is using PRS for induced pluripotent stem cell lines to identify and study cell lines which are at risk extremes. Selecting polygenic extremes can increase confidence in the cell line developing disease or remaining a control (Baker and Escott-Price 2020).

AD GWAS have contributed a lot to our understanding of the genetic architecture of AD. In the future it is likely that AD GWAS will continue to grow with the aim of finding new loci however it has been argued that will result in diminishing returns as current AD GWAS are not without their limitations (Escott-Price and Hardy 2022). It is likely that the GWAS are contaminated with dementia samples rather than being solely AD. This is due to AD being challenging to diagnose due to a lack of available biomarkers. Additionally, the GWAX rely on individuals accurately reporting their parents' diagnosis of dementia, so is also likely to be diluted with other types of dementia or not having dementia at all. The controls are potentially diluted too as some could be preclinical AD or dementia cases leading to a reduction of power. The future of GWAS may achieve more in furthering understanding of disease by using

only pathologically confirmed samples or those confirmed using biomarkers. This is likely to result in much smaller sample sizes in comparison to the GWAS/GWAX previously described, but likely to be better at informing on loci relevant to disease pathology and progression (Escott-Price and Hardy 2022).

Another limitation is the large overlap of participants between the separate GWAS meaning that the discovered loci are not truly independent findings and may be biased. The GWAS discussed so far all include individuals of European descent, so findings are biased towards a European-centric AD. More GWAS in other ancestries are required to develop a fuller understanding of AD genetic architecture.

One final limitation is that GWAS alone cannot identify which is the causal SNP or gene. The next challenge is how to identify which functional genetic variants and causal genes at these risk loci contribute to the biology of disease.

1.1.11 Expression quantitative trait loci and AD

Candidate functional SNPs affect genes differently depending on whether they are located in a regulatory or coding region. About 80% of the genetic susceptibility loci detected by GWAS were located in the non-coding regions (Zhao et al. 2019). This suggests that the pathogenic variants at these loci may be regulatory variants, or in other words, genetic variants that regulate expression levels of genes. Expression quantitative trait locus (eQTL) studies determine associations of genetic variants with gene expression (mRNA), but it is not possible to pinpoint which eQTLs are functionally important through eQTL associations alone.

Many studies (including many of the GWAS discussed above) utilise eQTL discovery in order to identify candidate genes driving disease risk (Gallagher and Chen-Plotkin 2018). EQTLs can be characterised as cis-acting or trans-acting eQTLs. Cis-eQTLs are those that affect the expression of nearby genes as opposed to trans-eQTLs which affect the expression of distant genes or even genes on different chromosomes. Distance in this context is usually defined depending on the study and is variable between studies. Typically, cis- is within 1Mb and trans- is a distance greater than 1Mb from the eQTL to the gene end. Research in AD using eQTLs has suggested that altered gene expression plays a role in the aetiology of AD however most studies focus only on cis-eQTLs (Zou et al. 2010; Sieberts et al. 2020; Patel et al. 2021). This is because detecting trans-eQTLs is more computationally intensive in addition to requiring larger sample sizes and being affected by multiple hypothesis testing burden. Trans-eQTLs also tend to have weaker effect sizes in comparison to cis-eQTLs (Clyde 2017).

Many eQTLs are shared between contexts (such as tissue, time, sex) but some eQTLs can be context specific. One study of over 400 healthy individuals found that although most eQTLs are shared, AD susceptibility alleles were enriched for monocyte-specific cis-eQTLs. The authors suggest that their results provide a genetic underpinning to the idea that the myeloid compartment of the immune system is driving the inflammatory component of susceptibility to neurodegenerative diseases. Additionally, as they studied young and healthy individuals, their results provide support for a role of myeloid cells in the prodromal phase of neurodegenerative diseases like AD (Raj et al. 2014).

One study compared cell-type specific eQTLs (ct-eQTLS) with bulk eQTLs from whole blood and brain. The authors then investigated the association of eQTLs with AD in addition to performing a differential gene expression analysis. They used brain samples from the Religious Orders Study/Memory and Aging Project (ROSMAP) (n=475) and blood samples from the Framington Heart Study (n=5,257). 24,028 significant SNP-gene eQTL pairs were found to be shared between blood and brain

with 386 distinct eGenes (the gene of the SNP-gene eQTL pair). 308 of these eGenes were differentially expressed between AD cases and controls. Six AD-associated genes (*CR1*, *ECHDC3*, *HLA-DRB1*, *HLA-DRB5*, *LRRC2*, and *WWOX*) were shared eQTL genes between brain and blood. *CR1*, *HLA-DRB1*, *HLA-DRB5* and *ECHD3* showed different patterns of association between AD cases and controls whereas *WWOX* and *LRRC2* were differentially expressed in AD cases versus controls (Patel et al. 2021). Additional significant gene-SNP eQTL pairs in the brain (n = 11,649) and blood (n = 2,533) were observed in ct-eQTL analysis that were not detected in the bulk eQTL analysis (Patel et al. 2021). These findings demonstrate that eQTL analysis in AD can help identify candidate genes involved in AD in a context specific manner which may provide mechanistic insights.

An area that has been understudied in AD is eQTL mapping of non-coding RNA such as microRNA and the study of microRNA's role in AD itself. MicroRNAs are known to be post-transcriptional regulators of gene expression by binding to target mRNA (Sonehara et al. 2021). It has been estimated that at least 1% of the human genome encodes microRNAs and each one could regulate up to 200 mRNAs. MicroRNA dysregulation has been implicated in AD and microRNAs are known to play a role in APP degradation and amyloid-beta metabolism through regulating gene expression. Dysregulation of microRNAs in AD is a relatively new area of research but is an important area of study as there is growing evidence that microRNAs may have utility as diagnostic markers in AD. Previous research has shown that combining between two and four microRNAs could distinguish AD cases and controls with an accuracy of between 75-82% (Angelucci et al. 2019; Wei et al. 2020).

1.1.12 Transcriptome-wide association studies and their application in AD

Large reference panels of eQTLs are now available for researchers to utilise, such as those from the Genotype-Tissue Expression (GTEx) project which has generated eQTLs from 54 different non-diseased tissues (GTEx-Consortium 2013). One method

that utilises reference panels of eQTLs is transcriptome-wide association studies (TWAS).

TWAS is a gene-based association method that identifies associations between genetically regulated gene expression and complex traits such as AD. The first stage of a TWAS imputes genetically regulated gene expression by combining individual-level genotype data or GWAS summary statistics with externally estimated eQTLs. These external eQTLs can come from reference panels such as those available from GTEx. The second stage assesses the association between imputed gene expression levels and a complex trait (Gusev et al. 2016; Li and Ritchie 2021).

TWAS have been applied in AD to try and disentangle the effect between genotype, transcript and disease status. A recent TWAS study in monocytes found an association between differences in gene expression and AD in seven genes in known AD risk loci, three of which (*PVR, PTK2B* and *MS4A6E*) were replicated (Harwood et al. 2021). Another study utilising the ROSMAP, MayoRNASeq and MSBB cohorts performed a TWAS across six different brain regions and found six candidate AD genes that were replicated: *APOC1*, *EED*, *CD2AP*, *CEACAM19*, *CLPTM1*, and *TREM2* (Gockley et al. 2021).

One benefit of TWAS is that the multiple hypothesis testing burden is much lower in comparison to GWAS. Another benefit is that analyses can be tissue specific meaning that it is possible to find associations specific to disease-relevant tissues (Li and Ritchie 2021).

Although TWAS is informative, it does have limitations. It is important to emphasise that TWAS is still a test of association, and the method does not confirm causality. Additionally TWAS-significant loci can contain multiple associated genes analogous to a GWAS identifying blocks of associated variants in LD (Wainberg et al. 2019). At present, TWAS studies only focus on gene expression that can be explained by common cis-eQTLs. This is in part due to the technical difficulties in assessing trans-

eQTLs. For trans-eQTLs the number of statistical tests is orders of magnitude higher than for cis-eQTLs leading to a huge multiple hypothesis testing burden. Therefore, increased sample sizes are required to ensure there is enough statistical power in any analysis (Li and Ritchie 2021). Furthermore, TWAS may be limited when there are overlapping effects between cis- and trans-eQTLs for the same gene (Li and Ritchie 2021) . Finally, the computational resources for calculating LD when including trans-eQTLs is prohibitive (Li and Ritchie 2021).

It has been estimated that common cis-eQTLs explain only 10% of genetic variance in gene expression (Grundberg et al. 2012). Trans-eQTLs on the other hand have been estimated to contribute around 70% of the genetic heritability of gene expression levels (Boyle et al. 2017). More reliable methods and more powerful datasets are required to detect rare cis-eQTLs and trans-eQTLs which are likely to explain more of the genetic basis of gene expression.

Some novel methods to extend the TWAS methodology have been suggested. A Bayesian genome-wide TWAS (BGW-TWAS) approach can gain insights from both cis- and trans-eQTLs. The authors performed BGW-TWAS in AD case-control data, neurofibrillary tangle density, global AD pathology and amyloid-beta. Using ROSMAP individual-level data they found *ZC3H12B* was a significant gene for the AD, global AD pathology and tangles phenotypes. *KCTD12* was significant for the amyloid-beta phenotype. Their data suggests that the association between *ZC3H12B* with all the included AD phenotypes was driven by trans-eQTLs, with the top four trans-eQTLs being located in the known risk gene *APOC1*. Using IGAP GWAS (Lambert et al. 2013) summary statistics they identified 13 significant genes including *HLA-DRB1*, *APOC1,* and *ZC3H12B* (Luningham et al. 2020).

1.1.13 Transcriptomics in AD

TWAS tests for association with genetically predicted expression using cis-eQTLs which only explains a very small part of total gene expression. Total expression also includes expression of rare cis-eQTLs, trans-eQTLs, the environment, and technical components (Wainberg et al. 2019). To further understanding of the molecular mechanisms underpinning the association between a risk gene, gene expression and AD, researchers have explored transcriptome profiling. This can be performed using microarray hybridisation or ribonucleic acid-sequencing (RNA-seq) technologies. Researchers then utilise these technologies with the aim of identifying genes that are differentially expressed between different phenotypes of AD, such as cases versus controls. Differentially expressed genes (DEGs) and their over- or under- expression may result in perturbations of biological pathways which could result in disease.

Differential gene expression (DGE) analysis is used to identify whether individual genes are expressed differently between phenotypes. In bioinformatic research it is also common to perform some form of pathway enrichment analysis as individual genes are not particularly informative as to underlying biology of disease. Sources of pathways include using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2008) and Reactome (Fabregat et al. 2016) databases.

Gene expression data is also often interpreted using gene ontology (GO) enrichment analysis. The Gene Ontology is a system for classifying sets of genes into terms. These GO terms fall under three categories: biological processes, molecular functions and cellular components. Differentially expressed genes are functionally annotated to biological processes, molecular functions or cellular components to discover pathways associated with disease. Identifying and modifying pathways associated with gene expression may be useful for disease diagnosis and discovering potential therapies (Bagyinszky et al. 2020). Tools for finding enriched GO terms include DAVID (Huang da et al. 2009b,a), GOrilla (Eden et al. 2009), and PANTHER (Mi et al. 2020) as well as many others.

The first step of an enrichment analysis typically involves defining a gene list of interest and depends on the type of analysis being performed. Ranked methods take as input a ranked gene list. Other methods take as input a gene list of interest, such as genes that are differentially expressed between two groups of samples from an RNA-seq experiment. The second step utilises a statistical method to identify pathways that are enriched in the gene list more than what one would expect by chance alone. Statistical methods do vary with non-ranked methods and require a definition of a background set, such as all genes included in the sequencing panel. A statistical test such as the Fisher's exact test based on the hypergeometric distribution is then used. More sophisticated methods do also exist as an alternative to non-ranked methods as they rely on arbitrary cut-offs. The alternative ranked methods include the entire gene list and often utilise statistical methods such as the Wilcoxon rank sum test. The third step is correction for multiple hypothesis testing correction and data visualisation (Reimand et al. 2019).

Transcriptomic studies in AD are often small due to the difficulty of collecting brain tissue samples. Therefore, gene expression studies in AD frequently utilise the AMP-AD gene expression data. This data is the most comprehensive publicly available expression data from AD brains and comprises data from the ROSMAP, MSBB and MayoRNASeq cohorts. One study downloaded microarray and RNA-seq data from the ROSMAP, MSBB and MayoRNASeq cohorts and analysed each dataset individually. They found 828 genes to be differentially expressed between AD cases and individuals with no cognitive impairment. The genes included *CD2AP*, *CD33* and *CR1*. A KEGG pathway analysis implicated the calcium signalling pathway and further analysis demonstrated the whole pathway barring 10 genes to be downregulated in AD in the ROSMAP RNA-seq data (Bihlmeyer et al. 2019).

Another study utilising AMP-AD RNA-seq data compared gene expression levels in different AD brain regions. The authors found alterations in gene expression are

highly prominent in samples obtained from the temporal lobe in comparison to the frontal lobe. Their interpretation for this was that the temporal lobe harbours the first brain regions to be affected by AD pathogenesis whereas the frontal lobe is affected only in advanced stages of AD. This study also demonstrated that the following AD genes were differentially expressed between AD cases and controls: *ABCA1*, *ABCA2*, *CALB1*, *C1R*, *C1S*, *GAD1/2*, *PVALB*, *REST*, *SLC32A1*, *SST*, *VGF*, and *VIP* (Marques-Coelho et al. 2021).

A study just using gene expression data from only the ROSMAP cohort performed differential gene expression analysis in blood and brain samples individually then together and stratified by *APOE* genotype. They found that established AD genes *INPP5D* and *HLA-DQA1* were differentially expressed in both blood and brain. The authors' findings also suggested that AD genes that are differentially expressed in both blood and brain were often associated with vascular markers, and their effects were also dependent on *APOE* genotype (Panitch et al. 2022). This study provides evidence of a potential benefit of evaluating differential gene expression data from different tissue types together. A limitation of this study is that the sample size of the study was small. Only 140 individuals had both blood and brain expression data and was further segregated by *APOE* genotype meaning each of the groups were very small.

Post-mortem bulk-cell transcriptomics show vast changes in gene expression throughout the brain, yet it is challenging to deduce which changes are driven by the cause of disease or are the result of disease. One recent article utilising Mendelian randomization provided evidence that differentially expressed genes between case and control status may reflect disease-induced changes in the transcriptome as opposed to disease-causing changes (Porcu et al. 2021). Analyses comparing gene expression may still have utility in pointing to useful biomarkers and further understanding of disease biology. It is important to consider this limitation when interpreting results as interventions that alter gene expression to normal levels may not necessarily be disease-modifying (Porcu et al. 2021).

1.2 Aims of this thesis

The aim of this thesis was to use a variety of bioinformatic approaches to identify potential genes involved with AD and identify potential biological mechanisms implicated in AD. This was achieved by initially integrating publicly available genetic and RNA-seq data with phenotypic data. Then by performing differential gene expression analysis and gene ontology enrichment analysis to identify disrupted pathways potentially implicated in AD. The next aim was to utilise eQTLs of AD differentially expressed genes to identify potential genes implicated in AD. The final results chapter aims to test if TWAS signals from three TWAS are enriched in differentially expressed genes for AD.

1.3 Outline of thesis

Chapter two gives a brief overview of the AMP-AD consortium and the three studies (ROSMAP, MSBB and MayoRNASeq) from which the data used in this thesis originates. This chapter also gives a brief overview of some of the computational methods used throughout this thesis.

Chapter three describes the work to produce a unified RNA-seq dataset from combining the three AMP-AD RNA-seq datasets together. The work followed an extensive quality control (QC) pipeline and used linear mixed-effect models (LMEM) in combination with principal component analysis to combine the three RNA-seq datasets into a larger, unified dataset. This chapter also describes the work to process and QC the accompanying genetic data and the work to define the phenotypic variables of interest, such as AD case-control status, Braak score and CERAD scores.

Chapter four investigates differentially expressed genes between AD cases and controls. Initially this is performed in only the ROSMAP data in order to see how the

LMEM and logistic regression method for determining differentially expressed genes performs against two frequently used DGE packages: Limma-voom and DESeq2. A DGE analysis was also performed on the combined AMP-AD RNA-seq data which was generated in chapter three for AD case-control, Braak and CERAD score phenotypes. A GO enrichment analysis was also performed to find potential pathways of biological interest.

In chapter five, a cis-eQTL analysis was performed to find associations between index SNPs from five AD GWAS and GWAX, and the AD case-control differentially expressed genes identified in chapter four. Additionally, a trans-eQTL analysis is performed to identify any associations between AD GWAS/GWAX index SNPs and the differentially expressed genes from chapter four.

Chapter six compares the results of the DGE analysis from chapter four to three existing AD TWAS results to identify if these two methods produce comparable results in AD research.

Chapter seven is the final chapter and is a discussion of the results of this thesis, its limitations, future directions and conclusions.

Chapter 2 – General Methods

The purpose of this chapter is to describe the origin of the data used in this thesis and some of the bioinformatic methodologies used. Finally, a summary of the programming languages and software used is given with their respective versions.

2.1 Cohort overview

2.1.1 Religious Orders Study and the Memory and Aging Project

The Religious Orders Study (ROS) and the Memory and Aging Project (MAP) are two cohorts that together are known as ROSMAP. Both ROS and MAP are longitudinal cohort studies of ageing and Alzheimer's disease from Rush University. Participants agreed to annual clinical evaluations and brain donation upon death (Bennett et al. 2012a; Bennett et al. 2012b).

Participants into ROS were recruited from religious communities such as nuns, priests, and brothers from all across the United States and all participants were without known dementia at enrolment (Bennett et al. 2012a). MAP was similar in study design to ROS but aimed to enrol participants with a much wider range of life experiences and socioeconomic status than ROS (Bennett et al. 2012b). For both ROS and MAP cohorts, upon death, dorsolateral prefrontal cortex samples were collected (Bennett et al. 2012a; Bennett et al. 2012b).

From the collected samples, the ROSMAP study has generated a large variety of omics data and made the data publicly available to researchers. This consists of whole genome sequencing, epigenomic, metabolomic, proteomic and transcriptomic data. Imaging and phenotype data has also been made available to allow for extensive analyses in dementia research.

2.1.2 Mount Sinai Brain Bank study

The Mount Sinai Brain Bank (MSBB) study is a cohort study that generated large-scale multi-omics data from AD, mild cognitive impaired (MCI) and AD control brains. Specifically, they generated whole genome sequencing, whole exome sequencing, transcriptome sequencing and proteome profiling data from multiple regions of the brain. Samples were taken from four areas of the brain: the parahippocampal gyrus (Brodmann area 36 – BM36); inferior frontal gyrus (BM44); superior temporal gyrus (BM22); and the frontal pole (BM10). Both raw and processed data were released and made publicly available in order to allow researchers to investigate the molecular underpinnings of AD (Wang et al. 2018).

2.1.3 MayoRNAseq study

The MayoRNAseq study has genomic, transcriptomic and proteomic data generated from cerebellum and temporal cortex samples from North American participants. In the original study, data was made publicly available to researchers to investigate transcriptional mechanisms contributing to neurodegenerative diseases.

Participants were recruited to the study if they had a neuropathological diagnosis of AD, progressive supranuclear palsy (PSP), pathological ageing (PA) in addition to elderly controls who were free of neurodegenerative diseases (Allen et al. 2016).

2.2 Data availability and the Accelerating Medicines Partnership for Alzheimer's Disease

2.2.1 Overview

The ROSMAP, MSBB and MayoRNAseq datasets are now hosted by the Accelerating Medicines Partnership for Alzheimer's Disease (AMP-AD). AMP-AD is a collaborative

program set-up between government, industry and non-profit organisations. The program launched in 2014 with two keys projects. The first was the Biomarkers in Clinical Trials project and the second was the Target Discovery and Preclinical Validation Project. This second project is the source of the data used throughout this thesis.

The aim of the Target Discovery and Preclinical Validation Project was to integrate the analyses of large-scale molecular data from human brain samples in order to discover potential drug targets. Genomic, epigenomic, transcriptomic and proteomic data were made available through the AD Knowledge Portal, a resource developed by Sage Bionetworks and hosted by Synapse. The data used throughout this thesis are the RNA-seq and genetic data which were made available on this platform and can be accessed through this website: https://adknowledgeportal.synapse.org/

2.2.2 RNA-sequencing data

Originally the ROSMAP, MayoRNAseq and MSBB studies were instigated separately and were completely independent of one another. Table 2-1 describes the sample generation, RNA extraction and processing each sample underwent which was performed by the original investigators.

As the aim of the Target Discovery and Preclinical Validation Project was to integrate data, the ROSMAP, MSBB and MayoRNAseq datasets underwent reprocessing as shown in Figure 2-1. This reprocessing was performed by researchers at the Mount Sinai Icahn School of Medicine for the AMP-AD consortium (Hodes and Buckholtz 2016).

I downloaded aligned bam files from the AMP-AD consortium using the Python *synapseclient* package via command line for quality control (QC) purposes (on 06

December 2019). Additionally, raw counts as generated by STAR were downloaded and used. In addition to this, phenotypic data were downloaded, and the variables included: sex, age at death, diagnosis at time of death, *APOE* carrier status, Braak score, CERAD score, race, RNA Integrity Number (RIN) and post-mortem interval (PMI). All data were downloaded from Synapse which hosts both unprocessed and reprocessed data for these studies

(https://www.synapse.org/#!Synapse:syn9702085). The reprocessed datasets for each study were downloaded and merged to form a single dataset. The QC process is further described in chapter 3.

	ROSMAP	**MSBB**	**MayoRNAseq**
Sample Source	Grey matter of the dorsolateral prefrontal cortex	Samples from Brodmann Areas 10, 22, 36 and 44	Cerebellum and Temporal Cortex
RNA Extraction	Qiagen miRNeasy	Qiagen RNeasy Lipid Tissue Mini Kit	Trizol and cleaned using Qiagen RNeasy with DNase treatment
Library Preparation	Strand specific dUTP method with poly-A selection followed by Illumina adapter ligation	Illumina TruSeq RNA Sample Preparation Kit: Non-strand specific method with poly-A selection followed by ligation with Illumina compatible adapters.	Illumina TruSeq RNA Sample Preparation Kit : Non-strand specific method with poly-A selection. No other details given.

Sequencing	Illumina HiSeq with 101bp paired-end reads with coverage of 150M for first 12 samples and remaining samples with coverage of 50M reads	Illumina HiSeq 2500 system with 100 nucleotide single end-reads with a coverage of 80M reads	Illumina HiSeq 2000 with 101bp paired end reads. Coverage not stated.
Quality Control of Reads	Trimming beginning and end bases of each read, identifying and trimming adapter sequences from reads, detecting and removing rRNA reads	Genes with at least 1 read count in at least 10 libraries were considered present, otherwise removed. Low RIN score (<4), or relatively large rRNA rate (>5%) were removed.	Low % mapped reads (<85%) or sex discrepant gene counts (Y chromosome gene expression) removed. Samples with a 3' bias excluded.
Aligning Reads to Genome	Non-gapped aligner Bowtie to gencode v14 in hg19 human genome reference (GRCh37)	Star aligner (v2.3.0e) to human genome hg19 (GRCh37)	SNAPR to GRCh38 reference and Ensembl v77
Estimation of Expression Levels	RSEM to produce FPKM	Quantified by featureCounts (v1.4.4)	SNAPR
Normalisation	Combat package to remove potential batch effect	TMM normalisation method using edgeR in R to calculate CPM	TMM normalisation method using edgeR in R to calculate CPM

Table 2-1 Sample processing methods used by original investigators in each of the three original studies (ROSMAP, MSBB and MayoRNASeq) to generate gene expression counts.

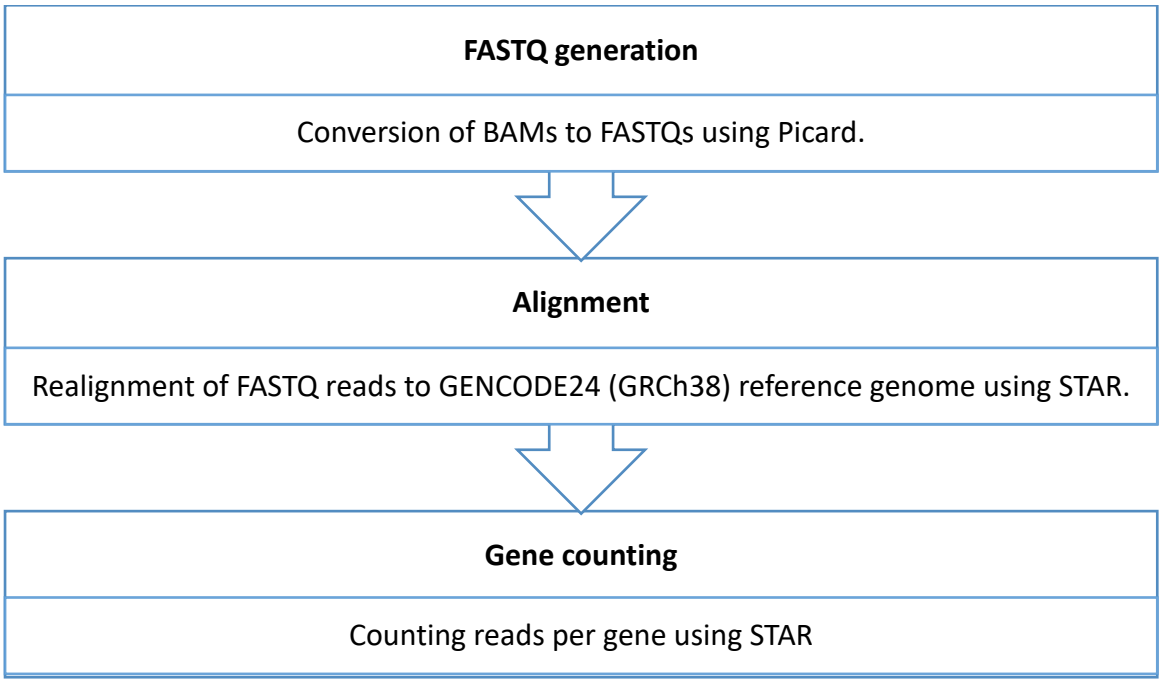


Figure 2-1 – The reprocessing strategy performed by the Mount Sinai Icahn School of Medicine for the AMP-AD consortium.

This reprocessing strategy was used on the three independent ROSMAP, MayoRNAseq and MSBB RNA-seq datasets. This was performed to reduce some of the technical variation in order for the datasets to be analysed together.

2.2.3 Genetic data

AMP-AD had performed whole genome sequencing (WGS) for the ROSMAP, MayoRNAseq and MSBB studies. Sequencing had been performed on an Illumina HiSeq X sequencer and data aligned to the GRCh37 (hg19) human reference.

VCF files for the use in this these were downloaded from the synapse website (https://www.synapse.org/#!Synapse:syn22264775). The QC process is further described in chapter 3.

2.3 Methodology

2.3.1 Linear mixed-effect models

Linear mixed-effect models build on linear models but allow the user to analyse data that show non-independence, are correlated, or show some form of hierarchical structure. The key component of linear mixed-effect models is that they allow the inclusion of both fixed and random effects. When considering RNA-seq data, certain variables are not independent, such as samples sequenced in the same sequencing batch in comparison to the rest of the samples from the study. The linear mixed-effect models that were used in this thesis were implemented using the *lme4* R package (Bates et al. 2015).

2.3.2 Limma-Voom

The limma R package was originally developed for differential expression analysis of microarray data. The voom function within the limma package modifies RNA-seq data for use with limma. Voom works by these steps:

1. Voom firstly transforms counts to log2 counts per million reads (CPM), where "per million reads" is defined based on the normalisation factors.
2. A linear model is fitted to the log2 CPM counts for each gene and residuals are then calculated.
3. A smoothed curve is fitted to the sqrt(residual standard deviation) by average expression
4. The smoothed curve is used to obtain weights for each gene and sample that are passed into limma along with the log2 CPMs.

Limma then fits models using weighted least squares for each gene. Limma-voom was implemented using the *limma* package in R (Law et al. 2014; Ritchie et al. 2015).

2.3.3 DESeq2

DESeq2 is an R package which allows the user to test for differential expression by use of negative binomial generalised linear models of the form:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$
$$\mu_{ij} = s_j q_{ij}$$
$$log2(q_{ij}) = x_j \beta_i \quad (1)$$

where counts K_{ij} for gene i, sample j are modelled using a negative binomial distribution with a fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean consists of a sample-size specific factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold change for gene i. The DESeq function in R performs three steps:

1. Estimation of size factors s_j
2. Estimation of dispersion α_i
3. Fitting of the negative binomial GLM for β_i and Wald statistics (Love et al. 2014).

2.3.4 MatrixEQTL

MatrixEQTL is an R package designed for computationally efficient eQTL analysis (Shabalin 2012). An analysis that used to take days or weeks can be performed in minutes. The fast performance is achieved by utilising large matrix operations when performing the most computationally intensive part of the algorithm. EQTL studies usually perform separate testing for each gene-SNP pair. The work presented in this thesis used a linear regression model with covariates for eQTL analysis and assumes that genotype only has an additive effect on gene expression. Each sample is encoded by 0, 1 or 2 corresponding to the number of minor alleles a SNP has. The

linear association between gene expression g and genotype s, considering covariate *x* is assumed to be:

$$g = \alpha + \gamma x + \beta s \quad (2)$$

Where α is the intercept of the linear model (the value of g when s = 0), γ is the slope of covariate x and β is the slope of genotype s.

For a faster computation, the equation can be reduced to testing of a simple linear regression model by orthogonalising g and s with respect to x through the following steps:

1. Centre variables g, x, and s to remove constant α from the model.
2. Orthogonalize g and s with respect to x:

$$\tilde{g} = g - \langle g, x \rangle x \, , \tilde{s} = s - \langle s, x \rangle x \quad (3)$$

3. Perform the analysis for the simple linear regression using one less degree of freedom for the test statistic to account for the removed covariate

$$\tilde{g} = \beta \tilde{s} + \varepsilon \quad (4)$$

2.4 Software, programming and data storage

2.4.1 Computing

All of the analyses presented in this thesis were performed either on an Apple iMac desktop or on the Supercomputing Wales' supercomputer based at Cardiff University called Hawk. Hawk is Linux-powered, and jobs were submitted to the slurm scheduler using Bash shell scripts.

2.4.2 R

R is a free software and programming language that is widely used for statistical programming and data visualisation (R Core Team 2021) . R version 3.6.2 was the default statistical tool used throughout this thesis unless otherwise stated.

2.4.3 Python

Python version 3.6.*5* was used for the execution of Go-Figure! and the use of the synapse command line client which was used to download WGS and RNA-seq count data and BAM files (Greenwood et al. 2020; Reijnders and Waterhouse 2021).

2.4.4 PLINK

PLINK is an open-source toolset for genome-wide association studies, research in population genetics and other analyses requiring the use of very large genetic datasets (Purcell et al. 2007; Chang et al. 2015). The genetic data utilised throughout this thesis was largely stored in the PLINK binary file format as -bed, -bim and -fam files. All analyses using PLINK were executed using PLINK version 1.9 which had been pre-installed on Hawk.

2.4.5 SAMtools

SAMtools is a widely used program for processing and analysing high-throughput sequencing data (Danecek et al. 2021). SAMtools version 1.9 was used for file processing such as sorting and indexing BAM files and had been pre-installed on Hawk.

2.4.6 Crossmap

The software Crossmap is a program for genomic coordinate conversion between different assemblies (Zhao et al. 2014). It was used to convert the WGS data and GWAS summary statistics from GRCh37 to GRCh38. Crossmap version 0.2.8 was used and installed using python onto Hawk.

2.4.7 RNASeQC

RNASeQC is a program that takes BAM files as input and provides RNA-seq quality metrics for use in QC pipelines (DeLuca et al. 2012). RNASeQC version 2.3.5 was used, and the program was obtained from: https://github.com/getzlab/rnaseqc/ and analyses were performed on Hawk.

2.4.8 VerifyBamID

VerifyBamID matches individual RNA-seq data with the data from genetic VCF files to identify sample mix-ups (Jun et al. 2012). This was used to confirm that the genetic data and RNA-seq data belonged to the same individuals and any samples with discrepancies were not included in the analysis. Version 1.1.3 of the software was obtained from:

https://github.com/statgen/verifyBamID/releases and implemented on Hawk using command line.

2.4.9 CATMAP

CATMAP is a program that was originally produced to functionally annotate microarray data (Breslin et al. 2004). It allows the user to perform a Wilcoxon rank sum test in order to perform Gene Ontology (GO) enrichment analysis. The software

and coding scripts to use CATMAP were provided by Dobril Ivanov (see contributions section).

2.4.10 False discovery rate

Throughout this thesis, in order to control for the false discovery rate, the Benjamini-Hochberg procedure was used. This was implemented in R using the function *p.adjust* (Benjamini and Hochberg 1995*)*. Unless otherwise stated, statistical significance was defined to be FDR p-value of less than 0.05 (FDR <0.05).

Chapter 3 – Quality control of RNA-seq data

3.1 Introduction

3.1.1 Public repositories and the Accelerating Medicines Partnership – Alzheimer's Disease (AMP-AD)

As DNA and RNA sequencing (DNA-seq and RNA-seq) technologies have advanced and reduced in cost, researchers are now able to generate vast amounts of data. Much of this data is now publicly available to computational researchers. This could drive forward new understanding of disease aetiology without having to spend time and money repeating data generation.

Some datasets are utilised often and are well known such as the UK Biobank (Sudlow et al. 2015). Many publicly available data sets are under-utilised in comparison. This is because interrogating existing data is not a straightforward enterprise. This is particularly true of RNA-seq data, as many of the individual studies are small in sample size. As shown by GWAS, larger datasets are imperative for finding novel gene associations. To enlarge RNA-seq datasets, combining datasets to create a larger dataset is an option - but this is not a straightforward process. Figure 3-1 demonstrates a simple outline of the processing RNA-seq data often undergoes. Each stage can differ in the method used, which is a problem as RNA-seq data are very sensitive to batch effects. Batch effects are sources of unwanted variation that could be due to technical artefacts or differences between the samples such as sequencing method or technician. Failure to remove this unwanted variability or combining data without proper correction could lead to spurious findings (Peixoto et al. 2015). Depending on some of the methods used, attempting to combine datasets could be inappropriate without reprocessing. For example, datasets that have aligned their

reads to different reference genomes would require a lot of time-consuming and resource intensive reprocessing.

The Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) knowledge portal hosts three independent RNA-seq datasets. These are the Religious Orders Study and Memory and Aging Project (ROSMAP), Mount Sinai Brain Bank (MSBB) study, and the MayoRNAseq study. An overview of these studies can be found in the previous chapter. Originally these three data sources were generated independently. Recently these three datasets have been reprocessed by researchers at the Mount Sinai Icahn School of Medicine to create a uniformly processed RNA-seq dataset. In addition to these three datasets, AMP-AD hosts phenotypic information such as Braak scores (measures severity of neurofibrillary changes such as neurofibrillary tangles) and CERAD (Consortium to Establish a Registry for Alzheimer's Disease) scores (a semiquantitative measure of neuritic plaques) allowing for a diverse range of analyses to be performed to potentially further our understanding of AD aetiology (Braak et al. 2006; Fillenbaum et al. 2008).

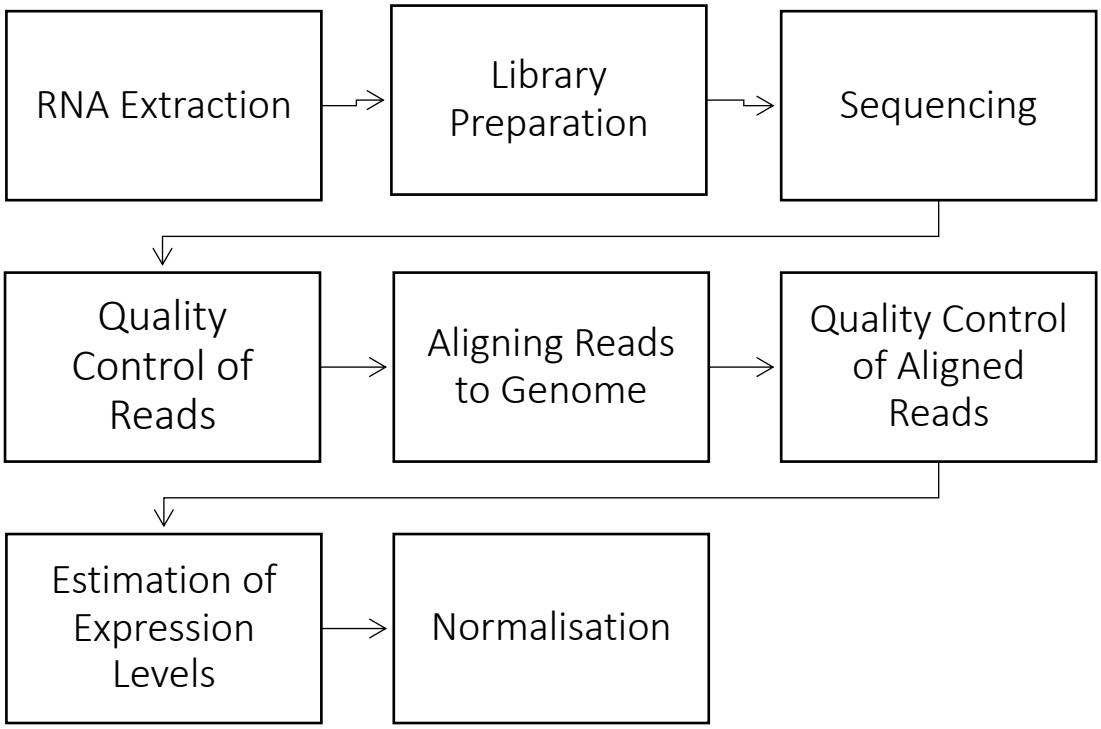


Figure 3-1 – A simple outline of the processing workflow of a typical RNA-seq experiment.

3.1.2 Aims

The over-arching aim of this chapter was to produce a single, cohesive dataset by taking advantage of the reprocessed files available from the AMP-AD consortium. This data consisted of phenotypic, RNA-seq and genetic data from three different studies.

The first aim of this work was to define the phenotypic variables of interest with the data that were available.

The second aim was to quality control (QC), pre-process and normalise the RNA-seq data.

The third aim was to investigate the presence of unwanted variation in the three individual RNA-seq datasets.

The final aim of this analysis was to combine the RNA-seq data available from the three studies. Linear mixed-effect models (LMEM) in combination with principal component analysis (PCA) were used to overcome batch effects and hidden confounders when combining RNA-seq data. The outcome was the production of a larger dataset with increased power where the normalised residuals were saved for use in downstream analyses such as differential gene expression analysis.

3.2 Methods

3.2.1 Overview of the methods and steps taken to QC and produce the single dataset

For each of the three studies, the original investigators performed different QC measures. This meant that samples from each study had been processed differently.

In addition to this, QC details on the reprocessed data have been sparse. To overcome the lack of clarity surrounding the QC of this data, an extensive QC procedure was executed in this thesis on both the reprocessed aligned BAM files and the counts. This QC process consisted of using multiple existing software and R packages including VerfiyBamID, RNASeQC, and conditional-quantile normalisation (CQN) (DeLuca et al. 2012; Hansen et al. 2012; Jun et al. 2012). Figure 3-2 provides an overview of the process. The resulting dataset from these QC processes includes a total of 627 individuals (379 AD cases vs 248 AD controls) with 930 samples from six different cortical brain tissues.

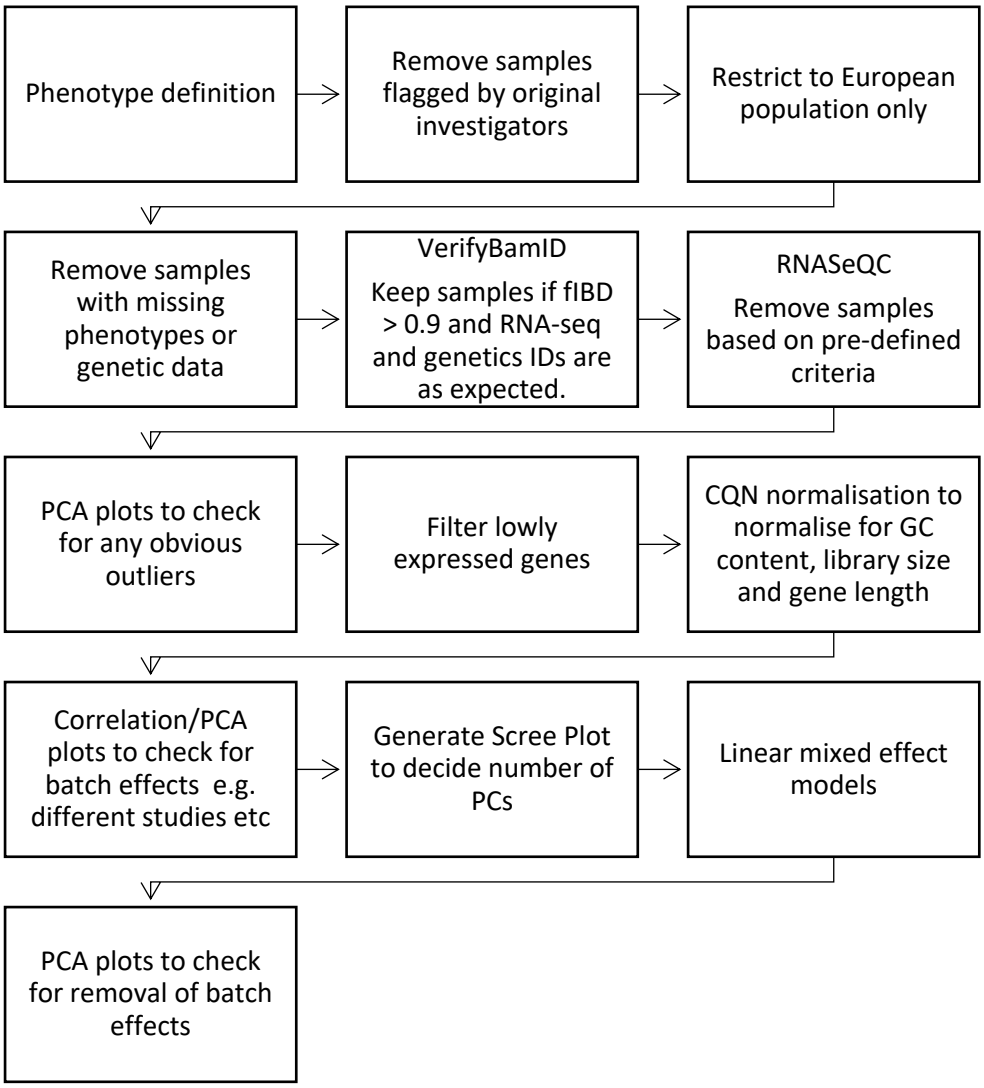


Figure 3-2 – An overview of the steps taken to produce a combined RNA-seq dataset from the three reprocessed ROSMAP, MSBB and MayoRNASeq datasets. fIBD = identity by descent ; RNA-seq = Ribonucleic acid sequencing; PCA = principal component analysis; CQN = conditional quantile normalization; GC = guanine-cytosine; PC = principal components

3.2.2 Disease and variable definition

3.2.2.1 AD diagnosis

Of the three studies, only the MayoRNAseq study had an AD case or control diagnosis. Therefore, defining AD case or control status for ROSMAP and MSBB had to be based on the variables that were available.

For ROSMAP the final consensus diagnosis was provided in the available metadata. It was determined by a neurologist with expertise in dementia who reviewed all available clinical data to determine which category the individual fell into. Every individual was given a value of 1-6 which is summarised in Table 3-1 (Schneider et al. 2007). In this study, individuals with a diagnosis of 1 were considered controls and those with a score of 4 or 5 were considered AD cases.

Value	Coding
1	NCI: No cognitive impairment (no impaired domains)
2	MCI: (one impaired domain and no other cause of CI
3	MCI: One impaired domain and no other cause of CI
4	AD: AD and no other cause of CI (NINCDS Probable AD)
5	AD: AD and another cause of CI (NINCDS Possible AD)
6	Other dementia: Other primary cause of dementia

Table 3-1 ROSMAP data final consensus diagnosis

Table amended from https://www.radc.rush.edu/docs/var/detail.htm?category=Clinical+Diagnosis&subcategory=Final+consensus+diagnosis&variable=cogdx and Schneider et al.(Schneider et al. 2007). (AD: Alzheimer's disease; CI: cognitive impairment; MCI: mild cognitive impairment; NCI: no cognitive impairment; NINCDS: National Institute of Neurological and Communicative Diseases and Stroke)

For the MSBB data, AD case-control status was based on the clinical dementia rating (CDR) and CERAD score. The CDR is derived from a semi-structured interview and rates impairment in six cognitive categories. These are: memory, orientation, judgement and problem solving, community affairs, home and hobbies and personal care. CDR was originally based on a five-point scale (0, 0.5, 1, 2, 3). The MSBB study used a validated extended version to account for the fact a significant proportion of individuals included in the study were living in care homes at the time of interview. Therefore, the investigators introduced additional scores of 4 and 5 (Morris 1993; Wang et al. 2018). I describe CERAD in more depth in the next section and an overview of all definitions for cases and controls for each study is given in Table 3-2. The final case-control definitions for the MSBB data used in this work were determined through discussions between me, DI, GL, AM and the AD field team (see contributions section of thesis). A final decision was made to define AD case based on a CDR rating of 2 or greater (capturing moderate, severe, profound and terminal dementia) and a CERAD score of 2,3,4 (capturing possible AD, probably AD and definite AD). This decision was made on the basis that it captured both clinical and pathological information.

Study	Variable used	Final case and control definition
ROSMAP	Final consensus diagnosis 1-NCI, 4 – AD-no other cause, 5-AD with another CI	1-Control 4,5- Cases
MayoRNAseq	AD, PSP, Pathological Ageing, Control	Control – control AD – AD case
MSBB	CDR: 0 - no cognitive deficits, 0.5 – questionable dementia 1 – mild dementia, 2 – moderate dementia, 3 – severe dementia, 4 – profound dementia, 5 - terminal dementia CERAD: 1, 2, 3, 4	Control – CERAD 1 AND CDR 0 or 0.5 CERAD 2, 3, 4 and CDR 2 or greater

Table 3-2 A summary of the case and control definitions used for the three datasets

(AD: Alzheimer's disease; CDR: clinical dementia rating; CERAD: Consortium to Establish a Registry for Alzheimer's Disease; CI: cognitive impairment; PSP: Progressive supranuclear palsy)

The original data files from the ROSMAP and MSBB studies had coded CERAD scores differently, for example the MSBB study had coded definite AD as a 2 whereas this was a 1 in the ROSMAP study. In this analysis I recoded them so that the score was consistent between studies. I have summarised this in Table 3-3 and the coding used in this analysis is shown in the harmonised CERAD scoring column. To demonstrate this point, an individual in the MSBB study with an original score of 2 (definite AD) would be recoded as a 4 (and the stage definition remains as definite AD as per the harmonised CERAD scoring in Table 3-3). In contrast, an individual from the ROSMAP study with an original score of 4 (No AD) would be recoded as a 1 (Normal/No AD).

Original MSBB CERAD scoring		Original ROSMAP CERAD scoring		Harmonised CERAD scoring used in this thesis	
Normal:	1	No AD:	4	Normal/No AD:	1
Possible AD:	4	Possible AD:	3	Possible AD:	2
Probable AD:	3	Probable AD:	2	Probable AD:	3
Definite AD:	2	Definite AD:	1	Definite AD:	4

Table 3-3 – A summary of how CERAD score was initially coded in both the MSBB and ROSMAP studies. The harmonised scoring shows how each of the CERAD scores were recoded in order to harmonise the variable across studies for use in this thesis. Only the label was changed, the CERAD stage of normal, possible AD, probable AD and definite AD remained the same**.**

3.2.2.2 Braak and CERAD phenotypes

Three phenotypes of interest were chosen for the analyses presented in this thesis: Braak stage, CERAD neuropathological stage and AD case-control status. Braak stage or score is a measure of the presence of hyperphosphorylated tau protein which is central to the AD process. The deposition of hyperphosphorylated tau occurs at predisposed cortical and subcortical sites in a predictable manner. This allows the deposition to be characterised by six stages of pathology (I – VI). The six stages are often further grouped into four further units: 0, I-II, III-IV, V-VI, with 0 being the absence of any hyperphosphorylated tau. Elevated tau is often seen in tandem with increased amyloid pathology and deterioration of cognition as Braak stages increase (Braak et al. 2006; Lowe et al. 2018). In the literature each Braak stage is referred to using either Roman or Arabic numerals. Throughout this thesis I will refer to them using Arabic numerals (0,1,2,3,4,5 and 6).

CERAD categories or scores are the evaluation of neuropathology post-mortem. They assess the frequency of neuritic and diffuse plaques in a semi-quantitative manner. This age-related plaque categorisation indicates the level of certainty of the diagnosis of AD. The categories range from definite, probable, possible or no evidence of AD (and are often indicated by a number running 1 to 4 or 4 to 1). CERAD provides a validated measure that permits comparison across studies and settings (Mirra 1997; Fillenbaum et al. 2008). I chose CERAD scores as a phenotype in this analysis in addition to Braak score. The reasoning is that although the two have a positive correlation, it is not yet clear if these two occur through separate mechanisms.

The three studies provided different phenotypic information. The MayoRNAseq study data had the least available to download. Individuals had a diagnosis of 'AD', 'Progressive Supranuclear Palsy', 'Pathological ageing' or 'control'. I excluded any individual with a diagnosis other than ‘AD’ or ‘control’ from the analysis. All individuals with a label of ‘AD’ had a definite diagnosis according to the National Institute of Neurological and Communicative Diseases and Stroke – Alzheimer’s Disease and Related Disorders Association (NINCDS-ADRDA) criteria and had a Braak stage of 4 or greater. Individuals identified as being ‘controls’ had a Braak score of 3 or less, CERAD neuritic and cortical plaque densities of none or sparse and lacked a pathological diagnosis of AD. They also lacked a pathological diagnosis of 11 other neurodegenerative disorders including Parkinson’s disease, motor neuron disease (MND), Huntington’s disease (HD), and dementia lacking distinctive histology (DLDH).

As Braak scores were unavailable for some MayoRNASeq controls, a “neutral” Braak score of 3 was assigned to those controls to maintain sample size. Some individuals in the MayoRNAseq study had Braak scores of 2.5, 4.5 and 5.5 (totalling 4, 3, and 4 samples respectively). After initial investigation of the MayoRNAseq data, these were recoded to 2, 4 and 5 respectively. The decision to recode was made to harmonise the MayoRNAseq data with the ROSMAP and MSBB data which did not use half Braak scores. The decision to round down the Braak score was made as the pathology

indicated that the lower score pathology was present but not quite enough pathology was present to be characterised as the higher Braak score.

Individual-level data for CERAD score was unavailable for all MayoRNAseq samples. Samples originating from this study were excluded from any analysis utilising the CERAD score. For individuals from the ROSMAP and MSBB studies, I only included samples in the analysis if information on both Braak score and CERAD score were available as no other information was given for inference. Table 3-4 summarises this information for clarity.

Braak Measures		CERAD measures	
ROSMAP & MSBB	MayoRNAseq	ROSMAP & MSBB	MayoRNAseq
Scores 0-6 available for all individuals	AD cases had Braak scores available. AD controls without individual level data were given a score of 3 as individual level data for some controls were unavailable	Scores 1-4 available for all individuals	Individual level data unavailable so excluded from analysis

Table 3-4 A summary of Braak and CERAD measures available in each study.

Arabic numerals have been used instead of roman numerals for the purposes of clarity.

Some variables were recoded to ensure harmony between the studies. Originally in the supplied metadata, individuals aged above 90 at death were coded as 90+. The age was recorded this way by the owners of the original data to comply with data privacy regulations. I changed those individuals with a recorded age at death of 90+ to 90 so that this variable was numerical rather than categorical. I also changed post-mortem interval (PMI) from minutes to hours, and I assigned a CERAD score of 1, 2, 3 or 4 to indicate no AD, possible AD, probable AD and definite AD respectively.

3.2.3 Tissues

The ROSMAP, MSBB and MayoRNAseq generated gene expression data from samples across the cerebral cortex and the cerebellum. The original ROSMAP RNA-seq investigators collected samples from the dorsolateral prefrontal cortex (DLPFC) (Wang et al. 2013; Bennett et al. 2018). The MSBB study included samples from four brain regions located in the cerebral cortex: Brodmann areas 10, 22, 36 and 44 (referred to as BM10, BM22, BM36 and BM44 respectively) (Wang et al. 2018). The MayoRNAseq study took samples from the temporal cortex (TCX) and the cerebellum (Allen et al. 2016). Cerebellum samples were excluded from this analysis as the cerebellum's role in AD is still debated. In addition, gene expression of the cerebellum in comparison to cortex brain samples is vastly different (Chappell et al. 2018). Thus, this analysis explored the gene expression of cortical samples only.

Anatomist Korbinian Brodmann defined regions of the cerebral cortex based on subtle differences in cortical structure. These regions are known as Brodmann areas and remain one of the most widely used systems for identifying regions of the brain (Johns 2014). Figure 3-3 shows where the cerebral cortex samples from the three studies are located in terms of Brodmann areas.

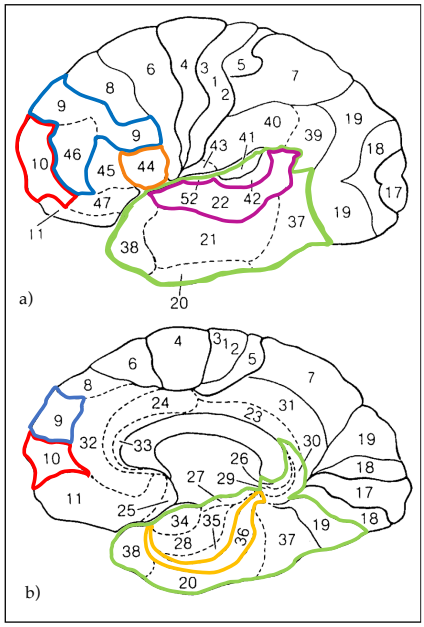


Figure 3-3 – A schematic of the brain regions sampled in the AMP-AD data.

The red, magenta, yellow and orange (Brodmann areas 10, 22, 36 and 44 respectively) represent areas MSBB samples originated from; the blue area represents the dorsolateral prefrontal cortex from which the ROSMAP samples were sampled from; the green area represents the temporal cortex from with the MayoRNAseq samples were sampled from. Figure 3a is the sagittal view of the brain and figure 3b is the mid-sagittal view. Image amended from (Cabeza and Nyberg 2000).

3.2.4 Initial sample exclusion

Samples were excluded if they had been flagged by the original investigators for exclusion in any future analyses. This could have been for reasons such as sex mismatch, sample ID duplication or discordance between RNA-seq and genetic data or due to the sample being of a low quality.

3.2.5 European ancestry

VCF files containing WGS data were downloaded from the AMP-AD website. SNPs were then excluded if their minor allele frequency (MAF) < 0.01; they deviated from Hardy-Weinberg equilibrium at $p \leq 1x10^{-06}$; or had a missingness of greater than 2%. Individuals were excluded if they had increased or decreased heterozygosity of $|F| > 0.1$; high pairwise relatedness (pi-hat > 0.22); or were population outliers as

identified by principal component analysis (PCA) in a joint analysis of 2,000 subjects from three different populations taken from the 1000 Genomes project. The three populations used were EUR, ASN, AFR ((http://www.1000genomes.org/) (Auton et al. 2015). This resulted in 8,717,089 SNPs and 1772 genetic samples. Not all genetic samples had corresponding RNA-seq data. VCF files were downloaded by me, but this QC analysis was performed by Ganna Leonenko (see contributions section of thesis).

3.2.6 Samples excluded for missing phenotype data

Samples were excluded if they lacked the data to be defined as a case or a control according to Table 3-2. Additional samples were excluded if they had missing phenotypic data. This included post-mortem interval (PMI), RNA integrity number (RIN), age at death and sex, and for MSBB and ROSMAP studies Braak score and CERAD score.

3.2.7 VerifyBamID

The software VerifyBamID was implemented to confirm that the genetic data and gene expression data were from the same individual (Jun et al. 2012). VerifyBamID checks that the reads in a BAM file for an individual against the polymorphisms for each genotype for all individuals in the sample. It also reports whether the reads are contaminated as a mixture of two samples or if a sample swap has occurred.

The aim of using VerifyBamID was to detect mislabelled samples and/or sample swaps rather than to detect the presence of DNA contamination. As a result, a genetic contamination level of up to 10% was deemed an acceptable contamination cut-off. Samples which were identified as being potential sample swaps were excluded from the analysis as it would not be possible to tell which was the true sample. This amounted to two individuals being excluded from the MayoRNASeq cohort, 18 from the ROSMAP cohort and 1 from the MSBB cohort.

VerifyBamID accepts VCF and BAM files as input for the analysis. As the genetic data are build 37 but the RNA-seq data are build 38, the software Crossmap was used to first convert the genetic data from build 37 to build 38 (Zhao et al. 2014).

Version 1.1.3 of the software was obtained from https://github.com/statgen/verifyBamID/releases and implemented on a Linux server using command line. Only genetic samples with a corresponding RNA-seq sample were retained for future analyses.

3.2.8 RNASeQC

The original investigators for each of the three studies performed different QC measures. In addition to this, QC details on the reprocessed data were sparse. In order to overcome this and to ensure more homogenous data, QC was performed on the reprocessed aligned BAM files. Initially, samples were excluded if they had been identified by the original investigators for previous QC failure. Some MSBB samples had been sequenced twice and recorded as two samples so the sample with the most mapped reads was retained for analysis. The remaining aligned BAM files were then sorted and indexed using Samtools (v. 1.9) and then processed using RNASeQC (v. 2.3.5).

RNASeQC is a program which computes a series of quality control metrics for RNA-seq data (DeLuca et al. 2012). Table 3-5 describes the metrics used for QC purposes and pass/fail criteria. Typically for the samples, the mean metric was calculated, and samples were excluded if they fell above or below four standard deviations (SD) from the mean, which was calculated using R (v3.6.1.).

QC measure as defined by the software	Description	Exclusion Criteria
Mapping rate	Proportion of reads in the BAM that were mapped	Any sample that is 4 SD from the mean in the lower tail of distribution
Duplicate rate of mapped	The proportion of all reads that were marked as PCR/Optical Duplicates out of all “Mapped Reads”	Any sample that is 4 SD from the mean in the upper tail of distribution
Duplicate rate of mapped excluding *Globin* genes	Similar to duplicate rate of mapped but excludes reads that did not align to HBA1, HBA2, HBB or HBD	Any sample that is 4 SD from the mean in the upper tail of distribution
Expression profiling efficiency	The proportion of “Exonic Reads” out of all reads which were not secondary alignments or platform/vendor QC failing reads	Any sample that is 4 SD from the mean in the lower tail of distribution
High quality rate	The proportion of properly paired reads with less than 6 mismatched bases and a perfect mapping quality	Any sample that is 4 SD from the mean in the lower tail of distribution
Exonic rate	The proportion of mapped reads for which all aligned segments unambiguously aligned to exons of the same gene	Any sample that is 4 SD from the mean in the lower tail of distribution

Intronic rate	The proportion of mapped reads for which all aligned segments unambiguously aligned to the same gene but none of which intersected any exons.	Any sample that is 4 SD from the mean in the upper tail of distribution
Intergenic rate	The proportion of mapped reads for which none of the aligned segments intersected any genes.	Any sample that is 4 SD from the mean in the upper tail of distribution
Ambiguous alignment rate	The proportion of mapped reads where the aligned segments unambiguously aligned to exons of more than one gene	Any sample that is 4 SD from the mean in the upper tail of distribution
High quality exonic rate	The proportion of exonic reads out of high quality reads as defined in high quality rate	Any sample that is 4 SD from the mean in the lower tail of distribution
High quality intronic rate	The proportion of intronic reads out of high quality reads as defined in high quality rate	Any sample that is 4 SD from the mean in the upper tail of distribution
High quality intergenic rate	The proportion of intergenic reads out of high quality reads as defined in high quality rate	Any sample that is 4 SD from the mean in the upper tail of distribution
High quality ambiguous alignment rate	The proportion of ambiguous alignment reads out of high quality reads as defined in high quality rate	Any sample that is 4 SD from the mean in the upper tail of distribution

rRNA rate	The proportion of mapped reads which at least partially intersected with an annotated rRNA gene	Any sample that is 4 SD from the mean in the upper tail of distribution
End 1 sense rate	The proportion of first mate reads which intersected with a sense strand feature out of all first or second mate reads which intersected with any features respectively.	Any sample that is 4 SD from the mean in the upper tail of distribution
End 2 sense rate	The proportion of second mate reads which intersected with a sense strand feature out of all first or second mate reads which intersected with any features respectively	Any sample that is 4 SD from the mean in the lower tail of distribution
Genes detected	The number of genes which had at least 5 unambiguous reads.	Less than 15,000 or greater than 30,000
Median 3’ bias	These aggregate statistics are based on the total coverage in 100 bp windows on both the 3' and 5' ends of a gene. The windows are both offset 150 bases into the gene. This computation is only performed on genes at least 600bp long and with at least 5 unambiguous reads. A gene with even coverage in both its 3' and 5' windows would have a bias of 0.5; bias near 1 or 0 may indicate degradation	Any sample that is 4 SD from the mean in either the lower or upper tail of distribution

3’ bias std	As Median 3’ bias	Any sample that is 4 SD from the mean in either the lower or upper tail of distribution
3’ bias MAD std	As Median 3’ bias	Any sample that is 4 SD from the mean in either the lower or upper tail of distribution
3’ bias 25th percentile	As Median 3’ bias	Any sample that is 4 SD from the mean in either the lower or upper tail of distribution
3’ bias 75th percentile	As Median 3’ bias	Any sample that is 4 SD from the mean in either the lower or upper tail of distribution
Median transcript coverage coefficient of variation	The statistics are the median of a given aggregate statistic of transcript coverage. Transcript coverage is computed by dropping the first and last 500bp of each gene and measuring the high-quality coverage over the remainder of the gene.	Any sample that is 4 SD from the mean in the upper tail of distribution
Median exon coefficient of variation	The median coefficient of variation of exon coverage. Exon coverage is computed by dropping the first and last 500bp of each gene and measuring	Any sample that is 4 SD from the mean the upper tail of distribution

End 1 antisense rate *	Number of End 1 reads that were sequenced in the antisense direction divided by total mapped reads	Any sample that is 4 SD from the mean in either the lower tail of distribution
End 2 antisense rate *	Number of End 2 reads that were sequenced in the antisense direction divided by total mapped reads	Any sample that is 4 SD from the mean in either the upper tail of distribution
Low mapping quality rate *	Number of low mapping quality reads divided by total mapped reads	Any sample that is 4 SD from the mean in the upper tail of distribution
Non globin reads rate *	Number of reads excluding reads which aligned to Globin genes divided by total mapped reads	Less than 0.9
Unique mapping vendor QC passed reads rate #	The count reads without the secondary or QC fail flags set. For a true count of total alignments use total reads divided by total mapped reads	Less than 0.5

Table 3-5 – RNASeQC quality control (QC) measures used to QC ROSMAP, MSBB and MayoRNASeq RNA-seq data

*All QC measures were as generated by RNASeQC apart from metrics with * which were calculated by dividing the metric by total mapped reads to produce rate, and # which were calculated by dividing the metric by total reads to produce rate.*

3.2.9 Read count filtering and normalisation

Raw RNA-seq counts were converted to counts per million (CPM) using EdgeR (Robinson et al. 2010) and the function filterByExpr was used to filter out lowly expressed genes. The EDASeq package (Risso et al. 2011) was used to determine gene length and GC content. Conditional Quantile Normalization (CQN) (Hansen et al. 2012) was performed on the raw counts and used to normalise for gene length, guanine-cytosine (GC) content and library size using the CQN package. This was performed on counts for each study individually, then data sets were combined only keeping genes in common between the three datasets. This resulted in 16,485 genes in common between the three datasets. This was all performed in R version 3.6.1.

3.2.10 PCA plots and scree plots

The data was further explored via PCA and plotting principal component (PC) biplots to investigate batch effects. Additionally, scree plots were generated to determine how many principal components to include in the LMEM analysis.

3.2.11 Linear mixed-effect models

Linear mixed-effect models (LMEMs) were implemented to correct for batch effects and hidden confounders when combining the three datasets together. This was achieved using the lmer function in the lme4 package in R. Equation 1 is the LMEM for the combined AMP-AD data:

$$GE \sim Sex + Age + PC1 + PC2 + PC3 + (1|Batch) + (1|ID) \qquad (1)$$

where GE is gene expression and the response variable, age is age at death in years, and PC1, PC2 and PC3 are the first three multivariate principal components. All these

variables are fixed effects. Batch is sequencing batch and ID is individual ID to reflect that some individuals in this data have multiple samples. Sequencing batch and individual ID have been included in the model as random effects which are indicated by the bar symbol (|). The residuals of this regression were then saved for use in further analyses in future chapters.

3.2.12 Checking for batch effects

To determine the effects of LMEM on the normalisation of batch effects, principal component biplots were inspected. The purpose of this was to identify any obvious remaining batch effects or sources of unwanted variation and that they had been removed from the gene expression data to allow it to be used in further analyses.

3.3 Results

3.3.1 Sample demographics

The number of samples remaining after each stage of QC can be seen in Table 3-6. The MayoRNAseq and ROSMAP studies only included one sample per individual whereas the MSBB had multiple samples from different tissues for some individuals which is reflected in Table 3-6. The total number of individuals and samples retained for analysis and their demographics can be seen in Table 3-7.

	MayoRNAseq	ROSMAP	MSBB (Individuals/Samples)
Initial sample number	278	632	315/1277
Samples after removal based on diagnosis	160	452	239/969
Samples after removal as flagged by original investigators or duplicates	150	409	219/797
Samples after removal as no WGS/RNA-seq sample or European ancestry	147	395	170/503
Samples after removal due to missing phenotype data	106	394	170/503
Samples after removal due to VerifyBamID	104	376	169/497
Samples after removal due to RNASeQC	90	369	168/471

Table 3-6 - Number of samples and individuals remaining after each stage of the QC process for the MayoRNAseq, ROSMAP and MSBB studies.

	MayoRNAseq	ROSMAP	MSBB	TOTAL
Individuals in cohort	90	369	168	627
Sex	F: 50 (55.6%) M: 40 (44.4%)	F: 242 (65.6%) M: 127 (34.4%)	F: 111 (66.1%) M: 57 (33.9%)	F: 403 (64.3%) M: 224 (35.7%)
Age at death (years)	Mean: 82.7 SD: 7.6	Mean: 86.4 SD: 4.9	Mean: 83.8 SD: 7.3	Mean: 85.1 SD: 6.2
Diagnosis	AD: 42 (61.9% F) Control: 35 (68.6% F)	AD: 204 (69.1% F) Control: 165 (61.2 % F)	AD: 133 (67.7% F) Control: 35 (60.0% F)	AD: 379 (613 samples) Control: 248 (317 samples)
Tissue	TCX: 90	DLFRC: 369	BM10: 135 BM22: 105 BM36: 113 BM44: 118	Total individuals: 627 Total samples: 930

Table 3-7 – Sample demographics for the MayoRNAseq, ROSMAP, and MSBB QCed datasets and their combined totals

3.3.2 Initial investigation of MayoRNAseq dataset

Initially boxplots were plotted for PMI, RIN, Braak score and age at death by diagnosis as per Figure 3-4. Boxplots were also plotted for age at death, RIN and PMI by Braak score (Figure 3-5). A correlation analysis was performed to understand the relationship between covariates (PMI, RIN, Braak score and age at death) as seen in Figure 3-6 for the MayoRNAseq dataset. RIN differed between case and control status and correlated significantly with Braak score albeit weakly (r=0.04, p-value =1.58×10^{-05}). PMI also correlated with age at death (r=0.32; p-

value =1.58x10^{-05}). The reason for this correlation cannot be determined through the available data, however it has previously been shown that PMI can be influenced by a range of factors. These include grieving time required by families, arrangement of tissue recovery by the brain bank and time required for legal processes to occur (White et al. 2018). PMI and Braak score were not significantly correlated (r = -0.05; p-value = 0.58). CERAD score was not included in this analysis as individual-level data was unavailable for the participants in the MayoRNAseq study.

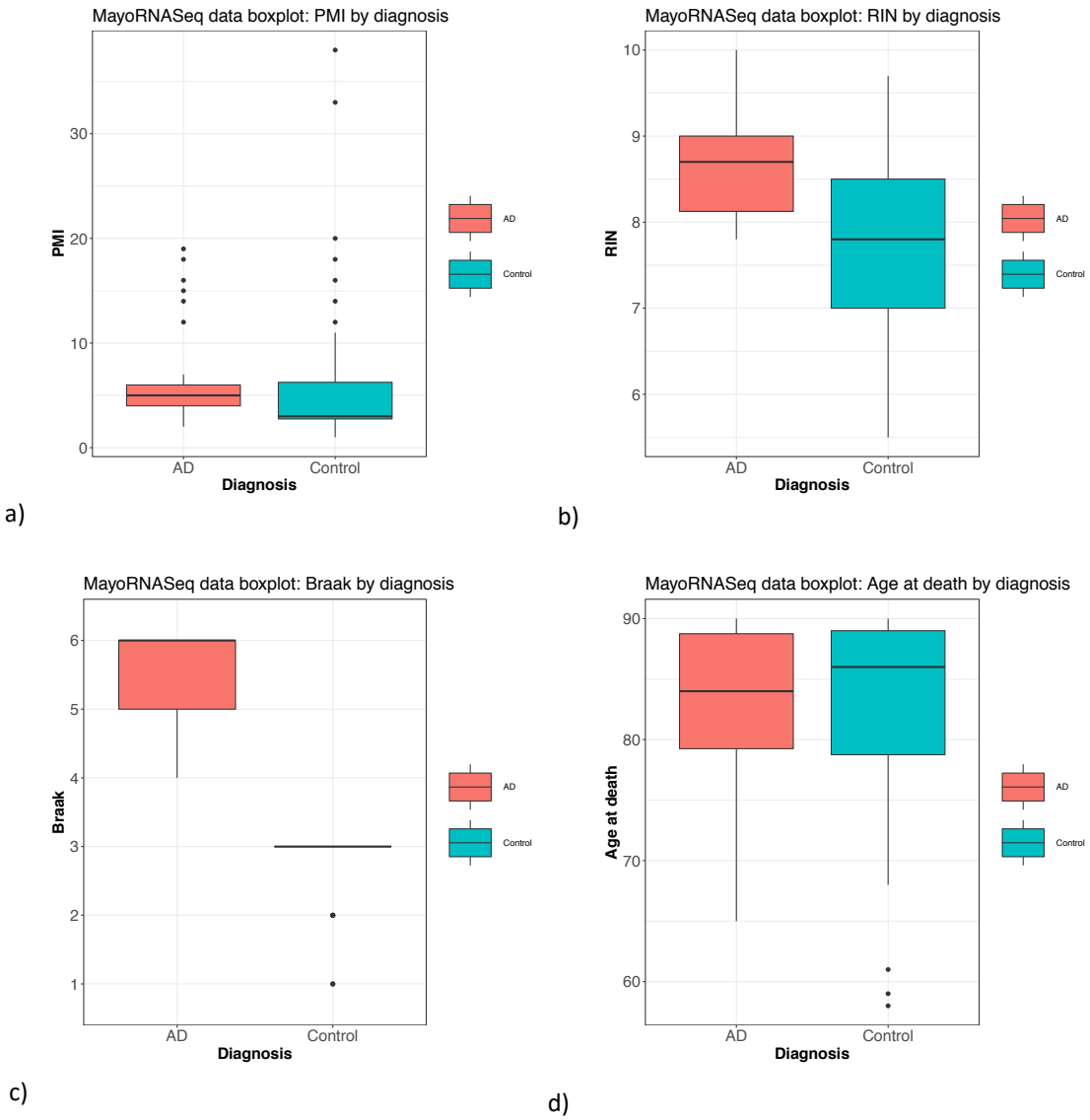
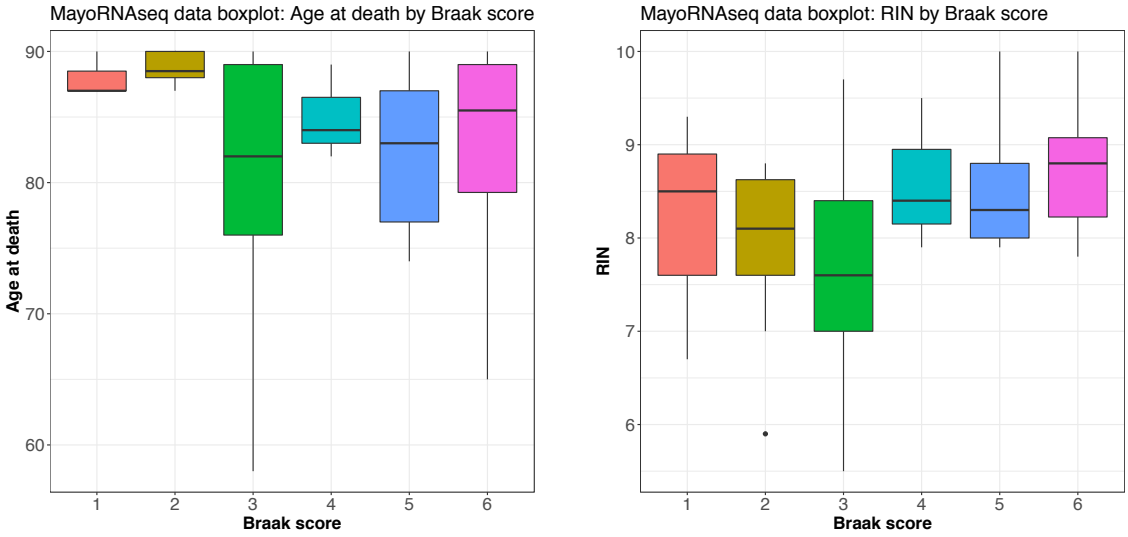
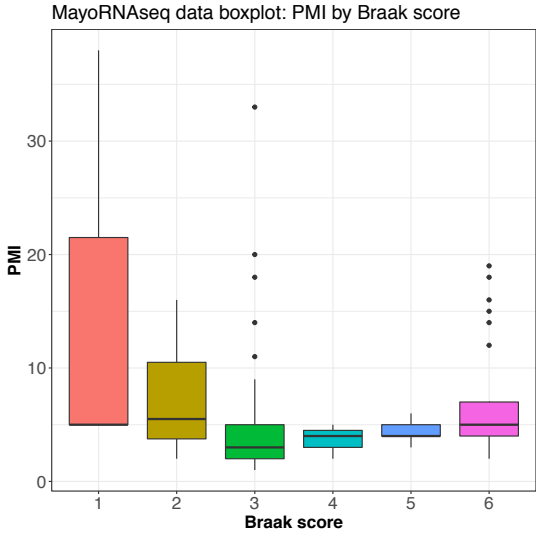


Figure 3-4 – Boxplots for a) post-mortem interval (PMI) in hours by diagnosis, b) RNA integrity number (RIN) by diagnosis, c) Braak score by diagnosis and d) age at death in years by diagnosis for the MayoRNAseq dataset



a) b)



c)

Figure 3-5 – Boxplots for a) age at death in years by Braak score, b) RNA integrity number (RIN) by Braak score and c) post-mortem interval (PMI) in hours by Braak score for the MayoRNAseq dataset.

	Braak score	PMI	Age at death	RIN
Braak score		-0.06 0.58	-0.04 0.69	0.44 $1.58x10^{-5}$
PMI			0.32 $2.22x10^{-3}$	0.04 0.69
Age at death				0.09 0.38
RIN				

Figure 3-6 – Correlation between Braak score, post-mortem interval number in hours (PMI), age at death in years, and RNA integrity number (RIN) for the MayoRNAseq dataset (top value is correlation and bottom value is p-value)

3.3.3 Initial investigation of ROSMAP dataset

For the ROSMAP dataset, age at death differed between cases and controls, and there were six significant covariate correlations. Braak score and CERAD score were positively correlated (r=0.61; p-value = 2.41×10^{-39}) and this is consistent with the wider literature (Boluda et al. 2014). Additionally, Braak scores were positively significantly correlated with age at death (r=0.44; p-value = 8.06×10^{-19}) and negatively with RIN (r=-0.18; p-value = 5.78×10^{-04}). CERAD scores were also significantly correlated with age at death (r=0.32, p-value = $3.63x10^{-10}$), and RIN (r=-0.14; p-value=$6.37x10^{-03}$) (Figure 3-10). PMI and Braak score were not significantly correlated (r = 0.08; p-value = 0.15). From the data available we cannot explain the reason as to why there are negative correlations between Braak and RIN and CERAD and RIN. A previous study has shown that brain samples with AD show greater RNA degradation than brain samples with no disease pathology. No difference in RNA degradation was seen when comparing brain samples with Parkinson's disease or Huntington's disease to controls samples, suggesting that this degradation is seen specifically in AD (Highet et al. 2021).

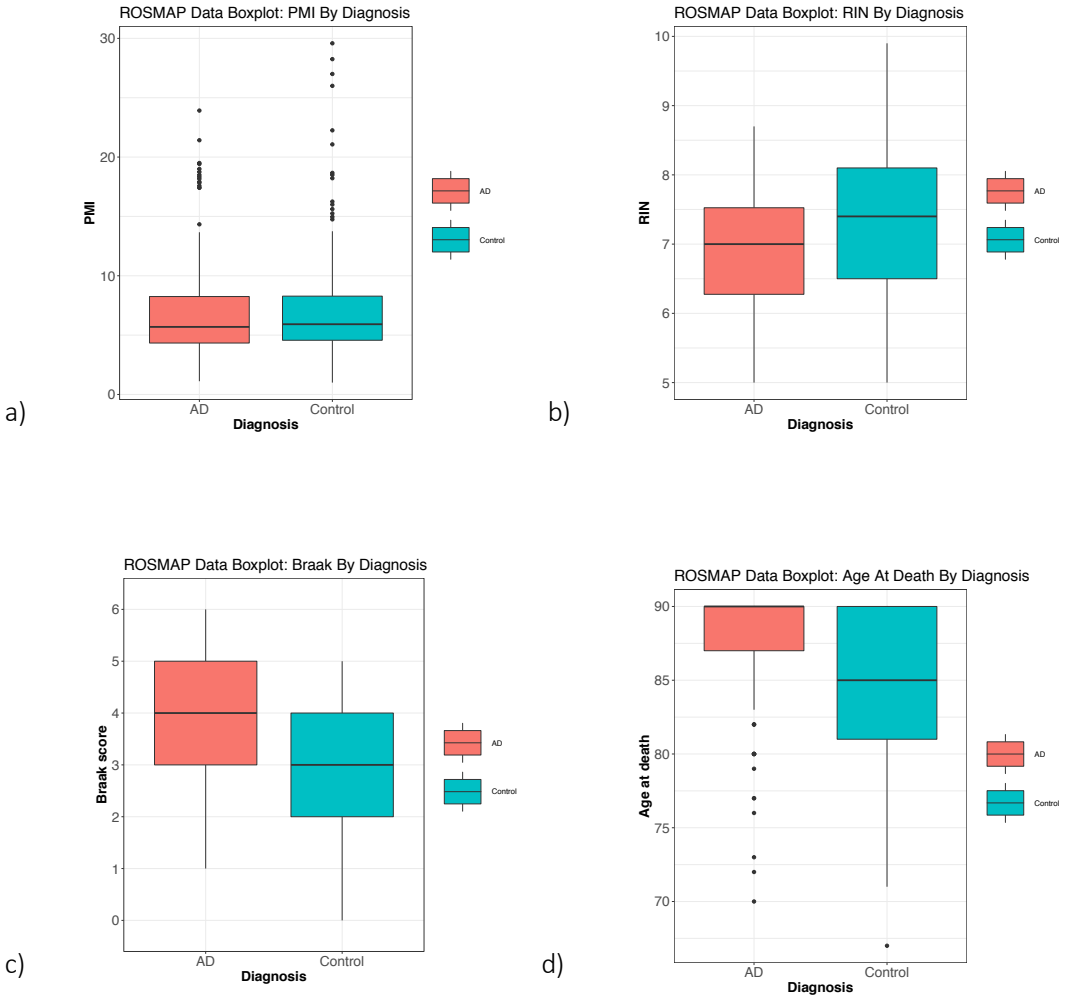
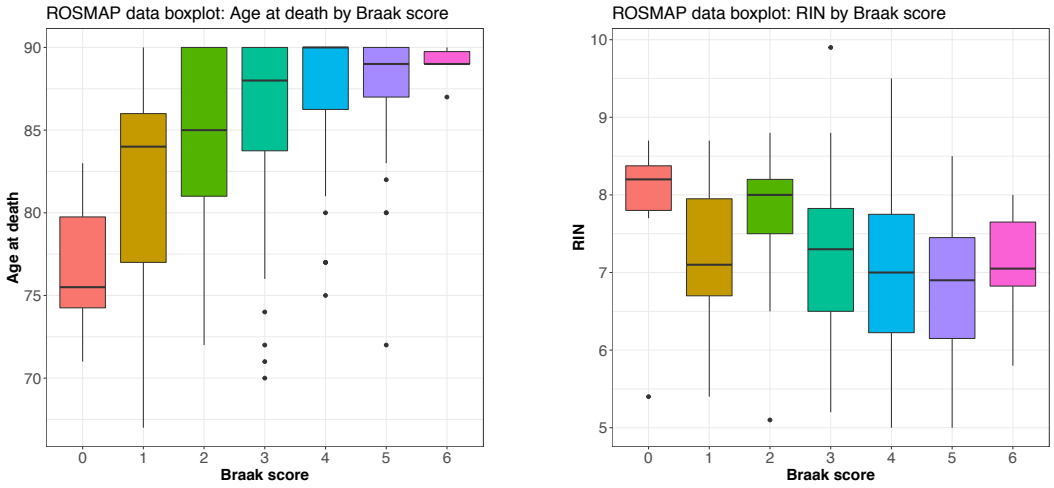
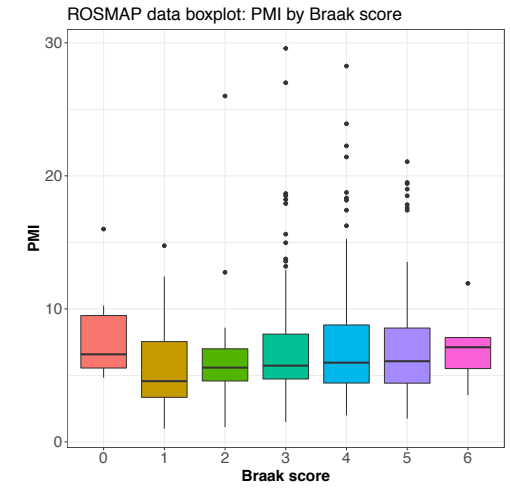


Figure 3-7 Boxplots for a) post-mortem interval (PMI)in hours by diagnosis, b) post-mortem interval (PMI) in hours by diagnosis, c) Braak score by diagnosis, d) age at death in years by diagnosis for the ROSMAP dataset



a) b)



c)

Figure 3-8 - Boxplots for a) age at death in years by Braak score, b) RNA integrity number (RIN) by Braak score and c) post-mortem interval (PMI) in hours by Braak score for the ROSMAP dataset

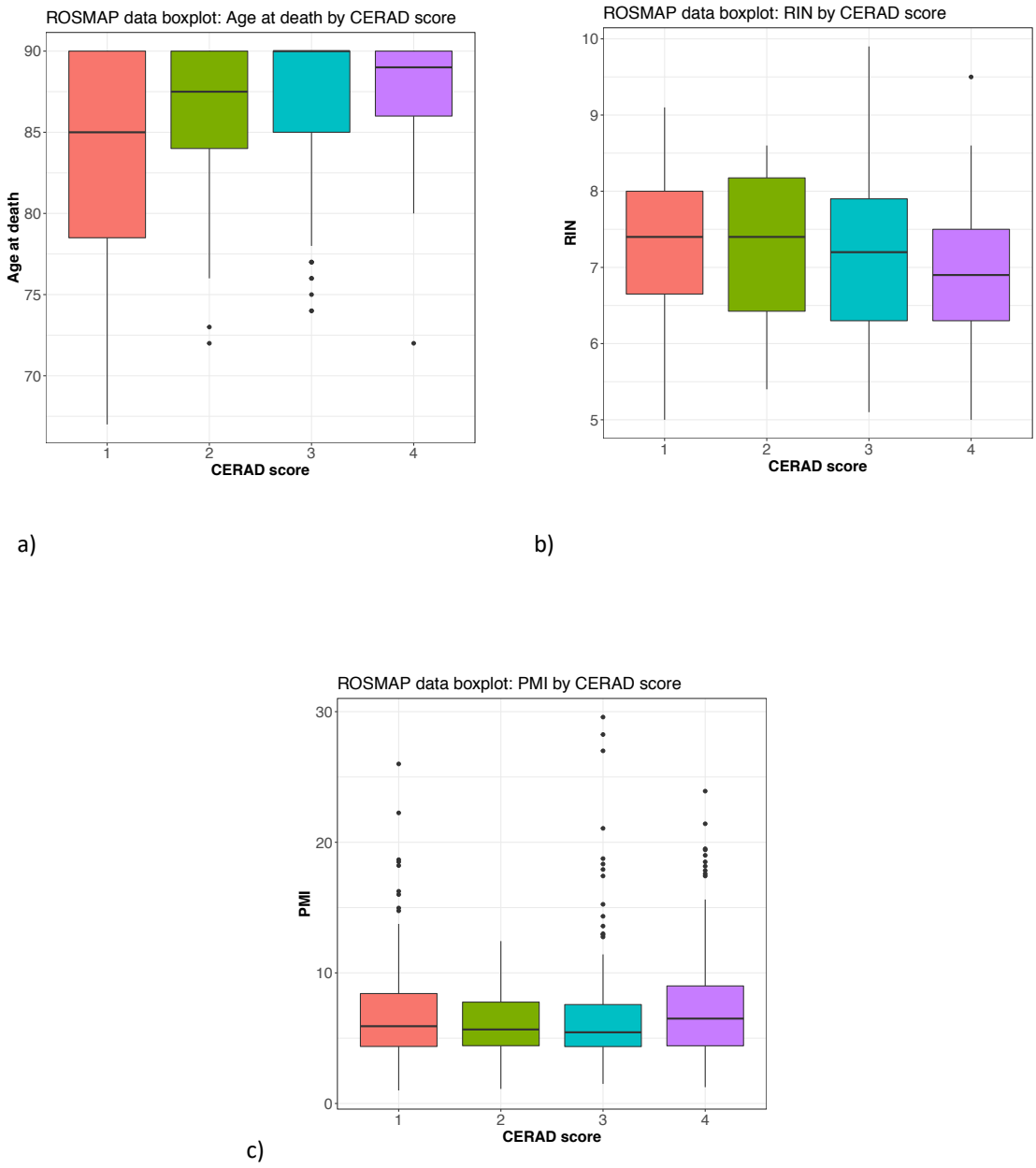


Figure 3-9 Boxplots for a) age at death in years by CERAD score, b) RNA integrity number (RIN) by CERAD score and c) post-mortem interval (PMI) in hours by CERAD score for the ROSMAP dataset

	Braak score	PMI	Age at death	RIN	CERAD score
Braak score		0.08 0.15	0.44 8.06×10^{-19}	-0.18 5.78×10^{-4}	0.61 2.41×10^{-39}
	PMI		-0.01 0.81	-0.13 0.01	0.05 0.39
		Age at death		-0.09 0.08	0.32 3.63×10^{-10}
			RIN		-0.14 6.37×10^{-3}
				CERAD score	

Figure 3-10 Correlation between Braak score, post-mortem interval number in hours (PMI), age at death in years, RNA integrity number (RIN) and CERAD score for the ROSMAP dataset *(top value is correlation and bottom value is p-value)*

3.3.4 Initial investigation of MSBB dataset

For the MSBB dataset Braak and CERAD score were also strongly correlated (r=0.81; p-value = $4.93x10^{-40}$. Braak scores were also significantly correlated with PMI (r=0.61; p-value = 2.41×10^{-39}) and RIN (r=0.61; p-value = 2.41×10^{-39}). CERAD score was also significantly correlated with PMI and RIN as well as PMI and age at death (Figure 3-13).

As a result, these relationships may need to be taken into account when utilising the full dataset to make sure any association is with the effect of interest and not a potentially related covariate.

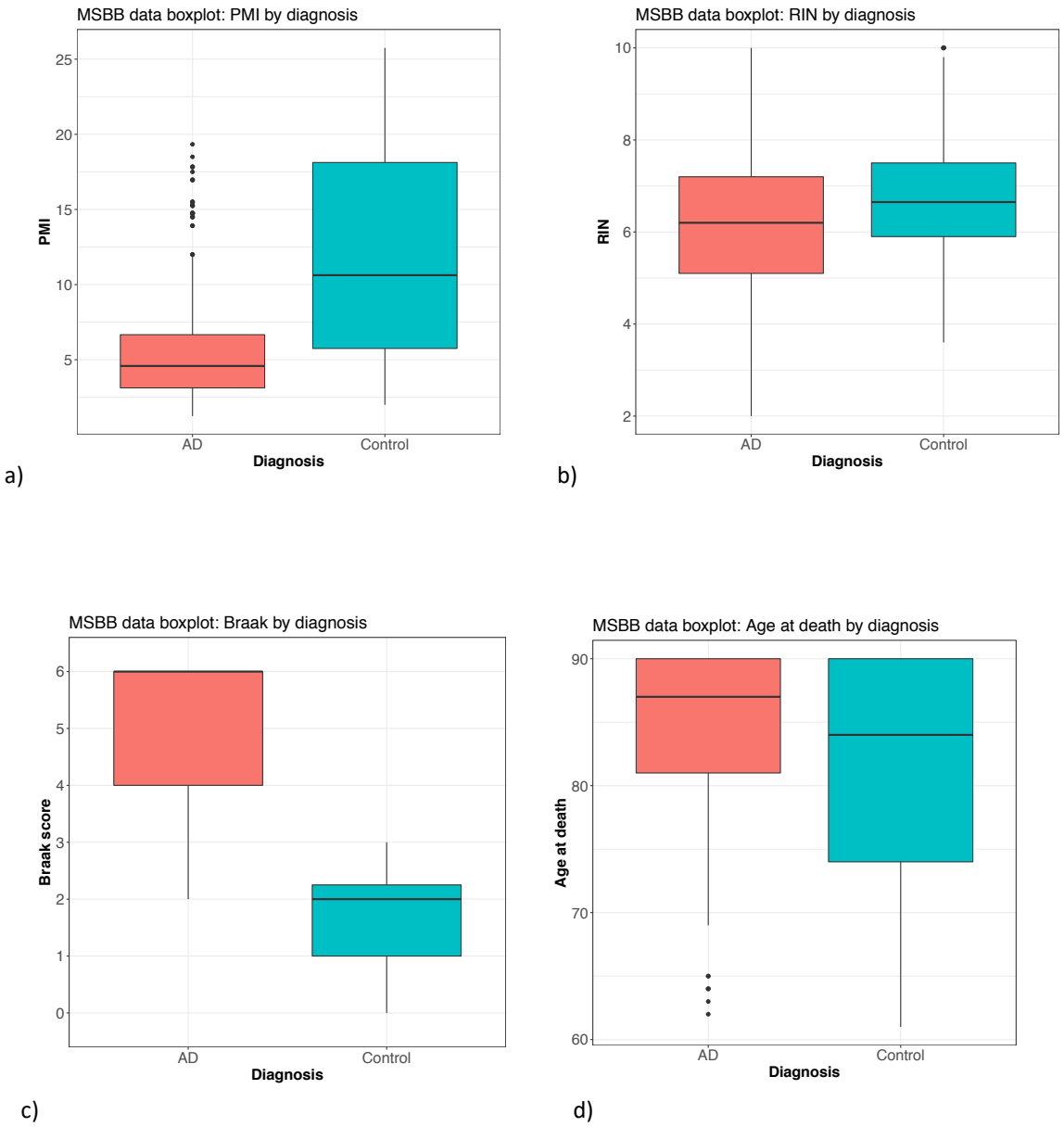


Figure 3-11 Boxplots for a) post-mortem interval (PMI) in hours by diagnosis, b) RNA integrity number (RIN) by diagnosis c) Braak score by diagnosis, d) age at death in years by diagnosis for the MSBB dataset

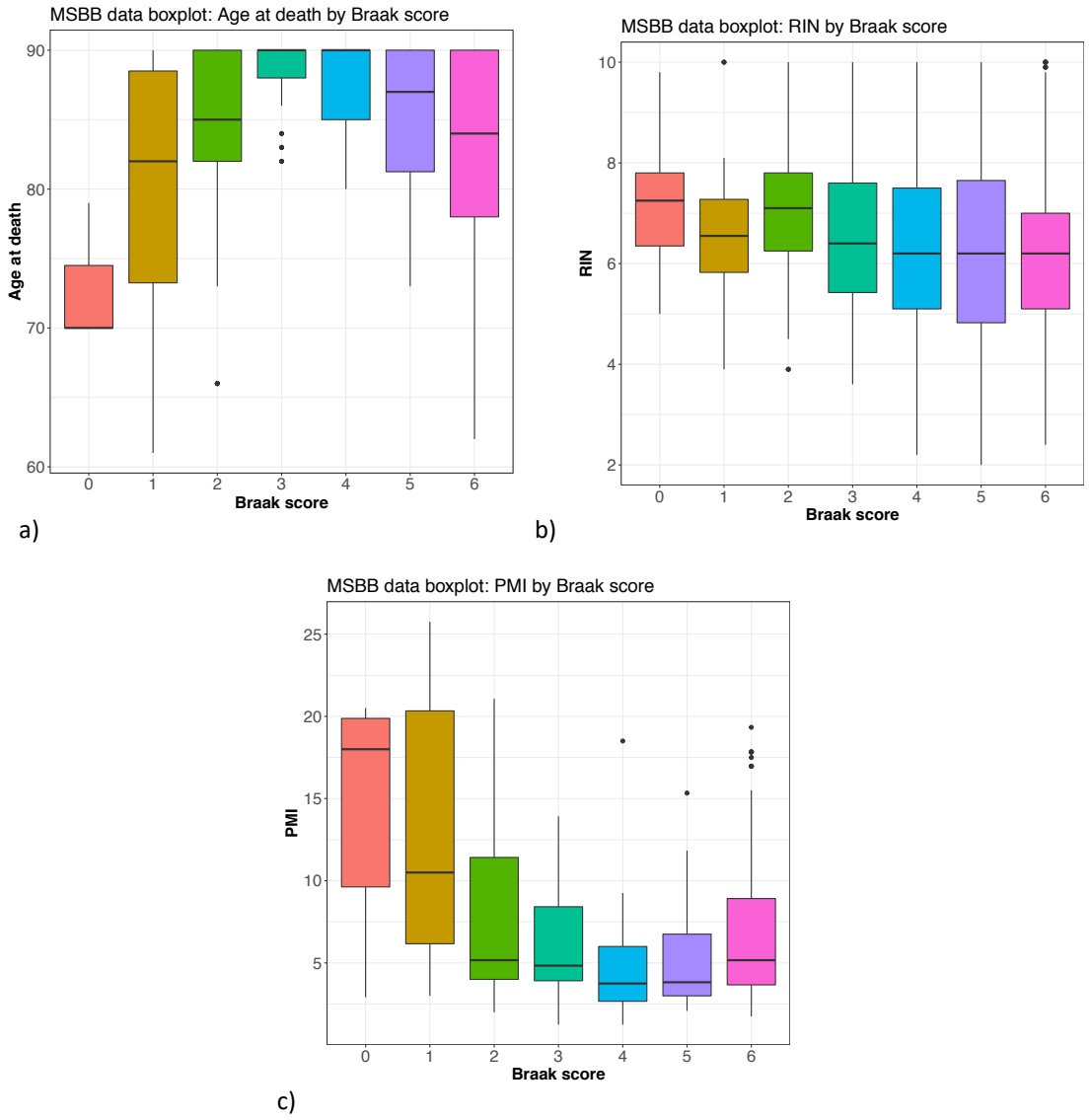


Figure 3-12 Boxplots for a) age at death in years by Braak score, b) RNA integrity number (RIN) by Braak score and c) post-mortem interval (PMI) in hours by Braak score for the MSBB dataset

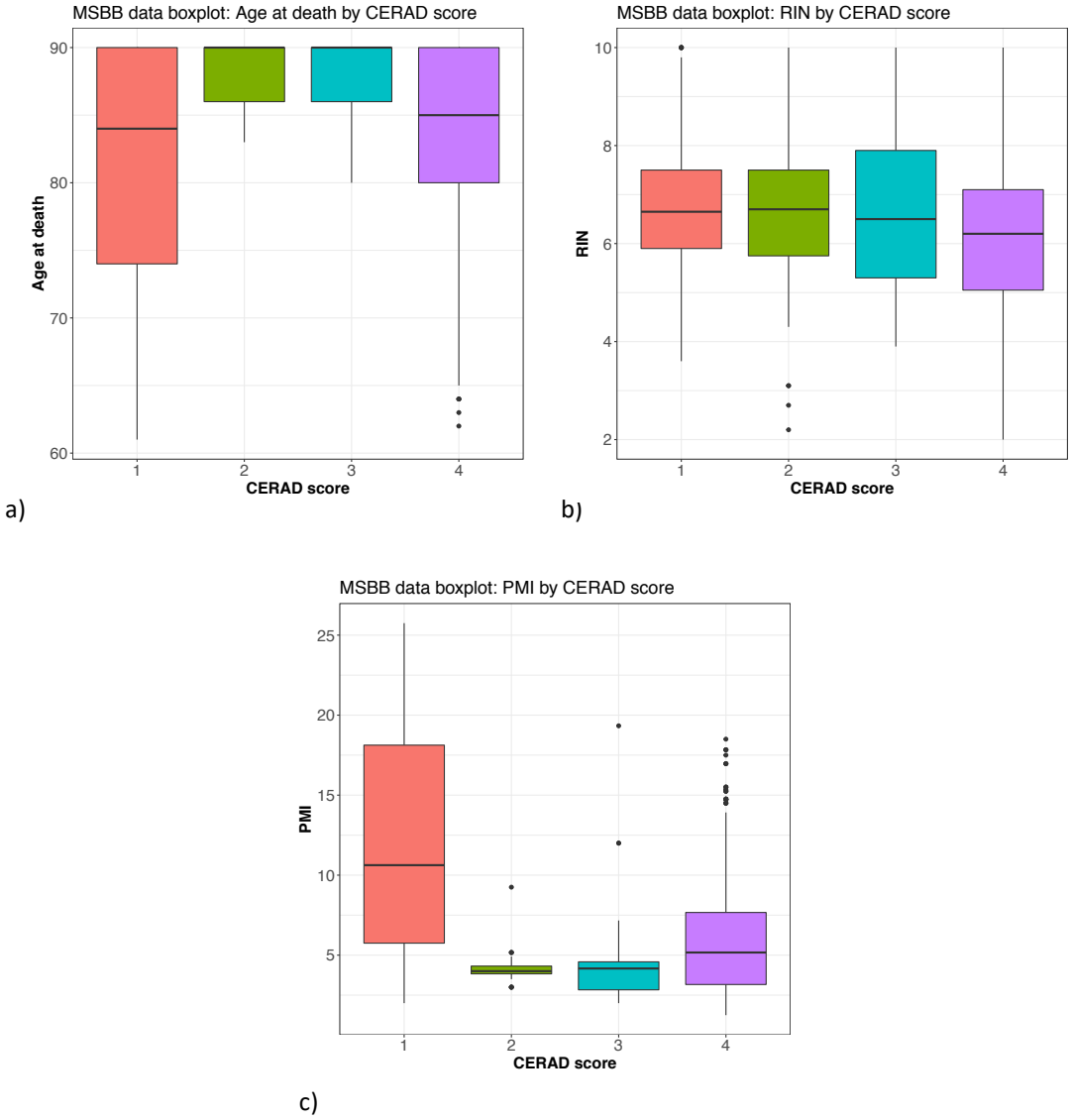


Figure 3-13 Boxplots for a) age at death in years by CERAD score, b) RNA integrity number (RIN) by CERAD score and c) post-mortem interval (PMI) in hours by CERAD score for the MSBB dataset

	Braak score	PMI	Age at death	RIN	CERAD score
Braak score		-0.26 7.43×10^{-4}	0.02 0.83	-0.28 2.17×10^{-4}	0.81 4.93×10^{-40}
PMI			-0.29 1.40×10^{-4}	-0.09 0.27	-0.31 5.05×10^{-5}
Age at death				0.07 0.35	0.04 0.58
RIN					-0.23 2.46×10^{-3}
CERAD score					

Figure 3-14 Correlation between Braak score, post-mortem interval number in hours (PMI), age at death in years, RNA integrity number (RIN) and CERAD score for the MSBB dataset (top value is correlation and bottom value is p-value)

3.3.5 Analysis of MayoRNAseq, ROSMAP, and MSBB biplots to identify potential confounding

For each of the three studies, PC biplots were inspected to detect any obvious batch effects or sources of unwanted variation. PC biplots were generated and inspected for the three datasets for the first 10 principal components and individuals labelled for the following variables: age at death, *APOE* genotype, sequencing batch, diagnosis, PMI, RIN, sex, sample source and additionally tissue type for the MSBB data.

The normalisation process of the MayoRNAseq RNA-seq data resulted in 17,392 genes. Initial investigation of the PCs on the normalised gene expression data identified no obvious known batch effects as per Figure 3-15 which suggests slight separation on AD case or control diagnosis.

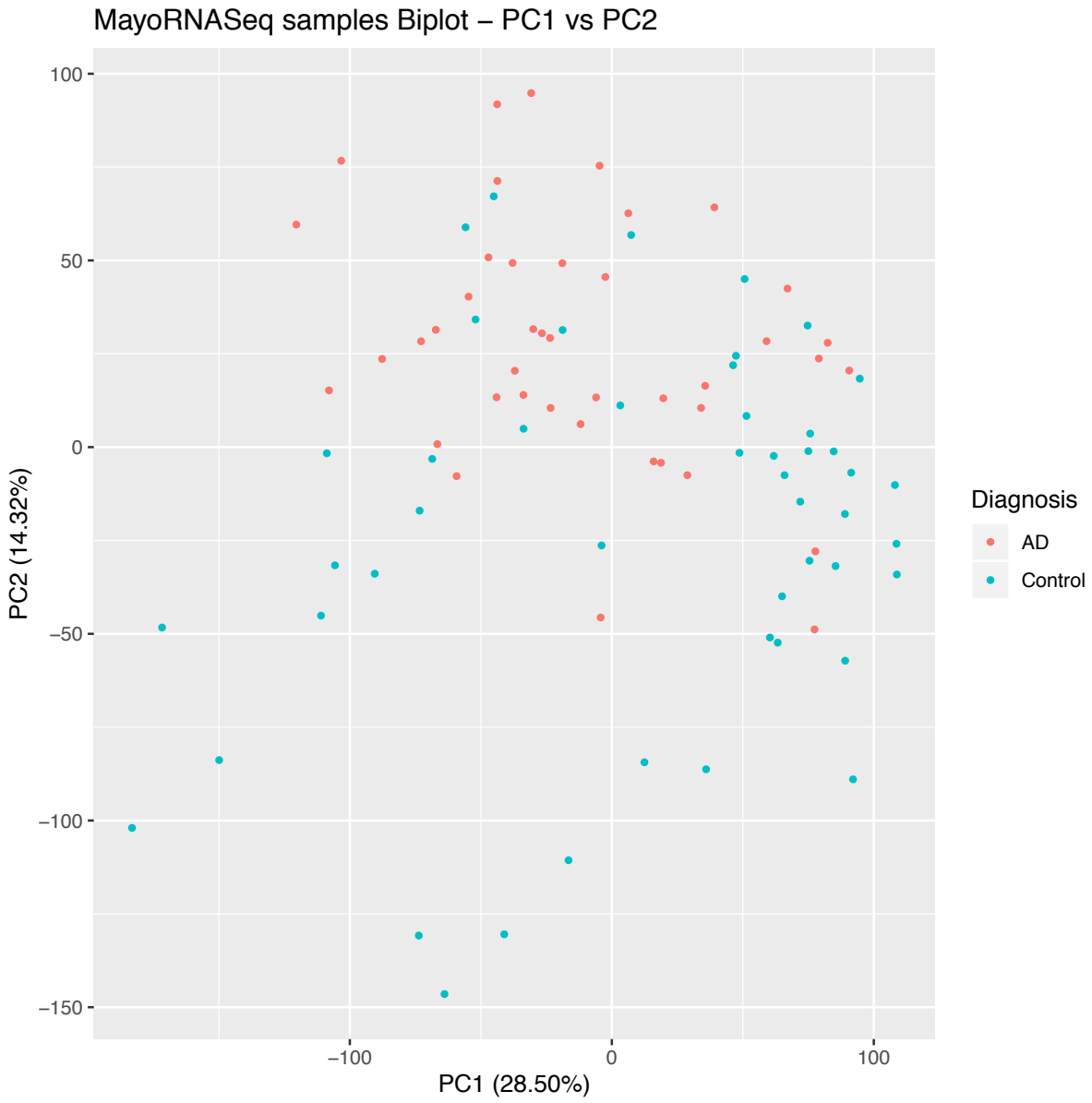


Figure 3-15 A PCA biplot of the first principal component (PC1) vs the second principal component (PC2) with the points coloured by diagnosis for normalised gene expression data from the MayoRNAseq study. The percentage figures refer to the variance that each principal component captures.

The normalisation process on the ROSMAP data resulted in 16,960 genes. Initial analysis of the PCs on the normalised data identified that sequencing batch (Figure 3-16) was an unwanted batch effect. From the PC biplot, it is possible to see that sequencing batch 7 forms a cluster which is completely segregated. Additionally there is some segregation within the larger cluster by sequencing batch.

Figure 3-16 - A PCA biplot of the first principal component (PC1) vs the second (PC2) with the sample points coloured by RNA sequencing batch for normalised gene expression data from the ROSMAP study. The percentage figures refer to the variance that each principal component captures.

The normalisation process on the MSBB data resulted in 17,810 genes. Initial analysis of the PCs on the normalised data also identified that sequencing batch was an unwanted source of variation as seen in Figure 3-17.

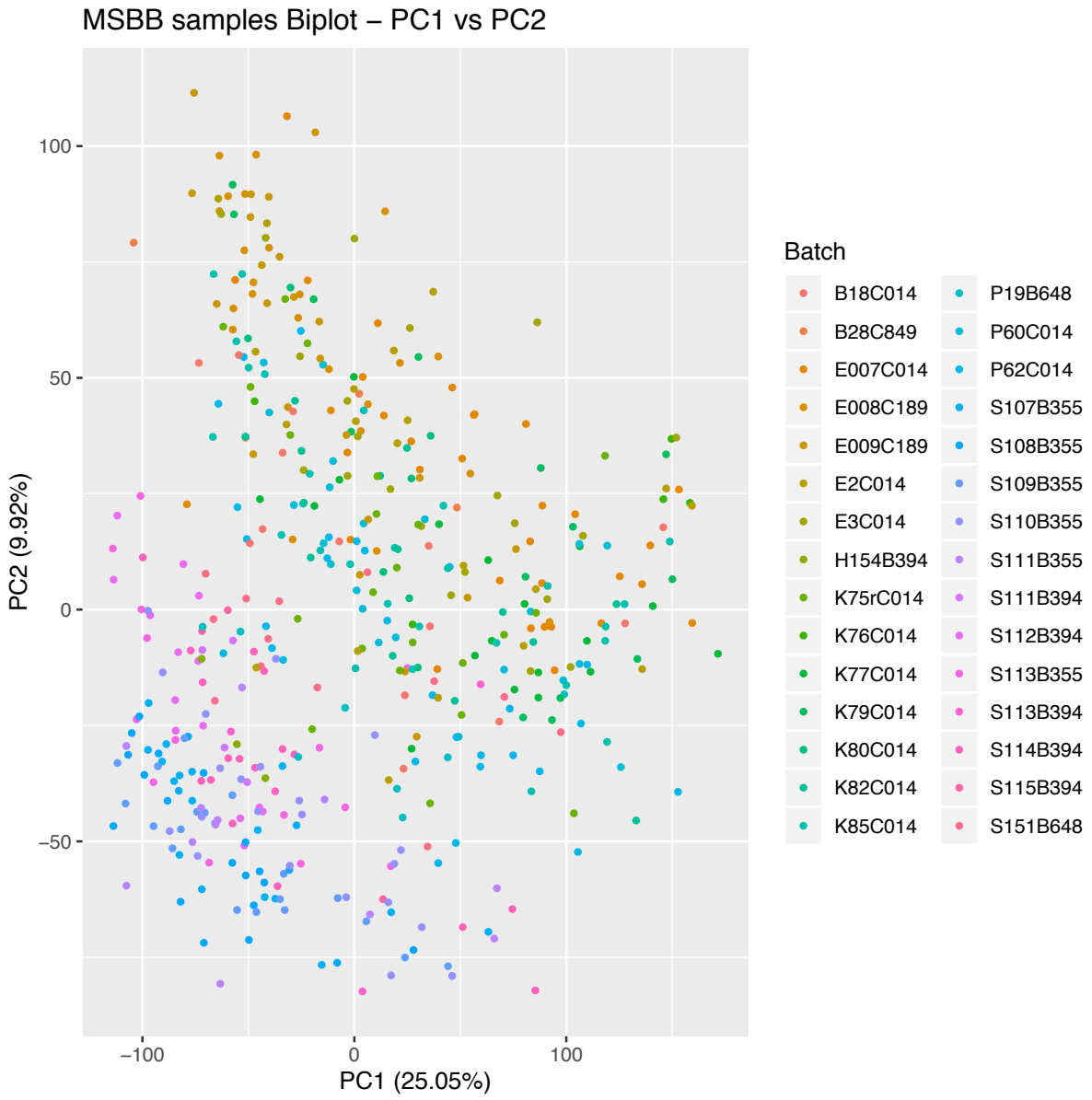


Figure 3-17 - A PCA biplot of the first principal component (PC1) vs the second (PC2) with the sample points coloured by RNA sequencing batch for normalised gene expression data from the MSBB study. The percentage figures refer to the variance that each principal component captures.

3.3.6 Merging datasets

After merging the three datasets, a total of 16,485 genes were left in common for analysis with 930 samples coming from 627 unique individuals. Once the datasets had been merged, PCA was performed to detect any batch effects. Unsurprisingly a large batch effect can be seen in Figure 3-18 as the samples clearly separate by study as three clusters have formed.

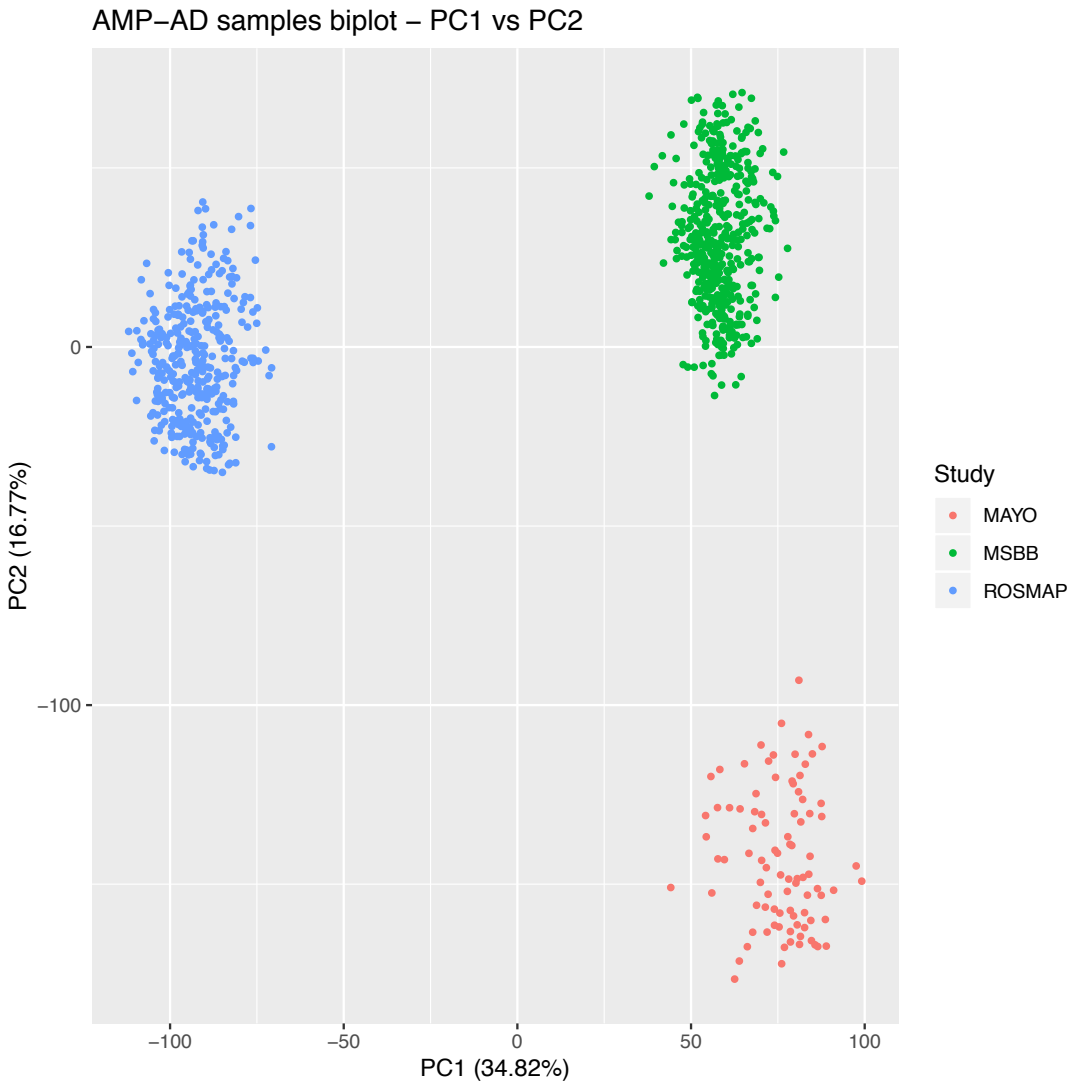


Figure 3-18 PCA biplot of the three studies combined showing segregation by study

Figure 3-19 (next page) also shows the segregation by sequencing batch is still present within the three clusters.

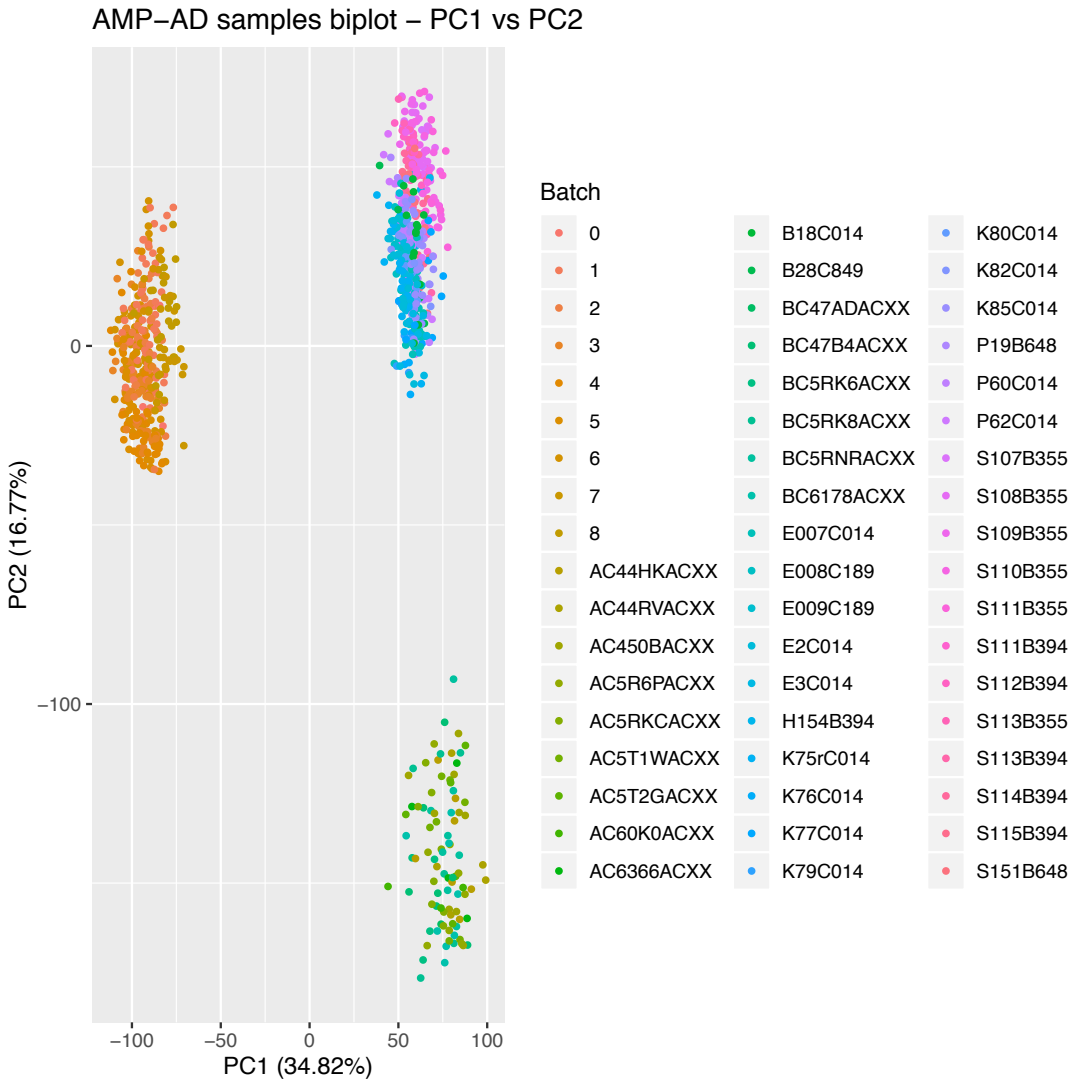


Figure 3-19 – PCA biplot of the three studies combined showing segregation by sequencing batch

A scree plot was generated to inspect the proportion of variance explained by the first 10 principal components which can be seen in Figure 3-20. The elbow of the scree plot was used to determine the number of PCs to obtain. Figure 3-20 shows that PC1 explains 35% of the variance, PC2 17%, PC3 12%, and PCs 4 and 5 both explain 3%. As the plot levels off from PC4 onwards, the first three PCs were retained to include in the LMEM.

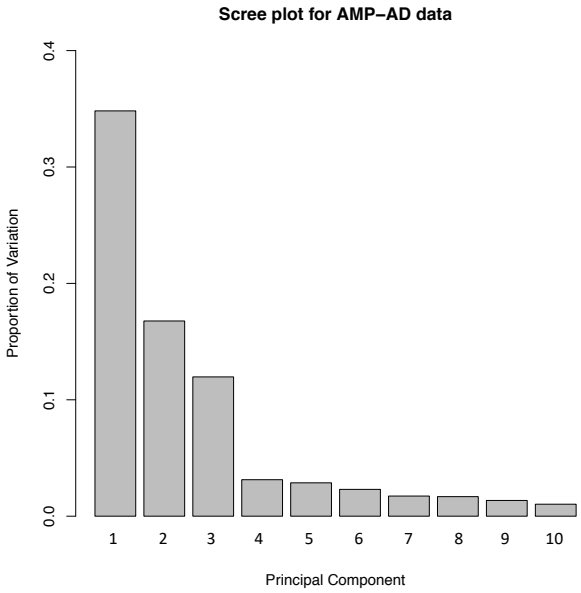


Figure 3-20 – A scree plot showing proportion of variation for each principal component for AMP-AD data

After performing the LMEM correcting for age at death, sex and the first three principal components as fixed effects in addition to individual ID and sequencing batch as random effects, the resulting biplot for study can be seen in Figure 3-21. Figure 3-22 shows the biplot of PC1 and PC2 using the residuals from the LMEM analysis and shows no segregation based on sequencing batch. Biplots for age at death, sex, RIN and PMI were also inspected and no segregation was seen based on these suggesting adequate correction for these variables (Figure 3-23).

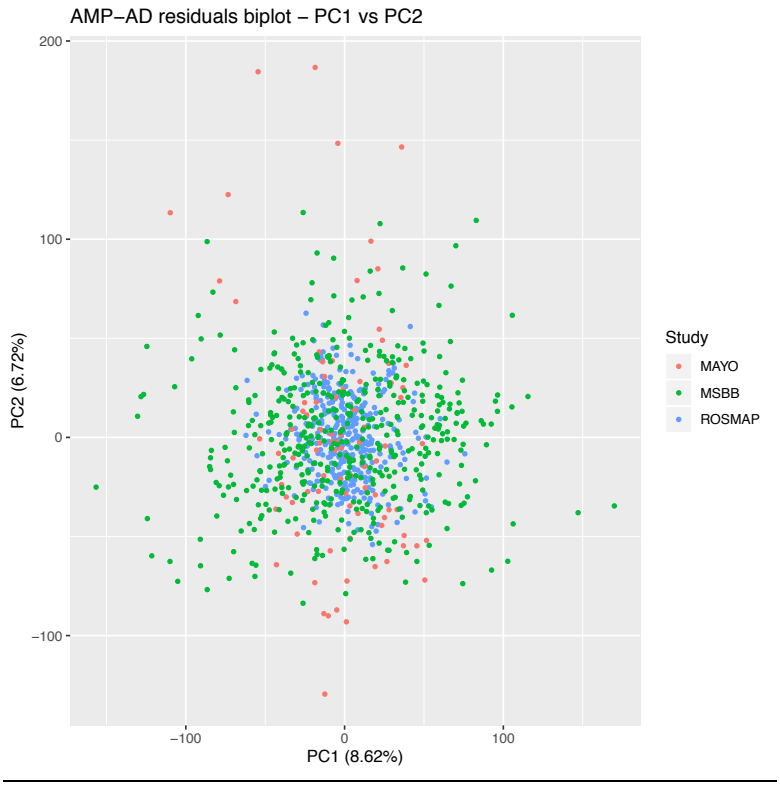


Figure 3-21 – Biplot of PC1 vs PC2 of the residuals from the linear mixed-effect model showing correction for originating study. The percentage figures refer to the variance that each principal component captures.

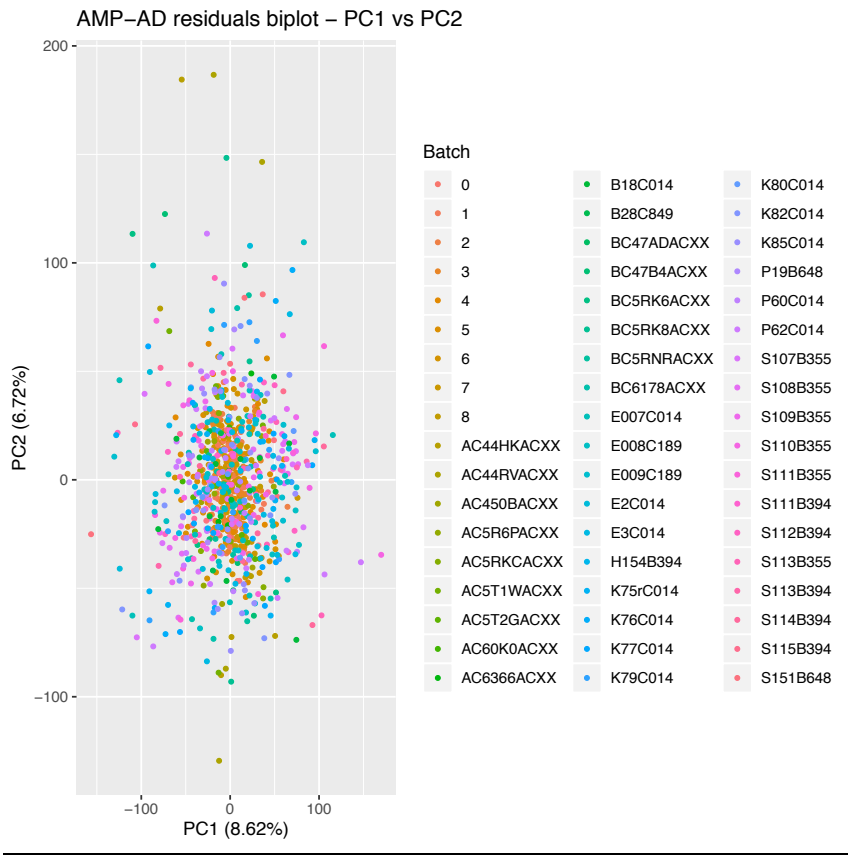


Figure 3-22- Biplot of PC1 vs PC2 of the residuals from the linear mixed-effect model showing correction for sequencing batch. The percentage figures refer to the variance that each principal component captures.

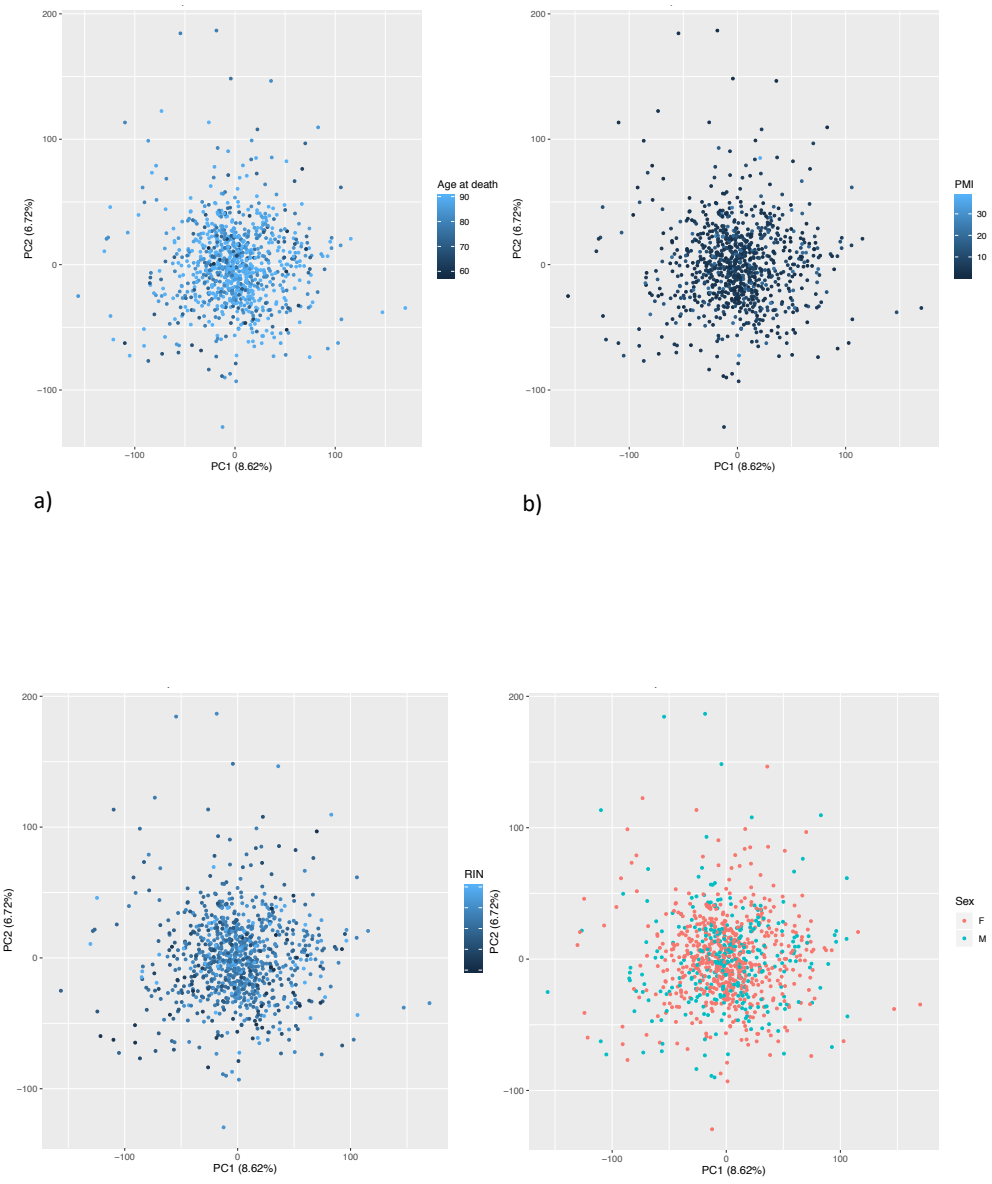


Figure 3-23 – PCA biplots of PC1 vs PC2 of the residuals from the linear mixed-effect model for a) age at death, b) post-mortem interval, C) RNA integrity number and d) sex

3.4 Discussion

The overall aim of this work was to produce a single, homogenous dataset consisting of RNA-seq data linked with genetic and phenotypic information using the data made available from the AMP-AD reprocessing initiative. This dataset can then be used in downstream analyses to hopefully help increase our understanding of the biology of Alzheimer's disease.

The first aim was to determine and define the phenotypic labels that can be used downstream. These were CERAD scores, Braak scores and AD case and control status. Defining these variables so that they were harmonious across the three studies was challenging as the three studies did not capture the phenotypic information in the same way.

The CERAD score which is an age-related measure of post-mortem neuropathology was a major source of disharmony. Individual-level data was not available to download in the MayoRNAseq data, although the authors confirm that control subjects had a CERAD neuritic and cortical plaque density of 0 (none) or 1 (sparse) (Allen et al. 2016). As a result, MayoRNAseq samples were excluded from any analyses using CERAD scores. Additionally, the ROSMAP study used a modified CERAD criteria in that it was an absolute measure of the pathology rather than an age-corrected measure of pathology (Mirra 1997; Bennett et al. 2012a; Bennett et al. 2012b). In contrast the MSBB study did follow the CERAD criteria for the generation of their CERAD scores (Mirra 1997; Wang et al. 2018).

The Braak score is a staging system for neurofibrillary tangles, and it has also been previously criticised for significant inter-rater variability (Boluda et al. 2014). The three studies used slightly different protocols to generate their Braak scores with the MayoRNAseq study being the only study to report half scores highlighting a phenotypic inconsistency.

The third phenotype was AD case-control status. The most up-to-date method of diagnosing AD post-mortem is to use the National Institute on Aging-Alzheimer's Association (NIA-AA)

criteria. These guidelines base a diagnosis on a combination of Thal Aβ amyloid phase, Braak score and CERAD score (DeTure and Dickson 2019).

MayoRNAseq used the NINCDS-ADRDA criteria in the diagnosis of AD case and AD control status which is the diagnosis criteria that was the predecessor to the NIA-AA criteria. The NINCDS-ADRDA criteria was based on the hypothesis that AD is a disease mostly comprising of both clinical and pathological symptoms and that these two are closely related (McKhann et al. 1984). As research has progressed it has become apparent that this clinical-pathological relationship is not consistent (Jack et al. 2011).

A limitation of this study is that the definition of AD case status presented in this thesis does not reflect the current gold standard of diagnosis. Another is that the phenotypes are not completely harmonious. However, these limitations are seen across AD research. For example, the ADGC consortium consists of samples from ROSMAP and the Mayo clinic along with 14 other cohorts and with often differing definitions of case-control status (Naj et al. 2011). The ADGC consortium regularly contributes to larger studies as seen in more recent GWAS and GWAS-by-proxy (GWAX) studies (Kunkle et al. 2019; Schwartzentruber et al. 2021). Thus, demonstrating that despite inconsistencies in phenotypes, our understanding of AD can still be advanced and henceforth the inclusion of these less than perfect variables.

The second aim was to extensively QC the available RNA-seq data. At present there is no best practise on how to pre-process and normalise RNA-seq data for downstream processes such as differential gene expression analysis. Studies often neglect to fully report their QC processes or perform them only on a limited basis. Extensive QC is vital for RNA-seq experiments as it helps to improve the reproducibility of the biological results even if it is a bioinformatic challenge to do so (Sheng et al. 2016).

The samples included in this analysis underwent extensive pre-processing and quality control using a multi-faceted approach. They underwent between-sample and within-sample

normalisation to remove as much technical variation as possible. RNA samples were also cross-referenced with their reported genetic sample to minimise the risk of sample swaps. This was all performed to produce a high-quality dataset that can be utilised in further analyses and minimise spurious findings in future analyses.

During the QC process, 159 MayoRNAseq, 181 ROSMAP and 76 MSBB individuals were excluded from the analysis due to diagnosis or missing phenotype data. This was a total of 374 across the three cohorts (33.59%). The main reason for such a high sample loss is that many of the samples were either MCI, so did not meet the threshold for AD case or were diagnosed with another disorder such as PSP. 29 MayoRNAseq, 82 ROSMAP and 71 MSBB individuals were excluded from the analysis due to sample QC. This was a total of 182 individuals across the three studies (14.9%). Factors contributing to this sample drop were mainly due to the original investigators flagging problems associated with these samples (such as sex mismatch), no genetic data being available, or the individuals were not of European ancestry.

After the QC workflow, initial investigation of the three datasets determined that there was likely unwanted technical variation. RIN and PMI were identified as potential confounders so needed to be considered during the LMEM and PCA step. In this analysis it was found that Braak score and CERAD score were correlated which is consistent with the literature (Boluda et al. 2014). Some of the other correlations could not be explained. For example, in the analysis of the MayoRNAseq cohort, PMI was correlated with age at death ($r=0.32$, p-value=1.58×10^{-05}). It is not possible to confirm the origin of this association given the data available. It could be a spurious association, but it could also be a result of some of the factors occurring during collection. For example, it could be that due to control brain samples being harder to recruit, these samples were prioritised for collection by the brain bank and expedited. This results in samples from individuals with a younger age at death having a lower PMI (as controls are generally younger than the AD individuals within this study). Further study would be required to determine the true nature of this relationship and therefore this comment is purely speculative.

The final aim was to use LMEM with PCA to combine the three datasets to create a more powerful one, to help enhance the biological signal in downstream analyses. Combining independently generated datasets is challenging and this partially stems from the fact that different studies will use different experimental methods. These and other unknown technical artefacts can remain in RNA-seq data. This unwanted source of variation can lead to spurious results in future analyses if not carefully corrected for.

Correcting for batch effects and unwanted variation in RNA-seq data is still a developing field and best practice has not yet been defined. RIN was determined to be a potential confounder and it is regularly corrected for in RNA-seq studies. One method to determine RIN is by placing RNA on a Bioanalyzer and obtaining a tracing of fragment sizes per sample. RIN ranges from 0 (completely degraded RNA) to 10 (high quality RNA). RNA quality biases can affect downstream differential expression analyses. It has been shown that standard RIN correction such as including RIN as a covariate in a model may not be the best approach. One paper found that adjusting for RIN largely fails to remove RNA degradation bias (Jaffe et al. 2017). The approach taken in this thesis was to allow for general PCA to correct for confounding. In the same paper above, the authors propose their own method of quality surrogate variable analysis (qSVA) to correct for such biases. The qSVA method is available under the SVA Bioconductor package. The authors claim that the qSVA approach may have an advantage over PCA as there is less risk of removing true signals along with noise. However, a comparison was not performed and so the best approach remains to be seen (Jaffe et al. 2017).

All sources of unwanted variation cannot be accounted for. The use of including PCs in the model as performed in this analysis, helps to remove some of this unwanted variation. Using LMEM in combination with PCA, it was possible to account for some of the known and unknown unwanted technical variation in the three individual datasets. After performing the LMEM, no obvious batch effects in the first 10 PCs were seen including for RIN and PMI, each of which was identified as a potential confounder.

In order to determine the number of PCs to include in the model, scree plots were used. Scree plots have been criticised for being ambiguous, however they have the advantage of being simple to implement (Ledesma et al. 2015). For the combined analysis, the scree plot suggested that 3 PCs may be enough to include in the model. The 3PC model identified no obvious batch effect in the first 10 PCs. Models beyond 3PC were also explored and little to no batch effects were noticed. Due to diminishing variance importance, including these additional principal components would be unnecessary and potentially start to take away from any actual explanatory effect in future analyses. The downside to correcting with PCs is that it can be conservative as the PCs could also comprise or capture some of the main effect of interest.

Sequencing batch and original study were sources of unwanted variation seen in this analysis. The rationale behind using LMEMs was that these obvious batch effects are non-independent. Using LMEMs allowed me to include sequencing batch and individual ID in a model as random effects and thus overcome the sequencing batch grouping factor and the repeated measures seen in the MSBB data.

One important point to make of the work presented here is that gene expression data from different brain regions have been combined. They have been combined in order to increase sample size with the caveat that this may increase the heterogeneity of the disease-relevant biology. Each of the three studies contributed samples from different brain regions with no region overlapping between studies. Therefore, it is challenging to disentangle any effect seen for region as it is correlated with the study effects.

All the brain regions included in this analysis are known to be implicated in AD albeit to varying extents. The ROSMAP study used DLPFC which is thought to contribute to abstract thought, working memory and complex cognition all of which can be affected in AD (Cieslik et al. 2012). The MSBB study contributed tissue from BM10, BM22, BM36 and BM44 areas. BM10 constitutes part of the prefrontal cortex but its function is not well understood but thought to contribute to episodic and working memory and decision making (Gilbert et al. 2006; Soon

et al. 2008; Knowlton et al. 2012). There is evidence that *APOE* ε4 carriers show different brain activity patterns in this area in comparison to non-carriers during working memory tasks (Wishart et al. 2006). BM22 is part of Wernicke's area, which is an area thought to be vital for speech production and known to be especially vulnerable in AD (Haroutunian et al. 2009; Binder 2015; Wang et al. 2016). BM36 has been found to be affected by tau-pathology in AD (Berron et al. 2021). BM36 and BM44 have previously been described as two of the most vulnerable regions to AD (Wang et al. 2018) . The MayoRNAseq study utilised temporal cortex (TCX) which is one of the first regions affected with AD neuropathology (Allen et al. 2016). The study also utilised the cerebellum, however as its involvement in AD is not well understood and its distinct gene expression profile in comparison to brain cortex, the cerebellum was excluded from this analysis (Allen et al. 2016; Jacobs et al. 2017).

The purpose of the analysis presented in this thesis is to produce a dataset that will be used to identify quantitative changes in expression levels in the brain cortex between AD cases and controls. A limitation of this approach is that the resolution for region specific gene expression differences are reduced, and it may not be able to detect differences that vary extensively among the different brain regions as their effects may be diluted. However, AD gene expression profiles have previously been found to be similar between different cortical regions of the brain (Chappell et al. 2018). Additionally, the approach taken in this analysis is in line with other published work that has utilised these datasets together. One is an eQTL analysis where the ROSMAP and MayoRNAseq studies were combined with the Common Mind Consortium data (Sieberts et al. 2020). Initially the studies were analysed individually and then meta-analysed. Focusing on the cortex could still offer novel insights into the pathophysiology and aetiology of AD.

LMEMs have previously been applied individually to the MayoRNASeq and MSBB datasets. LMEMs have been used to include individual ID as a random effect to adjust for the inclusion of multiple samples from different tissues from the same individual (Wan et al. 2020). The work presented in this chapter is the first time that LMEMs have been applied to the three AMP-AD studies to produce a single dataset ready for future analyses such as differential expression or eQTL analysis. Both of these will be discussed in the following chapters.

Chapter 4 – Differential Expression and Gene Ontology enrichment analysis of the combined AMP-AD dataset

4.1 Introduction

4.1.1 Differential gene expression and gene ontology enrichment analysis

The first big genomic studies in Alzheimer's disease (AD) were case-control genome-wide association studies (GWAS). GWAS have identified many genetic loci associated with AD increasing our knowledge of the genetic architecture of AD, but they do not account for the total heritability of AD (Lambert et al. 2013; Kunkle et al. 2019). Twin studies have estimated the heritability to be between 58-79% but the heritability based on common genome-wide SNPs is estimated to only be between 27 and 55% (Gatz et al. 2006; Cuyvers and Sleegers 2016; Escott-Price et al. 2017b). Additionally, translation of these loci to therapeutics or biomarker discoveries has been disappointing. The next challenge is to identify risk genes and functional variants at these risk loci and the role they may play in the development of AD. Expanding into other data from other omics could hold the key to new understanding of the aetiology of AD.

Genetic and environmental risk factors can influence gene expression (Fenoglio et al. 2018) and in doing so could perturb biological pathways contributing to the aetiology of AD. Detecting potential perturbed biological pathways is possible through bioinformatic analyses and one approach is to use differential gene expression (DGE) analysis. Using DGE analysis, it is possible to measure and compare gene expression between phenotypic groups to identify differentially expressed genes. Common tools for this approach are DESeq2 (Love et al. 2014), EdgeR (Robinson et al. 2010) and Limma-Voom (Ritchie et al. 2015), which are all available as R packages. Identified pathways can then be investigated further to determine if they are implicated in the aetiology of disease (Ertekin-Taner 2017).

4.1.2 Aims

The first aim of this chapter was to initially normalise gene expression data from the ROSMAP study using LMEM and then peform a logistic regression (LR) on the resulting normalised residuals to identify differentially expressed genes. This approach will herein be referered to as: LMEM + LR. The next step was to then compare the overlap of differentially expressed genes from the LMEM + LR approach to two well-known DGE packages, namely DESeq2 and Limma-Voom to find differentially expressed genes using the ROSMAP dataset only.

The second aim of this chapter was to perform DGE analyses for Braak and CERAD score phenotypes. To achieve this, a DGE analysis was performed using the residuals from the combined and normalised AMP-AD gene expression data (ROSMAP, MSBB and MayoRNASeq) from the previous chapter. A comparison of the use of logistic regression vs ordinal regression for each of the phenotypes was also performed.

The third aim was to produce a list of differentially expressed genes for case-control status using a logistic regression model. This made use of the residuals from the combined and normalised AMP-AD gene expression data that was generated in the previous chapter.

The fourth aim was to use lists of differentially expressed genes to perform gene ontology (GO) enrichment analysis to determine potential pathways of biological interest.

The final aim was to identify if significant GO terms as identified by a previous GWAS through their MAGMA analysis (Kunkle et al. 2019) were also significant GO terms in the GO enrichment analysis as identified in the previous aim.

4.2 Methods

4.2.1 Differential gene expression analysis of ROSMAP data

Initially, a DGE analysis was performed on the ROSMAP dataset. This was performed in order to identify the overlap of differentially expressed genes as a result of the LMEM + LR method compared to two more established DGE analysis tools (limma-voom and DESeq2). This analysis only used the ROSMAP dataset as opposed to using the combined AMP-AD data (of ROSMAP, MSBB and MayoRNAseq). This was due to the other two tools not being designed to handle combining RNA-seq datasets. Therefore, the larger of the three cohorts (ROSMAP) was selected to identify differentially expressed genes.

The details of the QC methods performed on the ROSMAP data were discussed in the previous chapter. A brief recap of this is that samples were excluded if they: had been flagged by original investigators, were not of a European ancestry, had missing phenotype or genotype data, VerifyBamID indicated genetic and RNA-seq sample mismatch or contamination over 10%, or PCA plots indicated sample outlier. Low count genes were excluded, and the remaining genes underwent CQN normalisation for GC content, library size and gene length were performed.

Correlation and PCA plots were generated to identify batch effects and unwanted confounding and scree plots were generated to identify the number of PCs to include in the LMEM model.

This procedure resulted in 16,960 genes and 369 samples (204 cases and 165 controls). Residuals from the individual LMEM analysis were used in a logistic regression to find differentially expressed genes between cases and controls.

4.2.2 Comparison of LMEM + logistic regression method to DESeq2 and limma-voom in ROSMAP data

DGE analyses using three separate methods were performed to identify differentially expressed genes. These were LMEM + LR method, DESeq2 and limma-voom. For all three methods, low count genes were excluded and counts underwent conditional quantile normalization (CQN). CQN is a method of applying a robust regression algorithm to correct for GC content, gene length and library size (Hansen et al. 2012). Each DGE analysis was performed on the same genes as described above (16,960 genes and 369 samples). All methods included sex, sequencing batch, age at death and the first four principal components of the expression data as covariates in the analysis to overcome batch effects and confounders. The decision to include four principal components was determined through interpretation of scree plots and PC biplots.

4.2.3 Differential expression analysis of AMP-AD data using logistic and ordinal regressions

The combined AMP-AD normalised residuals from the LMEM model discussed in the previous chapter were used to perform logistic regressions using the *glm* function in R and ordinal regressions using the *polr* function in R to assess the statistical significance of each gene.

For each of the Braak and CERAD phenotypes, three regression models were performed (six in total). The primary analyses were logistic regressions using the whole dataset with Braak scores of 0, 1, 2, 3 vs 4, 5, 6 (coded as 0 vs 1) or CERAD scores of 1, 2 vs 3, 4 (coded as 0 vs 1). Braak scores were segregated into a binary phenotype on the same basis that the MayoRNAseq study had included Braak scores into their AD case or control phenotype (Allen et al. 2016) . CERAD score was split into 1 and 2 vs 3 and 4 in keeping with recommendation on the RADC website which hosts the ROSMAP data (https://www.radc.rush.edu/docs/var/detail.htm?category=Pathology&subcategory=Alzheimer%27s+disease&variable=ceradsc) (Mirra et al. 1991; Bennett et al. 2006).

The second logistic regression pair used a reduced dataset where Braak scores of either 0, 1, 2 vs 5 or 6 (coded 0 or 1) or CERAD scores of 1 vs 4 (coded 0 or 1) were included. This reduced dataset was used for two reasons. The first was to exclude the MayoRNAseq samples that were given a score of 3 based on inference, so only known Braak scores were included at the cost of a lower sample size. The second was to exclude the intermediate scores of Braak and CERAD to find differentially expressed genes between more extreme phenotypes. An overview of the Braak and CERAD regressions is given in Table 4-1.

The final pair of regressions were ordinal regressions performed using the whole dataset and full sets of Braak and CERAD scores. The ordinal regression was performed to identify changes in gene expression associated with the degree of Braak and CERAD pathology present.

Finally, a logistic regression was used for a case-control analysis where cases were coded as 1 and controls as 0.

	Braak score	**CERAD score**
Logistic regression (0 vs 1)	0, 1, 2 and 3 vs 4, 5 and 6	1 and 2 vs 3 and 4
Reduced logistic regression (0 vs 1)	0, 1 and 2 vs 5 and 6 (3 and 4 excluded)	1 vs 4
Ordinal regression	0 vs 1 vs 2 vs 3 vs 4 vs 5 vs 6	1 vs 2 vs 3 vs 4

Table 4-1 – Summary of Braak scores and CERAD scores used in logistic and ordinal regression analyses

4.2.4 Gene ontology enrichment analysis

The GO terms chosen to be included in this analysis were obtained from the "gene2go" file, downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/) on March 11th, 2020.

"Parent" GO terms were assigned to genes using the ontology file downloaded from the Gene Ontology website (http://geneontology.org/docs/download-ontology/) on the same date. GO terms were assigned to genes based on experimental or curated evidence of a specific type, so evidence codes IEA (electronic annotation), NAS (non-traceable author statement), RCA (inferred from reviewed computational analysis) were excluded. Analysis was restricted to GO terms containing between 10 and 2000 genes. This was performed by Peter Holmans and distributed by Ioanna Katzourou (see contributions section). GO terms that were obsolete as of 02 December 2020 were removed prior to analysis.

The GO enrichment analysis was performed using the software CATMAP (Breslin et al. 2004). CATMAP was originally designed for ranked gene lists from microarrays, and provides an alternative to methods that require an arbitrary cut-off of top or significant differentially expressed genes when performing the gene-set enrichment analysis as it uses the Wilcoxon rank sum test. Gene lists were ranked based on the p-value from either the logistic or ordinal regressions.

For all phenotypes, three sets of gene lists were produced. Gene list one comprised of differentially expressed genes ranked based on the p-value from the significance of the DGE analysis without considering direction (termed no-direction). Gene list two consisted of the most differentially up-regulated genes (based on log-fold > 0 and p-value) at the top of the list and most differentially down-regulated genes at the bottom of the list (termed up-to-down). Gene list three consisted of the most differentially down-regulated (based on log-fold < 0 and p-value) at the top of the list and the most differentially up-regulated at the bottom of the list (termed down-to-up). The second and third lists are inverted copies of each other.

As large lists of GO terms can be produced from an analysis such as this, the software "*GO-Figure!"* was used and implemented using Python to reduce the GO terms to a more simplified list for easier interpretation. This is achieved by grouping terms based on semantic similarity. Scatterplots are generated which group together terms with similar functions and are coloured by significance of the representative GO term (log10 p-value) and the size indicates

the number of GO terms in the group (Reijnders and Waterhouse 2021). Plots were generated from lists for no-direction, up-to-down and down-to-up GO categories (as previous paragraph) across the three gene ontologies (Biological Process, Molecular Function, Cellular Component). This resulted in nine plots per regression, so 45 in total. As there was a large overlap between the pathways, only 18 are presented in this chapter and the remaining are in the appendix. The 18 presented are from the analysis using the binomial Braak score (0,1,2,3 vs 4,5,6), the binomial CERAD score (1,2 vs 3,4) and the case-control analysis.

4.2.5 Results from MAGMA pathway analysis based on genetic data and their significance in gene ontology enrichment analysis using gene expression data

The largest case-control GWAS also published a list of significant GO terms as a result of their pathway analysis using MAGMA (Kunkle et al. 2019). MAGMA performs a SNP-wise gene analysis of summary statistics (with LD correction) to test whether sets of genes are jointly associated with a phenotype compared to other genes across the genome (de Leeuw et al. 2015). The authors produced two lists of GO terms within their table 3. One consisting of FDR significant results from an analysis using only common variants (MAF $\geq$ 0.01) and another set of results using only rare variants (MAF < 0.01) (Kunkle et al. 2019). FDR significant results from their common variant analysis were taken from their published table 3 and those selected GO terms were checked to identify if they were significant GO terms using gene expression data in the AD case-control analysis described in the previous section of this thesis.

4.3 Results

4.3.1 QC and production of the ROSMAP dataset

After QC, the ROSMAP dataset consisted of 16,960 genes and 369 samples. Sample demographics can be seen in Table 4-2. In order to determine the number of PCs to include in any model to account for confounding, a scree plot was generated (Figure 4-1). From inspection of the scree plot, it was determined that the first four principal components from

the multivariate data would be included in the LMEM + LR, DESeq2 and limma-voom models to account for confounding.

	AD Cases	AD Controls	All samples
Sex	F: 141 M: 63 (69.1% F)	F: 101 M: 64 (61.2% F)	F: 242 M: 127 (65.6% F)
Age at death (years)	Mean: 88.0 SD: 3.6	Mean: 84.4 SD: 5.5	Mean: 86.4 SD: 4.9
Total	204	165	369

Table 4-2 – Sample demographics for the ROSMAP only dataset

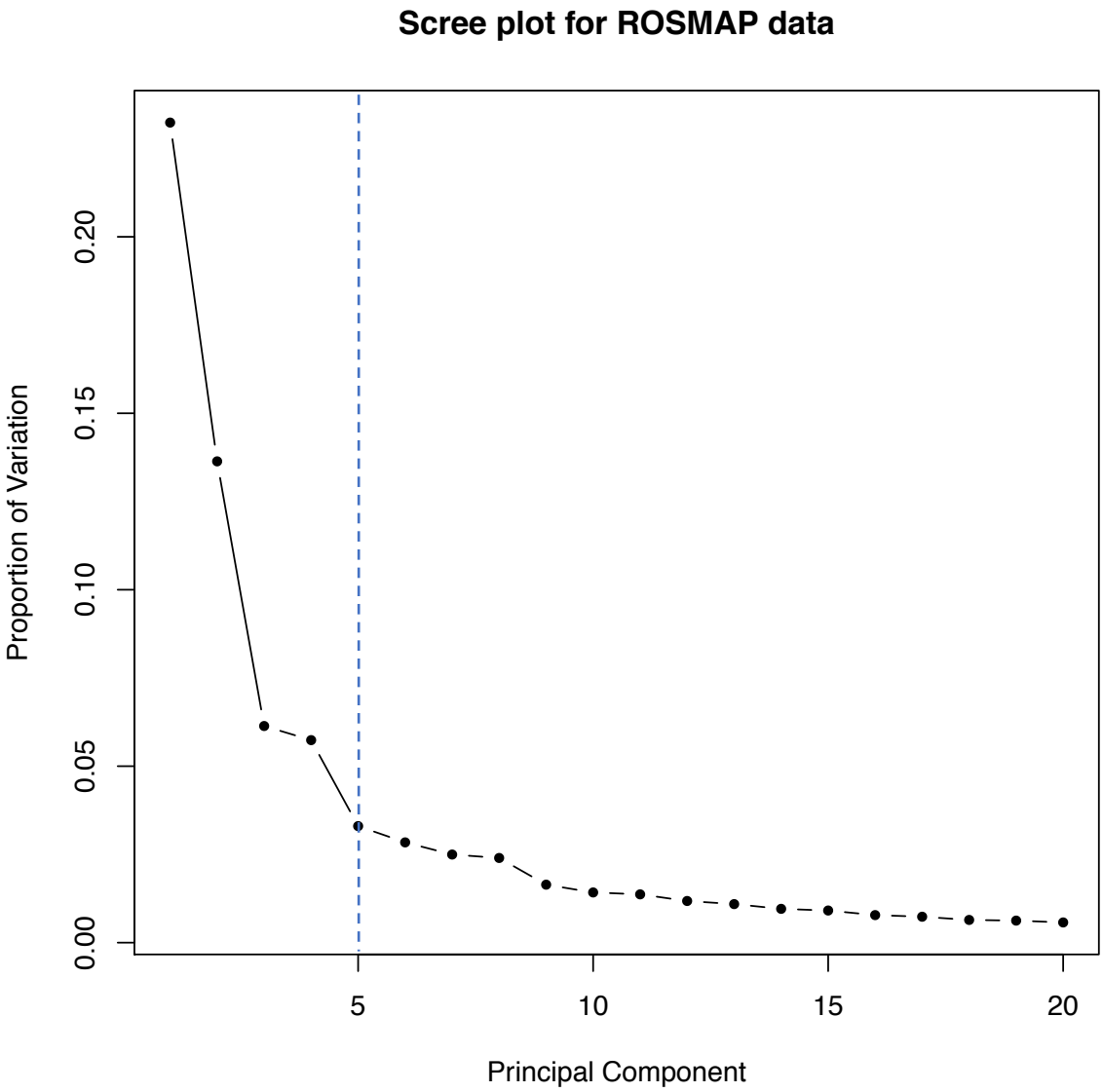


Figure 4-1 – A scree plot of the ROSMAP data to determine the number of principal components to include in the linear mixed effect model to account for hidden confounding. First four principal components were included as at the fifth principal component showed a levelling off in the proportion of variation explained. The beginning of this is indicated by the dashed blue line and all PCs included in the model are left of this line.

4.3.2 Overlap of differentially expressed genes identified using LMEM vs limma-voom and DESeq2 using ROSMAP data

To identify how many differentially expressed genes identified by the LMEM + LR method overlapped with those identified by more conventional DGE methods, the *DESeq2* and *Limma-Voom* packages were utilised. Case-control status was the phenotype of interest with

AD controls coded as 0 and AD cases coded as 1. The DGE analysis was performed on the ROSMAP dataset only, which was described in the previous section.

DESeq2 determined 1054 differentially expressed genes after FDR (<0.05) correction with 3693 genes nominally significant (p-value < 0.05). Limma-Voom determined 1260 differentially expressed genes after FDR (<0.05) correction with 3943 genes nominally significant. The LMEM followed by logistic regression identified 352 differentially expressed genes after FDR (<0.05) correction with 3182 nominally significant. 265 of the 352 FDR significant differentially expressed genes identified by the LMEM + logistic regression method were also identified as FDR significant by both DESeq2 and Limma-Voom, while all 352 were FDR significant with Limma-Voom alone (Figure 4-2b). For LMEM + logistic regression, only 25 nominally significant differentially expressed genes did not have a consensus with one of the other tools (0.05%) as can be seen in the red portion of the Venn diagram in Figure 4-2a.

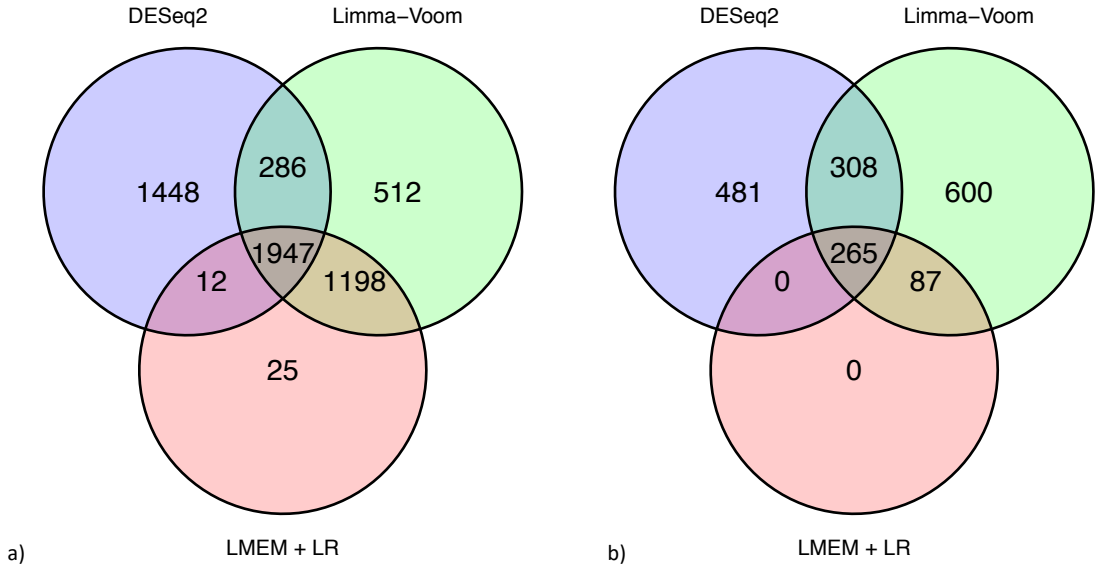
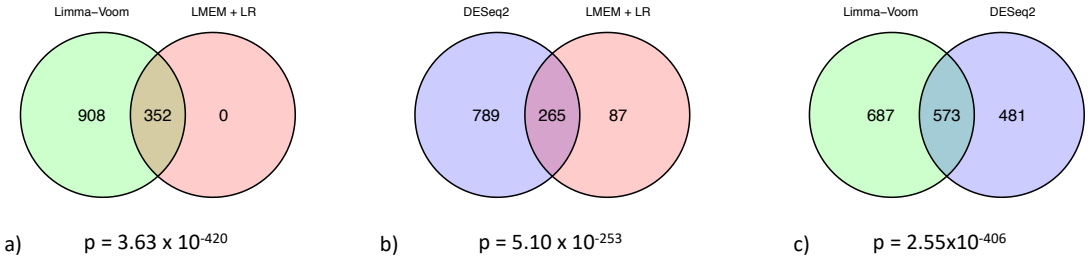


Figure 4-2 a) The overlap of nominally significant (p-value < 0.05) differentially expressed genes (DEGs) between three methods of limma-voom, DESeq2 and linear mixed-effect models + logistic regression (LMEM + LR); b) The overlap of FDR (<0.05) significant DEGs between the three methods of limma-voom, DESeq2 and LMEM + LR.

DESeq2 is in blue, limma-voom is in green and LMEM + LR is in red.

Pairwise hypergeometric tests were performed to determine if the overlaps of FDR significant differentially expressed genes between methods were more than what one would expect by chance. Comparisons of overlap were performed for: FDR-significant genes (Figure 4-3), FDR-significant genes which were up-regulated (Figure 4-4), FDR-significant genes which were down-regulated (Figure 4-5). All overlaps were more than expected through chance alone indicating that all three methods were finding similar differentially expressed genes. To confirm that the direction of effect was consistent between methods, a down-regulated vs up-regulated comparison was performed and then an up-regulated vs down-regulated comparison. All comparisons showed no overlap of differentially expressed genes meaning that the three methods were reporting genes with a consistent direction of effect.



a) $p = 3.63 \times 10^{-420}$ b) $p = 5.10 \times 10^{-253}$ c) $p = 2.55 \times 10^{-406}$

Figure 4-3 – Pairwise overlap of FDR (< 0.05) significant differentially expressed genes (DEG) between the three methods.

a) shows the overlap of DEG between Limma-Voom and LMEM + LR; b) shows the overlap of DEG between DESeq2 and LMEM + LR; c) shows the overlap of DEG between Limma-Voom and DESeq2. P-values resulting from hypergeometric test of gene overlap. Limma-Voom in green, LMEM + LR in red and DESeq2 in blue.

a) $p = 3.26 \times 10^{-322}$ b) $p = 6.73 \times 10^{-197}$ c) $p = 1.63 \times 10^{-336}$

Figure 4-4 - Pairwise overlap of FDR (< 0.05) significant up-regulated differentially expressed genes (DEG) between the three methods.

a) shows the overlap of up-regulated DEG between Limma-Voom and LMEM + LR; b) shows the overlap of up-regulated DEG between DESeq2 and LMEM + LR; c) shows the overlap of up-regulated DEG between Limma-Voom and DESeq2. P-values resulting from hypergeometric test of up-regulated gene overlap. Limma-Voom in green, LMEM + LR in red and DESeq2 in blue.

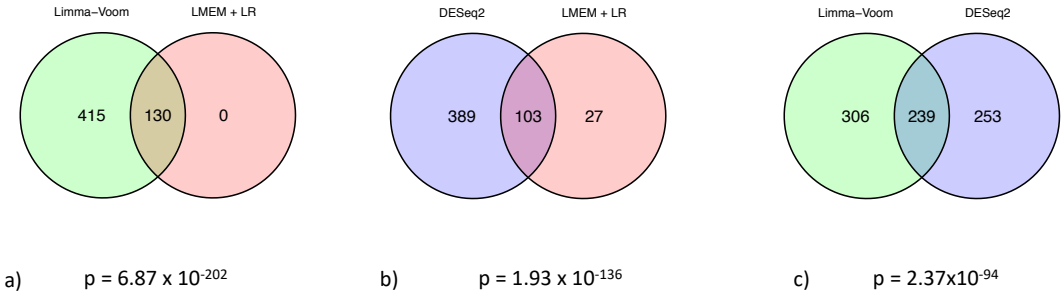


Figure 4-5 - Pairwise overlap of FDR (< 0.05) significant down-regulated differentially expressed genes (DEG) between the three methods.

a) shows the overlap of down-regulated DEG between Limma-Voom and LMEM + LR; b) shows the overlap of down-regulated DEG between DESeq2 and LMEM + LR; c) shows the overlap of down-regulated DEG between Limma-Voom and DESeq2. P-values resulting from hypergeometric test of down-regulated gene overlap. Limma-Voom in green, LMEM + LR in red and DESeq2 in blue.

QQ plots for the three methods were generated to compare the distribution of p-values generated from three different DGE analysis tools to the null model (Figure 4-6). As can be seen, all methods have an excess of small p-values compared to the null model as represented by the blue lines on each plot. The QQ plot for the LMEM + LR method (Figure 4-6c) provides evidence that the p-values for the LMEM + LR are more conservative in comparison to the other two tools (Figure 4-6a and b).

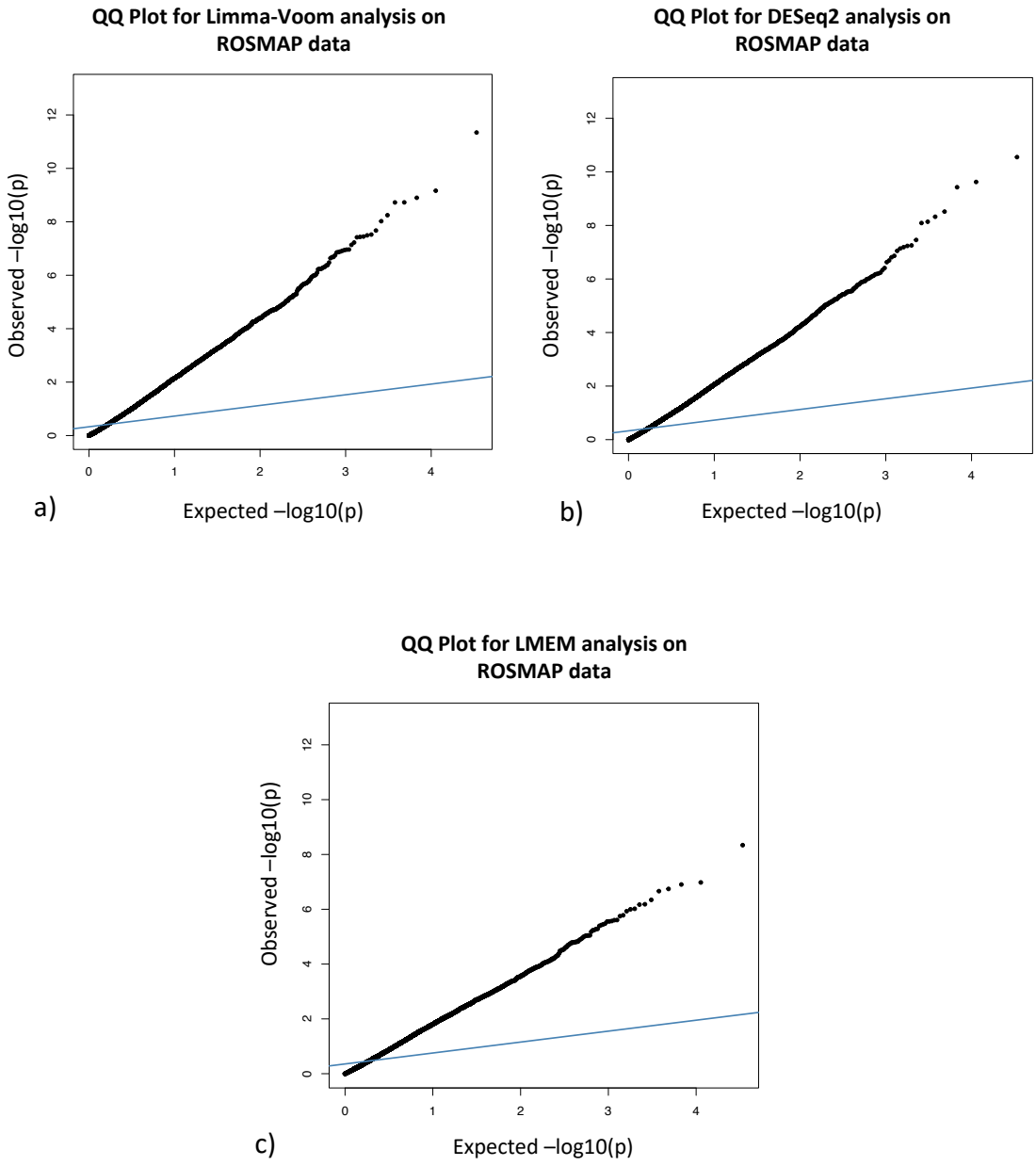


Figure 4-6 – QQ plots of p-values from the three differential gene expression analyses utilising a) limma-voom, b)DESeq2 and c) linear mixed-effect models (LMEM) + logistic regression. Each point represents the p-value (log-scale) from a test. Expected values are plotted on the x-axis and observed values on the y-axis. The blue lines in each plot are created using the function 'qqline' within R. 'qqline' adds a theoretical line for the expected values using a normal distribution.

The overlap of differentially expressed genes between methods, and the QQ plots provide evidence that the LMEM + LR method may be more conservative but can identify differentially

expressed genes with a consensus to another two tools. Therefore, for future experiments the LMEM + LR method was used for DGE analysis as the approach also allows the use of samples from the MSBB and MayoRNASeq studies.

4.3.3 Differential expression and GO enrichment analysis of Braak data

Using normalised residuals from LMEM, three regression models were performed to determine differentially expressed genes. The first, a logistic regression with Braak scores of 0, 1, 2, 3 vs 4, 5, 6 (coded 0 vs 1) respectively, resulted in 2196 significant (FDR < 0.05) differentially expressed genes. The second, also a logistic regression but with only a subset of the data contrasting Braak scores of 0, 1, 2 vs 5, 6 (coded 0 vs 1) resulted in 514 significant (FDR < 0.05) differentially expressed genes. The third, an ordinal regression including all Braak scores (0-6), resulted in 2049 significant (FDR < 0.05) differentially expressed genes. The top 10 differentially expressed genes from these three analyses, can be seen in Table 4-3, Table 4-4 and Table 4-5. Seven genes appeared in the top ten results from at least two analyses (*LPO*, *LINC01844*, *OCRL*, *ANKRD18DP*, *CBX5*, *NCDN*, *KCNK9*).

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
OCRL	X:129,539,849-129,592,561	0.58	$1.32x10^{-10}$	$1.09x10^{-06}$
LINC01844	5:142,716,229-142,761,035	-0.61	$7.86x10^{-11}$	$1.09x10^{-06}$
NCDN	1:35,557,473-35,567,274	0.57	$2.22x10^{-09}$	$5.47x10^{-06}$
NCOA1	2:24,491,254-24,770,702	0.56	$2.32x10^{-09}$	$5.47x10^{-06}$
CREB3L1	11:46,277,662-46,321,409	-0.54	$2.07x10^{-09}$	$5.47x10^{-06}$
KCNK9	8:139,600,838-139,704,109	0.52	$1.89x10^{-09}$	$5.47x10^{-06}$
DRD1	5:175,440,036-175,444,182	0.60	$1.08x10^{-09}$	$5.47x10^{-06}$
PAFAH1B3	19:42,297,033-42,303,546	0.54	$2.99x10^{-09}$	$5.48x10^{-06}$
CRH	8:66,176,376-66,178,464	-0.55	$2.77x10^{-09}$	$5.48x10^{-06}$
ANKRD18DP	3:198,053,522-198,080,737	-0.51	$4.13x10^{-09}$	$6.80x10^{-06}$

Table 4-3 - Top 10 differentially expressed genes from LMEM model including 3PCs after logistic regression with Braak scores 0, 1, 2, 3 vs 4, 5, 6 (coded 0 vs 1). *Genes were ranked based on their FDR corrected p-value and the top 10 most significant are reported with their beta coefficient from regression analyses and each gene's p-value and FDR corrected p-value (FDR >0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
CIC	19:42,268,537-42,295,797	0.63425239	$1.88x10^{-07}$	$1.65x10^{-03}$
EPB41L1	20:36,091,504-36,232,799	0.60735661	$3.99x10^{-07}$	$1.65x10^{-03}$
LINC01844	5:142,716,229-142,761,035	-0.6063251	$3.66x10^{-07}$	$1.65x10^{-03}$
PEG13	8:140,094,894-140,100,543	0.59688507	$3.93x10^{-07}$	$1.65x10^{-03}$
CBX5	12:54,230,942-54,280,133	0.59533403	$7.99x10^{-07}$	$2.64x10^{-03}$
OCRL	X:129,539,849-129,592,561	0.55043237	$2.02x10^{-06}$	$3.03x10^{-03}$
ATP6V1C1	8:103,021,063–103,073,051	0.56635713	$1.71x10^{-06}$	$3.03x10^{-03}$
LPO	17:58,218,548-58,268,518	-0.4996428	$1.95x10^{-06}$	$3.03x10^{-03}$
MECP2	X:154,021,573-154,137,103	0.52359314	$1.47x10^{-06}$	$3.03x10^{-03}$
KCNK9	8:139,600,838-139,704,109	0.56733674	$1.26x10^{-06}$	$3.03x10^{-03}$

Table 4-4 - Top 10 differentially expressed genes from LMEM model including 3PCs after logistic regression on reduced data with Braak scores 0, 1, 2 vs 5, 6 (coded 0 vs 1). *Genes were ranked based on their FDR corrected p-value and the top 10 most significant are reported with their beta coefficient from regression analyses and each gene's p-value and FDR corrected p-value (FDR < 0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
LPO	17:58,218,548-58,268,518	-0.50	$8.26x10^{-11}$	$6.81x10^{-07}$
LINC01844	5:142,716,229-142,761,035	-0.50	$5.63x10^{-11}$	$6.81x10^{-07}$
OCRL	X:129,539,849-129,592,561	0.47	$1.43x10^{-10}$	$7.87x10^{-07}$
ANKRD18DP	3:198,053,522-198,080,737	-0.44	$3.12x10^{-10}$	$1.28x10^{-06}$
VPS26B	11:134,224,671-134,247,788	0.45	$7.17x10^{-10}$	$2.37x10^{-06}$
CBX5	12:54,230,942-54,280,133	0.46	$1.29x10^{-09}$	$3.03x10^{-06}$
POU6F1	12:51,186,936-51,218,062	0.45	$1.10x10^{-09}$	$3.03x10^{-06}$
SCAMP5	15:74,957,219-75,021,495	0.48	$1.81x10^{-09}$	$3.74x10^{-06}$
NCDN	1:35,557,473-35,567,274	0.45	$4.42x10^{-09}$	$8.10x10^{-06}$
COL17A1	10:104,031,286-104,085,880	-0.41	$5.03x10^{-09}$	$8.25x10^{-06}$

Table 4-5 Top 10 differentially expressed genes from LMEM model including 3PCs after ordinal regression with all Braak scores 0 – 6 included. *Genes were ranked based on their FDR corrected p-value and reported with their beta coefficient from regression analyses and each gene's p-value and FDR corrected p-value (FDR >0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

A list of prioritised genes from the largest AD case-control GWAS was taken from figure 2 as published in their paper (Kunkle et al. 2019). In this paper, the prioritisation of genes was based on eight criteria. These were: (1) deleterious coding, loss of function or splicing variant in the gene; (2) significant gene-based tests; (3) expression in an AD relevant tissue; (4) a microglial-enriched gene; (5) having an eQTL effect; (6) being involved in a biological pathway

enriched in AD; (7) expression correlated with Braak stage; (8) DGE evidence in an AD case-control study.

I then checked to see if these genes were differentially expressed in the three Braak differential expression analyses. The full list of prioritised genes from the Kunkle et al. paper and results from my DGE analysis can be seen in Table 4-6. Only one gene was found to be significantly differentially expressed (FDR < 0.05) across all three regression analyses. This gene was *CLU* with p-values $2.88x10^{-03}$, $1.07x10^{-03}$, and $4.49x10^{-04}$ and FDR corrected p-values of 0.03, 0.04 and 0.01 for the Braak logistic, reduced Braak logistic and ordinal models respectively. The direction of effect was consistent suggesting that *CLU* is upregulated in cases in comparison to controls.

PSMB9 and *AGFG2* were both FDR significant in the Braak logistic and Braak ordinal regressions. They were only nominally significant in the reduced Braak logistic model. *PSMC5* was FDR significant in the Braak logistic and nominally significant in the Braak ordinal and reduced Braak logistic models. *ACP2*, *FAM131B*, *PILRA*, *WDR18*, and *YOD1* were nominally significant in at least one of the models. Full results can be seen in Table 4-6 where results that were at least nominally significant have been marked in bold.

The Kunkle et al. GWAS prioritised genes were used as a gene set and a one-sided Wilcoxon rank sum test used to see if these Kunkle et al. GWAS prioritised genes ranked higher in the DGE analysis than expected by chance for each of the three regression analyses. This analysis was non-directional and did not specify nor differentiate between up or down regulation. None of the three tests for the Braak score logistic regression (p-value: 0.83), reduced Braak score logistic regression (p-value: 0.82), nor the ordinal Braak score regression (p-value: 0.84) were statistically significant. This indicates that the Kunkle et al. GWAS prioritised genes were not enriched in the Braak score differential gene expression analysis. Boxplots demonstrating the difference in p-value between GWAS prioritised and non-prioritised genes can be seen in **Figure 4-7**.

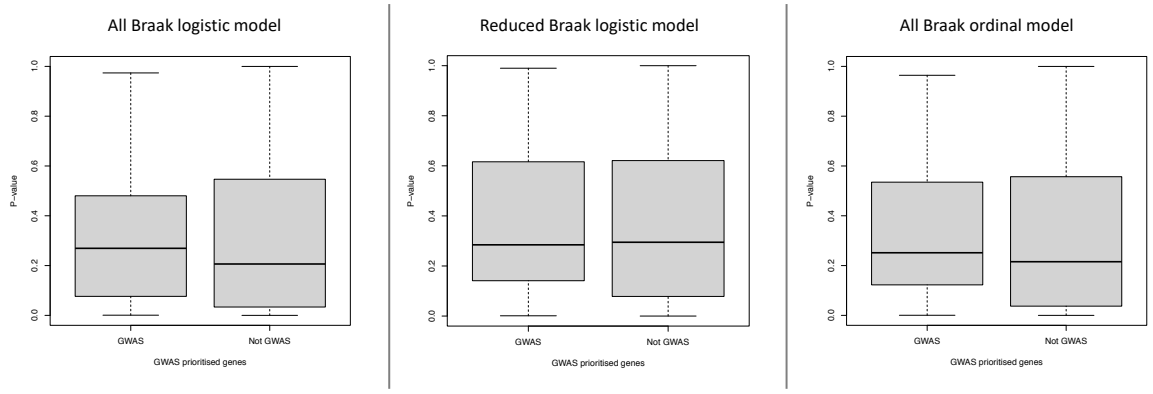


Figure 4-7 - Boxplots demonstrating the differences in p-value resulting from the AD Braak score gene expression analyses. 'GWAS' refers to the GWAS prioritised genes and 'Not GWAS' refers to any gene not in this set

	Chr:start-end	Braak score Logistic regression			Reduced Braak score logistic regression			Braak score ordinal regression		
		Beta	p-value	FDR p-value	Beta	p-value	FDR p-value	Beta	p-value	FDR p-value
ABCA7	19:1,039,997-1,065,572	0.17	0.09	0.24	0.13	0.29	0.58	0.14	0.13	0.31
ACP2	11:47,239,302-47,248,906	-0.20	0.05	0.17	-0.19	0.14	0.42	-0.18	0.05	0.18
ADAM10	15:58,588,809-58,749,791	0.14	0.11	0.28	0.07	0.48	0.74	0.10	0.16	0.36
ADAMTS1	21:26,835,755-26,845,409	0.07	0.55	0.73	0.01	0.91	0.97	0.04	0.71	0.84
AGFG2	**7:100,539,203-100,568,220**	**0.30**	**$4.69x10^{-03}$**	**0.04**	**0.34**	**0.02**	**0.15**	**0.26**	**$5.27x10^{-03}$**	**0.05**
BIN1	2:127,048,027-127,107,288	-0.12	0.17	0.36	-0.14	0.20	0.50	-0.10	0.19	0.40
C1QTNF4	11:47,589,667-47,594,411	0.007	0.92	0.96	0.003	0.97	0.99	0.04	0.57	0.75
C4A	6:31,982,052-32,002,681	0.10	0.38	0.59	0.18	0.18	0.47	0.12	0.26	0.48

C7orf43	7:100,154,420-100,158,723	0.05	0.48	0.68	0.08	0.40	0.68	0.07	0.32	0.54
CASS4	20:56,412,112-56,460,387	0.11	0.22	0.43	0.18	0.14	0.41	0.13	0.12	0.31
CD2AP	6:47,477,789-47,627,263	0.16	0.07	0.20	0.15	0.20	0.50	0.13	0.08	0.25
CD55	1:207,321,519-207,386,804	0.01	0.91	0.96	-0.08	0.54	0.78	-0.006	0.94	0.97
CELF1	**11:47,465,933-47,565,569**	**0.25**	**$4.89x10^{-03}$**	**0.04**	0.17	0.16	0.45	**0.18**	**0.02**	**0.11**
CLU	**8:27,596,917-27,614,700**	**0.28**	**$2.88x10^{-03}$**	**0.03**	**0.39**	**$1.07x10^{-03}$**	**0.04**	**0.29**	**$4.49x10^{-04}$**	**0.01**
CNN2	19:1,026,586-1,039,068	-0.11	0.22	0.43	-0.12	0.28	0.58	-0.07	0.38	0.60
CR1	1:207,496,147-207,641,765	-0.03	0.78	0.88	-0.02	0.84	0.94	-0.02	0.82	0.91
ECHDC3	10:11,742,366-11,764,070	0.06	0.48	0.68	0.07	0.58	0.81	0.06	0.46	0.67
EED	11:86,244,753-86,278,813	0.003	0.97	0.99	-0.02	0.89	0.96	-0.02	0.78	0.88

EPHB4	7:100,802,565-100,827,523	-0.006	0.94	0.97	0.04	0.75	0.89	0.03	0.73	0.85
FAM131B	**7:143,353,400-143,362,770**	**0.21**	**0.02**	**0.08**	**0.26**	**0.03**	**0.19**	**0.17**	**0.02**	**0.10**
GAL3ST4	7:100,159,244-100,168,617	-0.12	0.18	0.38	-0.02	0.89	0.96	-0.09	0.31	0.54
GPSM3	6:32,190,766-32,195,523	0.08	0.30	0.52	0.05	0.59	0.81	0.05	0.52	0.72
HLA-DPA1	6:33,064,569-33,080,775	0.03	0.76	0.87	0.04	0.74	0.89	0.01	0.91	0.96
HLA-DQA1	6:32,628,179-32,647,062	0.13	0.20	0.41	0.14	0.27	0.57	0.11	0.23	0.45
HLA-DRB1	6:32,577,902-32,589,848	0.08	0.42	0.63	0.11	0.42	0.70	0.06	0.49	0.69
HLA-DRB5	6:32,517,353-32,530,287	0.03	0.79	0.89	0.01	0.93	0.97	0.005	0.96	0.98
HMHA1	19:1,065,923-1,086,628	0.10	0.29	0.51	0.06	0.65	0.84	0.04	0.59	0.76
INPP5D	2:233,059,967-233,207,903	0.19	0.06	0.19	0.22	0.09	0.34	0.15	0.10	0.27

IQCK	16:19,716,456-19,858,467	0.10	0.23	0.44	0.13	0.24	0.54	0.07	0.30	0.52
MAF	16:79,585,843-79,600,737	0.10	0.30	0.52	0.22	0.09	0.34	0.11	0.19	0.40
MS4A4	11:60,185,657-60,318,080	0.13	0.20	0.41	0.08	0.55	0.78	0.06	0.52	0.72
MS4A6A	11:60,172,015-60,184,666	0.11	0.30	0.52	0.08	0.57	0.80	0.06	0.52	0.72
MS4A7	11:60,378,485-60,395,951	0.20	0.06	0.18	0.17	0.21	0.51	0.14	0.13	0.33
MTCH2	11:47,617,315-47,642,607	-0.10	0.33	0.54	-0.06	0.64	0.84	-0.09	0.32	0.55
NDUFS3	11:47,565,336-47,584,562	-0.04	0.62	0.78	-0.13	0.23	0.53	-0.09	0.24	0.47
NUP160	11:47,778,087-47,848,555	0.10	0.26	0.48	0.07	0.55	0.79	0.10	0.18	0.39
PICALM	11:85,957,175-86,069,882	0.00	0.96	0.98	-0.02	0.87	0.95	-0.01	0.86	0.93
PILRA	**7:100,367,530-100,400,096**	**0.20**	**0.03**	**0.13**	**0.27**	**0.03**	**0.19**	**0.18**	**0.03**	**0.13**

PSMB8	6:32,840,717-32,844,679	-0.11	0.19	0.40	-0.16	0.15	0.43	-0.12	0.13	0.32
PSMB9	**6:32,844,136-32,859,851**	**-0.27**	**$1.18x10^{-03}$**	**0.02**	**-0.35**	**$1.76x10^{-03}$**	**0.05**	**-0.26**	**$4.21x10^{-04}$**	**$9.44x10^{-03}$**
PSMC3	11:47,418,769-47,426,473	-0.08	0.36	0.57	-0.17	0.15	0.42	-0.10	0.19	0.39
PSMC5	**17:63,827,152-63,832,026**	**0.28**	**$7.24x10^{-04}$**	**0.01**	**0.21**	**0.04**	**0.23**	**0.20**	**$8.21x10^{-03}$**	**0.06**
PTK2B	8:27,311,482-27,459,391	0.12	0.22	0.42	-0.001	0.99	1.00	0.05	0.55	0.73
RIN3	14:92,513,781-92,688,994	0.05	0.57	0.75	0.06	0.60	0.81	0.03	0.68	0.83
SORL1	11:121,452,314-121,633,763	-0.05	0.56	0.74	0.02	0.86	0.95	-0.05	0.55	0.74
SPI1	11:47,354,860-47,378,547	0.09	0.33	0.55	0.11	0.36	0.65	0.05	0.51	0.71
STYX	14:52,730,166-52,774,989	-0.09	0.28	0.50	-0.08	0.44	0.71	-0.08	0.28	0.50
TREM2	6:41,158,506-41,163,186	0.16	0.13	0.31	0.20	0.13	0.40	0.14	0.15	0.35

WDR18	19:984,332-998,438	-0.17	0.06	0.19	-0.25	0.03	0.20	-0.20	0.01	0.07
WWOX	16:78,099,400-79,212,667	0.09	0.30	0.51	0.18	0.12	0.39	0.11	0.15	0.35
YOD1	1:207,043,849-207,052,980	0.21	0.02	0.09	0.24	0.04	0.22	0.19	0.02	0.09
ZKSCAN1	7:100,015,572-100,041,689	-0.12	0.27	0.49	-0.16	0.23	0.53	-0.14	0.14	0.34

Table 4-6 – Results from the Braak score differential gene expression analysis for top-prioritised genes from the largest AD case-control GWAS (Kunkle et al. 2019) Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).

4.3.4 GO enrichment analysis of AMP-AD Braak data

GO enrichment analysis on the Braak score logistic regression differentially expressed genes resulted in 1026 statistically significant GO categories that are up-regulated and 153 statistically significant GO categories that are down-regulated. 27 were significant in the analysis that did not define direction.

The analysis on the reduced logistic regression differentially expressed genes resulted in 1045 statistically significant GO categories that are up-regulated and 95 statistically significant GO categories that are down-regulated. 31 were significant in the non-directional analysis.

Additionally, the analysis on the ordinal regression differentially expressed genes resulted in 1015 statistically significant GO categories that are up-regulated and 147 statistically significant GO categories that are down-regulated. 38 were significant in the non-directional analysis. For all, statistically significant refers to an FDR-corrected p-value of less than 0.05 and the categories include biological process, molecular function and cellular component.

The python package *GO-Figure!* was used to reduce these large lists of GO terms to a summarised list of terms based on semantic similarity to make them easier to comprehend (Reijnders and Waterhouse 2021). This was performed across the three GO categories of biological process, molecular function and cellular component. GO terms have been reduced based on semantic similarity using GoFigure! (Reijnders and Waterhouse 2021). Colour refers to log10 p-value, size of circle refers to number of GO terms clustered (larger the circle, more GO terms clustered) and only the top 20 similar terms are labelled on scatterplots.

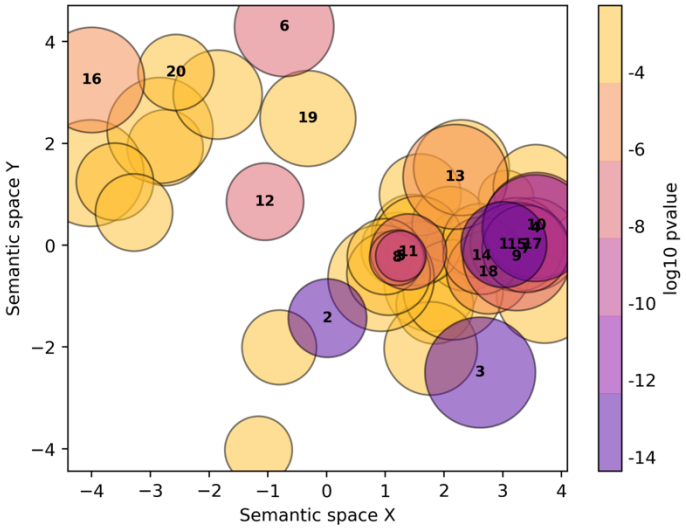
Across all three regression analyses, enriched biological process GO terms related to cell-signalling and response to stimulus and they were up-regulated as demonstrated in an example plot from the logistic regression analysis in Figure 4-9. Enriched GO categories such as SRP-dependent cotranslational protein targeting to membrane, viral transcription,

apoptosis and mitochondrial pathways were found to be down-regulated as can be seen in an example in Figure 4-10.

Molecular function GO terms that were enriched for up-regulated differentially expressed genes were related to protein and transcription factor binding and transporter activity (Figure 4-12) whereas ribosome and catalytic activity were GO terms that were enriched for down-regulated differentially expressed genes (Figure 4-13).

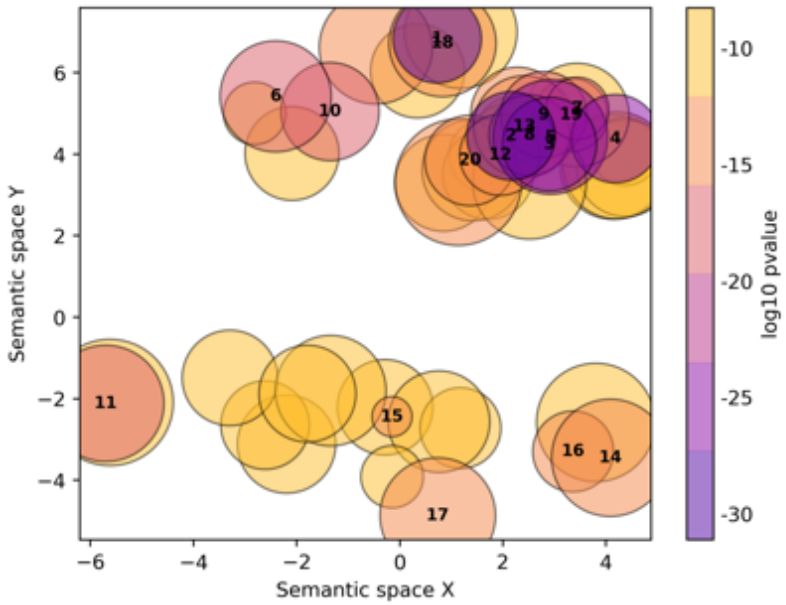
Cellular component GO terms that were enriched for up-regulated differentially expressed genes mainly related to the synapses and neurons (Figure 4-15) whereas down-regulated GO terms were related to the mitochondria, ribosome and endoplasmic reticulum (Figure 4-16).

Amongst all the plots, the cellular component GO term plots showed the clearest clustering and segregation (Figure 4-14 - Figure 4-16). The software works by grouping GO terms together based on semantic similarity or in other words based on terms with similar functions. The GO terms that were enriched from this analysis were identifying clear clusters of cellular components with distinct functions whereas for molecular function and biological process these were less distinct.



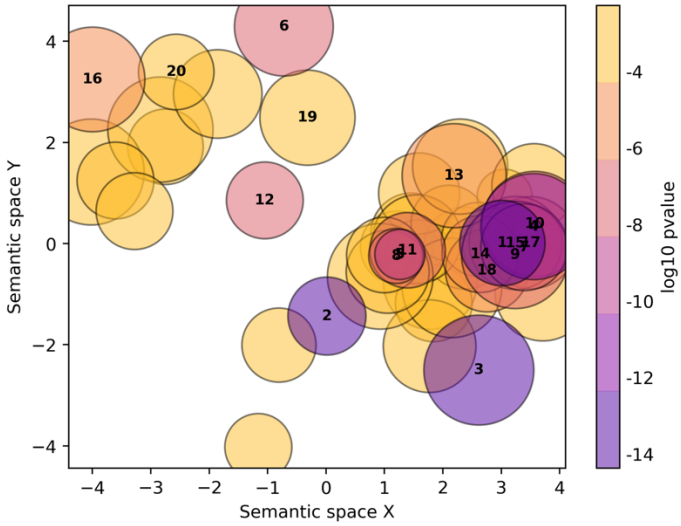
1. amide biosynthetic process
2. viral transcription
3. SRP-dependent cotranslational protein targeting to membrane
4. carboxylic acid catabolic process
5. translational initiation
6. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
7. organic cyclic compound catabolic process
8. drug metabolic process
9. ncRNA metabolic process
10. fatty acid beta-oxidation
11. respiratory electron transport chain
12. hormone-mediated apoptotic signaling pathway
13. mitochondrial respiratory chain complex I assembly
14. translational elongation
15. alpha-amino acid metabolic process
16. positive regulation of glucocorticoid secretion
17. acetyl-CoA metabolic process
18. peptidyl-glutamic acid modification
19. negative regulation of mitochondrial membrane permeability involved in apoptotic process
20. positive regulation of somatostatin secretion

Figure 4-8 Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score logistic regression (0,1,2,3 vs 4,5,6).



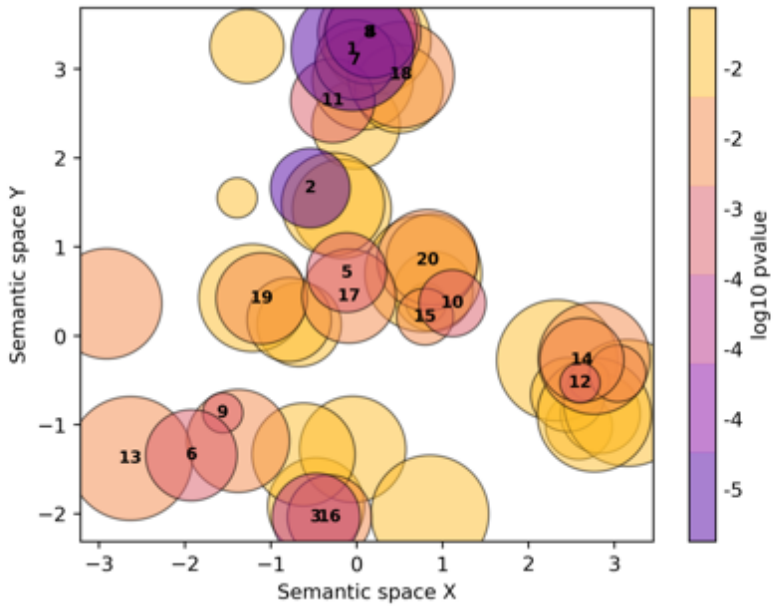
1. positive regulation of biological process
2. signal transduction
3. regulation of signaling
4. regulation of localization
5. regulation of cell communication
6. regulation of multicellular organismal process
7. negative regulation of cellular process
8. regulation of biological quality
9. regulation of response to stimulus
10. regulation of developmental process
11. developmental process
12. regulation of molecular function
13. regulation of cellular component organization
14. cellular response to organic substance
15. cellular process
16. cellular response to stimulus
17. cellular protein modification process
18. positive regulation of RNA metabolic process
19. negative regulation of nitrogen compound metabolic process
20. regulation of transcription by RNA polymerase II

Figure 4-9 Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)



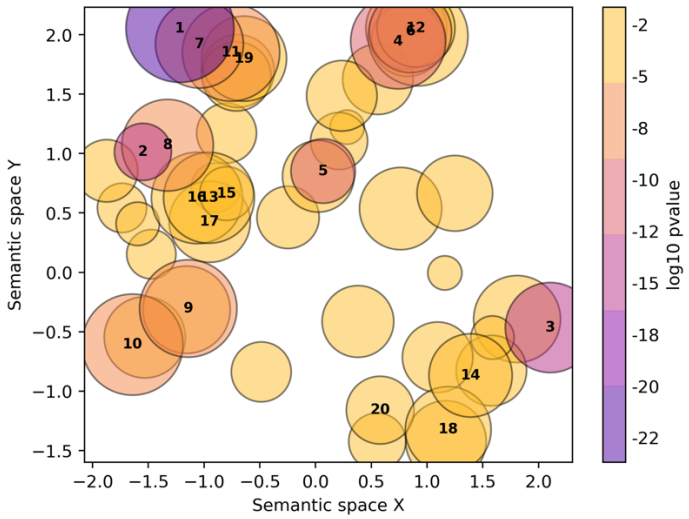
1. amide biosynthetic process
2. viral transcription
3. SRP-dependent cotranslational protein targeting to membrane
4. carboxylic acid catabolic process
5. translational initiation
6. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
7. organic cyclic compound catabolic process
8. drug metabolic process
9. ncRNA metabolic process
10. fatty acid beta-oxidation
11. respiratory electron transport chain
12. hormone-mediated apoptotic signaling pathway
13. mitochondrial respiratory chain complex I assembly
14. translational elongation
15. alpha-amino acid metabolic process
16. positive regulation of glucocorticoid secretion
17. acetyl-CoA metabolic process
18. peptidyl-glutamic acid modification
19. negative regulation of mitochondrial membrane permeability involved in apoptotic process
20. positive regulation of somatostatin secretion

Figure 4-10 Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)



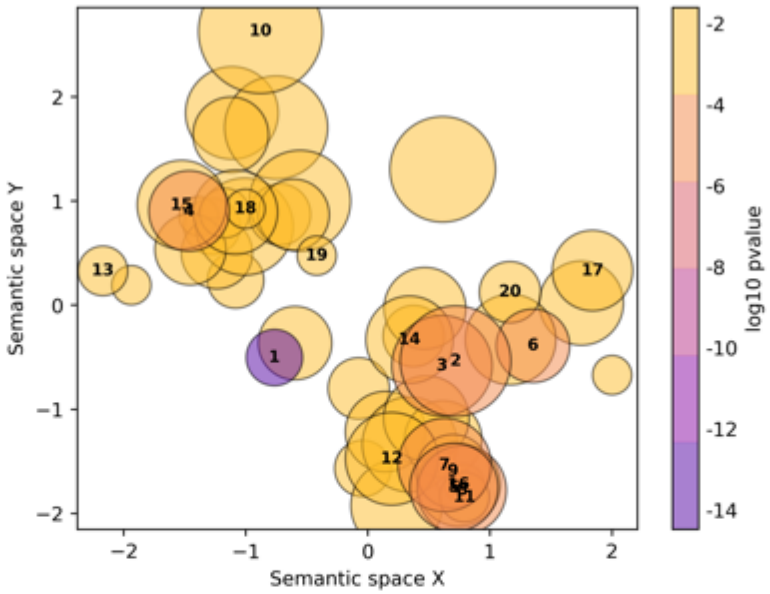
1. voltage-gated cation channel activity
2. cAMP-dependent protein kinase regulator activity
3. 3',5'-cyclic-GMP phosphodiesterase activity
4. cation channel activity
5. structural constituent of ribosome
6. protein C-terminal carboxyl O-methyltransferase activity
7. voltage-gated sodium channel activity
8. metal ion transmembrane transporter activity
9. aminomethyltransferase activity
10. cAMP binding
11. ion gated channel activity
12. phosphatidylinositol 3-kinase binding
13. protein serine/threonine kinase activity
14. myosin light chain binding
15. NADH binding
16. 3'-flap endonuclease activity
17. oxygen carrier activity
18. arginine transmembrane transporter activity
19. NAD(P)H oxidase H2O2-forming activity
20. neuropeptide binding

Figure 4-11 Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score logistic regression (0,1,2,3 vs 4,5,6).



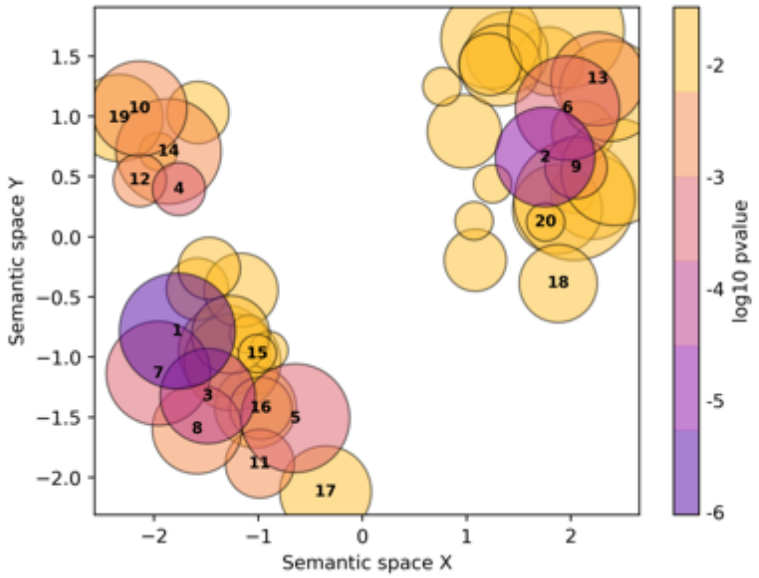
1. protein binding
2. binding
3. kinase activity
4. voltage-gated cation channel activity
5. DNA-binding transcription factor activity, RNA polymerase II-specific
6. potassium ion transmembrane transporter activity
7. transcription factor binding
8. RNA polymerase II transcription regulatory region sequence-specific DNA binding
9. molecular function regulator
10. molecular transducer activity
11. ephrin receptor binding
12. transporter activity
13. phospholipid binding
14. protein C-terminal carboxyl O-methyltransferase activity
15. chromatin binding
16. protein-containing complex binding
17. peptide binding
18. calcium-dependent protein serine/threonine phosphatase activity
19. adrenergic receptor binding
20. inositol-1,3,4,5-tetrakisphosphate 5-phosphatase activity

Figure 4-12 – Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)



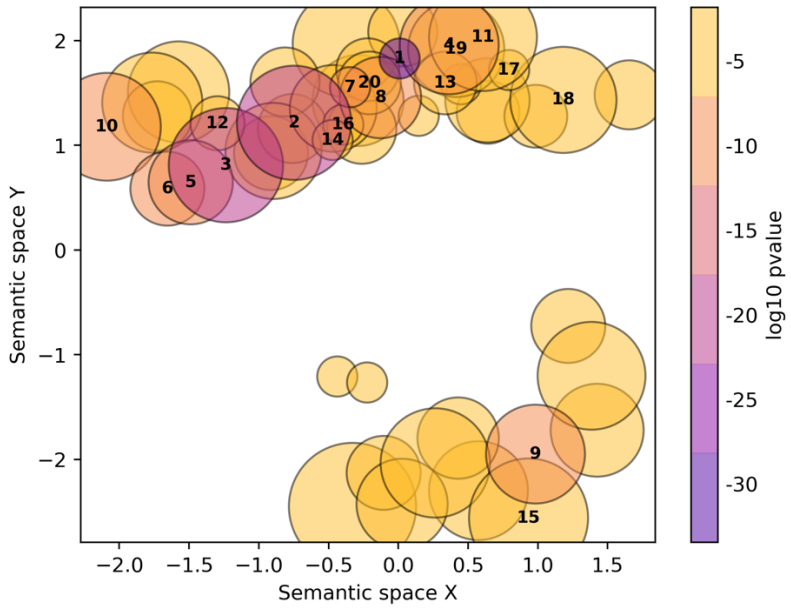
1. structural constituent of ribosome
2. catalytic activity, acting on a tRNA
3. ligase activity
4. flavin adenine dinucleotide binding
5. oxidoreductase activity, acting on the CH-CH group of donors
6. aminomethyltransferase activity
7. 3-hydroxyacyl-CoA dehydrogenase activity
8. oxidoreductase activity, acting on NAD(P)H
9. peroxidase activity
10. hormone activity
11. electron transfer activity
12. 3'-flap endonuclease activity
13. MutSalpha complex binding
14. enoyl-CoA hydratase activity
15. rRNA binding
16. oxidoreductase activity, acting on the aldehyde or oxo group of donors
17. acetyl-CoA C-acyltransferase activity
18. tetrapyrrole binding
19. ubiquinone binding
20. catechol O-methyltransferase activity

Figure 4-13 – Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)



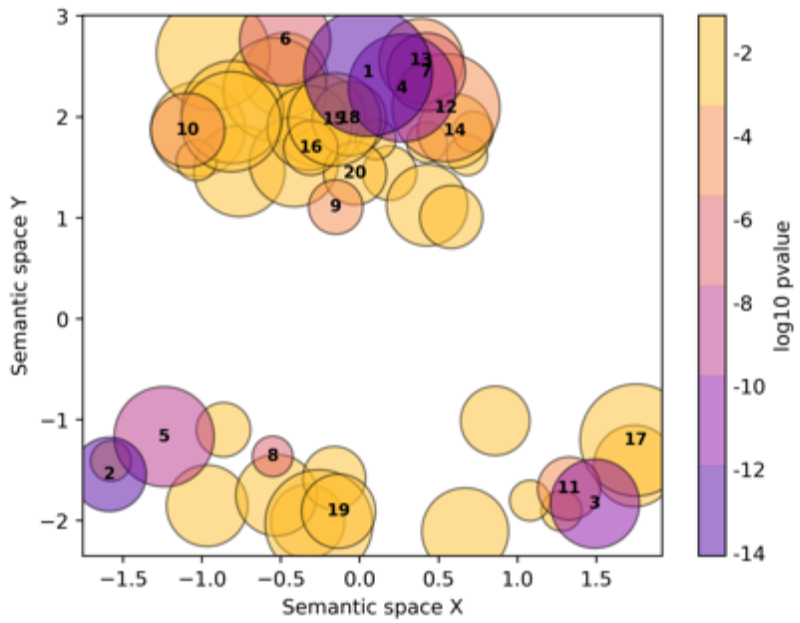
1. synapse
2. cytosolic large ribosomal subunit
3. integral component of postsynaptic density membrane
4. postsynaptic density
5. synaptic membrane
6. cation channel complex
7. cerebellar mossy fiber
8. apical dendrite
9. CERF complex
10. contractile fiber
11. cell projection
12. mitotic spindle
13. RNA polymerase II, holoenzyme
14. chromaffin granule
15. neuronal cell body
16. Golgi cis cisterna
17. mitochondrial inner membrane
18. mitochondrial intermembrane space protein transporter complex
19. ribosome
20. DNA ligase IV complex

Figure 4-14 Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score logistic regression (0,1,2,3 vs 4,5,6).



1. plasma membrane
2. synapse
3. neuron projection
4. plasma membrane region
5. intrinsic component of plasma membrane
6. integral component of postsynaptic density membrane
7. presynapse
8. chromatin
9. voltage-gated potassium channel complex
10. cytoplasmic vesicle
11. cytoplasmic vesicle membrane
12. postsynaptic density
13. cytosol
14. neuronal cell body
15. plasma membrane protein complex
16. postsynapse
17. intrinsic component of membrane
18. bounding membrane of organelle
19. dendrite membrane
20. neuron projection terminus

Figure 4-15 Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)



1. ribosomal subunit
2. mitochondrial inner membrane
3. mitochondrion
4. mitochondrial protein-containing complex
5. mitochondrial matrix
6. oxidoreductase complex
7. respiratory chain complex
8. respirasome
9. Lsm2-8 complex
10. methylosome
11. peroxisome
12. endoplasmic reticulum membrane protein complex
13. proton-transporting ATP synthase complex, coupling factor F(o)
14. mitochondrial proton-transporting ATP synthase complex
15. U2-type precatalytic spliceosome
16. Pwp2p-containing subcomplex of 90S preribosome
17. motile cilium
18. mitochondrial alpha-ketoglutarate dehydrogenase complex
19. Golgi cis cisterna
20. DNA recombinase mediator complex

Figure 4-16 – Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the Braak score logistic regression (0,1,2,3 vs 4,5,6)

4.3.5 Differential gene expression analysis of AMP-AD CERAD data

Using normalised residuals from the LMEM normalisation step as performed in chapter 3, three regression models were performed to determine differentially expressed genes. The first was a logistic regression with CERAD scores of 1, 2 vs 3, 4, (coded 0 vs 1) respectively and resulted in 76 significant differentially expressed genes (FDR <0.05). The second, also a logistic regression with only a subset of the data contrasting CERAD scores of 1 vs 4 (coded 0 vs 1) resulted in 31 significant (FDR < 0.05) differentially expressed genes. The third, an ordinal regression including all CERAD scores (1-4), resulted in 253 significant (FDR < 0.05) differentially expressed genes. The top 10 differentially expressed genes from these three analyses can be seen in Table 4-7, Table 4-8 and Table 4-9. Seven genes appeared in the top ten results from at least two analyses (*DNAJC19*, *LPO*, *DRD1*, *GCSH*, *TIMM8B*, *FAM19A2*, *KCTD8*).

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
DNAJC19	3:180,983,697-180,989,774	-0.46	$4.15x10^{-07}$	$6.83x10^{-03}$
LPO	17:58,218,548-58,268,518	-0.45	$1.21x10^{-06}$	$9.94x10^{-03}$
DRD1	5:175,440,036-175,444,182	0.51	$2.17x10^{-06}$	0.01
GIPC2	1:77,979,542-78,138,444	-0.42	$8.36x10^{-06}$	0.02
GCSH	16:81,081,945-81,096,395	-0.41	$5.97x10^{-06}$	0.02
TIMM8B	11:112,084,800-112,086,798	-0.40	$4.31x10^{-06}$	0.02
KCNK9	8:139,600,838-139,704,109	0.43	$1.01x10^{-05}$	0.02
IQGAP3	1:156,525,405-156,572,604	-0.47	$7.61x10^{-06}$	0.02
PABPC1L2A	X:73,077,276-73,079,512	0.38	$9.84x10^{-06}$	0.02
LINC01844	5:142,716,229-142,761,035	-0.45	$8.49x10^{-06}$	0.02

Table 4-7 - Top 10 differentially expressed genes from LMEM model including 3PCs after logistic regression with CERAD scores 1, 2 vs 3, 4 (coded 0 vs 1). *Genes were ranked based on their FDR corrected p-value and reported with their beta coefficient from regression analyses and each gene's FDR corrected p-value (FDR <0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
TIMM8B	11:112,084,800-112,086,798	-0.49	$2.59x10^{-06}$	0.02
FAM19A2	12:61,708,273-62,279,150	0.57	$3.09x10^{-06}$	0.02
DNAJC19	3:180,983,697-180,989,774	-0.53	$1.11x10^{-06}$	0.02
GCSH	16:81,081,945-81,096,395	-0.47	$9.41x10^{-06}$	0.02
LPO	17:58,218,548-58,268,518	-0.45	$9.23x10^{-06}$	0.02
KCTD8	4:44,173,903-44,448,809	0.50	$8.29x10^{-06}$	0.02
DRD1	5:175,440,036-175,444,182	0.59	$6.17x10^{-06}$	0.02
UBE2E1	3:23,805,955-23,891,640	-0.49	$1.15x10^{-05}$	0.02
UBE2K	4:39,698,109-39,782,792	0.48	$2.93x10^{-05}$	0.03
PPP2R5C	14:101,761,709-101,927,989	-0.52	$2.11x10^{-05}$	0.03

Table 4-8 - Top 10 differentially expressed genes from LMEM model including 3PCs after logistic regression with the reduced CERAD score dataset 1 vs 4 (coded 0 vs 1). *Genes were ranked based on their FDR corrected p-value and reported with their beta coefficient from regression analyses and each gene's FDR corrected p-value (FDR <0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
LPO	17:58,218,548-58,268,518	-0.42	$1.47x10^{-07}$	$2.42x10^{-03}$
EPB41L1	20:36,091,504-36,232,799	0.41	$1.58x10^{-06}$	$3.82x10^{-03}$
CBX5	12:54,230,942-54,208,133	0.40	$2.55x10^{-06}$	$3.82x10^{-03}$
GCSH	16:81,081,945-81,096,395	-0.35	$2.10x10^{-06}$	$3.82x10^{-03}$
TIMM8B	11:112,084,800-112,086,798	-0.35	$1.41x10^{-06}$	$3.82x10^{-03}$
COA4	11:73,872,667-73,876,901	-0.37	$1.87x10^{-06}$	$3.82x10^{-03}$
KCTD8	4:44,173,903-44,448,809	0.37	$2.32x10^{-06}$	$3.82x10^{-03}$
POU6F1	12:51,186,936-51,218,062	0.39	$1.87x10^{-06}$	$3.82x10^{-03}$
DRD1	5:175,440,036-175,444,182	0.43	$1.24x10^{-06}$	$3.82x10^{-03}$
FAM19A2	12:61,708,273-62,279,150	0.44	$8.44x10^{-07}$	$3.82x10^{-03}$

Table 4-9 - Top 10 differentially expressed genes from LMEM model including 3PCs after ordinal regression with CERAD scores 1, 2 vs 3, 4 (coded 0 vs 1. *Genes were ranked based on their FDR corrected p-value and reported with their beta coefficient from regression analyses and each gene's FDR corrected p-value (FDR < 0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Like the Braak score analysis, prioritised genes from Kunkle et al.'s AD GWAS were used, and their differential expression were investigated in the three CERAD score regressions. No differentially expressed genes survived multiple hypothesis testing correction. *CELF1*, *FAM131B*, *PSMB8*, and *PSMB9* were nominally significant across all three CERAD score regression DGE analyses. The direction of effect was consistent for both *CELF1* and *FAM131B*,

which was overexpressed in cases in comparison to controls. Both were at least nominally significant in the previous Braak logistic analysis. The direction of effect for *PSMB8* and *PSMB9* was consistent for both, which was down-regulated in cases in comparison to controls. *PSMB9* was significantly differentially expressed in the Braak score analysis but *PSMB8* was not.

The following genes were nominally significant in at least one of the CERAD analyses: *ADAM10*, *C4A*, *C7ORF43*, *MTCH2*, *PSMC5*, *WDR18*. *PSMC5* and *WDR18* were also found to be at least nominally significant in at least one of the previous Braak score analyses. Results from the Braak score analyses can be seen in Table 4-6 and CERAD score results can be seen in Table 4-10. Results that were at least nominally significant are marked in bold.

The Kunkle et al. GWAS prioritised genes were then used as a gene set and a one-sided Wilcoxon rank sum test used to see if these GWAS genes were differentially expressed more than we would expect by chance for each of the three regression analyses. The tests for the CERAD score logistic regression (p-value = 0.996), reduced CERAD score logistic regression (p-value: 0.969) and the ordinal CERAD score regression (p-value = 0.986) were not statistically significant. This suggests that when considering the rank of the differentially expressed genes, there is no evidence for GWAS-prioritised genes being enriched in this DGE analysis of CERAD scores. Boxplots demonstrating the difference in p-value between GWAS prioritised and non-prioritised genes can be seen in Figure 4-17. In general, Kunkle et al. GWAS prioritised genes were less significant than non-prioritised genes.

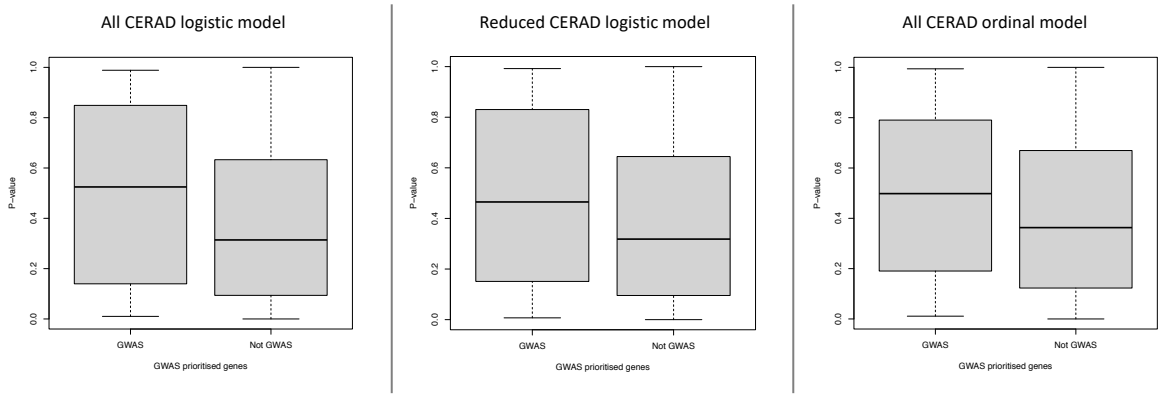


Figure 4-17 - Boxplots demonstrating the differences in p-value resulting from the AD CERAD score gene expression analyses. 'GWAS' refers to the GWAS prioritised genes and 'Not GWAS' refers to any gene not in this set.

Gene	Chr:start-end	CERAD score Logistic regression			Reduced CERAD score logistic regression			CERAD score ordinal regression		
		Beta	p-value	FDR p-value	Beta	p-value	FDR p-value	Beta	p-value	FDR p-value
ABCA7	19:1,039,997-1,065,572	0.11	0.32	0.63	0.11	0.39	0.74	0.12	0.21	0.54
ACP2	11:47,239,302-47,248,906	-0.21	0.06	0.31	-0.16	0.22	0.61	-0.12	0.21	0.54
ADAM10	**15:58,588,809-58,749,791**	**0.23**	**0.03**	**0.23**	0.17	0.14	0.51	0.17	0.05	0.28
ADAMTS1	21:26,835,755-26,845,409	-0.02	0.88	0.96	0.01	0.95	0.99	0.02	0.88	0.96
AGFG2	7:100,539,203-100,568,220	0.14	0.23	0.54	0.17	0.18	0.57	0.16	0.12	0.42
BIN1	2:127,048,027-127,107,288	0.07	0.48	0.75	0.06	0.62	0.87	0.01	0.87	0.96
C1QTNF4	11:47,589,667-47,594,411	0.05	0.56	0.80	0.02	0.79	0.94	0.01	0.91	0.97
C4A	6:31,982,052-32,002,681	0.01	0.93	0.98	0.02	0.90	0.97	0.01	0.90	0.97

C7ORF43	**7:100,154,420-100,158,723**	**0.17**	**0.04**	**0.26**	0.19	0.06	0.37	0.13	0.07	0.32
CASS4	20:56,412,112-56,460,387	0.05	0.63	0.84	0.09	0.44	0.77	0.08	0.36	0.67
CD2AP	6:47,477,789-47,627,263	0.03	0.78	0.92	0.06	0.60	0.86	0.08	0.33	0.65
CD55	1:207,321,519-207,386,804	-0.002	0.98	0.99	-0.01	0.92	0.98	0.01	0.93	0.98
CELF1	**11:47,465,933-47,565,569**	**0.22**	**0.03**	**0.24**	**0.30**	**0.01**	**0.21**	**0.23**	**0.01**	**0.13**
CLU	8:27,596,917-27,614,700	0.07	0.51	0.77	0.16	0.18	0.56	0.10	0.28	0.61
CNN2	19:1,026,586-1,039,068	-0.07	0.49	0.76	-0.05	0.63	0.87	-0.03	0.74	0.91
CR1	1:207,496,147-207,641,765	-0.04	0.73	0.89	-0.03	0.77	0.93	0.00	0.99	1.00
ECHDC3	10:11,742,366-11,764,070	-0.06	0.58	0.82	-0.02	0.85	0.96	-0.04	0.68	0.88
EED	11:86,244,753-86,278,813	0.11	0.25	0.57	-0.02	0.88	0.97	-0.02	0.81	0.94

EPHB4	7:100,802,565-100,827,523	-0.08	0.40	0.70	-0.05	0.67	0.89	-0.04	0.67	0.87
FAM131B	7:143,353,400-143,362,770	0.20	0.04	0.25	0.24	0.04	0.33	0.21	0.01	0.16
GAL3ST4	7:100,159,244-100,168,617	-0.02	0.87	0.95	-0.05	0.67	0.89	-0.04	0.64	0.86
GPSM3	6:32,190,766-32,195,523	0.07	0.37	0.67	0.07	0.44	0.77	0.05	0.47	0.75
HLA-DPA1	6:33,064,569-33,080,775	-0.04	0.70	0.88	-0.06	0.64	0.88	-0.03	0.73	0.90
HLA-DQA1	6:32,628,179-32,647,062	0.02	0.85	0.95	0.07	0.56	0.84	0.06	0.56	0.81
HLA-DRB1	6:32,577,902-32,589,848	0.01	0.91	0.97	0.03	0.80	0.94	0.02	0.80	0.93
HLA-DRB5	6:32,517,353-32,530,287	-0.01	0.95	0.98	-0.01	0.94	0.98	-0.01	0.91	0.98
HMHA1	19:1,065,923-1,086,628	-0.02	0.83	0.94	-0.03	0.81	0.94	-0.01	0.91	0.98
INPP5D	2:233,059,967-233,207,903	0.11	0.34	0.65	0.17	0.18	0.57	0.14	0.16	0.48

IQCK	16:19,716,456-19,858,467	0.02	0.85	0.95	0.04	0.71	0.91	0.04	0.62	0.85
MAF	16:79,585,843-79,600,737	0.16	0.14	0.44	0.12	0.35	0.71	0.09	0.33	0.65
MS4A4	11:60,185,657-60,318,080	-0.03	0.77	0.91	-0.02	0.87	0.96	-0.02	0.84	0.95
MS4A6A	11:60,172,015-60,184,666	0.002	0.99	1.00	0.02	0.86	0.96	0.01	0.89	0.97
MS4A7	11:60,378,485-60,395,951	0.14	0.24	0.56	0.17	0.20	0.59	0.15	0.15	0.47
MTCH2	11:47,617,315-47,642,607	-0.16	0.14	0.44	-0.15	0.23	0.61	-0.10	0.28	0.61
NDUFS3	11:47,565,336-47,584,562	-0.16	0.08	0.36	-0.18	0.08	0.43	-0.15	0.06	0.32
NUP160	11:47,778,087-47,848,555	0.04	0.71	0.88	0.12	0.30	0.68	0.08	0.34	0.66
PICALM	11:85,957,175-86,069,882	0.08	0.42	0.71	0.14	0.28	0.66	0.10	0.27	0.59
PILRA	7:100,367,530-100,400,096	0.06	0.54	0.79	0.14	0.24	0.63	0.13	0.14	0.45

PSMB8	**6:32,840,717-32,844,679**	**-0.21**	**0.03**	**0.23**	**-0.23**	**0.04**	**0.33**	**-0.17**	**0.04**	**0.25**
PSMB9	**6:32,844,136-32,859,851**	**-0.24**	**0.01**	**0.16**	**-0.25**	**0.01**	**0.23**	**-0.20**	**0.01**	**0.15**
PSMC3	**11:47,418,769-47,426,473**	**-0.20**	**0.04**	**0.25**	-0.13	0.26	0.65	-0.09	0.26	0.59
PSMC5	17:63,827,152-63,832,026	0.13	0.15	0.45	0.19	0.07	0.40	**0.18**	**0.02**	**0.20**
PTK2B	8:27,311,482-27,459,391	0.01	0.89	0.96	0.05	0.67	0.89	0.05	0.58	0.83
RIN3	14:92,513,781-92,688,994	0.01	0.94	0.98	0.02	0.84	0.95	0.002	0.98	1.00
SORL1	11:121,452,314-121,633,763	-0.05	0.64	0.85	-0.05	0.67	0.89	-0.04	0.66	0.87
SPI1	11:47,354,860-47,378,547	0.01	0.91	0.97	0.03	0.79	0.94	0.01	0.96	0.99
STYX	14:52,730,166-52,774,989	-0.002	0.98	0.99	-0.09	0.40	0.75	-0.06	0.46	0.75
TREM2	6:41,158,506-41,163,186	0.21	0.08	0.35	0.18	0.17	0.55	0.16	0.11	0.41
WDR18	19:984,332-998,438	-0.17	0.09	0.37	-0.25	0.03	0.32	-0.16	0.06	0.30

WWOX	16:78,099,400-79,212,667	0.02	0.85	0.94	0.0008	0.99	1.00	0.02	0.82	0.94
YOD1	1:207,043,849-207,052,980	0.20	0.06	0.30	0.13	0.25	0.64	0.12	0.15	0.47
ZKSCAN1	7:100,015,572-100,041,689	-0.03	0.83	0.94	0.05	0.75	0.92	0.01	0.95	0.99

Table 4-10 - Results from the CERAD score differential gene expression analysis for top-prioritised genes from the largest AD case-control GWAS (Kunkle et al. 2019). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).

4.3.6 GO enrichment analysis of AMP-AD CERAD data

GO enrichment analysis on the CERAD score logistic regression differentially expressed genes resulted in 337 statistically significant GO categories that are enriched for up-regulated genes and 331 statistically significant GO categories that are enriched for down-regulated genes. 57 GO terms were significant in the non-directional analysis.

The analysis on the reduced logistic regression differentially expressed genes resulted in 348 statistically significant GO categories that are enriched for up-regulated differentially expressed genes and 387 statistically significant GO categories that are enriched for down-regulated genes. 102 GO terms were significant in the non-directional analysis.

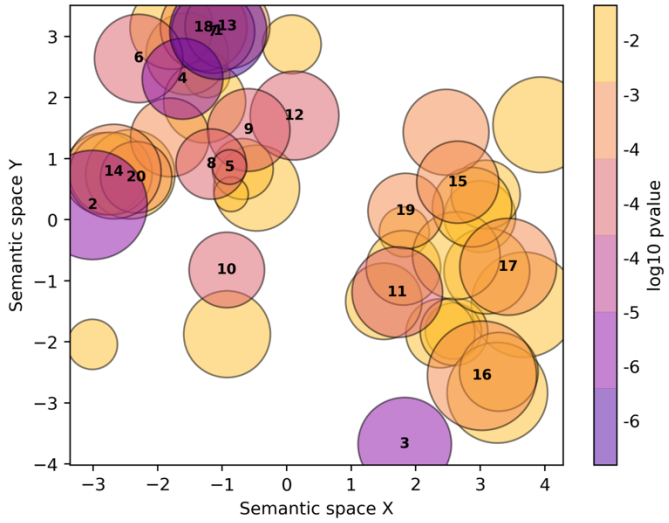
Additionally, the analysis on the ordinal regression differentially expressed genes resulted in 398 statistically significant GO categories that are enriched for up-regulated genes and 365 statistically significant GO categories that are enriched for down-regulated genes. 84 GO terms were significant in the non-directional analysis. For all, statistically significant refers to an FDR-corrected p-value of less than 0.05 and the categories include biological process, molecular function and cellular component.

The python package *GO-Figure!* was used to reduce these large lists of GO terms to a summarised list of terms across the three GO categories of biological process, molecular function and cellular component (Reijnders and Waterhouse 2021).

Across all three regression analyses, biological process GO terms that were enriched for up-regulated differentially expressed genes related to synaptic processes and transcription (Figure 4-19). GO categories enriched for down-regulated genes included terms such as SRP-dependent cotranslational protein targeting to membrane, viral transcription, and mitochondrial pathways (Figure 4-20). This is very similar to what was seen for the Braak phenotype.

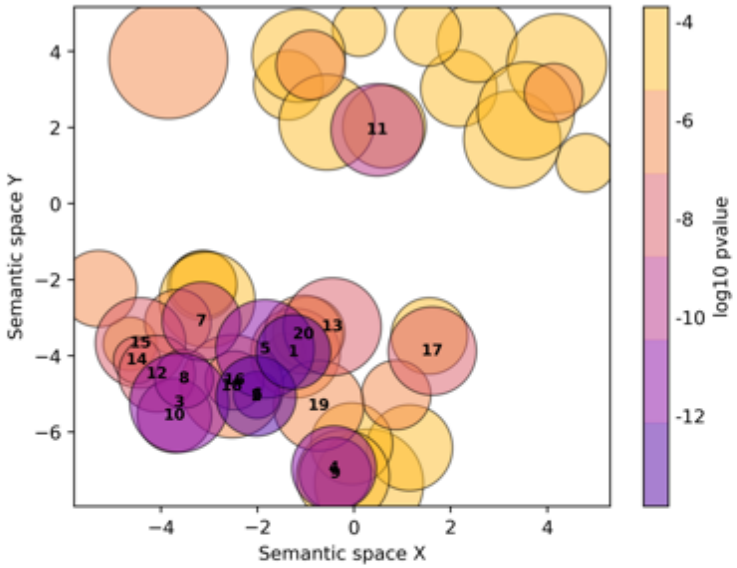
Molecular function GO terms that were enriched for up-regulated differentially expressed genes were related to glycine binding, DNA binding, transcription and transporter activity (Figure 4-21) whereas ribosome and catalytic activity were GO terms that were down-regulated (Figure 4-22) both of which are quite similar to the Braak phenotype.

Cellular components GO terms that were up-regulated mainly related to the synapses and neurons and plasma membrane (Figure 4-25) whereas down-regulated GO terms were relating to the mitochondria, ribosome and endoplasmic reticulum (Figure 4-26). Again, these results for the CERAD phenotype were similar to the Braak phenotype.



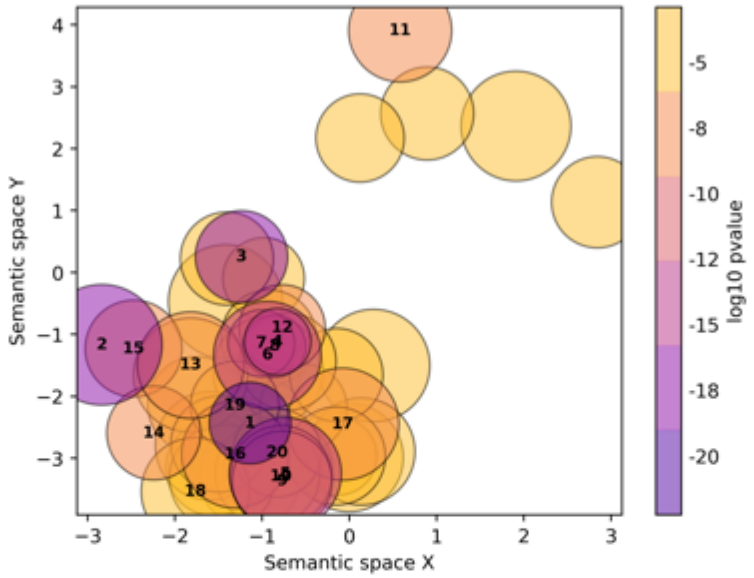
1. monocarboxylic acid catabolic process
2. cotranslational protein targeting to membrane
3. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
4. amide biosynthetic process
5. translational initiation
6. mitochondrial ATP synthesis coupled proton transport
7. glycine decarboxylation via glycine cleavage system
8. respiratory electron transport chain
9. lipid modification
10. viral transcription
11. membrane depolarization during action potential
12. calcium ion-regulated exocytosis of neurotransmitter
13. organic cyclic compound catabolic process
14. mitochondrial respiratory chain complex assembly
15. regulation of receptor localization to synapse
16. positive regulation of substrate adhesion-dependent cell spreading
17. regulation of synaptic plasticity
18. acetyl-CoA catabolic process
19. nerve growth factor signaling pathway
20. neurofilament cytoskeleton organization

Figure 4-18 Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD score logistic regression (1,2 vs 3,4)



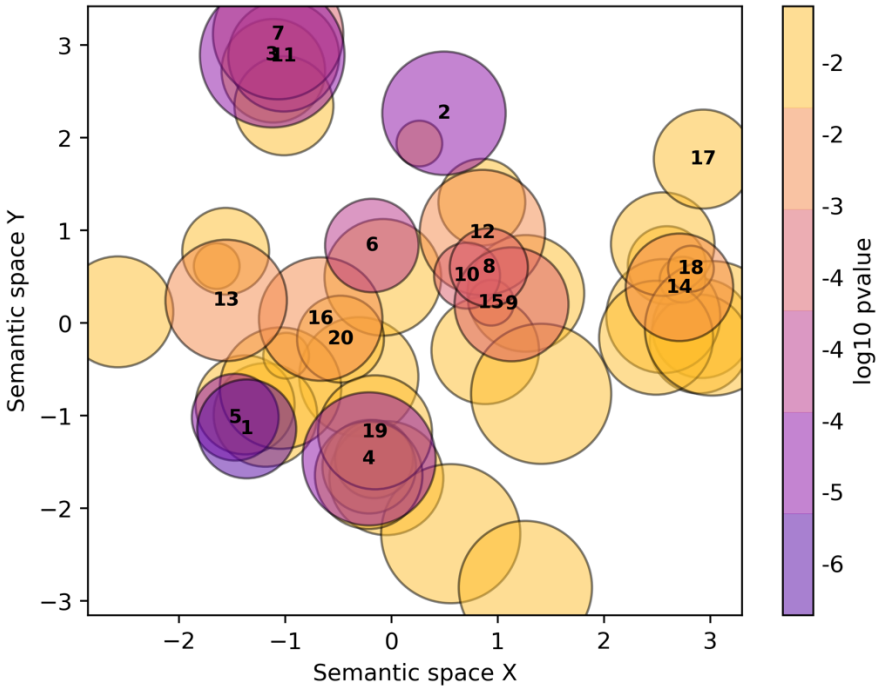
1. regulation of transcription by RNA polymerase II
2. regulation of RNA metabolic process
3. negative regulation of transcription, DNA-templated
4. positive regulation of RNA metabolic process
5. modulation of chemical synaptic transmission
6. regulation of cellular macromolecule biosynthetic process
7. regulation of plasma membrane bounded cell projection organization
8. regulation of nervous system development
9. positive regulation of cellular process
10. negative regulation of cellular metabolic process
11. chemical synaptic transmission
12. regulation of localization
13. regulation of membrane potential
14. regulation of ion transmembrane transport
15. regulation of vesicle-mediated transport
16. regulation of signaling
17. regulation of neuron differentiation
18. regulation of cell communication
19. regulation of cellular component organization
20. regulation of cellular component biogenesis

Figure 4-19 Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)



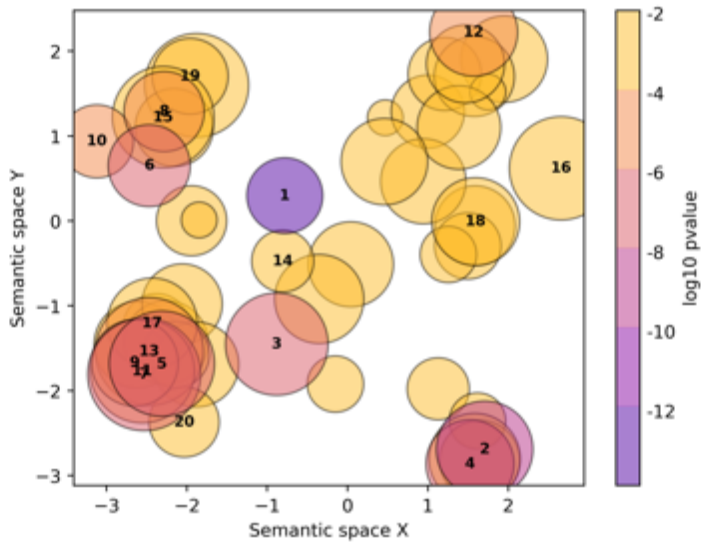
1. cellular amide metabolic process
2. SRP-dependent cotranslational protein targeting to membrane
3. viral transcription
4. drug metabolic process
5. monocarboxylic acid catabolic process
6. electron transport chain
7. metabolic process
8. translational initiation
9. aromatic compound catabolic process
10. organic cyclic compound catabolic process
11. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
12. cellular detoxification
13. sulfur compound metabolic process
14. mitochondrial ATP synthesis coupled proton transport
15. cristae formation
16. alpha-amino acid metabolic process
17. mitochondrial respiratory chain complex I assembly
18. ATP synthesis coupled proton transport
19. translational elongation
20. branched-chain amino acid catabolic process

Figure 4-20 Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)



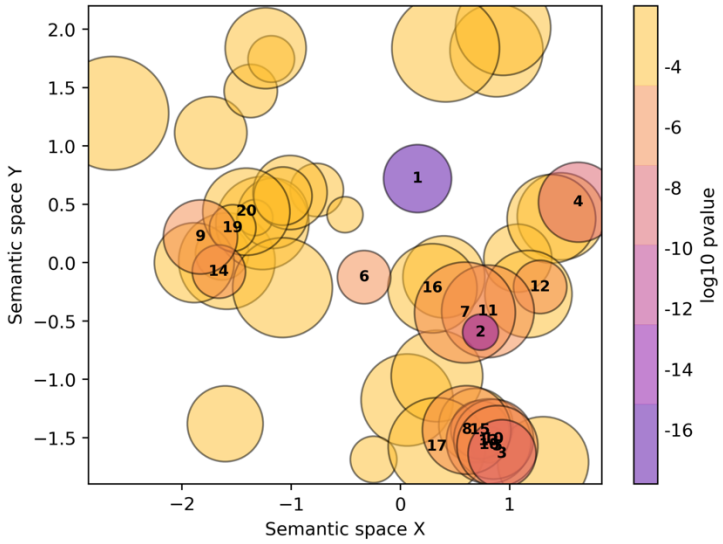
1. acetyl-CoA C-acyltransferase activity
2. cAMP-dependent protein kinase regulator activity
3. voltage-gated cation channel activity
4. 3-hydroxyacyl-CoA dehydrogenase activity
5. aminomethyltransferase activity
6. structural constituent of ribosome
7. cation channel activity
8. cAMP binding
9. large ribosomal subunit rRNA binding
10. NADH binding
11. voltage-gated sodium channel activity
12. glycine binding
13. lysozyme activity
14. myosin light chain binding
15. DNA binding, bending
16. enoyl-CoA hydratase activity
17. opioid peptide activity
18. protein C-terminus binding
19. 2,4-dienoyl-CoA reductase (NADPH) activity
20. methylcrotonoyl-CoA carboxylase activity

Figure 4-21 Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD score logistic regression (1,2 vs 3,4)



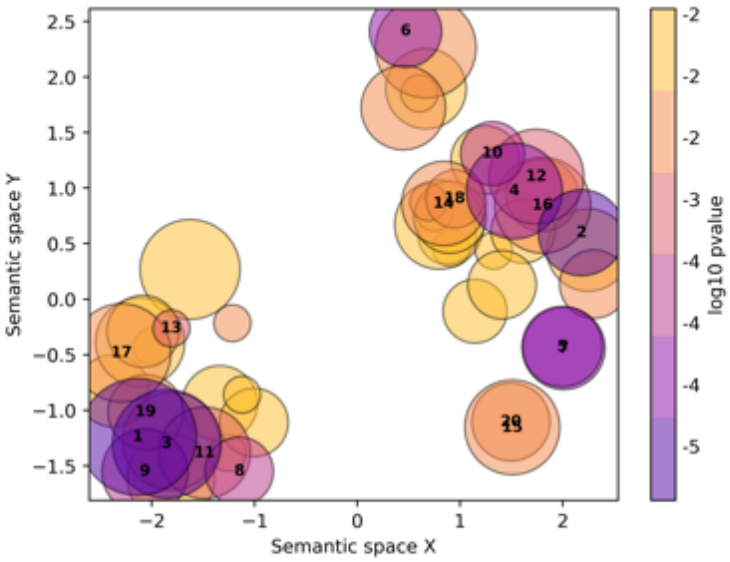
1. transcription regulator activity
2. voltage-gated cation channel activity
3. RNA polymerase II transcription regulatory region sequence-specific DNA binding
4. potassium channel activity
5. protein domain specific binding
6. ion gated channel activity
7. cytoskeletal protein binding
8. transcription coregulator activity
9. transcription factor binding
10. metal ion transmembrane transporter activity
11. enzyme binding
12. kinase activity
13. ephrin receptor binding
14. chromatin binding
15. GTPase activator activity
16. protein serine/threonine phosphatase activity
17. alpha-2A adrenergic receptor binding
18. lysozyme activity
19. G protein-coupled neurotransmitter receptor activity
20. opioid peptide activity

Figure 4-22 Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)



1. structural constituent of ribosome
2. catalytic activity
3. electron transfer activity
4. acetyl-CoA C-acyltransferase activity
5. oxidoreductase activity, acting on NAD(P)H
6. antioxidant activity
7. ligase activity
8. 3-hydroxyacyl-CoA dehydrogenase activity
9. flavin adenine dinucleotide binding
10. oxidoreductase activity, acting on the aldehyde or oxo group of donors
11. catalytic activity, acting on RNA
12. aminomethyltransferase activity
13. oxidoreductase activity, acting on the CH-CH group of donors
14. large ribosomal subunit rRNA binding
15. oxidoreductase activity, acting on peroxide as acceptor
16. enoyl-CoA hydratase activity
17. helicase activity
18. oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
19. NAD binding
20. vitamin binding

Figure 4-23 Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)



1. ribosomal subunit
2. mitochondrial inner membrane
3. mitochondrial protein-containing complex
4. mitochondrial matrix
5. integral component of postsynaptic density membrane
6. ribosome
7. integral component of synaptic membrane
8. mitochondrial proton-transporting ATP synthase complex
9. proton-transporting ATP synthase complex, coupling factor F(o)
10. neurofilament
11. sodium channel complex
12. glutamatergic synapse
13. apolipoprotein B mRNA editing enzyme complex
14. terminal cisterna
15. cerebellar mossy fiber
16. synaptic membrane
17. NuRD complex
18. chromosome, centromeric region
19. glycine cleavage complex
20. basal dendrite

Figure 4-24 Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD score logistic regression (1,2 vs 3,4)

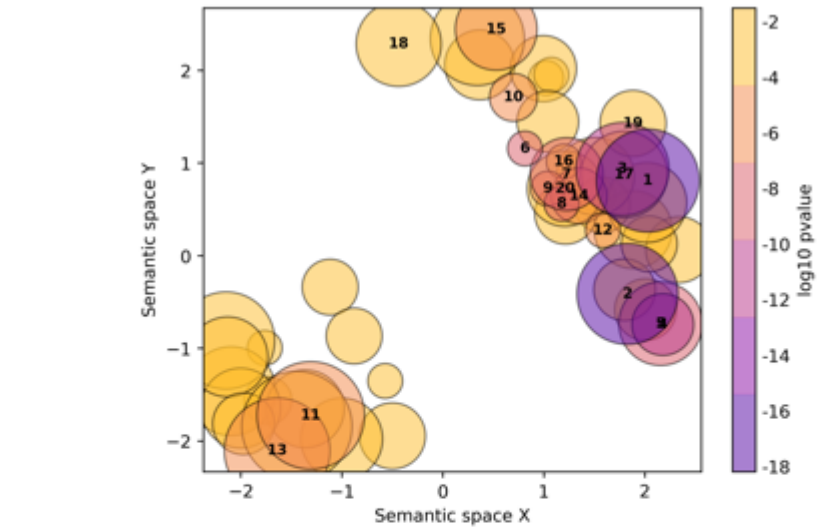
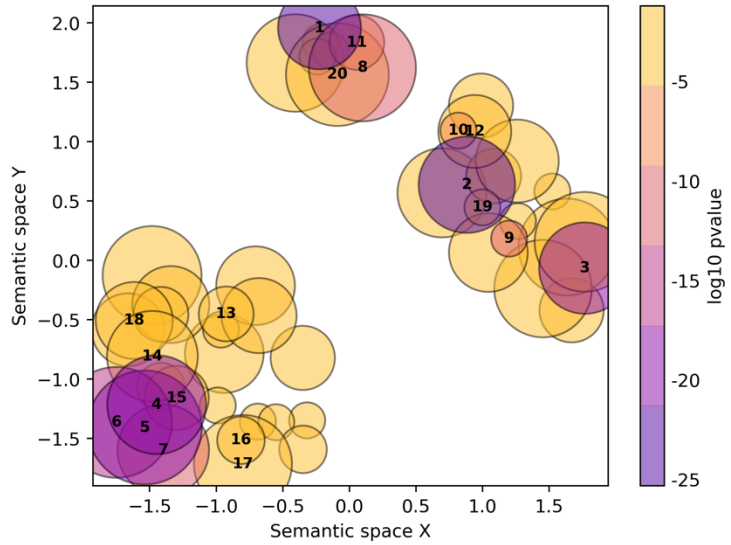


Figure 4-25 Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)



1. mitochondrion
2. intracellular organelle lumen
3. mitochondrial inner membrane
4. mitochondrial protein-containing complex
5. ribosomal subunit
6. oxidoreductase complex
7. respiratory chain complex
8. extracellular exosome
9. respirasome
10. extracellular region
11. peroxisome
12. collagen-containing extracellular matrix
13. Lsm2-8 complex
14. methylosome
15. tricarboxylic acid cycle enzyme complex
16. MHC class I protein complex
17. protein kinase CK2 complex
18. GINS complex
19. blood microparticle
20. motile cilium

Figure 4-26 Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values and betas from the CERAD score logistic regression (1,2 vs 3,4)

4.3.7 Differential gene expression analysis of AMP-AD case-control data

Using residuals from the LMEM normalisation step in the previous chapter, one regression model was performed to determine differentially expressed genes. This was a logistic regression of controls vs cases (coded 0 vs 1 respectively) and resulted in 1270 significant (FDR < 0.05) differentially expressed genes. The top 10 differentially expressed genes can be seen in Table 4-11.

Gene	Chr:start-end	Beta	p-value	FDR corrected p-value
PAFAH1B3	19:42,297,033-42,303,546	0.67	$3.18x10^{-12}$	$5.24x10^{-08}$
CLTB	5:176,392,501-176,416,539	0.58	$4.68x10^{-09}$	$3.86x10^{-05}$
ADAMTS2	5:179,110,853-179,345,461	0.64	$1.17x10^{-08}$	$5.84x10^{-05}$
DRD1	5:175,440,036-175,444,182	0.57	$1.77x10^{-08}$	$5.84x10^{-05}$
SCGN	6:25,652,201-25,701,783	0.48	$1.50x10^{-08}$	$5.84x10^{-05}$
NGB	14:77,265,483-77,271,206	0.52	$2.18x10^{-08}$	$5.98x10^{-05}$
OCRL	X:129,539,849-129,592,561	0.51	$2.66x10^{-08}$	$6.26x10^{-05}$
KREMEN2	16:2,964,216-2,968,383	-0.55	$3.37x10^{-08}$	$6.94x10^{-05}$
CBX5	12:54,230,942-54,280,133	0.51	$6.86x10^{-08}$	$9.16x10^{-05}$
FBXO44	1:11,654,375-11,663,327	0.50	$6.27x10^{-08}$	$9.16x10^{-05}$

Table 4-11 - Top 10 differentially expressed genes from LMEM model including 3PCs after logistic regression with controls vs cases (coded 0 vs 1). *Genes were ranked based on their FDR corrected and reported with their beta coefficient from regression analyses and each gene's FDR corrected p-value (FDR < 0.05). Gene Chr:start-end refers to gene chromosome and start and end base position. All in build GRCh38 (www.gencodegenes.org/human/release_24.html).*

Similar to previous analyses, Kunkle et al. AD GWAS prioritised genes inspected for evidence for differential expression. *AGFG2* and *WDR18* were the only genes that were differentially expressed, and that significance survived FDR correction (*AGFG2*: p-value: 2.02×10^{-03}; FDR p-value: 0.04 and *WDR18*: p-value: 2.12×10^{-03}; FDR p-value: 0.04).

The following genes were at least nominally significant: *ACP2*, *C1QTNF4*, *CD2AP*, *CLU*, *PSMB8*, *PSMC5*, *WWOX* and *YOD1*. *PSMB9* was the only gene that had been differentially expressed with at least nominal significance across all Braak, CERAD and case-control analyses. *C1QTNF4*, *CD2AP*, and *WWOX* were the only nominally significant genes in the case-control analysis that did not have a significant result in the previous analyses. The results of this case-control analysis can be seen in Table 4-12.

Kunkle et al. AD GWAS prioritised genes were also tested for enrichment for differentially expressed genes . This test was not significant (p-value: 0.95), indicating that GWAS prioritised genes were not more enriched in this DGE analysis than would be expected through chance alone. A boxplot demonstrating the difference in p-value between GWAS prioritised and non-prioritised genes can be seen in Figure 4-27. The boxplot shows that the non-GWAS prioritised genes tended to have lower p-values.

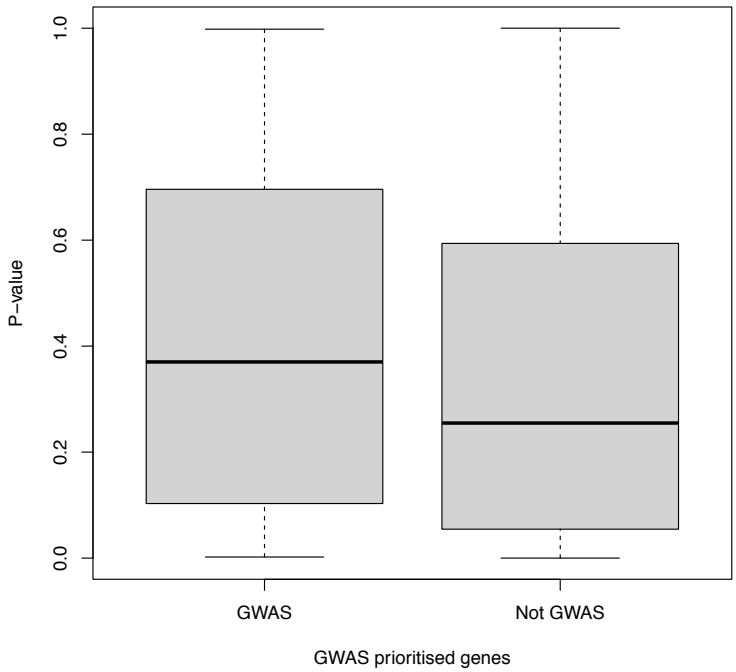


Figure 4-27 - A boxplot demonstrating the differences in p-value resulting from the AD case-control differential expression analysis. 'GWAS' refers to the GWAS prioritised genes and 'Not GWAS' refers to any gene not in this set.

Gene	Chr:start-end	**DGE case-control score Logistic regression**		
		Beta	p-value	FDR p-value
ABCA7	19:1,039,997-1,065,572	0.18	0.08	0.28
ACP2	**11:47,239,302-47,248,906**	**-0.23**	**0.03**	**0.16**
ADAM10	15:58,588,809-58,749,791	0.17	0.06	0.24
ADAMTS1	21:26,835,755-26,845,409	0.08	0.46	0.70
AGFG2	7:100,539,203-100,568,220	0.34	$2.02x10^{-03}$	0.04
BIN1	2:127,048,027-127,107,288	-0.14	0.12	0.34
C1QTNF4	11:47,589,667-47,594,411	0.04	0.61	0.80
C4A	6:31,982,052-32,002,681	0.09	0.40	0.66
C7ORF43	7:100,154,420-100,158,723	0.10	0.20	0.45
CASS4	20:56,412,112-56,460,387	0.06	0.55	0.77
CD2AP	**6:47,477,789-47,627,263**	**0.18**	**0.04**	**0.18**
CD55	1:207,321,519-207,386,804	0.06	0.57	0.78
CELF1	11:47,465,933-47,565,569	0.17	0.07	0.25
CLU	**8:27,596,917-27,614,700**	**0.26**	**0.01**	**0.06**
CNN2	19:1,026,586-1,039,068	-0.09	0.34	0.59

CR1	1:207,496,147-207,641,765	-0.04	0.63	0.81
ECHDC3	10:11,742,366-11,764,070	0.03	0.79	0.90
EED	11:86,244,753-86,278,813	0.07	0.42	0.67
EPHB4	7:100,802,565-100,827,523	0.07	0.50	0.71
FAM131B	7:143,353,400-143,362,770	0.26	0.07	0.25
GAL3ST4	7:100,159,244-100,168,617	-0.05	0.59	0.79
GPSM3	6:32,190,766-32,195,523	0.01	0.92	0.97
HLA-DPA1	6:33,064,569-33,080,775	-0.0003	1.00	1.00
HLA-DQA1	6:32,628,179-32,647,062	0.05	0.61	0.80
HLA-DRB1	6:32,577,902-32,589,848	-0.03	0.78	0.90
HLA-DRB5	6:32,517,353-32,530,287	-0.004	0.97	0.99
HMHA1	19:1,065,923-1,086,628	-0.04	0.69	0.85
INPP5D	2:233,059,967-233,207,903	0.15	0.14	0.37
IQCK	16:19,716,456-19,858,467	0.09	0.31	0.57
MAF	16:79,585,843-79,600,737	0.13	0.18	0.43
MS4A4	11:60,185,657-60,318,080	0.04	0.70	0.85

MS4A6A	11:60,172,015-60,184,666	0.03	0.82	0.92
MS4A7	11:60,378,485-60,395,951	0.12	0.28	0.54
MTCH2	11:47,617,315-47,642,607	-0.08	0.46	0.70
NDUFS3	11:47,565,336-47,584,562	0.03	0.70	0.86
NUP160	11:47,778,087-47,848,555	0.09	0.33	0.58
PICALM	11:85,957,175-86,069,882	-0.01	0.88	0.95
PILRA	7:100,367,530-100,400,096	0.03	0.75	0.88
PSMB8	6:32,840,717-32,844,679	-0.13	0.15	0.38
PSMB9	**6:32,844,136-32,859,851**	**-0.21**	**0.01**	**0.10**
PSMC3	11:47,418,769-47,426,473	-0.05	0.57	0.78
PSMC5	17:63,827,152-63,832,026	0.22	$9.75x10^{-03}$	0.08
PTK2B	8:27,311,482-27,459,391	0.03	0.75	0.88
RIN3	14:92,513,781-92,688,994	0.003	0.97	0.99
SORL1	11:121,452,314-121,633,763	-0.12	0.20	0.45
SPI1	11:47,354,860-47,378,547	-0.01	0.91	0.96
STYX	14:52,730,166-52,774,989	-0.10	0.21	0.47

TREM2	6:41,158,506-41,163,186	0.12	0.26	0.52
WDR18	19:984,332-998,438	**-0.29**	**$2.12x10^{-03}$**	**0.04**
WWOX	16:78,099,400-79,212,667	**0.22**	**0.02**	**0.13**
YOD1	1:207,043,849-207,052,980	**0.20**	**0.04**	**0.17**
ZKSCAN1	7:100,015,572-100,041,689	-0.13	0.25	0.50

Table 4-12 - Results from the case-control differential gene expression analysis for top-prioritised genes from the largest AD case-control GWAS (Kunkle et al. 2019). All in build GRCh38 (www.gencodegenes.org/human/release_24.html).

4.3.8 GO enrichment analysis of AMP-AD case-control data

GO enrichment analysis of the case-control logistic regression differentially expressed genes resulted in 1092 statistically significant GO categories that are enriched for up-regulated genes and 79 statistically significant GO categories that are enriched for down-regulated genes. 33 were significant in the non-directional analysis. For all, statistically significant refers to an FDR-corrected p-value of less than 0.05 and the categories include biological process, molecular function and cellular component.

The python package *GO-Figure!* was used to reduce the lists of GO terms to a summarised list of terms across the three GO categories of biological process, molecular function and cellular component.

For the case-control data, biological process GO terms enriched for upregulated genes related to signalling which is more in line with the results for the Braak phenotype than the CERAD phenotype (Figure 4-29). Enriched GO categories such as SRP-dependent cotranslational protein targeting to membrane, viral transcription, and mitochondrial pathways were found

to be down-regulated (Figure 4-30). This is very similar to what was seen for the Braak and CERAD phenotypes.

Enriched molecular function GO terms that were up-regulated were related to binding, transcription and transporter activity (Figure 4-32) whereas ribosome and catalytic activity were GO terms that were down-regulated (Figure 4-33), again reinforcing what was found with the Braak and CERAD phenotypes.

Cellular components GO terms that were up-regulated mainly related to the synapses and neurons and plasma membrane (Figure 4-35) whereas down-regulated GO terms were relating to the mitochondria, ribosome and endoplasmic reticulum (Figure 4-36). Again, supporting the results from the Braak and CERAD phenotypes.

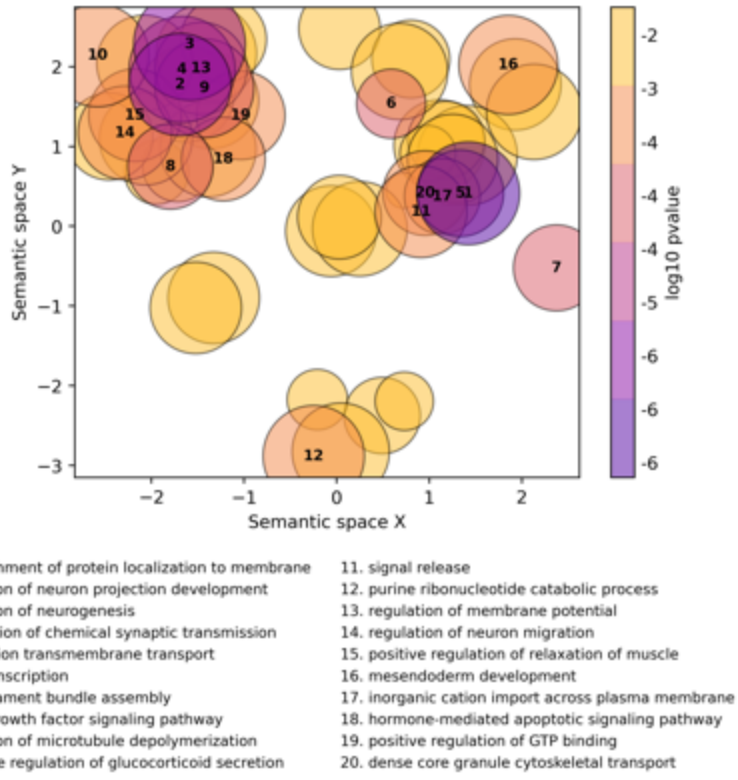
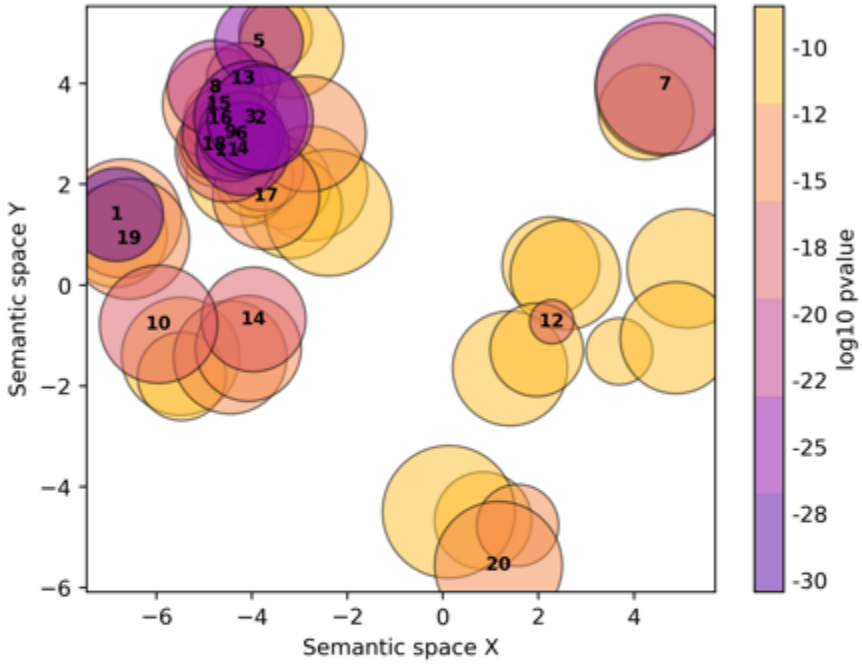
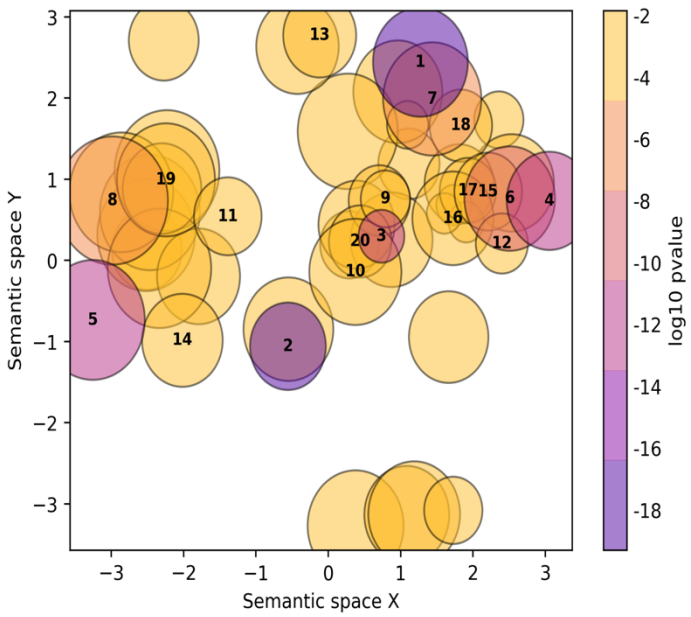


Figure 4-28 Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the case-control logistic regression



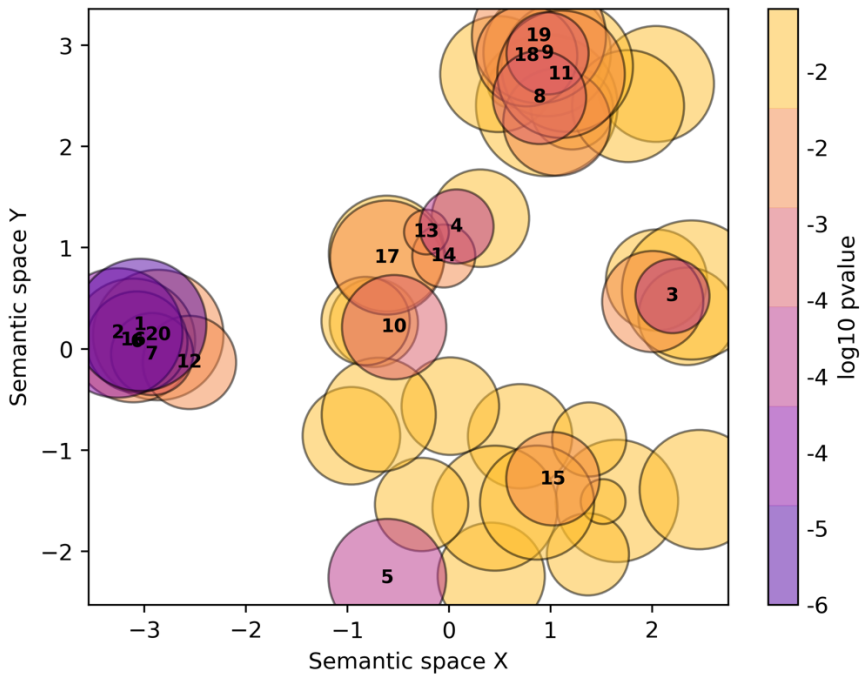
1. positive regulation of cellular process
2. regulation of signaling
3. regulation of cell communication
4. signal transduction
5. regulation of localization
6. regulation of biological quality
7. developmental process
8. negative regulation of cellular process
9. regulation of developmental process
10. regulation of multicellular organismal process
11. regulation of cellular component organization
12. cellular process
13. regulation of nervous system development
14. regulation of plasma membrane bounded cell projection organization
15. negative regulation of nitrogen compound metabolic process
16. regulation of response to stimulus
17. regulation of transcription by RNA polymerase II
18. regulation of primary metabolic process
19. positive regulation of developmental process
20. cellular response to organic substance

Figure 4-29 Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the case-control logistic regression



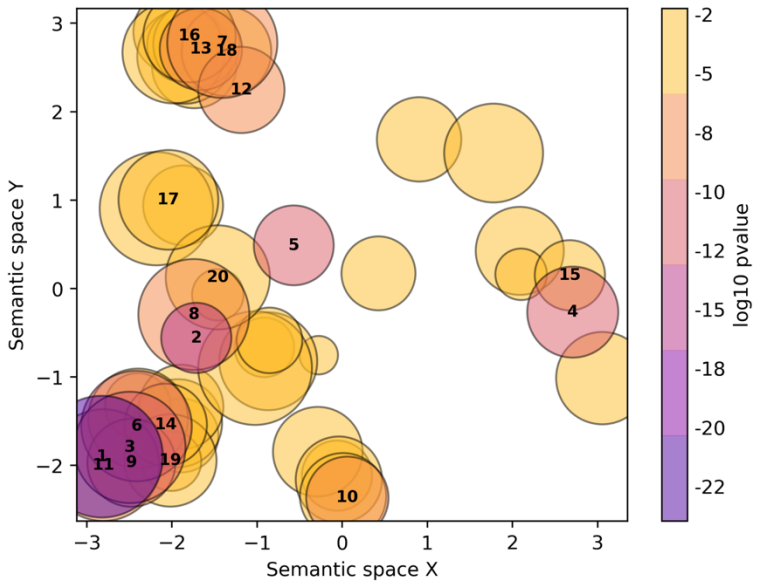
1. SRP-dependent cotranslational protein targeting to membrane
2. viral transcription
3. translational initiation
4. translation
5. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
6. ncRNA metabolic process
7. ribonucleoprotein complex assembly
8. positive regulation of glucocorticoid secretion
9. drug catabolic process
10. attachment of mitotic spindle microtubules to kinetochore
11. hormone-mediated apoptotic signaling pathway
12. peptidyl-glutamic acid modification
13. insemination
14. negative regulation of mitochondrial membrane permeability involved in apoptotic process
15. aromatic amino acid family metabolic process
16. indolalkylamine metabolic process
17. leucine metabolic process
18. transcription-coupled nucleotide-excision repair
19. positive regulation of immature T cell proliferation in thymus
20. cellular oxidant detoxification

Figure 4-30 Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the case-control logistic regression



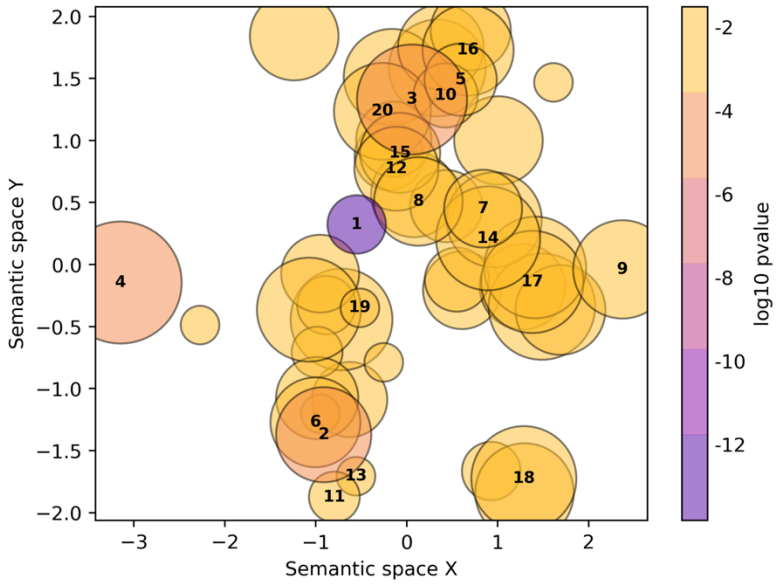
1. voltage-gated cation channel activity
2. cation channel activity
3. cAMP-dependent protein kinase regulator activity
4. cAMP binding
5. 3',5'-cyclic-nucleotide phosphodiesterase activity
6. potassium ion transmembrane transporter activity
7. voltage-gated sodium channel activity
8. myosin light chain binding
9. calmodulin binding
10. oxygen carrier activity
11. insulin-like growth factor receptor binding
12. ion gated channel activity
13. DNA binding, bending
14. NADH binding
15. aminomethyltransferase activity
16. sodium ion transmembrane transporter activity
17. neuropeptide binding
18. peptide hormone receptor binding
19. GTPase binding
20. arginine transmembrane transporter activity

Figure 4-31 Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the case-control logistic regression



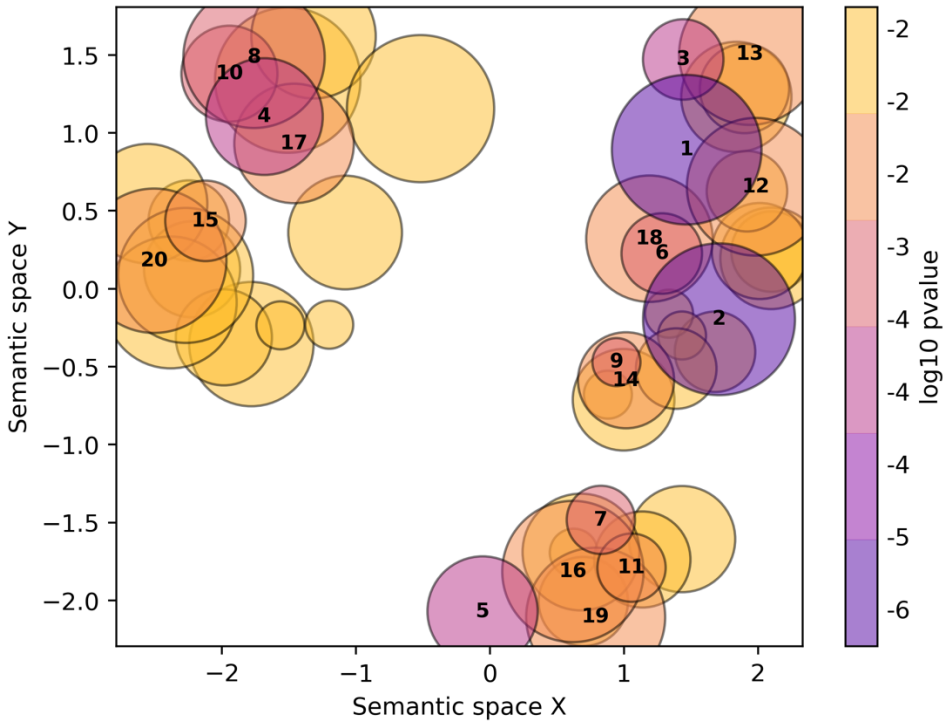
1. protein binding
2. binding
3. enzyme binding
4. kinase activity
5. DNA-binding transcription factor activity, RNA polymerase II-specific
6. protein domain specific binding
7. voltage-gated cation channel activity
8. RNA polymerase II transcription regulatory region sequence-specific DNA binding
9. transcription factor binding
10. transcription coregulator activity
11. signaling receptor binding
12. ion gated channel activity
13. potassium ion transmembrane transporter activity
14. cadherin binding
15. protein tyrosine kinase activity
16. monovalent inorganic cation transmembrane transporter activity
17. enzyme activator activity
18. channel activity
19. RNA polymerase II transcription factor binding
20. phospholipid binding

Figure 4-32 Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the case-control logistic regression



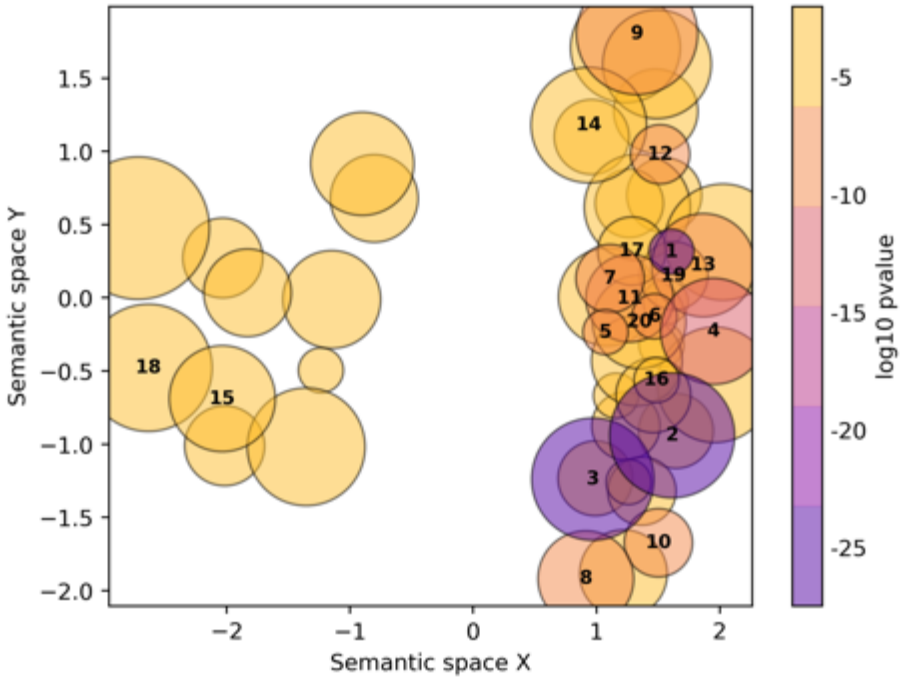
1. structural constituent of ribosome
2. rRNA binding
3. catalytic activity, acting on a tRNA
4. hormone activity
5. aminomethyltransferase activity
6. flavin adenine dinucleotide binding
7. peroxidase activity
8. trans-2-enoyl-CoA reductase (NADPH) activity
9. endoribonuclease activity
10. catechol O-methyltransferase activity
11. MutSalpha complex binding
12. enoyl-CoA hydratase activity
13. proteasome core complex binding
14. ligase activity
15. ammonia-lyase activity
16. protein-arginine omega-N monomethyltransferase activity
17. dITP diphosphatase activity
18. malate transmembrane transporter activity
19. ubiquinone binding
20. oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxyge...

Figure 4-33 Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the case-control logistic regression



1. neuron projection
2. synapse
3. integral component of postsynaptic density membrane
4. cytosolic large ribosomal subunit
5. integral component of postsynaptic membrane
6. chromosome, centromeric region
7. postsynaptic density
8. cation channel complex
9. neuronal cell body
10. sodium:potassium-exchanging ATPase complex
11. mitotic spindle
12. synaptic membrane
13. transport vesicle membrane
14. calyx of Held
15. CERF complex
16. chromaffin granule
17. kinesin complex
18. cytoplasmic side of lysosomal membrane
19. contractile fiber
20. SCF ubiquitin ligase complex

Figure 4-34 - Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the case-control logistic regression



1. plasma membrane
2. neuron projection
3. synapse
4. plasma membrane region
5. cell body
6. presynapse
7. cytosol
8. integral component of synaptic membrane
9. cytoplasmic vesicle
10. integral component of postsynaptic specialization membrane
11. chromatin
12. postsynaptic density
13. dendrite membrane
14. cytoplasmic vesicle membrane
15. voltage-gated potassium channel complex
16. postsynapse
17. neuron projection terminus
18. plasma membrane protein complex
19. membrane raft
20. cytoplasmic region

Figure 4-35 Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the case-control logistic regression

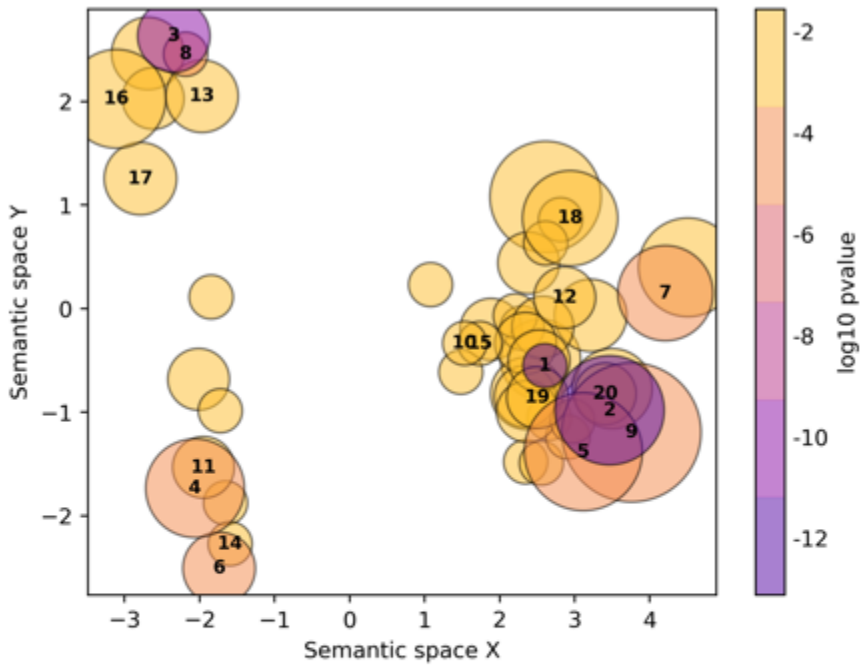


Figure 4-36 Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the case-control logistic regression

4.3.9 MAGMA pathways and comparison with results from CATMAP

Results from a MAGMA pathway analysis from the largest AD GWAS (Kunkle et al. 2019) were used to check if a similar enrichment of terms was found in the GO enrichment analysis using AD case-control gene expression data.

From the MAGMA pathway analysis, nine GO terms were found to be statistically significant after FDR correction as published in their paper. These GO terms related to APP metabolism/amyloid-beta formation, tau protein binding, lipid metabolism and immune response. None of the nine GO terms were statistically significant after FDR correction or at a nominal p-value of 0.05 in my non-directional GO enrichment analysis using case-control

gene expression data. The was also true of the GO enrichment analysis that ranked down-regulated genes higher (down-to-up analysis). Five GO terms were nominally significant in the GO enrichment analysis that gave upregulated genes a higher rank (up-to-down analysis) and two of which were also FDR significant (<0.05). The two FDR significant GO terms included tau protein binding (GO:0048156) and activation of immune response (GO:0002253). The results of the MAGMA analysis (from the Kunkle et al. paper) and my case-control analysis are presented in Table 4-13.

GO term	Description	GWAS P-value	GWAS FDR corrected p-value	Non-directional CATMAP p-value	Non-directional FDR corrected p-value	Up-to-down CATMAP p-value	Up-to-down FDR corrected p-value	Down-to-up CATMAP p-value	Down-to-up FDR corrected p-value
GO:0065005	Protein-lipid complex assembly	$1.4x10^{-07}$	$9.5x10^{-04}$	0.49	0.99	0.03	0.21	0.97	1.00
GO:1902003	Regulation of amyloid-beta formation	$4.5x10^{-07}$	$1.4x10^{-03}$	0.58	0.99	0.10	0.39	0.90	1.00
GO:0032994	Protein-lipid complex	$1.1x10^{-06}$	$2.5x10^{-03}$	0.12	0.89	0.44	0.75	0.56	1.00
GO:1902991	Regulation of amyloid precursor protein catabolic process	$3.5x10^{-06}$	$5.8x10^{-03}$	0.40	0.99	0.02	0.16	0.98	1.00
GO:0043691	Reverse cholesterol	$5.5x10^{-06}$	$6.7x10^{-03}$	0.38	0.99	0.90	0.99	0.10	1.00
GO:0071825	Protein-lipid complex	$6.1x10^{-06}$	$6.7x10^{-03}$	0.36	0.99	0.41	0.72	0.60	1.00
GO:0034377	Plasma lipoprotein	$1.6x10^{-05}$	$1.5x10^{-02}$	0.59	0.99	0.03	0.22	0.96	1.00
GO:0048156	Tau protein binding	$3.1x10^{-05}$	$2.6x10^{-02}$	0.08	0.86	$1.2x10^{-03}$	0.02	1.00	1.00
GO:0002253	Activation of immune response	$6.3x10^{-05}$	$4.6x10^{-02}$	0.99	0.99	$1.3x10^{-03}$	0.02	1.00	1.00

Table 4-13 - A list of significant GO terms as published in the largest AD case-control GWAS (Kunkle et al. 2019).

4.4. Discussion

In this chapter, RNA-seq data that had been quality controlled and normalised were used in logistic and ordinal regressions to perform DGE analysis. These lists of differentially expressed genes were then used to perform GO enrichment analysis.

Initially the aim was to utilise only the ROSMAP dataset to determine if using LMEM to normalise gene expression data along with logistic regression (to perform a DGE analysis) produced differentially expressed genes which overlapped with those identified using more conventional methods. These two other methods were limma-voom and DESeq2. The results presented in this chapter showed that the LMEM + LR method produced significant differentially expressed genes that largely overlapped with the two tools. Genes from LMEM + LR showed greater overlap to those from the limma-voom analysis than the DESeq2 analysis. This could be explained in part as DESeq2 assumes a negative binomial distribution of the RNA-seq data. In contrast limma-voom assumes a normal distribution so as a method its more comparable to LMEM + LR as that also assumes a normal distribution.

In a previous review of available software tools for DGE analysis, DESeq2 and Limma-Voom were the best performing parametric tools. They achieved true positive rates of 84% and 81% resepectively when compared with results from qRT-PCR. The authors also identified that differentially expressed genes identified through multiple approaches were more likely to be true differentially expressed genes (Costa-Silva et al. 2017). The LMEM + LR method identified 352 significant differentially expressed genes (FDR < 0.05) all of which were also differentially expressed in the limma-voom and/or DESeq2 analyses. This was fewer than both the DESeq2 and limma-voom methods (1054 and 1260 respectively). It is possible that the LMEM + LR method is more conservative than the other two methods explored. However, it has the added benefit of being easier to implement. It also enables multiple datasets to be combined, which cannot easily be done using the other two packages. This benefit was exploited in combining the ROSMAP data with data from the MSBB and MayoRNASeq cohorts to form a

single dataset (refered to as combined AMP-AD) to perform DGE analyses for the Braak, CERAD and case-control phenotypes.

In the analysis of the combined AMP-AD data, the analysis initially focused on the Braak score phenotype. The following seven genes were in the list of top 10 differentially expressed genes at least twice across the three regression models: *LPO*, *LINC01844*, *OCRL*, *ANKRD18DP*, *CBX5*, *NCDN*, *KCNK9.* Two of these have been implicated in neurodegeneration before. Deletion of the *LPO* gene in model mice has been found to result in multisystem inflammation and degenerative changes in neuropathology. It has also been proposed to play a role in the pathogenesis of Parkinson's disease (Fernández-Espejo et al. 2021; Yamakaze and Lu 2021). *CBX5* has not been found previously to be implicated in AD. It has been found to be an upregulated gene in drug-naive patients with Parkinson's disease in comparison to controls (Calligaris et al. 2015). *CBX5* was found to also be upregulated in my Braak analysis (fold-change=0.52 ; FDR p-value = 1.63×10^{-05}).

For the CERAD score phenotype, the following seven genes were in the list of top 10 differentially expressed genes at least twice across the three regression models: *DNAJC19*, *LPO*, *DRD1*, *GCSH*, *TIMM8B*, *FAM19A2*, *KCTD8*. Some of these have been associated with neurodegeneration before. *DNAJC19* is a gene which encodes for a protein with mitochondrial functions and mutations in which can result in dilated cardiomyopathy and ataxia (Zarouchlioti et al. 2018). This gene has been reported to be a differentially expressed gene in hippocampus RNA-seq data in AD cases and controls from The Banner Sun Health Research Insitute. (Dharshini et al. 2019). It is important to note that the controls used in this study potentially overlap with those included in the MayoRNASeq data as they were also sourced from the Banner Sun Health Research Institute (Allen et al. 2016). Additionally, variations in *DRD1* have been previously associated with behaviour changes in AD. These include aggression and psychotic symptoms and poorer cognition (Holmes et al. 2001; Tsang et al. 2015).

For the case-control phenotype, the following seven genes were in the list of top 10 differentially expressed genes: *DNAJC19*, *LPO*, *DRD1*, *GCSH*, *TIMM8B*, FAM19A2, *KCTD8*. All of which have featured in either the top 10 Braak or CERAD DGE list.

PSMB9 was the only Kunkle prioritised GWAS gene to be at least nominally significant across all analyses of Braak, CERAD and case-control phenotypes. *PSMB8* (chr6:32,840,717-32,844,679) and *PSMB9* (chr6:32,844,136-32,859,851) were both downregulated in AD cases in comparison to controls, at least with nominal significance across all regressions for *PSMB9* and CERAD regressions for *PSMB8*. The role these genes play in AD is unclear. PSMB9 and PSMB8 are subunits of the immunoproteasome. Under inflammatory conditions, PSMB9 and PSMB8 replace the constitutively expressed subunits PSMB5 and PSMB6. This leads to the creation of different peptides during inflammation (Kloetzel 2001; Kalaora et al. 2020). An association between over-expression of both *PSMB8* and *PSMB9* and improved survival and enhanced response to immune-checkpoint inhibitors such as anti-programmed cell death protein 1 (anti-PD1) in melanoma patients has been shown (Kalaora et al. 2020). Immune-checkpoint inhibitors have been proposed as a potential therapy for AD (Schwartz et al. 2019). In mouse models, exposure to with anti-PD1 resulted in clearance of amyloid and tau pathology and improved cognition (Rosenzweig et al. 2019; Schwartz et al. 2019). However, these findings have failed to replicate (Lin et al. 2020).

Six genes that were identified as prioritised genes in the Kunkle et al. GWAS were found to be differentially expressed after FDR correction in at least one of the phenotypes. These were *AGFG2*, *CELF1*, *CLU*, *PSMB9*, *PSMC5*, *WDR18*. *CLU* is considered to be the third most significant common variant implicated genetic risk factor for AD after *APOE* and *BIN1* (Foster et al. 2019). However, when GWAS prioritised genes were tested as a gene-set, there was no evidence for this group of genes being enriched in any of the DGE analyses.

There are a number of possible reasons as to why the GWAS prioritised genes were not enriched in the DGE analysis. GWAS identified common variants are significantly more likely to be eQTLs. Therefore gene expression has been suggested as a intermediary between DNA

variation and complex phenotypes (Nicolae et al. 2010; Porcu et al. 2021). GWAS trait associated SNPs largely fall in the non-coding region. Connecting causal variants with their probable causal gene is not clear-cut. This analysis used GWAS prioritised genes, but it is not known if they are the true causal genes. It could be that the prioritised genes are not the true causal genes and therefore not enriched in the DGE analysis. Conversely, it could be that the DGE analyses are identifying differentially expressed genes that are not a cause of disease but a result of disease so not reflected in the GWAS findings. Additionally the DGE analysis was performed in bulk brain cortex tissue, there may be more disease specific tissues that would identify genes associated with cause of disease better than result of disease. Another explanation could be that the identified genes possibly have another form of mechanistic involvement in AD if they are involved in the aetiology of disease.

The next aim of this chapter was to perform GO enrichment analyses. Across all phenotypes and regression analyses GO terms such as SRP-dependent cotranslational protein targeting to membrane, the ribosome, the endoplasmic reticulum, viral transcription, and mitochondrial pathways were found to be enriched. These findings are consistent with previous reports. One study used downloaded microarray data from Gene Expression Omnibus (GEO) database. In this study the authors performed weighted gene co-expression network analysis to explore the relationship between gene sets (modules) associated with AD and MCI. They performed functional enrichment analysis on the AD and MCI modules and identified that AD module genes showed an enrichment of SRP-dependent cotranslational protein targeting to membrane, protein targeting to ER and and cytosolic ribosome (Tao et al. 2020).

The analysis presented in this chapter has now also provided evidence for pathways that are also implicated in the Braak and CERAD phenotypes. There was some discrepancy between the CERAD phenotype and the Braak and case-control phenotypes in the up-regulated biological process GO pathways. The Braak and case-control phenotypes' up-regulated pathways mainly referred to cell signalling whereas for CERAD it was more related to synaptic processes and transcription. Molecular function and cellular component GO terms were consistent across all phenotypes. One would expect a consensus between the results from

the Braak and CERAD phenotypes given that these phenotypes are so closely correlated. Whether the discrepancy between Braak and CERAD is because different pathways play a role in how these different pathologies develops needs to be further explored.

Finally GO terms from the DGE analysis were compared to the nine significant GO terms from the MAGMA pathway analysis on genetic data published in the largest case-control GWAS (Kunkle et al. 2019). No significant GO terms from the Kunkle et al. analysis overlapped with FDR significant (< 0.05) GO terms from the non-directional and down-regulated analyses using RNA-seq data. Two GO terms were FDR significant in both the Kunkle et al. GWAS analysis and the up-regulated analysis. These were: Tau Protein Binding (GO:0048156) and activation of immune response (GO:0002253).

The work presented in this chapter has shown that LMEM + LR can be used as a method to not only combine datasets, but also integrate multiple brain areas to discover differentially expressed genes with a consensus to other tools. The DGE analysis has found genes that have been associated with AD before and implicated some new genes such as *CBX5* (chr12:54,230,942-54,280,133*)*. The analysis suggests that GWAS prioritised genes as a group were not particularly differentially expressed in AD cases and controls. This suggests that they possibly have another form of mechanistic involvement in AD if they are involved in the aetiology of disease. Finally, pathway enrichment analysis has suggested differentially expressed genes play a role in cell death, cell signalling, neuronal and synaptic processes.

Results from DGE analyses are often criticised for being vague and harbouring too many false positives (Li et al. 2022b). A limitation of the DGE approach is that it is not possible to tell which differentially expressed genes and associated pathways may be a cause of AD, may be a consequence of AD, or reflect uncontrolled confounders. The results of this analysis are tentative and firm conclusions cannot be drawn from them yet. Nevertheless, the results in this analysis have plausible biological relevance, and supported by existing evidence and are thus encouraging.

AD is likely due to the outcome of perturbations of gene networks by the co-action of genetic and environmental risk factors. So, it is necessary to look beyond genotypes in isolation to understand these perturbations and mechanisms of disease. The next chapter aims to build on this work by integrating genetic and gene expression data in an eQTL analysis and will investigate the relationship between genotypes and differentially expressed genes.

Chapter 5 – Expression Quantitative Trait loci (eQTL) analysis of AMP-AD

5.1 Introduction

5.1.1 From GWAS to expression quantitative loci

There are many factors that can contribute to gene expression variation. These include environmental exposures, genetics, technical artefacts, stochastic variation as well as sex and age differences (Leek and Storey 2007). When considering lists of differentially expressed genes, it is not easy to disentangle which sources of variation may be contributing to gene expression changes. The heritable component of gene expression is considered by some to be a crucial bridge in the linking of genomic variation to disease biology (Albert et al. 2018).

Many GWAS have identified SNPs that are associated with the AD phenotype. Translating these findings into understanding of the mechanisms behind how these variants contribute to disease risk has been difficult. GWAS identified SNPs are representatives for often a large number of SNPs in the region (tagging or index SNPs) and are not necessarily causal. Usually, the nearest gene to the index SNP is reported. It is possible that other SNPs (genes) in high LD with the array-identified SNPs are causal for the disease. Many risk-associated SNPs are located in non-coding regions and are likely to exert their biological function by modulating gene expression (Ni et al. 2020). Improved understanding of the relationship between non-coding variation and clinical AD is vital to increase understanding of disease biology and identify potential therapeutic targets.

One method to try and disentangle the genetic cause of gene expression and prioritise SNPs associated with disease is expression quantitative trait loci (eQTL) analysis. An eQTL analysis is a genetic/transcriptomic association approach for identifying genetic variants (such as SNPs) associated with the differential expression of a gene. EQTLs have frequently been used

to highlight candidate causal variants, genes, and provide a potential link between them and the biological processes they affect (Albert and Kruglyak 2015).

Research in AD using eQTLs has suggested that altered gene expression plays a role in the aetiology of AD however most studies focus only on cis-eQTLs (Zou et al. 2010; Sieberts et al. 2020; Patel et al. 2021). Cis-eQTLs (also known as local eQTLs) are those that are located near the gene-of-origin (gene which produces the transcript or protein) as opposed to trans-eQTLs which are distant from their gene-of-origin (Goswami and Sanan-Mishra 2022). There is no standard quantification of what is meant by local or distant. Researchers will normally specify these distances on the basis of the experiment being performed. The focus in AD research to date has mainly been on cis-eQTLs because detecting trans-eQTLs is more difficult. Trans-eQTLs tend to have weak effect sizes and therefore require larger sample sizes to detect than cis-eQTLs (GTEx-Consortium 2013; Clyde 2017; Võsa et al. 2021). Additionally trans-eQTL analyses involve a greater amount of tests of association between SNPs and genes than a cis-eQTL analysis resulting in a greater computational burden.

5.1.2 Aims

The first aim of this chapter is to perform a cis-eQTL analysis using MatrixEQTL to find AD GWAS index SNPs that are associated with genes that are differentially expressed between cases and controls which were identified in chapter four.

The second aim of this chapter is to perform a cis-eQTL analysis expanded to the 100kb region either side of significant eQTLs from aim one. This is to locate other potentially causal SNPs associated with differentially expressed genes. This 100kb region was selected as at least 80% of common variants identified in published GWAS that use imputed data were within 33.5 kb of causal variants, and over 90% within 100kb (Wu et al. 2017).

The final aim of this chapter is to then perform a trans-eQTL analysis using AD GWAS index SNPs to find associations with AD case-control differentially expressed genes.

5.2 Methods

5.2.1 An overview

The overall strategy for analysis was to use 594 samples with genetic and gene expression information, and perform a cis- and trans- eQTL analysis to identify AD candidate genes. An overview of the methodology is given in Figure 5-1 and described in more detail in each of the following sections.

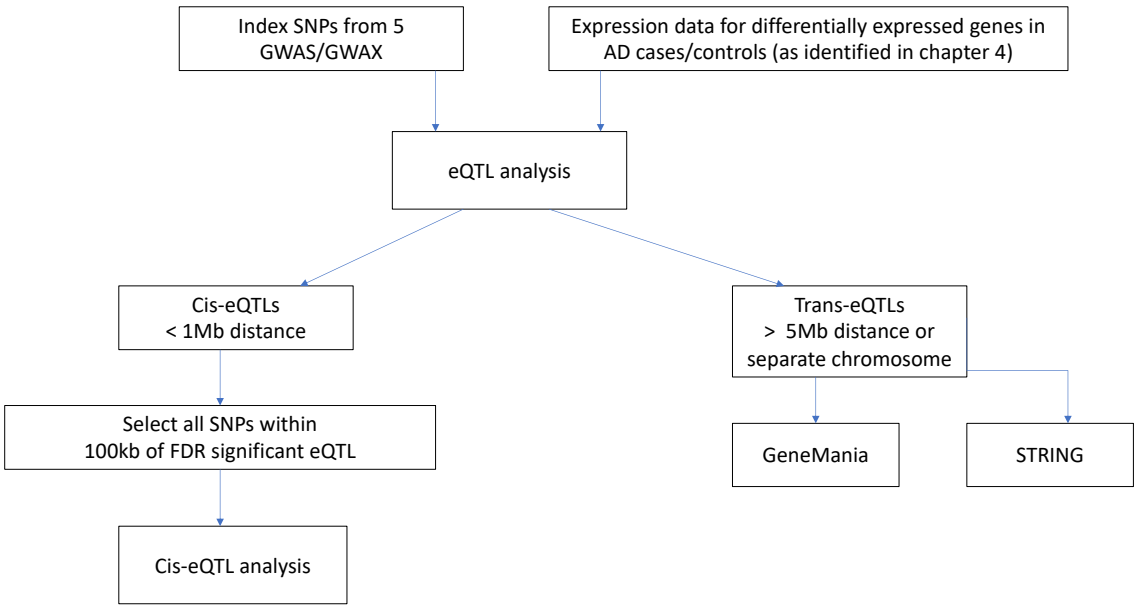


Figure 5-1 – An overview of the methodology used for the eQTL analysis applied to AMP-AD data

5.2.2. Gene expression data

The RNA-seq data used in the eQTL analysis is the same normalised pre-processed data as in previous chapters. As some individuals from the MSBB cohort have multiple samples from

different tissues, only BM10 samples were included in this analysis. To maintain power in the analysis, BM10 was selected over BM22, BM36 and BM44 as this retained the largest number of samples. This resulted in a total of 594 samples from 594 distinct individuals in the analysis. In the previous chapter 1270 genes were identified as differentially expressed as they had an FDR corrected p-value of less than 0.05.

5.2.3 Genotype QC, SNP selection and genotype

The genetic data have been described in previous chapters. SNPs for this analysis were selected for inclusion if they had a MAF ≥ 0.01 and were identified as an index SNP in at least one of five GWAS and GWAS-by-proxy (GWAX) studies. The five studies included one GWAS from 2013, a GWAX from 2018, a GWAS and a GWAX from 2019 and a GWAX from 2021. A summary of these can be found in Table 5-1.

It is important to note that the GWAS/GWAX studies mentioned in Table 5-1 are not independent from one another. Samples from Lambert et al. were included in the Kunkle et al., Jansen et al., and the Wightman et al. studies. Samples from Marioni et al, were included in the Jansen et al. and Wightman et al. studies. Samples from the Kunkle et al. study were also included in the Wightman et al. GWAX (Escott-Price and Hardy 2022). It is impossible to delineate the true sample overlap without the raw data and the fact that these studies are not independent is an important limitation.

Study type	Author	Year	Sample number
GWAS	Lambert et al.	2013	74,046
GWAX	Marioni et al.	2018	388,324
GWAX	Jansen et al.	2019	534,403
GWAS	Kunkle et al.	2019	94,437
GWAX	Wightman et al.	2021	1,126,563

Table 5-1 – The five genome-wide association studies (GWAS) and genome-wide association by proxy studies (GWAX) used to identify index SNPs for inclusion in this analysis. Sample number refers to total number of individuals included in the study.

SNP genotypes were coded as the number of minor alleles (0, 1 or 2). Files were generated using the software PLINK, and 103 AD index SNPs were included in the cis-eQTL and trans-eQTL analyses. A full list of the SNPs included in the analysis can be found in the appendix.

After an initial cis-eQTL analysis, a second cis-eQTL analysis was performed. Significant eQTLs and all SNPs 100kb either side of it were included in the second cis-eQTL analysis after removing SNPs with a MAF < 0.05.

5.2.4 Cis- and trans-eQTL generation using MatrixEQTL

MatrixEQTL is an R package which is designed for fast analysis of large datasets with no loss of precision and is able to perform both cis- and trans-eQTL analysis (Shabalin 2012). The tool is widely used and the package is an official tool of the GTEx project (GTEx-Consortium 2013). MatrixEQTL works by performing a separate linear regression model for each gene-SNP pair

and reporting the test statistic, the effect size estimate and p-value. MatrixEQTL performs its analysis fast by avoiding unnecessary calculations and thus does not calculate residuals or the significance and effect sizes of any covariates (Shabalin 2012). As a result, the package was chosen to be used to generate the eQTLs presented in this chapter.

For all cis-eQTL analyses, the cis-eQTL mapping window was defined as 1Mb upstream of the transcription start site to 1Mb downstream of the gene end. The trans-eQTL mapping window was defined to be either at least 5Mb if on the same chromosome or on separate chromosomes. EQTLs that resided on the same chromosome but were less than 5Mb from their eGenes (the gene in the eQTL SNP-gene pair) were not included in this analysis to avoid confounding by long-range LD patterns.

For the eQTL analysises, a linear regression model was chosen as is as follows:

$$Expression \sim Genotype + Diagnosis + Age\ at\ death + Sex + PC1 + \cdots + PC10 \quad (1)$$

where expression is the normalised expression level of the gene, genotype is in respect to the number of minor alleles. Diagnosis, age at death, sex, and the first 10 ancestry principal components were included as covariates.

After the initial cis-eQTL results, three of the seven significant results were based on chromosome 19. As a result, the analysis was rerun to include APOE $\varepsilon 4$ allele status (coded as 0, 1 or 2 indicating the number of $\varepsilon 4$ alleles present) into the cis-eQTL model as follows:

$$Expression \sim Genotype + Diagnosis + Age\ at\ death + Sex + PC1 + \cdots + PC10 + \text{APOE E4 status} \quad (2)$$

5.2.5 Multiple hypothesis testing correction and comparison of results

To correct for multiple hypothesis testing the Benjamini-Hochberg method was also used to calculate FDR corrected p-values (Benjamini and Hochberg 1995). This was performed by using the p.adjust function in R and using the 'BH' option. Only results meeting these thresholds were considered statistically significant.

5.2.6 Analysis of trans-eQTL results

Results from the trans-eQTL analysis were input into GeneMANIA (Warde-Farley et al. 2010). GeneMANIA aims to predict the function of a network of genes and gene-gene interactions. GeneMANIA was accessed through its web-based platform: https://genemania.org/ (Accessed 29/04/2022).

STRING (version 11.5) was used to identify protein-protein interactions (Szklarczyk et al. 2018). The search performed was "multiple proteins". This was accessed through its web platform: https://string-db.org/ (accessed 29/04/2022).

5.3 Results

5.3.1 Sample demographics

Some MSBB tissue samples were removed to reduce the need for LMEM in the eQTL analysis. The reduced sample demographics can be seen in Table 5-2. It led to a total of 594 samples being included in the eQTL analyses.

	MayoRNAseq	**ROSMAP**	**MSBB (BM10 only)**	**Total**
Sex	F: 50 M: 40	F: 242 M: 127	F: 88 M: 47	F: 380 M: 214
Age at death (years)	Mean: 82.7 SD: 7.6	Mean: 86.4 SD: 4.9	Mean: 84.0 SD: 7.16	Mean: 84.8 SD: 6.5
Diagnosis	AD: 42 (61.9% F) Control: 35 (68.6% F)	AD: 204 (69.1% F) Control: 165 (61.2 % F)	AD: 104 (67.3% F) Control: 31 (58.1% F)	AD: 350 (67.7% F) Control: 244 (58.6% F)
Total samples:	90	369	135	594

Table 5-2 – The summary statistics for samples that were included in the cis- and trans- eQTL analysis

5.3.2 Cis-eQTL analysis of index SNPs and differentially expressed genes from an AD case-control study

A total of 220 cis-eQTL-gene pair associations were found, with eight of them being FDR significant and 24 being nominally significant. The results for the FDR significant eQTLs can be seen in Table 5-3. Three of the seven results (rs12151021 and rs2452170 - which was a significant eQTL twice for genes *SEC1P* and *NTN5*) were on chromosome 19, leading to

concerns that this might be due to linkage with APOE. The SNP rs12151021 had a D' score of 0.26 and 0.11 with rs249358 and rs7412 respectively (the two common SNPs of *APOE*). The SNP rs2452170 had a D' score of 0.06 and 0.02 with rs429358, and rs7412 respectively. As there was some evidence for LD, the analysis was rerun with *APOE* ε4 allele added as a covariate. The addition of *APOE* ε4 as a covariate did not affect the existing results but did increase the significance of rs9381040. The results can be seen in Table 5-4.

GWAS study	SNP	SNP chr:pos:allele	Gene	Gene chr:pos:allele	beta	eQTL p-value	eQTL FDR p-value	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
W21	rs2452170	19:48710247:G	SEC1P	19:48638071-48682245	-0.32	$1.17x10^{-13}$	$2.57x10^{-11}$	1.85	0.06	4.57	$1.72x10^{-08}$
W21	rs2452170	19:48710247:G	NTN5	19:48661407-48673081	-0.30	$2.32x10^{-10}$	$2.55x10^{-08}$	1.85	0.06	4.57	$1.72x10^{-08}$
W21	rs708382	17:44364976:C	FAM171A2	17:44353215-44363853	-0.19	$2.23x10^{-06}$	$1.64x10^{-04}$	-2.07	0.04	4.93	$1.98x10^{-09}$
X19	rs113260531	17:5235685:A	CHRNE	17:4897771-4934438	0.36	$1.24x10^{-05}$	$6.81x10^{-04}$	3.78	$1.56x10^{-04}$	5.82	$5.72x10^{-09}$
X18	rs7225151	17:5233752:A	CHRNE	17:4897771-4934438	0.35	$2.18x10^{-05}$	$9.58x10^{-04}$	3.82	$1.31x10^{-04}$	5.78	$7.29x10^{-09}$
W21	rs12151021	19:1050875:A	WDR18	19:984332-998438	-0.17	$1.25x10^{-04}$	$4.57x10^{-03}$	3.78	$1.56x10^{-04}$	5.82	$5.72x10^{-09}$
W21	rs7209200	17:5066645:T	CHRNE	17:4897771-4934438	0.18	$1.40x10^{-03}$	$4.39x10^{-02}$	1.33	0.18	5.54	$3.11x10^{-08}$

Table 5-3 - Significant Benjamini-Hochberg FDR corrected cis-eQTL results of GWAS index SNPs and AD case-control differentially expressed genes

Genome-wide association study (GWAS): W21 = (Wightman et al. 2021); X19 = (Jansen et al. 2019); X18 =(Marioni et al. 2018) . Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

GWAS study	SNP	SNP chr:pos:allele	Gene	Gene chr:pos:allele	beta	eQTL p-value	FDR p-value	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
W21	rs2452170	19:48710247:G	SEC1P	19:48638071-48682245	-0.33	$4.69x10^{-14}$	$1.03x10^{-11}$	1.85	0.06	4.57	$1.72x10^{-08}$
W21	rs2452170	19:48710247:G	NTN5	19:48661407-48673081	-0.30	$1.87x10^{-10}$	$2.05x10^{-08}$	1.85	0.06	4.57	$1.72x10^{-08}$
W21	rs708382	17:44364976:C	FAM171A2	17:44353215-44363853	-0.19	$1.14x10^{-06}$	$8.39x10^{-05}$	-2.07	0.04	4.93	$1.98x10^{-09}$
X19	rs113260531	17:5235685:A	CHRNE	17:4897771-4934438	0.37	$1.15x10^{-05}$	$6.30x10^{-04}$	3.78	$1.56x10^{-04}$	5.82	$5.72x10^{-09}$
X18	rs7225151	17:5233752:A	CHRNE	17:4897771-4934438	0.35	$2.00x10^{-05}$	$8..80x10^{-04}$	3.82	$1.31x10^{-04}$	5.78	$7.29x10^{-09}$
W21	rs12151021	19:1050875:A	WDR18	19:984332-998438	-0.17	$8.94x10^{-05}$	$3.28x10^{-03}$	3.78	$1.56x10^{-04}$	5.82	$5.72x10^{-09}$
W21	rs7209200	17:5066645:T	CHRNE	17:4897771-4934438	0.18	$1.37x10^{-03}$	0.042	1.33	0.18	5.54	$3.11x10^{-08}$
X18	rs3752231	19:1043639:T	WDR18	19:984332-998438	-0.15	$1.75x10^{-03}$	0.046	5.39	$7.40x10^{-08}$	6.62	$3.62x10^{-11}$
X18	rs9381040	6:41186912:T	TAF8	6:42050513-42087461	0.13	$1.90x10^{-03}$	0.046	-3.73	$1.87x10^{-04}$	-5.85	$5.05x10^{-09}$

Table 5-4 Significant Benjamini-Hochberg FDR corrected cis-eQTL results of GWAS index SNPs and AD case-control differentially expressed genes with APOE E4 allele status added to the MatrixEQTL model. Genome-wide association study (GWAS): W21 = (Wightman et al. 2021); X19 = (Jansen et al. 2019); X18 =(Marioni et al. 2018) . Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

The next step in the analysis was to search the 100kb around each of the index SNPs identified in Table 5-4. None of the index SNPs were the top eQTL in the 100kb surrounding region. Results for each of the SNPs are summarised below:

100kb around rs2452170

A cis-eQTL analysis was re-run including SNPs in the 100kb region either side of rs2452170, to determine associations of these SNPs with *SEC1P* and *NTN5*. 274 SNP-gene pairs were found for SNPs associated with *SEC1P* with 163 being FDR (<0.05) significant.

274 SNP-gene pairs were found for SNPs associated with *NTN5* with 135 being FDR (<0.05) significant. The top eQTLs for *SEC1P* and *NTN5* were jointly rs601338 and rs516246 (Table 5-5) and differed from the previously identified index SNP rs2452170. Both rs601338 and rs516246 were in high LD with the index SNP rs2452170 with an r^2 of 0.85 and 0.84 respectively.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs601338	19:48703417:A	SEC1P	19:48638071-48682245	0.33	$2.34x10^{-14}$	$2.11x10^{-12}$	0.85	1.40	0.16	3.24	0.62
rs516246	19:48702915:T	SEC1P	19:48638071-48682245	0.33	$2.34x10^{-14}$	$2.11x10^{-12}$	0.84	1.44	0.15	3.23	$1.21x10^{-03}$
rs601338	19:48703417:A	NTN5	19:48661407-48673081	0.32	$1.15x10^{-11}$	$1.06x10^{-09}$	0.85	1.40	0.16	3.24	0.62
rs516246	19:48702915:T	NTN5	19:48661407-48673081	0.32	$1.15x10^{-11}$	$1.06x10^{-09}$	0.84	1.44	0.15	3.23	$1.21x10^{-03}$

Table 5-5 – Top eQTLs for SNPs in the 100kb region either side of rs2452170 associated with SEC1P and NTN5

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs708382

130 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs708382 that were associated with *FAM171A2*. 19 of them were FDR (<0.05) significant. The most significant eQTLs for association with *FAM171A2* can be found in Table 5-6. The top eQTL for *FAM171A2* was rs850732 which had an r^2 of 0.92 with rs708382.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs850732	17:44376875:T	FAM171A2	17:44353215-44363853	-0.20	$1.03x10^{-06}$	$1.81x10^{-05}$	0.92	2.50	0.01	4.46	$5.07x10^{-06}$
rs5910	17:44372421:A	FAM171A2	17:44353215-44363853	-0.20	$1.03x10^{-06}$	$1.81x10^{-05}$	0.93	2.48	0.01	4.57	$4.90x10^{-06}$
rs850733	17:44373937	FAM171A2	17:44353215-44363853	-0.20	$1.03x10^{-06}$	$1.81x10^{-05}$	0.93	2.49	0.01	4.58	$4.73x10^{-06}$
rs5911	17:44375697	FAM171A2	17:44353215-44363853	-0.20	$1.03x10^{-06}$	$1.81x10^{-05}$	0.92	-2.50	0.01	4.20	$2.611x10^{-05}$

Table 5-6 - Top eQTLs for SNPs in the 100kb region either side of rs708382 associated with FAM171A2

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs113260531

247 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs113260531 that were associated with *CHRNE*. 81 of them were FDR (<0.05) significant. The top eQTL for association with *CHRNE* can be found in Table 5-7. The top eQTL for *CHRNE* was rs113260531 which had an r^2 of 0.63 with rs113260531.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs2289101	17:5138409:G	CHRNE	17:4897771-4934438	0.42	$7.84x10^{-07}$	$2.34x10^{-05}$	0.63	-2.86	$4.28x10^{-03}$	3.01	$2.64x10^{-03}$

Table 5-7 - Top eQTL for SNPs in the 100kb region either side of rs113260531 associated with CHRNE

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs7225151

The results for rs7225151 were similar to those for rs113260531. 244 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs7225151 that were associated with *CHRNE*. 81 of them were FDR (<0.05) significant. The top eQTL for association with *CHRNE* can be found in Table 5-8. The top eQTL for *CHRNE* was rs2289101 which had an r^2 of 0.61 with rs7225151.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs2289101	17:5138409:G	CHRNE	17:4897771-4934438	0.42	$7.84x10^{-07}$	$2.34x10^{-05}$	0.61	-2.86	$4.28x10^{-03}$	3.01	$2.64x10^{-03}$

Table 5-8 - Top eQTL for SNPs in the 100kb region either side of rs7225151 associated with CHRNE

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs12151021

418 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs113260531 that were associated with *WDR18*. 184 of them were FDR (<0.05) significant. The top eQTL for association with *WDR18* can be found in Table 5-9. The top eQTL for *WDR18* was rs11667292 which had an r^2 of 0.04 with rs12151021.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs11667292	19:998687:T	WDR18	19:984332-998438	-0.35	$1.27x10^{-21}$	$5.31x10^{-19}$	0.04	1.48	0.14	3.84	$1.24x10^{-04}$

Table 5-9 - Top eQTL for SNPs in the 100kb region either side of rs12151021 associated with WDR18

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs7209200

236 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs7209200 that were associated with *CHRNE*. 102 of them were FDR (<0.05) significant. The top eQTL for association with *CHRNE* can be found in Table 5-10. The top eQTL for *CHRNE* was rs7222708 which had an r^2 of 0.14 with rs7209200.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs7222708	17:4984816:T	CHRNE	17:4897771-4934438	0.79	$2.20x10^{-17}$	$5.19x10^{-15}$	0.14	2.94	$3.26x10^{-03}$	3.72	$1.98x10^{-04}$

Table 5-10 - Top eQTL for SNPs in the 100kb region either side of rs7209200 associated with CHRNE

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs3752231

423 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs3752231 that were associated with WDR18. 184 of them were FDR (<0.05) significant. The top eQTL for association with WDR18 can be found in Table 5-11. The top eQTL for WDR18 was rs11667292 which had an r^2 of 0.006 with rs3752231.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs11667292	19:998687:T	WDR18	19:984332-998438	-0.35	$1.27x10^{-21}$	$5.31x10^{-19}$	0.006	1.48	0.14	3.84	$1.24x10^{-04}$

Table 5-11 - Top eQTL for SNPs in the 100kb region either side of rs7209200 associated with WDR18

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

100kb around rs9381040

252 SNP-gene pairs were identified in the cis-eQTL analysis including SNPs 100kb either side of rs9381040 that were associated with *TAF8*. 144 of them were FDR (<0.05) significant. The top eQTL for association with *TAF8* can be found in Table 5-12. The top eQTL for *TAF8* was rs35047410 which had an r^2 of 0.13 with rs9381040.

SNP	chr:pos:allele	Gene	Gene chr:pos:allele	eQTL beta	eQTL p-value	eQTL FDR p-value	LD r^2 with index SNP	Kunkle Z-score	Kunkle p-value	Wightman Z-score	Wightman p-value
rs35047410	6:41203216:A	TAF8	6:42050512-42087461	-0.16	$3.00x10^{-05}$	$1.71x10^{-03}$	0.13	1.86	0.06	3.37	$7.55x10^{-04}$

Table 5-12** - **Top eQTL for SNPs in the 100kb region either side of rs7209200 associated with WDR18

Kunkle refers to Kunkle et al. GWAS (Kunkle et al. 2019), Wightman refers to Wightman et al Genome-wide association study by proxy (Wightman et al. 2021)

In summary, in an analysis of AD GWAS index SNPs and AD differentially expressed genes, seven cis-eQTLs were identified. When adjusting for *APOE* ε4 carrier status, all seven remained statistically significant and an additional eighth became statistically significant. For all eight eQTLs, when refining the signal within a 100kb region, none of the GWAS index SNPs remained the strongest signal. For example, eQTL rs2452170 two stronger signals were found which were rs601338 with an r^2 of 0.85 and rss16246 with an r^2 of 0.84. For other eQTLs, pseudo independent SNPs were found such as for rs3752231 a stronger signal was found at rs11667292 (r^2 = 0.006) and rs9381040 a stronger signal was found at rs35047410 (r^2 = 0.13).

5.3.3 Trans-eQTL analysis of GWAS/GWAX index SNPs and previously identified differentially expressed genes between AD cases and controls

For this analysis, trans-eQTLs were defined to be a GWAS index SNP associated with a differentially expressed gene (as identified in chapter 4) located at least 5Mb apart or on separate chromosomes. Using MatrixEQTL, 102,150 SNP-gene pair associations were found. After FDR correction, four SNP-gene associations remained significant. These all include the eQTL rs5011436 being associated with the eGenes *SST*, *TAC1*, *MAF1* and *SCGN*. The SNP rs5011436 is an intron variant in *TMEM106B* which has been previously identified as a key ageing human brain transcriptome regulator (Yang et al. 2020).

SNP	SNP chr:pos:allele	Gene	Gene pos	eQTL beta	eQTL p-value	eQTL FDR p-value
rs5011436	7:12229132:C	SST	3:187668912-187670394	-0.24	$6.63x10^{-10}$	$6.77x10^{-05}$
rs5011436	7:12229132:C	TAC1	7:97732084-97740472	-0.27	$2.03x10^{-09}$	$1.04x10^{-04}$
rs5011436	7:12229132:C	MAF1	8:144104461-144107611	0.19	$8.37x10^{-08}$	$2.85x10^{-03}$
rs5011436	7:12229132:C	SCGN	6:25652201-25701783	0.22	$1.21x10^{-06}$	0.03

Table 5-13 FDR significant trans-eQTL results for GWAS index SNPs and differentially expressed genes

Chr = Chromosome; pos = position

Inputting *SST*, *TAC1*, *MAF1* and *SCGN* into GeneMania using default settings (Warde-Farley et al. 2010) gave a predicted gene-gene interaction network of 24 genes which can be seen in Figure 5-3. As only four genes were input into GeneMania, the predicted network of genes was based on GO annotation patterns. This included terms such as neuropeptide receptor activity and positive regulation of cilium movement and the results are summarised in Table 5-14.

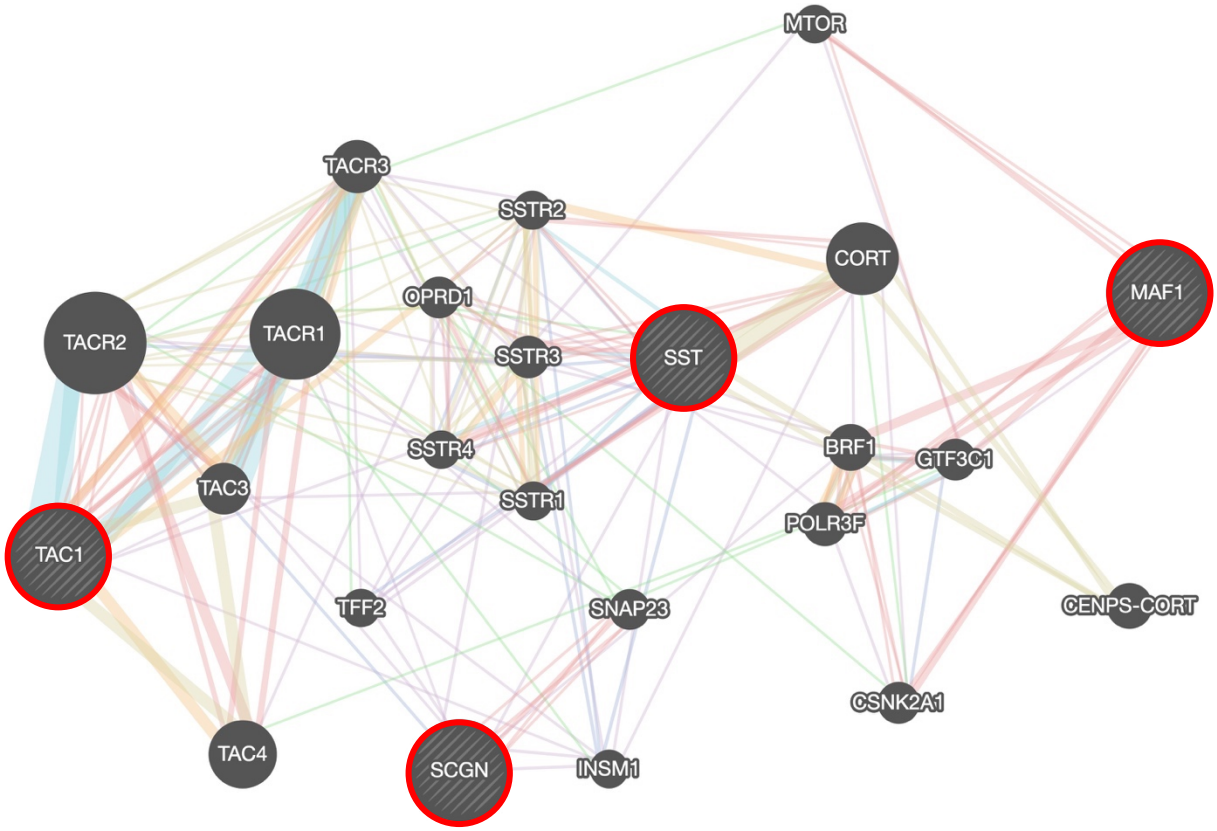


Figure 5-3 Results of GeneMania predicted gene-gene interactions using significant genes from trans-eQTL analysis of GWAS index SNPs and AD case-control differentially expressed genes.

Red circles indicate input genes from the trans-eQTL analysis. Colour of lines: red = physical interaction; purple = co-expression; orange= predicted; dark blue = co-localization; green = genetic interactions; light blue = pathway; yellow = shared protein domain. Colours are produced by software.

GO term enriched in genes in GeneMANIA network	FDR p-value	Coverage
neuropeptide receptor activity	$1.75x10^{-09}$	6/21
G protein-coupled peptide receptor activity	$2.56x10^{-07}$	7/102
peptide receptor activity	$1.06x10^{-06}$	7/132
positive regulation of cilium movement	$1.39x10^{-06}$	4/10
positive regulation of cilium-dependent cell motility	$1.39x10^{-06}$	4/10
transcription by RNA polymerase III	$2.36x10^{-06}$	5/38
regulation of flagellated sperm motility	$2.36x10^{-06}$	4/12
regulation of cilium movement involved in cell motility	$2.67x10^{-05}$	4/22
regulation of cilium-dependent cell motility	$2.67x10^{-05}$	4/22
G protein-coupled receptor activity	$2.92x10^{-05}$	7/252
flagellated sperm motility	$5.22x10^{-05}$	4/27
regulation of cilium movement	$5.58x10^{-05}$	4/28
regulation of microtubule-based movement	$2.28x10^{-04}$	4/40
positive regulation of reproductive process	$7.23x10^{-04}$	4/54
sperm motility	$1.72x10^{-03}$	4/68
cilium movement involved in cell motility	$2.15x10^{-03}$	4/73
cilium-dependent cell motility	$2.92x10^{-03}$	4/80
cilium or flagellum-dependent cell motility	$4.61x10^{-03}$	4/91
cilium movement	$6.17x10^{-03}$	4/99
regulation of reproductive process	$6.29x10^{-03}$	4/101
peptide binding	$9.76x10^{-03}$	5/247
sperm flagellum	0.02	3/47

Table 5-14 Results of the predicted functional enrichment provided by GeneMANIA of the genes as per figure 5-3

Changing the settings of GeneMANIA to understand the interactions of only the four genes and not the predicted genes shows that SST is co-expressed with SCGN and TAC1. No GO terms were enriched for these four genes alone (Figure 5-4).

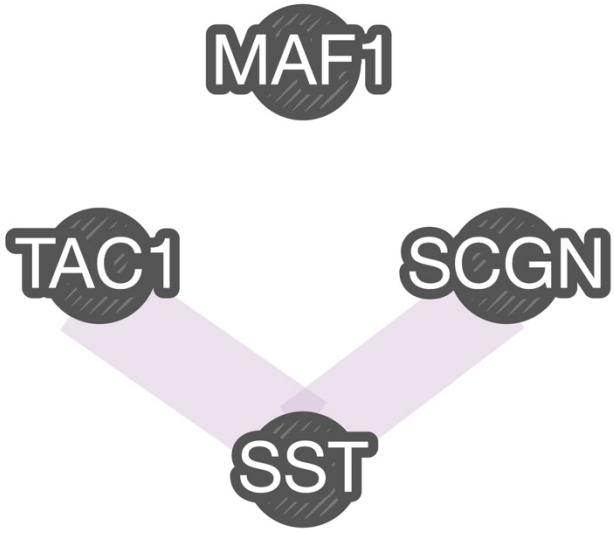


Figure 5-4 – Restricting GeneMANIA analysis to only the four significant eGenes: MAF1, TAC1, SCGN and SST. The light purple lines indicate correlated expression patterns (co-expression) between genes SST and SCGN and SST and TAC1.

Inputting the four proteins into STRING under default options, the protein-protein interaction (PPI) network identified that the proteins SCGN, SST and TAC1 were at least partially connected as a group (p-value = 1.25×10^{-04}) with an average local clustering coefficient of 0.75 (Figure 5-5).

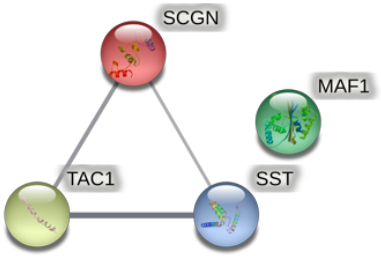


Figure 5-5 – Protein-protein interaction network of four genes from trans-eQTL analysis under default settings.

Adjusting the settings of STRING to remove “textmining” as a source of evidence for interaction resulted in TAC1, SCGN and SST no longer forming a connected network as shown in Figure 5-6.

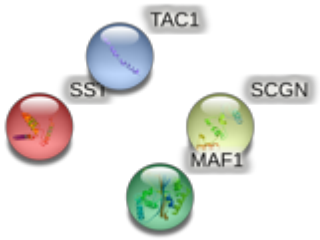


Figure 5-6 - Protein-protein interaction network of four genes from trans-eQTL analysis with “textmining” as a source of evidence for interaction removed.

The paper that indicated an association between SCGN, TAC1 and SST through text-mining included a review article that briefly described work that found that secretagogin^{+} neurons in mice were found to express mRNA transcripts for somatostatin and Tac1 (Maj et al. 2019).

5.4 Discussion

The overall aim of this work was to investigate index SNPs as identified from five AD GWAS/GWAX studies and their relationship with genes that had previously been shown to be differentially expressed in AD cases and controls (in chapter 4) using eQTL analysis.

The first aim was to identify which index SNPs were cis-eQTLs for AD differentially expressed genes. This identified eight SNP-gene pair associations. None of the eight SNPs (comprising the SNP-gene pairs), were identified as index SNPs through GWAS, only GWAX (Table 5-3). Six of these eight SNPs were at least nominally associated to AD in the GWAS. This could be due to the GWAX having larger sample numbers and therefore increased power in comparison to the GWAS, due to the inclusion of proxy cases (GWAX included 1,125,563 individuals whereas the GWAS included 94,437 individuals).

In the cis-eQTL analysis searching the 100kb region either side of the index SNP none of the top eQTLs were GWAS/GWAX index SNPs. Each of the eQTL SNPs were checked for their association with AD in both the Kunkle et al. GWAS and Wightman et al. GWAX summary statistics. Each are the largest of their respective study types at the time of study. Across the board, the Kunkle et al. GWAS p-values were less significant in comparison to the Wightman et al. GWAX p-values. Like the previous paragraph, this could be due to GWAX having larger sample numbers in comparison to the GWAS and therefore increased power.

Overall, few cis-eQTLs for differentially expressed genes were found through this analysis. The search for eQTLs was expanded to the 100kb region either side of the index SNP, as it is hypothesised that it is likely that the true causal variant is within this distance (Wu et al. 2017). This lack of identified cis-eQTLs could be either because the eQTL analysis was underpowered or that AD index SNPs (or their tagged SNPs) are not associated with the genes identified as differentially expressed (through mRNA profiling).

For the cis-eQTL analysis for both index SNPs and the expanded 100kb region either side of the index SNPs, Z-scores and p-values from the Kunkle et al. GWAS and Wightman et al. GWAX were included in the results tables. The association of eQTLs to AD was generally inconsistent between the two studies. As previously described, the GWAX is by far the larger of the two studies, but it is not yet clear if the differences are due to an increase in statistical power or if adding the proxies is increasing the heterogeneity of the samples and leading to the differences.

GWAS in AD have been criticised for being too heterogenous. In GWAS, the AD clinical phenotype is used and often lacks inclusion of specific biomarkers or neuropathologically defined phenotypes based on neurofibrillary tangles or amyloid plaques. Therefore, individuals which appear to have clinical AD may be included when they have non-AD related neuropathologies such as vascular disease pathology or Lewy body pathology (Andrews et al. 2020). It is impossible to discern whether the AD cases included in the analysis are true AD cases as relatively few have been confirmed pathologically. Additionally, it is impossible to know if the controls are true controls. As AD occurs mainly in later life, it could be that controls later go on to develop disease. Including proxy cases in GWAX could be adding to the heterogeneity of samples included.

A natural extension of the cis-eQTL work presented in this chapter would have been a summary-mendelian randomization (SMR) analysis. Briefly, an SMR tests if the effect size of a SNP on the phenotype is mediated by gene expression using summary statistics from GWAS and independent eQTL study data (Zhu et al. 2016). A transcriptome-wide SMR analysis has previously been performed which uses brain cortex eQTL data from the ROSMAP cohort meta-analysed with eQTL data from GTEx and the CommonMind Consortium (CMC) for a sample size of 1194, which is approximately double that available in the analysis for this thesis. Using those data, SMR analysis identified 12 genes with a significant association with AD. Four of these genes passed the heterogeneity in dependent instruments (HEIDI) test (*NDUFS2*, *RP11-385F7.1*, *PRSS36* and *AC012146.7*) (Lee et al. 2022). *NDUFS2*, *PRSS36* and *AC012146.7* were not differentially expressed between AD cases and controls in the analysis in chapter four (*RP11-385F7.1* was not included in the analysis).

Another SMR was reported using the same AMP-AD cohorts as the work in this thesis (ROSMAP, MayoRNAseq, and MSBB) so would have had a large sample overlap with the samples included in this chapter. They found eight AD candidate risk genes *APOC*, *EED*, *CD2AP*, *CEACAM19*, *CLPTM1*, *MTCH2*, *TREM2*, and *KNOP1* when using Kunkle et al. summary statistics. They also performed a replication using Jansen et al. summary statistics, and only two of these eight genes did not replicate (*MTCH2* and *KNOP1*) (Gockley et al. 2021). None of these were differentially expressed so these genes were not included in the cis-eQTL analysis in this chapter.

The final aim was to perform a trans-eQTL analysis to investigate the relationship between GWAS index SNPs and AD differentially expressed genes. Four SNP-gene pairs remained significant after FDR correction. All four genes (*SST*, *TAC1*, *MAF1* and *SCGN)* were associated with the C allele of SNP rs5011436, which is an intron variant in *TMEM106B.* The SNP rs5011436 was identified as an AD index SNP in the Wightman et al. GWAX (Wightman et al. 2021). The C allele of rs5011436 has previously been shown to be associated with AD and changes in brain morphology in another study (Monereo-Sánchez et al. 2021).

The *TMEM106B* locus had already been identified as a risk factor for frontotemporal degeneration with TDP-43 inclusions (Van Deerlin et al. 2010). It has also previously been identified as a key ageing human brain transcriptome regulator (Yang et al. 2020). Some have considered the role of the TMEM106B locus and protein in neurodegeneration to be controversial (Fan et al. 2022). Amyloid filaments of TMEM106B have been found in human brains, but distinct TMEM106B folds do not characterise different diseases. TMEM106B filaments have also been found in individuals without a neurodegenerative disease. Additionally it has been shown that number of TMEM106B filaments increase with age so it has been suggested that TMEM106B filaments form in an age-dependent manner with no clear mechanism to disease (Fan et al. 2022; Schweighauser et al. 2022). Although the loci have previously been associated with neurodegeneration, more work is needed to understand this relationship fully.

The variant in the *TMEM106B* locus was found to be associated with four genes: *MAF1*, *TAC1*, *SCGN* and *SST*. *MAF1* is a protein coding gene, and little is known of its potential role in AD. *TAC1* is a protein coding gene that encodes four products of the tachykinin peptide hormone family. *TAC1* has previously been found to be differentially expressed between AD patients and controls in an experiment using microarray data from the Gene Expression Omnibus database. The authors of this study identified *TAC1* as a hub gene associated with cognitive decline using tools including GO and KEGG enrichment analysis and PPI network analysis. They validated this finding using RT-qPCR and found that downregulation of *Tac1* was associated with cognition in the hippocampus of APP/PS1 mice leading the authors to conclude that downregulation of *TAC1* is associated with cognition in the hippocampus of AD patients (Liu et al. 2021). *SCGN* is a protein coding gene that encodes for secretagogin, a secreted calcium sensor. A significantly reduced level of SCGN has been reported in the hippocampus of a mouse model of AD and in Parkinson's patients and is predicted to lead to an accumulation of toxic fibrils (Chidananda et al. 2019). *SST* encodes the neuropeptide hormone somatostatin (SST) which is expressed throughout the brain. Key functions of SST include modulating cortical circuits, and cognitive function. It has been repeatedly reported that *SST* expression is reduced in AD patients and mouse models both in the brain and cerebrospinal fluid (Song et al. 2021). Additionally, two independent GWAS in Finnish and Chinese cohorts identified the *SST* gene as a genomic region associated with AD risk (Vepsäläinen et al. 2007; Xue et al. 2009).

The findings from this trans-eQTL analysis could indicate a potential mechanism for the involvement of these four genes in AD. The evidence for *SST*, *TAC1*, *MAF1* and *SCGN* being candidate genes are that they are trans-regulated by an AD locus, and they are differentially expressed in AD cases and controls. In addition, the PPI network using STRING identified that that SST, SCGN and TAC1 were related through text-mining scientific literature. Previous genetic work has also implicated the *SST* locus in AD risk before (Vepsäläinen et al. 2007; Xue et al. 2009). Functional follow-up work would be required to confirm these findings.

One limitation of the approach taken in this analysis is that only differentially expressed genes, identified by this study, were investigated. This means that many other potential genes

of interest would have been excluded from this analysis. The eQTL data set available for use in this analysis is relatively small (n=594). Therefore, focusing on the differentially expressed genes helps reduce the multiple hypothesis testing burden, especially for the trans-eQTL analysis.

A second limitation is that the gene expression data are sourced from brain cortex. Although the brain is relevant to AD it could be that it is too broad a region, and eQTLs from other more as specific brain regions, or specific cellular populations such as microglia or monocytes, may be more informative.

A limitation of the trans-eQTL analysis is that a window between SNP and gene of at least 5Mb was used to define a trans-eQTL. This is a common definition used in trans-eQTL studies. It does mean information about eQTLs in the space between 1 to 5 Mb and their associations are not captured. A final limitation is that the eQTLs were generated from a relatively small dataset, especially for the trans-eQTL analysis. It is highly likely that the trans-eQTL analysis was underpowered.

In conclusion, very few cis-eQTLs for AD case-control differentially expressed genes were identified. This could be due to the study being underpowered or that AD index SNPs (or their tagged SNPs) are not associated with the genes identified as differentially expressed through mRNA profiling. Findings also provide evidence that the association of the intron variant rs5011436 in *TMEM106B* to AD may be mediated through *SST*, TAC1, and *MAF1*. As *TMEM106B* itself was unchanged, the mechanism by which trans-effects may be acting is unclear. The TMEM106B variant rs5011436 is an intronic variant and is in high LD ($r2 > 0.8$) with many other intronic variants. It could be that one of these variants is impacting on splicing by interfering with splice site recognition. Alternatively, the variant rs5011436 is also in high LD with rs3173615 which is a missense variant so may also be a clue to a potential mediating mechanism. Firm conclusions cannot be drawn at present and functional follow-up studies are needed to confirm findings.

Chapter 6 – A comparison of transcriptome-wide association studies and differential gene expression analysis in Alzheimer's disease.

6.1 Introduction

6.1.1 An overview of transcriptome-wide association studies

One family of methods that aims to prioritise causal genes at GWAS loci are transcriptome-wide association studies (TWAS). TWAS can utilise either individual-level GWAS data through methods such as PrediXcan (Gamazon et al. 2015) or with GWAS summary statistics using methods such as Fusion (Gusev et al. 2016) and S-PrediXcan (Barbeira et al. 2018). In essence, TWAS works by using eQTL reference panels to train a predictive model of gene expression from genotype. This model is then used to predict expression for individuals in the GWAS. A TWAS using individual-level data will predict expression directly into genotyped samples using effect sizes from the eQTL reference panel and measure the association between predicted expression and trait. In contrast, a TWAS using GWAS summary statistics will indirectly estimate the association between predicted expression and trait as a weighted linear combination of SNP-trait standardised effect sizes while accounting for LD among SNPs (Gusev et al. 2016). Both approaches are summarised in Figure 6-1.

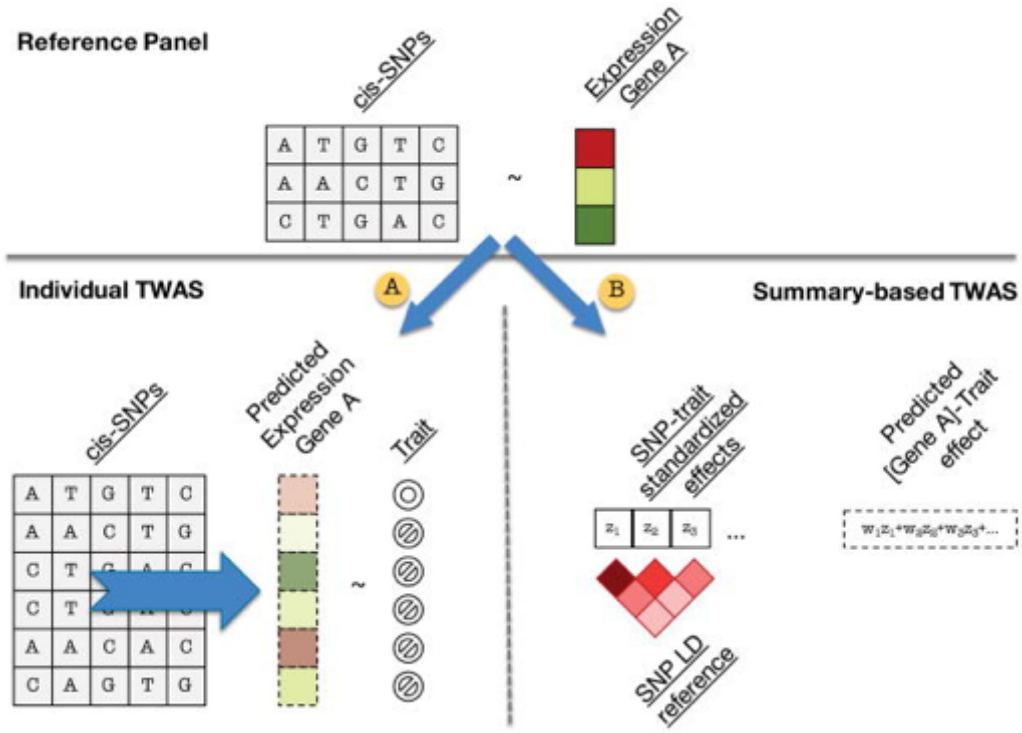


Figure 6-1 – An overview of the individual and summary-based TWAS approaches. Reproduced from: (Gusev et al. 2016)

TWAS have become a popular tool for prioritising candidate causal genes (genes mediating the phenotypic effects of causal genetic variants) and exploring gene expression with a genetic basis (Wainberg et al. 2019). One benefit of TWAS is that the multiple hypothesis testing burden is smaller in comparison to other variant based approaches. This is because TWAS aggregates the effects of multiple eQTLs and directly tests associations between genes and diseases. Therefore in a TWAS, only the number of genes needs to be accounted for, which is orders of magnitude lower than variant-based tests (Li and Ritchie 2021).

Summary-level TWAS are more commonly used than individual-level TWAS. Individual-level TWAS provide more accurate estimates of gene-trait associations but have a higher computational burden in comparison to summary-level TWAS. Individual-level GWAS data are not as easily accessible to the research community in comparison to GWAS summary statistics. Due to these two factors, summary-level TWAS are the more common type of TWAS in the literature (Li and Ritchie 2021). One disadvantage to summary-level TWAS is that extra noise can be introduced due to mismatch between the reference LD panel and the LD

structure of the GWAS cohort (Wainberg et al. 2019). Therefore, the GWAS needs to have a large enough sample size to achieve enough statistical power to overcome this additional noise (Li and Ritchie 2021). Due to these limitations, it is important that TWAS are interpreted with care and supported with additional validation (Wainberg et al. 2019; Li and Ritchie 2021).

6.1.2 TWAS in Alzheimer's disease

Several TWAS have been performed to identify candidate genes for AD risk. Raj et al. used the (Lambert et al. 2013) IGAP GWAS summary statistics, dorsolateral prefrontal cortex (DLPFC) samples from the ROSMAP study for the gene expression data, and FUSION to perform their TWAS. They identified 21 genes associated with AD (Raj et al. 2018).

Hao et al. also performed a TWAS using the same 2013 Lambert et al. IGAP summary statistics but used gene expression data from whole blood, adipose and brain tissues from GTEx. They identified 15 AD risk associated genes after Bonferroni correction including 11 known AD risk genes such as *BIN1*, *TOMM40*, *PICALM, CR1* and *CLU* (Lambert et al. 2013; Hao et al. 2019).

A study by Hu et al. used their software UTMOST to perform a TWAS using (Lambert et al. 2013) IGAP summary statistics and used gene expression data from 44 tissues available in GTEx version 6 (Hu et al. 2019). This was in addition to liver gene expression data from STARNET and eQTL data from three immune cell types (CD14+ monocytes, CD16+ neutrophils and naïve CD4+ T cells) from the BLUEPRINT consortium. They performed single-tissue association tests and then a cross-tissue analysis, where they identified 68 putative AD risk associated genes. They performed two replications using Alzheimer's Disease Genetics Consortium (ADGC) summary statistics that were not included in the IGAP analysis and a GWAS-by-proxy (GWAX) (Liu et al. 2017). 17 and 15 genes of the identified 68 replicated in the ADGC and GWAX respectively (Hu et al. 2019).

Gerring et al. used S-PrediXcan to analyse (Lambert et al. 2013) IGAP summary statistics but with eQTL data from 48 tissues using GTEx version 7 (Gerring et al. 2020). They identified 50 tissue-specific AD risk associated genes and a meta-analysis of the tissue-specific associations identified 73 genes associated with AD (Gerring et al. 2020).

Sun et al. also used an UTMOST framework with S-PrediXcan approach to perform their TWAS. They used 10 different brain tissues and spleen from GTEx version 8 to build their gene expression genetic prediction models. They used the larger Jansen et al. GWAX in their analysis (Jansen et al. 2019) and they found 53 genes associated with AD risk (Sun et al. 2021).

The AMP-AD consortia used a FUSION pipeline to perform their TWAS. They utilised eQTLs from the ROSMAP, MayoRNAseq and MSBB cohorts to produce trained expression weights. These weights were then used in their TWAS with the Kunkle et al. AD GWAS summary statistics (Kunkle et al. 2019). They found eight genes associated with AD (*APOC1*, *EED*, *CD2AP*, *CEACAM19*, *CLPTM1*, *MTCH2*, *TREM2* and *KNOP1*) which were also supported by SMR evidence (Gockley et al. 2021).

Harwood et al. used FUSION for their TWAS analysis. They utilised monocyte data and summary statistics from the Kunkle et al. GWAS, with a replication using Marioni et al. GWAX summary statistics (Marioni et al. 2018; Kunkle et al. 2019; Harwood et al. 2021). They also compared whether AD-associated changes in gene expression were specific to monocytes or across tissues. The authors did this by performing TWAS analyses on the Kunkle et al. GWAS summary statistics using expression weights from the GTEx (version 7) consortium, the young Finns study (YFS) whole blood, the Netherlands twin register peripheral blood, and the Common Mind Consortium. They found an association between changes in gene expression in both naïve and induced CD14+ monocytes and AD for nine genes. Three of these genes replicated (*PVR*, *PTK2B* and *MS4A6E*) when using the independent Marioni et al. GWAX summary statistics (Marioni et al. 2018; Harwood et al. 2021). The authors also found that the *PTK2B* signal was specific to blood, and the *MS4A6E* signal was specific to monocytes (Harwood et al. 2021).

TWAS studies in both AD and other disorders can only utilise cis-eQTLs at present. This is partly due to a lack of trans-eQTL resources in disease-relevant tissue. This is because a resource of this type would require very large sample sizes to have enough statistical power to detect trans-eQTLs. If such a resource were available, it is not clear how well TWAS would perform using trans-eQTLs. This is due to cis- and trans-eQTLs potentially having overlapping effects such as a cis-eQTL affecting the expression of a nearby gene that is a transcription factor, which then regulates the transcription of a distant gene (Yang et al. 2017; Võsa et al. 2018)(Li and Ritchie 2021).

This long list of TWAS shows that many studies have been performed in AD, with many suggested candidate causal genes. Differentially expressed genes identified through TWAS only use common cis-eQTLs to inform differential expression. It has been estimated that common cis-eQTLs explain only around 10% of genetic variance in expression (Grundberg et al. 2012). In contrast, differentially expressed genes from mRNA profiling will be influenced by environment, genetics, stochastic factors and technical artefacts. DGE results are often criticised for being a result of disease rather than a cause of disease (Porcu et al. 2021). At present, it is not clear how the two methods of TWAS and DGE analysis compare in AD.

Studies in other diseases have compared their significant TWAS results with the results from mRNA expression profiling to check for any DGE overlap. This is often performed to provide additional evidence for the involvement of TWAS-identified genes in disease risk. This has been performed in osteomyelitis (Zhang et al. 2020a), attention deficit hyperactivity disorder (ADHD) (Qi et al. 2019), and sporadic amyotrophic lateral sclerosis (ALS) (Li et al. 2022a). The study on osteomyelitis identified 86 candidate genes through TWAS, eight of which that were also differentially expressed between cases and controls (Zhang et al. 2020a). The ADHD TWAS identified 148 candidate genes and eleven of which were also differentially expressed between cases and controls (Qi et al. 2019). The ALS TWAS identified 761 candidate genes, 107 of which were also differentially expressed in their mRNA expression profiling study (Li et al. 2022a). However none of these studies performed any statistical analysis to identify if these overlaps were more than would be expected through chance alone. This comparative approach has not been done in AD and it is not yet clear how the genes suggested from TWAS,

which identifies the cis- component of gene expression compares to those genes suggested from a DGE analysis using mRNA data.

6.1.3 Aims

The overall aim of this chapter is to compare differentially expressed genes through the AD case-control mRNA profiling study performed in chapter four to the statistically significant genes from three AD TWAS.

The first aim is to compare which genes were significant across all four resulting datasets, to see if any genes replicated across methods.

The second aim was to look at the correlation of Z-scores of genes from the four studies to compare the similarity of results. Initially Z-scores from the different TWAS were compared to identify which TWAS were the most similar. Then the Z-scores from the DGE analysis were compared to TWAS.

The final aim was to perform statistical tests (hypergeometric tests and Spearman rank correlation) to determine if any observed overlap of candidate genes was more than expected through chance alone.

6.2 Methods

6.2.1 Differential gene expression

I described the generation of the AD case-control differentially expressed genes in Chapter 4. There were 16,485 genes which were tested for statistical significance of differential expression between AD cases and controls. Differentially expressed genes were defined as having a Benjamini-Hochberg FDR corrected p-value < 0.05 and this amounted to 1270 differentially expressed genes in total.

6.2.2 Selection of TWAS

Three TWAS in AD were selected for this analysis: a TWAS by Harwood et al. and the two AMP-AD TWAS. The two AMP-AD TWAS differ as one uses Kunkle et al. GWAS summary statistics and the other uses Jansen et al. GWAX summary statistics. These three studies were selected as they are the most recent TWAS in AD. The TWAS by Harwood et al. and the AMP-AD (GWAS) TWAS both use the largest AD case-control GWAS summary statistics (Kunkle et al.) in their TWAS analysis. The Harwood et al. TWAS that was selected for this analysis used the latest GTEx eQTL reference panel generated from brain cortex samples. This was chosen as the samples from my DGE analysis came from six different regions of the brain cortex, so this tissue is the most similar match. The AMP-AD TWAS were selected as an additional comparator as the AMP-AD TWAS utilise the same ROSMAP, MayoRNAseq and MSBB cohort samples as my DGE analysis. As a result, there is likely to be a large sample overlap between those individuals included in my DGE analysis and the AMP-AD TWAS. Therefore, it will be possible to compare the TWAS methods between one another but also compare results from the TWAS method to the DGE method with a large overlap of samples. A summary of the TWAS methods can be found in Table 6-1.

	AMP-AD (GWAS) TWAS	**AMP-AD (GWAX) TWAS**	**Harwood et al. TWAS**
GWAS summary statistics	Kunkle et al. (Kunkle et al. 2019)	Jansen et al. (Jansen et al. 2019)	Kunkle et al. (Kunkle et al. 2019)
eQTL reference panel	AMP-AD	AMP-AD	GTEx brain cortex

Table 6-1 – An overview of the summary statistics and eQTL panels used in the three TWAS.

6.2.3 Harwood et al. TWAS

As described in their paper, Harwood et al. used GWAS stage 1 summary statistics from Kunkle et al. for use in their TWAS (Kunkle et al. 2019; Harwood et al. 2021). GTEx (version 7) weights for 3883 genes were also used in the TWAS along with the 1000 genomes reference panel (European population) which were both downloaded from the FUSION website: (http://gusevlab.org/projects/fusion/ . Results for the brain cortex TWAS used for the analysis in this chapter were obtained directly from the authors (Harwood et al. 2021). FDR corrected p-values were not included in the dataset, so I calculated Benjamini-Hochberg corrected p-values using the *p.adjust* function in R.

6.2.4 AMP-AD TWAS

TWAS study methods were originally described in their paper (Gockley et al. 2021). This study used 'CEU' ancestrally matched genotype and RNA-seq expression profiling across the MayoRNAseq, MSBB and ROSMAP AMP-AD cohorts to train predictive weights for 6780 genes. This enabled the authors to impute the genetic component of expression in patients directly from genotype. The training set consisted of 790 genotypes paired to 888 RNA-seq profiles across six cortical tissues including temporal cortex, DLPFC, and Brodmann areas 10, 22, 36 and 44. A modified FUSION pipeline was implemented to support some individuals having multiple samples from different tissues. For their analyses, they performed an analysis using Kunkle et al.'s GWAS summary statistics and another using the Jansen et al. GWAX summary statistics (Jansen et al. 2019; Kunkle et al. 2019). Results for both AMP-AD TWAS used for the analyses in this chapter were downloaded from Synapse https://www.synapse.org/#!Synapse:syn22231399 (accessed: 07 April 2022).

6.3 Results

Initially a comparison of Z-scores from the four studies were compared pair-wise and Pearson correlations were calculated. Plots are shown in Figure 6-2.

The two studies that were the most closely correlated were the AMP-AD TWAS (which used the Kunkle et al. GWAS summary statistics) and the Harwood et al. TWAS (which also used the Kunkle et al. GWAS summary statistics (Figure 6-2d) but used different gene expression reference panels. The Z-score from my DGE analysis were not significantly correlated with any of the TWAS, as seen in Figure 6-2 a, b and c.

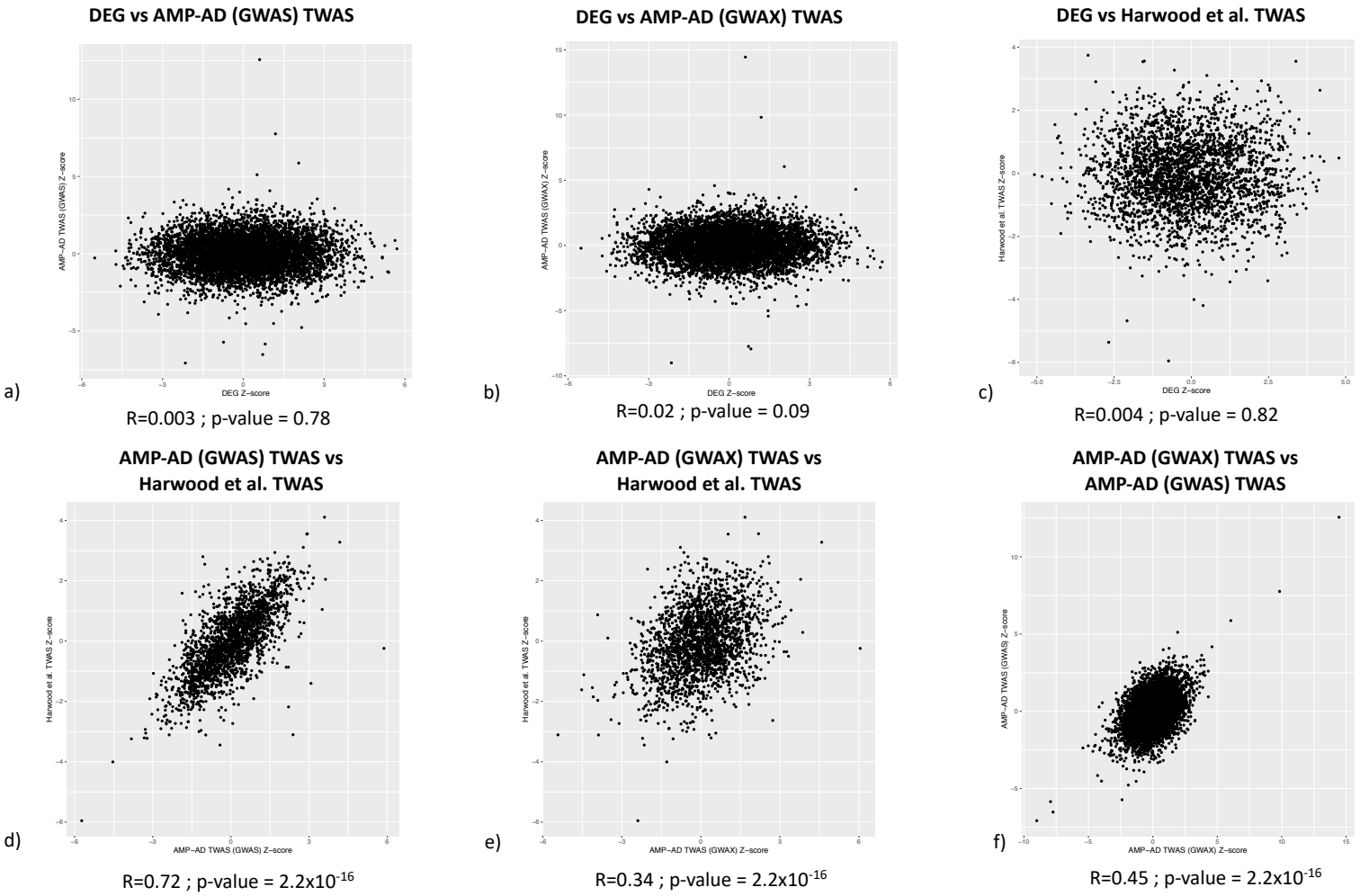
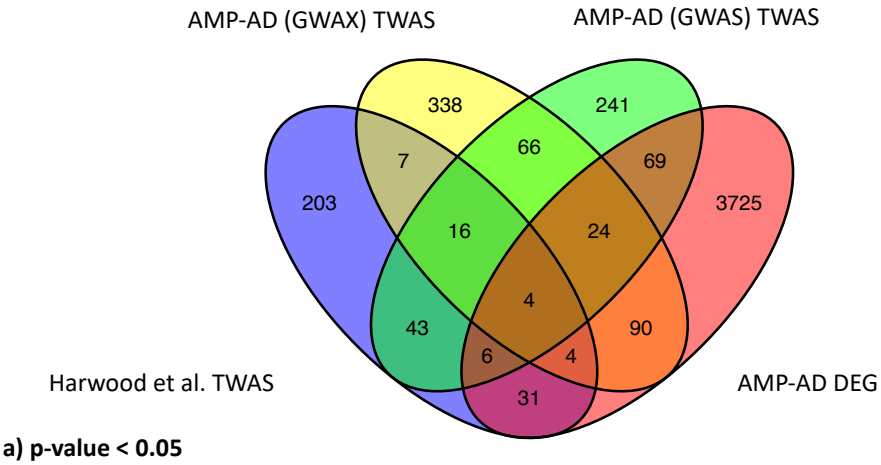


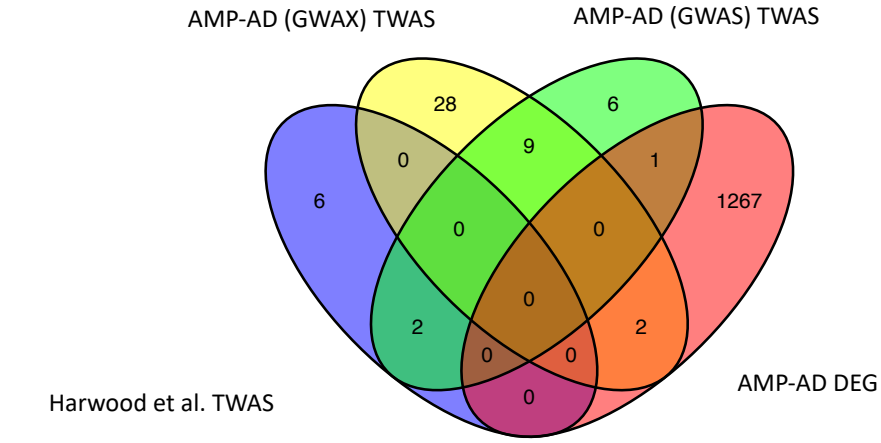
Figure 6-2 – A comparison of Z-scores from three AD TWAS and differential gene expression (DEG) analysis with Pearson correlations.

The results from the AD case-control mRNA differential expression analysis were compared with three sets of TWAS results. No genes were significant across all four studies at FDR < 0.05 nor FDR < 0.1. Only four genes were nominally significant across the four studies. These were *NPEPPS*, *QPCT*, *PRKD3* and *DECR2*. *NPEPPS* and *QPCT* are protein coding genes that have been previously implicated in AD outside of these DGE/TWAS analyses (Kudo et al. 2011; Song et al. 2015; de Rojas et al. 2021). Although four were found to be nominally significant, the z-score direction differs when comparing results from TWAS versus DGE analysis for *QPCT* and *PRKD3* which can be seen in Table 6-2. Differing z-scores and insignificance after multiple hypothesis testing correction indicates that the overlap is more likely due to chance than a true association to AD in this case.

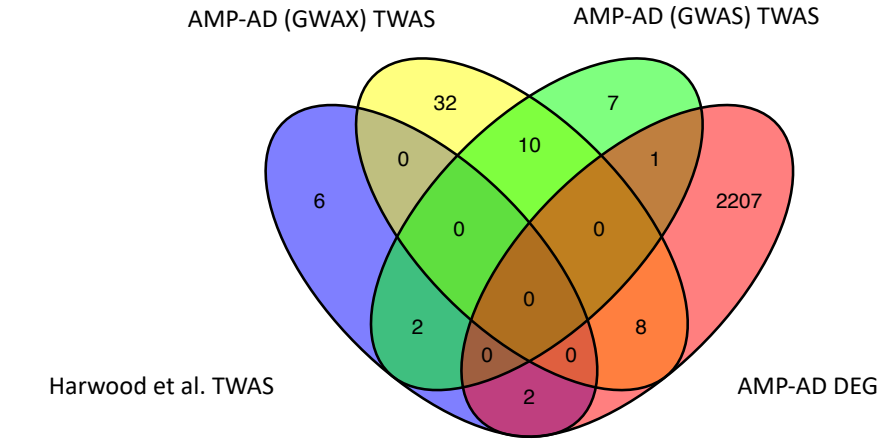
The DGE analysis had the largest overlap of differentially expressed genes with the AMP-AD TWAS which used the GWAX (Jansen et al. 2019) summary statistics. There was an overlap of two genes at FDR p-value < 0.05 (Figure 6-3b) and 90 genes with uncorrected p-value < 0.05 (Figure 6-3a). Between the three TWAS, the largest overlap of genes was seen between the two AMP-AD TWAS with nine genes shared FDR < 0.05 (Figure 6-3b) and 110 shared at the nominal significance level (Figure 6-3a). A summary of overlaps between all four studies can be seen in Figure 6-3 a-c.



a) p-value < 0.05



b) FDR p-value < 0.05



c) FDR p-value < 0.1

Figure 6-3 – Venn diagrams showing overlap of significant genes identified by the AMP-AD (GWAS) and AMP-AD (GWAX) TWAS (Gockley et al. 2021), the Harwood et al. TWAS and my AMP-AD DGE analysis (DEG). A) significance of genes set at a nominal p-value of less than 0.05. b) significance set at an FDR corrected p-value of less than 0.05 and c) a significance level set at an FDR corrected p-value of less than 0.1.

Gene Symbol	Harwood et al. TWAS Z-score	Harwood et al. TWAS p-value	Harwood et al. TWAS FDR p-value	AMP-AD GWAS TWAS Z-score	AMP-AD GWAS TWAS p-value	AMP-AD GWAS TWAS FDR p-value	AMP-AD GWAX TWAS Z-score	AMP-AD GWAX TWAS p-value	AMP-AD GWAX TWAS FDR p-value	DGE Z-score	DGE p-value	DGE FDR p-value
NPEPPS	3.56	$3.73x10^{-04}$	0.11	2.93	$3.38x10^{-03}$	0.34	2.19	0.03	0.52	3.39	$7.05x10^{-04}$	0.02
QPCT	-2.65	$8.04x10^{-03}$	0.47	-2.64	$8.26x10^{-03}$	0.42	-2.08	0.04	0.57	2.01	0.04	0.20
PRKD3	-2.60	$9.20x10^{-03}$	0.47	-2.80	$5.10x10^{-03}$	0.37	-3.41	$6.50x10^{-04}$	0.09	2.93	$3.36x10^{-03}$	0.05
DECR2	-2.17	0.03	0.59	-2.52	0.01	0.46	-2.12	0.03	0.54	-3.26	$1.10x10^{-03}$	0.03

Table 6-2 – Genes that were nominally significant in Harwood et al. TWAS, AMP-AD (Kunkle et al.) GWAS TWAS, AMP-AD (Jansen et al.) GWAX TWAS and differentially gene expression (DGE) analysis.

With the overlaps found, hypergeometric tests were performed to determine if the overlap of genes were more than would be expected by chance. None of the overlaps between differentially expressed genes and the three TWAS were statistically significant (Table 6-3).

Comparison	**Hypergeometric (p-value < 0.05)**	**Hypergeometric (FDR < 0.1)**
DGE & Harwood TWAS	0.86	0.19
DEG & AMP-AD (GWAS) TWAS	0.66	0.95
DEG & AMP-AD (GWAX) TWAS	0.72	0.34
AMP-AD (GWAS) TWAS & Harwood TWAS	4.24×10^{-39}	2.54×10^{-05}
AMP-AD (GWAX) TWAS & Harwood et al. TWAS	2.98×10^{-06}	No overlap at this threshold
AMP-AD TWAS vs AMP-AD TWAS	2.81×10^{-26}	3.62×10^{-17}

Table 6-3 Results from hypergeometric tests when comparing differential gene expression (DGE) data and TWAS p-values.

A Spearman rank corelation analysis was also performed to measure the relationship between the rank-ordered genes. The results from TWAS were significantly correlated to one another but not when TWAS was compared with DEGs. Results are summarised in Table 6-4.

Comparison	Spearman rho	Spearman p-value
DGE & Harwood TWAS	0.013	0.48
DGE & AMP-AD (GWAS) TWAS	0.017	0.17
DGE & AMP-AD (GWAX) TWAS	-0.001	0.94
AMP-AD (GWAS) TWAS & Harwood TWAS	0.715	$2.2x10^{-16}$
AMP-AD (GWAX) TWAS & Harwood et al. TWAS	0.342	$2.2x10^{-16}$
AMP-AD TWAS vs AMP-AD TWAS	0.393	$2.2x10^{-16}$

Table 6-4 Results of pairwise Spearman rank correlation analysis.

6.4 Discussion

TWAS and DGE analysis are both supposed to identify differentially expressed genes associated with AD. TWAS identifies cis-regulated expression, whereas DGE analysis identifies differentially expressed genes associated with AD due to disease related (potentially downstream) issues, influences from the environment, genetic factors, stochastic events and technical artefacts. The evidence presented in this chapter suggests there is little overlap between the two approaches.

The first aim of this chapter was to compare which genes were significant across all three TWAS and the DGE analysis to see if any results replicated across methods. Only four genes replicated at a nominal p-value with none overlapping after multiple hypothesis testing correction. These were: *NPEPPS*, *QPCT*, *PRKD3*, and *DECR2*. Two of these have additional functional evidence for involvement in AD. It has been shown that elevation of Npepps activity blocks accumulation of hyperphosphorylated TAU protein in mice and may be a therapeutic target in AD (Kudo et al. 2011) . QPCT is known to play a role in human amyloid-beta formation and its inhibition is considered a potential treatment mechanism in AD (Vijayan and Zhang 2019). Although it is possible to functionally speculate on these results, the value of it is limited as they did have opposing Z-scores. Studies in other disorders have often presented their work in a similar way i.e. comparing overlap between TWAS and DGE analysis and speculating on functional involvement on the overlap. Generally the overlap is small and further statistical tests are not performed. This analysis shows that care does need to be taken when speculating on overlapping results from TWAS compared to DGE analysis as the overall evidence for consistent involvement is weak and that this small overlap is possibly due to chance.

The second aim of this chapter was to determine if there was a correlation of Z-scores between the TWAS methods and DGE analysis. Z-scores were significantly correlated between the three TWAS but not between any TWAS and DGE analysis. The two TWAS with the strongest correlation ($r = 0.72$, p-value = 2.2×10^{-16}) was between that of the AMP-AD (GWAS)

TWAS and Harwood et al. TWAS. These two studies used the same GWAS summary statistics (Kunkle et al.) but different eQTL panels.

The two AMP-AD TWAS which used the same reference panel of eQTLs but different summary statistics (GWAS vs GWAX) were moderately corelated (r = 0.45, p-value = $2.2x10^{-16}$). The weaker correlation could be the result of the Jansen et al.GWAX having more statistical power than the Kunkle GWAS. The GWAX included 534,403 individuals versus 94,437 individuals for the GWAS (Andrews et al. 2020). Alternatively, it could be due to phenotypic and/or genetic heterogeneity between the two GWAS methods which then affect the results of the TWAS. In the GWAX they used AD cases and controls as well as a self-reported proxy phenotype of whether their parents were affected by AD. The self-reported proxy is at risk of individuals misremembering or not knowing the difference between dementia and an AD diagnosis. It has previously been hypothesised that phenotypic heterogeneity due to misdiagnosed AD could lead to genetic heterogeneity and reduced statistical power for GWAS discovery (Andrews et al. 2020). It is not clear if by including large numbers of people self-reporting their parental history if it leads to an increase in statistical power or just an increase in noise. Despite slight differences in correlation, the three TWAS performed similar to one another as would be expected.

The final aim of this chapter was to perform hypergeometric tests and Spearman rank correlation tests to determine if any observed overlap of candidate genes is more than expected through chance alone. On a pairwise comparison basis, any gene overlap between any TWAS and DGE analysis had occurred not more than expected by chance (the enrichment was not statistically significant). Many studies in other disorders have compared the results of their TWAS to lists of differentially expressed genes in order to provide additional support to their TWAS findings. However, the findings are often a very small list of genes and rarely are further statistical tests performed to determine if these overlaps were likely through chance alone. The findings in this chapter highlight the prudency of performing additional tests to diminish chance findings.

When looking at the whole distribution of Z-scores, enrichment of TWAS signals in differentially expressed genes was not found. When focusing on only significant results (FDR > 0.1 or p-value > 0.05), no enrichment of TWAS signals in differentially expressed genes was found either. This is the case even when there is a large sample overlap. This leads to the conclusion that in AD, differentially expressed genes are not enriched for TWAS signals, and the two approaches are not comparable.

TWAS is often a misunderstood method for prioritising candidate genes and has been mistakenly used as a causal gene test (Wainberg et al. 2019). TWAS is useful for isolating the cis- component of gene expression. DGE analysis on the other hand, is used for finding differences between AD cases and controls. It will be subject to a lot more noise in comparison to TWAS, as DGE analysis will be identifying genes whose expression is also sensitive to the environment, technical artefacts, as well as cis- and trans regulation. The results of DGE analyses could well be due to consequences of disease as opposed to causes, so may not be as informative to the underlying biology.

A recent study has identified that many gene expression profiles are generic and are highly predictable with many of the same genes being differentially expressed across a wide variety of phenotypes. These highly predictable differentially expressed genes (referred to as DE prior) implicate the immune response, inflammation, the extracellular matrix, stress responses and sex. They are likely to be biologically relevant to disease processes in general but are not necessarily specific to individual phenotypes (Crow et al. 2019). Future work could make use of these predictable genes, to account for some of the redundancy in a DGE analysis.

One limitation of the analysis in this chapter is that the TWAS used for comparison in the analysis all used the summary-based method FUSION. It is possible that other methodologies may produce results with more consistency to the DGE analysis. Another limitation is that the DGE analysis and TWAS were all in bulk tissue. Although the brain cortex is disease relevant for AD, it is still a very broad tissue type. Performing these methods in specific brain regions,

or cell types, may find different results, and the TWAS and DGE analysis may show a greater overlap.

In conclusion, the results in this chapter suggest that there is little overlap in the differentially expressed genes from DGE analysis and TWAS in AD in brain cortex. The little overlap that was found is no more than would be expected through chance. Future work could explore the differences in results when using gene expression data from other samples such as monocytes or specific brain regions. Future work could also explore the inclusion of the DE prior in the DGE analysis and then compare results to TWAS to see if this affects the overlap between the two methods.

Chapter 7 General discussion

7.1 Thesis overview

Alzheimer's disease (AD) is a devastating neurodegenerative disorder that currently has no cure. Studies of genetics have shed light on some of the potential mechanisms involved in AD such as the role of amyloid-beta, immunity and inflammation. However, there is a need to move beyond pure genomics to try and shed light on the relationship between genetic variants and function in AD.

The work presented in this thesis aims to build on our understanding of AD risk that findings from genetics have provided. The result is the application of a range of bioinformatic approaches to genetic and transcriptomic data to elucidate underlying mechanisms and biology of AD. This thesis also looked to explore if differential gene expression has utility in prioritising AD GWAS and TWAS findings.

7.2 Summary of findings

Chapter one gave an overview of AD including the pathology, progression and epidemiology of the disease. This was followed by a discussion of the risk factors associated with disease including genetics and how various bioinformatic analyses have been applied to try and further our understanding of the disease. This included discussion of methods such as GWAS, eQTL analysis, TWAS and DGE analysis.

Chapter two describes the data and methods used throughout this thesis. The chapter began with a description of how the three cohorts of ROSMAP, MayoRNASeq and MSBB were originally generated and then reprocessed and hosted by AMP-AD. The remainder of the chapter describes the various bioinformatic analyses used throughout this thesis.

Chapter three describes the work to produce a single cohesive RNA-seq dataset from combining the three AMP-AD RNA-seq datasets. An extensive QC pipeline was used to produce a high-quality dataset that could be used for downstream analyses. Initial investigation of the datasets highlighted that there were batch effects and sources of unwanted variation present in the data which could increase the risk of spurious findings if left uncorrected. These included sequencing batch and originating study. Linear mixed-effect models (LMEM) in combination with principal component analysis (PCA) were used to combine the three RNA-seq datasets into a unified dataset. Investigation of the combined datasets did not identify any obvious batch effects. This is the first time this approach has been used to combine these three datasets together into a single dataset which was then used for multiple downstream analyses.

Another aim of this chapter was to define the phenotypic variables reflecting AD pathology. The work to define case-control status highlighted the challenge of defining phenotypes in the AD field. The three datasets provided differing amounts of information on the phenotypes available. The MayoRNASeq study provided information on case-control status and Braak scores. The ROSMAP study provided data on diagnosis based on clinical data, Braak scores and non-age modified CERAD scores. The MSBB study provided data on clinical dementia rating, CERAD and Braak scores. This is a problem in the AD and dementia field at large, as both can be subject to a considerable amount of phenotypic heterogeneity. Failing to capture this adequately could lead to misclassification bias or a reduction in power (Ryan et al. 2018). The phenotypic definitions included in this analysis do have the added advantage of including measures of pathology and clinical characteristics meaning that it is more likely to capture true AD than clinical information alone.

Chapter four built on the work in the previous chapter by using the residuals from the combined normalised AMP-AD RNA-seq data to perform a DGE analysis. Initially this was performed in only the ROSMAP data between cases and controls. This was in order to see how the LMEM and PCA method for combining data followed by logistic regression to determine differentially expressed genes performs against two frequently used DGE packages: limma-voom and DESeq2. A comparative analysis showed that the use of logistic

regression following the use of LMEM and PCA to correct for batch effects identified differentially expressed genes with a significant overlap to two other tools. Previous work has shown that differentially expressed genes which are identified through multiple tools are more likely to be truly differentially expressed (Costa-Silva et al. 2017).

A DGE analysis was also performed on the combined AMP-AD RNA-seq data which was generated in chapter three for AD case-control, Braak and CERAD score phenotypes. Overall, the results for the DGE analysis and GO enrichment analysis were similar for the Braak and CERAD score phenotypes. This could be due to the two being highly correlated and/or sharing underlying mechanisms. The biological processes involved included GO terms associated with mitochondrial processes, the ribosome and endoplasmic reticulum. Results from this analysis suggest the involvement of these processes in AD as described by the Braak and CERAD pathology. Although these processes have been previously shown to be involved in AD, they are not that well characterised in the disease (Weidling and Swerdlow 2020; Iatrou et al. 2021; Shi et al. 2022). Studying these further could enhance overall understanding of biological mechanisms for AD development and progression.

An investigation of whether GWAS prioritised genes are enriched in the DGE analysis was also performed, and it was found that they were not for statistically significant differentially expressed genes. The DGE analysis was derived from bulk brain tissue. Bulk brain tissue has the caveat that it is dominated by the most abundant cell type and does not capture information about cell type or composition (Trapnell 2015; Cano-Gamez and Trynka 2020). More specific cell-types, with other temporal resolutions may have more utility for GWAS gene prioritisation in AD based on the gene expression

Lastly in chapter four, MAGMA pathway analysis results from the largest AD GWAS (Kunkle et al. 2019) were checked to see if they overlapped with my gene ontology enrichment analysis results from my analysis. No overlap of FDR-significant terms was found between the GWAS and my non-directional and downregulated analysis GO terms (as identified through using CATMAP). In the upregulated analysis five of the nine GO terms overlapped at nominal

significance (Protein-lipid complex assembly, regulation of amyloid-beta formation, regulation of amyloid precursor protein catabolic process, tau protein binding and activation of immune response) of which the latter two still remained significant after FDR correction. Although the GWAS prioritised genes were not enriched as a set in the differentially expressed genes, there is limited evidence for some convergence of pathways identified through both genetic and gene expression data.

In chapter five, a cis-eQTL analysis was performed to find associations between index SNPs from five AD GWAS and GWAX, and the AD case-control differentially expressed genes identified in chapter four. As index SNPs are not necessarily the causal variant, the 100kb region either side of the index SNP was investigated, as that is where over 90% of causal variants are located (Wu et al. 2017). The initial cis-eQTL analysis identified seven SNP-gene associations. As three of them were located on chromosome 19 there were concerns that this might be due to the long-range effects of *APOE*. Rerunning the analysis with *APOE* carrier status as a covariate resulted in a slight increase in the significance of results with an additional SNP-gene pair becoming statistically significant. Therefore, there was evidence that the results were not due to *APOE*. The next step in the analysis was to search the 100kb region around each index SNP. For the eight regions searched, none of the index SNPs were the top eQTL in the 100kb surrounding region.

A trans-eQTL analysis was also performed to identify associations between AD GWAS/GWAX index SNPs and the differentially expressed genes from chapter four. Only four trans-eQTLs were identified, which were *MAF1*, *TAC1, SCGN* and *SST*. These were all associated with the C allele of rs5011436 in the *TMEM106B* locus. The evidence presented suggests a potential mechanism for the association on chromosome 7. The AD associated SNP could be trans-regulating these AD case-control differentially expressed genes. A STRING analysis identified that at least *SST*, *SCGN* and *TAC1* were related through text-mining (p-value = 1.25×10^{-04}). Previous genetic work has also implicated the *SST* locus in AD risk in Finnish and Chinese cohort (Vepsäläinen et al. 2007; Xue et al. 2009). Although firm conclusions cannot be drawn, this may offer an avenue for future functional work.

Chapter six compares the results of the DGE analysis (generated in chapter 4) to three existing AD TWAS results to identify if these two methods produce comparable results. This has not yet been investigated in the field of AD research. The results of this analysis identified that AD TWAS produce results with a significant overlap to one another, but not to DGE analysis. Although there were some overlapping genes between TWAS and DGE analysis methods, it was no more than what was expected by chance and differentially expressed genes are not enriched for TWAS signals. A recent study has suggested that differentially expressed genes are more likely to be the result of disease processes rather than the cause (Porcu et al. 2021). It could also be that current transcriptomic datasets are not using the most informative tissue type and that transcriptomic data from microglia or prodromal AD may be more informative or are more enriched for TWAS signals.

7.3 Limitations of thesis

One challenge with researching Alzheimer's disease is that one of the most disease relevant tissues to utilise is the brain. Brain tissue is hard to come by, and sample sizes remain relatively small. The work presented in this thesis tried to overcome this by combining samples from different areas of the brain into a single dataset. This has the benefit of increasing the sample size but at the cost of region specificity.

A further limitation is that bulk cortical brain tissue may not be the most disease relevant tissue. Although the cortex is affected in AD, it may be that these global changes occur later in disease progression. Therefore, using bulk brain cortex tissue may be identifying changes due to disease rather than cause of disease or even capturing end-stage of disease which would not benefit as much from therapeutic intervention in comparison to earlier stages of disease.

Bulk brain tissue is dominated by the most abundant cell type and does not capture information about cell type or composition (Trapnell 2015). There is evidence that microglia

plays a significant role in AD pathogenesis (Hemonnot et al. 2019) so may be a more relevant cell-type for the study of AD.

Building on this, another limitation is that the transcriptomic data comes from post-mortem brain tissue. Gene expression can differ between life and death. Post-mortem effects are understudied, and little is known as to how expression is affected in the brain upon death (Ferreira et al. 2018). The work presented in this thesis tried to overcome this limitation by investigating post-mortem interval (PMI) and found no evidence of PMI batch effects when investigating PCA biplots.

Another limitation of the work presented in this thesis is that only individuals of European descent were included in the analysis. The decision for this was to avoid population stratification bias. Additionally, the data available from non-Europeans in the AMP-AD datasets were very small. Therefore, the findings from this thesis may only be applicable to those with a European ancestry. Another limitation is that focusing on a homogenous population could mean that biological processes that occur in AD in other populations will be missed due to a lack of genetic diversity being included in the analysis. Studying other populations are likely to inform more about disease processes (Carress et al. 2021).

The work presented in this thesis made use of publicly available RNA-seq, genetic and phenotypic data. Many of the limitations of the work in this thesis reflect problems in the wider AD field. Larger datasets with deeper phenotyping are required to further understanding of AD.

Another limitation is that shortly after the analysis for this thesis was complete, another GWAX was published identifying an additional 42 novel loci at the time of publication (Bellenguez et al. 2022). This new GWAX prioritised 55 genes in the 42 new loci. Of these 55, *ATP8B3* was the only differentially expressed gene in my case-control analysis (chapter 4) with a p-value of 2.38×10^{-03} and an FDR corrected p-value of 0.04. Therefore, the same conclusion

remains that there is limited enrichment of GWAS prioritised genes in AD case-control differentially expressed genes. The Bellenguez et al. GWAX identified novel index SNPs which would not have been included in my eQTL analysis which is a further limitation.

One final limitation is that the work in this thesis used a bioinformatics approach to study Alzheimer's disease. Although many potential mechanisms and genes have been identified in the work presented in this thesis, conclusions of causality cannot be drawn from these results. Instead, it offers avenues for future directions for functional follow-up in laboratory studies.

7.4 Future work and directions

The future work of bioinformatic analyses to improve understanding of the biological underpinnings of AD will rely on increasing sample sizes especially in disease relevant cell-types. GWAS have been a major contributor to identifying variants associated with disease. Increasing sample sizes, expansion into using whole genome sequencing data and deeper phenotyping will lead to new discoveries in the genetics of AD. Larger GWAS will lead to the discovery of new variants which will typically have smaller effect sizes potentially increasing resolution. Identifying causal variants from these findings will continue to be a major challenge.

Variants identified from GWAS are used in polygenic risk scores, which is an approach that could be integrated into future work. Although polygenic risk scores alone may not have the prediction accuracy for personalised medicine they may be beneficial in identifying those at the PRS extremes (individuals more than two standard deviations from the mean) (Baker and Escott-Price 2020). Recent work has shown that standardising PRS against the population mean as opposed to the sample mean makes the individuals' scores comparable between studies (Leonenko et al. 2021). As phenotypic information is often variable between studies, polygenic risk scores could be used to identify individuals with high polygenic risk versus low polygenic risk burden as a phenotype alternative to case-control and would be comparable across studies.

AD GWAS have already implicated microglial and immune genes as important factors in the development and progress of disease (Efthymiou and Goate 2017). The transcriptomic data used in this thesis were derived from bulk brain tissue which is known to biased by the most abundant cell types (Trapnell 2015). The proportion of microglia in the cortex is approximated to be 5% so likely to be underrepresented in bulk analyses (Ochocka and Kaminska 2021). Studying microglia and other disease relevant tissues at high resolution is needed to increase insights into disease. At present the ROSMAP study has a sample size of 10 for RNA-seq data of prefrontal cortex microglia and 13 of single-cell RNA-seq of prefrontal cortex microglia (www.radc.rush.edu/docs/omics.htm). With time, it is likely that these resources will grow and allow researchers to further understand the role microglia has in AD and if microglia are viable therapeutic targets.

Increasing sample sizes of tissue expression data will also be of benefit to eQTL studies and DGE analyses. At present, studies mainly focus on cis-eQTLs due to limited sample sizes and statistical power. However cis-eQTL effects are thought to contribute only a small fraction of the heritability of gene expression. The majority of this heritability is mainly thought to come from the combination of many weak trans-eQTL effects and thought to have greater impact on the phenotype than those regulated by strong eQTL effects (Võsa et al. 2021). Furthermore tissue-specific eQTLs with larger sample sizes may identify novel genes associated with AD.

The findings in this thesis, such as genes identified through DGE analysis or in the eQTL analysis could also be taken up for functional follow-up in model systems such as *drosophila* to help further understanding of AD biology.

The work produced in this thesis made extensive use of differentially expressed genes. DGE analysis often produces a large list of disease associated genes that will be the result of genetics, the environment, and stochastic and technical factors. Recent work has suggested that many differentially expressed genes are highly predictable and are not necessarily specific to individual phenotypes but to general disease processes such as inflammation. This

can be used to inform a DE prior which can be used to account for this (Crow et al. 2019). Future work could use this DE prior to try and elucidate specific AD differentially expressed genes and pathways.

Future work could also integrate the current data with the other -omics data which is available on the Synapse platform in a multi-omics approach. This could include DNA methylation data to try and predict biological age which could be integrated into analysis. Future work could also look at microRNA data. Increasing evidence is building for the involvement on microRNAs. This includes not only dysregulation of microRNA in AD but also as a potential diagnostic biomarker and so is an important area of research (Wei et al. 2020). The ROSMAP dataset does have microRNA data available for download (Zhang et al. 2013). This data could be downloaded and integrated with the genetic and mRNA data used in this thesis to further our understanding of AD disease biology.

7.5 Implications

The work presented in this thesis has identified and provided additional evidence for the involvement of *SST*, *SCGN*, *MAF1* and *TAC1* with an association on chromosome 7 in AD. Additionally the work has highlighted the potential involvement of mitochondrial and ribosomal processes, in addition to the endoplasmic reticulum. These are under characterised in AD, so may offer new avenues for functional follow-up.

Finally, the work from the last results chapter has shown that TWAS signals are not enriched in case-control differential gene expression derived from RNA sequencing of bulk brain tissue. TWAS are often a misunderstood method, and the work in chapter 6 highlights that TWAS can only inform on associations between the common cis- component of gene expression whereas differential gene expression will be sensitive to not only genetically controlled factors, but environmental, stochastic and technical factors too. Therefore, in AD, the two methods at present are not comparable.

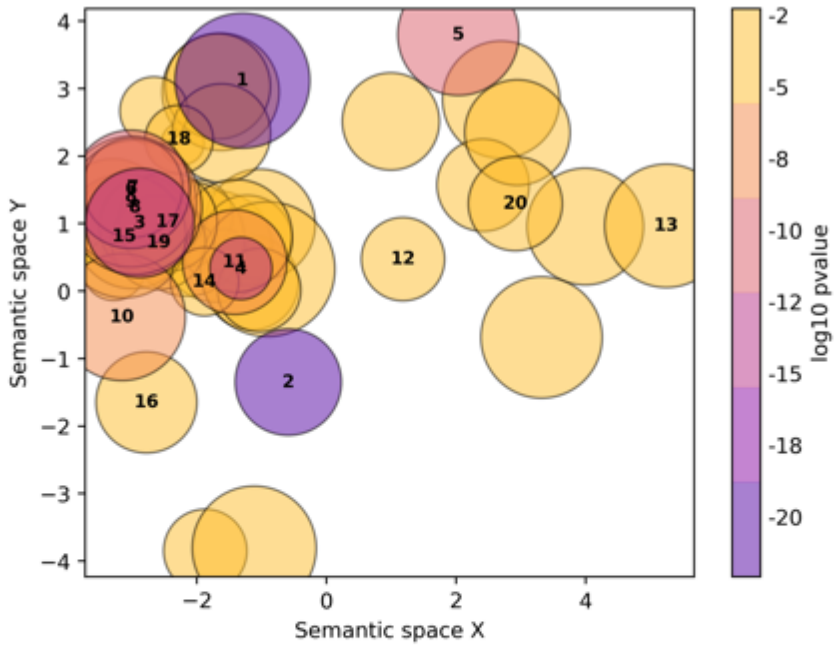
7.6 Conclusions

The work presented in this thesis largely were from the investigation of RNA-seq data from three different cohorts combined into a single dataset. The findings of this thesis would suggest that case-control differential gene expression data from bulk cortical tissue is not informative for follow-up of prioritised genes from GWAS nor TWAS. It could be that this is due to gene expression being the result of disease rather than the cause of disease. However, pathway analysis of transcriptomic data has identified involvement of mitochondrial and endoplasmic reticulum processes which are biologically plausible. Additionally, eQTL analysis of differentially expressed genes has highlighted some candidate genes with a good amount of evidence that are also biologically plausible, such as for *SST*.

The onset of AD is hypothesised to start in middle-age. As transcriptomic datasets grow and become more diverse both temporally and with originating tissue, more will be discovered about the underlying processes of AD. Integration with genetics and other -omics data will help improve understanding of AD and hopefully lead to new treatments.

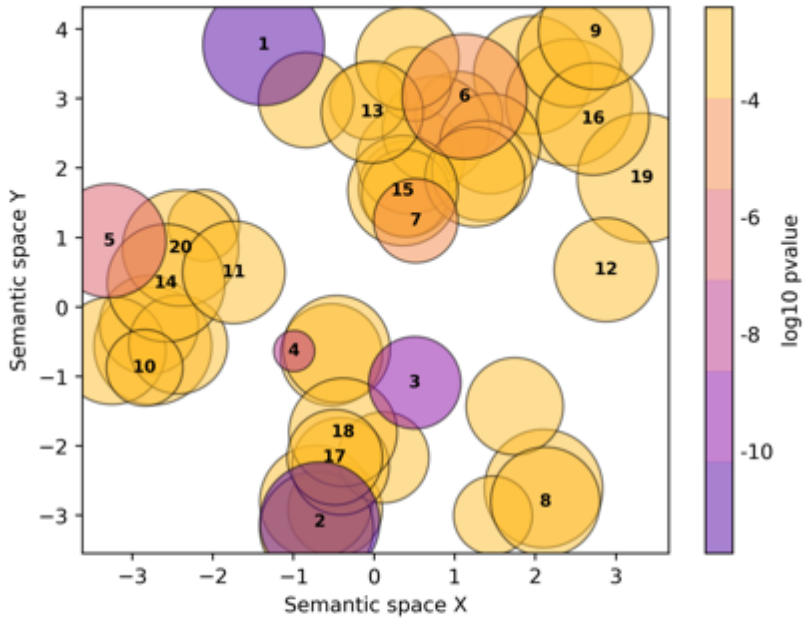
Appendix 1

Go-Figure! results for reduced Braak score logistic regression analysis.



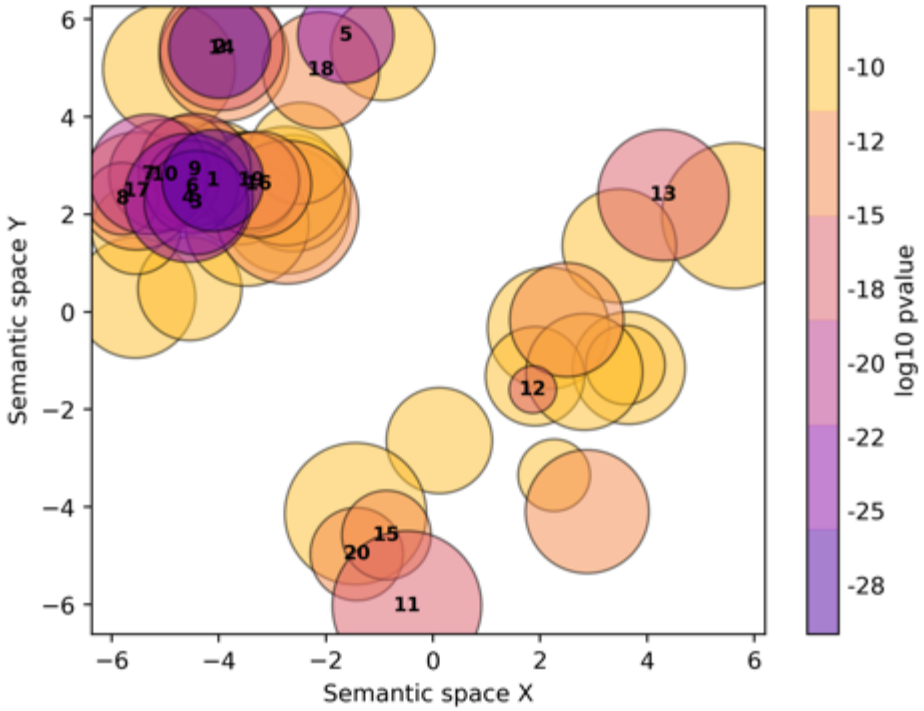
1. SRP-dependent cotranslational protein targeting to membrane
2. viral transcription
3. amide biosynthetic process
4. translational initiation
5. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
6. heterocycle catabolic process
7. organic cyclic compound catabolic process
8. ncRNA metabolic process
9. monocarboxylic acid catabolic process
10. mitochondrial respiratory chain complex assembly
11. respiratory electron transport chain
12. hormone-mediated apoptotic signaling pathway
13. positive regulation of glucocorticoid secretion
14. drug catabolic process
15. fatty acid metabolic process
16. insemination
17. translational elongation
18. transcription-coupled nucleotide-excision repair
19. GDP-L-fucose metabolic process
20. negative regulation of negative chemotaxis

Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



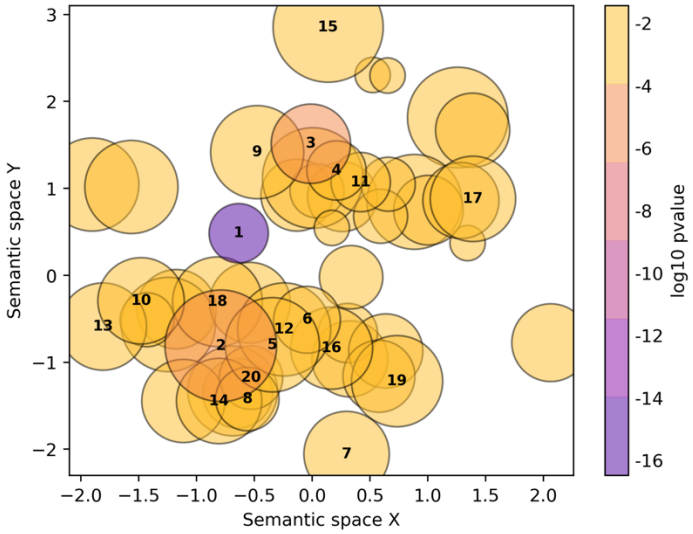
1. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
2. SRP-dependent cotranslational protein targeting to membrane
3. viral transcription
4. translational initiation
5. intracellular protein transport
6. modulation of chemical synaptic transmission
7. hormone-mediated apoptotic signaling pathway
8. mesendoderm development
9. positive regulation of glucocorticoid secretion
10. peptide biosynthetic process
11. potassium ion transmembrane transport
12. neurotransmitter secretion
13. regulation of receptor localization to synapse
14. neurofilament bundle assembly
15. nerve growth factor signaling pathway
16. positive regulation of neuron projection development
17. sodium ion transmembrane transport
18. calcium-ion regulated exocytosis
19. regulation of neurogenesis
20. response to DNA damage checkpoint signaling

Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



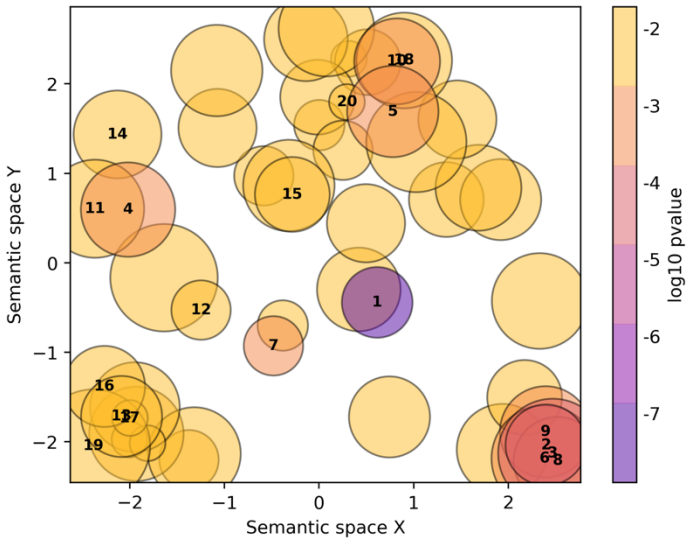
1. signal transduction
2. positive regulation of biological process
3. regulation of signaling
4. regulation of cell communication
5. regulation of localization
6. regulation of biological quality
7. regulation of multicellular organismal process
8. negative regulation of cellular process
9. regulation of developmental process
10. regulation of response to stimulus
11. developmental process
12. cellular process
13. cellular response to chemical stimulus
14. positive regulation of RNA metabolic process
15. cellular response to stimulus
16. regulation of transcription by RNA polymerase II
17. negative regulation of nitrogen compound metabolic process
18. regulation of cellular component organization
19. regulation of molecular function
20. response to chemical

Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



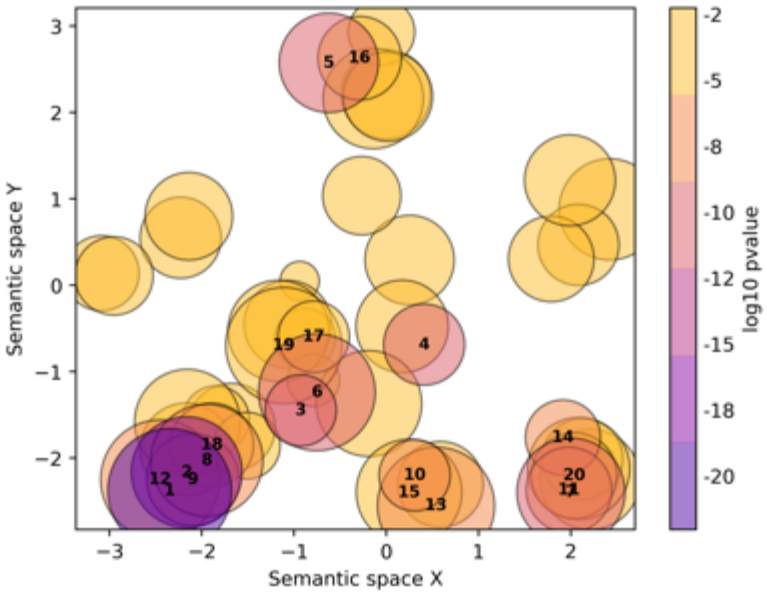
1. structural constituent of ribosome
2. catalytic activity, acting on a tRNA
3. flavin adenine dinucleotide binding
4. large ribosomal subunit rRNA binding
5. NADH dehydrogenase (ubiquinone) activity
6. trans-2-enoyl-CoA reductase (NADPH) activity
7. C-acetyltransferase activity
8. aminomethyltransferase activity
9. MutSalpha complex binding
10. 3'-flap endonuclease activity
11. iron-sulfur cluster binding
12. oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxyge...
13. DNA-dependent ATPase activity
14. tRNA (cytosine) methyltransferase activity
15. neuropeptide hormone activity
16. aldehyde dehydrogenase [NAD(P)+] activity
17. translation elongation factor activity
18. DNA ligase (ATP) activity
19. 3-hydroxyacyl-CoA dehydrogenase activity
20. catechol O-methyltransferase activity

Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



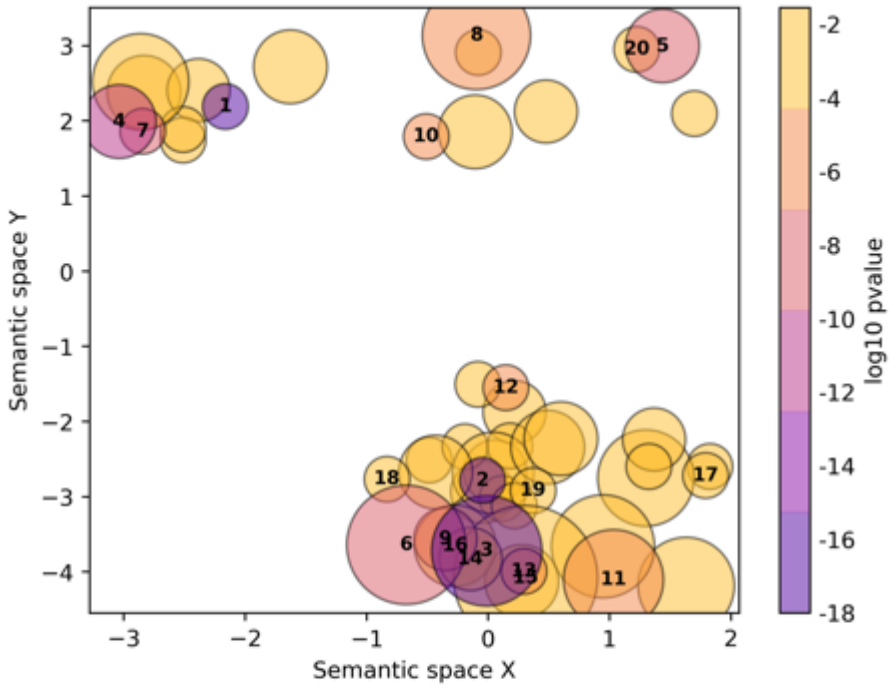
1. structural constituent of ribosome
2. potassium ion transmembrane transporter activity
3. voltage-gated potassium channel activity
4. cAMP-dependent protein kinase regulator activity
5. Rab geranylgeranyltransferase activity
6. monovalent inorganic cation transmembrane transporter activity
7. cAMP binding
8. cation channel activity
9. sodium channel activity
10. protein C-terminal carboxyl O-methyltransferase activity
11. antiporter activity
12. large ribosomal subunit rRNA binding
13. myosin light chain binding
14. sphingosine-1-phosphate receptor activity
15. trans-2-enoyl-CoA reductase (NADPH) activity
16. MRF binding
17. S100 protein binding
18. alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase activity
19. ephrin receptor binding
20. aminomethyltransferase activity

Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



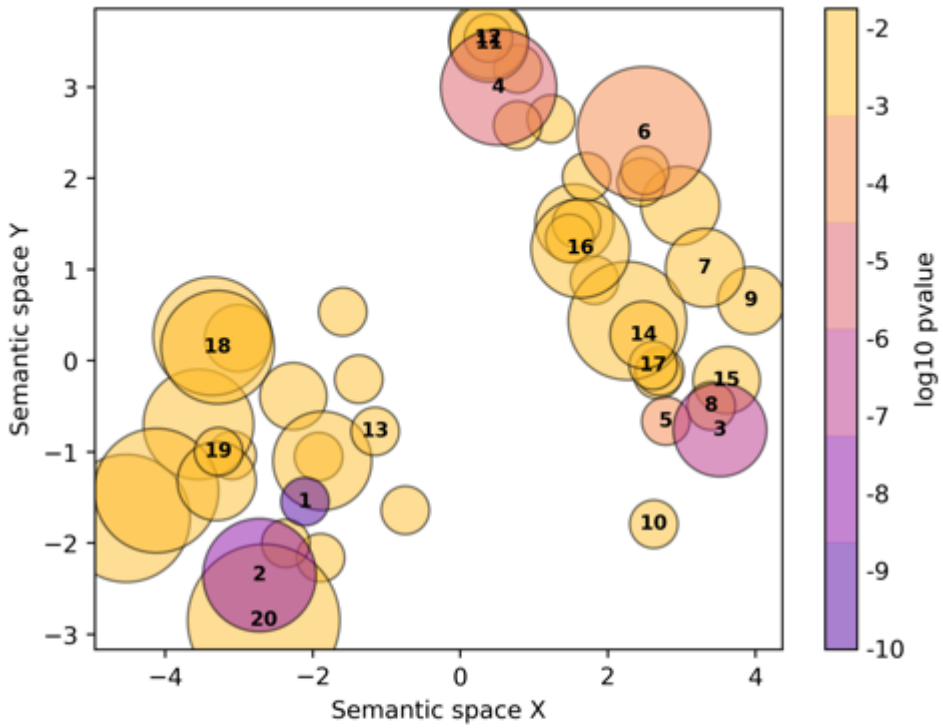
1. protein binding
2. enzyme binding
3. binding
4. DNA-binding transcription factor activity, RNA polymerase II-specific
5. kinase activity
6. RNA polymerase II transcription regulatory region sequence-specific DNA binding
7. voltage-gated cation channel activity
8. protein domain specific binding
9. transcription factor binding
10. transcription coregulator activity
11. potassium ion transmembrane transporter activity
12. signaling receptor binding
13. molecular transducer activity
14. ion gated channel activity
15. enzyme activator activity
16. protein tyrosine kinase activity
17. chromatin binding
18. cell adhesion molecule binding
19. protein-containing complex binding
20. transporter activity

Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



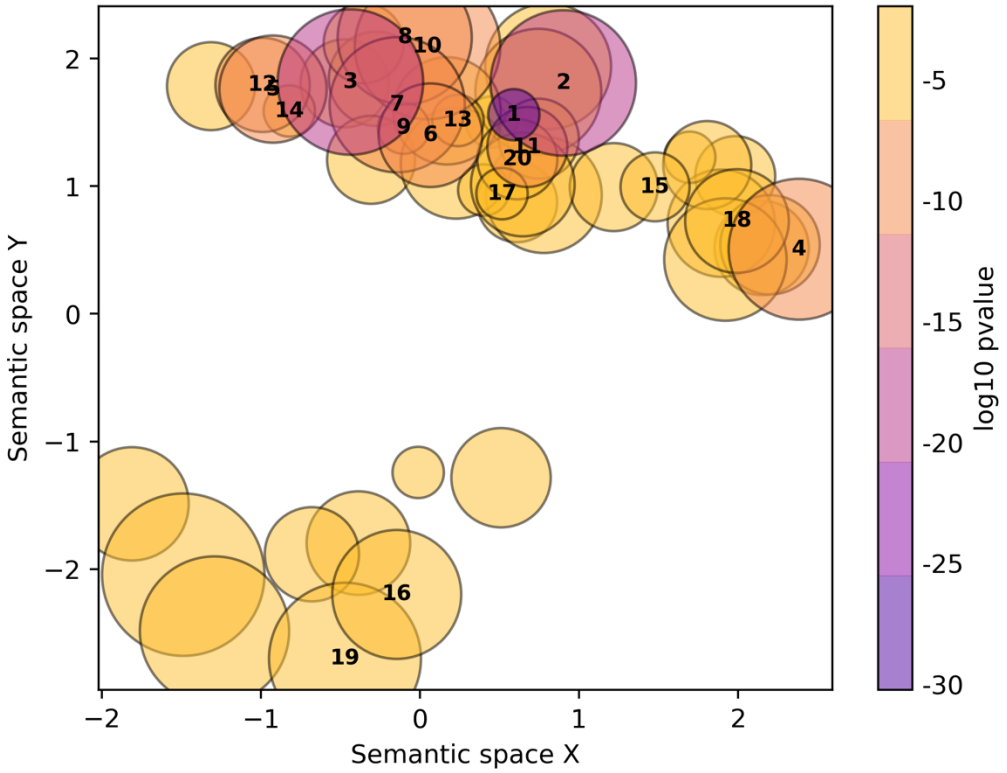
1. intracellular anatomical structure
2. cytosolic large ribosomal subunit
3. ribosomal subunit
4. ribosome
5. mitochondrial inner membrane
6. inner mitochondrial membrane protein complex
7. mitochondrion
8. mitochondrial matrix
9. respiratory chain complex
10. respirasome
11. oxidoreductase complex
12. Lsm2-8 complex
13. U6 snRNP
14. mitochondrial respiratory chain complex I
15. U4/U6 x U5 tri-snRNP complex
16. proton-transporting ATP synthase complex, coupling factor F(o)
17. methylosome
18. mitochondrial proton-transporting ATP synthase complex
19. GINS complex
20. smooth endoplasmic reticulum membrane

Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



1. cytosolic large ribosomal subunit
2. ribosomal subunit
3. ribosome
4. synapse
5. intracellular anatomical structure
6. neuron projection
7. mitochondrial inner membrane
8. chromaffin granule
9. synaptic vesicle membrane
10. intrinsic component of postsynaptic density membrane
11. intrinsic component of presynaptic membrane
12. intrinsic component of synaptic membrane
13. Lsm2-8 complex
14. postsynaptic density
15. contractile fiber
16. ciliary basal body
17. neurofilament cytoskeleton
18. voltage-gated potassium channel complex
19. Rab-protein geranylgeranyltransferase complex
20. U6 snRNP

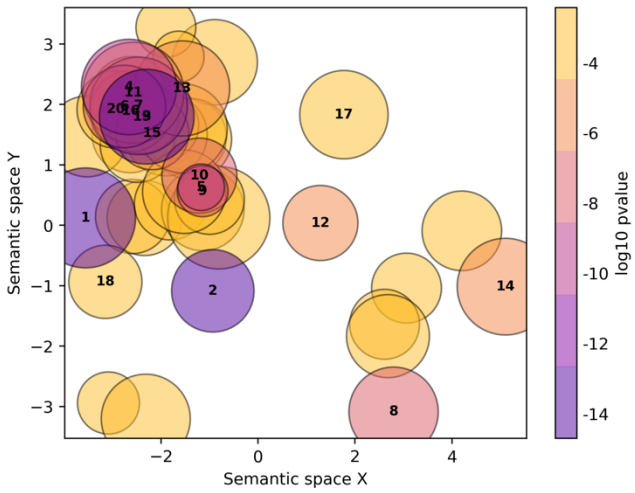
Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced Braak score logistic regression



1. plasma membrane
2. neuron projection
3. synapse
4. cytoplasmic vesicle
5. intrinsic component of plasma membrane
6. chromatin
7. plasma membrane region
8. vesicle membrane
9. presynapse
10. bounding membrane of organelle
11. cytosol
12. integral component of postsynaptic density membrane
13. cell body
14. intrinsic component of membrane
15. postsynaptic density
16. voltage-gated potassium channel complex
17. perinuclear region of cytoplasm
18. Golgi apparatus
19. plasma membrane protein complex
20. neuron projection terminus

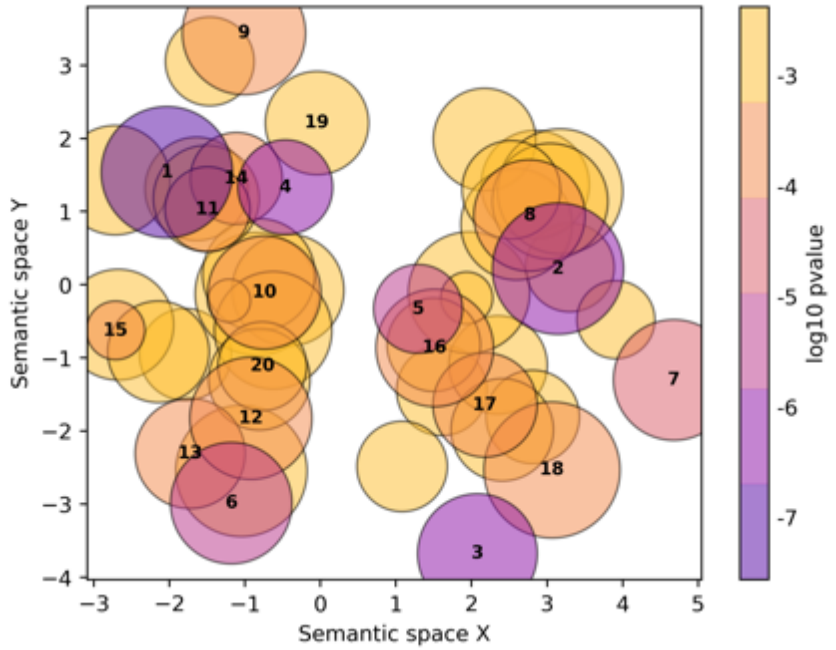
Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced Braak score logistic regression

Go-Figure! results for Braak score ordinal regression analysis.



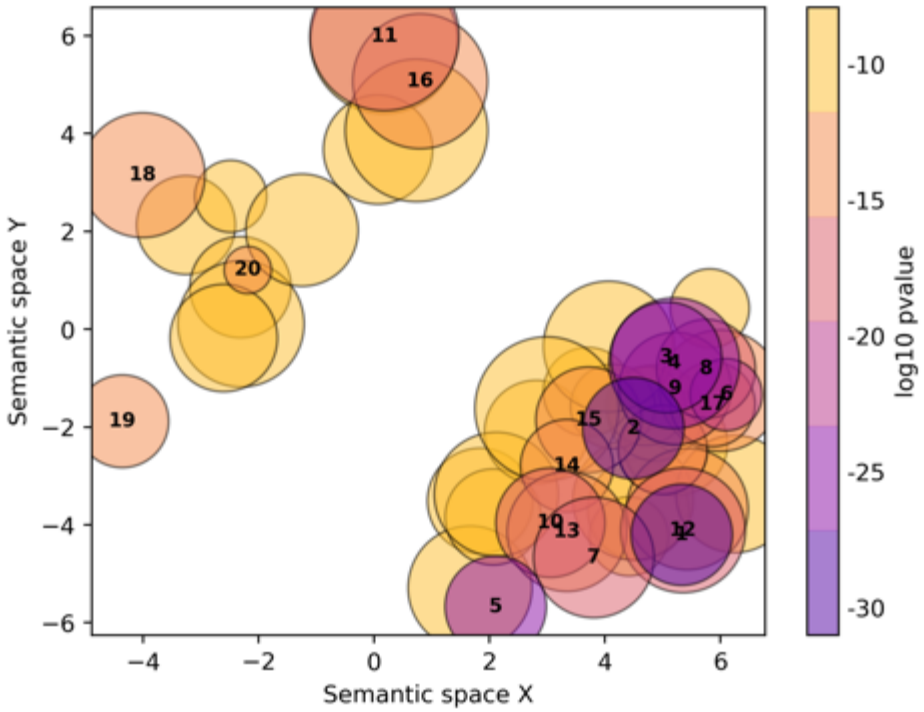
1. SRP-dependent cotranslational protein targeting to membrane
2. viral transcription
3. amide biosynthetic process
4. carboxylic acid catabolic process
5. translational initiation
6. organic cyclic compound catabolic process
7. ncRNA metabolic process
8. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
9. drug metabolic process
10. respiratory electron transport chain
11. fatty acid beta-oxidation
12. hormone-mediated apoptotic signaling pathway
13. mitochondrial respiratory chain complex I assembly
14. positive regulation of glucocorticoid secretion
15. translational elongation
16. alpha-amino acid metabolic process
17. negative regulation of mitochondrial membrane permeability involved in apoptotic process
18. insemination
19. aromatic amino acid family metabolic process
20. acetyl-CoA metabolic process

Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the Braak score ordinal regression



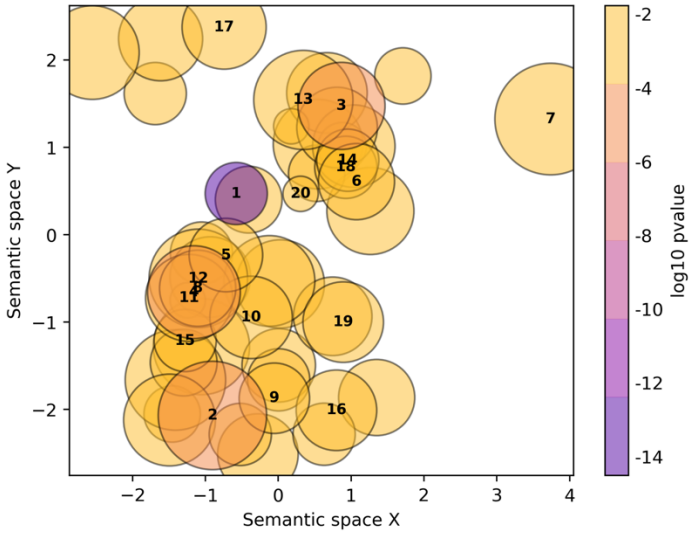
1. establishment of protein localization to membrane
2. modulation of chemical synaptic transmission
3. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
4. viral transcription
5. hormone-mediated apoptotic signaling pathway
6. intracellular protein transport
7. positive regulation of glucocorticoid secretion
8. regulation of neuron projection development
9. mesendoderm development
10. signal release from synapse
11. calcium-ion regulated exocytosis
12. response to DNA damage checkpoint signaling
13. neurofilament bundle assembly
14. protein-containing complex localization
15. peptidyl-threonine phosphorylation
16. nerve growth factor signaling pathway
17. regulation of receptor localization to synapse
18. regulation of neurogenesis
19. insemination
20. sodium ion transmembrane transport

Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score ordinal regression



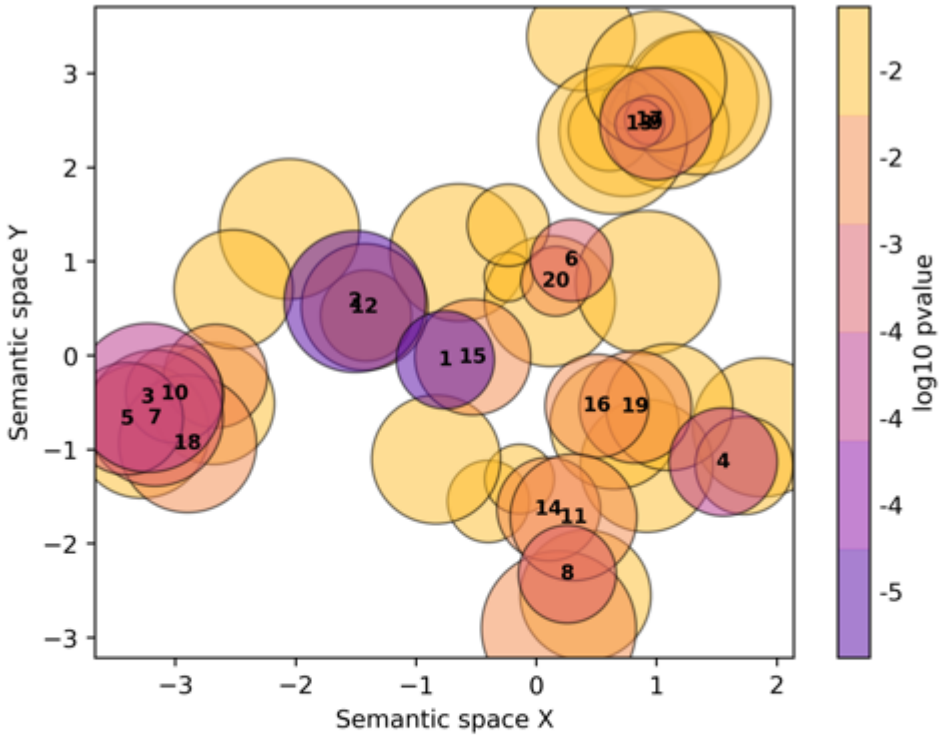
1. positive regulation of biological process
2. signal transduction
3. regulation of signaling
4. regulation of cell communication
5. regulation of localization
6. negative regulation of cellular process
7. regulation of multicellular organismal process
8. regulation of response to stimulus
9. regulation of developmental process
10. regulation of biological quality
11. developmental process
12. positive regulation of RNA metabolic process
13. regulation of cellular component organization
14. regulation of molecular function
15. regulation of transcription by RNA polymerase II
16. cellular response to organic substance
17. negative regulation of nitrogen compound metabolic process
18. cellular protein modification process
19. cellular response to stimulus
20. cellular process

Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the Braak score ordinal regression



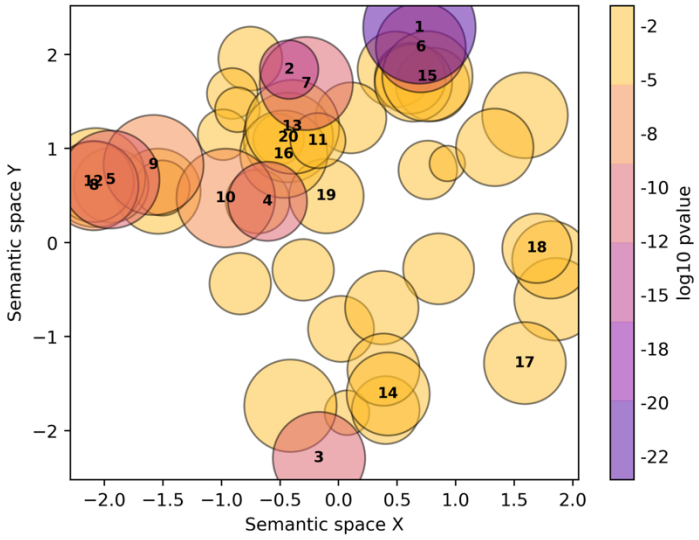
1. structural constituent of ribosome
2. catalytic activity, acting on a tRNA
3. flavin adenine dinucleotide binding
4. NADH dehydrogenase (ubiquinone) activity
5. trans-2-enoyl-CoA reductase (NADPH) activity
6. large ribosomal subunit rRNA binding
7. hormone activity
8. peroxidase activity
9. aminomethyltransferase activity
10. ligase activity
11. oxidoreductase activity, acting on the aldehyde or oxo group of donors
12. 3-hydroxyacyl-CoA dehydrogenase activity
13. MutSalpha complex binding
14. tetrapyrrole binding
15. 3'-flap endonuclease activity
16. acetyl-CoA C-acyltransferase activity
17. endoribonuclease activity
18. iron-sulfur cluster binding
19. oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxyge...
20. ubiquinone binding

Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the Braak score ordinal regression



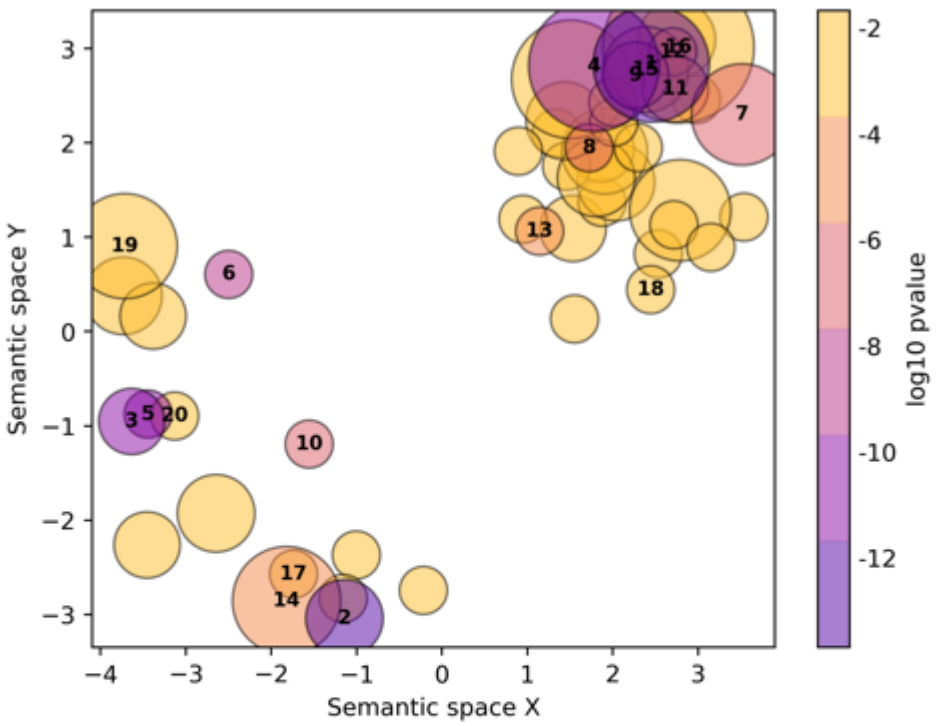
1. structural constituent of ribosome
2. cAMP-dependent protein kinase regulator activity
3. voltage-gated cation channel activity
4. 3',5'-cyclic-GMP phosphodiesterase activity
5. cation channel activity
6. cAMP binding
7. potassium ion transmembrane transporter activity
8. protein C-terminal carboxyl O-methyltransferase activity
9. myosin light chain binding
10. voltage-gated sodium channel activity
11. Rab geranylgeranyltransferase activity
12. calcium channel regulator activity
13. S100 protein binding
14. aminomethyltransferase activity
15. oxygen carrier activity
16. trans-2-enoyl-CoA reductase (NADPH) activity
17. phosphatidylinositol 3-kinase binding
18. arginine transmembrane transporter activity
19. NAD(P)H oxidase H2O2-forming activity
20. NADH binding

Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score ordinal regression



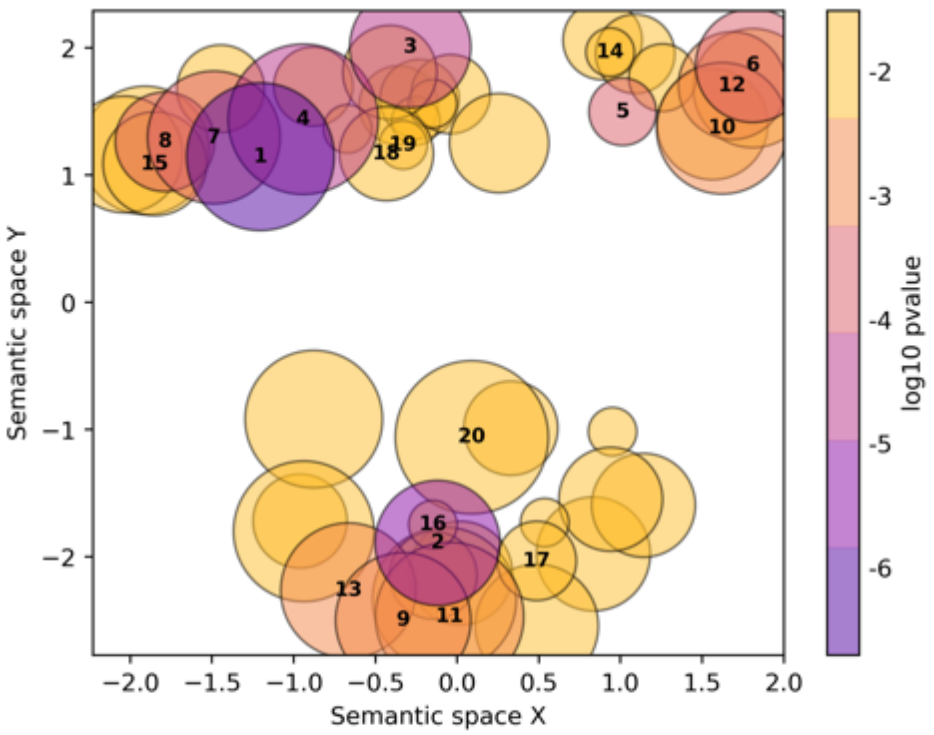
1. protein binding
2. binding
3. kinase activity
4. DNA-binding transcription factor activity, RNA polymerase II-specific
5. voltage-gated cation channel activity
6. transcription factor binding
7. RNA polymerase II transcription regulatory region sequence-specific DNA binding
8. potassium ion transmembrane transporter activity
9. molecular transducer activity
10. enzyme activator activity
11. chromatin binding
12. transporter activity
13. protein-containing complex binding
14. protein C-terminal carboxyl O-methyltransferase activity
15. adrenergic receptor binding
16. signal sequence binding
17. calcium-dependent protein serine/threonine phosphatase activity
18. inositol-1,3,4,5-tetrakisphosphate 5-phosphatase activity
19. protein-macromolecule adaptor activity
20. phospholipid binding

Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the Braak score ordinal regression



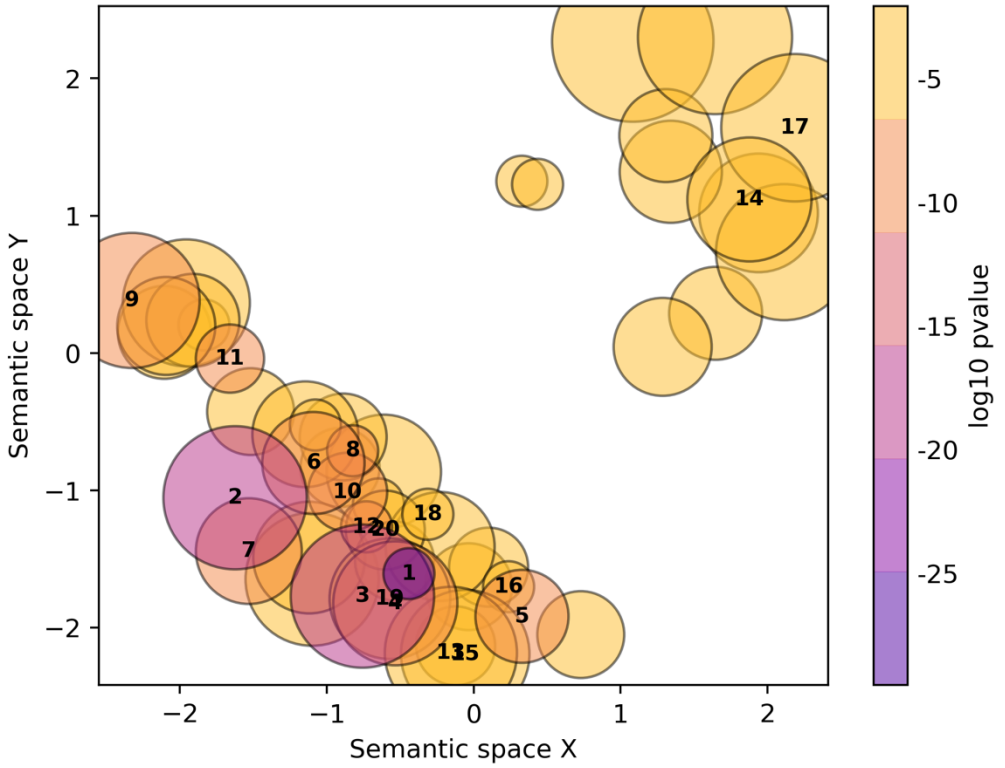
1. ribosomal subunit
2. mitochondrial inner membrane
3. ribosome
4. mitochondrial protein-containing complex
5. mitochondrion
6. mitochondrial matrix
7. oxidoreductase complex
8. cytosolic small ribosomal subunit
9. respiratory chain complex
10. respirasome
11. mitochondrial respiratory chain complex I
12. U6 snRNP
13. Lsm2-8 complex
14. organelle lumen
15. proton-transporting ATP synthase complex, coupling factor F(o)
16. U4/U6 x U5 tri-snRNP complex
17. peroxisomal matrix
18. mitochondrial proton-transporting ATP synthase complex
19. motile cilium
20. peroxisome

Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the Braak score ordinal regression



1. synapse
2. cytosolic large ribosomal subunit
3. integral component of postsynaptic membrane
4. neuron projection
5. postsynaptic density
6. ribosome
7. synaptic membrane
8. mitochondrial inner membrane
9. transporter complex
10. chromaffin granule
11. SCF ubiquitin ligase complex
12. contractile fiber
13. cation channel complex
14. neurofilament cytoskeleton
15. clathrin-sculpted glutamate transport vesicle membrane
16. prefoldin complex
17. CERF complex
18. calyx of Held
19. photoreceptor connecting cilium
20. Lsm2-8 complex

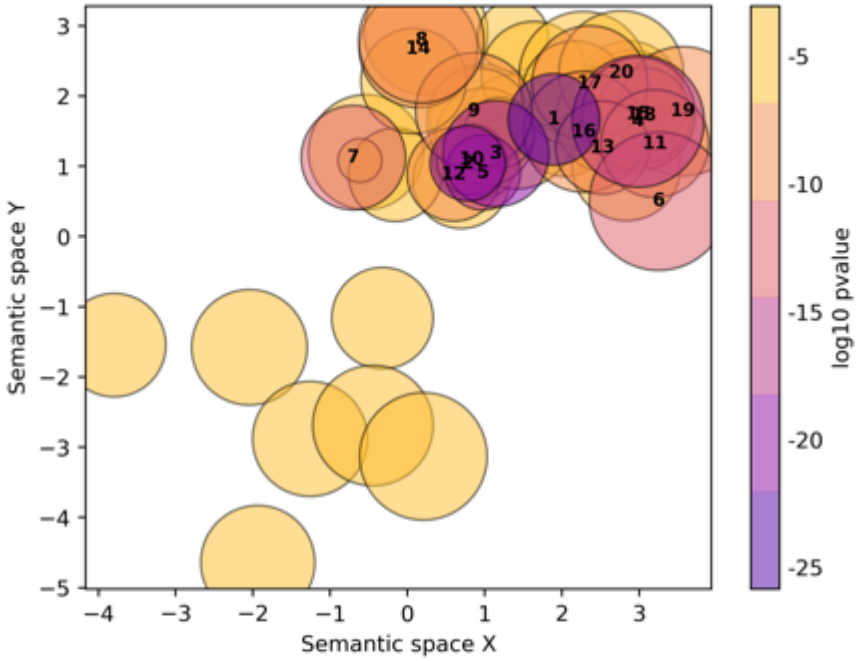
Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the Braak score ordinal regression



1. plasma membrane
2. neuron projection
3. synapse
4. plasma membrane region
5. integral component of postsynaptic density membrane
6. chromatin
7. intrinsic component of plasma membrane
8. presynapse
9. cytoplasmic vesicle
10. cytosol
11. postsynaptic density
12. cell body
13. vesicle membrane
14. voltage-gated potassium channel complex
15. bounding membrane of organelle
16. intrinsic component of membrane
17. plasma membrane protein complex
18. postsynapse
19. dendrite membrane
20. neuron projection terminus

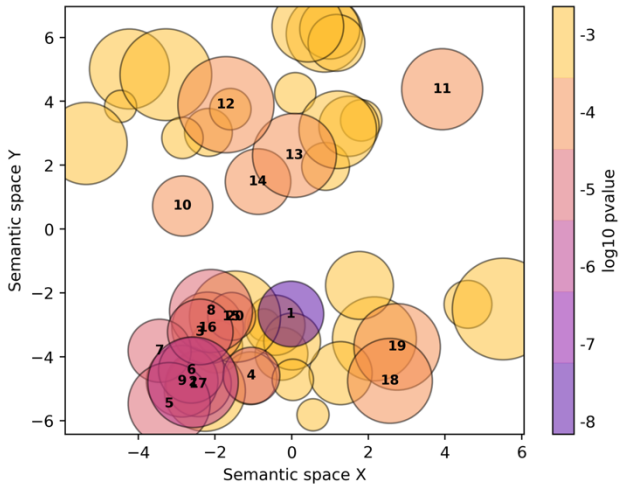
Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the Braak score ordinal regression

Go-Figure! results for reduced CERAD logistic regression analysis.



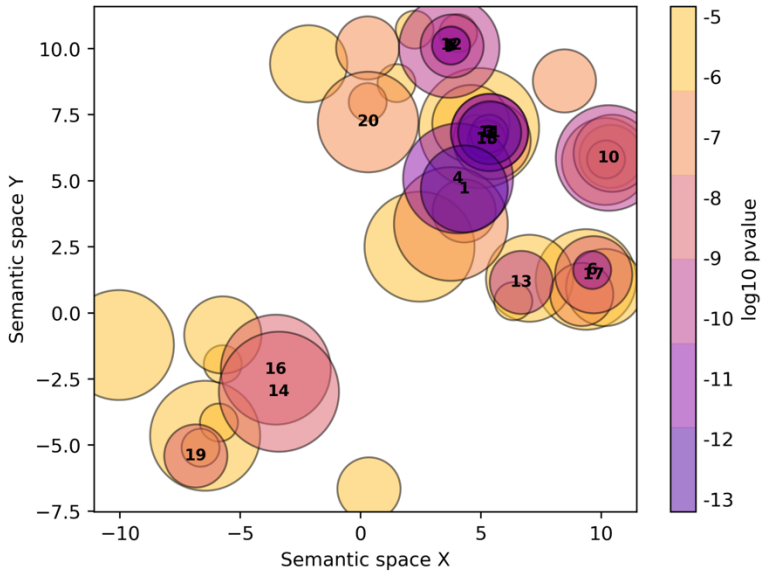
1. cellular amide metabolic process
2. drug metabolic process
3. generation of precursor metabolites and energy
4. monocarboxylic acid catabolic process
5. metabolic process
6. SRP-dependent cotranslational protein targeting to membrane
7. viral transcription
8. mitochondrial respiratory chain complex assembly
9. sulfur compound metabolic process
10. translational initiation
11. mitochondrial ATP synthesis coupled proton transport
12. cellular detoxification
13. hydrogen peroxide catabolic process
14. inner mitochondrial membrane organization
15. organic cyclic compound catabolic process
16. lipid metabolic process
17. alpha-amino acid metabolic process
18. heterocycle catabolic process
19. ATP synthesis coupled proton transport
20. nucleotide metabolic process

Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



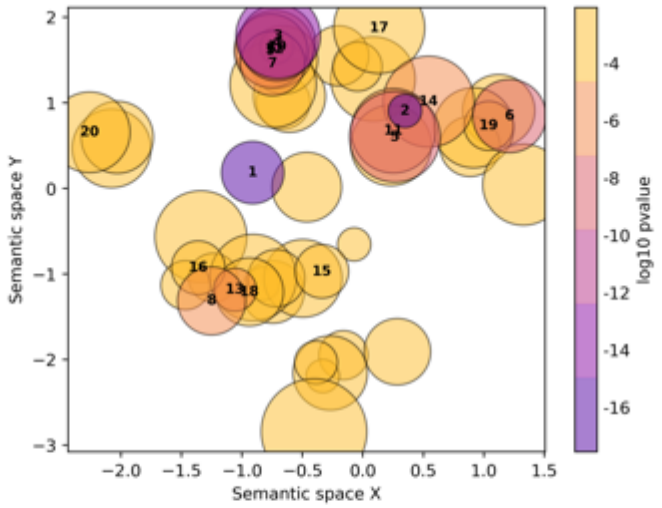
1. respiratory electron transport chain
2. monocarboxylic acid catabolic process
3. mitochondrial respiratory chain complex assembly
4. amide biosynthetic process
5. ATP synthesis coupled proton transport
6. cellular catabolic process
7. cotranslational protein targeting to membrane
8. calcium ion-regulated exocytosis of neurotransmitter
9. mitochondrial ATP synthesis coupled proton transport
10. negative regulation of mitochondrial membrane permeability involved in apoptotic process
11. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
12. modulation of chemical synaptic transmission
13. membrane depolarization during action potential
14. nerve growth factor signaling pathway
15. neurofilament cytoskeleton organization
16. NADH dehydrogenase complex assembly
17. glycine decarboxylation via glycine cleavage system
18. protein targeting to ER
19. establishment of localization in cell
20. neuron projection development

Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



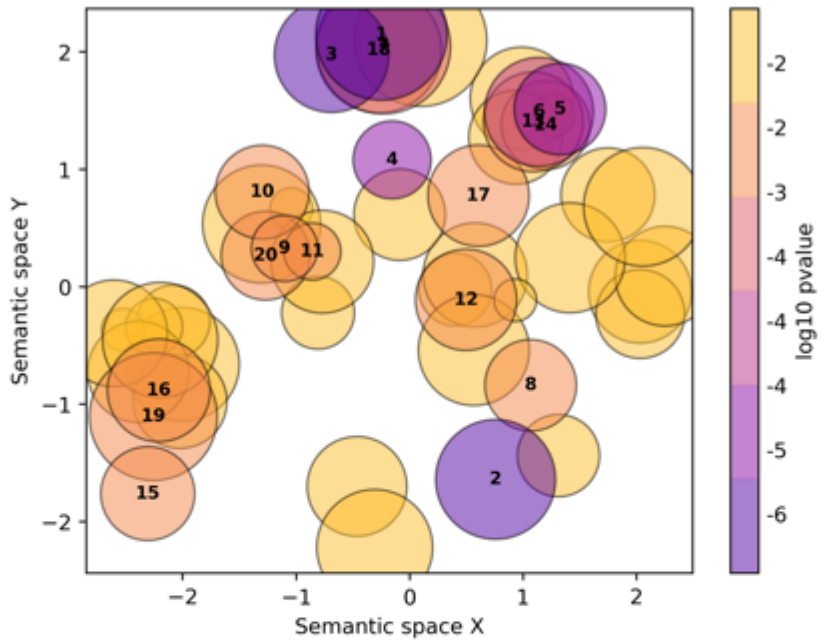
1. regulation of transcription by RNA polymerase II
2. regulation of RNA metabolic process
3. regulation of transcription, DNA-templated
4. modulation of chemical synaptic transmission
5. negative regulation of cellular macromolecule biosynthetic process
6. regulation of nervous system development
7. regulation of cellular macromolecule biosynthetic process
8. negative regulation of RNA metabolic process
9. negative regulation of transcription, DNA-templated
10. positive regulation of RNA metabolic process
11. regulation of cellular biosynthetic process
12. negative regulation of cellular biosynthetic process
13. regulation of plasma membrane bounded cell projection organization
14. chemical synaptic transmission
15. regulation of nitrogen compound metabolic process
16. signaling
17. regulation of localization
18. regulation of primary metabolic process
19. chromatin organization
20. regulation of neuron differentiation

Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



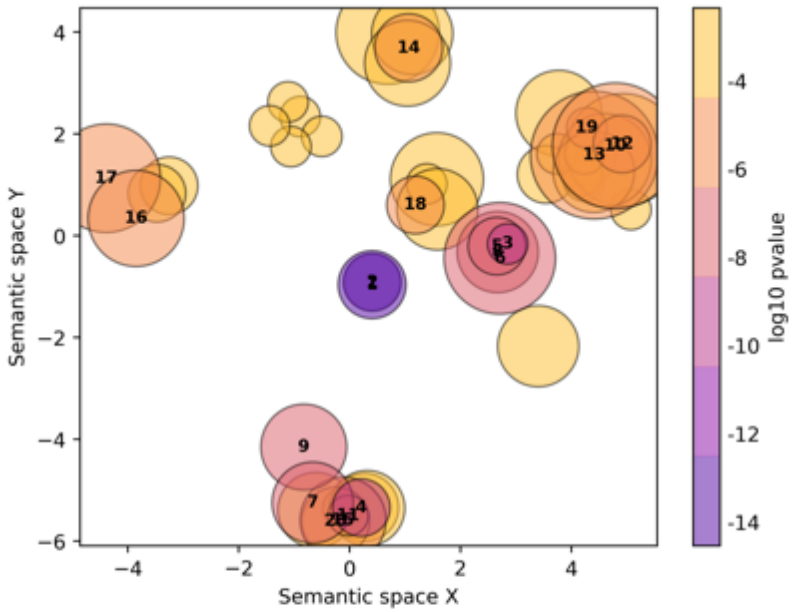
1. structural constituent of ribosome
2. catalytic activity
3. electron transfer activity
4. oxidoreductase activity, acting on NAD(P)H
5. ligase activity
6. acetyl-CoA C-acyltransferase activity
7. peroxidase activity
8. flavin adenine dinucleotide binding
9. oxidoreductase activity, acting on the aldehyde or oxo group of donors
10. oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
11. lyase activity
12. oxidoreductase activity, acting on the CH-CH group of donors
13. NAD binding
14. catalytic activity, acting on a tRNA
15. 2 iron, 2 sulfur cluster binding
16. large ribosomal subunit rRNA binding
17. nucleoside-triphosphatase activity
18. vitamin binding
19. aminomethyltransferase activity
20. pyrophosphate hydrolysis-driven proton transmembrane transporter activity

Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



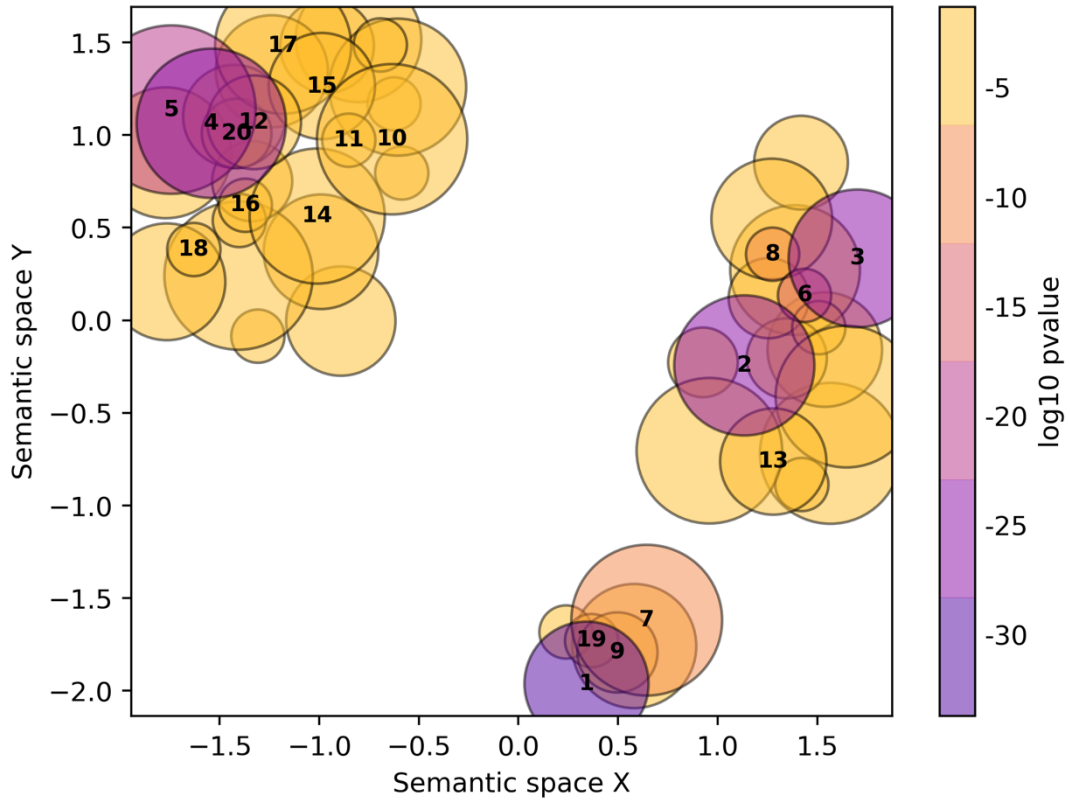
1. voltage-gated cation channel activity
2. acetyl-CoA C-acyltransferase activity
3. cAMP-dependent protein kinase regulator activity
4. structural constituent of ribosome
5. electron transfer activity
6. 3-hydroxyacyl-CoA dehydrogenase activity
7. voltage-gated sodium channel activity
8. aminomethyltransferase activity
9. cAMP binding
10. large ribosomal subunit rRNA binding
11. NADH binding
12. enoyl-CoA hydratase activity
13. succinate dehydrogenase activity
14. peroxidase activity
15. opioid peptide activity
16. myosin light chain binding
17. pyruvate dehydrogenase (NAD+) activity
18. P-P-bond-hydrolysis-driven protein transmembrane transporter activity
19. olfactory receptor binding
20. manganese ion binding

Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



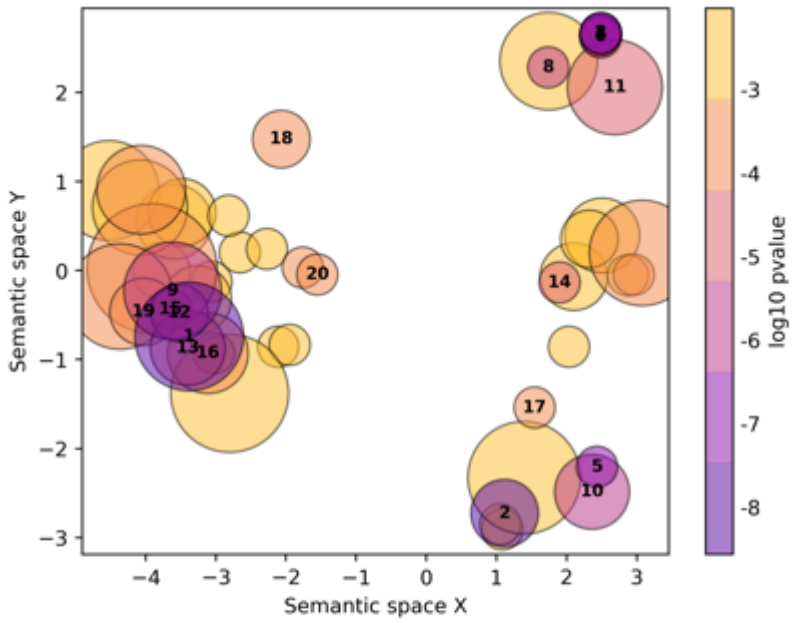
1. DNA-binding transcription factor activity, RNA polymerase II-specific
2. transcription regulator activity
3. DNA binding
4. voltage-gated cation channel activity
5. sequence-specific DNA binding
6. RNA polymerase II transcription regulatory region sequence-specific DNA binding
7. voltage-gated ion channel activity
8. double-stranded DNA binding
9. ion gated channel activity
10. cytoskeletal protein binding
11. potassium channel activity
12. transcription factor binding
13. protein domain specific binding
14. transcription coregulator activity
15. cation channel activity
16. kinase activity
17. protein serine/threonine kinase activity
18. chromatin binding
19. ephrin receptor binding
20. metal ion transmembrane transporter activity

Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



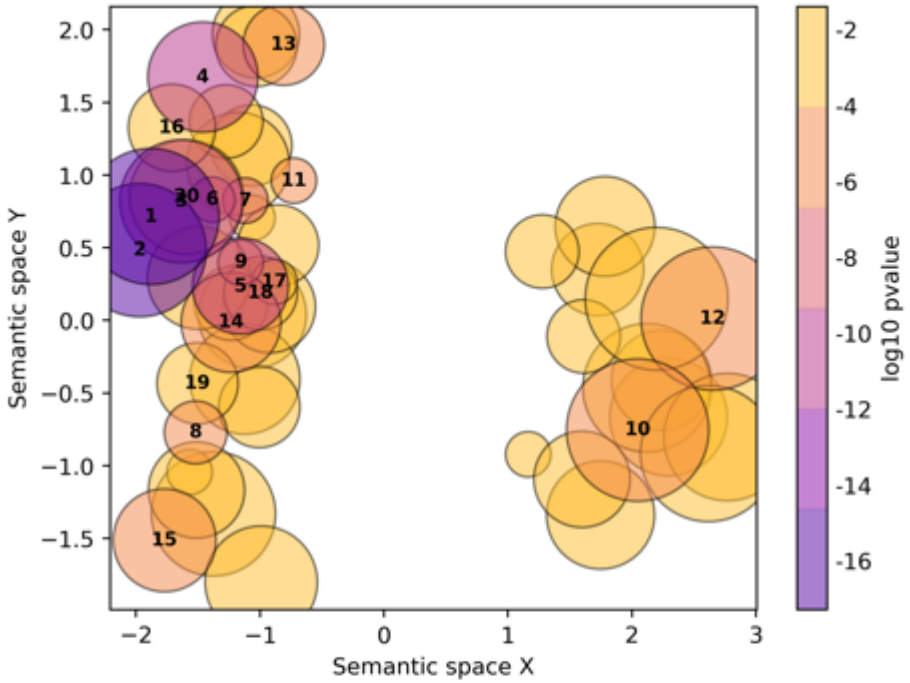
1. mitochondrion
2. intracellular organelle lumen
3. mitochondrial inner membrane
4. mitochondrial protein-containing complex
5. ribosomal subunit
6. respirasome
7. extracellular exosome
8. extracellular region
9. peroxisome
10. Lsm2-8 complex
11. Lsm1-7-Pat1 complex
12. tricarboxylic acid cycle enzyme complex
13. collagen-containing extracellular matrix
14. endoplasmic reticulum chaperone complex
15. methylosome
16. apolipoprotein B mRNA editing enzyme complex
17. protein kinase CK2 complex
18. MHC class I protein complex
19. mitochondrial nucleoid
20. eukaryotic translation initiation factor 3 complex

Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



1. mitochondrial protein-containing complex
2. mitochondrial inner membrane
3. intrinsic component of postsynaptic membrane
4. intrinsic component of synaptic membrane
5. mitochondrion
6. intrinsic component of postsynaptic density membrane
7. intrinsic component of postsynaptic specialization membrane
8. mitochondrial matrix
9. ribosomal subunit
10. ribosome
11. synapse
12. respiratory chain complex
13. cation channel complex
14. respirasome
15. proton-transporting ATP synthase complex, coupling factor F(o)
16. voltage-gated calcium channel complex
17. neurofilament
18. mitochondrial proton-transporting ATP synthase complex
19. transmembrane transporter complex
20. Lsm2-8 complex

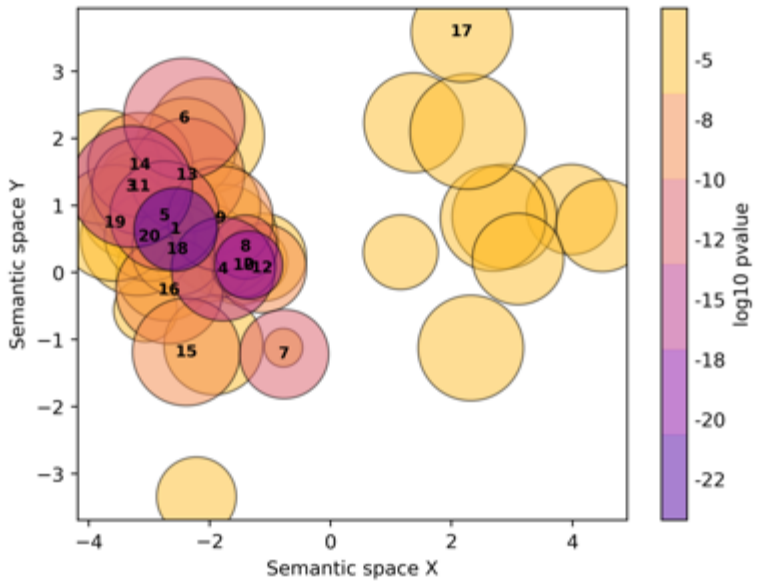
Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the reduced CERAD logistic regression



1. neuron projection
2. synapse
3. synaptic membrane
4. intrinsic component of postsynaptic density membrane
5. chromatin
6. plasma membrane
7. neuronal cell body
8. postsynaptic density
9. presynapse
10. voltage-gated potassium channel complex
11. postsynapse
12. transmembrane transporter complex
13. integral component of presynaptic membrane
14. nuclear body
15. nucleus
16. clathrin-sculpted glutamate transport vesicle membrane
17. neuronal cell body membrane
18. neuron projection terminus
19. neurofibrillary tangle
20. dendrite membrane

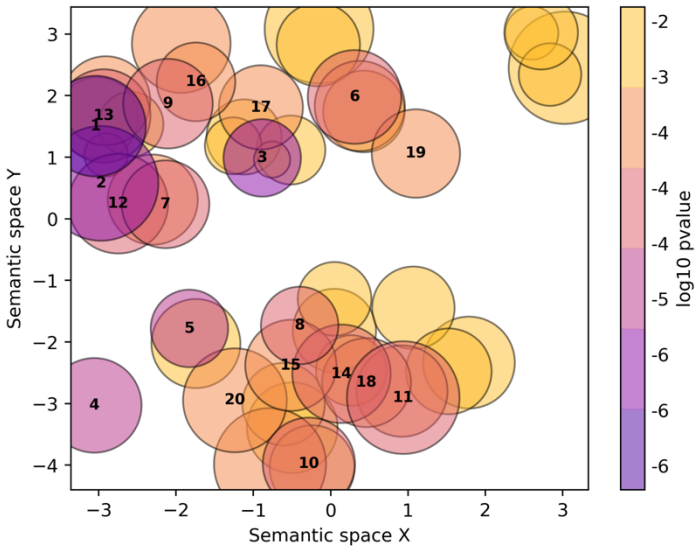
Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the reduced CERAD logistic regression

Go-Figure! results for CERAD ordinalregression analysis.



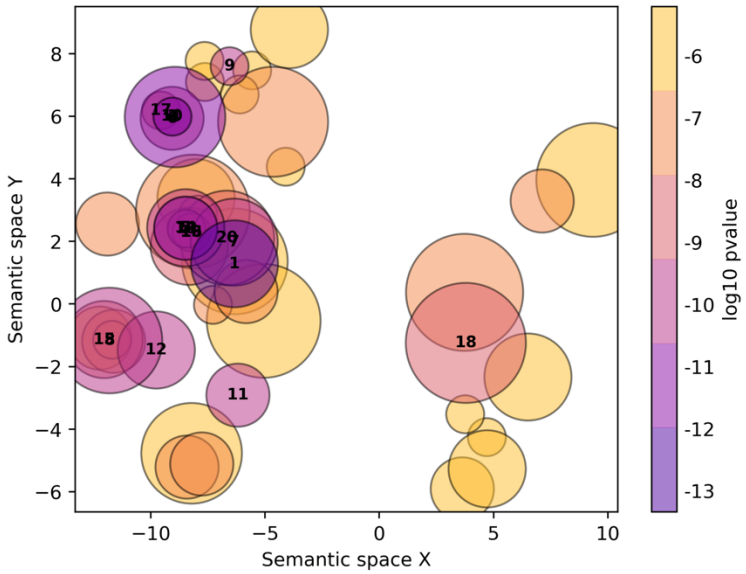
1. cellular amide metabolic process
2. drug metabolic process
3. carboxylic acid catabolic process
4. generation of precursor metabolites and energy
5. monocarboxylic acid metabolic process
6. SRP-dependent cotranslational protein targeting to membrane
7. viral transcription
8. metabolic process
9. sulfur compound metabolic process
10. translational initiation
11. organic cyclic compound catabolic process
12. cellular detoxification
13. mitochondrial respiratory chain complex I assembly
14. mitochondrial ATP synthesis coupled proton transport
15. inner mitochondrial membrane organization
16. lipid metabolic process
17. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
18. translational elongation
19. nucleotide metabolic process
20. purine-containing compound metabolic process

Scatterplot of biological process gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the CERAD ordinal regression



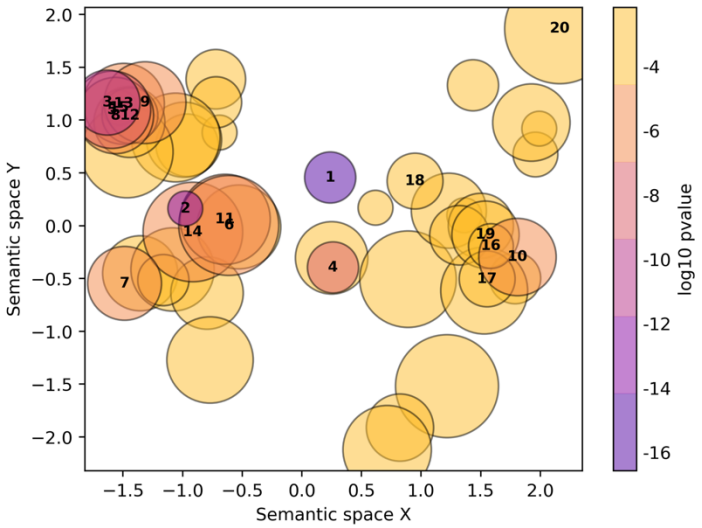
1. monocarboxylic acid catabolic process
2. cotranslational protein targeting to membrane
3. respiratory electron transport chain
4. nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
5. negative regulation of mitochondrial membrane permeability involved in apoptotic process
6. calcium ion-regulated exocytosis of neurotransmitter
7. neurofilament cytoskeleton organization
8. nerve growth factor signaling pathway
9. amide biosynthetic process
10. positive regulation of sodium ion transport
11. modulation of chemical synaptic transmission
12. mitochondrial respiratory chain complex assembly
13. glycine decarboxylation via glycine cleavage system
14. regulation of membrane potential
15. regulation of receptor localization to synapse
16. peptidyl-threonine phosphorylation
17. lipid modification
18. regulation of microtubule depolymerization
19. viral transcription
20. regulation of heart contraction

Scatterplot of biological process gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD ordinal regression



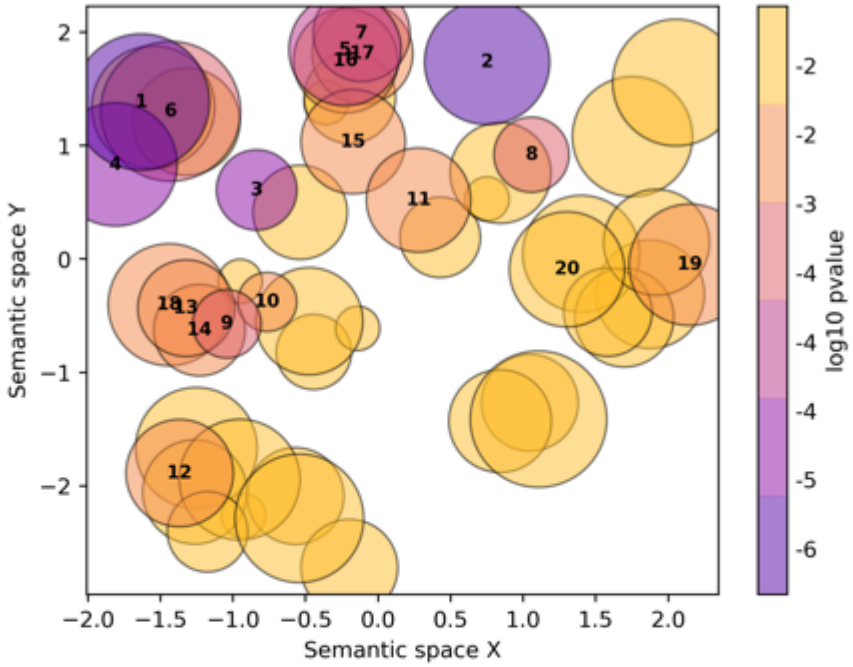
1. regulation of transcription by RNA polymerase II
2. regulation of RNA metabolic process
3. negative regulation of cellular macromolecule biosynthetic process
4. negative regulation of transcription, DNA-templated
5. regulation of transcription, DNA-templated
6. negative regulation of RNA metabolic process
7. modulation of chemical synaptic transmission
8. positive regulation of RNA metabolic process
9. regulation of nervous system development
10. negative regulation of cellular biosynthetic process
11. regulation of plasma membrane bounded cell projection organization
12. regulation of localization
13. regulation of cellular macromolecule biosynthetic process
14. regulation of cellular biosynthetic process
15. positive regulation of cellular process
16. regulation of nitrogen compound metabolic process
17. negative regulation of cellular metabolic process
18. chemical synaptic transmission
19. regulation of primary metabolic process
20. regulation of synaptic plasticity

Scatterplot of biological process gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the CERAD ordinal regression



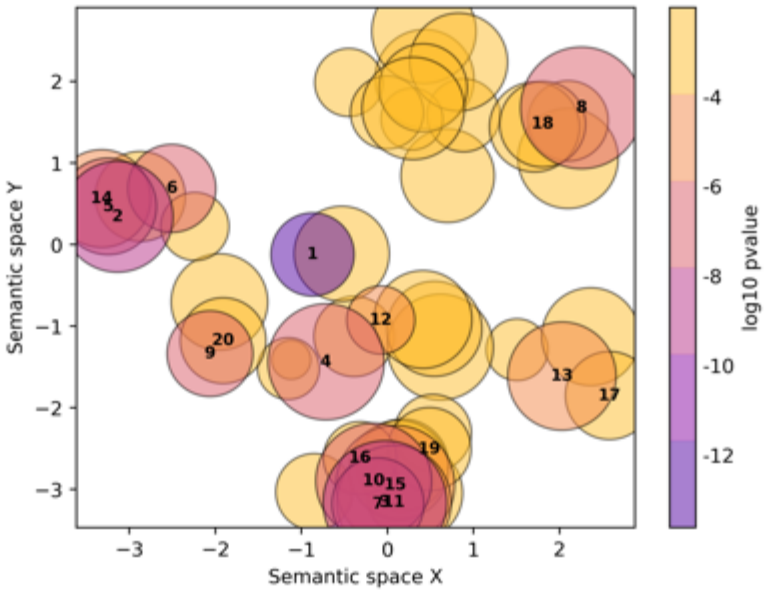
1. structural constituent of ribosome
2. catalytic activity
3. electron transfer activity
4. antioxidant activity
5. oxidoreductase activity, acting on NAD(P)H
6. ligase activity
7. acetyl-CoA C-acyltransferase activity
8. oxidoreductase activity, acting on the aldehyde or oxo group of donors
9. 3-hydroxyacyl-CoA dehydrogenase activity
10. flavin adenine dinucleotide binding
11. lyase activity
12. oxidoreductase activity, acting on peroxide as acceptor
13. oxidoreductase activity, acting on the CH-CH group of donors
14. catalytic activity, acting on RNA
15. oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
16. NAD binding
17. large ribosomal subunit rRNA binding
18. 2 iron, 2 sulfur cluster binding
19. vitamin binding
20. hormone activity

Scatterplot of molecular function gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the CERAD ordinal regression



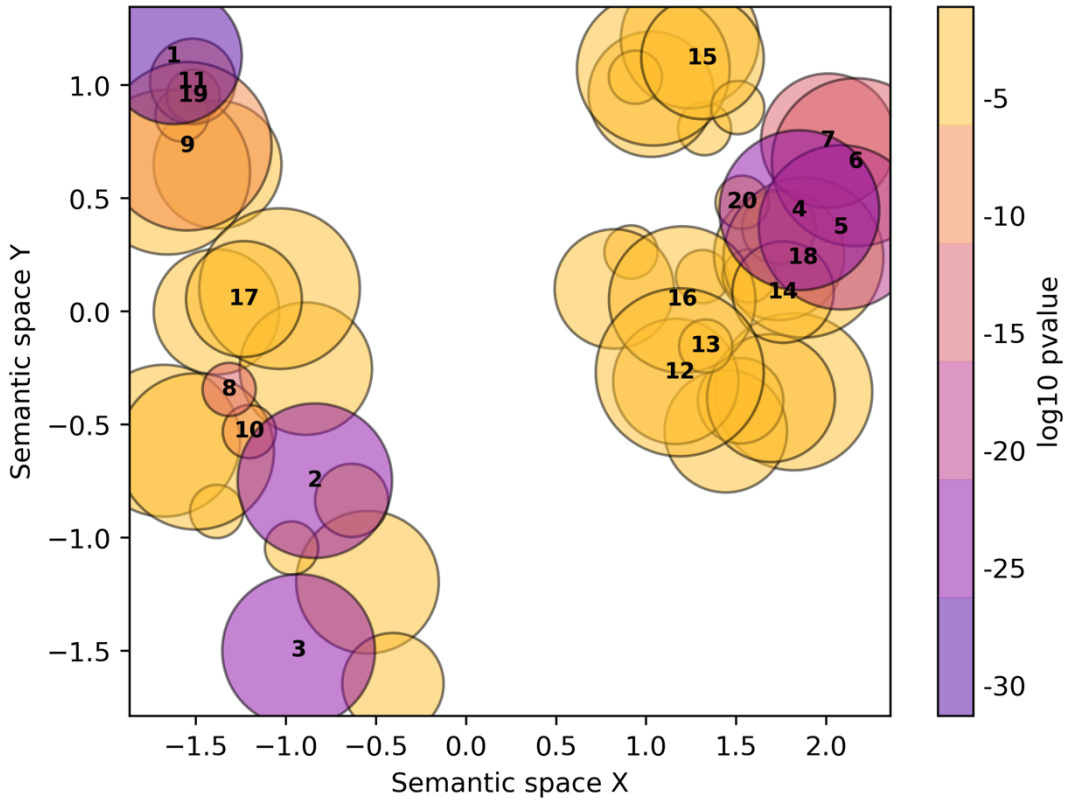
1. voltage-gated cation channel activity
2. acetyl-CoA C-acyltransferase activity
3. structural constituent of ribosome
4. cAMP-dependent protein kinase regulator activity
5. 3-hydroxyacyl-CoA dehydrogenase activity
6. voltage-gated sodium channel activity
7. electron transfer activity
8. aminomethyltransferase activity
9. cAMP binding
10. NADH binding
11. enoyl-CoA hydratase activity
12. myosin light chain binding
13. large ribosomal subunit rRNA binding
14. manganese ion binding
15. pyruvate dehydrogenase (NAD+) activity
16. succinate dehydrogenase activity
17. peroxidase activity
18. fatty-acyl-CoA binding
19. calcium-dependent protein serine/threonine phosphatase activity
20. lysozyme activity

Scatterplot of molecular function gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD ordinal regression



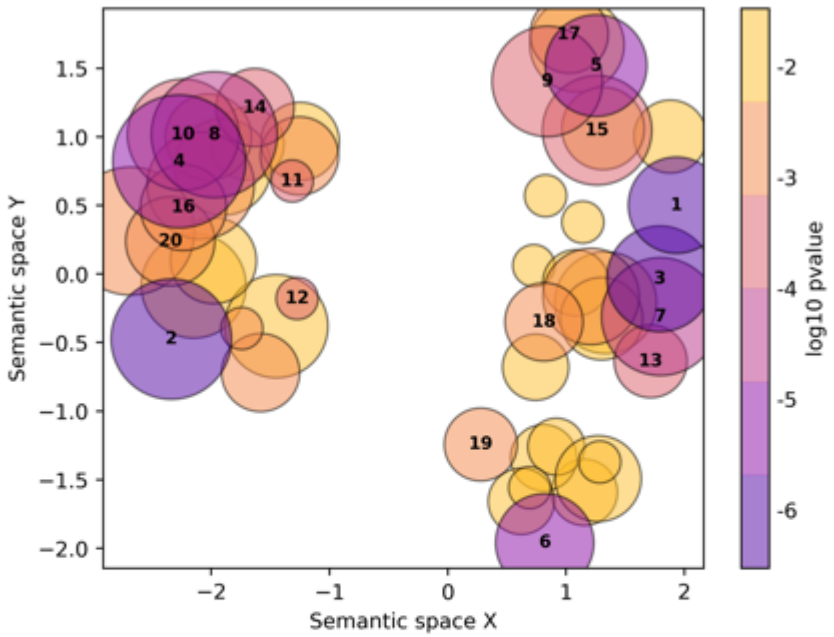
1. transcription regulator activity
2. voltage-gated cation channel activity
3. cytoskeletal protein binding
4. RNA polymerase II transcription regulatory region sequence-specific DNA binding
5. potassium channel activity
6. ion gated channel activity
7. transcription factor binding
8. protein serine/threonine kinase activity
9. transcription coregulator activity
10. protein domain specific binding
11. enzyme binding
12. chromatin binding
13. GTPase activator activity
14. metal ion transmembrane transporter activity
15. ephrin receptor binding
16. bHLH transcription factor binding
17. amino acid transmembrane transporter activity
18. histone methyltransferase activity (H3-K4 specific)
19. RNA polymerase II C-terminal domain binding
20. channel regulator activity

Scatterplot of molecular function gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the CERAD ordinal regression



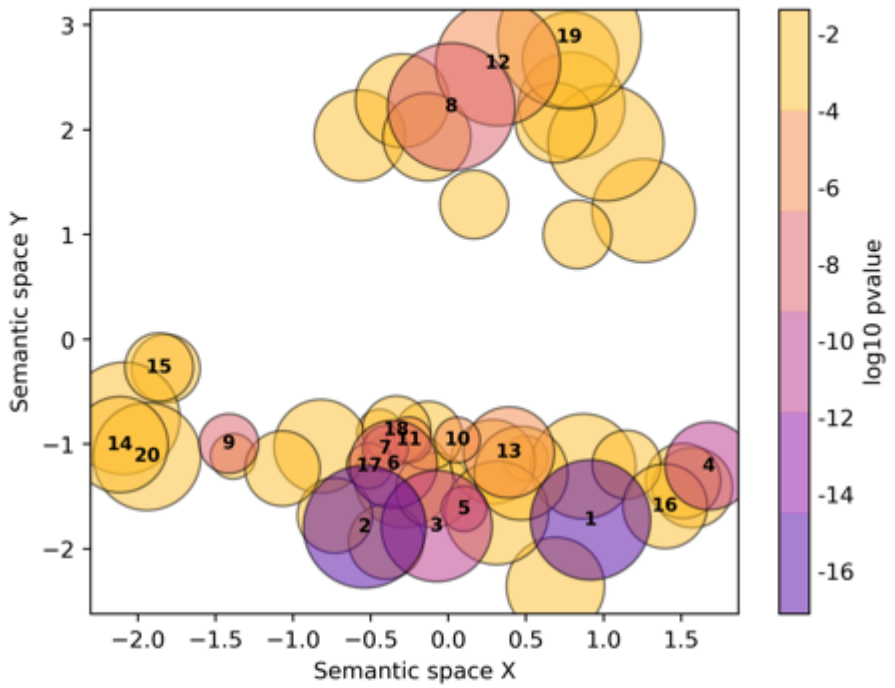
1. mitochondrion
2. intracellular organelle lumen
3. mitochondrial inner membrane
4. mitochondrial protein-containing complex
5. ribosomal subunit
6. oxidoreductase complex
7. respiratory chain complex
8. respirasome
9. extracellular exosome
10. extracellular region
11. peroxisome
12. Lsm2-8 complex
13. Lsm1-7-Pat1 complex
14. tricarboxylic acid cycle enzyme complex
15. methylosome
16. endoplasmic reticulum chaperone complex
17. collagen-containing extracellular matrix
18. protein kinase CK2 complex
19. mitochondrial nucleoid
20. Ragulator complex

Scatterplot of cellular component gene ontology (GO) terms from the down-to-up GO enrichment analysis using gene p-values from the CERAD ordinal regression



1. intrinsic component of postsynaptic membrane
2. mitochondrial protein-containing complex
3. mitochondrial matrix
4. ribosomal subunit
5. mitochondrial inner membrane
6. mitochondrion
7. synapse
8. cation channel complex
9. cerebellar mossy fiber
10. proton-transporting ATP synthase complex, coupling factor F(o)
11. Lsm1-7-Pat1 complex
12. Lsm2-8 complex
13. neurofilament
14. mitochondrial proton-transporting ATP synthase complex
15. postsynaptic membrane
16. glycine cleavage complex
17. clathrin-sculpted glutamate transport vesicle membrane
18. calyx of Held
19. basal dendrite
20. NuRD complex

Scatterplot of cellular component gene ontology (GO) terms from the non-directional GO enrichment analysis using gene p-values from the CERAD ordinal regression



1. neuron projection
2. synapse
3. synaptic membrane
4. integral component of postsynaptic density membrane
5. plasma membrane
6. chromatin
7. neuronal cell body
8. voltage-gated potassium channel complex
9. postsynaptic density
10. presynapse
11. postsynapse
12. transmembrane transporter complex
13. nuclear body
14. nucleus
15. clathrin-sculpted glutamate transport vesicle membrane
16. integral component of presynaptic membrane
17. neuronal cell body membrane
18. neuron projection terminus
19. histone acetyltransferase complex
20. cytoskeleton

Scatterplot of cellular component gene ontology (GO) terms from the up-to-down GO enrichment analysis using gene p-values from the CERAD ordinal regression

Appendix 2

The list of GWAS index SNPs that were selected from five GWAS and GWAX studies can be found below. S13 = GWAS (Lambert et al. 2013) ; X18 = GWAX (Marioni et al. 2018); X19 = GWAX (Jansen et al. 2019); S19 = GWAS (Kunkle et al. 2019) ;W21 = GWAX (Wightman et al. 2021). All in genome build GRCh38 (www.gencodegenes.org/human/release_24.html).

GWAS	SNP	chr:position
X19	rs4575098	1:161185602
S13, X18	rs6656401	1:207518704
W21	rs679515	1:207577223
X19	rs2093760	1:207613483
S19	rs4844610	1:207629207
X19, W21	rs4663105	2:127133851
S13, X18, S19	rs6733839	2:127135234
X19, S19	rs10933431	2:233117202
S13, X18	rs35349669	2:233159830
W21	rs7597763	2:233173931
W21	rs4504245	4:11013198
X19	rs6448453	4:11024404
S13	rs190982	5:88927603
W21	rs871269	5:151052827
W21	rs6891966	5:157099320
X18	rs34855541	6:32592048
S19	rs9271058	6:32607629
S13	rs111418223	6:32610753

X19	rs6931277	6:32615580
W21	rs1846190	6:32616036
S19	rs114812713	6:41066261
X18	rs9381040	6:41186912
S19	rs9473117	6:47463548
X18, X19	rs9381563	6:47464901
S13	rs10948363	6:47520026
W21	rs9369716	6:47584444
W21	rs5011436	7:12229132
S13	rs2718058	7:37801932
W21	rs7384878	7:100334426
X19	rs1859788	7:100374211
S13, X18	rs1476679	7:100406823
S19	rs12539172	7:100494172
X18, S19	rs10808026	7:143402040
W21	rs3935067	7:143407238
X19	rs7810606	7:143411065
S13	rs11771145	7:143413669
X18, X19	rs4236673	8:27607412
W21	rs1532278	8:27608798
S13, S19	rs9331896	8:27610169
W21	rs61732533	8:144053248
X19	rs11257238	10:11675398
W21	rs7912495	10:11676714
X18, S19	rs7920721	10:11678309
W21	rs7902657	10:59978394

S19, W21	rs3740688	11:47358789
X18	rs12292911	11:47427521
S13	rs10838725	11:47536319
S13	rs983392	11:60156035
S19	rs7933202	11:60169453
X19	rs2081545	11:60190907
X18, W21	rs1582763	11:60254475
X19	rs867611	11:86065502
W21	rs561655	11:86089237
S13, X18	rs10792832	11:86156833
S19	rs3851179	11:86157598
S13, X18, X19, S19, W21	rs11218343	11:121564878
W21	rs7146179	14:52832135
X18, S19	rs17125924	14:52924962
S13	rs17125944	14:52933911
S13	rs10498633	14:92460608
S19	rs12881735	14:92466484
X18, X19	rs12590654	14:92472511
X18	rs59685680	15:50709337
X19	rs442495	15:58730416
X18, S19	rs593742	15:58753575
W21	rs602602	15:58764824
X18, W21	rs117618017	15:63277703
S19	rs7185636	16:19796841
X18	rs889555	16:31111250
X19	rs59735493	16:31121779

X18	rs4985556	16:70660097
S19	rs62039712	16:79321960
X18	rs12444183	16:81739604
W21	rs7209200	17:5066645
X18	rs7225151	17:5233752
X19	rs113260531	17:5235685
W21	rs708382	17:44364976
X19, W21	rs28394864	17:49373413
X19	rs2526380	17:58320645
W21	rs2632516	17:58331728
X18, S19	rs138190086	17:63460787
W21	rs6504163	17:63468418
X19	rs76726049	18:58522227
X19	rs111278892	19:1039324
X18	rs3752231	19:1043639
W21	rs12151021	19:1050875
S19	rs3752246	19:1056493
S13	rs4147929	19:1063444
S13, X18, X19	rs41289512	19:44848259
W21	rs429358	19:44908684
S19	rs12691088	19:44915229
W21	rs2452170	19:48710247
X19	rs3865444	19:51224706
X18	rs12459419	19:51225221
W21	rs1354106	19:51234736
W21	rs1761461	19:54313903

W21	rs6069736	20:56408019
W21	rs6069737	20:56420643
S19	rs6024870	20:56422512
X19	rs6014724	20:56423488
S13	rs7274581	20:56443204
W21	rs2154482	21:26148613
S19	rs2830500	21:26784537

References

Albert, F. W., Bloom, J. S., Siegel, J., Day, L. and Kruglyak, L. 2018. Genetics of trans-regulatory variation in gene expression. *Elife* 7, doi: 10.7554/eLife.35471

Albert, F. W. and Kruglyak, L. 2015. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* 16(4), pp. 197-212. doi: 10.1038/nrg3891

Allen, M. et al. 2016. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data* 3, p. 160089. doi: 10.1038/sdata.2016.89

Alzheimer, A. 1*907*. Uber eine eigenartige Erkrankung

der Hirnrinde. *Allg Zeitschrift Psychiatr*

Alzheimer, A., Stelzmann, R. A., Schnitzlein, H. N. and Murtagh, F. R. 1995. An English translation of Alzheimer's 1907 paper, "Uber eine eigenartige Erkankung der Hirnrinde". *Clin Anat* 8(6), pp. 429-431. doi: 10.1002/ca.980080612

Andrews, S. J., Fulton-Howard, B. and Goate, A. 2020. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *The Lancet. Neurology* 19(4), pp. 326-335. doi: 10.1016/S1474-4422(19)30435-1

Angelucci, F., Cechova, K., Valis, M., Kuca, K., Zhang, B. and Hort, J. 2019. MicroRNAs in Alzheimer's Disease: Diagnostic Markers or Therapeutic Agents? *Front Pharmacol* 10, p. 665. doi: 10.3389/fphar.2019.00665

Atri, A. 2019a. Current and Future Treatments in Alzheimer's Disease. *Semin Neurol* 39(2), pp. 227-240. doi: 10.1055/s-0039-1678581

Atri, A. 2019b. The Alzheimer's Disease Clinical Spectrum: Diagnosis and Management. *Medical Clinics of North America* 103(2), pp. 263-293. doi: https://doi.org/10.1016/j.mcna.2018.10.009

Auton, A. et al. 2015. A global reference for human genetic variation. *Nature* 526(7571), pp. 68-74. doi: 10.1038/nature15393

Bagyinszky, E., Giau, V. V. and An, S. A. 2020. Transcriptomics in Alzheimer's Disease: Aspects and Challenges. *International journal of molecular sciences* 21(10), p. 3517. doi: 10.3390/ijms21103517

Baker, E. and Escott-Price, V. 2020. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. *Front Digit Health* 2, p. 14. doi: 10.3389/fdgth.2020.00014

Bali, J., Gheinani, A. H., Zurbriggen, S. and Rajendran, L. 2012. Role of genes linked to sporadic Alzheimer's disease risk in the production of β-amyloid peptides. *Proceedings of the National Academy of Sciences* 109(38), p. 15307. doi: 10.1073/pnas.1201632109

Barbeira, A. N. et al. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 9(1), p. 1825. doi: 10.1038/s41467-018-03621-1

Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *2015* 67(1), p. 48. doi: 10.18637/jss.v067.i01

Bellenguez, C. et al. 2022. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* 54(4), pp. 412-436. doi: 10.1038/s41588-022-01024-z

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), pp. 289-300.

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S. and Schneider, J. A. 2018. Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's disease : JAD* 64(s1), pp. S161-S189. doi: 10.3233/JAD-179939

Bennett, D. A., Schneider, J. A., Arvanitakis, Z., Kelly, J. F., Aggarwal, N. T., Shah, R. C. and Wilson, R. S. 2006. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* 66(12), pp. 1837-1844. doi: 10.1212/01.wnl.0000219668.47116.e6

Bennett, D. A., Schneider, J. A., Arvanitakis, Z. and Wilson, R. S. 2012a. Overview and findings from the religious orders study. *Curr Alzheimer Res* 9(6), pp. 628-645.

Bennett, D. A., Schneider, J. A., Buchman, A. S., Barnes, L. L., Boyle, P. A. and Wilson, R. S. 2012b. Overview and findings from the rush Memory and Aging Project. *Curr Alzheimer Res* 9(6), pp. 646-663.

Berron, D. et al. 2021. Early stages of tau pathology and its associations with functional connectivity, atrophy and memory. *Brain* 144(9), pp. 2771-2783. doi: 10.1093/brain/awab114

Bihlmeyer, N. A., Merrill, E., Lambert, Y., Srivastava, G. P., Clark, T. W., Hyman, B. T. and Das, S. 2019. Novel methods for integration and visualization of genomics and genetics data in Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 15(6), pp. 788-798. doi: 10.1016/j.jalz.2019.01.011

Binder, J. R. 2015. The Wernicke area: Modern evidence and a reinterpretation. *Neurology* 85(24), pp. 2170-2175.

Blennow, K. and Zetterberg, H. 2018. Biomarkers for Alzheimer's disease: current status and prospects for the future. *J Intern Med* 284(6), pp. 643-663. doi: 10.1111/joim.12816

Boluda, S. et al. 2014. A comparison of Aβ amyloid pathology staging systems and correlation with clinical diagnosis. *Acta neuropathologica* 128(4), pp. 543-550. doi: 10.1007/s00401-014-1308-9

Boyle, E. A., Li, Y. I. and Pritchard, J. K. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169(7), pp. 1177-1186. doi: 10.1016/j.cell.2017.05.038

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. and Del Tredici, K. 2006. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol* 112(4), pp. 389-404. doi: 10.1007/s00401-006-0127-z

Breijyeh, Z. and Karaman, R. 2020. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules (Basel, Switzerland)* 25(24), p. 5789. doi: 10.3390/molecules25245789

Breslin, T., Edén, P. and Krogh, M. 2004. Comparing functional annotation analyses with Catmap. *BMC bioinformatics* 5, pp. 193-193. doi: 10.1186/1471-2105-5-193

Brookmeyer, R., Johnson, E., Ziegler-Graham, K. and Arrighi, H. M. 2007. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia* 3(3), pp. 186-191. doi: https://doi.org/10.1016/j.jalz.2007.04.381

Bush, W. S. and Moore, J. H. 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology* 8(12), pp. e1002822-e1002822. doi: 10.1371/journal.pcbi.1002822

Cabeza, R. and Nyberg, L. 2000. Imaging cognition II: An empirical review of 275 PET and fMRI studies. *J Cogn Neurosci* 12(1), pp. 1-47. doi: 10.1162/08989290051137585

Calligaris, R. et al. 2015. Blood transcriptomics of drug-naive sporadic Parkinson's disease patients. *BMC Genomics* 16, p. 876. doi: 10.1186/s12864-015-2058-3

Cano-Gamez, E. and Trynka, G. 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* 11, doi: 10.3389/fgene.2020.00424

Carress, H., Lawson, D. J. and Elhaik, E. 2021. Population genetic considerations for using biobanks as international resources in the pandemic era and beyond. *BMC genomics* 22(1), pp. 351-351. doi: 10.1186/s12864-021-07618-x

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. and Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4(1), doi: 10.1186/s13742-015-0047-8

Chappell, S., Patel, T., Guetta-Baranes, T., Sang, F., Francis, P. T., Morgan, K. and Brookes, K. J. 2018. Observations of extensive gene expression differences in the cerebellum and potential relevance to Alzheimer's disease. *BMC Research Notes* 11(1), p. 646. doi: 10.1186/s13104-018-3732-8

Chidananda, A. H., Sharma, A. K., Khandelwal, R. and Sharma, Y. 2019. Secretagogin Binding Prevents α-Synuclein Fibrillation. *Biochemistry* 58(46), pp. 4585-4589. doi: 10.1021/acs.biochem.9b00656

Cieslik, E. C. et al. 2012. Is There "One" DLPFC in Cognitive Action Control? Evidence for Heterogeneity From Co-Activation-Based Parcellation. *Cerebral Cortex* 23(11), pp. 2677-2689. doi: 10.1093/cercor/bhs256

Clyde, D. 2017. Transitioning from association to causation with eQTLs. *Nature Reviews Genetics* 18(5), pp. 271-271. doi: 10.1038/nrg.2017.22

Costa-Silva, J., Domingues, D. and Lopes, F. M. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* 12(12), p. e0190152. doi: 10.1371/journal.pone.0190152

Cousins, K. A. Q. et al. 2021. ATN incorporating cerebrospinal fluid neurofilament light chain detects frontotemporal lobar degeneration. *Alzheimers Dement* 17(5), pp. 822-830. doi: 10.1002/alz.12233

Crow, M., Lim, N., Ballouz, S., Pavlidis, P. and Gillis, J. 2019. Predictability of human differential gene expression. *Proceedings of the National Academy of Sciences* 116(13), pp. 6491-6500. doi: doi:10.1073/pnas.1802973116

Cuyvers, E. and Sleegers, K. 2016. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *Lancet Neurol* 15(8), pp. 857-868. doi: 10.1016/s1474-4422(16)00127-7

Danecek, P. et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10(2), doi: 10.1093/gigascience/giab008

Davis, M., O Connell, T., Johnson, S., Cline, S., Merikle, E., Martenyi, F. and Simpson, K. 2018. Estimating Alzheimer's Disease Progression Rates from Normal Cognition Through Mild Cognitive Impairment and Stages of Dementia. *Current Alzheimer research* 15(8), pp. 777-788. doi: 10.2174/1567205015666180119092427

de Jong, D., Jansen, R. W., Pijnenburg, Y. A., van Geel, W. J., Borm, G. F., Kremer, H. P. and Verbeek, M. M. 2007. CSF neurofilament proteins in the differential diagnosis of dementia. *J Neurol Neurosurg Psychiatry* 78(9), pp. 936-938. doi: 10.1136/jnnp.2006.107326

de Leeuw, C. A., Mooij, J. M., Heskes, T. and Posthuma, D. 2015. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* 11(4), p. e1004219. doi: 10.1371/journal.pcbi.1004219

de Rojas, I. et al. 2021. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nature Communications* 12(1), p. 3417. doi: 10.1038/s41467-021-22491-8

DeLuca, D. S. et al. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28(11), pp. 1530-1532. doi: 10.1093/bioinformatics/bts196

DeTure, M. A. and Dickson, D. W. 2019. The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration* 14(1), p. 32. doi: 10.1186/s13024-019-0333-5

Dharshini, S. A. P., Taguchi, Y. h. and Gromiha, M. M. 2019. Investigating the energy crisis in Alzheimer disease using transcriptome study. *Scientific Reports* 9(1), p. 18509. doi: 10.1038/s41598-019-54782-y

Drew, L. 2018. An age-old story of dementia.*Nature*. Vol. 559. England, pp. S2-S3.

Drew, L. 2022. Turning back time with epigenetic clocks.*Nature*. Vol. 601. England, pp. S20-s22.

Dujardin, P., Vandenbroucke, R. E. and Van Hoecke, L. 2022. Fighting fire with fire: the immune system might be key in our fight against Alzheimer's disease. *Drug Discovery Today*, doi: https://doi.org/10.1016/j.drudis.2022.01.004

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, p. 48. doi: 10.1186/1471-2105-10-48

Efthymiou, A. G. and Goate, A. M. 2017. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Molecular Neurodegeneration* 12(1), p. 43. doi: 10.1186/s13024-017-0184-x

Ertekin-Taner, N. 2017. Identifying therapeutic targets for Alzheimer's disease with big data. *Neurodegener Dis Manag* 7(2), pp. 101-105. doi: 10.2217/nmt-2017-0008

Escott-Price, V. and Hardy, J. 2022. Genome-wide association studies for Alzheimer's disease: bigger is not always better. *Brain Commun* 4(3), p. fcac125. doi: 10.1093/braincomms/fcac125

Escott-Price, V., Myers, A. J., Huentelman, M. and Hardy, J. 2017a. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann Neurol* 82(2), pp. 311-314. doi: 10.1002/ana.24999

Escott-Price, V., Shoai, M., Pither, R., Williams, J. and Hardy, J. 2017b. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiology of Aging* 49, pp. 214.e217-214.e211. doi: https://doi.org/10.1016/j.neurobiolaging.2016.07.018

Escott-Price, V. et al. 2015. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 138(Pt 12), pp. 3673-3684. doi: 10.1093/brain/awv268

Fabregat, A. et al. 2016. The Reactome pathway Knowledgebase. *Nucleic Acids Res* 44(D1), pp. D481-487. doi: 10.1093/nar/gkv1351

Fan, Y. et al. 2022. Generic amyloid fibrillation of TMEM106B in patient with Parkinson's disease dementia and normal elders. *Cell Research*, doi: 10.1038/s41422-022-00665-3

Fenoglio, C., Scarpini, E., Serpente, M. and Galimberti, D. 2018. Role of Genetics and Epigenetics in the Pathogenesis of Alzheimer's Disease and Frontotemporal Dementia. *J Alzheimers Dis* 62(3), pp. 913-932.

Fernández-Espejo, E., Rodriguez de Fonseca, F., Suárez, J. and Martín de Pablos, Á. 2021. Cerebrospinal fluid lactoperoxidase level is enhanced in idiopathic Parkinson's disease, and correlates with levodopa equivalent daily dose. *Brain Res* 1761, p. 147411. doi: 10.1016/j.brainres.2021.147411

Ferreira, P. G. et al. 2018. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature communications* 9(1), pp. 490-490. doi: 10.1038/s41467-017-02772-x

Fillenbaum, G. G. et al. 2008. Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers Dement* 4(2), pp. 96-109.

Fortea, J., Zaman, S. H., Hartley, S., Rafii, M. S., Head, E. and Carmona-Iragui, M. 2021. Alzheimer's disease associated with Down syndrome: a genetic form of dementia. *The Lancet Neurology* 20(11), pp. 930-942. doi: https://doi.org/10.1016/S1474-4422(21)00245-3

Foster, E. M., Dangla-Valls, A., Lovestone, S., Ribe, E. M. and Buckley, N. J. 2019. Clusterin in Alzheimer's Disease: Mechanisms, Genetics, and Lessons From Other Pathologies. *Frontiers in Neuroscience* 13, p. 164.

Gallagher, M. D. and Chen-Plotkin, A. S. 2018. The Post-GWAS Era: From Association to Function. *Am J Hum Genet* 102(5), pp. 717-730. doi: 10.1016/j.ajhg.2018.04.002

Gamazon, E. R. et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 47(9), pp. 1091-1098. doi: 10.1038/ng.3367

Gatz, M. et al. 2006. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* 63(2), pp. 168-174. doi: 10.1001/archpsyc.63.2.168

Gerring, Z. F., Lupton, M. K., Edey, D., Gamazon, E. R. and Derks, E. M. 2020. An analysis of genetically regulated gene expression across multiple tissues implicates novel gene candidates in Alzheimer's disease. *Alzheimer's research & therapy* 12(1), pp. 43-43. doi: 10.1186/s13195-020-00611-8

Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D. and Burgess, P. W. 2006. Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *J Cogn Neurosci* 18(6), pp. 932-948. doi: 10.1162/jocn.2006.18.6.932

Gockley, J. et al. 2021. Multi-tissue neocortical transcriptome-wide association study implicates 8 genes across 6 genomic loci in Alzheimer's disease. *Genome Medicine* 13(1), p. 76. doi: 10.1186/s13073-021-00890-2

Goswami, K. and Sanan-Mishra, N. 2022. Chapter 7 - RNA-seq for revealing the function of the transcriptome. In: Singh, D.B. and Pathak, R.K. eds. *Bioinformatics*. Academic Press, pp. 105-129.

Greenwood, A. K. et al. 2020. The AD Knowledge Portal: A Repository for Multi-Omic Data on Alzheimer's Disease and Aging. *Current protocols in human genetics* 108(1), pp. e105-e105. doi: 10.1002/cphg.105

Gregory, S., Saunders, S. and Ritchie, C. W. 2022. Science disconnected: the translational gap between basic science, clinical trials, and patient care in Alzheimer's disease. *Lancet Healthy Longev* 3(11), pp. e797-e803. doi: 10.1016/s2666-7568(22)00219-7

Grill, J. D., Nuño, M. M. and Gillen, D. L. 2019. Which MCI Patients Should be Included in Prodromal Alzheimer Disease Clinical Trials? *Alzheimer Dis Assoc Disord* 33(2), pp. 104-112. doi: 10.1097/wad.0000000000000303

Grundberg, E. et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics* 44(10), pp. 1084-1089. doi: 10.1038/ng.2394

GTEx-Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6), pp. 580-585. doi: 10.1038/ng.2653

Gusev, A. et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48(3), pp. 245-252. doi: 10.1038/ng.3506

Hansen, K. D., Irizarry, R. A. and Wu, Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13(2), pp. 204-216. doi: 10.1093/biostatistics/kxr054

Hao, S., Wang, R., Zhang, Y. and Zhan, H. 2019. Prediction of Alzheimer's Disease-Associated Genes by Integration of GWAS Summary Data and Expression Data. *Frontiers in genetics* 9, pp. 653-653. doi: 10.3389/fgene.2018.00653

Hardy, J. and Allsop, D. 1991. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol Sci* 12(10), pp. 383-388. doi: 10.1016/0165-6147(91)90609-v

Hardy, J. and Selkoe, D. J. 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics.*Science*. Vol. 297. United States, pp. 353-356.

Hardy, J. A. and Higgins, G. A. 1992. Alzheimer's disease: the amyloid cascade hypothesis. *Science* 256(5054), pp. 184-185. doi: 10.1126/science.1566067

Haroutunian, V., Katsel, P. and Schmeidler, J. 2009. Transcriptional vulnerability of brain regions in Alzheimer's disease and dementia. *Neurobiology of aging* 30(4), pp. 561-573. doi: 10.1016/j.neurobiolaging.2007.07.021

Harwood, J. C., Leonenko, G., Sims, R., Escott-Price, V., Williams, J. and Holmans, P. 2021. Defining functional variants associated with Alzheimer's disease in the induced immune response. *Brain Communications* 3(2), doi: 10.1093/braincomms/fcab083

Hemonnot, A.-L., Hua, J., Ulmann, L. and Hirbec, H. 2019. Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities. *Frontiers in Aging Neuroscience* 11, doi: 10.3389/fnagi.2019.00233

Highet, B., Parker, R., Faull, R. L. M., Curtis, M. A. and Ryan, B. 2021. RNA Quality in Post-mortem Human Brain Tissue Is Affected by Alzheimer's Disease. *Front Mol Neurosci* 14, p. 780352. doi: 10.3389/fnmol.2021.780352

Hodes, R. J. and Buckholtz, N. 2016. Accelerating Medicines Partnership: Alzheimer's Disease (AMP-AD) Knowledge Portal Aids Alzheimer's Drug Discovery through Open Data Sharing. *Expert Opinion on Therapeutic Targets* 20(4), pp. 389-391. doi: 10.1517/14728222.2016.1135132

Holmes, C., Smith, H., Ganderton, R., Arranz, M., Collier, D., Powell, J. and Lovestone, S. 2001. Psychosis and aggression in Alzheimer's disease: the effect of dopamine receptor gene variation. *J Neurol Neurosurg Psychiatry* 71(6), pp. 777-779. doi: 10.1136/jnnp.71.6.777

Horvath, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol* 14(10), p. R115. doi: 10.1186/gb-2013-14-10-r115

Hu, Y. et al. 2019. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics* 51(3), pp. 568-576. doi: 10.1038/s41588-019-0345-7

Huang da, W., Sherman, B. T. and Lempicki, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1), pp. 1-13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T. and Lempicki, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1), pp. 44-57. doi: 10.1038/nprot.2008.211

Iatrou, A., Clark, E. M. and Wang, Y. 2021. Nuclear dynamics and stress responses in Alzheimer's disease. *Molecular neurodegeneration* 16(1), pp. 65-65. doi: 10.1186/s13024-021-00489-6

Jack, C. R., Jr. et al. 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7(3), pp. 257-262. doi: 10.1016/j.jalz.2011.03.004

Jack, C. R., Jr. et al. 2018. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* 14(4), pp. 535-562. doi: 10.1016/j.jalz.2018.02.018

Jacobs, H. I. L., Hopkins, D. A., Mayrhofer, H. C., Bruner, E., van Leeuwen, F. W., Raaijmakers, W. and Schmahmann, J. D. 2017. The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline. *Brain* 141(1), pp. 37-47. doi: 10.1093/brain/awx194

Jaffe, A. E. et al. 2017. qSVA framework for RNA quality correction in differential expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* 114(27), pp. 7130-7135. doi: 10.1073/pnas.1617384114

Jansen, I. E. et al. 2019. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics* 51(3), pp. 404-413. doi: 10.1038/s41588-018-0311-9

Johns, P. 2014. Chapter 5 - Neurons and glial cells. In: Johns, P. ed. *Clinical Neuroscience*. Churchill Livingstone, pp. 61-69.

Jun, G. et al. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91(5), pp. 839-848. doi: 10.1016/j.ajhg.2012.09.004

Kalaora, S. et al. 2020. Immunoproteasome expression is associated with better prognosis and response to checkpoint therapies in melanoma. *Nature Communications* 11(1), p. 896. doi: 10.1038/s41467-020-14639-9

Kanehisa, M. et al. 2008. KEGG for linking genomes to life and the environment. *Nucleic acids research* 36(Database issue), pp. D480-D484. doi: 10.1093/nar/gkm882

Karikari, T. K. et al. 2021. Diagnostic performance and prediction of clinical progression of plasma phospho-tau181 in the Alzheimer's Disease Neuroimaging Initiative. *Molecular Psychiatry* 26(2), pp. 429-442. doi: 10.1038/s41380-020-00923-z

Kasuga, K. et al. 2022. Different AT(N) profiles and clinical progression classified by two different N markers using total tau and neurofilament light chain in cerebrospinal fluid. *BMJ Neurol Open* 4(2), p. e000321. doi: 10.1136/bmjno-2022-000321

Kloetzel, P. M. 2001. Antigen processing by the proteasome. *Nat Rev Mol Cell Biol* 2(3), pp. 179-187. doi: 10.1038/35056572

Knopman, D. S. et al. 2021. Alzheimer disease. *Nature reviews. Disease primers* 7(1), pp. 33-33. doi: 10.1038/s41572-021-00269-y

Knowlton, B. J., Morrison, R. G., Hummel, J. E. and Holyoak, K. J. 2012. A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences* 16(7), pp. 373-381. doi: https://doi.org/10.1016/j.tics.2012.06.002

Kudo, L. C. et al. 2011. Puromycin-sensitive aminopeptidase (PSA/NPEPPS) impedes development of neuropathology in hPSA/TAU(P301L) double-transgenic mice. *Hum Mol Genet* 20(9), pp. 1820-1833. doi: 10.1093/hmg/ddr065

Kunkle, B. W. et al. 2019. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* 51(3), pp. 414-430. doi: 10.1038/s41588-019-0358-2

König, T. and Stögmann, E. 2021. Genetics of Alzheimer's disease. *Wiener Medizinische Wochenschrift* 171(11), pp. 249-256. doi: 10.1007/s10354-021-00819-9

Lambert, J. C. et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45(12), pp. 1452-1458. doi: 10.1038/ng.2802

Lane, C. A., Hardy, J. and Schott, J. M. 2018. Alzheimer's disease. *Eur J Neurol* 25(1), pp. 59-70. doi: 10.1111/ene.13439

Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15(2), p. R29. doi: 10.1186/gb-2014-15-2-r29

Lawrenson, K. et al. 2015. Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. *Nature Communications* 6(1), p. 8234. doi: 10.1038/ncomms9234

Ledesma, R. D., Valero-Mora, P. and Macbeth, G. 2015. The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *The Spanish Journal of Psychology* 18, p. E11. doi: 10.1017/sjp.2015.13

Lee, B., Yao, X. and Shen, L. 2022. Integrative analysis of summary data from GWAS and eQTL studies implicates genes differentially expressed in Alzheimer's disease. *BMC Genomics* 23(Suppl 4), p. 414. doi: 10.1186/s12864-022-08584-8

Leek, J. T. and Storey, J. D. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3(9), pp. 1724-1735. doi: 10.1371/journal.pgen.0030161

Leonenko, G. et al. 2021. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nature Communications* 12(1), p. 4506. doi: 10.1038/s41467-021-24082-z

Lewis, C. M. and Vassos, E. 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* 12(1), p. 44. doi: 10.1186/s13073-020-00742-5

Li, B. and Ritchie, M. D. 2021. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Frontiers in Genetics* 12, doi: 10.3389/fgene.2021.713230

Li, P. et al. 2022a. Identifying Candidate Genes Associated with Sporadic Amyotrophic Lateral Sclerosis via Integrative Analysis of Transcriptome-Wide Association Study and Messenger RNA Expression Profile. *Cellular and Molecular Neurobiology*, doi: 10.1007/s10571-021-01186-0

Li, Y., Ge, X., Peng, F., Li, W. and Li, J. J. 2022b. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology* 23(1), p. 79. doi: 10.1186/s13059-022-02648-4

Li, Y., Laws, S. M., Miles, L. A., Wiley, J. S., Huang, X., Masters, C. L. and Gu, B. J. 2021. Genomics of Alzheimer's disease implicates the innate and adaptive immune systems. *Cellular and Molecular Life Sciences*, doi: 10.1007/s00018-021-03986-5

Lin, Y., Rajamohamedsait, H. B., Sandusky-Beltran, L. A., Gamallo-Lana, B., Mar, A. and Sigurdsson, E. M. 2020. Chronic PD-1 Checkpoint Blockade Does Not Affect Cognition or Promote Tau Clearance in a Tauopathy Mouse Model. *Frontiers in Aging Neuroscience* 11, doi: 10.3389/fnagi.2019.00377

Liu, C.-C., Liu, C.-C., Kanekiyo, T., Xu, H. and Bu, G. 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature reviews. Neurology* 9(2), pp. 106-118. doi: 10.1038/nrneurol.2012.263

Liu, J. Z., Erlich, Y. and Pickrell, J. K. 2017. Case-control association mapping by proxy using family history of disease. *Nat Genet* 49(3), pp. 325-331. doi: 10.1038/ng.3766

Liu, Y.-J., Liu, T.-T., Jiang, L.-H., Liu, Q., Ma, Z.-L., Xia, T.-J. and Gu, X.-P. 2021. Identification of hub genes associated with cognition in the hippocampus of Alzheimer's Disease. *Bioengineered* 12(2), pp. 9598-9609. doi: 10.1080/21655979.2021.1999549

Livingston, G. et al. 2017. Dementia prevention, intervention, and care. *Lancet* 390(10113), pp. 2673-2734. doi: 10.1016/s0140-6736(17)31363-6

Love, M. I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12), p. 550. doi: 10.1186/s13059-014-0550-8

Lowe, V. J. et al. 2018. Widespread brain tau and its association with ageing, Braak stage and Alzheimer's dementia. *Brain* 141(1), pp. 271-287. doi: 10.1093/brain/awx320

Luningham, J. M., Chen, J., Tang, S., De Jager, P. L., Bennett, D. A., Buchman, A. S. and Yang, J. 2020. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am J Hum Genet* 107(4), pp. 714-726. doi: 10.1016/j.ajhg.2020.08.022

Lyketsos, C. G. et al. 2011. Neuropsychiatric symptoms in Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7(5), pp. 532-539. doi: 10.1016/j.jalz.2011.05.2410

Mahase, E. 2021. Aducanumab: European agency rejects Alzheimer’s drug over efficacy and safety concerns. *BMJ* 375, p. n3127. doi: 10.1136/bmj.n3127

Maj, M., Wagner, L. and Tretter, V. 2019. 20 Years of Secretagogin: Exocytosis and Beyond. *Front Mol Neurosci* 12, p. 29.

Makin, S. 2018. The amyloid hypothesis on trial. *Nature* 559(7715), pp. S4-s7. doi: 10.1038/d41586-018-05719-4

Marioni, R. E. et al. 2018. GWAS on family history of Alzheimer's disease. *Transl Psychiatry* 8(1), p. 99.

Marques-Coelho, D. et al. 2021. Differential transcript usage unravels gene expression alterations in Alzheimer’s disease human brains. *npj Aging and Mechanisms of Disease* 7(1), p. 2. doi: 10.1038/s41514-020-00052-5

Mattsson-Carlgren, N. et al. 2020. The implications of different approaches to define AT(N) in Alzheimer disease. *Neurology* 94(21), pp. e2233-e2244. doi: 10.1212/wnl.0000000000009485

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

Mehta, D., Jackson, R., Paul, G., Shi, J. and Sabbagh, M. 2017. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert opinion on investigational drugs* 26(6), pp. 735-739. doi: 10.1080/13543784.2017.1323868

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T. and Thomas, P. D. 2020. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* 49(D1), pp. D394-D403. doi: 10.1093/nar/gkaa1106

Milicic, L. et al. 2022. Comprehensive analysis of epigenetic clocks reveals associations between disproportionate biological ageing and hippocampal volume. *GeroScience* 44(3), pp. 1807-1823. doi: 10.1007/s11357-022-00558-8

Mirra, S. S. 1997. The CERAD neuropathology protocol and consensus recommendations for the postmortem diagnosis of Alzheimer's disease: a commentary. *Neurobiol Aging* 18(4 Suppl), pp. S91-94. doi: 10.1016/s0197-4580(97)00058-4

Mirra, S. S. et al. 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), pp. 479-486. doi: 10.1212/wnl.41.4.479

Monereo-Sánchez, J. et al. 2021. Genetic Overlap Between Alzheimer's Disease and Depression Mapped Onto the Brain. *Frontiers in Neuroscience* 15, doi: 10.3389/fnins.2021.653130

Morris, J. C. 1993. The Clinical Dementia Rating (CDR). *Current version and scoring rules* 43(11), pp. 2412-2412-a. doi: 10.1212/WNL.43.11.2412-a

Naj, A. C. et al. 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet* 43(5), pp. 436-441. doi: 10.1038/ng.801

Ni, J. et al. 2020. Integration of GWAS and eQTL Analysis to Identify Risk Loci and Susceptibility Genes for Gastric Cancer. *Frontiers in Genetics* 11, doi: 10.3389/fgene.2020.00679

Nichols, E. 2022. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* 7(2), pp. e105-e125. doi: 10.1016/s2468-2667(21)00249-8

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. and Cox, N. J. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6(4), p. e1000888. doi: 10.1371/journal.pgen.1000888

Ochocka, N. and Kaminska, B. 2021. Microglia Diversity in Healthy and Diseased Brain: Insights from Single-Cell Omics. *Int J Mol Sci* 22(6), doi: 10.3390/ijms22063027

Panitch, R., Hu, J., Xia, W., Bennett, D. A., Stein, T. D., Farrer, L. A. and Jun, G. R. 2022. Blood and brain transcriptome analysis reveals APOE genotype-mediated and immune-related pathways involved in Alzheimer disease. *Alzheimer's Research & Therapy* 14(1), p. 30. doi: 10.1186/s13195-022-00975-z

Patel, D., Zhang, X., Farrell, J. J., Chung, J., Stein, T. D., Lunetta, K. L. and Farrer, L. A. 2021. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Translational Psychiatry* 11(1), p. 250. doi: 10.1038/s41398-021-01373-z

Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A. and Abel, T. 2015. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res* 43(16), pp. 7664-7674. doi: 10.1093/nar/gkv736

Porcu, E. et al. 2021. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications* 12(1), p. 5647. doi: 10.1038/s41467-021-25805-y

Prince, M. J., Wimo, A., Guerchet, M. M., Ali, G. C., Wu, Y.-T. and Prina, M. 2015. World Alzheimer Report 2015-The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends.

Purcell, S. et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3), pp. 559-575.

Qi, X. et al. 2019. An integrative analysis of transcriptome-wide association study and mRNA expression profile identified candidate genes for attention-deficit/hyperactivity disorder. *Psychiatry Research* 282, p. 112639. doi: https://doi.org/10.1016/j.psychres.2019.112639

Raj, T. et al. 2018. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature genetics* 50(11), pp. 1584-1592. doi: 10.1038/s41588-018-0238-1

Raj, T. et al. 2014. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science (New York, N.Y.)* 344(6183), pp. 519-523. doi: 10.1126/science.1249547

Reijnders, M. J. M. F. and Waterhouse, R. M. 2021. Summary Visualizations of Gene Ontology Terms With GO-Figure! *Frontiers in Bioinformatics* 1, p. 6.

Reimand, J. et al. 2019. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols* 14(2), pp. 482-517. doi: 10.1038/s41596-018-0103-9

Riedel, B. C., Thompson, P. M. and Brinton, R. D. 2016. Age, APOE and sex: Triad of risk of Alzheimer's disease. *The Journal of steroid biochemistry and molecular biology* 160, pp. 134-147. doi: 10.1016/j.jsbmb.2016.03.012

Risso, D., Schwartz, K., Sherlock, G. and Dudoit, S. 2011. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12, p. 480. doi: 10.1186/1471-2105-12-480

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7), p. e47. doi: 10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), pp. 139-140. doi: 10.1093/bioinformatics/btp616

Rosenzweig, N. et al. 2019. PD-1/PD-L1 checkpoint blockade harnesses monocyte-derived macrophages to combat cognitive impairment in a tauopathy mouse model. *Nature Communications* 10(1), p. 465. doi: 10.1038/s41467-019-08352-5

Ryan, J., Fransquet, P., Wrigglesworth, J. and Lacaze, P. 2018. Phenotypic Heterogeneity in Dementia: A Challenge for Epidemiology and Biomarker Studies. *Frontiers in Public Health* 6, doi: 10.3389/fpubh.2018.00181

Scheltens, P. et al. 2021. Alzheimer's disease. *The Lancet* 397(10284), pp. 1577-1590. doi: https://doi.org/10.1016/S0140-6736(20)32205-4

Schneider, J. A., Arvanitakis, Z., Bang, W. and Bennett, D. A. 2007. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology* 69(24), pp. 2197-2204. doi: 10.1212/01.wnl.0000271090.28148.24

Schwartz, M., Arad, M. and Ben-Yehuda, H. 2019. Potential immunotherapy for Alzheimer disease and age-related dementia. *Dialogues in clinical neuroscience* 21(1), pp. 21-25. doi: 10.31887/DCNS.2019.21.1/mschwartz

Schwartzentruber, J. et al. 2021. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics* 53(3), pp. 392-402. doi: 10.1038/s41588-020-00776-w

Schweighauser, M. et al. 2022. Age-dependent formation of TMEM106B amyloid filaments in human brains. *Nature*, doi: 10.1038/s41586-022-04650-z

Shabalin, A. A. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)* 28(10), pp. 1353-1358. doi: 10.1093/bioinformatics/bts163

Sheng, Q. et al. 2016. Multi-perspective quality control of Illumina RNA sequencing data analysis. *Briefings in Functional Genomics* 16(4), pp. 194-204. doi: 10.1093/bfgp/elw035

Shi, M., Chai, Y., Zhang, J. and Chen, X. 2022. Endoplasmic Reticulum Stress-Associated Neuronal Death and Innate Immune Response in Neurological Diseases. *Frontiers in immunology* 12, pp. 794580-794580. doi: 10.3389/fimmu.2021.794580

Shireby, G. L. et al. 2020. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain* 143(12), pp. 3763-3775.

Sieberts, S. K. et al. 2020. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Scientific Data* 7(1), p. 340. doi: 10.1038/s41597-020-00642-8

Silva, M. V. F., Loures, C. d. M. G., Alves, L. C. V., de Souza, L. C., Borges, K. B. G. and Carvalho, M. d. G. 2019. Alzheimer's disease: risk factors and potentially protective measures. *Journal of biomedical science* 26(1), pp. 33-33. doi: 10.1186/s12929-019-0524-y

Sims, R., Hill, M. and Williams, J. 2020. The multiplex model of the genetics of Alzheimer's disease.*Nat Neurosci*. Vol. 23. United States, pp. 311-322.

Soldan, A. et al. 2019. ATN profiles among cognitively normal individuals and longitudinal cognitive outcomes. *Neurology* 92(14), pp. e1567-e1579. doi: 10.1212/wnl.0000000000007248

Sonehara, K. et al. 2021. Genetic architecture of microRNA expression and its link to complex diseases in the Japanese population. *Human Molecular Genetics* 31(11), pp. 1806-1820. doi: 10.1093/hmg/ddab361

Song, H. et al. 2015. Inhibition of glutaminyl cyclase ameliorates amyloid pathology in an animal model of Alzheimer's disease via the modulation of γ-secretase activity.*J Alzheimers Dis*. Vol. 43. Netherlands, pp. 797-807.

Song, Y.-H., Yoon, J. and Lee, S.-H. 2021. The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. *Experimental & Molecular Medicine* 53(3), pp. 328-338. doi: 10.1038/s12276-021-00580-4

Soon, C. S., Brass, M., Heinze, H.-J. and Haynes, J.-D. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11(5), pp. 543-545. doi: 10.1038/nn.2112

Sperling, R. A. et al. 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7(3), pp. 280-292. doi: 10.1016/j.jalz.2011.03.003

Sudlow, C. et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12(3), p. e1001779. doi: 10.1371/journal.pmed.1001779

Sun, Y. et al. 2021. A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes. *Genome Medicine* 13(1), p. 141. doi: 10.1186/s13073-021-00959-y

Szklarczyk, D. et al. 2018. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 47(D1), pp. D607-D613. doi: 10.1093/nar/gky1131

Tao, Y., Han, Y., Yu, L., Wang, Q., Leng, S. X. and Zhang, H. 2020. The Predicted Key Molecules, Functions, and Pathways That Bridge Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). *Frontiers in Neurology* 11, p. 233.

Team, R. Core. 2021. R*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Trapnell, C. 2015. Defining cell types and states with single-cell genomics. *Genome Res* 25(10), pp. 1491-1498. doi: 10.1101/gr.190595.115

Tsang, J. et al. 2015. The relationship between dopamine receptor D1 and cognitive performance. *npj Schizophrenia* 1(1), p. 14002. doi: 10.1038/npjschz.2014.2

Uffelmann, E. et al. 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1(1), p. 59. doi: 10.1038/s43586-021-00056-9

Van Cauwenberghe, C., Van Broeckhoven, C. and Sleegers, K. 2016. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in medicine : official journal of the American College of Medical Genetics* 18(5), pp. 421-430. doi: 10.1038/gim.2015.117

Van Deerlin, V. M. et al. 2010. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat Genet* 42(3), pp. 234-239. doi: 10.1038/ng.536

Vepsäläinen, S., Helisalmi, S., Koivisto, A. M., Tapaninen, T., Hiltunen, M. and Soininen, H. 2007. Somatostatin genetic variants modify the risk for Alzheimer's disease among Finnish patients. *J Neurol* 254(11), pp. 1504-1508. doi: 10.1007/s00415-007-0539-2

Vijayan, D. K. and Zhang, K. Y. J. 2019. Human glutaminyl cyclase: Structure, function, inhibitors and involvement in Alzheimer's disease. *Pharmacological Research* 147, p. 104342. doi: https://doi.org/10.1016/j.phrs.2019.104342

Võsa, U. et al. 2018. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv.

Võsa, U. et al. 2021. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* 53(9), pp. 1300-1310. doi: 10.1038/s41588-021-00913-z

Wainberg, M. et al. 2019. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics* 51(4), pp. 592-599. doi: 10.1038/s41588-019-0385-z

Wan, Y. W. et al. 2020. Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. *Cell Rep* 32(2), p. 107908. doi: 10.1016/j.celrep.2020.107908

Wang, M. et al. 2018. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data* 5, p. 180185. doi: 10.1038/sdata.2018.185

Wang, M. et al. 2016. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer’s disease. *Genome Medicine* 8(1), p. 104. doi: 10.1186/s13073-016-0355-3

Wang, M. et al. 2013. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* 77(4), pp. 736-749. doi: 10.1016/j.neuron.2012.12.032

Warde-Farley, D. et al. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38(Web Server issue), pp. W214-220. doi: 10.1093/nar/gkq537

Watanabe, K., Taskesen, E., van Bochoven, A. and Posthuma, D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* 8(1), pp. 1826-1826. doi: 10.1038/s41467-017-01261-5

Wei, W., Wang, Z. Y., Ma, L. N., Zhang, T. T., Cao, Y. and Li, H. 2020. MicroRNAs in Alzheimer's Disease: Function and Potential Applications as Diagnostic Biomarkers. *Front Mol Neurosci* 13, p. 160. doi: 10.3389/fnmol.2020.00160

Weidling, I. W. and Swerdlow, R. H. 2020. Mitochondria in Alzheimer's disease and their potential role in Alzheimer's proteostasis. *Experimental neurology* 330, pp. 113321-113321. doi: 10.1016/j.expneurol.2020.113321

White, K., Yang, P., Li, L., Farshori, A., Medina, A. E. and Zielke, H. R. 2018. Effect of Postmortem Interval and Years in Storage on RNA Quality of Tissue at a Repository of the NIH NeuroBioBank. *Biopreserv Biobank* 16(2), pp. 148-157. doi: 10.1089/bio.2017.0099

Wightman, D. P. et al. 2021. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics* 53(9), pp. 1276-1282. doi: 10.1038/s41588-021-00921-z

Winblad, B. et al. 2016. Defeating Alzheimer's disease and other dementias: a priority for European science and society.*Lancet Neurol*. Vol. 15. England, pp. 455-532.

Wingo, T. S., Lah, J. J., Levey, A. I. and Cutler, D. J. 2012. Autosomal recessive causes likely in early-onset Alzheimer disease. *Archives of neurology* 69(1), pp. 59-64. doi: 10.1001/archneurol.2011.221

Wishart, H. A. et al. 2006. Increased brain activation during working memory in cognitively intact adults with the APOE epsilon4 allele.*Am J Psychiatry*. Vol. 163. United States, pp. 1603-1610.

Wittenberg, R. et al. 2020. Projections of care for older people with dementia in England: 2015 to 2040. *Age and Ageing* 49(2), pp. 264-269. doi: 10.1093/ageing/afz154

Wu, H., Huang, Q., Yu, Z., Wu, H. and Zhong, Z. 2020. The SNPs rs429358 and rs7412 of APOE gene are association with cerebral infarction but not SNPs rs2306283 and rs4149056 of SLCO1B1 gene in southern Chinese Hakka population. *Lipids in Health and Disease* 19(1), p. 202. doi: 10.1186/s12944-020-01379-4

Wu, Y., Zheng, Z., Visscher, P. M. and Yang, J. 2017. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biology* 18(1), p. 86. doi: 10.1186/s13059-017-1216-0

Xue, S., Jia, L. and Jia, J. 2009. Association between somatostatin gene polymorphisms and sporadic Alzheimer's disease in Chinese population. *Neurosci Lett* 465(2), pp. 181-183. doi: 10.1016/j.neulet.2009.09.002

Yamakaze, J. and Lu, Z. 2021. Deletion of the lactoperoxidase gene causes multisystem inflammation and tumors in mice. *Scientific Reports* 11(1), p. 12429. doi: 10.1038/s41598-021-91745-8

Yang, F., Wang, J., Pierce, B. L. and Chen, L. S. 2017. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res* 27(11), pp. 1859-1871. doi: 10.1101/gr.216754.116

Yang, H.-S. et al. 2020. Genetics of Gene Expression in the Aging Human Brain Reveal TDP-43 Proteinopathy Pathophysiology. *Neuron* 107(3), pp. 496-508.e496. doi: 10.1016/j.neuron.2020.05.010

Zarouchlioti, C., Parfitt, D. A., Li, W., Gittings, L. M. and Cheetham, M. E. 2018. DNAJ Proteins in neurodegeneration: essential and protective factors. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 373(1738), p. 20160534. doi: 10.1098/rstb.2016.0534

Zhang, B. et al. 2013. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 153(3), pp. 707-720. doi: 10.1016/j.cell.2013.03.030

Zhang, J., Sun, M., Zhao, Y., Geng, G. and Hu, Y. 2021. Identification of Gingivitis-Related Genes Across Human Tissues Based on the Summary Mendelian Randomization. *Frontiers in Cell and Developmental Biology* 8, doi: 10.3389/fcell.2020.624766

Zhang, L. et al. 2020a. Integrating transcriptome-wide association study and mRNA expression profiling identified candidate genes and pathways associated with osteomyelitis. *Scand J Rheumatol* 49(2), pp. 131-136. doi: 10.1080/03009742.2019.1653492

Zhang, L. et al. 2020b. Genome-wide analysis of expression quantitative trait loci (eQTLs) reveals the regulatory architecture of gene expression variation in the storage roots of sweet potato. *Horticulture Research* 7(1), p. 90. doi: 10.1038/s41438-020-0314-4

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P. and Wang, L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30(7), pp. 1006-1007. doi: 10.1093/bioinformatics/btt730

Zhao, T., Hu, Y., Zang, T. and Wang, Y. 2019. Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Frontiers in genetics* 10, pp. 1021-1021. doi: 10.3389/fgene.2019.01021

Zhu, Z. et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* 48(5), pp. 481-487. doi: 10.1038/ng.3538

Zou, F. et al. 2010. Gene expression levels as endophenotypes in genome-wide association studies of Alzheimer disease. *Neurology* 74(6), pp. 480-486. doi: 10.1212/WNL.0b013e3181d07654