2022

Statistical Learning for Gene Expression Biomarker Detection in Neurodegenerative Diseases

Kelly, Jack

UNIVERSITY OF
PLYMOUTH

STATISTICAL LEARNING FOR GENE EXPRESSION BIOMARKER DETECTION IN NEURODEGENERATIVE DISEASES

by

Jack Kelly

A thesis submitted to the University of Plymouth in partial fulfillment for the degree of

DOCTOR OF PHILOSOPHY

Peninsula Medical School

January 2022

Acknowledgements

I would like to give my special thanks to the two director of studies I have had during my PhD, Dr. Xinzhong Li and Prof. Shouqing Luo. Dr. Li has guided me through the beginning of my research career and taught me invaluable lessons in approaching my life inside and out of my work. Prof. Luo helped lead me through the final stretch of my thesis and has given me opportunities with research I would have not otherwise had.

I also thank my two other supervisors Dr. Camille Carroll and Dr. Rana Moyeed who used there expertise to guide me through some of the complex methodologies and interpretation of results. I appreciate their helpful insights into my work. It has been a great privilege and honour to work and study under my whole supervisory team.

I extend my thanks to Dr. Robert Belshaw who introduced me to bioinformatics and sparked my lifelong interest. Additionally, I thank Birbal Prasad, Yi Yang, Evelina Valionyte and everyone at the Derriford research facility for their support in and outside of work.

I would like to thank Kate, Abbygail and my mum and dad for their support and encouragement throughout the project and all through my studies. Finally, I would like to thank my friends for giving me support and a space to not think about my project.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was financed with the aid of a studentship from the University of Plymouth.

Publications:

Kelly, J., Moyeed R., Carroll C., Albani D. & Li X., 2019. Gene expression meta-analysis of Parkinson's disease and its relationship with Alzheimer's disease. *Molecular Brain*, 12: 16. DOI: 10.1186/s13041-019-0436-5.

Kelly, J., Moyeed R., Carroll C., Luo S. & Li X., 2020. Genetic networks in Parkinson's and Alzheimer's disease. *Aging*, 12(6): pp 5221-5243. DOI: 10.18632/aging.102943

Word count of main body of thesis: 37036

Signed: 

Date: 29/01/2022

Abstract

Statistical learning for gene expression biomarker detection in neurodegenerative disease

Jack Rees Kelly

In this work, statistical learning approaches are used to detect biomarkers for neurodegenerative diseases (NDs). NDs are becoming increasingly prevalent as populations age, making understanding of disease and identification of biomarkers progressively important for facilitating early diagnosis and the screening of individuals for clinical trials. Advancements in gene expression profiling has enabled the exploration of disease biomarkers at an unprecedented scale. The work presented here demonstrates the value of gene expression data in understanding the underlying processes and detection of biomarkers of NDs. The value of novel approaches to previously collected -omics data is shown and it is demonstrated that new therapeutic targets can be identified. Additionally, the importance of meta-analysis to improve power of multiple small studies is demonstrated. The value of blood transcriptomics data is shown in applications to researching NDs to understand underlying processes using network analysis and a novel hub detection method. Finally, after demonstrating the value of blood gene expression data for investigating NDs, a combination of feature selection and classification algorithms were used to identify novel accurate biomarker signatures for the diagnosis and prognosis of Parkinson's disease (PD) and Alzheimer's disease (AD). Additionally, the use of feature pools based on previous knowledge of disease and the viability of neural networks in dimensionality reduction and biomarker detection is demonstrated and discussed. In summary, gene expression data is shown to be valuable for the investigation of ND and novel gene biomarker signatures for the diagnosis and prognosis of PD and AD.

Contents

List of Figures

List of Tables

Abbreviations

ND ..... Neurodegenerative diseases

AD ..... Alzheimer's disease

PD ..... Parkinson's disease

HD ..... Huntington's disease

MS ..... Multiple sclerosis

ALS ..... Amyotrophic lateral sclerosis

Aβ ..... Amyloid-β

ACHE ..... Acetylcholinesterase

APP ..... Amyloid precursor protein

ICD ..... Intracellular domain

APOE ..... Apolipoprotein E

CSF ..... Cerebrospinal fluid

MRI ..... Magnetic resonance imaging

SPECT ..... Single photon emission computed tomography

PET ..... Positron emission tomography

T-tau ..... Total tau

P-tau ..... Phosphorylated tau

NfL ..... Neurofilament light chain protein

SVM ..... Support vector machine

ROC . . . . . Receiver operating characteristic

AUC . . . . . Area under the curve

LASSO . . . . . Least absolute shrinkage and selection operator

SN . . . . . Substantia nigra

MOA-B . . . . . Monoamine Oxidase Type B

SNCA . . . . . α-synuclein

LRRK2 . . . . . Leucine-rich repeat kinase 2

UCHL1 . . . . . Ubiquitin carboxy-terminal hydrolase L1

DJ-1 . . . . . Deglycase DJ-1

PRKN . . . . . Parkin

PINK-1 . . . . . PTEN-induced kinase 1

FBXO7 . . . . . F-box only protein 7

DAT . . . . . Dopamine transporter imaging

MSA . . . . . Multiple system atrophy

miR . . . . . MicroRNA

HTT . . . . . Huntingtin

mHTT . . . . . Mutant huntington

polyQ . . . . . Polyglutamine

ER . . . . . Endoplasmic reticulum

qEEG . . . . . Quantitative electroencephalography

sTNFR . . . . . Soluble tumour necrosis factor receptors

mRNA . . . . . Messenger RNA

cDNA . . . . . Complementary DNA

RNA-seq . . . . . RNA sequencing

NGS . . . . . Next-generation sequencing

RMA . . . . . Robust Multi-array Average

FPKM . . . . . Fragments per kilobase of exon model per million

RPKM . . . . . Reads per kilobase of exon model per million

TPM . . . . . Transcripts per million

DEGs . . . . . Differentially expressed genes

WGCNA . . . . . Weighted gene co-expression network analysis

TOM . . . . . Topological overlap matrix

LR . . . . . Logistic regression

RBF . . . . . Radial basis function

RF . . . . . Random forest

GBM . . . . . Gradient boosting machines

XGBoost . . . . . Extreme Gradient boosting

ANN . . . . . Artificial neural networks

MLP . . . . . Multi-layered perceptrons

ReLu . . . . . Rectified linear units

VAE . . . . . Variational autoencoder

CNN . . . . . Convolutional neural networks

PCA . . . . . Principal component analysis

t-SNE . . . . . t-Distributed Stochastic Neighbor Embedding

TP . . . . . True Positive

FP . . . . . False Positive

FN . . . . . False Negative

TN . . . . . True Negative

pr . . . . . Precision-recall

ANOVA . . . . . Analysis of variance

RFE . . . . . Recursive feature elimination

VSSRFE . . . . . Variable step size RFE

IHC . . . . . Immunohistochemistry

DAB . . . . . 3,3'-diaminobenzidine

HOX . . . . . Homeobox

BA9 . . . . . Brodmann area 9

ENA . . . . . European Nucleotide Archive

MAD . . . . . Median absolute deviation

PMI . . . . . Post-mortem interval

KNN . . . . . *k*-nearest neighbour

SSE . . . . . Sum of squared errors

RIN . . . . . RNA Integrity Number

IHW . . . . . Independent hypothesis weighting

GO . . . . . Gene ontology

KEGG . . . . . Kyoto encyclopedia of genes and genomes

TFs . . . . . Transcription factors

ENCODE . . . . . Encyclopedia of DNA Elements

ChIP . . . . . Chromatin immunoprecipitation

ChEA . . . . . ChIP Enrichment Analysis

PPIN . . . . . Protein-protein interaction network

HPRD . . . . . Human Protein Reference Database

CTL . . . . . Control

FNN . . . . . First neighbour nodes

miRNAs . . . . . MicroRNAs

NF-κB . . . . . Nuclear factor-κB

MAS5 . . . . . Microarray suite version 5

Q-Q . . . . . Quantile-Quantile

GTEx . . . . . Genotype-Tissue Expression

IPA . . . . . Ingenuity Pathway Analysis

URA . . . . . Upstream regulator analysis

FDR . . . . . False discovery rate

YWHAZ . . . . . 14–3-3 zeta

CJD . . . . . Creutzfeldt-Jakob Disease

ROS . . . . . Reactive oxygen species

REST . . . . . Repressor element 1-silencing transcription factor

ADAD . . . . . AD dataset, AD disease samples

ADHC . . . . . AD dataset, healthy control samples

ADMCI . . . . . AD dataset, mild cognitive impairment samples

PDPD . . . . . PD dataset, PD disease samples

PDHC . . . . . PD dataset, healthy control samples

SNP . . . . . Single nucleotide polymorphism

SCAN . . . . . SNP and Copy number Annotation

HC . . . . . Healthy control

MCI . . . . . Mild cognitive impairment

MM . . . . . Module membership

BC . . . . . Betweenness centrality

GWAS . . . . . Genome Wide Association Studies

PTEN . . . . . Phosphatase and tensin homolog

TRPC . . . . . Transient receptor potential canonical

BBB . . . . . Blood brain barrier

LATE . . . . . Limbic-predominant age-related TDP-43 encephalopathy

CV . . . . . Cross validation

prAUC . . . . . Precision-recall AUC

SGD . . . . . Stochastic gradient descent

Chapter 1

Introduction

1.1 Introduction to neurodegenerative diseases

Neurodegenerative diseases (ND) are a wide range of heterogeneous diseases characterised by a progressive loss of neurons. This results in the deterioration of a wide range of cognitive functions including memory, special cognition, learning, language, and judgment. NDs are incurable and typically lead to years of decline in quality of life and eventual death. The main known risk factor for many NDs is age, and in a world that is predicted to have an increase in over 65s of 120% between 2019 and 2050 [1], effective treatment and diagnosis of ND is needed more than ever.

There are a wide range of NDs which primarily impact different areas of the brain, resulting in a wide variety of symptoms. The most common NDs include Alzheimer's disease (AD), Parkinson's disease (PD) and Huntington's disease (HD). Less common but still important forms of ND include multiple sclerosis (MS) and Amyotrophic lateral sclerosis (ALS).

1.2 Alzheimer's disease

AD is the most common ND and dementia, accounting for 60-80% of dementia cases. AD is characterised pathologically by accumulation of extracellular amyloid-β1 (Aβ) and deposits of intracellular tau neurofibrillary tangles [2]. In the US the number of peo-

ple living with AD is projected to increase from 5.5 million in 2018 to 13.8 million by 2050 [3]. Gradual progressive memory loss is the most common clinical symptom of AD, which eventually affects other cognitive functions such as communication and movement. There are currently many promising advances in the understanding of AD, including discovery of novel biomarkers [4, 5] and analysis of underlying biological mechanisms.

1.2.1 Symptoms of AD

Frequency and intensity of symptoms allow for grouping of AD into three stages; early, middle and late AD. Early AD patients generally continue to live a normal life with little difficulty, and many may forgo seeking a diagnosis because symptoms are so mild. Symptoms at this stage include:

- Misplacing items
- Increased frequency in forgetting object names and recent events
- Increase in poor judgement

The middle stages of AD are when many AD symptoms become clear and more intense. Symptoms at the middle stages of AD include:

- Confusion and disorientation becoming noticeable and effecting day to day life
- Disruption in sleep
- Aphasia (impairment of language)
- Noticeable changes in mood, including depression and anxiety
- Impairment of spatial awareness
- Hallucinations

Late stages of AD have severe symptoms as large sections of brain tissue have died. Patients in the later stages generally need full time care. Including previous stages symptoms, at the late stages of AD symptoms include:

- Severe problems with short- and long-term memory
- Dysphagia (difficulty in swallowing)
- Severe weight loss
- Difficulty in mobility without assistance

- Progressive loss of speech
- Incontinence

1.2.2 Diagnosis and treatment of AD

Diagnosis of AD is based on clinical examination. Patients tend to initially present with memory difficulties and can have a wide range of other symptoms including mood changes, impaired visuospatial abilities and impaired reasoning [6]. Diagnosis of AD can be difficult as early stages of the disease can be mistaken for changes that take place in normal aging, and the patient may disregard symptoms in fear of an AD diagnosis. Misdiagnosis rates of AD range from 12% to 23% in pathologically confirmed studies [7], demonstrating the need for more effective diagnosis procedures.

There is a current lack of effective treatments for AD, and no new therapies have been approved for over 10 years [8]. Treatment of AD generally involves controlling the most severe symptoms as best as possible. Symptom management is tailored to individual patients, for example those that suffer from sleep disturbances can be treated using sleep aids [9]. Currently, Acetylcholinesterase (ACHE) inhibitors are one of the main drug treatments used to manage dementia in AD patients with varied response from patients [10].

ACHE inhibitors work by preventing ACHE from breaking down the neurotransmitter acetylcholine. Acetylcholine is important in the normal functioning of the peripheral and central nervous system, and is reduced in AD brains as neurons die. As ACHE inhibitors prevent the breakdown of acetylcholine, it means more is available to combat the reduction and slow down the progression of some disease symptoms. However, ACHE inhibitors do not work when the disease becomes severe, as the production of acetylcholine is so low that even preventing the breakdown of it does not maintain normal levels [11].

Currently there is a focus on developing new treatments for AD, both to better control individual symptoms and modify the underlying pathophysiology of the disease [8].

1.2.3 Pathophysiology of AD

Aβ and tau are considered the main two contributors to the pathophysiology of AD [12]. Additionally, there are a several causative genes that are associated with AD.

1.2.3.1 Contribution of Aβ to AD

Amyloid plaques are an accumulation of Aβ peptides between nerve cells that form as a result of unusual processing of amyloid precursor protein (APP). The cleavage of APP is shown in figure 1.1. In normal non-amyloidogenic pathways α-secretase cleaves APP creating a long secreted form of APP called sAPPα which has a neuroprotective function [13]. Additionally, a C-terminal fragment C83 is generated, which is then cleaved by γ-secretase to create APP intracellular domain (ICD), which helps regulates neurogenesis [14] and p3. In the healthy brain, the function of p3 is not well understood, however can be found within amyloid plaques in AD patients [15].

Within the amyloidogenic pathway, cleavage by β-secretase generate a shorter secreted APP (sAPPβ) and a longer C-terminal fragment C99. Cleavage of C99 by γ-secretase generate APP ICD and Aβ [16]. Amyloidegenic pathway, in combination with impairment in clearance of Aβ frequently due to mutations in apolipoprotein E (APOE), lead to an increase in Aβ [17]. Aβ aggregates to form amyloid plaques in the brain, which disrupt the function of synapses and induce neurotoxicity [18]. Length of Aβ can also play a role in AD. Soluble Aβ40 is present in healthy blood plasma and cerebrospinal fluid (CSF), however a combination of Aβ40 and Aβ42 are present in amyloid plaques [19].

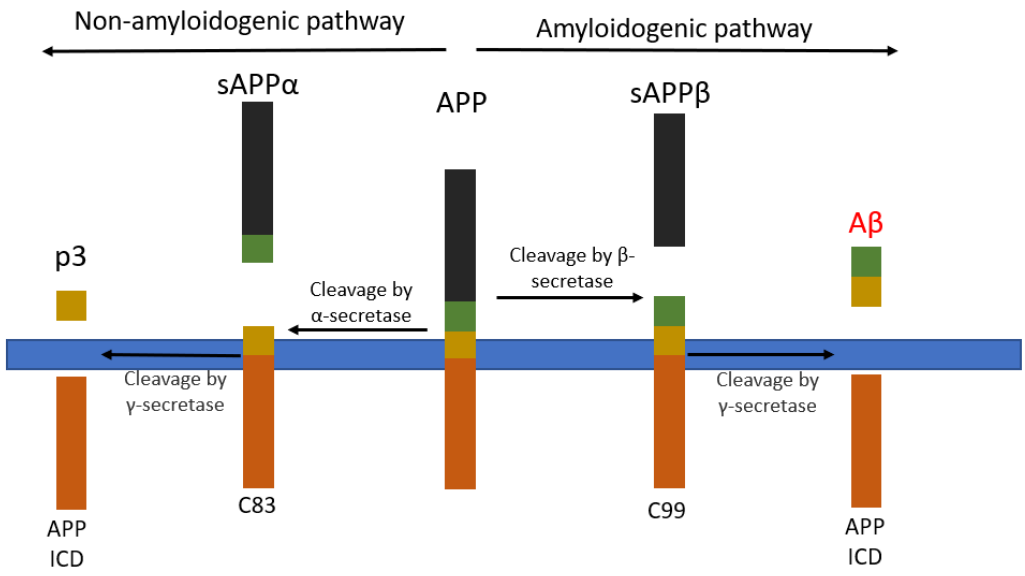


Figure 1.1: **Cleavage of APP to produce Aβ.** In the normal non-amyloidogenic pathway, APP is cleaved by α-secretase and γ-secretase to generate long form sAPPα, C83, p3 and APP ICD. In the amyloidogenic pathway present in AD, β-secretase and γ-secretase cleave APP to generate shorter sAPPβ, C99, ICD and Aβ, which is the main component of amyloid plaques. Adapted from [20]

1.2.3.2 Contribution of Tau to AD

Neurofibrillary tangles are the accumulation of abnormal hyperphosphorylated tau neurons [21]. In healthy neurons, tau stabilised microtubules assist in transporting molecules around the cell, however, abnormal tau instead attaches to one another instead of microtubules which forms thread like tangles which disrupt normal neuron functioning. This tau present in AD brains is abnormal as it is hyperphosphorylated. Hyperphosphlyation of tau can be as a result of increased Aβ, impaired brain glucose metabolism and dysregulation of processes causing phosphorylation [22].

Amyloid plaques initially seed the formation of neurofibrillary tangles, and thereafter toxic tau increases the toxicity of Aβ accumulation [21]. This suggests amyloid plaques and neurofibrillary tangles self-propagate one another in a feedback loop that leads to dysfunction of neurons and ultimately cell death.

1.2.3.3 Causative genes associated with AD

There are a number of causative genes associated with autosomal dominant AD (*APP, PSEN1*, and *PSEN2*) [23]. Around 10-15% of early-onset familial AD patients are accounted for by APP mutations [23], generally occurring in or around the Aβ peptide. This mutation contributes to the abnormal cleavage of APP leading to increased amyloid plaques. Mutations in PSEN1 lead to very severe forms of AD that can onset as early as 30 year old with almost complete penetrance [24]. PSEN2 mutations have reduced penetrance compared to PSEN1 mutations and a higher age of onset (45–88 years) [23]. Both PSEN1 and 2 are components of γ-secretase, which is responsible for cleavage of Aβ so mutations in these two genes can lead to an imbalance in Aβ42 to Aβ40 ratio. Additionally, individuals carrying the APOE ε4 allele have an increased risk for AD [25]. APOE has a role in clearing Aβ in the brain, so it is likely that the ε4 allele is less efficient at this than the common ε3 allele [26].

1.2.4 Biomarkers for AD

A biomarker is an objective measure that can be used to diagnose or track the progress of the disease over time. There are particular characteristics that make an ideal biomarker. Non-invasive, cheap and fast biomarkers would make them more widely accessible to be used. High sensitivity and specificity would reduce misdiagnosis and early detection would allow for quick treatment to improve effectiveness of available treatments. Additionally, biomarkers that give prognostic information are fantastic for understanding the progression of a disease and the effectiveness of treatments. Very few, if any, biomarkers meet these criteria.

There are currently no effective biomarkers for diagnosis for many NDs, which require complex and difficult diagnosis processes. Poor diagnosis leads to many patients not being diagnosed until the disease symptoms have manifested, by which time neuron cell death is extensive and treatment effectiveness is reduced. In addition, misdiagnosis of ND can be common, for example 17% of vascular dementia patients are misdiagnosed with Alzheimer's disease (AD) [7], which can cost the patient time of proper treatment and the health system substantial money.

1.2.4.1 Imaging biomarkers of AD

Imaging is un-invasive approach to biomarker detection, with magnetic resonance imaging (MRI), single photon emission computed tomography (SPECT) and positron emission tomography (PET) all having been investigated for potential imaging biomarkers in ND previously. In a clinical setting, neuroimaging has become a standard tool used by specialists with ND patients to assist with diagnosis and to understand the progression of disease in an individual, however identifying an imaging biomarker would improve accuracy and speed of diagnosis and reduce need for specialist personnel.

Reduction in the hippocampus volume detected using MRI is one of the primary biomarkers for AD [27]. However, other dementias exhibit the same loss in hippocampus volume, including frontotemporal dementia [28] and Lewy body dementias [29]. As a result of this, hippocampus volume is used by medical professionals as one of many tools to diagnose AD, if used at all. Whole brain, white matter, grey matter and cortex volume have been investigated as prognostic markers of AD under the assumption that volume of tissue reflects number of neurons. It is known that neuron number reduces in brain tissue impacted by AD, however it has been shown that this is a result of dementia and not linked to specific pathophysiology of AD [30], meaning these approaches to neuroimaging are not the best to identify a specific AD biomarker.

Several studies have investigated Aβ accumulation in the brain using neuroimaging. Although neurodegeneration occurs in AD without Aβ accumulation, the presence of increased Aβ does lead to worse cognition [31]. Elevated Aβ detected by PET is associated with worse cognition and changes in daily function, even in patients who possess the APOE ε4 allele and have not been diagnosed with AD, highlighting its potential as a very early biomarker for cognitive decline and AD [32]. Recent PET studies investigating tau have shown that atrophy and deposition of tau are correlated [33]. Additionally, tau PET could effectively predict brain atrophy in the later stages of AD, performing better than Aβ PET [34]. These results suggest that tau PET may be suited to predicting the progression of AD and play an important role in future clinical trials and investigations into disease therapies. However, testing of Aβ and tau PET in much larger cohorts is needed before they can be effectively used in the clinic.

1.2.4.2 Cerebrospinal fluid biomarkers of AD

CSF biomarkers have been the most targeted to date in ND. They better represent the neurological and pathological changes in the central nervous system (CNS) than other peripheral tissues due to being in direct contact with the brain.

Currently, CSF levels of total tau (T-tau), phosphorylated tau (P-tau) and Aβ42 (including Aβ42/40 ratio [35]) are the most evaluated biomarkers for AD, having been evaluated with consistent result in hundreds of clinical studies [36]. However, these biomarkers still have a way to go before they can be consistently used as a tool for diagnosis as reference and cut-off values for diagnosis need to be established [36]. Studies have consistently identified high concordance between Aβ42 CSF levels and brain amyloid plaques detected by PET [37]. More recent studies have actually shown Aβ42/40 ratio to be a more accurate CSF biomarker for AD patients [35]. Although the relationship between Aβ42 and Aβ40 is unclear, it is likely the ratio is more successful as Aβ40 acts as a proxy for total Aβ, and so effectively normalises the Aβ42 levels across patients. CSF T-tau and P-tau have been proposed as a marker of ND in AD [38]. The ratio of P-tau and Aβ42 measurements in AD has been shown to be particularly successful in diagnosis [39], and helps distinguish AD diagnosis from other similar NDs [40].

1.2.4.3 Blood biomarkers of AD

CSF biomarkers have been the most targeted to date in ND, as they better represent the neurological and pathological changes in the CNS than other peripheral tissue, however CSF can be difficult to access requiring specialist training and sampling is commonly regarded as a minor surgical procedure. Blood, however, is cheap and un-invasive to sample, though tends to be less accurate [41]. Many blood biomarker studies have focused on Aβ40 and Aβ42 levels. Decrease in blood plasma levels of Aβ40/42 and a correlation between the Aβ42/40 ratio levels in plasma and CSF have been shown in AD [42]. However, this has only been shown consistently using highly sensitive assay, so as these assays become cheaper and more sensitive techniques are developed Aβ plasma biomarkers could become more dependable. Neurofilament light chain protein (NfL) concentrations in blood has been proposed as a consistent biomarker for many NDs [43, 44].

NfL is a cytoplasmic protein expressed in axons and is released into CSF and blood as a result of axonal damage [45]. Plasma NfL change is associated with cognitive decline [46], however plasma NfL levels are not related to levels of tau or Aβ in the brain detected using PET imaging [43, 46]. It is clear that plasma NfL levels are associated with AD, however more works need to be done to understand this relationship and how it can be used as a biomarker for neurodegeneration.

There is a poor correlation between CSF T-tau levels and plasma T-tau [47], likely explaining the lack of consistent results regarding the use of T-tau as a plasma biomarker for AD. P-tau181 has potential as a diagnostic and prognostic blood biomarker of AD. Plasma P-tau181 can effectively classify AD against other NDs [44], correlates with tau and Aβ levels in the brain detected by PET [43, 46], and is associated with severity of AD symptoms [48].

Blood gene expression levels have been proposed as potential biomarkers of AD. As transcriptomics data measures the expression of thousands of genes, and approaches become more accessible, a panel of genes where the expression can be objectively measured to classify control and disease is potentially more obtainable than measurement of individual proteins or molecules. Gene expression can measure changes that directly lead to disease or the response to a disease. When used for biomarker detection, this distinction becomes less important than when using gene expression to investigate the underlying cause of disease, as it is just of interest if these changes can be used for diagnosis.

Blood gene expression analysis is generally performed in whole blood samples, and so contains a large variety of cells including B-cells, T-cells, lymphocytes and granulocytes [49]. Using whole blood data reduces the number of steps that can artificially alter gene expression levels and allows the measurements to include information even from relatively rare cell types including dendritic cells and eosinophils. However, whole blood gene expression can be impacted by cellular composition of an individuals blood at the time of drawing blood. As blood is not the main area of effect of NDs, accuracy and specificity have previously been a problem with blood gene expression biomarkers, however as dataset sample sizes increase and transcriptomics becomes more accessible, this problem should be reduced [50].

As the number of genes included in a transcriptomics dataset is very large, machine learning is often employed to identify biomarkers of disease. Using the known disease state of a sample, the data can be used to train a machine learning algorithm to then classify new unseen samples into disease and control. Reducing the dimensionality of data to reduce the complexity and computational time by selecting the most important features [51] is frequently performed.

Long *et al.* [5] used a novel feature selection approach of support vector machine (SVM) forward selection followed by classification using SVM. They identified a panel of two genes (*ECH1* and *ERBB2*) which returned a receiver operating characteristic (ROC) area under the curve (AUC) of 89.5%. This model, however, was trained on a small dataset, comprised of only 30 AD and 30 control samples. The team addressed this in a later paper which used a much larger sample size of 143 AD patients and 104 controls [52]. Here, they used least absolute shrinkage and selection operator (LASSO) feature selection to identify a panel of four genes (*NDUFA1, MRPL51, RPL36AL* and *RPL36AL*) which can classify AD from control patients with a ROC AUC of 0.87% using a SVM classifier. Although the AUC was lower, the result had a greater power due to the much larger sample size. This small set of features has the potential as a diagnosis panel of AD if validated in the future. Testing multiple different feature selection and classification algorithms may potentially improve these results in the future.

1.3 Parkinson's disease

PD is the second most prevalent ND effecting approximately 145,000 people in the UK [53]. It is predicted that with an aging population the number of PD patients in the UK will increase 18.1% between 2015 and 2065 [53] and in the US PD cases will increase to 1,238,000 from 680,000 by 2030 [54]. PD is characterised by tremor and bradykinesia [55], primarily effecting motor systems of the CNS as a result of the death of dopamine generating cells in the substantia nigra (SN) in the midbrain [56]. Non-motor symptoms are also common in PD [57] including hyposmia, sleep disorders, depression and constipation and as many as 30% of PD patients go on to develop dementia [58].

1.3.1 Symptoms of PD

PD symptoms have a progressive onset that begin very mildly. Patients symptoms are varied between individuals, as is the speed and severity of the onset. Parkinson's is characterised by three main symptoms:

- Tremors - uncontrollable movement of muscles
- Bradykinesia – slowness of movement
- Rigidity – stiffness of muscles

In addition to these main symptoms there are various physical symptoms that PD patients can develop, including:

- Olfactory loss (loss of sense of smell and taste)
- Incontinence
- Erectile dysfunction in men and sexual dysfunction in women
- Hyperhidrosis (excessive sweating)
- Dysphagia (difficulty in swallowing)
- Insomnia
- Balance problems that make a person more prone to injury from falling

Cognitive symptoms can also develop, including:

- Dementia
- Depression and anxiety

1.3.2 Diagnosis and treatment of PD

Diagnosis of PD is initially based on the presence of motor symptoms, primarily bradykinesia and either resting tremor or rigidity [59]. This is supported by presence of other symptoms, especially olfactory loss or a response to PD treatments. Diagnosis of PD is difficult as it relies on subtle details to separate it from other parkinsonism's, including Vascular parkinsonism. In fact, it has been shown that up to 20% of PD diagnosed patients have a different diagnosis at autopsy, and the rate of correct clinical diagnosis has not recently improved [60].

PD has some effective treatments that can mitigate symptoms, unlike most other ND [55]. Levodopa, dopamine agonists and MAO-B inhibitors are all commonly used to successfully alleviate motor symptoms which greatly improves the quality of life in many patients. Levodopa is the precursor to dopamine that can, unlike dopamine, pass through the blood-brain barrier [61]. This allows it to be used to combat the low levels of dopamine in PD brains. To reduce symptoms caused by conversion of Levodopa to dopamine outside of the CNS, a DOPA decarboxylase inhibitor is often administered along with Levodopa in PD patients [62]. Although Levodopa is the most effective treatment in most PD patients, its effects are often reduced in those that take it for the long-term. Natural levels of dopamine get so low in the brain that the administering of the drug cannot make up for the low levels, leading to re-emerging symptoms. Additionally, there is some evidence that DNA methylation changes that occur when taking Levodopa reduce the sensitivity of striatial neurons to the drug [63].

Early in PD, dopamine agonists are often used to delay or reduce the dosage of Levodopa treatment. Dopamine agonists bind to dopamine receptors to reduce symptoms caused by a reduction in dopamine available to neurons. Monoamine Oxidase Type B (MAO-B) is an enzyme that breaks down neurotransmitters, including dopamine [64]. MAO-B inhibitors are used to treat moderate symptoms of PD early, and are often used in combination with Levodopa treatment to increase the time that Levodopa treatment is successful [65]. However, these treatments are not effective in some patients and only work to treat symptoms, having no disease-modifying effect [66].

1.3.3 Pathophysiology of PD

The primary neuropathological hallmark of PD is α-synuclein accumulation in neurons, in the form of Lewy bodies [56]. α-synuclein is involved in normal synapse activity by regulating many processes, including release of neurotransmitter and vesicle docking [67]. However, its full physiological function is not yet understood. α-synuclein has three domains; a N-terminal lipid-binding alpha-helix, a non-amyloid-component and an acidic C-terminal tail [68]. α-synuclein has an increased tendency to aggregate when phosphorylated at Serine 129 in the C-terminal domain, and this phosphorylation is responsible

for the aggregation present in the brains of patients with PD and other synucleinopathies [69].

Lewy bodies are abnormal aggregations of proteins that develop inside nerve cells and interfere with normal functioning of the cell. They are visible under a microscope, so their presence can be identified through histopathology. Although α-synuclein is one of the major constituents of Lewy bodies, they are composed of a heterogeneous blend of over 90 molecules [70]. These include gene products of *SNCA*(α-synuclein), *DJ-1*, *LRRK2*, *parkin*, and *PINK-1* and proteins and molecules associated with mitochondria and ubiquitin-proteasome pathways. The relationship between Lewy bodies and neuronal cell death is complex, with some evidence suggesting that Lewy bodies are cytotoxic [71] and contrary evidence that nonfibrillar α-synuclein is cytotoxic and that fibrillar Lewy bodies aggregates of α-synuclein may actually be a cytoprotective mechanism in PD [70].

There are number of genes in which mutations can carry a significant risk for PD. α-synuclein (*SNCA*), leucine-rich repeat kinase 2 (*LRRK2*) and Ubiquitin carboxy-terminal hydrolase L1 (*UCHL1*) are linked to autosomal dominant PD and mutations in proteins deglycase DJ-1 (*DJ-1*), parkin (*PRKN*), PTEN-induced kinase 1 (*PINK-1*), and F-box only protein 7 (*FBXO7*) are linked to autosomal recessive Parkinsonism [72].

1.3.4 Biomarkers for PD

Biomarkers for PD are critically needed to ensure early diagnosis and treatment of those with PD but are currently lacking. The most well researched biomarkers for PD are α-synuclein and DJ-1 [73].

1.3.4.1 Imaging biomarkers of PD

As PD is characterised by loss of dopaminergic neurons, neuroimaging of the dopaminergic system in the brain has a strong potential for diagnosis and progressive biomarkers. The large neuropathological and disease presentation overlap between PD and other Parkinsonisms make differential diagnosis using neuroimaging difficult. Many imaging biomarkers for PD have low reproducibility and studies are performed in small patient cohorts giving inconsistent results [74].

SPECT has been used with dopamine transporter imaging (DAT) in the past to detect dopaminergic denervation in PD brains [75] and extrastriatal DAT uptake is correlated with the severity of motor symptoms [76]. MRI studies have shown that measuring volumetric differences in the SN between control and PD patients is very inconsistent, with some studies noting no differences [77, 78], others showing loss of volume [79] and some even a gain in volume [80].

1.3.4.2 CSF biomarkers of PD

α-synuclein is the most consistent CSF PD marker, being shown in many studies to have 61-94% sensitivity and 25-64% specificity at classifying PD from control patients [81], however there is little evidence that it is a useful biomarker for disease severity. DJ-1 has been proposed as a potential biomarker for PD due to its link to autosomal recessive Parkinsonism [72]. CSF levels of DJ-1 are higher in early PD than controls and late stage PD patients, making it a prominent potential biomarker for those who are exhibiting mild PD symptoms to confirm diagnosis [82].

Deposits of Aβ42 [83] and tau [84] have been shown to be present in PD brains. As they have shown promise in AD, CSF levels of Aβ42 and tau have been tested in PD patients [85]. Levels of tau were not significantly different in PD CSF, however CSF Aβ42 levels were progressively reduced as PD cognitive symptoms increased, which in combination with AD studies [86] suggests CSF Aβ42 is associated with increased deposition of Aβ in the brain.

1.3.4.3 Blood biomarkers of PD

Studies investigating α-synuclein in blood plasma are very conflicting, with some studies identifying α-synuclein to be reduced in PD patients [87, 88], some identifying it as being increased [89, 90] and others showing no difference between PD and controls [91, 92, 93]. Over 99% of blood α-synuclein is present in red blood cells [94] and so even residual levels in samples could affect the levels of α-synuclein in plasma, likely explaining the large variance in study results. α-synuclein levels in red blood cells has been proposed as a biomarker for PD in the past, however studies have had contradictory results [81], likely

due to variations in red blood cell count between individuals.

Much like in CSF, blood DJ-1 levels have been proposed as prognostic biomarkers for PD. However, in blood DJ-1 levels are higher in later stages of the disease [82]. DJ-1 protein levels have been shown to be potential biomarkers in the past [95], however further investigation has shown that DJ-1 levels in PD and controls are not significantly different [96, 97, 98].

Additionally, levels of circulating cell-free microRNA (miR) 124 in blood has been proposed as a potential biomarker for PD. miR-124 is one of the most abundantly expressed miRs in the brain and it is involved in neurogenesis, synapse morphology, neurotransmission, inflammation, autophagy and mitochondrial function [99]. As miR-124 is so abundant in brain tissue and involved in many processes that are dysregulated in PD, it has been investigated as a biomarker. miR-124 has a neuroprotective role in PD, and can distinguish PD patients from controls, though sample sizes in studies have been low and it is likely it would not be able to identify PD patients from other NDs, including multiple system atrophy (MSA) [99].

1.4 Huntington's disease

HD is a neurodegenerative disease caused by autosomal dominant inheritance of CAG trinucleotide repeat expansion within the huntingtin (HTT) gene on chromosome 4, characterised by movement and cognitive dysfunction [100]. CAG expansion results in mutant huntington (mHTT) protein being produced with a long polyglutamine (polyQ) repeat [101], and genetic and transgenic data suggest that the mutation in HTT causes the disease predominantly by gain-of-function mechanisms. If a person has greater than 39 CAG repeats, they are certain to develop HD, and less than 36 repeats they will not develop the disease. Reduced penetrance is seen between 36 and 39 repeats and a person may or may not develop the disease. The age of onset for HD is inversely correlated to the CAG repeat length, occurring on average at age 45 [102]. The prevalence of HD in Western populations is around 10.6-13.7 per 100,000 [103], however the prevalence in East Asia is substantially lower, only effecting 1-7 per million individuals.

1.4.1 Symptoms of HD

The symptoms of HD vary between patients but include motor, psychological and cognitive symptoms that develop progressively. Early symptoms of HD include:

- Lapses in memory
- Depression and mood disorders
- Moments of uncoordinated movement and stumbling
- Difficulty concentrating

Later symptoms include:

- Chorea (involuntary and irregular muscle movement)
- Difficulty in clear speaking
- Problems in swallowing that can lead to infections
- Changes in personality
- Loss of mobility

1.4.2 Diagnosis and treatment of HD

Around 40% of HD patients exhibit cognitive and behavioural changes up to 15 years before the development of motor symptoms [104]. HD can be diagnosed using genetic tests, usually because of developing symptoms or family history of disease. The monogenic nature and penetrance of HD, however, makes it one of the most treatable neurodegenerative diseases. Targeting chorea using medication that reduces dopaminergic neurotransmission has been successful [105] and treatments for psychiatric symptoms including anxiety and depression are available. However, there are no current treatments that reduce many other symptoms, and there has been little success in identifying disease modifying therapies. Lowering levels of mHTT has been shown to ameliorate mHTT toxicity in disease models [106, 107] and in mouse models CRISPR/Cas9-mediated gene editing can ameliorate neurotoxicity [108]. However, genome editing and lowering mHTT levels in patients' brain is extremely difficult and to date has been unsuccessful.

1.4.3 Pathophysiology of HD

In HD, mHTT neurotoxicity occurs in various brain regions, but striatal medium spiny neurons and cortical neurons undergo the greatest degeneration [109], however other regions of the brain are also effected, including SN, hippocampus, lateral tuberal nuclei of the hypothalamus and parts of the thalamus [110]. mHTT is believed to induce neurodegeneration through its aggregation, which leads to neurotoxicity through endoplasmic reticulum (ER) stress [111], perturbation of Ca2+ signaling, inhibition of protein clearance pathways and alterations of gene transcription [112].

The function of normal HTT is not fully understood. HTT is known, however, to be associated with the cytoplasmic surface of a number of organelles (including mitochondria, microtubules, transport vesicles and synaptic vesicles) [100] suggesting that the most important changes that occur in HD may be dysregulation in normal cellular interaction.

1.4.4 Biomarkers for HD

As HD can be reliably diagnosed using genetic testing and family history of disease, a particular importance for HD biomarkers is identification of disease progression to evaluate therapies [113]. Biomarkers of mHTT levels in the brain are also important as they can act as an assessment of therapies which aim to lower brain mHTT levels.

1.4.4.1 Imaging biomarkers for HD

Neuroimaging biomarkers have shown promise in HD [114]. HD gene carriers and controls can be classified by measuring striatum and white matter atrophy using volumetric MRI, even before patients become symptomatic [115, 116, 117]. This change has been confirmed in longitudinal studies [118, 119, 120]. Fluorodeoxyglucose PET has identified a reduction in glucose metabolism prior to the onset of disease and over time [121]. Pilot studies have shown that quantitative electroencephalography (qEEG) can classify HD from control patients with a sensitivity and specificity of 83% [122].

mHTT has a large toxic effect on striatal medium spiny neurons which highly express PDE10A, and PDE10A has a key role in cAMP/cGMP signalling regulation that promotes

neuronal survival [123]. This has made it a potential target biomarker and PET has been used to investigate reduction of PDE10A as an imaging biomarker [124]. This appears to be particularly noticeable in the caudate, striatum and globus pallidus of the brain [125]. More recent studies suggest PDE10A progressively decrease across all stages of HD [114]. No one single neuroimaging target can be used as a HD biomarker, but it is likely that in the future a combination of different imaging targets will allow for good identification of pathogenesis.

1.4.4.2 CSF biomarkers for HD

Many investigations into CSF biomarkers begin with proteins or markers that would be expected to be dysregulated with previous knowledge of disease pathophysiology and biological pathways. In HD research, many of these expected markers were found to be poor CSF biomarkers. Neurogranin, a marker of postsynaptic damage and TREM2, a marker of microglial function, are altered in HD CSF despite each being important in HD pathogenesis [126].

CSF levels of mHTT were difficult to study before 2015, when the first quantification of soluble mHTT levels in CSF was performed [127]. mHTT has consistently been shown to be present in HD patients CSF since then using more sensitive approaches to mHTT detection [128]. There is no shortage of potential CSF biomarkers of HD, however sensitivity of assays to quantify levels of biomarkers is increasing over time, particularly in the past few years. With more therapeutic approaches entering clinical trials the need for an accurate CSF biomarker is increasing along with the ability to detect them. For these reasons, although currently there is not an accurate biomarker, it is likely that CSF biomarkers will be identified in the future.

1.4.4.3 Blood biomarkers for HD

Measurement of mHTT in circulating monocytes, T cells and B cells has been shown to be a successful biomarker for brain levels of mHTT [129]. NfL is a potential prognostic blood plasma marker for HD [130]. mHTT carriers have significantly higher Nfl concentration than controls, and concentration increases from each disease stage to the next. Nfl

concentrations are also correlated with cognitive decline.

As both innate and adaptive immunity are important in HD pathology [131], measuring peripheral immune response has been suggested as a potential biomarker. Proteomics of HD blood plasma has identified elevated proinflammatory cytokine IL-6 [132], soluble tumour necrosis factor receptors (sTNFR) [133] and neopterin [133]. No follow-up studies in early stages of HD have been performed on immune markers as HD biomarkers.

1.5 Project aims

This thesis aims to investigate approaches to gene expression data to elucidate underlying processes of NDs and use statistical learning to identify potential biomarkers. One goal of this work is to identify novel genes and processes that are important in the underlying pathogenesis of AD, PD and HD. Another goal is to identify disease biomarkers to predict disease in patients.

The aims of this thesis are completed in the following order: Introduce the background of NDs, in particular AD, PD and HD, and discuss previous gene expression research done (chapter 1) and introduce the methodologies that will be used within this thesis (chapter 2). Then I will discuss applications of differential expression analysis to RNA-seq data to identify novel HD associated genes (chapter 3), and using a meta-analysis approach to improve on individual transcriptomics datasets in PD microarray data and compare results to AD (chapter 4). Univariate methods like differential expression analysis of gene expression data are great at identifying individual gene changes, however these methods are expanded on using a gene co-expression network analysis approach to have a more systemic view of disease including novel approaches to identifying key hub genes in disease networks (chapter 5). This network analysis approach is used on AD and PD blood data to compare the two and identify similarities and differences in gene expression. Statistical learning is then applied to the same AD and PD blood gene expression datasets to identify a model that can be used as a biomarker to categorise disease from healthy cohorts (chapter 6). Finally, the contribution to the literature is summarised and future work and questions that it creates are discussed (chapter 7).

Chapter 2

Methodologies

2.1 Abstract

In this chapter the methodologies used throughout this work are introduced and discussed. The two main approaches to investigating the genome, DNA microarray and RNA-sequencing, are described. Differential expression analysis, meta-analysis and network analysis are all important ways to explore this data and are introduced. Additionally, cutting edge statistical learning algorithms are described, as well as their applications to gene expression biomarker detection.

2.2 Introduction to gene expression data

Gene expression profiling has become extremely widely used as it has become cheaper and more accessible over time. It allows for biologists to study and monitor genome wide expression levels of genes. There are two main approaches to investigating the activity of the genome: DNA microarray and RNA-sequencing.

2.2.1 DNA Microarray

Microarray technology is the most widely used way to investigate gene expression due to its high accessibility and relatively low cost. The first whole-genome microarray study was performed in 1997 by placing the whole yeast genome on a microarray [134]. Figure

2.1 shows the principles of microarray techniques. Most commonly they are used to investigate the gene expression changes between two conditions or cohorts to understand the difference between the two. In studies of NDs, for example, cells from patients with a ND are compared to cells from controls to understand the differences in gene expression that are present in disease.

A microarray platform consists of DNA molecules attached onto a solid surface at specific known positions called spots [135]. A microarray contains thousands of spots each of which correspond to a predetermined gene or DNA sequence and each spot contains millions of a known DNA sequence (known as probes). To perform single-channel microarray analysis messenger RNA (mRNA) are collected from the sample being investigated and converted to complementary DNA (cDNA) and fluorescently labelled. These labelled cDNA are then hybridised to the microarray platform and any cDNA that is not hybridised washed away.

Hybridisation is the pairing of complementary nucleic acid sequences to one another forming tight non-covalent bonding between the two. High levels of hydridisation of cDNA from samples to microarray probes generates a greater fluorescence signal at a spot when excited by a laser. Fluorescence can be quantified and intensity at each spot can be compared across conditions to identify any changes in gene expression.

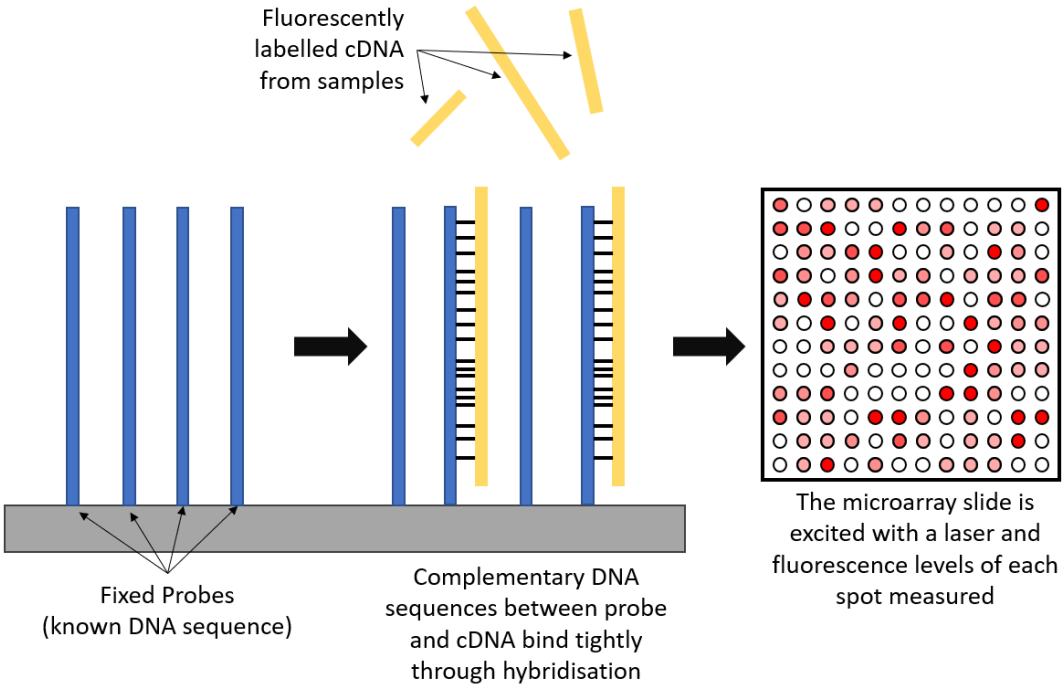


Figure 2.1: **Principles of single-channel microarray analysis.** Each DNA microarray contains 10s of thousands of spots which all contain millions of copies of the same known fixed DNA molecules (probes). Fluorescently labelled cDNA from samples hybridise to complementary probes. The microarray slide is excited with a laser and flourescence at each spot is quantified.

2.2.1.1 Advantages of Microarray

DNA microarray techniques offer multiple advantages [135]:

- **User-friendly**: Microarrays are user-friendly so can be performed by many labs with limited training
- **Speed**: As arrays have already sequenced DNA probes, no large-scale DNA sequencing is required which means microarrays are fast.
- **High coverage**: Well sequenced genomes, such as the human genome, have high coverage using microarray analysis.
- **Low cost**: As the technology has developed the cost of running microarrays has reduced, making them more accessible.

2.2.1.2 Limitations of Microarray Techniques

Although they are widely used and have been important in the study of genetics for over 20 years, they still have some limitations [135]:

- **Not quantifiable between arrays**: Microarray data expression values are only relative to other signals in the same array so directly comparing between arrays is difficult
- **Technical noise**: Although microarrays have greatly improved since they started to be widely used they still yield noisy outputs. They generally require a large sample size to reduce the impact of this noise on results.
- **Limited to predetermined transcripts**: Microarrays only measure known gene sequences which limits analysis. This can miss novel transcripts and potentially important changes like single nucleotide variants.
- **Low dynamic range**: Microarrays are limited by background noise at low expression levels and signal saturation at high expression levels.

2.2.2 RNA-sequencing

RNA sequencing (RNA-seq) measures gene expression via RNA using next-generation sequencing (NGS). Figure 2.2 shows the principles of RNA-seq analysis. Briefly, RNA is extracted from samples as is done with microarray analysis, however this RNA is fragmented before being converting to cDNA. These fragments are sequenced using NGS which has high throughput, scalability and speed compared to previous DNA sequencing technologies [136]. Each gene in the human genome is sequenced multiple times so results have high depth and any DNA variants can be detected with high accuracy. These sequenced fragments are aligned to the reference genome or transcriptome of the organism the sample is from. The number of overlapping reads is the 'count', a quantification of gene expression.

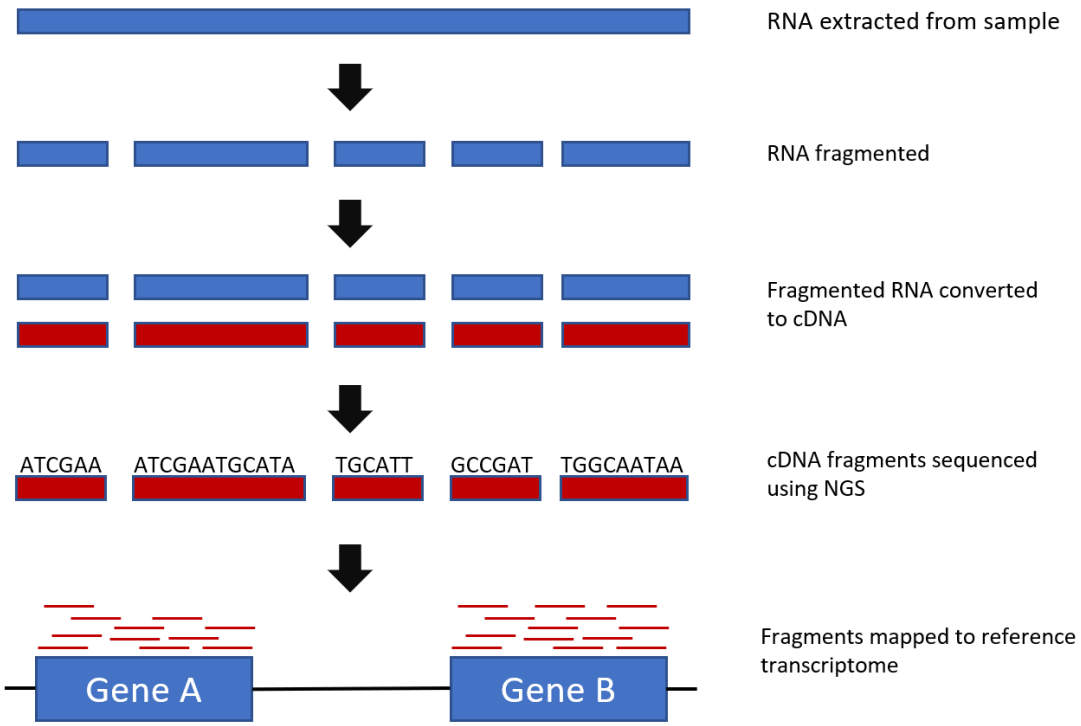


Figure 2.2: **Principles of RNA-seq analysis.** RNA is extracted from the sample of interest and fragmented before being converted to cDNA via reverse transcription. Millions of these fragments are sequenced using NGS and aligned to a known trancriptome. The number of reads that are mapped to each gene gives the level of gene expression, called the 'count'.

Although it has been around for over a decade, the techniques used in RNA-seq are still developing. Sequencing methods are beginning to trend towards taking longer reads in plant and animal studies [137], and longer reads are considered the gold standard in *de novo* assembly of microbial genomes. Single cell RNA-seq is becoming increasingly used in the study of diseases to understand gene expression in different cell type, particularly in cancer where cell specific changes are key to understanding the disease [138]. Methods in the *in silico* steps of RNA-seq, particularly alignment, are constantly being improved meaning old data can be re-analysed using new technologies to obtain more accurate results [139].

2.2.2.1 Advantages of RNA-seq

RNA-seq offers a large number of advantages:

- **High quantifiability**: Microarray values are relative to other signals on the array,

wheras RNA-seq counts are not.

- **High dynamic range**: Unlike microarrays, RNA-seq is good at detecting very low expression values and very high expression values.
- **Can detect novel transcripts**: RNA-seq requires no transcript specific probes as microarray does. As a result of this it can detect novel transcripts, single nucleotide variants, gene fusions, insertions and deletions.

2.2.2.2 Limitations of RNA-seq

Despite it being the gold standard for studying gene expression, there are still limitations to RNA-seq:

- **Lack of reference genomes and transcriptomes**: Some organisms may have no available reference genomes and transcriptomes and so alignment can be extremely difficult and slow. This is becoming less of a problem as it becomes easier to produce quality genomes and transcriptomes.
- **Hard to quantify repetitive sequences**: Genomes that have high repetitive sequences are difficult to map as it cannot be determined which repeat on the genome the cDNA fragment is from.
- **Cost**: For some RNA-seq can have a prohibitively high cost, and cost can limit the sample size in some studies.

2.3 Data pre-processing

Microarray and RNA-seq generate a large quantity of data. Preprocessing is required to get quantification of gene expression data so results can be analysed. Additionally, if poor-quality and noisy data is used in analysis, results will be of little value.

2.3.1 Microarry pre-processing

2.3.1.1 Processing of microarray images

Microarray data is generated as an image, with intensity of fluorescence of spots on the image representing relative gene expression. Platform manufacturers provide their own

software for converting images to data files, for example the Affymetrix© DNA microarray image analysis software coverts microarray images to CEL files. These software have the same basic steps: identification of the spots on the microarray surface, determining the background intensity for spots, subtracting this background intensity from spot intensities and giving relative intensity values for each spot.

Most platform manufacturers provide their own software for processing the data files created from microarray images. For example, CEL files created by Affymetrix© DNA microarray image analysis software mentioned above can be imported using the *ReadAffy()* function in the affy R package [140].

2.3.1.2 Normalisation of microarray data

Microarray experiments have many sources of systematic variation that can affect the measured gene expression levels, and normalisation is required to remove these variations, including batch effects, and allow for comparisons to be made within datasets. Sources of this variation can include efficiencies in fluorescent dye incorporation, heat and light conditions and how the technician performs the work.

Robust Multi-array Average (RMA) is the most widely used approach to normalise microarray data [141], and consists of three steps: background correction, normalisation and summarisation.

The background correction in RMA fits probe intensity to a normal-exponential convolution distribution and then normalises data using quantile normalisation. Data is then $\log_2$ transformed and fit to a linear model which has its parameters estimate using median polish to reduce impact of outlier probes on the data. An in depth view on these steps can be seen in Do *et al.* [142].

2.3.2 RNA-seq pre-processing

2.3.2.1 Quality control of RNA-seq

RNA-seq experiments generate fastq files which contain the raw sequencing reads and quality scores of each base. The presence of quality scores allows for efficient quality

control and pre-processing of RNA-seq data. Assessing the quality of sequences is important to identify any problems that occurred during the sequencing process, as well as during handling of samples, extraction of RNA and preparation of libraries. Removing low quality reads leads to better alignment of reads to reference genomes and transcriptomes.

RNA-seq reads can contain an *N* in place of a base, indicating a poor base call. Removing low quality or *N* bases at the beginning and end of reads, known as trimming, is important in improving alignment of reads. However, if trimming is done to aggressively it can reduce the number of reads available and make those still available much shorter, making them harder to align to a trancriptome [143]. This means trimming needs to be done with care to ultimately improve results.

There is software which can perform both quality control of sequence data and trimming of RNA-seq reads, for example FastqPuri [144]. FastqPuri is very fast and has reduced memory usage compared to existing widely used tools such as FastQC and afterQC, making it a very strong up-to-date approach to RNA-seq quality control.

2.3.2.2 RNA-seq read alignment and quantification

Reads that are quality controlled are mapped to either a genome or transcriptome. Mapping to the genome requires no knowledge of how genes are transcribed and is good for identification of new transcripts, however mapping to the transcriptome allows for better quantification of transcript expressions, especially in organisms such as humans which have a very in depth and comprehensive reference transcriptome. There are a large number of alignment tools for RNA-seq data and a lot of resources are spent improving this step in RNA-seq analysis. As alignment tools improve, old and new RNA-seq datasets can be reanalysed to give improved results. Tools such as STAR [145], TopHat2 [146] and *Salmon* have become extremely efficient at aligning RNA-seq reads.

Once the RNA-seq reads are mapped, tools such as htseq-count [147] and featureCounts [148] are used to count reads mapped to genes to get a quantification of gene expression. These transcript quantification should be checked for transcript length and GC content bias, to see if they will have an effect on read count values. To reduce the

impact of transcript length, sequencing biases and the number of reads, a within sample normalisation approach is generally used. Examples of normalisation approaches are FPKM (fragments per kilobase of exon model per million mapped reads), RPKM (reads per kilobase of exon model per million reads) and TPM (transcripts per million). These normalised values give gene expression values that can be compared across samples reducing the impact of biases in the RNA-seq reads.

There are many similar transcripts within the human transcriptome, and so there can be problems with reads matching multiple transcripts (known as multi-mapping). Advancements to the more basic approaches to quantification of gene expression have been developed that assign multi-mapping reads among matching transcripts and then perform within-sample normalisation. Sailfish [149] and kallisto [150] are examples of these more advanced transcript quantification approaches.

Salmon [151] is a particularly strong approach to aligning and quantifying RNA-seq data. Along with being very fast, *Salmon* does not generate an intermediate alignment file meaning it requires little computational space. As *Salmon* uses a quasi-mapping approach to align reads to transcriptomes, it is not necessary to obtain full alignments to get accurate quantification [149, 151], instead identifying the transcripts where reads may have originated from, which is more than enough to get accurate transcript quantification. Additionally, it allows for correction of sequence-specific biases and fragment-level GC biases during its alignment and quantification.

2.4 Differential gene expression analysis

Differential gene expression analysis is an important technique to identify differentially expressed genes (DEGs) between two sample sets. Genes are differentially expressed if they are statistically significantly different in expression levels or read counts between the two sample sets, and thus likely play a role in the condition being investigated. There are multiple approaches to identifying DEGs which all share two common steps; estimating the size of the differential expression of a gene between two sample sets and then determining the significance of this difference. The most common approaches to iden-

tifying DEGs are *edgeR* [152] *DESeq* [153] and *DESeq2* [154] as they show excellent performance [155, 156].

As well as excellent performance, *DESeq2* has been shown to give consistent results [157]. *DESeq2* performs internal normalisation to correct for library size and sequencing depth and then gets gene-wise dispersion estimates across all samples. A negative binomial generalised linear model is then fit for each gene and a Wald test used for significance testing. This is performed using the DESeq() wrapper from the *DESeq2* R package [158]. This wrapper combines three functions from the *DESeq2* package (*estimateSizeFactors()*, *estimateDispersions()* and *nbinomWaldTest()*).

2.4.1 Meta-analysis

A meta-analysis is a statistical approach that combines data from multiple independent studies to identify a overall effect. Meta-analysis approaches can be used to combine different gene expression studies to increase the statistical power of detecting DEGs. Meta-analysis of gene expression data was initially described by Choi *et al* in 2003 [159], and since then has been widely adapted and improved [160]. They are particularly useful for investigating diseases with many available datasets that have small sample sizes, which is common with NDs [161].

Li *et al* [162] have previously proposed a novel approach to meta-analysis using gene expression data. This meta-analysis method calculates the combined effect size across studies to identify DEGs with the assumption of a normal distribution of the data. This approach works on the combined gene sets from all the studies included in the meta-analysis, rather than the genes that are common between all datasets as other approaches to gene expression data have done [163, 164].

2.5 Network analysis

Gene co-expression relationships contain a wealth of information that univariate methods like differential expression analysis cannot detect [165]. Weighted gene co-expression network analysis (WGCNA) is a popular tool used in systems biology to construct co-

expression gene networks which can detect gene modules as well as identify key genes and hubs within these modules [166].

WGCNA measures the correlation between genes and interprets them as connectivity. Genes with similar expression profiles are often functionally related, for example being controlled by the same transcription factors or involved in the same biological pathways [167, 168]. For this reason, genes with a high connectivity are clustered into modules so that specific biological functions and transcription factors can be identified. It can then be determined if these modules are preserved and reproducible in another set of data to see if a biological function or transcription factor is important in both. WGCNA has been used successfully to identify significant biological pathways and hub genes previously in many diseases, including cancer [169], AD [170] and abdominal aortic aneurysm [171].

The R package WGCNA [166] performs gene co-expression network analysis as follows: A matrix of pairwise correlations between all pairs of genes across each sample group (e.g. case and control groups separately) are created and each raised to a soft-thresholding power to achieve a scale-free topology R^2 of a chosen threshold. From this, a topological overlap matrix (TOM) is calculated, which takes correlation between genes expression as well as connections the genes share into consideration. This TOM is then converted to topological overlap dissimilarities to be used with hierarchical clustering. A dynamic tree-cutting algorithm is then used to determine initial module assignments of genes [166].

There are multiple downsides to using hierarchical clustering to identify modules in networks. The final results of hierarchical clustering is dependent on how the distances are compared, and importantly, once a gene has been assigned a branch on the dendrogram, it cannot be undone [172]. To combat this, Botía *et al* [172] have proposed an additional *k*-means clustering step to improve the results of the hierarchical clustering in WGCNA. This approach has been reported to be able to reduce the number of misplaced genes and improve the enrichment of GO pathway terms.

2.6 Statistical learning

2.6.1 Overview of statistical learning

Statistical learning, otherwise known as machine learning, allows a computer to detect patterns or make decisions based on previous data. Machine learning has becoming increasingly important in bioinformatics to get new meaningful knowledge from complex biological data which is being collected at an exponential rate [173]. These data include imaging (MRI, PET, etc.), omics profiling, clinical data and even data collected by technology such as fitness wearables and social media [174].

Each sample has a number of features, which are the individual measurements or characteristics of the sample. For example, with gene expression data the expression levels of each gene are the features of each sample. Gene expression data contains the expression level of 10s of thousands of genes and are extremely complex, very often having fewer samples than features, making them good targets for use with machine learning [175]. Datasets are typically split into training and test datasets. Training a machine learning model means determining the good values for weights and bias from given training data. The machine learning model is trained and optimised on the training dataset before it is evaluated using the test set. Evaluating the model on this test set allows for appraisal using previously unseen data and avoids the model just fitting exactly to the training data. Model evaluation allows for the researcher to understand how useful the model is and whether it can be improved.

Supervised learning is the most often used approach to gene expression data. Each sample is assigned a label which can be categorical (eg. disease or control sample) or continuous (eg. disease severity) and model is trained using these labels. This trained model is then used to to predict the label of new samples based on their feature set. Alternatively, unsupervised machine learning datasets are not assigned labels and instead models work to identify previously undetected patterns within the data.

2.6.2 Supervised learning

Supervised learning is important in applications to genetics as it allows for classification. This can range in use from predicting the function of a gene based on its network connectivity [176] to identifying biomarkers that can accurately distinguish disease from control patients based on gene expression data [174].

Here, the general supervised learning approaches used in this work are discussed: logistic regression, support vector machines, decision trees and artificial neural networks.

2.6.2.1 Logistic Regression

Logistic regression (LR) aims to separate two classes from one another by creating a linear decision boundary that separate the two [177]. From this decision boundary, the probability of new samples belonging to each class can be predicted. The formula for LR is written as

$$P(Y = 1|X) = \frac{1}{1 + e^{-(B_0 + B_1 X)}} \tag{2.1}$$

where Y is the outcome class and X is the predictor variable. B_0 is the intercept and B_1 is the regression coefficient of X. Logistic regression assumes that the log of the odds of $Y = 1$ occurring is linearly related to the predictor variables. The estimation of the parameters of a logistic regression model are done using an iterative approach to maximum likelihood estimation, usually Newton's method.

Logistic regression is a popular approach to classification due to their good performance on simple datasets, speed of training and ease of implementation. However, the major limitation of logistic regression is the assumption that the data can be linearly separated and so if this is not the case they will not perform well.

2.6.2.2 Support vector machines (SVM)

SVMs aim to generate the hyperplane that has the largest margin separating sets of objects that belong to different classes (shown in figure 2.3). This hyperplane is built from the closest data points from each class, known as support vectors.

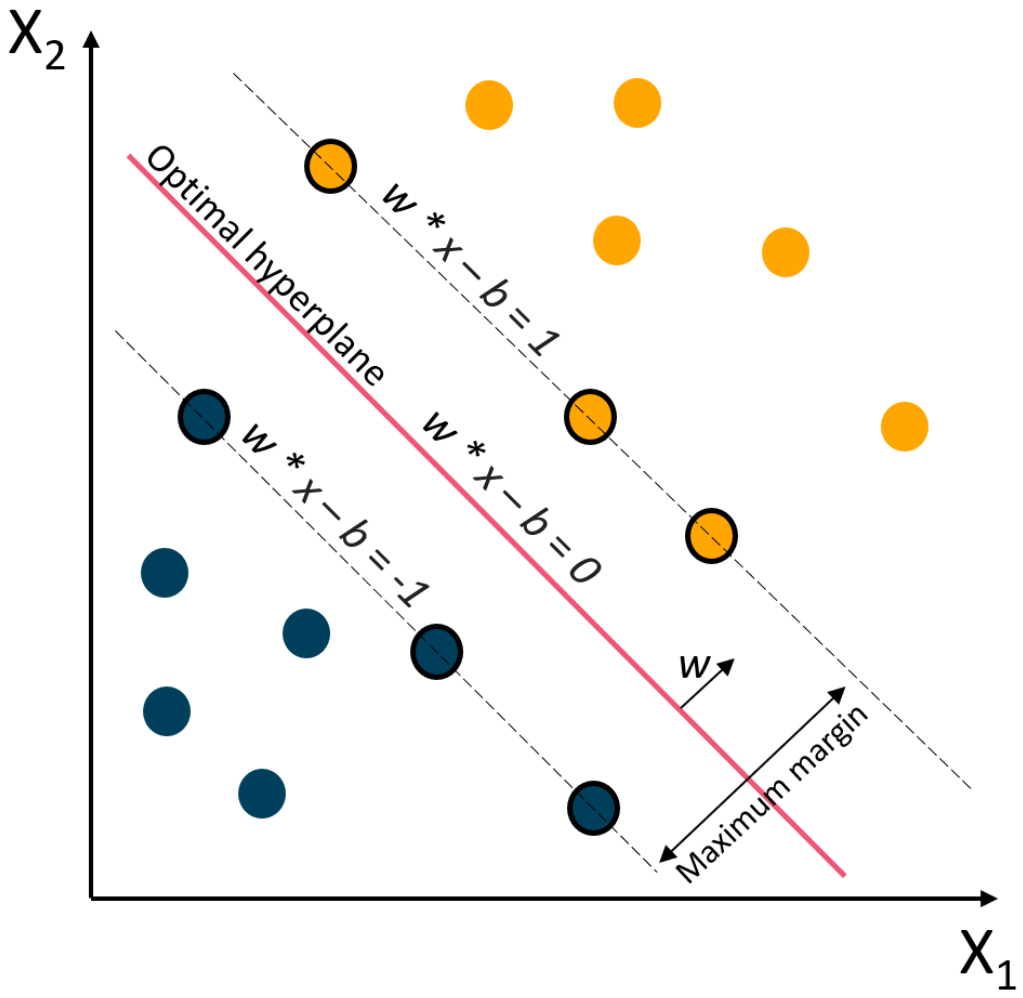


Figure 2.3: **Demonstration of how hyperplanes are found using SVM.** Here, two groups are separated by the optimal hyperplane that exists to maximise the margin between closest samples from each group. The samples on this margin are called support vectors. Adapted from [178].

The optimal hyperplane is defined as:

$$wx^T + b = 0 \tag{2.2}$$

where w is the weight vector, x is the input feature vector, and b is the bias. This satisfies

$$\begin{aligned} w^T x_i + b \geq +1 \text{ when } y_i = +1 \\ w^T x_i + b \leq -1 \text{ when } y_i = -1, \end{aligned} \tag{2.3}$$

which can be combined into

$$y_i(x_i w) \geq 1. \tag{2.4}$$

SVM models aim to identify w and b so that the hyperplane separates the classes within the data and maximises the margin $1/||w||^2$. The support vectors are the vectors x_i of which $|y_i|(wx_i^T + b) = 1$.

Most real life data is complex and difficult to separate without overfitting and so a soft margin, which allows for some misclassifications, is often used [179]. The soft margin SVM is controlled by the C parameter, which determines the trade-off between maximising the margin and maximising the number of points that are correctly classified. The C parameter is carefully tuned in individual models to get the best trade-off for the particular dataset. Additionally, the soft margin is realised by the ξ_i slack variable in the constraints

$$\begin{aligned} w^T x_i + b \geq +1 - \xi, \text{ for } y_i = +1 \text{ and } \xi \geq 0 \\ w^T x_i + b \leq -1 + \xi, \text{ for } y_i = -1 \text{ and } \xi \geq 0. \end{aligned} \tag{2.5}$$

Lagrange Multipliers are used to optimise the SVM with these constraints

$$L_p = \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i\{y_i(x_i w - b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i. \tag{2.6}$$

SVMs were initially developed for linear classification and can be altered to enable non-linear separation of class labels using the radial basis function (RBF) kernel method. The RBF kernel finds similarity between two points X_1 and X_2 using

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2}) \tag{2.7}$$

where σ is the variance and $||X_1 - X_2||$ is the Euclidean (L_2-norm) distance between X_1 and X_2.

2.6.2.3 Decision trees

Decision trees are a classification approach that uses observations from a samples features to identify the class of the sample (shown in fig 2.4). Each node in a decision tree is a feature and progresses to the next node based on feature value. This process starts from the root node and progresses until it reaches a 'leaf' node, which gives a prediction.

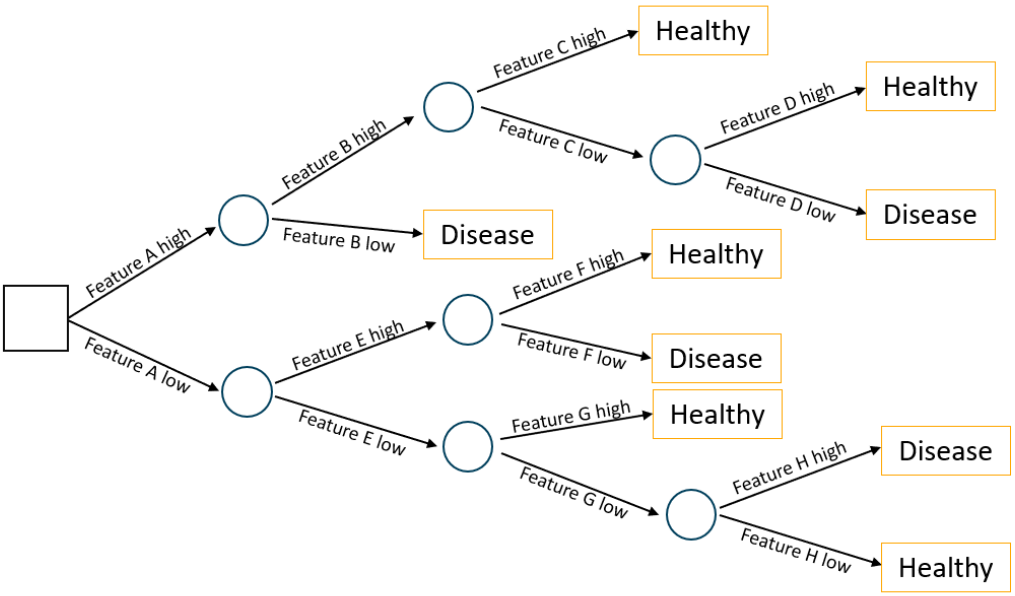


Figure 2.4: **A simple example of a decision tree for classifying healthy and disease patients.** A sample with an unknown class label progresses through nodes in the tree. The feature level (high or low) in the sample determine which node it will progress to. This is repeated until it reaches a prediction.

Decision trees have been used in biology for over 60 years [180] due to their simplicity and ease of interpretation, however they are prone to overfitting on training data. They can be built to be extremely complicated which has allowed for many developments and advancements based on this simple concept.

Random Forest (RF)

Random forest (RF) is an extension of decision trees that works to reduce overfitting by constructing multiple decision trees and having final classification be the mode output of the individual trees [181]. Each tree is constructed using a random sampling of the full dataset with replacement, meaning some observations may be repeated. This is known as bootstrap aggregating.

Additionally, RF selects a subset of features for every tree that is trained to reduce correlations between individual trees. All trees are unpruned and fully grown to reduce the bias of trees. This results in a model that has low bias and low variance as classification is performed from a voting strategy of a large number of low-bias, high-variance trees

[178]. The main benefits of RF is its inherent robustness to avoid overfitting, ease of use and interpretability [182].

Gradient boosting machines (GBM) and Extreme Gradient boosting (XGBoost)

Much like RF, gradient boosting machines (GBM) form a prediction based on an ensemble of weak decision tree models [183]. However, rather than classifying based on a voting strategy of multiple trees each tree is built sequentially. A gradient descent approach is used that adds trees to the model that minimise the error or loss of classification. Gradient boosting is prone to overfitting as it is a greedy algorithm, meaning GBMs require tuning of regularisation and tree parameters to increase the performance of classification.

Extreme Gradient Boosting (XGboost) is an implementation of GBMs that is quickly becoming one of the most popular algorithms for machine learning problems [184]. To control for overfitting that can be present in normal GBM, XGBoost introduces regularisation which reduces feature weights of unimportant features and penalises complex models. Additionally, XGBoost has been built with great scalibility and optimisation that allows for parallel processing, which means it run extremely quickly.

2.6.2.4 Artificial neural networks (ANN)

Artificial neural networks (ANNs) are an approach to machine learning that were inspired by the biological neural networks that constitute the human brain. ANNs are based around connected nodes called artificial neurons that process information from input to output in a series of layers that only allow consecutive layers to be connected. At each node the weight is adjusted to regulate the signal of connection to nodes in the next layer. Inclusion of multiple layers in a neural networks is often known as deep learning. This has become possible as computers have increased in power and speed, and the size and complexity of data increases.

Single layered perceptrons

Single layer perceptrons are the most basic approaches to ANNs [185] that can only learn linearly separable patterns. A perceptron learns a binary classifier using the threshold

function

$$f(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{m} w_i x_i + b > 0, \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

with an input x to a single binary output $f(x)$. w is a vector of real-valued weights, m is the number of inputs to the perceptron, and b is the bias. The perceptron is trained iteratively on the training dataset one at a time. Perceptrons are very simple however have been developed and built upon over time to form other more advanced and complex models.

Multi-layered perceptrons (MLP)

Single layer perceptrons have been expanded on to create multi-layered perceptrons (MLPs) that can handle data that is not linearly separable. MLPs are comprised of multiple layers of neurons that pass information from input to output through a series of hidden layers, as shown in figure 2.5. Activation values are initialised randomly in the input layer and sent to each node in the first hidden layer. Weights are applied to all inputs into each node and an activation function applied to the sum of these weighted inputs which is then passed to all nodes in the next layer. This is repeated until it reaches the output layer where a simple activation function is applied to classify data.

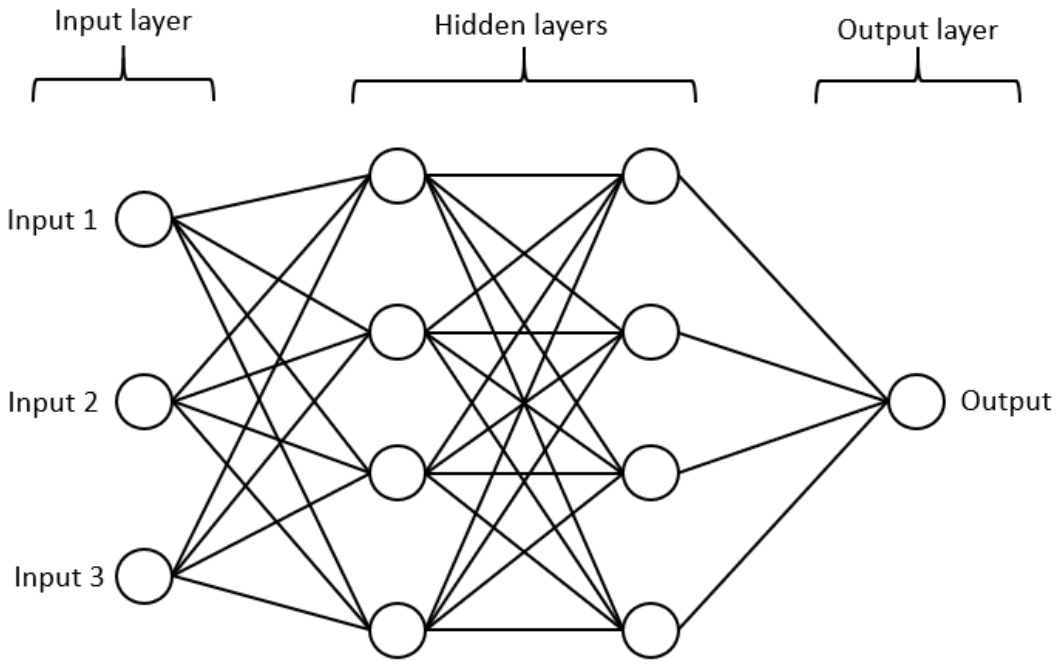


Figure 2.5: **Basic architecture of MLPs with two hidden layers.** MLPs are an interconnected group of nodes organised into layers. Inputs are fed to hidden layers through each circular node which has a weighted connection to the nodes in the next layer. Adapted from [186].

MLPs are trained by iteratively applying training data to the model and comparing the output results of the MLP to the known results. The weights in the MLP are adjusted to improve the model. This is done through the process of backpropagation. Briefly, weights are initialised with small random values and error function value calculated based on known labels of samples. The gradients of the error function with respect to each weight are calculated so that weights can be adjusted to minimise the error. Each weight is iteratively updated by

$$w \leftarrow w - \eta \frac{\partial E}{\partial w} \tag{2.9}$$

where the new weight w is the value of the previous weight minus the value that is proportional to the gradient of the error function. η is the learning rate of the model, which adjusts the size of the changes when updating the weights and $\frac{\partial E}{\partial w}$ is the partial derivative of the error function E with respect to each weight of the array w.

There are multiple activation functions that can be used with MLPs. The most often

used approach is rectified linear units (ReLu) [187]. The ReLu activation function is simply

$$f(x) = \max(0, x) \tag{2.10}$$

where x is the input to the node. As it is so simple it makes using MLP much faster and less computationally expensive than alternatives such as sigmoid and tanh functions [187].

The size of the hidden layers and complexity of the model is important when training MLPs [188]. If they are too simple, models can have weak approximation and generalisation for the problem, whereas models that are too complex with undue hidden layers and nodes will overfit and become excessively computationally expensive and time consuming to train [189]. Tuning this parameter in MLPs can be difficult and leads to inefficient models if not carefully tuned.

The largest disadvantage of MLPs is the long training times as backpropagation adjusts weights across the whole model a number of times. Multiple approaches have been developed to increase the speed of training MLPs, including estimating the optimal initial node weights rather than using random values [190] and the development of the ReLU activation function. Another limitation of MLPs is that they disregard spatial information in data which can impact pattern identification in imaging data, but would have little effect on gene expression data.

Variational autoencoder (VAE)

Variational Autoencoders (VAE) use dimensionality reduction to create a representation for a set of data, called encoding, and then reconstructs the data as close as possible to its original output, called decoding. The encoder and decoder are both ANNs. The basic architecture of VAEs is shown in figure 2.6. The aim of a VAE is to create a latent space that best represents the data for reconstruction. VAEs can be used for unsupervised tasks such as anomaly detection [191, 192], however they perform very well for supervised classification tasks [193, 175].

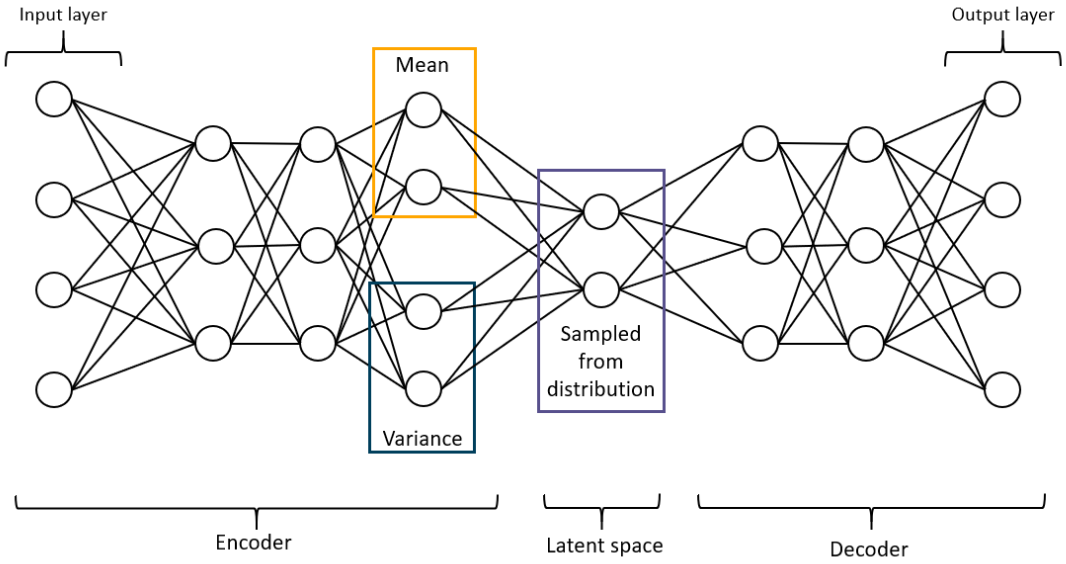


Figure 2.6: **Basic architecture of VAEs.** The encoder passes inputs through hidden layers and mean and variance is found. The latent space is sampled from the mean and variance to create a representation of the original data with reduced dimensionality. The decoder then reconstructs the data based on the latent space as close as possible to the original output. Adapted from [175].

Each variable in the latent space z is generated from the prior distribution $p(z)$ and each sample x has a likelihood $p(x \mid z)$ that is it conditioned on the latent variable z. The distribution of x is found using

$$p(x) = \int p(x|z)p(z)dz. \tag{2.11}$$

The true posterior $p(z \mid x)$ is usually intractable and computationally infeasible. Alternatively, the variational distribution $q(z \mid x)$ is used to approximate $p(z \mid x)$. To get a good approximation of $p(z \mid x)$ using $q(z \mid x)$ the Kullback-Leibler (KL) divergence loss which calculates how similar two distributions are is minimised

$$\min \mathrm{KL}\left(q\left(z|x\right)||p\left(z|x\right)\right). \tag{2.12}$$

This minimisation problem is simplified to be the equivalent of the maximisation problem

$$E_{q(z|x)} \log p\left(x|z\right) - \mathrm{KL}\left(q\left(z|x\right)||p\left(z\right)\right). \tag{2.13}$$

The total loss of the model is a combination of terms, the KL-divergence loss and classification loss

$$\text{Loss} = L(x,\hat{x}) + \sum_j KL\left(q_j(z|x)\,||\,p(z)\right) \tag{2.14}$$

where the L is the cross-entropy loss between true and predicted labels of the data [175].

In many areas of research, VAEs are useful as they allow for construction of new datasets using the representative data in the latent space. For research into gene expression, the value of VAEs lies in the latent space. Gene expression data has high dimensionality, containing a large number of features and low sample numbers [175]. VAEs have the potential to reduce the dimensionality of data [194] so classification can be performed on a much smaller representation of the data.

Convolutional neural networks (CNN)

Convolutional neural networks (CNN) are neural networks that are similar to MLPs, with some changes. They are sparsely connected rather than fully connected and share weights across layers, making them more efficient and better at identifying patterns across a dataset [195]. This makes them particularly adept at applications in computer vision [196]. Importantly, these changes also allow CNNs to work well when built with more layers than MLPs and just as good at reducing data dimensionality.

CNNs use kernel layers that perform the dot product of a sub-region of the input data to return a smaller matrix of output. This reduces the number of features and maintains the information available in these sub-regions. Additionally, CNNs have pooling layers which reduce the spatial size of kernel layers, primarily to decrease the computational cost and time. There are two approaches to pooling, max pooling and average pooling of which max pooling is generally preferred over average pooling as it reduces noise in the data. Max pooling returns the maximum value from sub-region of the data covered by the kernel layer.

CNNs have been used extensively in biomedical imaging studies [197] with very good performance. Imaging data is generally very complex, so CNNs aim to reduce the data to a form that is easier to process while maintaining features that still allow for good pre-

dictions. This makes CNNs particularly scalable for large datasets with a large number of features. CNNs have been used to predict gene expression values by feature extraction from histone modification data in a way that outperforms other machine learning approaches including RF and SVMs [198]. They have also successfully been applied to gene expression data to predict cancer type [196], showing they have potential in classification studies of gene expression.

2.6.3 Unsupervised learning

Unsupervised learning is used to perform cluster analysis and dimensionality reduction of data without the use of class labels [199]. Clustering analysis groups unlabeled data to identify patterns. Clustering is extremely important in the study of gene expression as it allows researchers to identify groups of similarly expressed genes, as well as identify groups of genes that are dissimilar [200, 170]. Dimensionality reduction aims to reduce the dimensional space of a set of data while maintaining important information. This is useful in data that has very high dimensionality like gene expression data which has low sample sizes but large feature numbers, as previously discussed [175].

2.6.3.1 Hierarchical clustering

Hierarchical clustering is a simple but effective approach to clustering data [178]. Briefly, each observation is initially treated as a separate cluster and the closest clusters are merged together. This is repeated iteratively until all clusters are merged together. The number of clusters is then found using a dendrogram (shown in figure 2.7). A threshold distance is set to cut the dendrogram and if the distance between clusters are above the threshold, they are not merged.

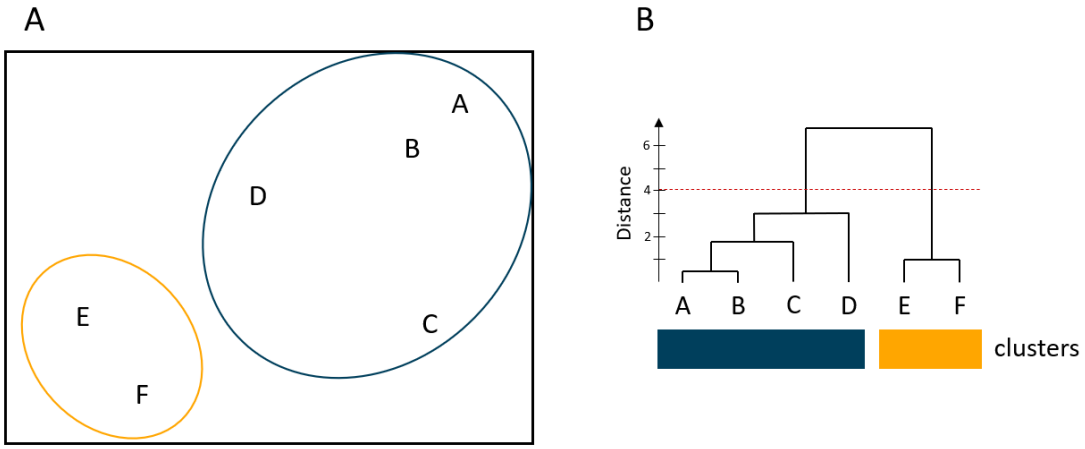


Figure 2.7: **Example of Hierarchical clustering.** Observations and their clusters found using hierarchical clustering (A). The dendrogram (B) shows the hierarchical relationships between the clusters. The dashed horizontal line indicates the chosen threshold distance of 4. Clusters are determined where vertical lines are intersected by the threshold.

There are many ways of measuring distance between clusters and this measure of similarity must be the most suitable for the particular problem. Most commonly Euclidean distance it used, which is simply the length of a straight line between two clusters. It is also important to define where the distances between clusters are measured. Mean-linkage calculates distance using the center of clusters, single-linkage calculates distance using the two closest parts of the clusters and complete-linkage uses the furthest points between clusters.

One of the main advantages of hierarchical clustering is that a number of clusters does not have to be chosen beforehand, and results are easy to interpret with the help of dendrograms. Additionally, hierarchical clustering will always generate the same clusters if it is run on the same data, as it does not begin with initialising any values. On the other hand, choosing the threshold which determines the number of clusters can be difficult and can sometimes be arbitrary [201]. In addition, once a data point is assigned to a cluster it cannot be assigned to another, even if one it is more suited to is found later in the process of hierarchical clustering.

Table 2.1: Pseudo-code for *k*-means clustering

k-means clustering
1. Choose the number of clusters (*k*)
2. Initialise *k* centroids randomly
3. **repeat**
(1) **expectation**: assign each observation to its closest centroid
(2) **maximisation:** calculate the new centroid of each cluster
4. **Until** the position of centroids do not change

2.6.3.2 *k*-means clustering

k-means clustering partitions observations into *k* clusters where each observation belongs to the cluster with the closest mean (centroid). The pseudo-code for *k*-means clustering is shown in table 2.1. Briefly, a *k* number of clusters is selected and centroids are randomly initialised. Then, each observation is assigned to the closest cluster and new centroids calculated. This is repeated until the position of centroids no longer change.

The biggest limitation to *k*-means clustering is selecting the number of clusters. One of the most common approaches is the elbow-method, in which the average within-cluster distance to the centroid is calculated for various *k* and plotted onto a graph. The value of *k* at which the average within-cluster distance to centroid reduction as *k* increases begins to slow down. In many cases it can be very difficult to determine where this point lies, and is up to the discretion of the investigator. *k*-means clustering also does not handle outlier data very well, as they can effect centroid position and sometimes end up in their own clusters.

2.6.3.3 Principal component analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that maps data to a lower-dimensional space so that the variance of the data is maximised. This removes the impact of the least important features while maintaining the most important information. The new features (components) created by PCA are independent. PCA does this by calculating the eigenvectors from the covariance matrix of the data. Much of the variance of the original data is reconstructed using the eigenvectors that correspond to the largest eigenvalues (the principal components). The theory behind PCA is explained in

detail in a review by Jolliffe and Cadima [202].

There are many advantages of PCA. Use of PCA reduces the number of features and removes correlated features in data which can improve the results of machine learning algorithms [202]. Additionally, PCA can transform data into two-dimensions which makes it easy to visualise. This is particularly useful in identifying outlier samples. However, there are some limitations to PCA. PCA assumes that the principle components are linear combinations of the original features and if this is not true results will not be a good representation of the data.

2.6.3.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction similar to PCA, however t-SNE operates non-linearly. Additionally, t-SNE aims to preserve small pairwise distances whereas PCA aims to preserve large pairwise distances. This makes t-SNE very suited to visualising very high-dimensional datasets, including imaging, speech processing and gene expression data [203].

t-SNE works by minimising the probability of distribution between the high-dimensional dataset and a lower dimensional space. The probability distribution of the high dimensional data is calculated with similar objects being assigned a higher probability. A similar probability distribution is defined in lower-dimensional data and difference between the probability of distribution in both spaces are calculated using KL-divergence. The KL-divergence is minimised using gradient descent [204].

As t-SNE is a non-linear dimensionality reduction technique, it can identify patterns in data better than PCA if the principle components of the data aren't linear combinations of the original features. As it is more complex than PCA, t-SNE is computationally expensive and there are several hyperparameters that need to be tuned, including perplexity and learning rate, to get good results. This makes t-SNE much less accessible to researchers working with high-dimensionality datasets.

Table 2.2: **An example of a confusion matrix for binary classification.** The predicted positive and predicted negative are labels predicted by a binary classifier, and true positive and negative are the actual labels of the observation.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

2.6.4 Methods of evaluating models

Once a model is trained, evaluating its performance is extremely important to see if it is constructed well. It is important that the performance of the model is evaluated on unseen data that was not used to initially train the model and so when building models, data is split into test and training datasets. A model may perform very well on the training data but when tested on unseen data perform very poorly, indicating there is a problem with the model. For example, the model may have overfit to the training data. A satisfactory model will have similar good performance on training and test datasets. Most methods of evaluating models are based around a confusion matrix, as shown in table 2.2. A confusion matrix is constructed by comparing the actual labels of observations compared to the label assigned by a binary classifier.

2.6.4.1 Classification Accuracy

Accuracy is a simple way of evaluating classification models. It is defined as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.15}$$

Accuracy works very well in many situations, however there are some disadvantages to using accuracy to evaluate a model. If there is a large imbalance between the number of class labels then a classifier may have a high accuracy just by predicting all observations to be in the majority class [205]. Additionally, in many situations there is a higher cost to misclassifications of the minor class which accuracy does not account for. For example, classifying a patient with a disease as healthy is much worse than a healthy person being

misdiagnosed and going on to have further tests.

2.6.4.2 Receiver operating characteristic (ROC) curve

Another approach to evaluating classification models is plotting the ROC curve and calculating the AUC. The ROC curve plots the 1-specificity ($1 - (\text{TN}/(\text{TN} + \text{FP}))$) of a model on the x-axis against sensitivity ($\text{TP}/(\text{FN} + \text{TP})$) on the y-axis for different values of a continuous test. An example of a ROC curve is shown in figure 2.8. If a ROC curve follows the line $y = x$ then the model produces true positive results at the same rate as false positives and so a curve above this line is expected for a reasonable model. The AUC of the ROC curve is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Effectively, the ROC curve represents a trade-off between sensitivity and specificity achieved with different settings when the method is kept constant. The great advantage of ROC curves is the ease of interpretation and visualisation [206], however ROC curves can be effected by class imbalances as the false positive rate is not effected when true negatives is large.

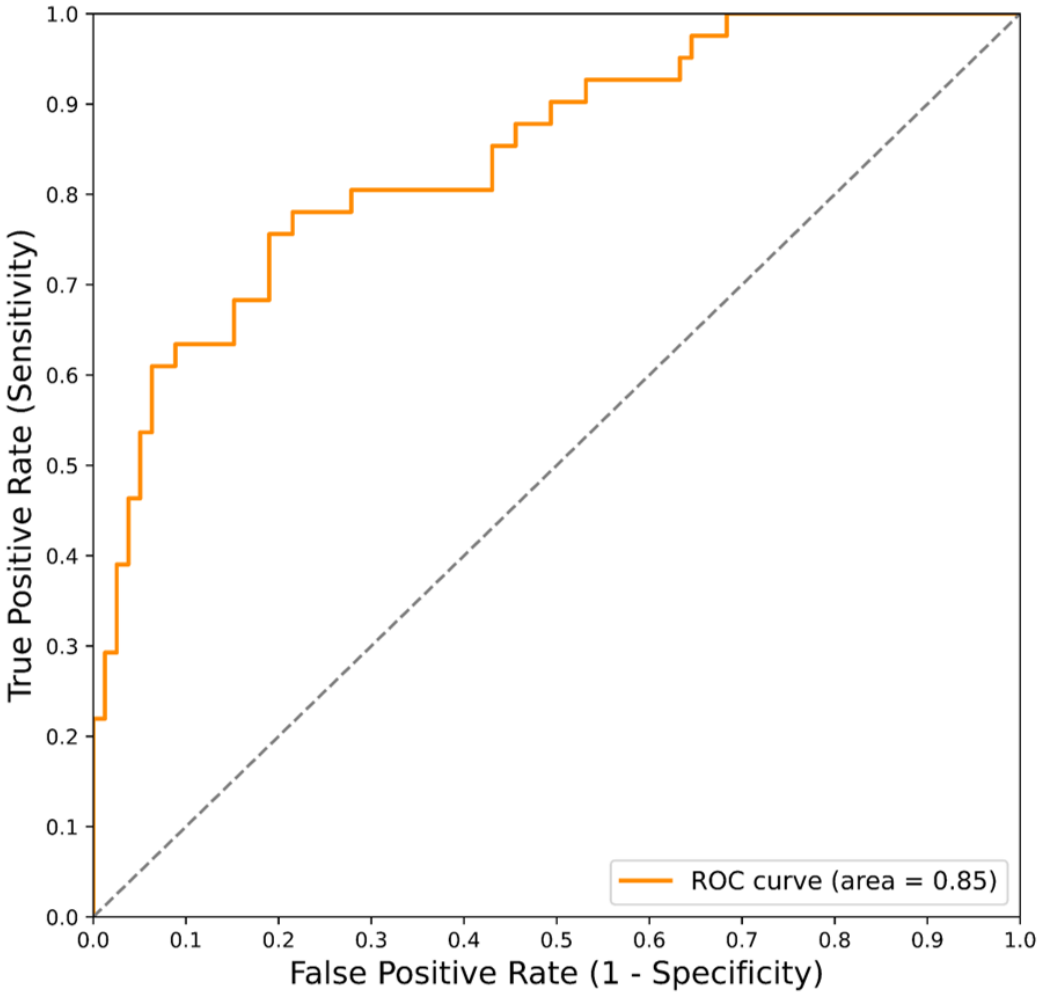


Figure 2.8: **An example of a ROC curve created using simulation data.** The true positive rate (sensitivity) is plotted against the false positive rate (1 - specificity) for different decision thresholds. Each point on the ROC curve is a pair of sensitivity and specificity for a decision threshold. The AUC is measured and determines how good a model is at binary classification.

2.6.4.3 Precision-recall (pr) curve

Considering recall (same as sensitivity) and precision (TP/TP + FP) gives a more complete picture of classification models with imbalanced class numbers. The precision-recall (pr) curve plots the precision of a model against its recall. An example of a pr-curve is shown in figure 2.9. The AUC for the pr-curve is a measurement of how good a model is at binary classification, in the same way as ROC-AUC. In datasets with class imbalances, pr-AUC gives a better indication of model quality than ROC-AUC [207].

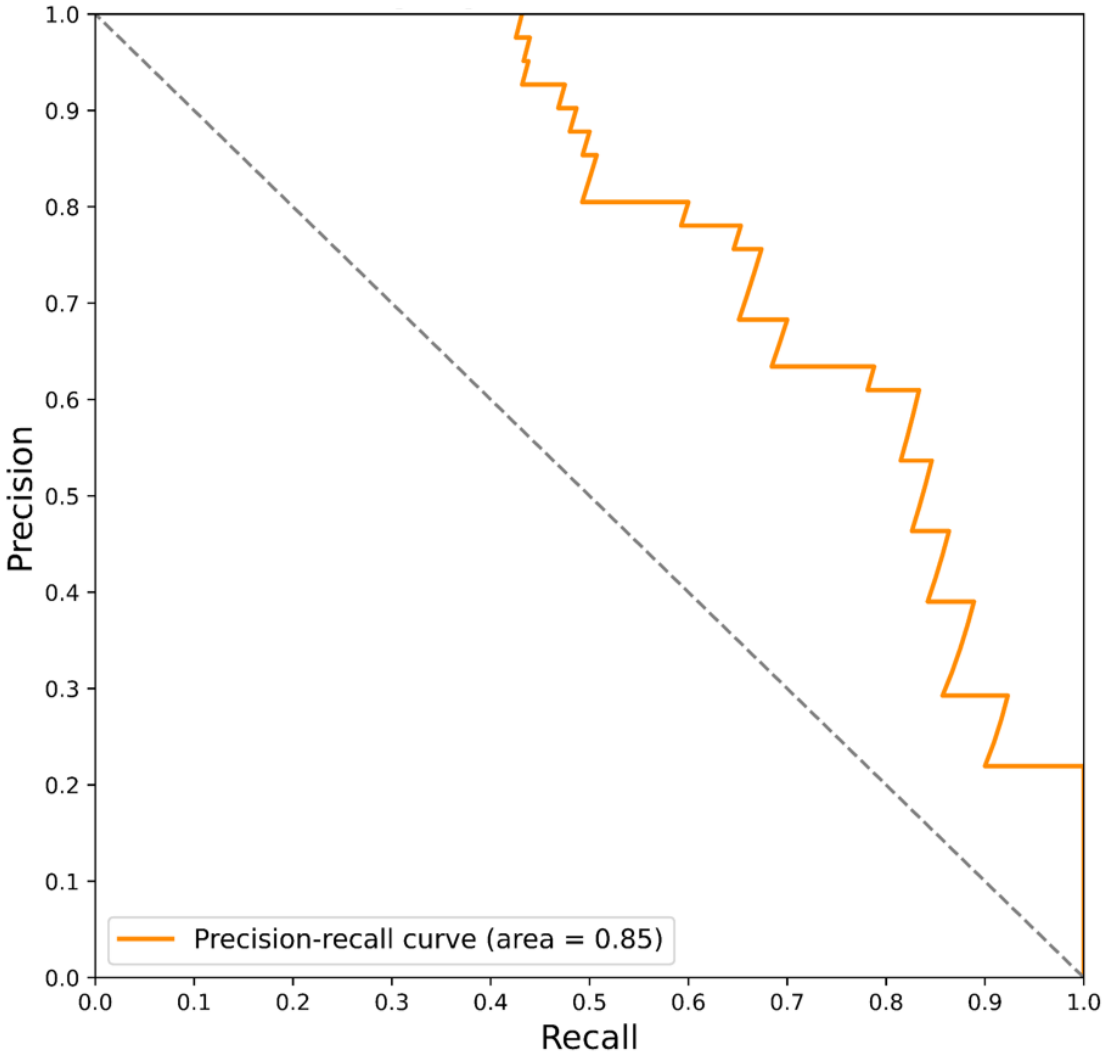


Figure 2.9: **An example of a precision-recall curve created using simulation data.** The model precision is plotted against the recall for different decision thresholds. The AUC is measured and determines how good a model is at binary classification.

2.6.5 Methods of optimising models

In addition to evaluating a trained model on test data, evaluation methods are used to optimise models on training data. When training a model, the aim is to balance it being not too complex to capture relationships between features and labels (underfitting) and being to complex so that only relationships that are present in the training dataset and not in additional data are identified (overfitting). The most common approach to evaluating performance of a model on the training dataset when optimising the model is the re-sampling technique k-fold cross validation [208].

In k-fold cross validation, samples are divided into k folds, where each fold is left

out of training the classifier and used as the test dataset to evaluate the model. The mean error across all folds is used to evaluate the quality of the model. An example of this is shown in figure 2.10. *k*-fold cross validation is used with model optimisation approaches to select the parameters for the classification model that give the best performance.

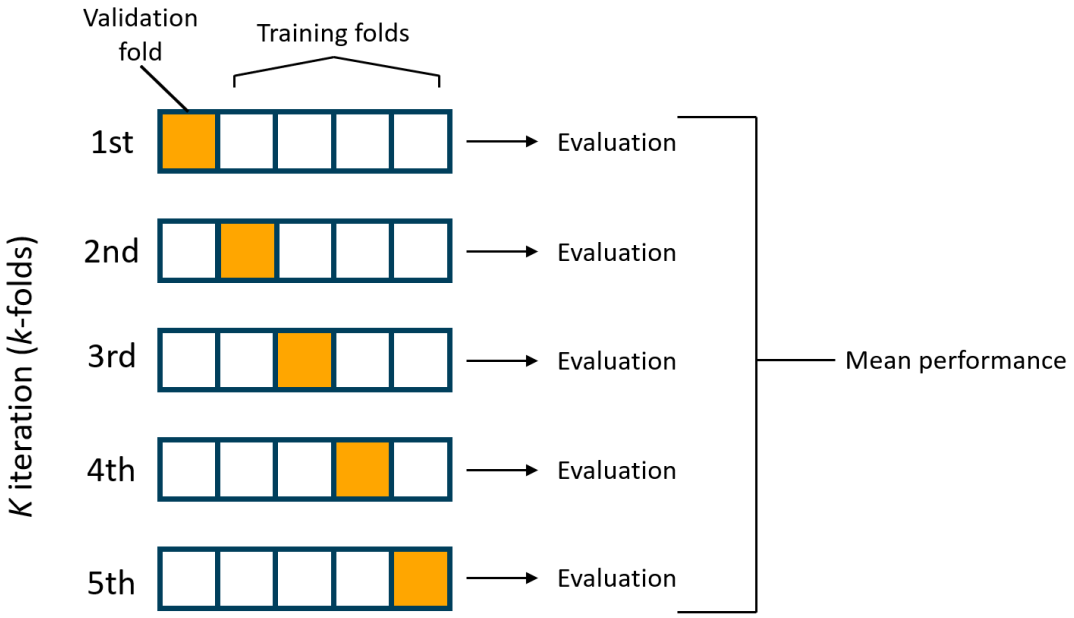


Figure 2.10: **An example of *k*-fold cross validation.** Samples are divided into *k* folds, where each fold is left out of training the classifier and used as the test dataset to evaluate the model. In this example, five folds are used. The model is evaluated on each training and validation fold combination and the mean performance is used to evaluate the quality of the model. Adapted from [209].

2.6.5.1 Grid search

Grid search is a very simple approach to optimising classification algorithms. A manually specified list of parameter options are given to be tested, and an exhausative search of each combination of these parameters is used to train the classification algorithm. The parameter combination that has the best mean performance is determined to be the best model. The biggest disadvantage to grid search is that in models with a large number of hyperparameters that need to be tuned computational time can be very high. Usually, each grid search is evaluated using *k*-fold cross validation, which can increase computational time further. Additionally, grid search can only identify optimal parameters from the selection that are given, so a good selection of hyperparameters need to be selected

initially.

2.6.5.2 Bayesian optimisation

Bayesian optimisation is an approach to optimising models that can search over a larger search space of hyperparameters in less time [210]. It does this by taking into account past evaluations when selecting the next set of hyperparameters. By choosing the parameters to test based on past performance, the time spent trying different combinations of hyperparameters is greatly reduced, and a wider range of values can be trialed.

Rather than a list of values to try for each parameter, a search space is provided and the parameters are sampled from the search space. The objective function of the bayesian optimisation is the cross validation evaluation score. If the chosen evaluation metric is better the larger it is, the aim of the algorithm is to maximise the objective function. It selects parameters to use based on Bayes' theorem, giving the probability of objective function score based on hyperparameters. A prior distribution is updated to a posterior distribution every time a score from the objective function becomes available and so each iteration of the algorithm it becomes a more accurate predictor of validation scores for parameters [211].

2.6.6 Feature selection

Feature selection is an important part of classification in datasets with a large number of features. Gene expression data has high dimensionality with a relatively low sample size making feature selection particularly important to reduce complexity of models and computational time [51]. There are three main approaches to feature selection: filter, wrapper and embedded methods.

Selection of features using filter methods is done independent of any machine learning algorithms and instead based on general characteristics of the data. Examples include Pearson's correlation, analysis of variance (ANOVA) and chi-square test. Although these approaches are fast, they inherently ignore interactions with classifiers, so the features selected may not be what is best to improve the classifier [212]. Wrapper methods involve testing feature sets on machine learning algorithms to see which sets returns the best

model. This approach returns a good feature set for use in classification, however has a high computational cost. Additionally, wrapper approaches have a risk of overfitting, however this can be reduced using cross validation. Embedded methods are implemented by machine learning algorithms that have inbuilt feature selection and are relatively robust to overfitting, however are computationally demanding much like wrapper methods.

2.6.6.1 Recursive feature elimination (RFE)

Recursive feature elimination (RFE) is a wrapper style feature selection approach. RFE works to recursively eliminate the most unimportant feature until a feature set remains. Briefly, an estimator is trained to find the importance of features in the dataset and the least important feature is removed. This is repeated recursively on the feature set until the data is pruned to the desired number of features, usually the number that gives the best performance evaluation scores from the estimator. RFE is computationally expensive and has a huge time consumption particularly in highly dimensional data like gene expression data. To combat this, Li *et al*. [213] proposed an approach called variable step size RFE (VSSRFE).

The pseudo code for VSSRFE is shown in table 2.3. Rather than eliminate features one at a time, an initial step size is set. This step size is the number of genes eliminated at each step in RFE. Once the number of features in the dataset has been halved the step size is also halved until the step size is one. With gene expression data only a small proportion of the huge number of genes will be related to the disease and so if a gene has a low importance when the number of features is large, it is not likely to become more important throughout RFE.

2.6.6.2 Least absolute shrinkage and selection operator (LASSO)

LASSO regression is a supervised learning approach with an embedded feature selection approach. LASSO is an adaptation of linear regression that uses shrinkage. All data points are shrunk to a central point to encourage a simple model with fewer parameters. Although it can be used for classification, LASSO does particularly well at feature selection, shrinking feature importance of unimportant features down to 0 so they can be

Table 2.3: Pseudo-code for VSSRFE. Adapted from Li *et al.* [213]

Recursive Feature Elimination with variable step size (VSSRFE)
1. Given a set of genes, X; labels of sample, Y; number of genes to select, *n_select*; initial step size, *s_initial*
2. Get total quantity of genes from X, *n_total*
3. Temp = *n_total*; N = *n_total*; S = *s_initial*
4. While N>*n_select*:
(1) N = N–S;
(2) If temp / N >= 2 and S >1:
Temp = N;
S = S / 2;
(3) Train classification algorithm with X and Y and get sorted weights vector W;
(4) Delete features according to W and S, and update X;
5. Return X.

removed for classification [214]. This makes it particularly suitable for high dimensionality data [175].

LASSO performs L1 regularisation, which adds a penalty to features which is equal to the absolute value of the magnitude of coefficients. The cost function (sum of squares of the difference between the actual and predicted value) that is minimised for LASSO is

$$RSS + \lambda \sum_{j=1}^{p} \mid w_j \mid \quad = \quad \sum_{i=1}^{n} \left(y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \mid w_j \mid \tag{2.16}$$

where a dataset has n samples and p features from x feature matrix, and RSS is residual sum of squares [178]. w represents the weights of importance of features. λ is the tuning parameter that controls the L1 penalty strength. The greater the value of λ the stronger the penalty and greater number of feature coefficients are set to zero. Setting λ too high will increase bias of results, and setting too low will increase the variance so it is important that λ is tuned to be optimal for the model.

2.7 Immunohistochemistry (IHC)

Immunohistochemistry (IHC) is a immunostaining technique used to identify expression of proteins in the cells of a tissue section. Antigens in cells are identified and bound to by antibodies that can be visualised through staining. Tissue sections are fixed to slides,

primary antibodies that bind to the antigen of interest are washed over the tissue and secondary antibodies are bound to the primary antibodies. An enzyme that catalyses the oxidation of substrate molecules to a coloured product is bound to the secondary antibody (shown in figure 2.11). For example, in 3,3'-diaminobenzidine (DAB) staining, DAB is oxidised by hydrogen peroxide catalysed by peroxidase to produce a dark brown product. Staining can then be observed under a microscope and quantified to calculate the optical density of protein in the tissue.

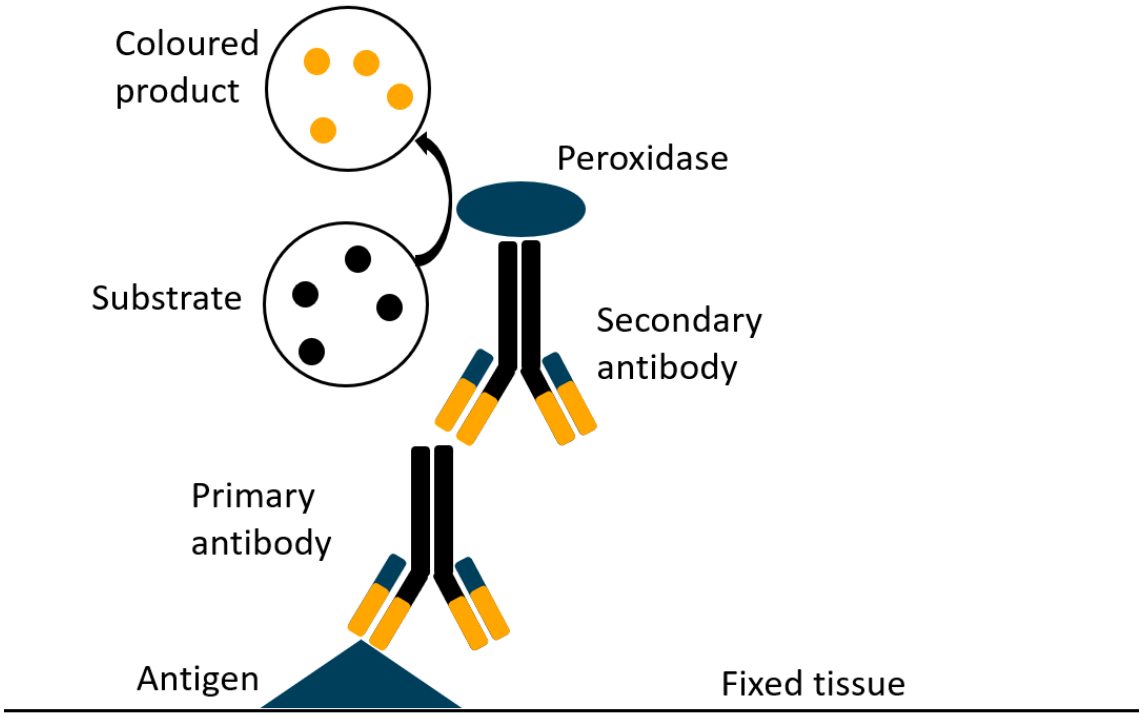


Figure 2.11: **The basic principle of immunohistochemistry (IHC).** A fixed tissue is washed using a primary antibody that is directed towards a specific antigen of interest. A secondary antibody is bound to the primary antibody and carries the peroxidase enzyme which catalyses the transformation of a substrate to a coloured product, which can be quantified.

IHC is very accessible, has a very low cost and is quick to perform making it a very effective and efficient way of investigating protein expression in tissue. However, there are some limitations to IHC. Investigation is limited to known proteins and antibodies that are available can vary in quality. Sourcing quality antibodies requires previous knowledge and research to ensure staining is possible and the research is reproducible.

Chapter 3

Gene expression analysis of Huntington's disease

3.1 Abstract

Although it is known that HD is caused by the CAG expansion in HTT, understanding transcriptional dysregulation in the HD brain will give an insight into how mHTT modulates gene expression. Gene expression changes in the prefrontal cortex were investigated to identify important dysregulated genes, pathways and protein interactions in HD. Results identified *DSP* as an important key gene in HD and potential therapeutic target, and this was confirmed through investigation using IHC. Immune response and inflammatory pathways were shown to be dysregulated in HD brains, particularly involving astrocytes and microglia. The genes *NFE2L2* and *PITX1* were identified as potential therapeutic targets in HD brain inflammation.

3.2 Background

Gene expression analysis in HD is important for multiple reasons. It gives an insight into the mechanism of the disease and can identify potential gene and pathway targets for treatment. Additionally, investigation into gene expression could elucidate the reason some people with between 36 and 39 CAG repeats develop HD while others do not and

elucidate confounding genetic causes of HD. Additionally, although the protein level of mHTT toxicity in various pathways was extensively studied, the role of mHTT in gene transcription level has not been well characterized.

Given mHTT toxicity is caused by the gain-of-function mechanism, it has been recognized that lowering mHTT level may be an effective approach to tackling HD [215]. In fact, this strategy has been used in current clinical trials for HD therapy [216]. Autophagy is important in the modulation of mHTT levels and enhancing autophagic flux has been demonstrated as an effective approach to tackle HD pathology [217]. As such it is important to perform network analysis to identify other processes or genes that could modulte mHTT levels.

Neueder and Bates [218] used network analysis to investigate transcriptional dysregulation in four HD brain regions (BA4 and BA9 frontal cortex regions, cerebellum and caudate nucleus). All brain regions shared common modules. Modules associated with mitochondrial function, glycolysis, intracellular protein transport, proteasome and synaptic vesicles were negatively correlated with HD. In addition, metallothioneins and genes involved in stress response pathways and angiogenesis were positively correlated with HD. This data has been re-analysed by Mina *et al.* [219] with similar results. To further investigate network analysis results, Neueder and Bates compared the human gene expression networks to HD mouse models networks [218]. These mouse models did not reflect some important aspects of HD, including inflammatory response, highlighting the importance of combining gene expression studies and disease models when investigating disease.

Labadorf *et al.* [220] used RNA-seq to investigate 20 HD and 49 neuropathologically normal brain tissue samples from the BA9 area of the frontal cortex. They identified DEGs that implicate dysregulation of transcription, developmental processes and immune response, highlighting the homeobox (HOX) gene family as deregulated. Some of their results were confirmed using RT-qPCR which showed that four out of six genes tested genes (*AHNAK, SLC38A7C, TP53INP2* and *PITX1*) were differentially expressed in the same direction in RNA-seq and RT-pPCR. However, in the last few years, mapping and gene expression quantification have improved [151] so it is possible these previous analy-

sis could be improved upon using new technologies. It remains a challenge to relate gene expression to understanding molecular cause of HD.

Understanding the alteration of gene expression in HD is critical for the understanding of the complex mechanisms that underlay such a devastating disease. In this chapter, bioinformatics analysis of RNA-seq data is employed to comprehensively characterize the gene expression modulated by mHTT in HD. Immunohistochemistry is also used to identify novel HD associated genes and confirm expression changes in brain tissue.

3.3 Materials and methods

3.3.1 Data collection and pre-processing

The publicly available prefrontal cortex Brodmann area 9 (BA9) dataset comprising 20 HD and 49 neurologically normal samples was downloaded from the ENA (European Nucleotide Archive) database (`https://www.ebi.ac.uk/ena`) from accession identifier PRJNA271929. This dataset is the largest RNA-seq of brain tissue for HD.

FastqPuri was used to preprocess the paired-end RNA-seq data. Low quality (below a default of 27) base callings at the beginning and end of reads were trimmed and any reads that had over 5% low quality nucleotides were removed (--trimQ ENDSFRAC). Any N's were trimmed if found at the ends of reads (--trimN ENDS) and reads were discarded if smaller than the minimum length of 31. Trimming was used in conjunction with minimum read length to minimise unpredictable changes in expression estimates that trimming can introduce [143].

Salmon (v1.2.1) [151] was used to quantify transcript expression of the human transcriptome GRCh38 (Ensembl release 100). This was performed with selective alignment enabled (--validateMappings), GC bias correction (--gcBias) and sequence-specific bias correction (--seqBias). Transcript-level abundance was imported and summarised for gene-level analysis using the R package *tximport* [221].

Ensembl ID were then mapped to gene symbol using Ensembl version 86 [222, 223]. If multiple Ensembl IDs mapped to one gene symbol the Ensembl ID with the highest

median absolute deviation (MAD) was kept. MAD was used as the Ensembl ID with the highest MAD had the greatest variability while being robust to outliers, and so likely had the most information[224].

There was one sample missing post-mortem interval (PMI) information, which was imputed using k-nearest neighbour (KNN). The optimum number of clusters was found using the elbow method. The elbow method involves calculating the sum of squared errors (SSE) for KNN with number of clusters from 1 to 15. The SSE is plotted and the number of clusters where the SSE is low and begins to show diminishing returns for the reduction of SSE for each extra cluster. The analysis workflow used in this study is shown in figure 3.1.

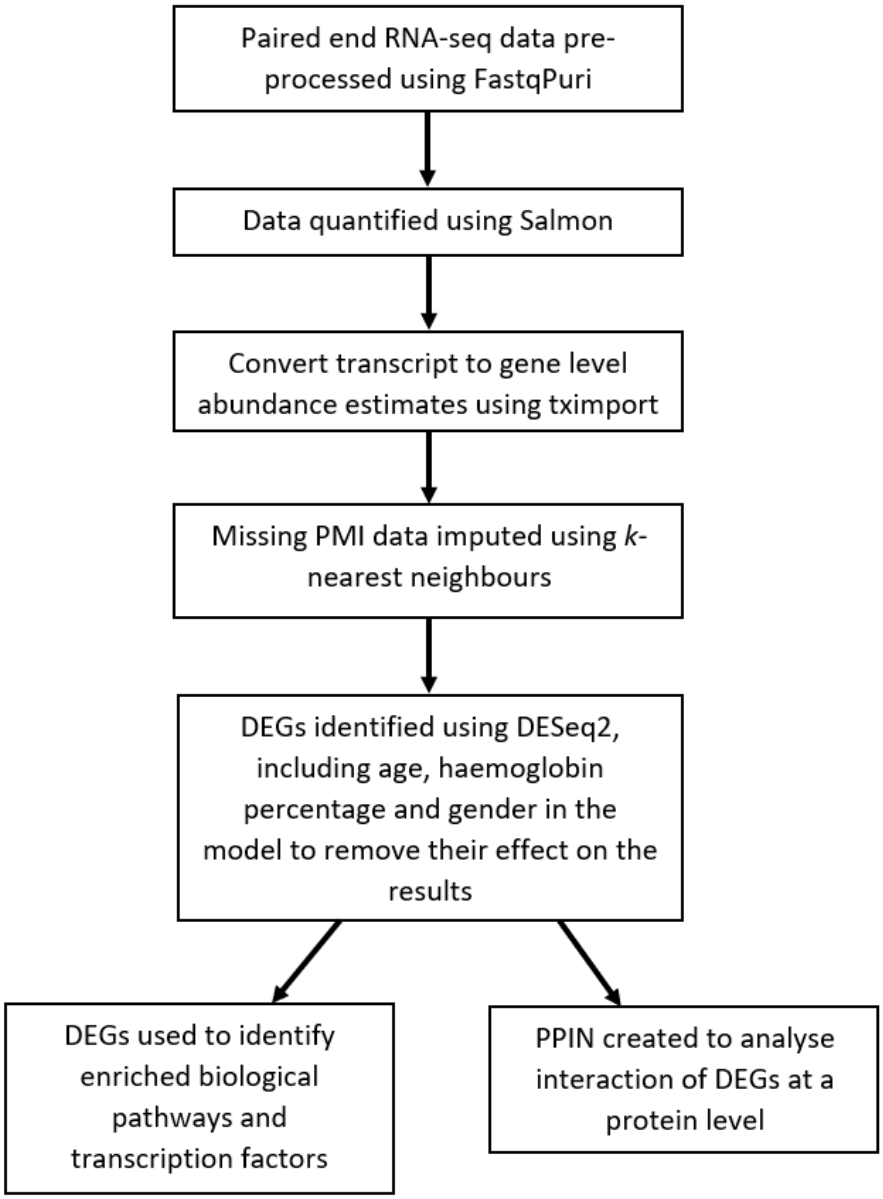


Figure 3.1: Workflow of RNA-seq data analysis. The RNA-seq data was preprocessed using FastqPuri [144] which trimmed read ends and then discards reads if they have over 5% low quality nucleotides and trims ends of reads with N's. RNA-seq transcripts were quantified using *Salmon* [151] and transcript level abundance converted to gene level using tximport [221].

3.3.2 Differential expression analysis

Differential gene expression analysis was conducted in R using DESeq2 v1.24.0 [154]. DEGs were identified adjusting for age at death binned into intervals (0–54, 55–70, 71–106), RNA Integrity Number (RIN) and PMI. The p-values from DESeq2 were adjusted using independent hypothesis weighting (IHW) and genes with an adjusted p-

value < 0.05 were considered differentially expressed [225]. The Nextflow [226] pipeline and docker used to reproduce the analysis is available at `https://doi.org/10.5281/zenodo.4268860`.

3.3.3 Identification of transcription factors, pathway analysis

It has been shown that choice of pathway databases can impact the results of pathway enrichment analysis [227] so to identify the biological pathways that the DEGs represent multiple databases were searched. The gene ontology (GO) biological process, Kyoto encyclopedia of genes and genomes (KEGG) and the WikiPathways databases were searched using the Enrichr web tool [228, 229] to perform pathway enrichment analysis. Pathways were considered significant if they had a Bonferroni corrected p-value < 0.01.

To identify transcription factors (TFs) the DEGs were used to query the Encyclopedia of DNA Elements (ENCODE) and chromatin immunoprecipitation (ChIP) Enrichment Analysis (ChEA) Consensus TFs from ChIP-X found using the Enrichr web tool [228, 229]. TFs with a Benjamini–Hochberg adjusted p-value < 0.01 were considered significant.

3.3.4 Protein-protein interaction network analysis

A protein-protein interaction network (PPIN) from the Human Protein Reference Database (HPRD) was used to analyse the interaction of DEGs at the protein level. The HPRD was built using known protein-protein interactions that were manually extracted from the experimental literature by trained biologists [230]. The HPRD is comprehensive and stringently selects interactions based on *in vivo* and *in vitro* research and includes protein interactions with nucleic acids and small molecules.

The PPIN from the HPRD (release 9) was downloaded and was visualized in Cytoscape v.3.6.1 [231] to create a whole human PPIN with 9617 unique protein entries (nodes) and 37,049 unique undirected interactions (edges). The top 30 most significant DEGs were mapped to build a subnetwork.

Table 3.1: Table showing the age and gender of samples.

Sample	Age	Gender
HD	70	F
HC	72	F
HD	67	F
HC	66	F
HD	66	M
HC	67	M

3.3.5 DSP immunohistochemistry

DSP levels were assessed by IHC staining. Age and gender matched prefrontal cortex sections of control (CTL, n = 3) and HD (n = 3) (shown in table 3.1) brain tissue underwent citric acid heat mediated antigen retrieval. Nonspecific background staining was blocked by a 1-hour incubation in a solution containing 5% horse serum. Tissue sections were then incubated overnight in Desmoplakin polyclonal antibody (PA5-89145; ThermoFisher). Sections were washed of primary antibody, then incubated with secondary antibodies for 1-hour (biotinylated horse anti-mouse/rabbit IgG Vector Laboratories PK-6100). After washing they were incubated in avidin–biotin complex (Vector Laboratories PK-6100) for 30 minutes. Peroxidase activity was visualized with DAB and the slides were counterstained with haematoxylin.

Sections were examined using a microscope equipped with a digital camera (Leica, Germany), and the intensity of the staining was measured using the Fiji ImageJ2 Program [232, 233]. Optical density of images was found using log (maximum intensity value/mean intensity value). An independent Student's t-test was applied to compare optical density of DSP staining between age and gender matched HD and control samples (statistical sig. = $p < 0.05$).

3.4 Results

3.4.1 Differential expression analysis

After IHW correction (IHW p-value < 0.05) 3106 DEGs were identified from the initial pool of 35,412 genes, of which 1912 were upregulated and 1194 were downregulated.

Table 3.2: Top 30 most significant differentially expressed genes found in HD.

Gene Symbol	Ensembl ID	Mean Counts	Fold Change	pval	IHW corrected pval
XKRYP6	ENSG00000237546	1.54	1.05e-9	8.92e-40	8.69e-35
CDC42P6	ENSG00000237350	1.77	1.37e-8	8.92e-40	1.25e-19
HOXA10	ENSG00000253293	3.39	35.01	4.43e-22	4.86e-18
POU4F2	ENSG00000151615	2.80	28.62	5.26e-21	3.88e-17
SLC16A12	ENSG00000152779	51.39	11.36	3.42e-19	1.25e-15
PITX1	ENSG00000069011	4.15	19.28	6.40e-18	3.07e-14
OR9H1P	ENSG00000228336	1.01	1.07e+9	3.54e-17	5.59e-13
F13A1	ENSG00000124491	406.67	9.23	3.01e-15	5.41e-12
HOXA7	ENSG00000122592	3.21	30.47	5.21e-15	1.71e-11
BMP5	ENSG00000112175	76.24	28.45	8.61e-15	1.71e-11
OGN	ENSG00000106809	476.13	20.18	1.31e-14	1.88e-11
NFKBIA	ENSG00000100906	1506.10	2.12	1.95e-14	3.57e-11
HOXA11	ENSG00000005073	1.89	27.86	2.97e-14	6.98e-11
SLC38A2	ENSG00000134294	4607.57	2.30	6.26e-14	7.87e-11
ALKBH6	ENSG00000239382	501.50	0.60	8.45e-14	9.56e-11
VNN2	ENSG00000112303	22.44	7.03	3.82e-13	4.22e-10
IL18	ENSG00000150782	134.16	4.22	3.76e-13	4.22e-10
PNRC1	ENSG00000146278	1604.10	1.75	4.11e-13	4.78e-10
OTP	ENSG00000171540	2.81	10.35	2.76e-13	4.78e-10
FCGR2B	ENSG00000072694	64.71	9.69	5.17e-13	4.86e-10
DSP	ENSG00000096696	351.19	6.81	7.18e-13	4.91e-10
CRABP1	ENSG00000166426	320.16	5.31	8.65e-13	7.24e-10
KRT17P2	ENSG00000186831	49.75	0.29	1.04e-12	8.29e-10
SLC6A20	ENSG00000163817	226.74	3.03	1.23e-12	9.25e-10
MT1M	ENSG00000205364	1056.70	3.35	1.47e-12	1.18e-09
MRC1	ENSG00000260314	134.02	5.30	2.19e-12	1.41e-09
SYTL4	ENSG00000102362	423.32	4.03	3.73e-12	1.98e-09
AL118520.1	ENSG00000231034	0.20	1.07e+9	3.82e-17	2.11e-09
C2orf66	ENSG00000187944	0.20	1.07e+9	3.72e-17	2.11e-09
HOXD10	ENSG00000128710	2.36	33.90	2.50e-12	2.55e-09

A full list of the 3106 DEGs are shown at `https://doi.org/10.6084/m9.figshare.13237253.v1`. Table 3.2 lists the top 30 most significant DEGs, sorted by IHW adjusted p-value. The fold changes smaller than 1e-3 represent very large downregulation of genes and fold changes larger than 1e+3 are massive upregulation of the gene. This could be due to a gene being expressed in controls and not in those with disease or a gene being only expressed in disease patients respectively.

3.4.2 Pathway analysis, identification of transcription factors

As choice of pathway database can affect the results of pathway enrichment analysis [227], multiple pathway databases were searched using Enrichr [228, 229]; GO biological processes, KEGG and WikiPathways pathways. The top 10 significant pathways identified in each database are shown in table 3.3.

Figure 3.2 shows the top 10 statistically significant GO biological processes for the 3106 DEGs by Benjamini-Hochberg corrected p-value. Pathways identified include Neutrophil mediated immunity (adjusted p-value = 5.87e-08, ratio = 129/487), Inflammatory response (adjusted p-value = 1.01e-06, ratio = 79/252) and Cellular response to cytokine stimulus (adjusted p-value = 3.60e-05, ratio = 115/456). Figure 3.3 shows the top 10 statistically significant KEGG pathways for the 3106 DEGs by Benjamini-Hochberg corrected p-value, including Cytokine-cytokine receptor interaction (adjusted p-value = 4.620e-4, ratio = 76/294) and Pathogenic Escherichia coli infection (adjusted p-value = 2.320e-3, ratio = 21/55). Figure 3.4 shows the top 10 statistically significant WikiPathways pathways for the 3106 DEGs by Benjamini-Hochberg corrected p-value, including TYROBP Causal Network (adjusted p-value = 1.732e-9, ratio = 33/61) and Microglia Pathogen Phagocytosis Pathway (adjusted p-value = 3.063e-4, ratio = 19/40).

Table 3.3: Top 10 most significant pathways identified using all HD pre-frontal cortex DEGs.

Pathway	**p-value**	**Adj. p-value**	**Genes in pathway**	**DEGs/gene ratio**
GO Biological Process 2018				
Neutrophil mediated immunity	2.301e-10	5.870e-7	129	0.26
Neutrophil degranulation	6.001e-10	7.656e-7	126	0.26
Neutrophil activation involved in immune response	5.276e-10	8.974e-7	127	0.26
Inflammatory response	1.984e-10	1.013e-6	79	0.31
Cytokine-mediated signaling pathway	1.063e-8	1.085e-5	152	0.24
Cellular response to cytokine stimulus	4.231e-8	3.598e-5	115	0.25
Positive regulation of cell differentiation	4.993e-8	3.640e-5	60	0.31
Response to molecule of bacterial origin	2.317e-7	1.314e-4	36	0.37
Positive regulation of cell proliferation	2.256e-7	1.439e-4	106	0.25
Regulation of cell proliferation	1.265e-6	6.458e-4	163	0.22
KEGG Human 2019				
Cytokine-cytokine receptor interaction	3.000e-6	4.620e-4	76	0.26
TNF signaling pathway	5.413e-6	5.557e-4	36	0.33
Transcriptional misregulation in cancer	2.064e-6	6.358e-4	54	0.29
Pertussis	4.478e-5	2.299e-3	26	0.34
Pathogenic Escherichia coli infection	3.767e-5	2.320e-3	21	0.38
Osteoclast differentiation	7.247e-5	2.790e-3	37	0.29
Legionellosis	3.767e-5	2.901e-3	21	0.38
NF-kappa B signaling pathway	6.767e-5	2.977e-3	30	0.32
Mineral absorption	1.284e-4	4.393e-3	19	0.37
TGF-beta signaling pathway	1.553e-4	4.784e-3	28	0.31
WikiPathways Human 2019				
TYROBP Causal Network	3.669e-12	1.732e-9	33	0.54
Microglia Pathogen Phagocytosis Pathway	1.947e-6	3.063e-4	19	0.48
Nuclear Receptors Meta-Pathway	3.058e-6	3.609e-4	81	0.25
Spinal Cord Injury	1.701e-6	4.014e-4	39	0.33
Adipogenesis	9.184e-6	7.225e-4	40	0.31
Complement and Coagulation Cascades	7.803e-6	7.366e-4	23	0.40
IL1 and megakaryocytes in obesity	1.367e-5	9.214e-4	13	0.54
Pathogenic Escherichia coli infection	3.767e-5	2.223e-3	21	0.38
Macrophage markers	5.829e-5	2.751e-3	7	0.78
Platelet-mediated interactions with vascular and circulating cells	5.445e-5	2.855e-3	10	0.59

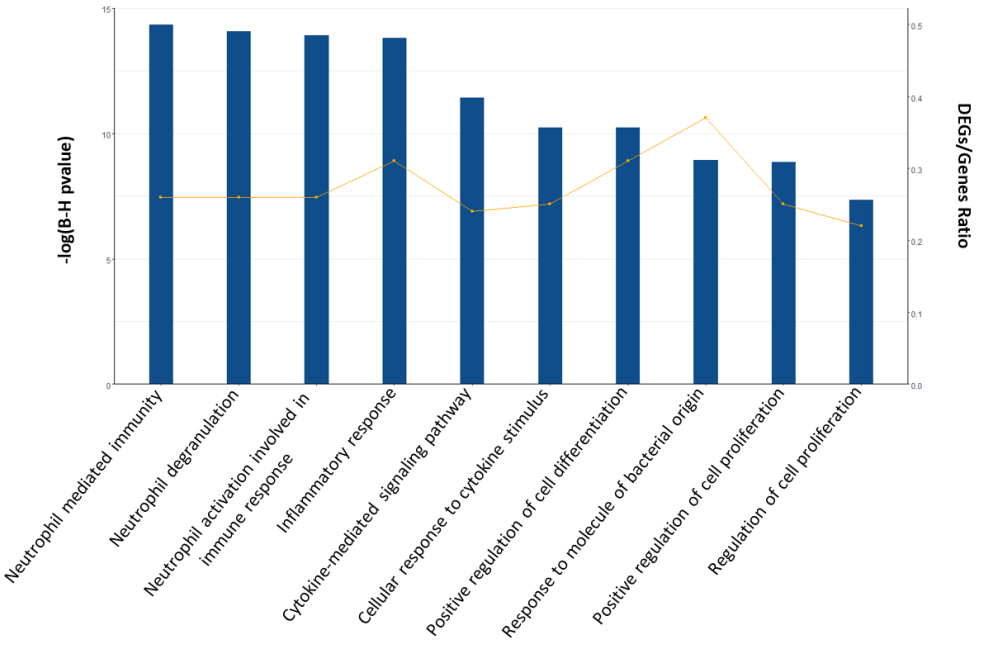


Figure 3.2: Top 10 most significant GO biological process pathways by Benjamini-Hochberg corrected p-value identified using the HD prefrontal cortex DEGs. The DEGs/gene ratio shows which proportion of genes in the pathway are differentially expressed in this work. Benjamini-Hochberg corrected p-value have been transformed to a minus log scale for better visualisation.

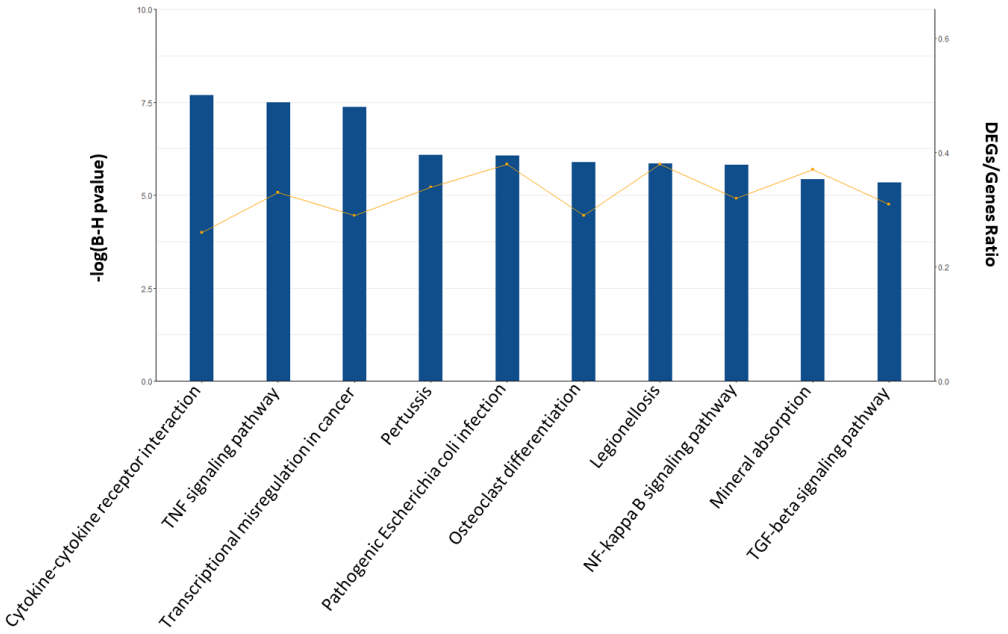


Figure 3.3: Top 10 most significant KEGG pathways by Benjamini-Hochberg corrected p-value identified using the HD prefrontal cortex DEGs. The DEGs/gene ratio shows which proportion of genes in the pathway are differentially expressed in this work. Benjamini-Hochberg corrected p-value have been transformed to a minus log scale for better visualisation.

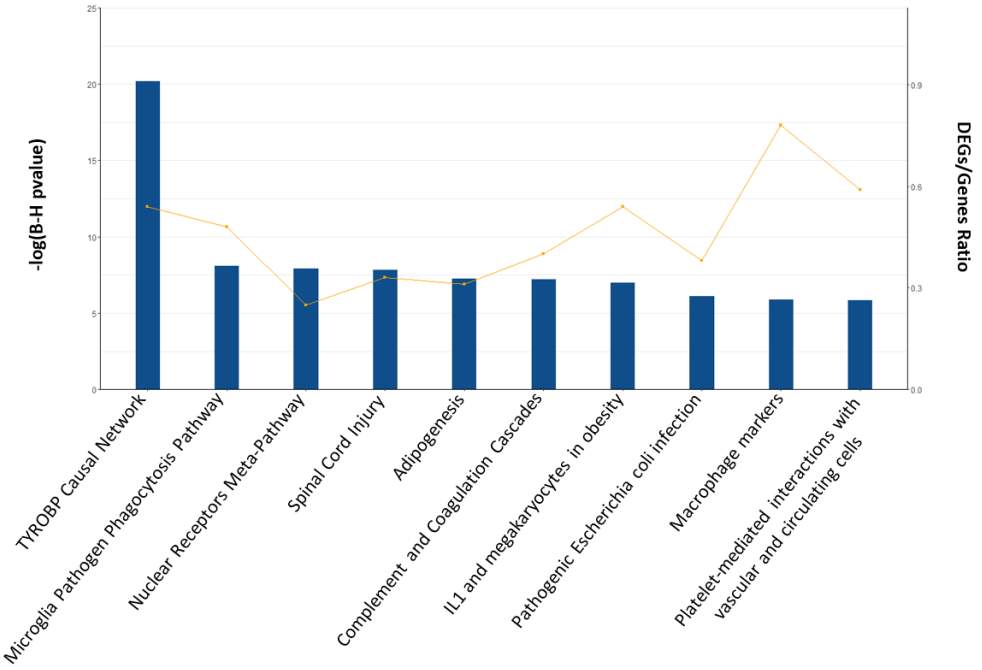


Figure 3.4: Top 10 most significant Wikipathways pathways by Benjamini-Hochberg corrected p-value identified using the HD prefrontal cortex DEGs. The DEGs/gene ratio shows which proportion of genes in the pathway are differentially expressed in this work. Benjamini-Hochberg corrected p-value have been transformed to a minus log scale for better visualisation.

The top 10 significant pathways identified in each database using the 1912 upregulated DEGs only are shown in table 3.4. Using the downregulated DEGs, no significantly perturbed pathways were identified by applying multiple testing. All DEGs, only downregulated DEGs and only upregulated DEGs were used to look for enriched TFs (BH p-value < 0.01) in the CHEA and ENCODE Consensus TFs from ChIP-X database using Enrichr [228, 229]. These are shown in table 3.5. Using all DEGs, the TF *SUZ12* from the CHEA database (adj. p-value = 2.324e-5, ratio = 67/334) was identified. Using upregulated genes, only one TFs was identfied, REST (adj. p-value = 5.23e-04, ratio = 46/383). Using upregulated DEGs *SPI1* in the CHEA database (adj. p-value = 1.162e-7, ratio = 24/159) was identified.

3.4.3 Protein-protein interaction network analysis

A PPIN was created to understand relationships among top DEGs at a protein level. From the top 30 DEGs, 18 were mapped to the PPIN and first neighbour nodes (FNN) extracted.

Table 3.4: Top 10 most significant pathways identified using upregulated HD pre-frontal cortex DEGs.

Pathway	p-value	Adj. p-value	Genes in pathway	DEGs/gene ratio
GO Biological Process 2018				
Neutrophil mediated immunity	6.822e-21	3.481e-17	116	0.24
Cytokine-mediated signaling pathway	5.511e-20	7.031e-17	136	0.21
Neutrophil activation involved in immune response	3.235e-20	8.254e-17	114	0.24
Neutrophil degranulation	4.960e-20	8.437e-17	113	0.24
Cellular response to cytokine stimulus	6.867e-18	5.841e-15	105	0.23
Inflammatory response	6.277e-18	6.406e-15	72	0.29
Positive regulation of transcription, DNA-template	6.100e-17	4.447e-14	194	0.17
Regulation of cell proliferation	3.935e-14	2.008e-11	136	0.18
Positive regulation of cell proliferation	3.676e-14	2.085e-11	92	0.22
Regulation of transcription from RNA Polymerase II promoter	3.553e-14	2.266e-11	229	0.15
KEGG Human 2019				
Cytokine-cytokine receptor interaction	4.233e-13	1.304e-10	70/294	0.24
Transcriptional misregulation in cancer	8.672e-12	1.335e-9	50/186	0.27
Osteoclast differentiation	5.509e-9	5.656e-7	35/127	0.28
Legionellosis	1.195e-8	9.204e-7	21/55	0.38
TNF signaling pathway	2.282e-8	1.406e-6	31/110	0.28
Pathways in cancer	4.707e-8	2.071e-6	90/530	0.17
Proteoglycans in cancer	4.652e-8	2.388 e-6	45/201	0.22
NF-kappa B signaling pathway	1.495e-7	5.756e-6	27/95	0.28
Complement and coagulation cascades	1.845e-7	6.314e-6	24/79	0.30
Malaria	2.617e-7	8.059e-6	18/49	0.37
WikiPathways Human 2019				
TYROBP Causal Network	2.314e-18	1.092e-15	33	0.54
Microglia Pathogen Phagocytosis Pathway	7.134e-10	1.684e-7	19	0.48
Nuclear Receptors Meta-Pathway	1.252e-9	1.969e-7	66	0.21
Adipogenesis	2.904e-9	3.427e-7	36	0.28
Human Complement System	1.504e-8	1.420e-6	29	0.30
Spinal Cord Injury	3.736e-8	2.519e-6	32	0.27
Complement and Coagulation Cascades	3.580e-8	2.816e-6	21	0.36
IL1 and megakaryocytes in obesity	4.863e-8	2.869e-6	13	0.54
Platelet-mediated interactions with vascular and circulating cells	6.453e-7	3.384e-5	10	0.59
PI3K-Akt Signaling Pathway	1.065e-6	5.027e-5	61	0.18

Table 3.5: Top five most significant TFs found using all DEGs, downregulated DEGs and upregulated DEGs. The CHEA and ENCODE Consensus TFs from ChIP-X database was searched using Enrichr [228, 229]. The bracketed database indicates which database the TF was identified in.

Transcription Factor	Pvalue	AdjPvalue	Number of target molecules	DEGs/gene ratio
All DEGs				
NFE2L2 (CHEA)	3.061e-8	3.183e-6	223	0.22
SUZ12 (CHEA)	4.470e-7	2.324e-5	334	0.20
KLF4 (CHEA)	3.948e-5	1.027e-3	199	0.20
SALL4 (CHEA)	3.538e-5	1.226e-3	84	0.24
ESR1 (CHEA)	1.295e-4	2.694e-3	42	0.27
Downregulated DEGs				
REST (ENCODE)	5.025e-6	5.226e-4	46	0.12
Upregulated DEGs				
NFE2L2 (CHEA)	8.955e-11	9.313e-9	161	0.16
SUZ12 (CHEA)	1.949e-9	1.013e-7	233	0.14
SPI1 (CHEA)	3.351e-9	1.162e-7	159	0.15
STAT3 (ENCODE)	1.136e-7	2.363e-6	113	0.18
SALL4 (CHEA)	1.007e-7	2.618e-6	66	0.19

This subnetwork contained 167 nodes and 439 edges, shown in figure 3.5. The top 10 hubs, which had the greatest number of first neighbour connections, are shown in table 3.6. Within the network, six HOX family genes were identified (*HOXA10, HOXD10, HOXA7, HOXB7, HOXA11, HOXA9*) all of which were upregulated. Figure 3.6 shows a subnetwork created using the FNN of the HOX protein family in the top 30 DEG PPIN.

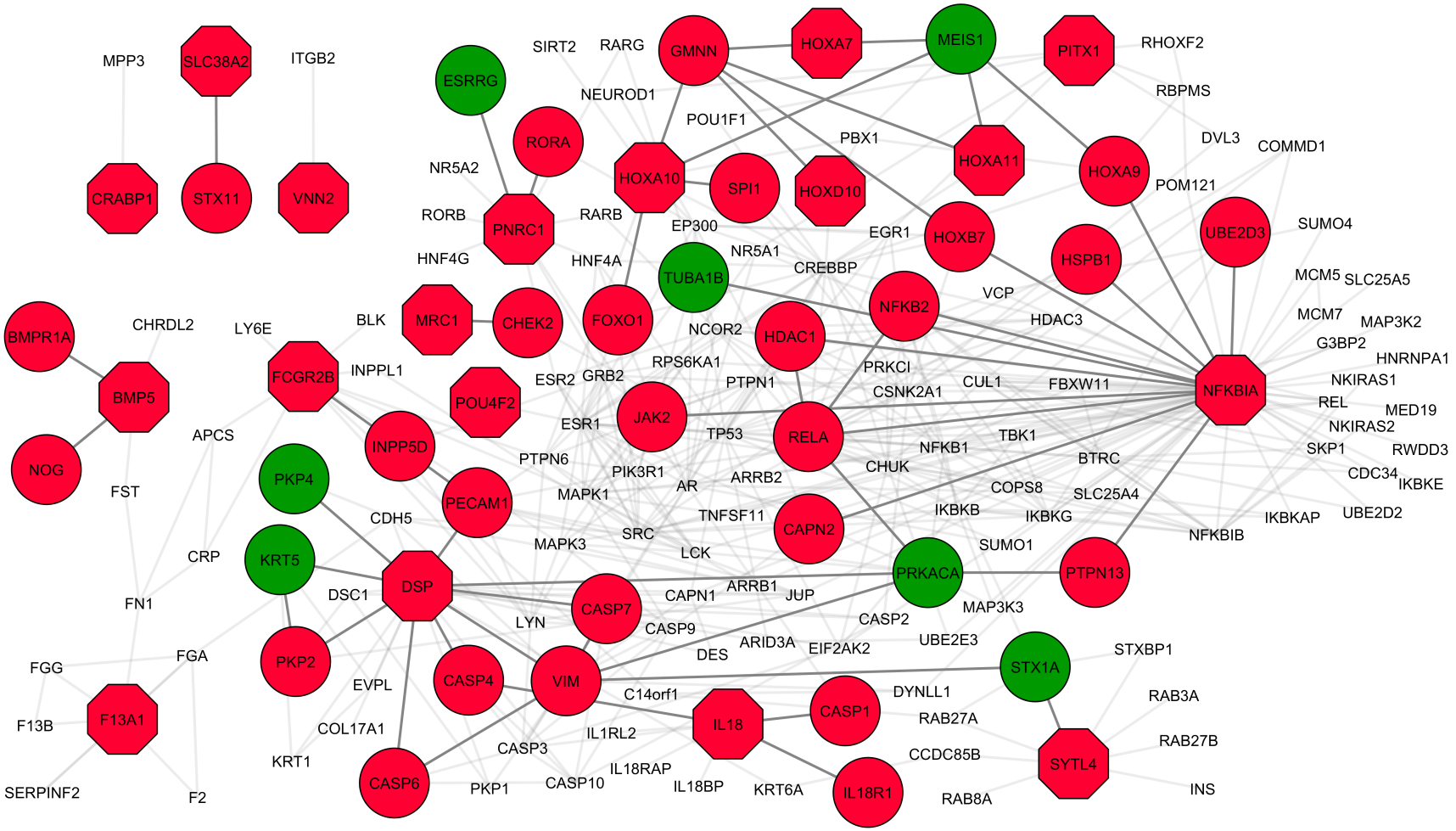


Figure 3.5: PPIN created using the first neighbour nodes of the top 30 DEGs. 18 of the top 30 DEGs were mapped to the PPIN and FNN extracted. This network contained 167 nodes and 439 edges. There were 52 DEGs mapped to the subnetwork, with red nodes indicating upregulated genes and green nodes indicating downregulated genes. Octagons denote genes that were in the top 30 DEGs. Darker edges indicate connection between two DEGs.

Table 3.6: Top 10 hubs found in the PPIN subnetwork created using the top 30 HD DEGs.

Gene name	Number of First neighbour nodes
NFKBIA	62
ESR1	24
DSP	23
CREBBP	22
RELA	21
NFKB1	20
TP53	20
SRC	19
AR	18
CHUK	18

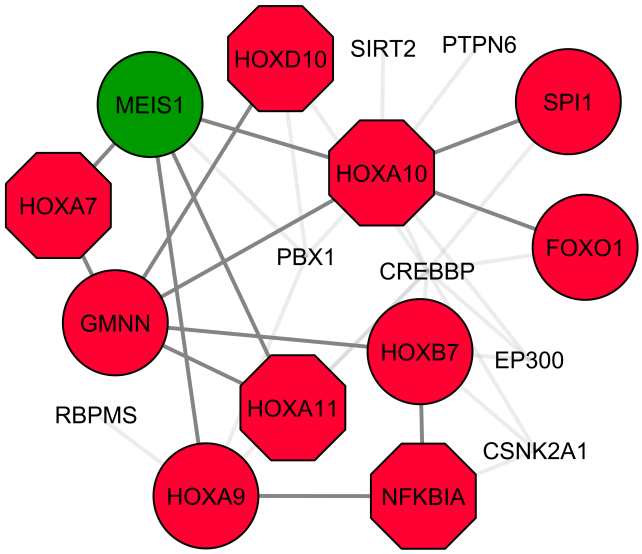


Figure 3.6: Protein-protein interaction subnetwork created using the first neighbour nodes of the HOX protein family in the DEG PPIN. There were 11 DEGs that mapped to this, with red nodes indicating upregulated genes and green nodes indicating downregulated genes. Octagons denote genes that were in the top 30 DEGs. This first neighbour network contains 18 nodes and 33 edges. Darker edges indicate connection between two DEGs.

3.4.4 DSP immunohistochemistry

DSP is within the top 30 DEGs and top 3 hubs found in the PPIN subnetwork created using the top 30 PD DEGs. Figure 3.7 shows a subnetwork created using the FNN of the DSP protein in the top 30 DEG PPIN. DSP has been shown to be expressed in brain endothelial

cells [234] and have an important role in maintaining the function of epithelial barriers [235]. It is a key component of restoration of lung epithelial function after injury [235] and it is possible it has a similar role in the brain during HD. Additionally, the shortening of leukocyte telomeres is associated with HD [236] and DSP has a role in protecting against telomere DNA damage. As DSP is present in the top DEGs and PPIN, has previous literature that supports a potential role in HD and is novel in HD research, IHC staining was used to further investigate whether DSP is expressed in HD brain samples.

Representative images of DSP staining in control and HD human prefrontal cortex samples is shown in figure 3.8. A Student's t-test was applied to compare optical density of DSP staining between age and gender matched control and HD samples, shown in figure 3.9. All age and gender matched samples showed a significant difference (p-value < 0.01) in optical density, however two showed higher intensity in HD samples, and one showed higher intensity in control samples.

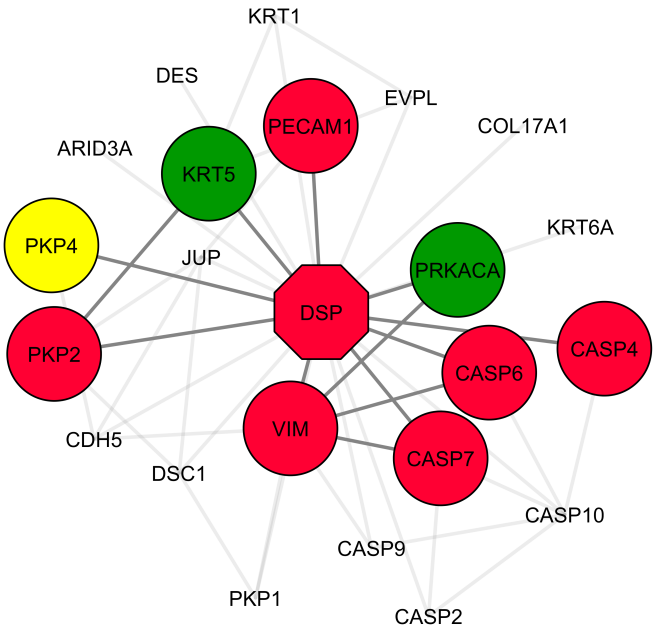


Figure 3.7: Protein-protein interaction subnetwork created using the first neighbour nodes of the DSP protein in the DEG PPIN. There were 10 DEGs that mapped to this, with red nodes indicating upregulated genes and green nodes indicating downregulated genes. Octagons denote genes that were in the top 30 DEGs. This first neighbour network contains 23 nodes and 45 edges. Darker edges indicate connection between two DEGs.

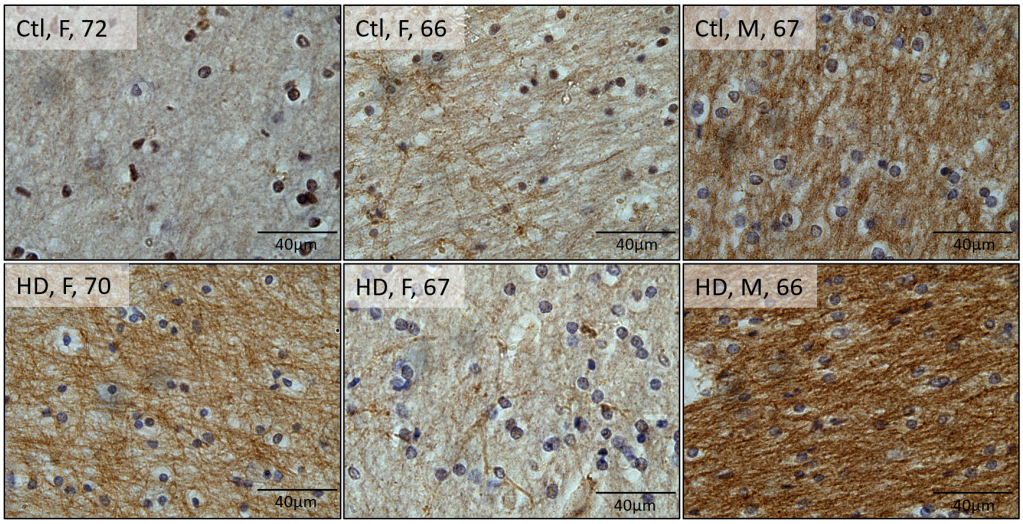


Figure 3.8: DSP protein expression in human prefrontal cortex. Representative images of control and HD. Tissue sections were immunostained with Desmoplakin polyclonal antibody. Scale bars represent 40 microns. Samples are labelled with gender (M= male, F = female) and age.

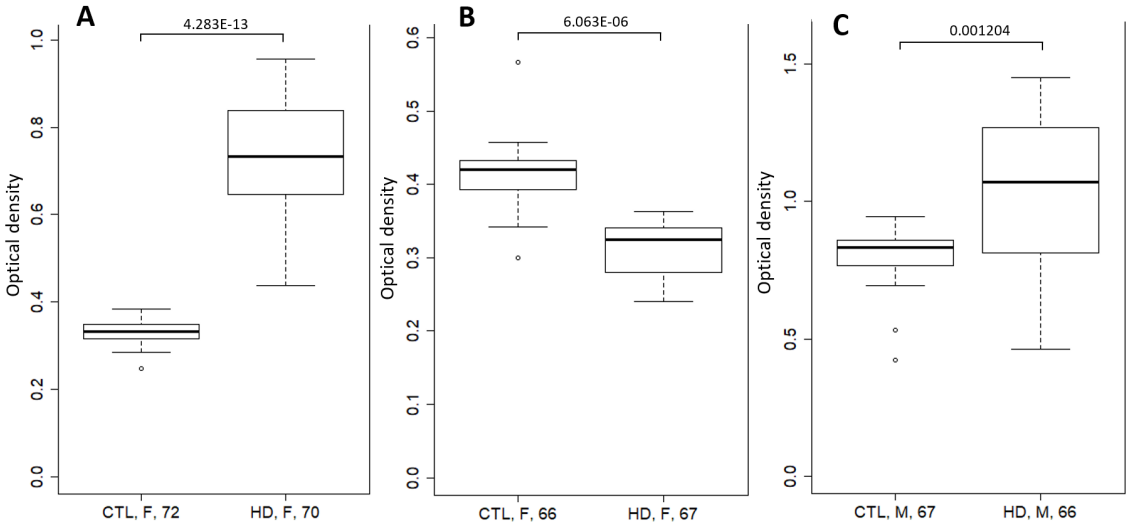


Figure 3.9: The optical density of DSP staining in prefrontal cortex samples from 3 sets of age and gender matched control and HD samples. DSP intensity was significantly (p-value < 0.01) higher in two (A and C) and significantly lower in one (B) of the control and HD pairs.

3.5 Discussion

Using 69 brain samples from prefrontal cortex samples, 3106 DEGs were identified in HD. Using stringent data-preprocessing and the most up to date RNA-seq analysis pipelines identified not only the HD DEGs, but also important pathways, upstream regulators and proteins. There is a significant dysregulation in genes associated with immune response, cell proliferation and cell signalling.

The gene *DSP* was identified within the top 30 most significant DEGs with a 6.81 fold increase in expression. In addition, it was one of the top hubs in the PPIN network created using DEGs. *DSP* (desmoplakin), located at chromosome 6p24, is a major desmosomal protein. Desmosomes are protein complexes that regulate cell to cell adhesion and maintain the mechanical integrity of tissues. DSP plays a role in both assembly and stabilisation of desmosomes and so is a critical component in sustaining tissue integrity. DSP has an important role in maintaining epithelial barrier function [235] and has been shown to be expressed in brain endothelial cells [234]. It has previously been associated with being a key component of lung repair and restoration of lung epithelial function after injury [235]. It is possible that it has a similar function in HD response, being overexpressed as a response to the disease. In addition, DSP has been previously identified as a potential telomere binding protein [237], protecting against telomere DNA damage and cell apoptosis. Shortening of leukocyte telomeres has been associated with HD [236] and so it is possible that DSP has a protective effect in HD.

There is further evidence that DSP plays a protective role in HD in the literature. DSP has been shown to regulate the Wnt/β-catenin signaling pathway in cancer [238], and Wnt Signaling modulation has been shown to correct for aberrant development in HD cell cultures [239]. *DSP* is expressed in proliferating microglia [240] and microglia activation is important in pathogenesis and progression of HD [241]. Additionally, *DSP* variants are associated with malignant arrhythmia [171] and emphysema [242] highlighting its importance in cell disorders. CRISPR/Cas9-mediated epigenome editing has been used previously to regulate expression of *DSP* [235] so it has a potential as a therapeutic target in HD.

IHC was used to confirm the protein expression of DSP in the prefrontal cortex of HD patients. DSP is expressed in the frontal cortex, and has a significantly higher staining intensity in some patients, however one of our age and gender matched samples had significantly lower expression. This demonstrates that DSP expression is deregulated in the HD brain. A possible explanation for the differing directions of staining intensity is linked to the progression of HD in these patients. Having information on the severity of their disease could give more information about the reason for differing *DSP* expression. For example, it may be that DSP is overexpressed early in disease in a protective capacity in HD, and expression reduces once severe cell death has taken place. This is supported by the fact that loss of *DSP* results in shortened telomere DNA and induces DNA damage response, leading to cell apoptosis [237].

Within the top 30 most significant DEGs there are four HOX family genes (*HOXA10, HOXA7, HOXA11, HOXD10*) and six within the PPIN (*HOXA10, HOXD10, HOXA7, HOXB7, HOXA11, HOXA9*), all of which are upregulated. This confirms the results of Labadorf *et al.* [220]. HOX genes are known to be important in regulating development [243] and processes that involve stem cells [244]. MicroRNAs (miRNAs) located in clusters of HOX genes are associated with pathogenesis of HD and exhibit protective effects [245].

These results show that immune response is a major pathomechanism in HD. Multiple neutrophil associated pathways are identified as enriched in the DEGs, including neutrophil activation (adj. pvalue = 8.974e-7, ratio = 127/488) and neutrophil degranulation (adj. pvalue = 7.656e-7, ratio = 126/485). Astrocytes and microglia are important immune cells in the brain and here are shown to be dysregulated. Microglia and astrocytes are crucial in regulating activity of neurons and maintaining an optimal environment for neuronal function. Astrocytes are a type of glial cells in the CNS that envelop synapses and regulate synapse formation, cell homeostasis and are integral in the blood brain barrier, among many other healthy brain functions. *PITX1* is identified as one of the top DEGs, and dysregulation of *PITX1* has been shown previously in HD [246]. As a transcription factor, *PITX1* is involved in the differentiation of astrocytes by regulating the *SOX9* gene [247]. The most significant TF NFE2L2 (adj. pvalue = 3.183e-6, ratio =

223/1014) encodes NRF2, a major regulator of cellular antioxidant defenses. NRF2 activates the metabolism of astrocytes and in AD astrocytes reduces the secretion of Aβ secretion and normalises the release of cytokines [248]. In animal models of HD, NRF2 exhibits a protective effect, however signalling is impaired by mHTT [249]. Investigation into NRF2 and its interactions with astrocytes in HD patients has been limited, and future research may illuminate its potential as a therapeutic target.

Microglia are glial cells responsible for maintenance of neuronal networks and repair of brain injury. In addition to *DSP* being expressed in proliferating microglia [240], the microglia pathogen phagocytosis pathway (adj. pvalue = 3.063e-4, ratio = 19/40) and the TYROBP causal network pathway (adj. pvalue = 1.732e-9, ratio = 33/61) were found to be enriched in the DEGs. In the brain, TYROBP is localised primarily in microglia [250] and is a key regulator of inflammation in AD, repressing microglia mediated cytokine production. There are previous studies that show that activated microglia and astrocytes are important in HD pathology as they transcriptionally activate pro-inflammatory pathways and contribute to inflammation in HD brain [251]. Indeed, enrichment of the inflammatory response pathway is identified (adj. pvalue = 1.013e-6, 79/255). *NFKBIA* is a top DEG and one of the top hubs identified in the PPIN. *NFKBIA* regulates Nuclear factor-κB (NF-κB), which plays a critical role in inflammatory response in NDs [252]. Normal HTT transports NF-κB out of dendritic spines and maintains high levels of NF-κB in the nucleus of neurons where it transcriptionally regulates many genes. This function of NF-κB is impaired by mHTT [253], potentially being a trigger for inflammation in HD brains. Targeting inflammation in the treatment of HD is a promising approach to therapeutics [254], and here potential targets for future research are identified.

There are some limitations to this work. Although RNA-seq has many advantages over alternative methods of detecting gene expression such as microarray, as discussed in chapter 2.2.2, the sample sizes in this study are still relatively small. However, using the best available mapping and qualification approaches to RNA-seq analysis will have extracted the most value from the available data.

The pre-frontal cortex is not the main region of the brain effected in HD, however the main brain areas effected in HD suffer from extreme tissue loss by the time symptoms

manifest. Striatum tissue suffers from extreme neurodegeneration, for the most severe stages of HD this can be over 90% tissue loss [109]. As tissue death is so high, gene expression changes could reflect the processes of cell death or cell type composition differences after tissue death as opposed to changes driven by disease mechanisms. The BA9 region of the prefrontal cortex is impacted in HD [255], and so using frontal cortex data allows investigation of gene changes without severe cell death effecting results. However, it is likely these results only reflect the changes at the point in time of disease and do little to show the changes that reflect pathogenesis of development of disease.

IHC analysis showed dysregulation of DSP expression levels in HD brain tissue, however the direction of this dysregulation was not consistent. Improving the sample size of the IHC analysis of DSP expression in HD brain tissue would give better indication as to the DSP levels in HD patients overall. Importantly, additional data on severity of disease in the patients and other phenotypic information would make interpretation of the results much clearer as this may impact levels of *DSP* expression.

3.6 Conclusion

Using up to date approaches to RNA-seq data elucidates some important underlying genes and pathways in HD. DSP gene expression is shown to be dysregulated in HD patients and disrupted protein levels in HD prefrontal cortex tissue. Testing DSP levels in larger cohorts could give more information as to its potential as a therapeutic target in HD.

In addition, dysregulation of immune response and inflammatory pathways in HD brain are shown. In particular the importance of astrocytes and microglia, and potential therapeutic targets such as NFE2L2 and PITX1, are highlighted.

Chapter 4

Meta-analysis of gene expression for Parkinson's disease and the crosstalk between Parkinson's and Alzheimer's diseases

4.1 Abstract

In this chapter, a novel meta-analysis approach to combine multiple gene expression datasets is used to identify important differentially expressed genes (DEGs) in PD microarray datasets comprising 69 PD and 57 control brain samples, the biggest cohort for such studies to date. Pathway, upstream and protein-protein interaction analysis were performed using identified DEGs. A total of 1046 DEGs were identified, of which a majority (739/1046) were downregulated in PD. YWHAZ and other genes coding 14–3-3 proteins are identified as important DEGs in signaling pathways and in PPIN. Perturbed pathways also include mitochondrial dysfunction and oxidative stress. Additionally, meta-analysis was used to investigate the common pathological and physiological links between PD and AD, as understanding the cross-talk between them could reveal potentials for the development of new strategies for early diagnosis and therapeutic intervention thus improving the quality of life of those affected. A significant overlap in DEGs between PD and AD

was identified, and over 99% of these were differentially expressed in the same up or down direction across the diseases. REST was identified as an upstream regulator in both diseases. This work demonstrates that PD and AD share significant common DEGs and pathways, and identifies novel genes, pathways and upstream regulators which may be important targets for therapy in both diseases.

The work in this chapter has been published in Molecular Brain in 2019 [256].

4.2 Background

There is increasing evidence that PD and AD both have several common characteristics [257]. Around 80% of PD patients develop dementia over time, with the average time from onset of PD to dementia being 10 years [258]. PD and AD are both age-related diseases that have hallmarks of protein aggregation, indeed α-synuclein is found as a non-amyloid component within AD amyloid plaques and over 60% of AD cases are accompanied by the formation of Lewy bodies [259].There are certain genetic variants that increase both PD and AD risk, for example the strong risk factor for AD, APOEε4, has been shown to be related to cognitive decline in PD [260]. There is evidence that molecular pathways, including mitochondrial function, oxidative stress and inflammation underlie the pathogenesis of both AD and PD, however, the pathogenic mechanisms of both diseases have not been entirely explained [257]. There has been found a co-occurrence of Aβ, tau and α-synuclein pathology within neurons and oligodendrocytes from post-mortem brain tissue derived from those with AD and PD [261]. Complex interactions between these proteins can seed the aggregation of each another, though the underlying cause of this is not yet understood [261].

In PD brain tissue, it has been shown that microarray and RNA-seq have a difference in DEG detection, however the enriched pathways shared between the two were similar [262]. This suggests that well performed microarray studies can elucidate the mechanisms of PD in the brain along with more advanced RNA-seq.

The largest RNA-seq study in the PD brain was performed using prefrontal cortex tissue, and subset of these samples were tested using proteomics [263]. This study gives

excellent insight into the transcriptomic and proteomic changes that occur within the frontal cortex of PD patients highlighting disruptions in protein folding, mitochondrial pathways and ubiquitin conjugation pathway, reflecting processes that are characteristic of PD. However, as the prefrontal cortex is not the primary brain region effected in PD, in some cases the PD could have had a minimal effect [264]. The sample size, although largest for the area of study, were still relatively small and imbalanced. In addition, the age of death of the PD samples was significantly later than the control groups and so some detected changes may be a result of age differences.

PD effects different areas of the brain in different ways and so ideally an understanding of the changes in each area would give a greater understanding of how the disease effects the brain. Zhang *et al.* [265] identified that *MKNK2* expression was significantly upregulated and a significant increase in the metallothionein gene group across multiple brain regions in PD, however this study only investigated three regions (SN, putamen, and BA9 area of frontal cortex) and using very small sample sizes with microarray, and so the study had very low statistical power. More recently, Riley *et al.* [262] investigated the striatum, cortex and SN regions of the brain using microarray and RNA-seq. Interestingly, much like Zhang *et al.* [265] they also identified upregulated causal pathways of metal homeostasis driven by robust expression of the metallothionein genes across all three brain regions, highlighting potential importance of metallothionein genes in PD brain. However, they also had an extremely small sample size.

A recent review has highlighted the previous transcriptomics studies published about PD [161]. This review highlights the limitation of small samples sizes in many transcriptomic studies of PD even when not restricted to the SN, demonstrating the need for meta-analysis to increase the power of these previous studies. In addition, it has been shown that there are low similarities between results of previous PD microarray studies in both human and animal tissues, due to the small sample sizes and differing microarray platforms used across studies [266]. Use of meta-analysis methods to increase the statistical power of studies as a result of increasing sample size has been successful in the past in identifying PGC-1α as a potential therapeutic target in PD [267]. Other previous brain microarray meta-analyses have used data from all brain regions available, ignoring

region differences in the brain. Making the data independent to a brain region is important as processes involved in PD can occur dependent on region. However, several previous meta-analysis studies have included repeated samples from patients being analysed using multiple different platforms [268] or multiple areas of the SN being analysed in the same patients. Including these, as previous meta-analyses have done [269, 270, 271], may introduce bias of results towards these individuals.

In this chapter an integrated study is performed to give insight into the genomics, genetics and molecular mechanisms that underlie the features of PD, and reveal the relationship with AD. A novel meta-analysis approach proposed by Li *et al.* [162] is applied to discover DEGs in PD and then a comparison is made to AD. This meta-analysis approach avoids relying exclusively on the genes that have expression data for each constituent study, as previous PD SN meta-analysis have done [163, 164], therefore may lead to novel discovery. The data of the SN was chosen for this meta-analysis as degeneration of neurons in the SN is a hallmark of the disease [257] and has the largest amount of microarray data available.

4.3 Materials and Methods

4.3.1 Data collection and pre-processing

The arrayExpress (`https://www.ebi.ac.uk/arrayexpress/`) and NCBI GEO (`http://www.ncbi.nlm.nih.gov/geo/`) databases were searched using the keywords "Parkinson AND substantia nigra" to find mRNA expression studies of human post-mortem brain tissue from the SN related to PD. Studies were included if they: (1) used clinically diagnosed idiopathic PD patients; (2) used brain tissue samples and (3) had cohorts with more than three samples in either disease or control conditions. If a patient had duplicate samples analysed using different platforms or multiple samples from within the SN, only one of them was used.

Data processing is shown in figure 4.1. All work was done in the R programming language [272]. The identified datasets were downloaded and raw CEL file data were loaded into R using the affy package available on bioconductor (`http://www.bioconductor.`

org) [273]. Boxplots and density plots were used to identify any outlier samples that were subsequently removed. The datasets were then normalized using the RMA approach in the affy R package. Probesets were first mapped to Entrez Gene IDs using manufacturer-supplied annotation files. Probesets that mapped to multiple genes were removed, and for any genes that mapped to multiple probesets only the probeset that had the largest absolute estimated effect size was kept [162].

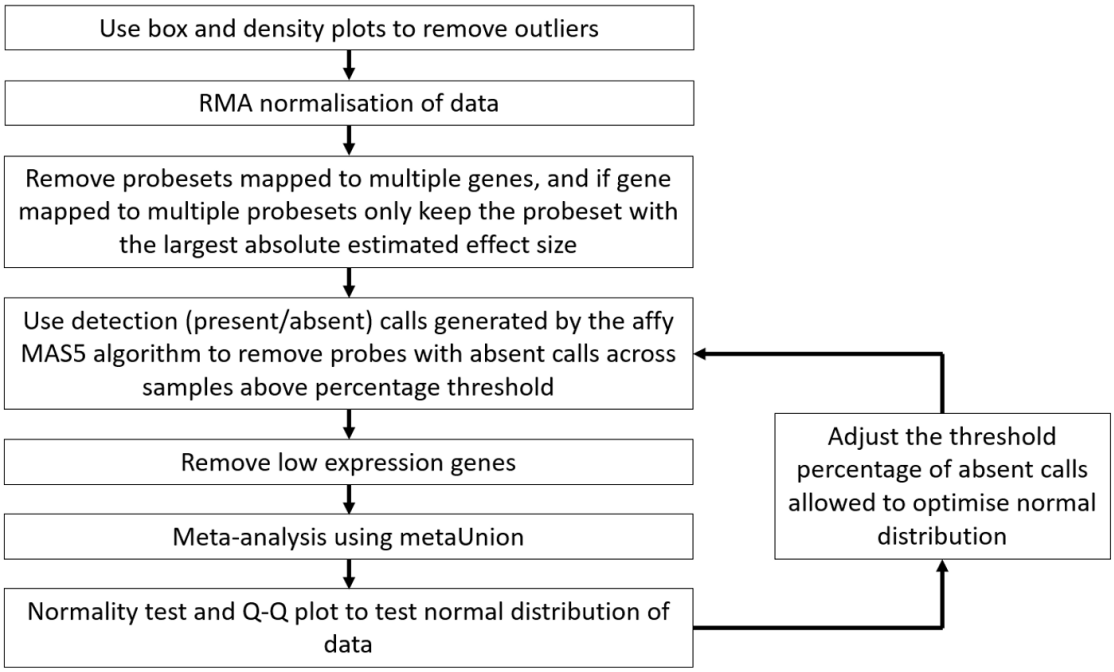


Figure 4.1: **Workflow of data processing**. Outlier samples were removed, and data normalized before the detection (Present/Absent) call algorithm was used to remove data that was not reliably detected. For each study, probesets with absent calls across a chosen percentage of samples were removed. This was repeated in 5% intervals removing probesets with 5% up to 95% of samples absent. The percentage absent cut-off used was set to optimize the normal distribution of the data. After this, the bottom 5% of average expression values across samples was removed and meta-analysis performed.

The first step of pre-filtering was using detection (Present/Absent) call generated by the affy microarray suite version 5 (MAS5) algorithm to remove data that was not reliably detected. For each study, probesets with absent calls across a chosen percentage of samples were removed. This was repeated in 5% intervals removing probesets with 5% up to 95% of samples absent. The percentage absent cut-off used was set to minimize the p-value of the Anderson-Darling normality test using the nortest R package [274] and give optimum Quantile-Quantile (Q-Q) plots of the meta-analysis z-score results. This was

done to reduce how arbitrary the selected filtering parameters are. After this, the bottom 5% of average expression values across samples was removed to reduce low expression data noise.

The Genotype-Tissue Expression (GTEx) database [275] contains RNA-seq data for SN tissue which were used to test robustness of the control data. The RNA-seq Gene TPM from GTEx analysis v7 were downloaded (available at `https://gtexportal.org/home/datasets`). Genes that mapped to more than one gene symbol and any duplicated gene symbols were removed. All RMA normalized microarray control data were merged using the ComBat function [276] from the sva R package [277]. The Pearson correlation coefficient between the average expression levels for the microarray and the average log2 TPM of the RNA-seq was then calculated.

4.3.2 Comparing microarray and RNA-seq data

To see if the microarray data used in this study had a similar quality to that of previous RNA-seq data, the gene expression values was compared to the healthy SN RNA-seq data in the GTEx database [275]. Average absolute expression level of RNA-seq log2(TPM) of SN tissue from GTEx database was correlated with the RMA normalised and filtered intensity of control and PD data separately to see if gene expression patterns between RNA-seq and microarray are similar.

4.3.3 Meta-analysis

Meta-analysis was performed using the novel metaUnion R package previously proposed by Li *et al* [162] (available at `https://github.com/chingtoe365/metaUnion`). This meta-analysis method calculates the combined effect size across studies to identify DEGs with the assumption of a normal distribution of the data. This approach works on the combined gene sets from all the studies included in the meta-analysis, rather than the genes that are common between all datasets as other approaches have done [163, 164]. The metaUnion package is adapted to include age and gender as covariates in the model, implemented using limma [278].

4.3.4 Identification of activated transcriptional regulators, pathway analysis and protein-protein interaction network analysis

The QIAGEN Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, `www.qiagen.com/ingenuity`) software was used to analyse canonical pathways and upstream regulator analysis (URA) [279] of the DEGs. The canonical pathways with Benjamini-Hochberg corrected p-values < 0.05 and upstream regulators with p-values < 0.01 are considered significant.

A PPIN was used to analyse the interaction of DEGs at the protein level. The PPIN from the Human Protein Reference Database (release 9) was downloaded and visualized in Cytoscape v.3.6.1 [231] to create a whole human PPIN with 9617 unique protein entries (nodes) and 39,240 unique undirected interactions (edges). The DEGs and known risk loci for PD identified by a recent GWAS meta-analysis were mapped to the PPIN to build a subnetwork [280].

4.3.5 Comparison to Alzheimer's data

These results were compared to a previous study using similar methodology on AD frontal cortex microarray data [162]. The significance of the DEGs shared between AD and PD was determined using a two-tailed Fisher's exact test and DEGs in common were tested for significant distribution of up or down regulation using a Sign test. Pathways perturbed in both PD and AD in addition to those unique to each disease were identified. Furthermore, pathway analysis was done on DEGs unique to each disease and DEGs shared between diseases.

4.4 Results

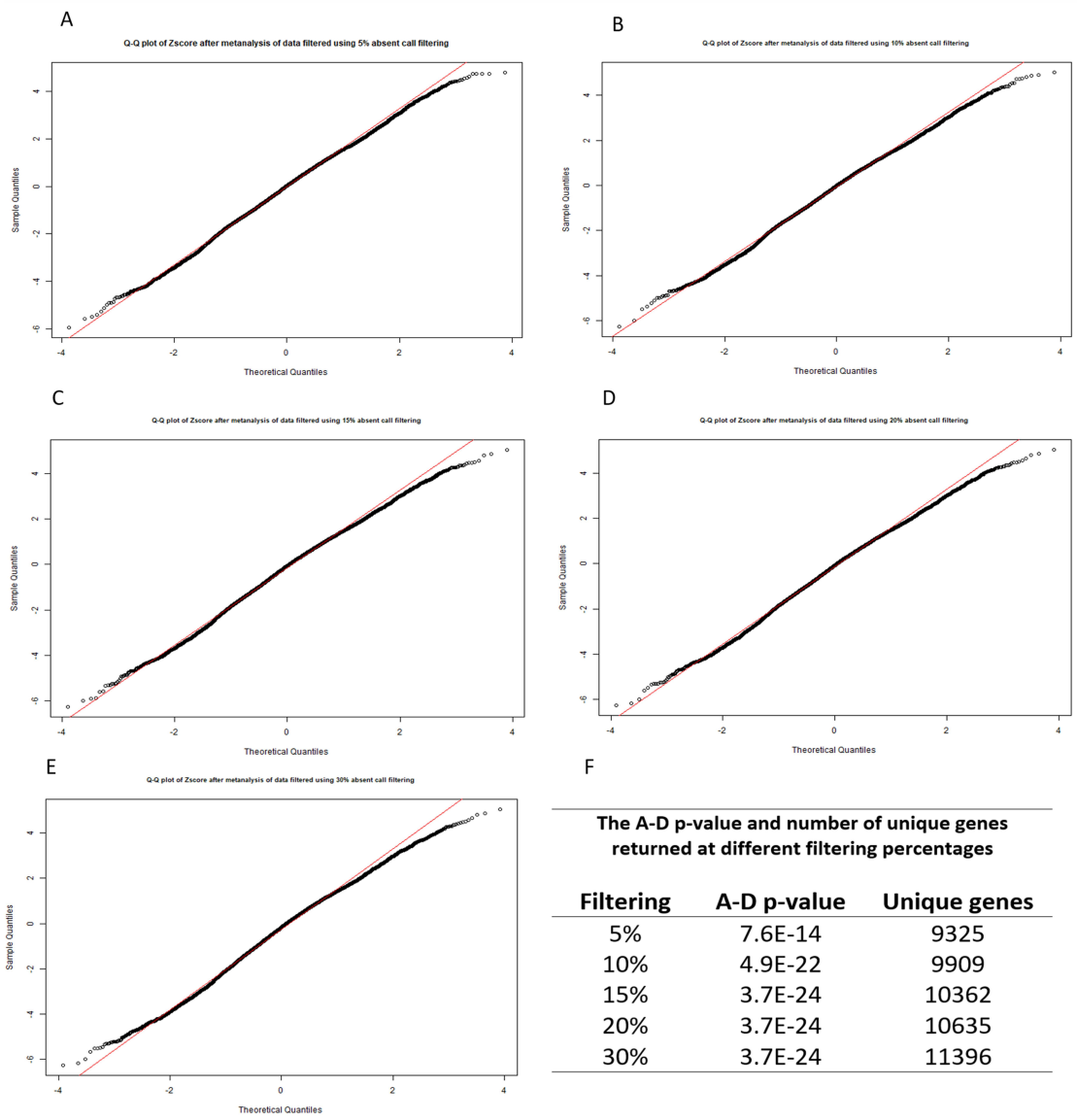
4.4.1 Data sets collected for this study

The search criteria identified 7 Affymetrix chip datasets which included 69 PD and 57 control samples. Information about the datasets is shown in table 4.1. After several rounds of calculation with different filtering threshold, the optimal detection call threshold was

identified as 15% absent as it gave data the closest to normal distribution (shown in figure 4.2).

Table 4.1: Information about each study used in the meta-analysis after removal of outlier samples.

GEO Accession number	Platform name	Platform ID		Male	Female	All	Age range (average)
GSE7621	Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	PD	13	3	16	N/A
			Control	4	5	9	N/A
GSE20141	Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	PD	N/A	N/A	9	N/A
			Control	N/A	N/A	6	N/A
GSE8397	Affymetrix Human Genome U133A Array	GPL96	PD	9	6	15	68-89 (80)
			Control	5	1	6	46-81 (68.2)
GSE20292	Affymetrix Human Genome U133A Array	GPL96	PD	6	5	11	67-84 (75.5)
			Control	13	5	18	41-94 (66.8)
GSE20163	Affymetrix Human Genome U133A Array	GPL96	PD	N/A	N/A	8	N/A
			Control	N/A	N/A	9	N/A
GSE20164	Affymetrix Human Genome U133A Array	GPL96	PD	N/A	N/A	6	N/A
			Control	N/A	N/A	3	N/A
GSE20333	Affymetrix Human HG-Focus Target Array	GPL201	PD	1	3	4	70-87 (77.3)
			Control	5	1	6	68-88 (79)



F

The A-D p-value and number of unique genes returned at different filtering percentages

Filtering	A-D p-value	Unique genes
5%	7.6E-14	9325
10%	4.9E-22	9909
15%	3.7E-24	10362
20%	3.7E-24	10635
30%	3.7E-24	11396

Figure 4.2: The percentage of studies called absent in a mas5 present absent call for each probe was calculated, and threshold determined by minimizing Anderson-Darling normality tests and giving optimal Q-Q plot of the Z-scores after meta-analysis. The Q-Q plot for (A) 5%, (B) 10%, (C) 15%, (D) 20% and (E) 30% filtering. After 15% filtering A-D p-values were minimized (F) and the 15% Q-Q plot gave closest values to normality. A-D is Anderson-Darling normality test.

4.4.2 Comparing microarray and RNA-seq data

SN RNA-seq data from the GTEx database [275] were correlated with the control and PD data to see if the microarray data was of a similar quality to RNA-seq data. Results are shown in figure 4.3. The Pearson correlation coefficient between the control microarray data and healthy RNA-seq data was 0.70 (pvalue $<$ 2.2e-16), between the PD microar-

ray and healthy RNA-seq was 0.73 (pvalue $<$ 2.2e-16). It would be expected that the PD microarray data would have a lower correlation to healthy RNA-seq than healthy microarray due to the DEGs. However, when using only DEGs, correlation between the healthy RNA-seq and the control and PD microarray data reduced to 0.65 (pvalue $<$ 2.2e-16) and 0.66 (pvalue $<$ 2.2e-16) respectively. This suggests the unexpected higher correlation between PD microarray and healthy RNA-seq was likely due to larger sample size of the PD data.

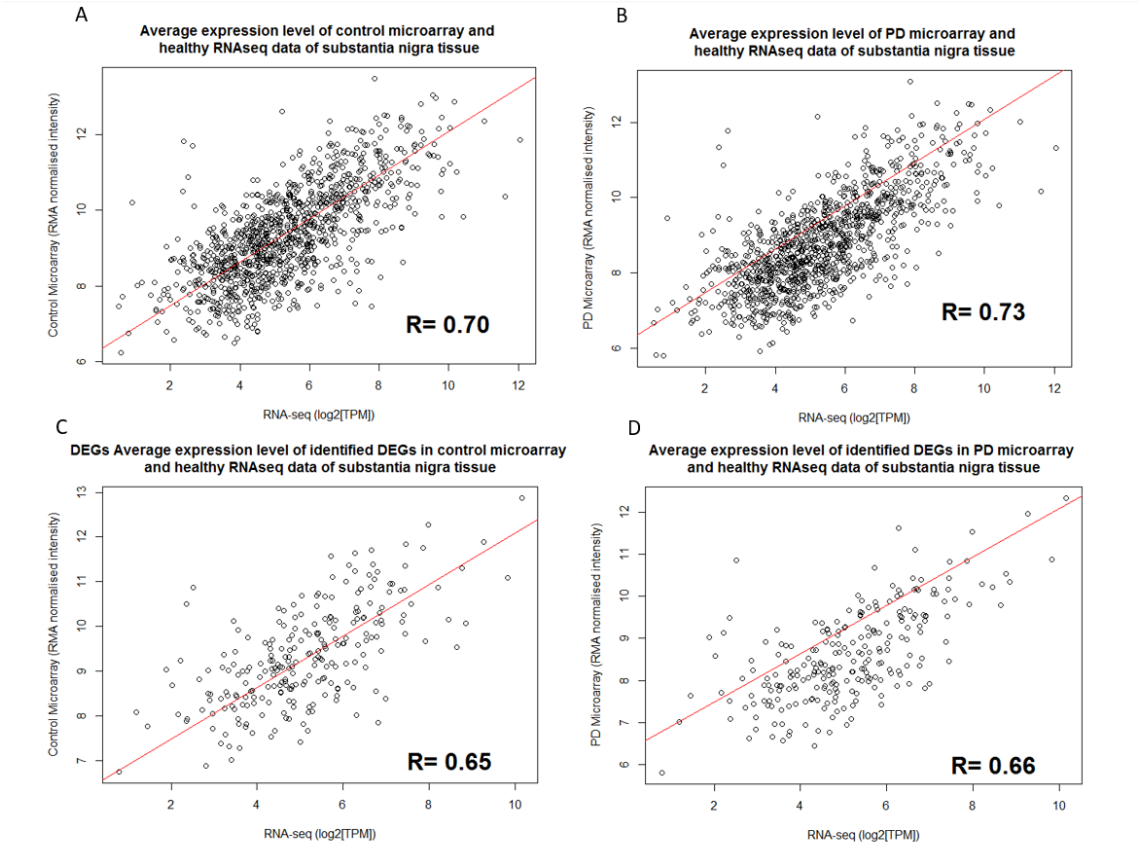


Figure 4.3: **RNA-seq vs microarray** Average absolute expression level of RNA-seq log2 (TPM) of SN tissue from GTEx database plotted against RMA normalised and filtered intensity of microarray control and PD data used in this meta-analysis. The Pearson correlation coefficient between the control microarray data and healthy RNA-seq data (A) is 0.70 (pvalue $<$ 2.2e-16) showing that the expression values of genes between microarray and RNA-seq are correlated and expression data distribution is similar. The Pearson correlation between the healthy RNA-seq and PD microarray data (B) is actually higher than between RNA-seq and control microarray at 0.73 (pvalue $<$ 2.2e-16), when it would be expected to be lower due to some genes being differentially expressed. When using only DEGs, correlation between healthy RNA-seq and control microarray (C) and PD microarray (D) data this difference in correlation is minimised to 0.65 (pvalue $<$ 2.2e-16) and 0.66 (pvalue $<$ 2.2e-16) respectively, suggesting that the difference in correlation could be due to the larger sample size of the PD data.

4.4.3 Meta-analysis

Meta-analysis identified 1046 DEGs from the initial pool of 10,362 genes after false discovery rate (FDR) correction (FDR p-value < 0.05), of which 307 were upregulated and 739 were downregulated. A full list of the 1046 DEGs are shown at `https://doi.org/10.6084/m9.figshare.7252145`. Table 4.2 lists the top 30 most significant DEGs, sorted by FDR adjusted p-value, of which only three are up-regulated.

The metaZscore is the Z score of the gene, it shows how many standard deviations there are in effect size between conditions across all studies a gene is included in. This is important as it is used to calculate p-value and gives information as to if the DEG is up or down regulated and the variability in gene expression. Full explanation of how metaZscore is calculated is given by Li *et al*. [162]. The metaZscore was used to calculate the meta-analysis p-value using $metaPval = 2 * (1 pnorm(abs(metaZscore)))$. As the metaZscore is used to calculate the p-value, for a gene to be differentially expressed at a statistically significant level it needs to be consistent within each condition group and these groups need to be quantitatively different. As a result of this, genes with a higher fold change at the same standard deviation between replicates will often have a lower p-value.

A recent meta-analysis of GWAS data identified 69 risk genes for PD [280] only 49 of which were present in the initial gene pool and 9 were identified as DEGs, including *SNCA*, *ANK2* and *MAPT* (shown in table 4.3). DEGs were more likely to contain disease associated variants than non-DEGs, however the significance of this is not very strong (OR=2.25, 95% CI 0.96 $\sim$4.72, p-value=0.041, Fisher Exact test).

Table 4.2: Top 30 most significant differentially expressed genes found by meta-analysis. metaZscore shows how many standard deviations there are in effect size between conditions across all studies a gene is included in.

Gene name	**Entrez ID**	**Average FC** [a]	**metaZscore**	**Effect** [b]	**FDR corrected pval**
YWHAZ	7534	0.52	-6.26	-------	4.09e-6
SNCA	6622	0.57	-6.00	-------	1.03e-5
DCLK1	9201	0.52	-5.91	-------	1.08e-5
GBE1	2632	0.43	-5.88	-?-----	1.08e-5
PAIP1	10605	0.53	-5.61	------?	4.06e-5
TMEM255A	55026	0.39	-5.58	-??---?	4.06e-5
OLFM1	10,439	0.48	-5.33	--?---?	1.31e-4
OPA1	4976	0.59	-5.32	------?	1.31e-4
HPRT1	3251	0.45	-5.30	-------	1.31e-4
PPP3CB	5532	0.54	-5.25	-------	1.41e-4
PDXK	8566	0.67	-5.24	-------	1.41e-4
SLC18A2	6571	0.31	-5.24	-?-?---	1.41e-4
MDH2	4191	0.60	-5.21	-------	1.50e-4
CHN1	1123	0.54	-5.17	-------	1.77e-4
RAB2A	5862	0.62	-5.10	-------	2.37e-4
RUFY1	80230	1.27	5.04	++?+++?	3.01e-4
CDH8	1006	0.47	-5.00	-????-?	3.47e-4
UBE2N	7334	0.66	-4.93	-------	4.55e-4
ENSA	2029	0.67	-4.93	-------	4.55e-4
SERINC3	10955	0.63	-4.89	-------	4.86e-4
FGF13	2258	0.41	-4.88	-------	4.86e-4
ATP6V1D	51382	0.57	-4.87	-------	4.86e-4
FRRS1L	23732	0.54	-4.87	--?---?	4.86e-4
CDK14	5218	0.67	-4.86	--?----	4.86e-4
LHPP	64077	1.43	4.86	++?++++	4.86e-4
AASDHPPT	60496	0.60	-4.81	-------	5.97e-4
SH3BP4	23677	1.34	4.80	++?+++	6.08e-4
REEP1	65055	0.45	-4.75	--?---?	7.41e-4
FBXO9	26268	0.65	-4.74	------?	7.47e-4
APLP2	334	0.72	-4.72	-------	8.04e-4

[a] Average Fold Change

[b] '+/-/?' indicates up/down and missing in each individual study

Table 4.3: Differentially expressed genes identified in the meta-analysis that have been identified as PD risk genes in a recent GWAS meta-analysis [280].

Gene name	Entrez ID	Average FC	metaZscore	FDR corrected Pval
SNCA	6622	0.57	-6.00	1.03e-5
ANK2	287	0.61	-4.21	2.33e-3
ALAS1	211	0.76	-3.52	1.12e-2
SH3GL2	6456	0.64	-3.46	1.31e-2
DLG2	1740	0.79	-3.34	1.68e-2
SCN3A	6328	0.56	-3.30	1.79e-2
MAPT	4137	1.23	3.15	2.45e-2
ATP6V0A1	535	0.85	-3.03	3.15e-2
VPS13C	54832	1.17	2.85	4.61e-2

4.4.4 Pathway analysis, identification of activated transcriptional regulators and PPIN analysis

After Benjamini-Hochberg correction IPA identified 54 canonical pathways that were significant for the 1046 DEGs, shown in Appendix table A.1. Pathways identified include Sirtuin Signalling pathway (adjusted p-value=2.18e-7, ratio=34/283) and 14–3-3 mediated Signalling (adjusted p-value=9.56e-7, ratio=21/130). Using the downregulated DEGs 81 significant pathways were found (shown in Appendix table A.2). The top ten IPA pathways by Benjamini-Hochberg corrected p-value identified using the downregulated DEGs are shown in figure 4.4.

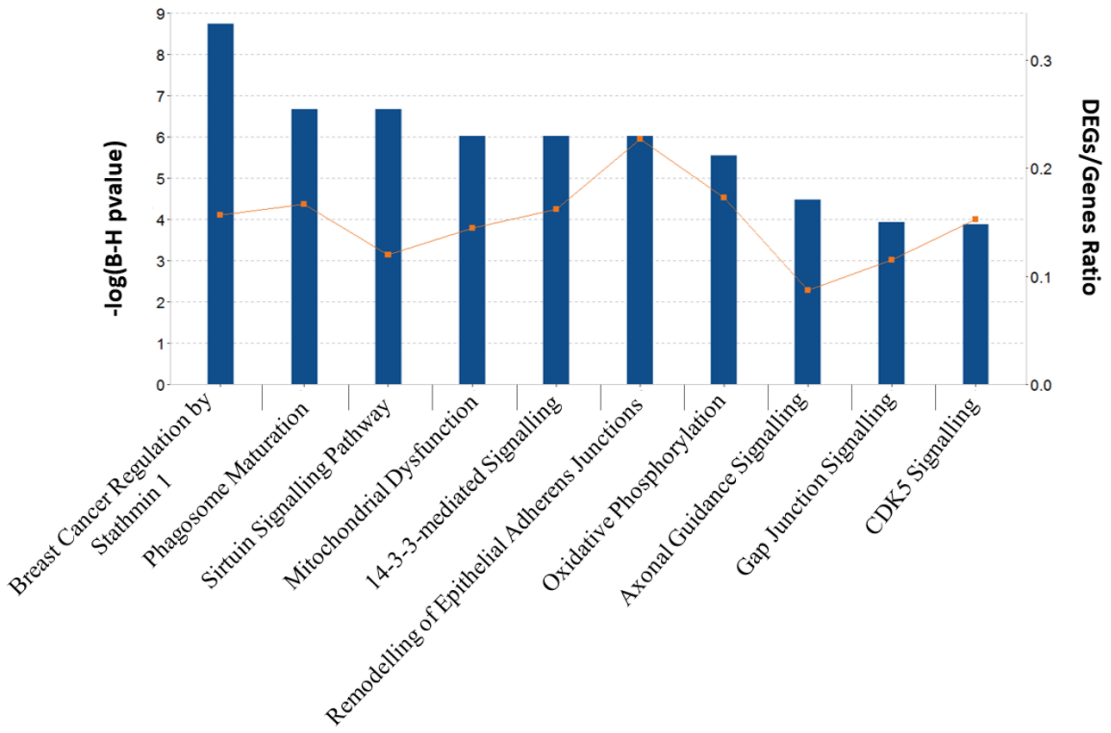


Figure 4.4: Top 10 most significant IPA pathways by Benjamini-Hochberg corrected p-value. Downregulated DEGs between PD and control patients in substantia nigra tissue were used. The DEGs/gene ratio shows which proportion of genes in the pathway are differentially expressed in this work. Benjamini-Hochberg corrected p-value have been transformed to a minus log scale for better visualisation.

Using the upregulated DEGs, no significantly perturbed pathways were identified by applying multiple testing. Using less stringent nominal p-value, ten pathways were identified (p-value$<$0.01), including Adipogenesis pathway (p-value=2.04e-4, ratio=9/132) and STAT3 pathway (p-value=7.41e-4, ratio=7/97). Using down-regulated DEGs IPA identified 17 upstream regulators (shown in table 4.4) including TF *REST* (p-value=2.91e-04), which regulates six down regulated genes (*GAP43, INA, SCG2, SNAP25, TUBB3, UCHL1*). Using up-regulated DEGs IPA identified 25 upstream regulators including *HSF1* (p-value=1.57e-4) which regulates 8 upregulated DEGs.

Table 4.4: IPA upstream regulator analysis for up and down regulated PD DEGs analyzed separately.

Upstream Regulator	**Molecule type**	**p-value**	**Number of target molecules**
Upstream regulators for Down-regulated DEGs			
Lh	complex	1.21e-08	27
FSH	complex	7.25e-07	28
HSP90B1	other	8.38e-05	7
CUL4B	other	2.03e-04	5
SBDS	other	2.67e-04	11
REST	transcription regulator	2.91e-04	6
SUZ12	enzyme	4.01e-04	11
LONP1	peptidase	6.46e-04	9
MMP12	peptidase	1.51e-03	7
INHBA	growth factor	3.64e-03	10
NMNAT1	enzyme	4.84e-03	3
PRKAR1A	kinase	4.84e-03	3
RBM5	other	5.10e-03	6
IL15	cytokine	5.16e-03	10
HNRNPA2B1	other	5.42e-03	12
CCND1	transcription regulator	5.87e-03	16
TP53	transcription regulator	7.91e-03	37
Upstream regulators for Up-regulated DEGs			
HSF1	transcription regulator	1.57e-04	8
TGFBR2	kinase	5.16e-04	6
miR-346 (and other miRNAs w/seed GUCUGCC)	mature microrna	6.73e-04	2
TP73	transcription regulator	8.09e-04	9
SP4	transcription regulator	1.09e-03	3
MTOR	kinase	1.20e-03	5
NPAT	transcription regulator	1.33e-03	2
AREG	growth factor	2.60e-03	5
COL18A1	other	3.18e-03	5
ZBTB10	other	3.27e-03	2
MYC	transcription regulator	4.07e-03	10
miR-22-3p (miRNAs w/seed AGCUGCC)	mature microrna	4.23e-03	3

CD24	other	4.50e-03	6
ZNF652	other	4.53e-03	2
CCND1	transcription regulator	5.50e-03	9
GATA6	transcription regulator	5.86e-03	4
Cdk	group	6.11e-03	3
SAFB	other	7.26e-03	4
mir-122	microrna	8.41e-03	5
miR-155-5p (miRNAs w/seed UAAUGCU)	mature microrna	8.42e-03	3
E2F1	transcription regulator	8.87e-03	8
KITLG	growth factor	9.28e-03	3
CASP8	peptidase	9.42e-03	2
DDIT3	transcription regulator	9.42e-03	2
CBL	transcription regulator	9.42e-03	2

Table 4.5: Top 10 hubs found in the PPIN subnetwork created using the top 30 PD DEGs.

Gene name	Number of First neighbour nodes
YWHAZ	122
YWHAB	62
YWHAG	62
YWHAE	42
YWHAQ	39
SNCA	38
YWHAH	32
TP53	25
AKT1	22
ABL1	19

A PPIN was created to understand relationships among top DEGs at a protein level. From the top 30 DEGs, 21 were mapped to the PPIN and FNN extracted. This subnetwork contains 248 nodes and 912 edges, and included 2 GWAS genes, *SNCA* and *MAPT*. The top 10 hubs, which have the greatest number of first neighbour connections, are shown in table 4.5. Of the top ten hubs, 6 belonged to the 14–3-3 family of proteins, including 14–3-3 zeta (*YWHAZ*) which is connected to 122 other genes in the subnetwork including 6 down and 4 upregulated DEGs. Figure 4.5 shows a subnetwork created using the FNN of the 14–3-3 protein family in the top 30 DEG PPIN.

Of the 69 GWAS genes previously identified, 37 mapped to the PPIN created. The subnetwork created had 331 nodes and 1245 edges that included 45 DEGs, including *SNCA*, *YWHAZ* and *MAPT*. DEGs were over-represented in the GWAS PPI sub-network (hypergeometric test, p-value=1.05e-6). The largest hub of the GWAS gene PPI subnetwork was *MAPT* which had 46 mapped genes, followed by *DLG4* and *SNCA*.

4.4.5 Comparison to Alzheimer's disease

The PD DEGs identified in this study were compared to the 3124 AD DEGs previously found [162]. Between PD and AD, there were 436 DEGs in common (shown at `https://doi.org/10.6084/m9.figshare.7252145`), an overlapping analysis showed that is not just a chance event (OR=4.32, 95% CI 3.79 $\sim$4.93, p-value=$<$2.2e-16, Fisher Exact test). This means around 42% of PD DEGs were found in AD and around 14% of AD DEGs were found in PD. Over 99% (432) of the shared DEGs were differentially expressed

in the same up or down direction. *PIK3R3, LIMK2, CD55* and *MAPT* were the only genes not dysregulated in the same direction between diseases. It is interesting that the majority of DEGs in common between AD and PD were significantly distributed towards downregulation (two-tailed sign test p-value<2.2e-16) as can be seen in table 4.6.

IPA identified 54 affected pathways in PD and 107 pathways in AD, with 27 shared between these two (shown in Appendix table A.1). Interestingly, many of the top pathways in PD were also dysregulated in AD, including Sirtuin Signalling pathway (AD adjusted

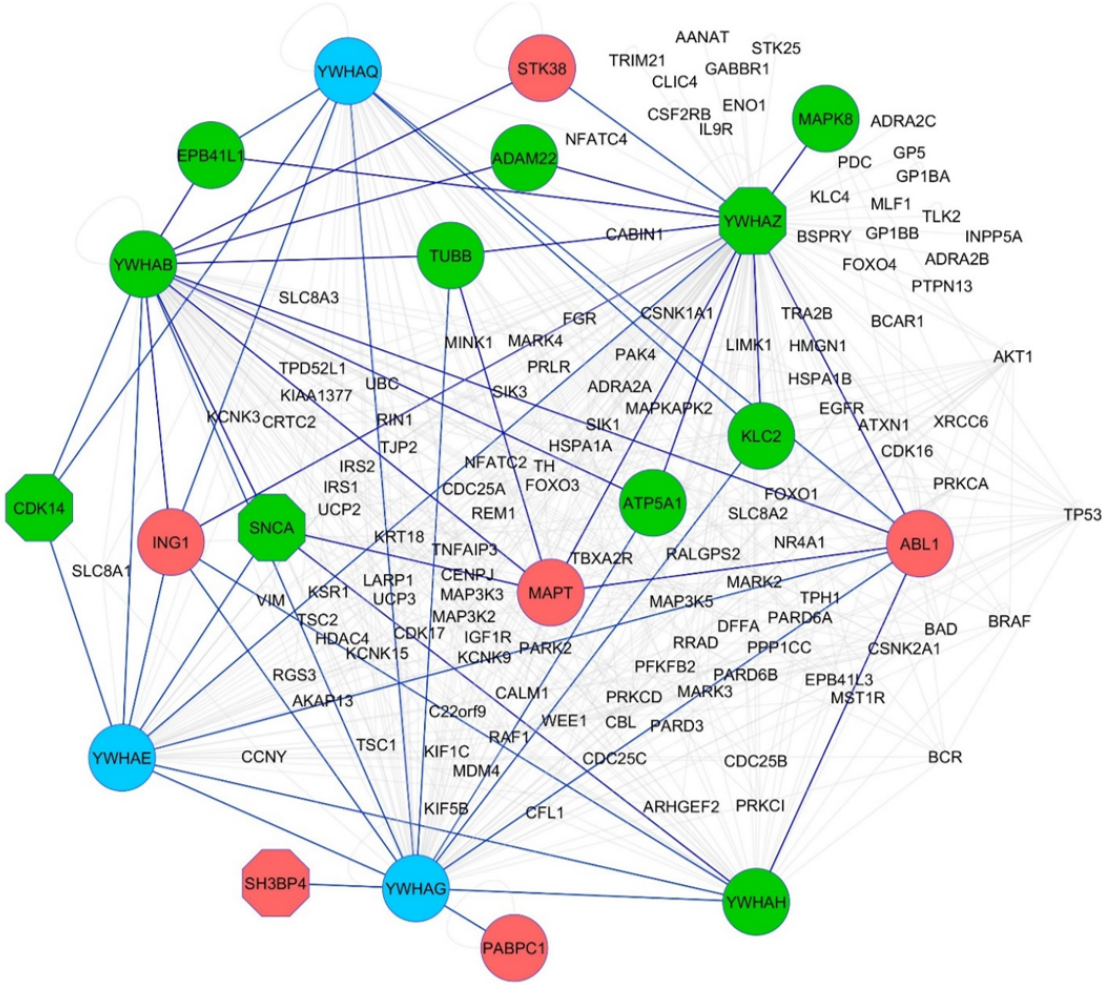


Figure 4.5: Protein-protein interaction subnetwork created using the first neighbour nodes of the 14–3-3 protein family in the DEG PPIN. Six 14–3-3 family genes, *YWHAZ, YWHAB, YWHAG, YWHAE, YWHAQ* and *YWHAH*, were in the top 10 hubs for the subnetwork created from the top 30 DEGs found in this PD meta-analysis. A subnetwork of these 14–3-3 family members and their first neighbours were created. There were 18 DEGs that mapped to this, with red nodes indicating upregulated genes and green nodes indicating downregulated genes. Blue nodes indicate 14–3-3 family members that are not PD DEGs. Octagons denote genes that were in the top 30 DEGs. Blue edges indicate connections between two DEGs. This first neighbour network contains 139 nodes and 539 edges.

Table 4.6: The direction of differential expression between the common DEGs found between AD and PD.

	PD upregulated	**PD downregulated**	**Total**
AD upregulated	114	3	117
AD downregulated	1	318	319
Total	115	321	436

p-value=3.39e-4) and 14–3-3-mediated Signalling (AD adjusted p-value=5.13e-3). The top five pathways identified using DEGs unique to PD were all among the common pathways between AD and PD. In contrast, only two of the top ten pathways identified by AD unique DEGs were also perturbed in PD (HIPPO Signalling and Sirtuin Signalling pathway). Of the top five perturbed pathways for the 2688 AD unique DEGs, neuroinflammation signalling pathway, complement system and NF-κB signalling were not perturbed in PD.

4.5 Discussion

By integrating 126 brain samples from seven microarray gene expression datasets, 1046 DEGs were identified in PD. To my knowledge this is the largest meta-analysis study on microarray SN data about PD. This approach allows inclusion of all the genes across all datasets included in this study. Only 267 out of the 1046 identified DEGs were included in all datasets. If only the common genes were used for meta-analysis, as applied in other previous gene expression meta-analysis about PD [164], it will have introduced many false negative results. This is because potentially interesting genes would not be identified when they are not common between studies. For instance, out of the top 30 identified DEGs, 14 would not have been identified, including *GBE1* [164] and *OPA1* [281] which have been associated with PD in previous studies.

The gene *YWHAZ*, coding for the 14–3-3 zeta protein, was the top DEG and six 14–3-3 family proteins were important hubs in the PPIN. The 14–3-3 protein family has seven isoforms that bind and regulate very diverse signaling proteins, including kinases, phosphatases, and transmembrane receptors [282]. They are highly expressed in brain tissue and are important in development of the nervous system. Previously 14–3-3 proteins

have been implicated in interactions with several proteins associated with PD including α-synuclein, Parkin and LRRK2 [283] and targeting 14–3-3 PPI using small molecules offers a promising strategy for PD and other NDs [284]. 14–3-3 theta phosphorylation at S232 is observed in human PD brains to be pathogenic and contributes to the neurodegenerative process [285]. In Creutzfeldt-Jakob Disease (CJD) phosphorylation levels of 14–3-3 proteins have been used as a diagnostic biomarker clinically [283]. As dysregulation of various 14–3-3 proteins are found in the post-mortem brain, further investigation into the potential of 14–3-3 protein dysregulation and phosphorylation levels as PD biomarkers in CSF and plasma is warranted.

The previous chapter identified neuroinflammation as one of the key pathways in HD, and previous work has identified that the extent of neuroinflammation is greater in PD and AD patients [286]. NF-κB was shown to have a critical role in inflammatory response in HD, and it is a known TF in AD, acting as a key regulator of reactive oxygen species (ROS) production [162]. This can trigger a pro-inflammatory response. In this present study, inflammation pathways and upstream TFs that are pro-inflammatory are not perturbed in PD, suggesting a reduced importance of inflammation in the brain of patients with developed PD in comparison to AD. Degradation of dopamine is a major source of ROS in nigral tissue in PD brains, and late into PD development a lot of the dopamine producing cells are lost, potentially reducing inflammation levels [287]. Previously it has been shown that particular inflammation markers are not present in Parkinson's disease dementia when compared to AD, suggesting that the neuroinflammatory mechanisms in PD and AD differ [288].

Although the DEGs between the two diseases were significantly overlapped, PD had a higher proportion that are also perturbed in AD. In addition, of the top five pathways perturbed in PD all were also perturbed in AD, however of the top 5 pathways perturbed in AD, only one was in PD. This suggests that the processes underlying the two diseases are similar, however this is more apparent with PD. Interestingly, the shared DEGs between PD and AD are almost always differentially expressed in the same up or down direction. This suggests that these genes could represent the crosstalk that is apparent between PD and AD. MAPT is one of four genes not differentially expressed in the same

direction between the two diseases, being downregulated in AD and upregulated in PD. MAPT encodes the tau protein, and tau pathologies are important in both diseases [2, 289]. It has been shown that in three brain regions of AD patients there is a reduction in MAPT expression [290], however for PD it has been proposed that brain regions expressing greater levels of MAPT are more susceptible to tau mediated neurodegeneration [291]. This difference in MAPT and the role of tau pathology in both diseases warrants further investigation as these processes are not greatly understood.

The transcriptional regulator REST was identified by IPA as an upstream regulator of down-regulated PD DEGs. REST is of particular interest in PD as it has been identified to be an important regulator in NDs, particularly in AD and HD [292], and there is established research into the mechanisms of REST and how they are effected in AD [293]. Additionally, REST has been identified using IPA as an upstream regulator of AD brain data in similiar meta-analysis previously [162]. Repressor element 1-silencing transcription factor (*REST*) has been implicated as an important regulator of neurons in the normal aging brain, closely correlating with cognitive longevity [293]. In AD and other dementias, REST is lost from the nucleus and is found with misfolded proteins in autophagosomes. In cell models of PD, abnormal levels of the REST neuronal splice form REST4 have been implicated in pathology of PD [294]. It has been suggested that overexpression of α-synuclein affects the histone maker distribution on REST complex associated genes and results in repression of the *SNAP25* and *L1CAM* genes in both Drosophila and cell line models [295]. Reduction in these genes has been implicated in contributing to synaptic dysfunction in PD [295]. Here both genes have shown to be downregulated DEGs in PD supporting this mechanism underlying human PD pathogenesis.

The Sirtuin Signalling pathway was revealed to be perturbed in AD and PD and modulating their activities can alter the course of both diseases in both cell and animal models [296]. In PD SIRT1 and SIRT3 have protective effects against degeneration of SN neurons by neurotoxins, whereas activity of SIRT2 worsens the degeneration [296]. It is likely that SIRT1 and SIRT3 modulate homeostasis of mitochondria and anti-oxidative mechanisms, whereas activity of SIRT2 could result in adverse microtubule dynamics that disrupt clearance of toxic waste including Lewy bodies. In AD, the pan-sirtuin activator resveratrol

has been shown to be safe, well-tolerated, and alter the trajectory of some biomarkers in a clinical trial [297]. Further research is needed to understand the therapeutic potential of sirtuins [296].

The SN was chosen as the brain region of interest in this study as neuron degeneration in this region is a hallmark of PD and it is the region with the most data for the meta-analysis [257]. SN microarray study GSE54282 was excluded from this meta-analysis due to low sample size and E-MEXP-1416 due to high variance in the data. There are also many studies using SN dopaminergic neurons, however including a number of these could lead the gene expression data to reflect these neuron types instead of the whole SN.

Although RNA-seq has demonstrated itself as a superior approach [298], there is not much data available for PD, although there is likely going to be further applications in the future. Microarrays are still very useful tools for measuring the gene expression and their power is further increased by using meta-analysis. Our microarray data has correlated gene expression values to the healthy SN RNA-seq data in the GTEx database [275], demonstrating that the microarray expression data used in this study has the similar quality to that of previous RNA-seq data.

For PD there has been a limited application of RNA-seq to identify DEGs, in fact for the analysis of the SN only one RNA-seq study has been completed [262]. There are minimal similarities between the results of this RNA-seq analysis and our meta-analysis results. Only 70 of their 2961 identified DEGs are identified in our results, and only three of our top 30 DEGs (*SLC18A2, FGF13, AASDHPPT*) are identified in their results. However, pathways associated with oxidative phosphorylation, cardiac hypertrophy and the cytoskeleton were shared. A possible explanation for this is the very low power of the RNA-seq study, which only used three control and three PD samples and the fact that these samples were not age and gender matched. This is particularly important as age and gender are some of the largest risk factors for PD. The control samples had an average age of 87.3 ($\pm$5.5) and were all females, and the PD samples had an average age of 79.0 ($\pm$5.6) and only one sample was female.

A limitation of this study is that the SN is affected early in PD development, and by

the time symptoms manifest much of the SN can be lost. This means our results reflect the perturbed genes and pathways present once the disease has been established, and not the changes that take place that lead to PD. To investigate early changes in disease more accessible tissues, such as blood and cerebrospinal fluid, would have to be investigated. Currently, there is no reliable way of diagnosing PD before it has had a substantial effect. As a result, investigating the perturbed pathways at this point in the disease would be difficult without development of effective early diagnosis biomarkers. Nonetheless identifying genes and pathways perturbed in the later stages of the disease can still help identify therapeutically important information and compare to similarly late stages of AD.

A large limitation in this meta-analysis is the limited number of PD samples. As only 69 PD and 57 control samples are included in this study, the statistical power would be lower than that of the previous meta-analysis for AD which included 450 AD and 212 control samples [162]. This relatively low sample size could also introduce false positive and false negative DEGs and pathways, nevertheless, meta-analysis will outperform individual microarray studies. Moreover, the PPINs would be best enriched by proteomics data of PD if such datasets were publicly available.

4.6 Conclusion

In conclusion, this meta-analysis is the largest study of its type in PD SN tissue to date. REST, which has been shown to perturb Wnt signalling [293], is highlighted as an important upstream regulator in PD and AD. The results reveal the importance of *YWHAZ* and 14–3-3 proteins in PD, through their down regulation, involvement in perturbed pathways and as hubs in PPIN. PD and AD are demonstrated to share a significant number of DEGs that are differentially expressed in the same direction and perturbed pathways that indicate some novel shared pathogenesis between the two diseases. These insights suggest several new areas for mechanistic research into PD and cross-talk between AD and PD.

Chapter 5

Network analysis to identify key dysregulated processes and hub genes in neurodegenerative diseases

5.1 Abstract

This chapter describes the largest network analysis of PD and AD based on gene expression in blood to date. In addition to identifying the deregulated processes that underpin these diseases, it aims to identify if these processes are reflected in blood gene expression data. Gene interaction networks for AD and PD transcriptomics data were built and modules that were not preserved between disease and healthy control networks were analysed. Within these non-preserved modules, important hub genes and transcription factors in these modules were identified. A module in the PD network associated with insulin resistance was not preserved in healthy control networks, and *HDAC6* was identified as a hub gene in this module. In AD, the AD module associated with regulation of lipolysis in adipocytes and neuroactive ligand-receptor interaction was not preserved in healthy and mild cognitive impairment networks and the key hubs TRPC5 and BRAP identified as potential targets for therapeutic treatments of AD. This research expands on previous work demonstrating that PD and AD share common disrupted genetics. In addition it identifies novel pathways, hub genes and transcription factors that may be new areas for

mechanistic research and important targets in both diseases.

The work in this chapter has been published in Aging in 2020 [299].

5.2 Background

As discussed in the previous chapter, there are many common characteristics that are shared between PD and AD. They share significant common DEGs, disturbed pathways including the sirtuin signaling pathway, and REST is an important upstream regulator in both diseases.

The analysis of gene co-expression networks can uncover considerable information that differential expression analysis cannot [165]. WGCNA has been used to find strong evidence for mitochondrial dysfunction and chronic low grade innate immune response in AD [170]. In addition, Chatterjee *et al.* [300] identified 11 hub genes by using WGCNA in frontal cortex and SN brain samples of PD patients.

Blood gene expression data has been used previously to reflect changes that take place in brain tissue in NDs. Microarray has been used to investigate blood gene expression differences between rapid and slow progression PD [301], with rapid progression being classed as patients with postural instability. The expression of the top seven DEGs (*RAD18*, *ABCA1*, *FOXP1*, *AGAP1*, *PPAT*, *NUB1*, *AKT2*, *ABI2*, *APC*, *FHL1*) were investigated in a dopaminergic-like cell model of PD [301]. Six of the seven DEGs were differentially expressed in cell models suggesting that dysregulation in blood reflects cell models of dopaminergic cell death in the brain.

Lunnon *et al.* [170] identified DEGs and built gene co-expression networks in blood AD datasets to see if dysregulated brain pathways are reflected in blood. They identify gene enrichment in mitochondrial dysfunction and inflammation pathways, known hallmarks of AD [302] and shown consistently by brain transcriptomics studies [162, 303, 304]. Dysfunction in mitochondrial respiratory chain activity increases cellular stress and increases the production of ROS which leads to cell apoptosis and ND. Immune response is considered another important hallmark of AD [305], and as immune cells, including lymphocytes and macrophages, are present in blood it is no surprise that it is dysregulated

in blood gene expression data. The gene *SORL1*, which interacts with APP reducing secretase activity and thus reducing Aβ levels [306], was identified as a hub within the blood network module associated with immune response in AD patients.

Li *et al.* [52] support use of blood gene expression data to investigate changes in the bran, as they identify a significant overlap in DEGs between blood data and various brain regions, including prefrontal cortex, superior temporal gyrus and the inferior temporal gyrus. Common pathways including mitochondrial dysfunction and oxidative phosphorylation were identified in concordance with previous studies [170].

To date there have been no studies investigating PD and AD using gene expression network simultaneously to reveal potential shared biological process and pathology.Here, gene co-expression networks are analysed based on PD and AD blood microarray data and common genetic networks between both diseases identified. The analysis workflow is illustrated in figure 5.1. Compared to brain tissues, blood tissue is easier to access from patients with ND, and publicly available AD and PD blood datasets have a large enough sample size to construct reliable and robust networks. This network analysis expands on standard WGCNA and hub detection approach which can robustly find key processes and genes that are associated with both PD and AD.

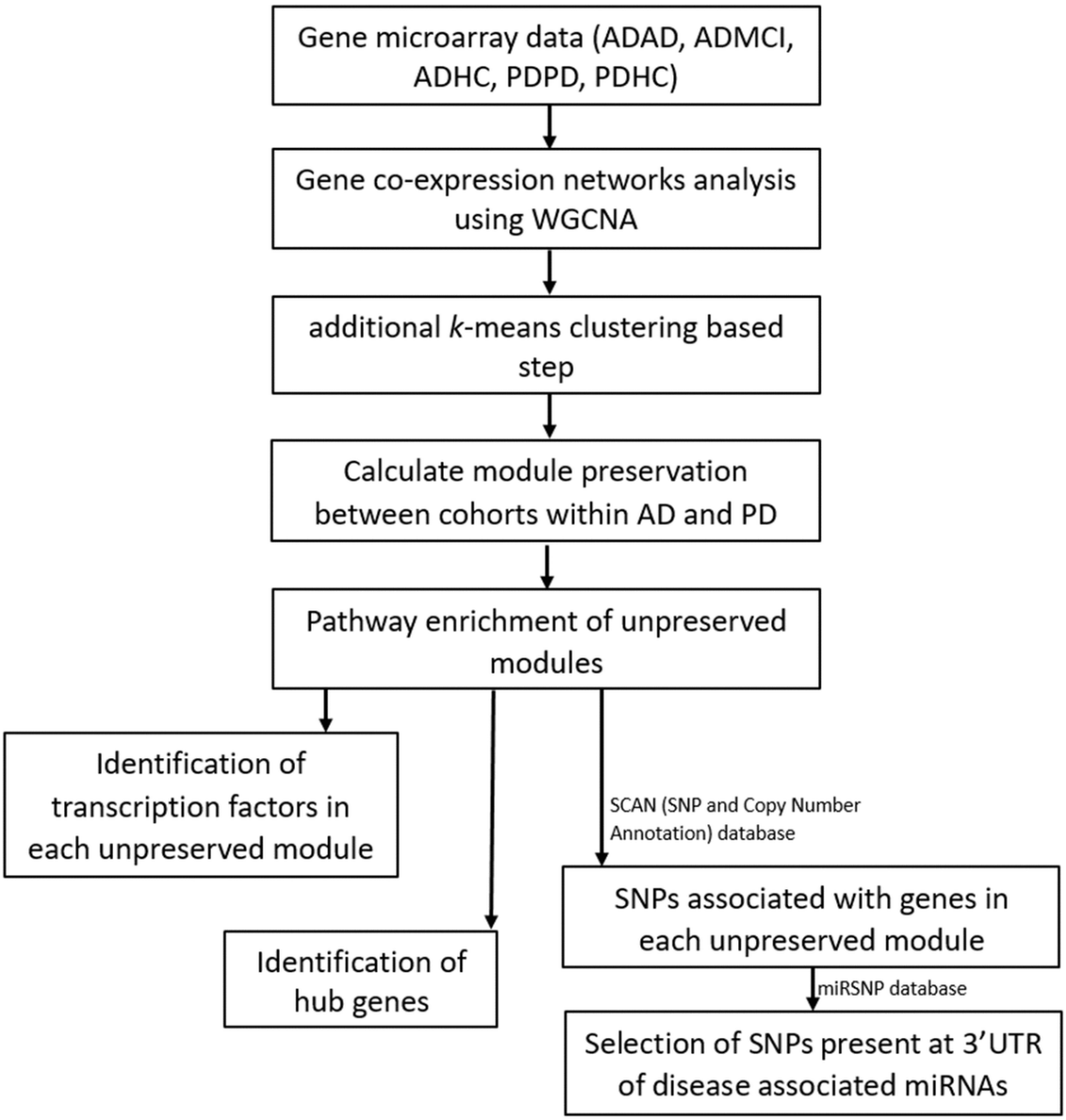


Figure 5.1: **Workflow of network analysis.** Filtered and normalized microarray data were separated into five datasets: AD disease (ADAD), healthy control (ADHC) and mild cognitive impairment (ADMCI) data from the AD dataset, and the PD disease (PDPD) and healthy control (PDHC) data from the PD dataset. On each dataset gene co-expression networks analysis was performed using the WGCNA R package [166]. An additional k-means correction step to reduce number of misplaced genes [172] was then performed and module preservation between cohorts within AD and PD was found using NetRep (v.1.2.1) [307]. The pathways associated with non-preserved modules were then found using the Enrichr web tool [229, 228] and hub genes and transcription factors in these non-preserved modules identified. The SCAN (single nucleotide polymorphism (SNP) and Copy number ANnotation) database [308] was used to find SNPs associated with the genes in each non-preserved module and these SNPs used to search the MiRSNP database to find the SNPs at 3' UTR of disease associated miRNAs.

5.3 Materials and Methods

5.3.1 Data preparation for PD and AD blood datasets

The publicly available peripheral venous whole blood dataset comprising 205 PD and 233 control samples was downloaded from the NCBI GEO database (`http://www.ncbi.nlm.nih.gov/geo/`) with accession identifier GSE99039. This dataset is the largest of its type and has a sample size enough to run WGCNA and reliably find hub genes [166]. Samples with known PD mutation genes (*Parkin, DJ-1 and PINK1, ATP13A2, LRRK2, SNCA*) were removed to reduce biases introduced by these genes (see section), and outlier samples were detected and removed based on box and density plots of probe intensities. This removed a total of one PD and three healthy control (HC) samples, leaving 204 PD and 230 HC samples. Data was then RMA normalized using the affy R package [273]. Samples missing gender information (35 samples) were assigned sex by using the massiR R package [309] which uses the information from microarray probes that represent genes in Y chromosome to perform k-medoids clustering to classify the samples into male and female groups. A probe-variation threshold of 4 was selected by inspecting a probe-variation plot (figure 5.2) to select the Y chromosome probes to be used in the sex classification process.

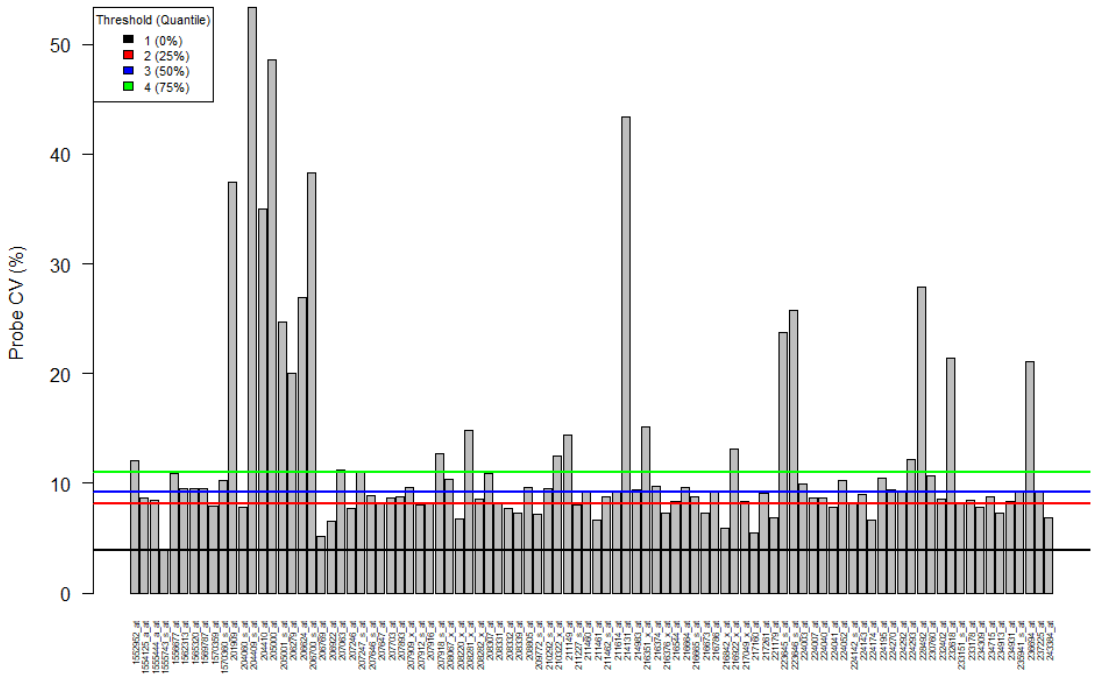


Figure 5.2: **The probe variation plot used to determine which genes to use in massiR R package** [309]. A threshold of 4 was selected as it encompassed the genes with the highest variation and ignores genes with low variation that may be useful in classifying samples

The ComBat function in the sva R package [277] was used to control the effect of gender and running batch of the samples. After this, control probes and those without Entrez gene annotation were removed. For any genes that mapped to multiple probes, the probe with the highest MAD was kept. MAD was used as, similarly to inter-quartile range, the probe with the highest MAD has the greatest variability and so likely has more information [224]. Finally, the bottom 5% probes by average expression values across all samples were removed.

For AD, the two independent peripheral venous whole blood datasets GSE63060 and GSE63061, from the AddNeuroMed Cohort [310], were used to construct the blood gene expression networks. As these two datasets were from the same cohort study and sample collection and analysis was carried out using the same methodologies, except using different biological samples and microarray platforms, they can be merged to produce a larger dataset that can improve the power of the study. The two normalized datasets (generated

Table 5.1: Information on number of samples, sex and age of samples in datasets.

GEO dataset		No. Samples	Sex (male/female)	Mean Age (± SD)
GSE99039	PD	204	97/107	NA
	HC	230	150/80	NA
	All	434	247/187	NA
GSE63060 +	AD	245	166/79	76.5 (± 6.6)
GSE63061	MCI	142	79/66	74.9(± 6.3)
	HC	182	110/72	73.6 (± 6.3)
	All	569	352/217	75.2 (± 6.5)

by different Illumina platforms) were merged using the inSilicoMerging R package [311], which removes the batch effects between these two, as has been done previously [52].

Patients of Western European and Caucasian ethnicity were extracted from the merged dataset leaving a total of 245 AD, 142 mild cognitive impairment (MCI) and 182 HC to reduce any potential genetic impact that ethnicity may have on AD. The effect of the age and gender were controlled for using the ComBat function in the sva R package [277]. As with the PD data, control probes and those without Entrez gene annotation were removed and for any genes that mapped to multiple probes, the probe with the highest MAD was kept. Finally, the bottom 5% probes by average expression values across samples were removed. Information on number of samples, gender and age of samples is shown in table 5.1.

5.3.2 PD blood and brain DEG overlap

To see if there was a significant overlap between PD gene expression in blood and brain as has been shown previously in AD [52], the PD blood data was compared to DEGs previously identified in PD substantia nigra [256]. Using the normalised and filtered PD data, DEGs were identified by applying limma with gender and running batch adjusted. Slightly stringent nominal Pvalue$<$0.01 was used for significance as only one DEG could pass multiple testing (FDR corrected Pvalue$<$0.05).

5.3.3 Gene co-expression network construction

Gene co-expression networks were built using the WGCNA R package [166] as discussed in section 2.5.The matrix of pairwise correlations were raised to a soft-thresholding power to achieve a scale-free topology R^2 of 0.85. Initial module assignments of genes was determined using a dynamic tree-cutting algorithm (cutreeHybrid, using default parameters except deepSplit of 3, minModuleSize of 10 and mergeCutHeight of 0.05) [166]. An additional k-means clustering step was applied to improve the results of the hierarchical clustering in WGCNA as proposed by Botía *et al* [172] which has been reported to be able to reduce the number of misplaced genes and improve the enrichment of GO pathway terms. All analysis was conducted in R3.5.2 [272].

5.3.4 Calculation of module preservation

NetRep (v1.2.1) [307] was applied to identify modules that are not preserved between conditions within datasets using a permutation test procedure on seven module preservation statistics. This was permuted 10,000 times. The 'alternative' parameter was set to 'less' to test whether each module preservation statistic is smaller than expected by chance in order to identify these non-preserved modules which are extremely different in the two networks. If all seven module preservation statistics had a Pvalue < 0.05 then that module was determined to be significantly non-preserved between conditions.

5.3.5 Pathway enrichment analysis

GO and KEGG pathway enrichment analysis (KEGG 2019) using the Enrichr web tool [229, 228] was performed to identify the biological pathways that the modules represented. Pathways and GO terms with a Pvalue < 0.05 were considered significant.

5.3.6 Hub gene identification

Generally, detecting hub genes in co-expression networks has been done using module membership (MM), which is the correlation of a gene to its eigengene (the first principle component calculated using the expression data of genes in each module) [312]. Be-

tweenness centrality (BC) of a gene is the number of shortest paths connecting all gene pairs that pass through that gene [313], and genes with high BC were considered as 'high traffic'.

Here hub detection has been expanded to include multiple other hub detection methods frequently used in network analysis. In addition to MM and BC, closeness centrality [314], Kleinberg's hub centrality score [315] and the PageRank algorithm [316] are used, which will reduce the chance of missing any important hub genes that regulate the network that may be missed by applying individual methods. Genes with high closeness centrality scores have the shortest path to all other genes in the module and are placed to influence the entire network quickly [314]. PageRank emphasizes nodes that are connected to other nodes with high Pagerank scores [316]. Kleinberg's hub centrality score [315] is similar to the PageRank algorithm, however, the small differences between the two widens the net for identifying important hubs.

A novel hub detection permutation test was developed to obtain Pvalues for each hub detection store and determine if they are statistically significant. Briefly, the gene ID labels on the adjacency matrix were randomly re-labelled and hub score recalculated 1000 times to obtain a statistical distribution. The Pvalue was calculated by dividing the number of recalculated permutation hub scores that are higher than the observed hub score in the original network by the number of permutations. Genes were considered significant hubs if any hub scores had a Pvalue < 0.01. This was performed for all modules not preserved between PD and HCs in the PD dataset, and the modules not preserved between any of the AD, MCI and HCs networks in the AD dataset. BC, closeness centrality, PageRank and Kleinberg's hub centrality scores were calculated using the igraph R package with default settings without normalization [317]. The R code used for the novel hub detection test is available at `http://dx.doi.org/10.5281/zenodo.3686007`.

5.3.7 Identifying transcription factors

To identify TFs that potentially regulate each module, the ENCODE and ChEA consensus TFs from ChIP-X database was used through the Enrichr web tool [229, 228]. TFs with a Pvalue < 0.01 were considered significant. If a TF was found significant in both

ENCODE and ChEA then the lower Pvalue was assigned to the TF.

5.3.8 SNP and microRNA analysis of significant WGCNA modules

A two-tailed Fisher's exact test was used to test the hypothesis that non-preserved modules were more likely to contain Genome Wide Association Studies (GWAS) identified genes than preserved modules. The risk loci for PD and AD were from recent GWAS, between which only one GWAS gene was shared (*KAT8*) [280, 318].

Further insight into SNPs associated with non-preserved modules was gained using a similar methodology to Chatterjee *et al.* [300]. The SCAN database [308] was used to find all SNPs that have been shown to predict the expression of each gene within non-preserved modules. For each non-preserved module, only SNPs that predicted gene expression with Pvalues $<$ 1.0e-4 and frequency $>$ 0.10 within the CEU human samples of European descent were selected.

Previous studies have revealed that differential expression of miRNAs were associated with PD [319] and AD [320]. In addition, SNPs have been identified as disease prognostic markers by association to miRNAs [321]. SNPs found to be associated with genes from the PD related modules were used to search the MirSNP [322] database in order to find which SNPs were associated with the 83 experimentally confirmed PD related miRNAs in the HMDD v3.0 database [323]. The same process was done for genes within the AD related modules and the 57 experimentally confirmed AD related miRNAs in the HMDD v3.0 database. The MirSNP database identified the SNPs that are present at the 3' untranslated region of miRNA target sites, and so narrowed down the selection of SNPs to those that likely effect known miRNAs associated with the disease.

5.3.9 Comparison of PD and AD results

The processes associated with non-preserved modules in AD and PD were compared to see if any processes were similar between diseases. Hub genes and TFs identified in non-preserved modules were also compared between AD and PD to see if any were shared. In addition, the hypothesis that AD and PD share SNPs associated with disease related

miRNAs identified in non-preserved modules was tested.

5.4 Results

5.4.1 Gene co-expression network construction

After quality control, there were 19176 genes in the PD dataset which included 204 PD and 230 HC samples, meanwhile there were 13661 genes in the AD dataset which included 245 AD, 142 MCI and 182 HC samples. WGCNA [166] was applied to build the networks and the soft threshold power to define the adjacency matrix of each dataset based on approximate scale-free topology R^2 of 0.85 was selected (Figure 5.3). In this method, highly correlated nodes are placed into a single module or cluster which are thought to be regulated by similar TFs and represent certain biological processes. These networks were constructed for the ADAD, ADHC and ADMCI data from the AD dataset, and the PDPD and PDHC data from the PD dataset separately. In total, there were 27, 54, 29, 32 and 58 modules in PDPD, PDHC, ADAD, ADMCI, ADHC networks respectively.

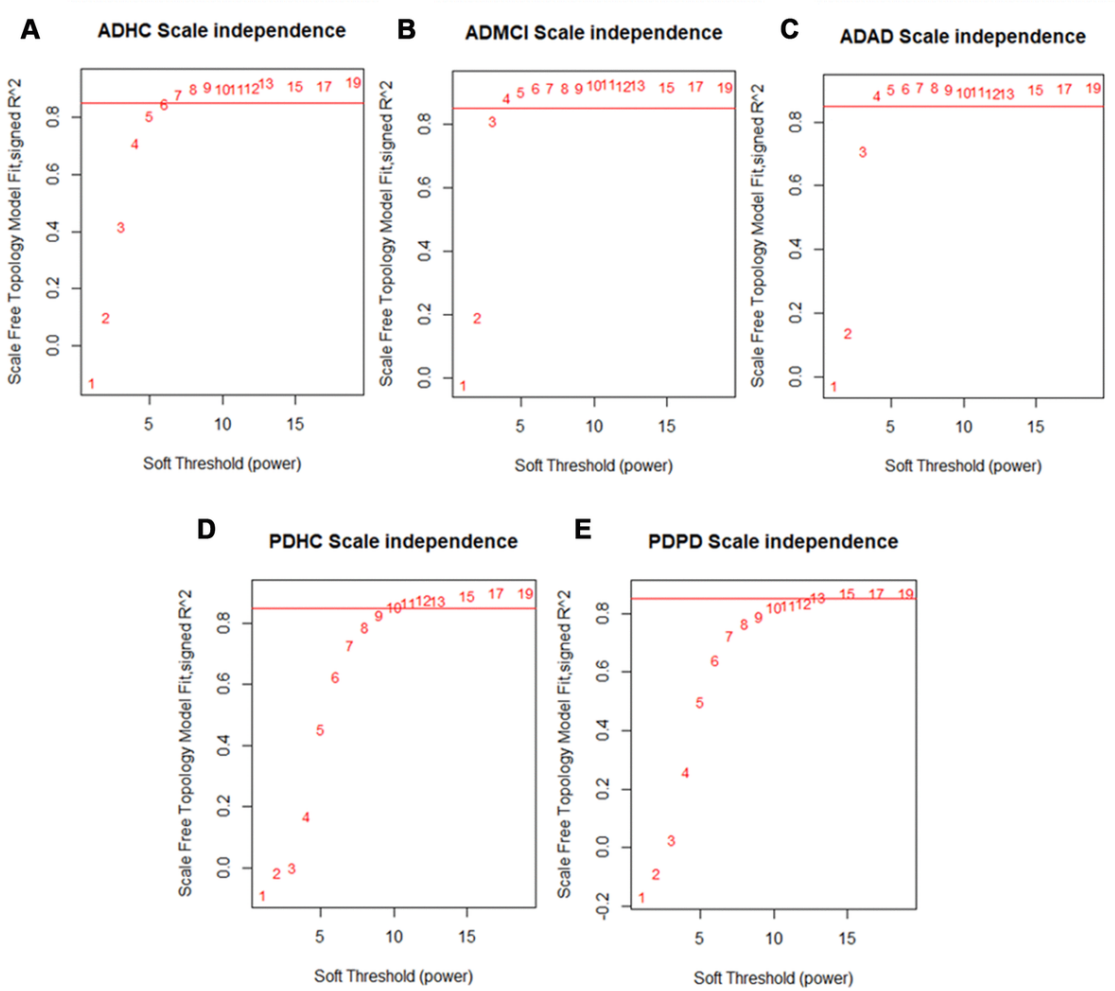


Figure 5.3: **Scale free network topology (signed R^2) for different soft-thresholding powers of data.** A soft thresholding power that achieved a scale-free topology of R^2 of 0.85 was chosen to define approximate scale-free topology. (A) ADHC data achieved approximate scale-free topology at a soft thresholding power of 6 and the (B) ADMCI and (C) ADAD data at a soft thresholding power of 4. The (D) PDHC data reached approximate scale-free topology at a soft thresholding power of 10 and (E) PDPD data at a soft thresholding power of 13.

5.4.2 PD blood and brain DEG overlap

In the PD blood dataset 360 DEGs were identified (nominal Pvalue<0.01, available at `https://doi.org/10.6084/m9.figshare.14512116`) and compared to the DEGs identified in the meta-analysis study about PD in substantia nigra region in chapter 4. An overlap of 21 genes were found including *LRRN3*, *BASP1* and *TPM3*. However, a Fisher Exact test was not significant for the overlap showing that this was likely by chance (OR = 1.08, 95% CI 0.65 1.72, Pvalue = 0.72, Fisher Exact test).

5.4.3 Identification of non-preserved modules

In the network analysis, if the relationships and correlation structure between nodes composing each module were not replicated, then they were considered non-preserved. In the case of healthy and disease networks, non-preserved modules suggested the expression pattern and regulation of the genes in these modules vary between disease and healthy conditions. On the other hand, modules preserved between disease and healthy networks represented processes that are not affected by disease status. Non-preserved modules which may help to reveal the disease mechanism and so are the main focus of results.

Table 5.2 shows the non-preserved modules between PDHC and PDPD networks and the biological processes associated with these modules. Three of the 54 modules in the PDPD network were not preserved in PDHC network, and one of those 27 PDHC modules was not preserved in the PDPD network. The GO and KEGG terms that were significantly enriched within non-preserved modules (Pvalue $<$0.01) were found using the Enrichr web tool [229, 228]. The PDPD salmon module was found to be associated with insulin signaling (KEGG pathway, Pvalue = 0.0030, 7/108 overlap). The PDPD darkseagreen4 module was found to be associated with antigen processing and presentation (KEGG pathway, Pvalue = 5.38e-16, overlap = 14/77) and natural killer cell mediated cytotoxicity (KEGG pathway, Pvalue = 2.94e-15, overlap = 10/41).

Table 5.3 shows the non-preserved modules between the ADHC, ADMCI and ADAD networks. Of the 29 ADAD modules, one was not preserved in both ADHC and ADMCI networks. In addition, one of the 32 ADMCI modules was not preserved in ADAD and ADHC networks. Moreover, three of the 58 ADHC modules were not preserved in both ADAD and ADMCI networks and one non-preserved in ADMCI networks. The ADAD blue module was not preserved in ADHC and ADMCI networks and was associated with regulation of lipolysis in adipocytes (KEGG pathway, Pvalue = 6.24e-4, overlap = 10/55) and neuroactive ligand-receptor interaction (KEGG pathway, Pvalue = 0.005070, overlap = 30/338). The ADHC darkolivegreen module was associated with sensory perception (GO biological process, Pvalue = 1.83e-4, overlap = 8/55).

Table 5.2: List of non-preserved modules found between PD and HC.

Module Colour	Pvalue of NetRep	Processes associated with module found using Enrichr	No. genes in module
PD modules not preserved in HC			
Darkseagreen4	9.99e-5	Antigen processing and presentation, Natural killer cell mediated cytotoxicity, cellular defense response, regulation of immune response	150
Navajowhite2	9.99e-5	Cellular response to misfolded protein	150
Salmon	9.99e-5	Insulin resistance, regulation of protein homooligomerization	351
HC modules not preserved in PD			
Purple	9.99e-5	Antigen processing and presentation, VEGF signaling pathway, regulation of intracellular transport	606

5.4.4 Identifying hub genes

Hubs are genes that are highly interconnected or important within a module and likely have functional significance [324]. Hubs have a role in maintaining the structure of the gene network of the module and the biological processes associated with the module. In this study, hub genes were identified using five approaches: BC, PageRank, MM, closeness centrality and Kleinberg's centrality. Any gene with a Pvalue$<$0.01 in any hub detection method was considered a significant hub gene. Using multiple methods for identifying hubs allowed for hub identification that may otherwise have been missed by use of just one method. To demonstrate hub score distribution, figure 5.4A shows an example of betweenness hub score distribution across all genes in the PDPD darkseagreen4 module which was non-preserved in PDHC network and the (figure 5.4B) distribution of the significant *GINS2* (Pvalue = 0.005) BC scores across the 1000 iterations of the hub permutation test.

Table 5.3: List of non-preserved modules found between AD and HC.

Module Colour	**Pvalue of NetRep**	**Processes associated with module found using Enrichr**	**No. genes in module**
AD modules not preserved in HC			
Blue	9.99e-5	Regulation of lipolysis in adipocytes, Neuroactive ligand-receptor interaction, detection of chemical stimulus involved in sensory perception of smell, extracellular matrix organization	1076
AD modules not preserved in MCI			
Blue	9.99e-5	Regulation of lipolysis in adipocytes, Neuroactive ligand-receptor interaction, detection of chemical stimulus involved in sensory perception of smell, extracellular matrix organization	1076
MCI modules not preserved in AD			
Sienna3	8.59e-3	Regulation of lipolysis in adipocytes, axonal fasciculation, hippo signaling	770
MCI modules not preserved in HC			
Sienna3	9.99e-5	Regulation of lipolysis in adipocytes, axonal fasciculation, hippo signaling	770
HC modules not preserved in AD			
Darkolivegreen	9.99e-5	sensory perception, regulation of potassium ion transmembrane transport	584
Darkorange2	0.011	Peroxisome, amide transport	248
Skyblue	0.015	establishment of epithelial cell polarity	187
HC modules not preserved in MCI			
Darkolivegreen	9.99e-5	sensory perception, regulation of potassium ion transmembrane transport	584
Red	9.99e-5	Regulation of lipolysis in adipocytes, bicellular tight junction assembly	704
Darkorange2	2.99e-4	Peroxisome, amide transport	248
Skyblue	0.022	establishment of epithelial cell polarity	187

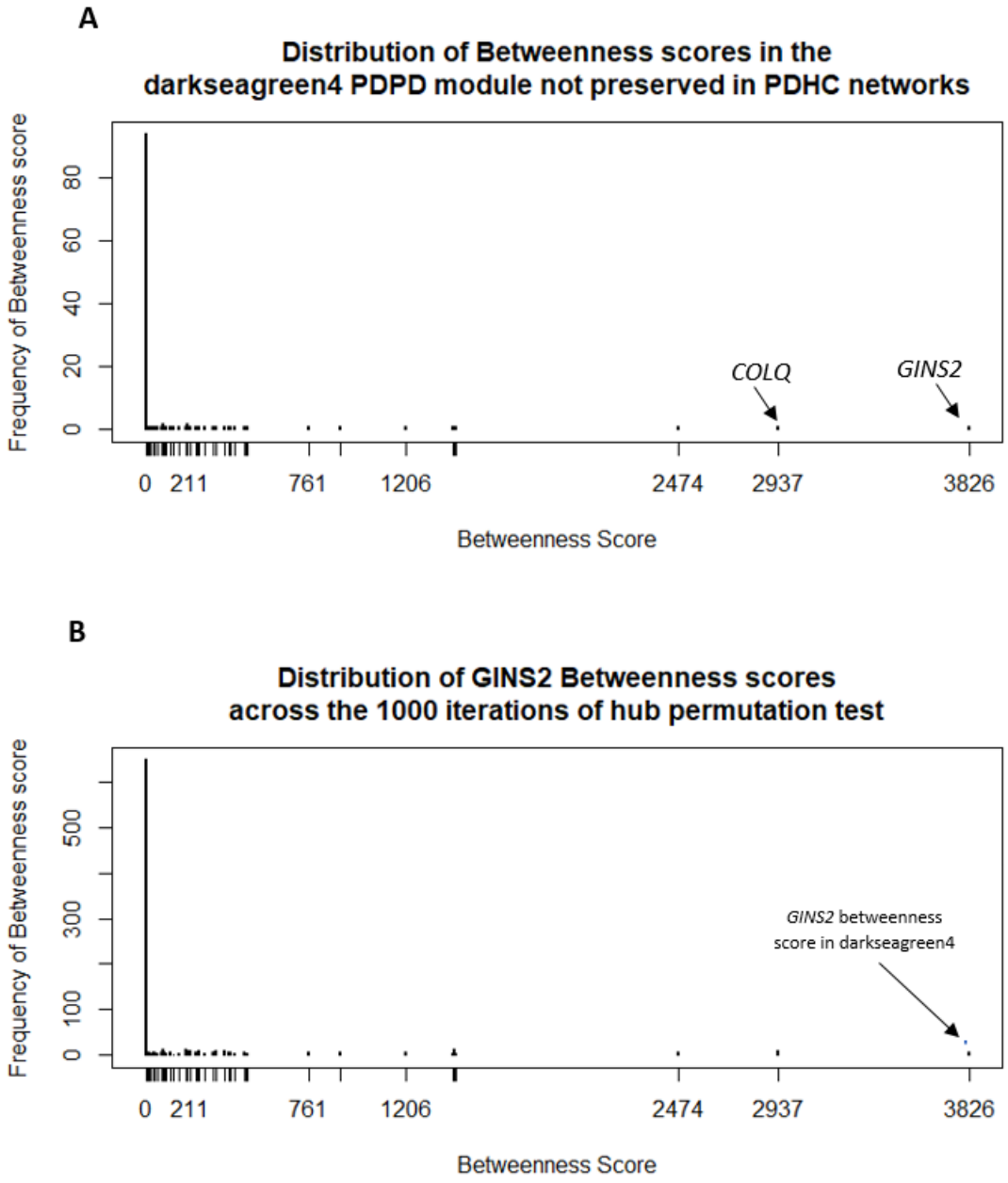


Figure 5.4: **An example of hub score distribution in networks.** (A) The distribution of betweenness scores for each gene in the darkseagreen4 module. Many genes have a betweenness score of 0 indicating they do not act as hubs in regard to betweenness in this module. After the hub permutation test, one gene was found to be significant (*GINS2*, Pvalue = 0.005). (B) The distribution of betweenness scores for *GINS2* over the 1000 iterations of the hub permutation test. The betweenness score of *GINS2* in the original darkseagreen4 module network is highlighted.

In modules not preserved between the PDPD and PDHC networks 34 hubs were identified (shown in Appendix Table B.1) and 92 hubs in the non-preserved modules between ADAD, ADMCI and ADHC networks (shown in Appendix Table B.2). It was expected that larger modules may have more hubs than smaller ones, for example the PDHC pur-

ple module contained 606 genes, of which 17 were found to be hubs (e.g. *FAM110C, PAK4, NEB*), and the smaller salmon PDPD module contained 351 genes, of which only 10 were hubs (e.g. *HDAC6, TYSND1*). The PD salmon module was associated with insulin resistance and was not preserved in PDHC network shown in Figure 5.5A, where hub genes are highlighted. Interestingly, it includes *HDAC6* which has been shown to influence tau phosphorylation and autophagic flux in AD [325]. The blue AD module which was associated with regulation of lipolysis in adipocytes and neuroactive ligand-receptor interaction and was not preserved in ADMCI and ADHC networks (Figure 5.5B) which included *TRPC5* and *BRAP* as hub genes. Networks were visualized in Gephi [326].

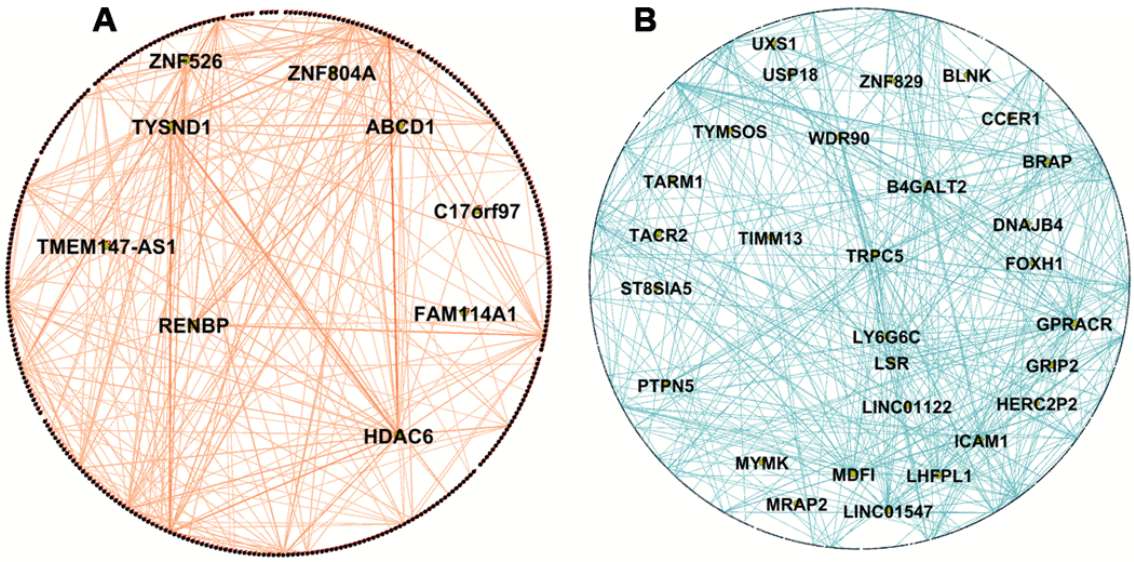


Figure 5.5: **Network visualization of PD and AD modules.** (A) Visualization of WGCNA network connections of the PDPD salmon network module found to be associated with insulin resistance and not preserved in the PDHC network. It shows network connections whose adjacency is above 0.2, including all 351 nodes and 595 of 61776 edges. (B) Visualization of WGCNA network connections of the ADAD blue module found to be associated with regulation of lipolysis in adipocytes and neuroactive ligand-receptor interaction and not preserved in ADHC and ADMCI networks. It shows network connections whose adjacency is above 0.55, including all 1076 nodes and 1458 of 1157776 edges. Hub genes are in the center of the network and are labelled with names. Networks visualized in Gephi [326].

5.4.5 Identifying transcription factors (TFs)

Genes that are clustered together by WGNCA likely are regulated in a similar way, thus which TFs potentially regulate the gene expression of each module were identified. The

Table 5.4: Significant TFs (Pvalue<0.01) associated with each non-preserved module between PD and healthy control networks found using Enrichr (ENCODE and ChEA Consensus TFs from ChIP-X) [229, 228].

Module Colour	Significant TFs	P-value	Gene overlap
PD modules not preserved in HC			
Darkseagreen4	FOXM1	4.004e-08	9/95
	E2F4	8.131e-08	21/710
Navajowhite2	RUNX1	0.008305	18/1294
Salmon	FOXM1	0.006578	6/95
HC modules not preserved in PD			
Purple	SIX5	0.0001626	55/1094
	ZBTB7A	0.0002814	94/2184
	SRF	0.0008434	20/299
	CREB1	0.001402	64/1444
	NFYB	0.004818	138/3715
	PBX3	0.007364	54/1269

TFs that potentially regulate each non-preserved module (Pvalue<0.01) were identified by using ENCODE and ChEA consensus TFs from the ChIP-X database by using the Enrichr web tool [229, 228]. A total of four TFs that regulated at least one of the three PDPD modules were identified, including FOXM1 which regulated 6 genes in the salmon modules (Pvalue = 0.0066) and 9 in the darkseagreen4 module (Pvalue = 4.00e-08). Within one PDHC module, there were a total of six TFs, including CREB1 which regulated 64 genes in the purple module (Pvalue = 0.001402). Table 5.4 shows the significant TFs found in modules that were not preserved between PD and HC networks.

Two TFs were identified (SUZ12, EZH2) regulating non-preserved ADAD modules, and one TF (SUZ12) regulating 115 genes in the ADMCI sienna3 module (Pvalue = 8.24e-10). Furthermore, 18 TFs that regulated at least one of four non-preserved ADHC modules were identified. This included REST which regulated 20 genes in the darkolivegreen (Pvalue = 0.0092) and SUZ12 which regulated 68 genes in the darkolivegreen (Pvalue = 0.0039) and 107 genes in the red module (Pvalue = 1.21e-09). In addition, CREB1 regulated 29 genes in the ADHC darkorange2 module (Pvalue = 0.007005). Table 5.5 shows the significant TFs for modules that were not preserved between ADAD, ADMCI and ADHC.

Table 5.5: Significant TFs (Pvalue<0.01) associated with each non-preserved module between AD, MCI and healthy control networks found using Enrichr (ENCODE and ChEA Consensus TFs from ChIP-X) [229, 228].

Module Colour	**Significant TFs**	**P-value**	**Gene overlap**
AD modules not preserved in HC and MCI			
Blue	SUZ12	3.36e-10	150/1684
	EZH2	0.0004579	26/237
MCI modules not preserved in AD and HC			
Sienna3	SUZ12	8.24e-10	115/1684
HC modules not preserved in AD and MCI			
Darkolivegreen	SUZ12	0.00392	68/1684
	REST	0.009205	20/383
Darkorange2	IRF3	0.000002884	24/663
	SP2	0.000006359	30/994
	NFYB	0.0000105	74/3715
	GABPA	0.00001689	48/2082
	BRCA1	0.0003388	61/3218
	CTCF	0.0003775	39/1790
	NFYA	0.0004409	46/2250
	PBX3	0.0005193	30/1269
	SIX5	0.00115	26/1094
	SMC3	0.003293	26/1181
	NR2C2	0.004466	11/350
	FOS	0.006121	16/637
	CREB1	0.007005	29/1444
Skyblue	RCOR1	0.002542	15/702
	BCLAF1	0.006338	16/851
HC modules not preserved in MCI			
Red	SUZ12	1.21e-09	107/1684
	EZH2	0.0001041	21/237

5.4.6 SNP analysis of significant WGCNA modules

As non-preserved modules contain genes which play a role in processes that were associated with AD or PD, they may have been more likely to contain disease associated variants than preserved modules. Each non-preserved PD module was searched for known GWAS genes associated with PD [280]. There are 69 known GWAS genes, of which four (*TMEM163, TLR9, ITIH4, TUBG2*) were in the salmon module and two (*TMEM175, STAB1*) were in the navajowhite2 module.

Significant enrichment of GWAS genes within modules that were not preserved compared to preserved networks was observed (OR = 2.96, 95% CI 1.04 6.88, Pvalue = 0.02, Fisher Exact test). Furthermore, the non-preserved PDHC purple network contained five GWAS gene (*KAT8, BIN3, TLR9, ITIH4, TUBG2*), however the non-preserved HC modules were not more likely to contain GWAS genes (OR = 2.61, 95% CI 0.08 6.47, Pvalue = 0.052, Fisher Exact test). The same analysis was performed for the non-preserved AD modules, however, no AD associated GWAS genes were found within any non-preserved modules.

In addition to searching for known GWAS genes in non-preserved modules, the SCAN database (`http://www.scandb.org/`) [308] was used to identify SNPs corresponding to the genes in each non-preserved module. These SNPs were used to search the MirSNP [322] database to identify SNPs associated with known PD or AD miRNAs dependent on the dataset of the module.

Across all non-preserved modules in the PD dataset 29 SNPs associated with 9 PD related miRNAs were identified (shown in table 5.4.7). Across the non-preserved modules in the AD dataset, 27 SNPs associated with 8 AD related miRNAs were identified (shown in table 5.4.7).

5.4.7 Comparison of AD and PD results

There is increasing evidence that PD and AD share several common characteristics [257], thus the shared processes associated with non-preserved modules were investigated in both the AD and PD dataset to see which were important in both diseases. The biological

Table 5.6: **SNPs associated with non-preserved PD modules.** SNPs in bold are shared between PD and AD.

Chromosome	SNPs	Associated PD related miRNAs	Modules with SNP associated gene	Genes
1	rs12140193	hsa-miR-495	PD darkseagreen4	*METTL13*
2	rs1138729	hsa-miR-495	PD salmon	*RRM2*
	rs12603	hsa-miR-543	HC purple	*EPB41L5*
	rs2058703	hsa-miR-1283	HC purple; PD salmon	*BCL11A*
	rs4852735	hsa-miR-4271	PD navajowhite2	*TEX261*
	rs707718	hsa-miR-543	HC purple	*CYP26B1*
3	**rs1135750**	hsa-miR-147a	PD navajowhite2	*IQCB1*
	rs11551405	hsa-miR-203	HC purple	*DCP1A*
4	**rs3805317**	hsa-miR-203	HC purple	*CLGN*
5	rs2561659	hsa-miR-543	HC purple	*AHRR*
6	**rs12528857**	hsa-miR-203	HC purple; PD darkseagreen4; PD salmon; PD navajowhite2	*TDRD6*
	rs1966	hsa-miR-543	HC purple; PD darkseagreen4	*PSORS1C1*
7	**rs1044718**	hsa-miR-147a	HC purple; PD darkseagreen4; PD salmon	*PARP12*
8	**rs2929969**	hsa-miR-133b; hsa-miR-203	PD darkseagreen4	*WISP1*
9	rs7047770	hsa-miR-133b	HC purple; PD navajowhite2	*C9orf139*
	rs818055	hsa-miR-147a	HC purple; PD navajowhite2	*LAMC3*
10	rs1042192	hsa-miR-376b	HC purple	*CYP2C18*
11	**rs10832733**	hsa-miR-543	HC purple	*PIK3C2A*
	rs2512676	hsa-miR-147a	PD darkseagreen4; PD salmon	*DLG2*
	rs7126647	hsa-miR-543	PD navajowhite2	*MRGPRX2*
	rs9444	hsa-miR-495	HC purple	*RNF169*
14	**rs1054195**	hsa-miR-543	PD navajowhite2	*CLMN*
16	rs1568391	hsa-miR-495	PD darkseagreen4	*IRF8*
17	**rs3744711**	hsa-miR-203	HC purple; PD salmon	*DHX33*
18	rs1790974	hsa-miR-203	HC purple	*DOK6*
	rs3745067	hsa-miR-4271	HC purple; PD darkseagreen4; PD salmon	*ONECUT2*
19	rs36621	hsa-miR-376b	PD navajowhite2	*TSEN34*
20	rs1060347	hsa-miR-134	HC purple	*PCMTD2*
22	rs712979	hsa-miR-203	HC purple	*C22orf39*

Table 5.7: **SNPs associated with non-preserved PD modules.** SNPs in bold are shared between AD and PD.

Chromosome	SNPs	Associated PD related miRNAs	Modules with SNP associated gene	Genes
1	rs6660019	hsa-miR-433	AD blue; HC darkolivegreen; MCI sienna3	*SASS6*
2	**rs12603**	hsa-miR-543	HC darkorange2	*EPB41L5*
	rs707718	hsa-miR-543	AD blue; HC darkolivegreen; HC red; MCI sienna3	*CYP26B1*
3	**rs1135750**	hsa-miR-147a	HC skyblue	*IQCB1*
	rs11551405	hsa-miR-203	AD blue; HC darkorange2; HC red	*DCP1A*
	rs340833	hsa-miR-433	HC skyblue	*IL5RA*
	rs6792607	hsa-miR-153	HC skyblue	*EIF5A2*
4	**rs3805317**	hsa-miR-203	AD blue; HC red; MCI sienna3	*CLGN*
	rs8336	hsa-miR-203	AD blue	*SMARCAD1*
6	rs10864	hsa-miR-433	AD blue; HC red; MCI sienna3	*BCKDHB*
	rs12528857	hsa-miR-203	AD blue; HC darkorange2; HC red; MCI sienna3	*TDRD6*
	rs1966	hsa-miR-543	AD blue; HC red; MCI sienna3	*PSORS1C1*
	rs4709266	hsa-miR-433	AD blue; HC red; MCI sienna3	*TAGAP*
7	**rs1044718**	hsa-miR-147a	HC red	*PARP12*
8	rs1042992	hsa-miR-495	HC darkorange2	*BNIP3L*
	rs2929969	hsa-miR-133b; hsa-miR-203	AD blue	*WISP1*
	rs732338	hsa-miR-134	AD blue; HC red; MCI sienna3	*LZTS1*
10	rs7071789	hsa-miR-495	HC darkolivegreen	*TRUB1*
11	**rs10832733**	hsa-miR-543	HC darkorange2	*PIK3C2A*
14	**rs1054195**	hsa-miR-543	AD blue; MCI sienna3	*CLMN*
16	rs7294	hsa-miR-147a	HC darkolivegreen	*VKORC1*
17	**rs3744711**	hsa-miR-203	HC darkorange2; HC skyblue	*DHX33*
18	rs1046699	hsa-miR-433	AD blue; HC red; MCI sienna3	*C18orf54*
	rs608823	hsa-miR-433	AD blue; HC red; MCI sienna3	*ONECUT2*
21	rs243609	hsa-miR-543	AD blue; HC red; MCI sienna3	*C21orf91*
22	rs137124	hsa-miR-134	AD blue	*CYB5R3*
	rs17032	hsa-miR-495	HC darkolivegreen	*SUN2*

processes found to be associated with significant modules in AD and PD were compared to see which were important in both diseases. No significant modules were found that were common between these two. However, some similarities between AD and PD were identified. The PDHC purple module and the ADHC darkorange2 module had four significant TFs which regulate both modules (SIX5, CREB1, NFYB, PBX3). Of the 29 PD SNPs and 27 AD SNPs identified, 12 were common between the two. The genes associated with these SNPs were: *EPB41L5, CYP26B1, IQCB1, DCP1A, CLGN, TDRD6, PSORS1C1, PARP12, WISP1, PIK3C2A, CLMN, DHX33* which are highlighted in tables and .

5.4.8 Data accession

The hub scores for each gene in PD modules not preserved in HC networks can be accessed and downloaded from `https://jack-kelly.shinyapps.io/pdpd_hubs/`. The same information for HC modules not preserved in PD networks can be found at `https://jack-kelly.shinyapps.io/pdhc_hubs/`.

The hub scores for each gene in the AD modules not preserved in HC or MCI networks can be found at `https://jack-kelly.shinyapps.io/adad_hubs/`. The same for MCI modules not preserved in HC or AD networks can be found at `https://jack-kelly.shinyapps.io/admci_hubs/` and for HC modules not preserved in MCI or AD networks at `https://jack-kelly.shinyapps.io/adhc_hubs/`.

5.5 Discussion

In this study, by using gene co-expression network analysis many important biological processes and key genes in PD and AD blood samples, and the common results between them, were identified. This is the largest network analysis of AD and PD blood to date. Insulin resistance was found to be associated with PD and *HDAC6* may play an important role in this process. The overlap in disease miRNA associated SNPs that are shared between PD and AD is highlighted, suggesting similarities in genetic risk factors between the diseases. This approach used blood data, as the available blood datasets have a large

enough sample size to construct robust and reliable networks and blood samples are easily accessible in ND patients. Previously DEGs in AD blood have been shown to be more likely to be DEGs in AD brain tissue [52]. However, in this study, DEGs in blood were not more likely to be DEGs in brain tissue for PD, nevertheless it has been shown that changes in blood gene expression did reflect changes in PD [301].

The PD network module associated with insulin resistance is not preserved in HCs. Insulin resistance is increasingly being shown to be important in PD as a potential therapeutic target [327] and insulin receptor signaling pathways are disturbed in PD as shown in chapter 4 [256]. Within this module *HDAC6* was identified as a hub gene which promotes the formation of inclusions from α-synuclein toxic oligomers [328]. *HDAC6* can promote insulin resistance by deacetylating phosphatase and tensin homolog (PTEN) in ovarian OVCAR-3 cells [329], and PTEN has in turn been shown to be involved in the pathophysiology of PD [330]. *HDAC6* has a role in influencing tau phosphorylation and autophagic flux in ND [325]. In addition, insulin signaling promotes the DNA-binding activity of FOXM1, identified as a significant TF in the insulin resistance module, which regulates pathways to promote adaptive pancreatic β cell proliferation [331], but its role in ND is not clear.

The PD module associated with cellular response to misfolded proteins was also not preserved in HC networks. PD is characterized by accumulation of misfolded α-synuclein and a failure of the proteasome to degrade these and other large protein aggregates [332]. The hub gene *SNRNP70* has been shown to be differentially expressed in PD blood previously [333]. Additionally, *SNRNP70* encodes the small nuclear ribonucleoprotein snRNP70 which co-localizes with tau in AD [334], and as tau aggregation is shown in 50% of PD cases snRNP70 may colocalize in PD cases [289].*MIR142* was also identified, which encodes miRNA-142, as a hub. miRNA-142 has been identified as an important miRNA in PD, regulating *GNAQ, TMTC2, BEND2*, and *KYNU* [335].

The AD module associated with regulation of lipolysis in adipocytes and neuroactive ligand-receptor interaction was not preserved in both MCI and HC networks. Aβ, a key molecule in AD brain pathology, can induce lipolysis within human adipose tissue [336]. In addition, lipolysis is promoted by insulin resistance and in turn lipolysis

generates ceramides further impairing insulin signaling, which is becoming increasingly more important in AD [337]. *TRPC5* was identified as a hub in this module, which along with other transient receptor potential canonical (TRPC) proteins assembles to form non-selective Ca2+ permeable channels. Another hub, *BRAP*, has a polymorphism associated with obesity and other metabolic traits, which can play a role in effecting insulin signaling and aging [338]. Interestingly, a module in the HC network that was not preserved in AD and MCI networks was also associated with regulation of lipolysis in adipocytes. This suggests that these processes are occurring in both healthy and AD conditions, however the enrichment pathways are different between the two. As no hubs are shared between the regulation of lipolysis in adipocytes modules in healthy and AD networks they are likely regulated differently.

The module associated with sensory perception in the HC network was not preserved in AD and MCI networks. Sensory dysfunction may precede the cognitive symptoms of AD [339], particularly olfactory impairment [340]. *OR5AS1* was identified as a hub gene within the module which encodes a member of the olfactory receptor family and plays a role in triggering response to smells [341]. The TF REST was identified as a regulator of the module and has been shown to regulate olfactory systems [342]. In chapter 4, REST was identified to be an important upstream TF for DEGs identified in both AD and PD previously, and as an important potential therapeutic target [256]. Future work to validate the identified hubs and TFs in both AD and PD disease models would further elucidate their potential as targets for disease treatment.

Although no common non-preserved modules in the AD and PD cohorts were identified, there were other similarities shared in the results. Four TFs were shared between the PDHC purple and the ADHC darkorange2 module (*CREB1, NFYB, PBX3, SIX5*). These two modules were associated with different transport pathways in HCs which were not preserved in the disease networks, suggesting that the roles of these TFs are dysregulated in both AD and PD. In addition to this, 12 SNPs that were shared between the 29 PD miRNAs associated SNPs and 27 AD miRNAs associated SNPs were identified. This number of shared SNPs is highly significant, which suggests that there are potential risk factors that underlie both diseases.

Several studies have applied WGCNA in ND studies for gene expression and proteomics analysis. For example, Seyfried and colleagues studied proteomic data of cortical tissue of asymptomatic and symptomatic AD [343]. They found that there was a modest overlap between networks at RNA and protein level. If a larger dataset becomes available, expanding the methods to proteomic data could give further understanding into the mechanisms of AD and PD and enable the investigation into the link between genomics and proteomics. Chatterjee *et al.* [300] have performed network analysis of PD brain tissue, however they only performed WGCNA on DEGs found in the data, which built very limited networks that removed potentially important gene interactions and disease regulators and introduced a bias of modules and hubs towards these DEGs. In addition, they used tissue from multiple brain regions which would all be affected differently by the disease [344].

A limitation of this study is that, although it has been shown that AD blood DEGs are more likely to be DEGs in the brain [52], these results suggest this is not the case for PD. Because of this, these results may not reflect major changes that take place in the brain. However, network analysis approach emphasizes the interactions of genes which univariate methods like differential expression does not. Similarly to AD, there is disruption that happens in the blood brain barrier (BBB) of PD patients [345]. Hence, it is likely that changes that take place in the brain could be reflected in the blood and vice versa. Additionally, a lot of the biological processes and genes found in the PD network have been implicated in the PD brain in the previous chapter. Tau and Aβ are hallmarks of both AD and PD in the brain and have potential as blood biomarkers in both diseases [346, 347], suggesting that changes in the brain are reflected in blood. Leukocytes have been shown to impact progression of NDs. An interaction between brain and systemic inflammation has been implicated in PD progression by an association between leukocyte apoptosis and central dopamine neuron loss [348]. Increased mitochondrial respiratory activity in leukocytes has been shown in PD patients, potentially impacting progression of neurodegeneration [349] and elevated leukocytes in CSF are significantly associated with shorter survival of patients [350]. Peripheral leukocytes have been discussed as potential biomarkers for AD previously [351], and gene expression changes in leukocytes

have been shown to be closely associated with AD progression [352]. In AD animal models circulating leukocytes have been shown to cross a dysfunctional blood brain barrier and impact brain integrity [353].

Recently limbic-predominant age-related TDP-43 encephalopathy (LATE) has been reported to be under-recognized and often misdiagnosed as AD as they share common pathogenetic mechanisms and present similarly in patients [354]. There is the potential that patients in the AD cohort may have been misdiagnosed and actually have LATE, however as LATE is seen with increasing frequency over the age of 85, and less than 6% of the AD samples were over the age of 85 this likely had little effect on the results.

The greatest risk factor for both AD and PD is age. Adjusting AD data by age before WGCNA ensured any changes found were reflective of disease state. The PD data, however, does not include samples' age information, thus the effect of age could not be removed technically. As a result of this, the PD results may have been biased towards changes as a result of aging if there was a significant difference in age between PD and HC cohorts. However, the samples were age matched in the original design which should reduce such biases [355].

From the PD dataset patient samples with known PD mutations were removed. Although the biological pathways underlying familial and sporadic forms of PD are likely to be shared, known PD mutations may impact pathways to disease or regulators of disease [356]. Removal of samples with known PD mutations prevented these mutations from having an impact on results, however had little impact on sample size due to the low number of samples with mutations. AD samples were not screened for known mutations, which could have had an impact on the results. For example, nearly 19% of the familial late onset AD population carry 2 APOE $\varepsilon 4$ alleles which only occurs in about 1% of normal Caucasian controls [23]. This and other known mutations may impact the progression and regulators of AD, and knowing which samples had these mutations could have improved the findings.

5.6 Conclusion

In conclusion, this network analysis is the largest study of its type using AD and PD blood data to date. The non-preserved module in PD associated with insulin resistance, and the hub *HDAC6* identified in this module is highlighted. These results reveal that a large proportion of disease miRNA associated SNPs are shared between PD and AD, suggesting similarities in genetic risk factors between the diseases. The hub genes that are identified have the possibility to be further investigated as potential biomarkers for disease. These insights suggest several new areas for mechanistic studies in PD and AD research fields.

Chapter 6

Identify blood biomarkers of neurodegenerative diseases by machine learning

6.1 Abstract

This chapter aims to identify blood-based biomarkers for AD and PD by applying machine learning approaches. Multiple feature selection methods are used including a knowledge-based feature pool incorporating genes identified in previous chapters of this thesis. These consequently derived feature sets are used with various classification algorithms to identify the best approach and biomarkers for AD and PD datasets individually. Additionally, deep learning algorithms are applied to reduce the feature dimensionality of data to evaluate their potential in future work on gene expression biomarkers.

To AD the best random forest model trained with 159 genes identified using VSSRFE with logistic regression (ROC AUC = 0.886) while to PD, the best random forest model is identified with all genes included in the dataset (ROC AUC = 0.743). CNN with a softmax classifier performs consistently well across both AD and PD datasets, suggesting its good potential in gene expression biomarker detection. Using knowledge-based feature pools did not inherently improve classification performance over using all genes in the dataset, suggesting that when looking for biomarkers of ND a genes importance in

pathophysiology of the disease does not translate to biomarker potential.

6.2 Background

Blood tissue is simple and easy to access and changes in blood gene expression reflect the disease processes that occur in the brain during disease as demonstrated in the previous chapter. Therefore, in addition to identifying DEGs and important pathways and processes in disease, transcriptomics data is an important omics data type for diagnostic study of human disease.

There is a lack of reliable blood-based biomarkers for both AD and PD diagnosis. α-synuclein and DJ-1 have been investigated as blood biomarkers for PD and demonstrated a high potential to be used in the clinic [82, 89, 90], however have both failed in further studies [97, 98, 93, 92]. No blood biomarkers for AD have been used clinically to date, with research identifying Aβ42/40 ratio levels [42] and NfL blood concentration [43, 44] as blood biomarkers not being validated further.

Blood gene biomarkers for NDs are particularly interesting as they have a high accessibility and are relatively cheap to perform. Identifying AD gene expression biomarkers in blood has been difficult in the past due to small sample sizes [357]. The use of statistical learning has been of particular interest for investigating blood gene biomarkers due to the high dimensionality of gene expression data. Machine learning algorithms can be use for feature selection to identify gene sets that can be best used to classify between disease and control patients. This panel of genes can then be used to train classification algorithms that can identify if new unlabelled data are from disease or control samples.

Long *et al.* [5] applied a SVM to small AD datasets and returned good results, and later used a larger dataset with LASSO feature selection and SVM classifier to get a good biomarker model with a ROC AUC of 0.87. More recently, Lee and Lee [194] used multiple feature selection and classification algorithms on multiple datasets and identified models that worked well within datasets, but performed poorly between datasets.

Shamir *et al.* [355] conducted the largest gene expression analysis of PD tissue in whole blood, including 205 PD patients and 233 health patients. They used a SVM ap-

proach to classify PD patients from healthy controls using an 87 gene signatures and obtained an ROC AUC of 0.79. Wang *et al*. [50] analyzed this data, taking a random forest approach to classify PD patients from healthy control, and achieved an ROC AUC of 0.74. With the limited approaches to classification used on such a relatively large dataset, there is a large potential for investigating other methodologies to see if this can be improved. As with AD, testing multiple different feature selection and classification algorithms may potentially improve these results.

This chapter applies a multitude of feature selection and machine learning approaches to AD and PD blood transcriptomics data. See the analysis workflow illustrated in figure 6.1. This includes a knowledge-based feature pool that includes important genes identified in chapter 4 and 5 of this thesis, to validate the hypothesis that context of existing biological knowledge improves classifying models of NDs. Additionally, deep learning approaches to dimensionality reduction and classification are applied in biomarker detection using microarray data.

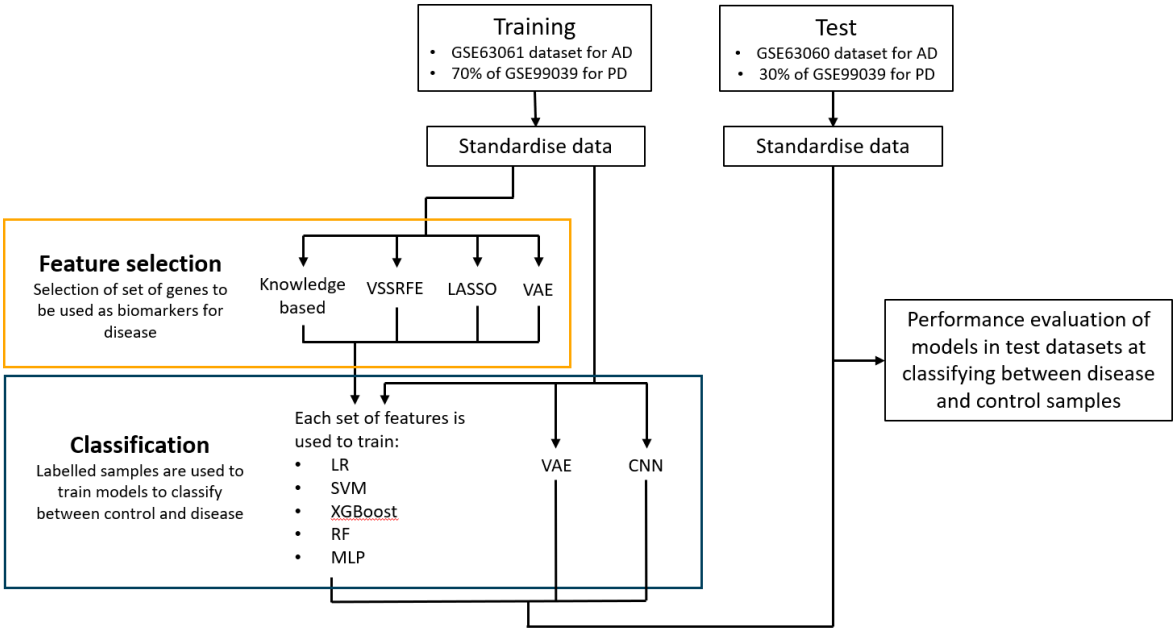


Figure 6.1: **Workflow for identification of blood biomarkers**. For the AD blood biomarker study, GSE63061 was used as the training dataset and GSE63060 as the test dataset. For the PD blood biomarker study, the GSE99039 dataset was randomly split into 70% training and 30% test data. Training and test datasets were standardised separately. Feature selection is applied to training data to generate five feature sets of genes (all genes, knowledge based genes, VSSRFE, LASSO and VAE). Each of these feature sets is used to train five classification models to identify control and disease patients (LR, SVM, XGBoost, RF and MLP). The feature set and classification model combinations are evaluated in test datasets. Additionally, a VAE and CNN, which have built in dimensionality reduction, are trained on the standardised training data and classification performance evaluated in test datasets.

6.3 Materials and Methods

6.3.1 Data processing

The GSE63061 and GSE63060 AD datasets were used as independent training and test datasets. They were processed separately using the same previous methodology however, the mild cognitive impairment patients were removed from each dataset before they were annotated. Additionally, since low expression genes have been shown to be important features in previous machine learning based microarray analyses [358], the bottom 5% of probes by average expression value in datasets was not discarded. The processed GSE63061 was used as the training dataset and the GSE66060 as the test.

The GSE99039 PD dataset was processed using the previous methodology without removing the bottom 5% of probes by average expression value. This dataset was randomly divided into a training and a testing dataset so that 70% of samples were for training and 30% for testing. This process was done by using the *train_test_split()* function in pythons sklearn library [359]. All AD and PD datasets were scaled using StandardScaler() in sklearn [359], which transforms each features distribution to a mean value of 0 and standard deviation of 1.

t-SNE was used to visualise local structures of the high dimensionality data and identify any clear groups by dimensionality reduction. t-SNE was created using the TSNE() package in sklearn [359] and 5 runs with perplexity set to 5, 15, 30, 40 and 50 run over 1000 iterations. They were then visualised using the tsneplot() function in bioinfokit (v0.9) [360].

6.3.2 Feature selection

The following multiple approaches for feature selection were considered in model training process:

- Knowledge-based feature selection
- Variable step size RFE with Logistic regression
- LASSO
- VAE

Bayesian optimization with 5-fold cross validation (CV) was applied in the feature selection process. CV was performed to optimise the precision-recall AUC (prAUC). prAUC was used for optimisation as it is less sensitive to unbalanced classes that may be present in the data [207]. This was done using *BayesSearchCV()* in the scikit-optimize python library. The subset of features that are selected are used in the training and test data to apply the machine learning algorithms later. The python code used for feature selection is available at `https://doi.org/10.5281/zenodo.4483751`.

6.3.2.1 Knowledge-based feature selection

To investigate whether feature selection under the context of existing biological knowledge can improve classification performance and yield better classifying models, a set of genes based on previous knowledge of the disease was included. In addition, genes with high variances across all samples are included as well.

For the PD dataset, the following sources were used to identify knowledge-based genes:

- DEGs identified in meta-analysis described in chapter 4 (1046 genes)
- Genes in control network modules not preserved in PD networks described in chapter 5 (606 genes)
- Genes in PD network modules not preserved in control networks described in chapter 5 (651 genes)
- PD GWAS genes (70 genes) [280]
- Genes from the KEGG 'KEGG_PARKINSONS_DISEASE' pathway [361, 362, 363] (128 genes)

In addition to these, the top 3000 genes by MAD in the PD training dataset were included.

For the AD dataset, the following sources were used to identify knowledge-based genes:

- DEGs identified in meta-analysis of AD frontal cortex performed by Li *et al.* [162] (3124 genes)
- Genes in control network modules not preserved in AD networks described in chapter 5 (1019 genes)
- Genes in AD network modules not preserved in control networks described in chapter 5 (1076 genes)
- AD GWAS genes (30 genes) [318]
- Genes from the KEGG 'KEGG_ALZHEIMERS_DISEASE' pathway [361, 362, 363] (165 genes)
- Risk genes from the Alzgene database (`Alzgene.org`) (680 genes)

In addition to these, the top 3000 genes by MAD in the AD training dataset were included.

6.3.2.2 Recursive feature elimination with variable step size

VSSRFE works to recursively eliminate the most unimportant feature until a feature set remains, as described in section 2.6.6.1. Briefly, an estimator is trained to find the importance of features in the dataset and the least important features are removed. The number of feature removed at the first step is determined by the initial step size, and as the number of features in the dataset is halved, the step size is also halved until the step size is one. This is repeated recursively on the feature set until the data is pruned to the desired number of features, usually the number that gives the best performance evaluation scores from the estimator.

As all of the datasets are a similar size, the initial step size is set to 100 for all. The feature weights used in VSSRFE are found using LR. The parameter controlling the strength of regularization of LR is tuned on the whole training datasets before VSSRFE using Bayesian optimisation with 5-fold CV.

6.3.2.3 Feature reduction using LASSO

LASSO and elastic net reduce number of features using regularisation. Regularisation approaches to feature selection are able to shrink some coefficients of features to zero and remove these features from the model.

The LASSO algorithm was applied with the sklearn python library [359] to reduce the dimensions of the data. The α constant that multiplies the L1 term was optimised so that the full feature set was reduced to best subset of features.

6.3.2.4 Variational autoencoder

As microarray data generally has a high dimensionality with a large number of features and relatively low sample numbers, VAE has great potential to reduce the dimensionality of data [175]. The basic VAE architecture based on the VAE from Zhang *et al*. [175] is shown in figure 6.2. The encoder reduces the number of features to 128 at the latent space, which was used with the machine learning classification algorithms.

The VAE is built using the keras module in python with each layer using a ReLU activation function and compiled using an Adam optimiser and categorical cross-entropy loss function with an early stopping of 10, so if loss function does not improve across for 10 epochs the training is stopped. The optimum VAE architecture was found using five-fold CV, identifying the model with best average accuracy from its softmax classifier. Accuracy is used to evaluate the model as opposed to prAUC as softmax does not give probabilities for classification.

Three architectures of VAE were tested:

- Basic VAE architecture based on the VAE from Zhang *et al.* [175] shown in figure 6.2.
- Basic VAE architecture including batch normalisation at each layer of the VAE
- Basic VAE architecture including batch normalisation at each layer of the VAE and dropout layers of 20% to prevent overfitting

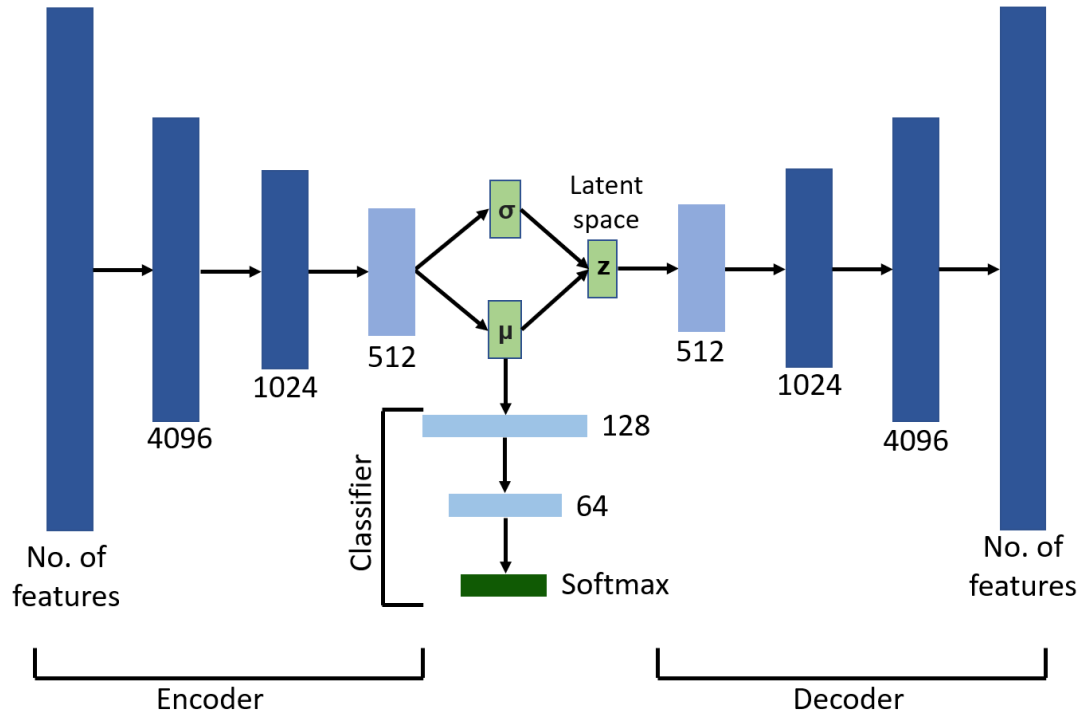


Figure 6.2: **Diagram showing the basic architecture of VAEs.** This particular example of number of layers and nodes for microarray data is based on Zhang *et al.* [175]. The VAE has three sections, the encoder, the classifier and the decoder. The input and output have the number of features in the dataset. The latent space has 128 features.

6.3.3 Machine learning for classification

Optimisation of classification algorithms was performed on training datasets using bayesian optimisation with 5-fold CV to optimise the prAUC. Table 6.1 shows the classification algorithms used on various feature sets for PD and AD training datasets. It also shows the base python code to run the algorithms and the parameters that are tuned to optimise the algorithm to training data. These algorithms were tuned and trained on all features in the training datasets and the feature sets found using the six feature selection methods discussed above to identify which feature set each classification method performs best on.

In addition to these approaches, neural network approaches that have built in dimensionality reduction and classification were used. The optimum VAE architecture found in section 6.3.2.4 was used to reduce the feature down to 128 and softmax classifier to assign samples as disease or controls.

A CNN model was built based on based on CNN applications in computer vision [196]. CNNs are similar to MLPs, however have some changes that allow CNNs to work well when built with more layers than MLPs and good at reducing data dimensionality [195]. The CNN inputs gene expression data is reshaped to a two-dimensional space that is similar to how image data is input. After a two-dimensional convolutional layer a ReLU activation function is applied to data. This data is then passed to a maxpooling layer and flattened before it is passed to a dense layer with a ReLU activation function. Softmax is then used as a classifier. This CNN is compiled using a stochastic gradient descent (SGD) optimiser and categorical crossentropy loss function.

The performance of all classification models was assessed using ROC-AUC (plotted using the roc_curve function in the sklearn [359] python package) and prAUC (plotted using precision_recall_curve function in sklearn [359]). The python code used for classification is available at `https://doi.org/10.5281/zenodo.4483751`.

Table 6.1: Classification models and the parameters that were tuned on training data

Classification algorithm	**Python library**	**Base python code**	**Parameters tuned**
LR	sklearn [359]	LogisticRegression(random_state=2, class_weight='balanced', penalty='l2', solver='liblinear')	• C
SVM with radial kernel	sklearn [359]	SVC(random_state=142, kernel = 'rbf', class_weight = 'balanced')	• C • gamma
XGBoost	xgboost [364]	XGBClassifier(random_state=42)	• scale_pos_weight • learning_rate • n_estimators • max_depth • min_child_weight • gamma • colsample_bytree • subsample • reg_alpha • reg_lambda
RF	sklearn [359]	RandomForestClassifier(random_state=10, class_weight='balanced')	• max_depth • min_samples_leaf • n_estimators • min_samples_split • max_features
MLP	sklearn [359]	MLPClassifier(random_state=10, max_iter = 10000, tol = 0.00001)	• activation • hidden_layer_sizes • solver
VAE	keras	model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=[metrics.AUC(name='PR', curve ='PR')])	• batch normalisation • dropout layers
CNN	keras	model.compile(loss = 'categorical_crossentropy', optimizer = 'sgd', metrics = ['categorical_accuracy'])	• dense layer sizes • filters • kernel size

Table 6.2: Information on number of samples in training and test datasets for AD and PD datasets

		GEO dataset	Disease	Control
AD	Train	GSE63061	137	131
	Test	GSE63060	143	104
PD	Train	GSE99039 (70% split)	141	162
	Test	GSE99039 (30% split)	68	63

6.4 Results

6.4.1 Data processing

After pre-processing the PD GSE99039 dataset is randomly split into a training dataset of 141 PD and 162 controls and test dataset of 68 PD and 63 controls, all of which have 20183 features. The GSE63061 dataset being used the training dataset for AD had 137 AD and 131 control samples, and the test dataset (GSE63060) has 143 AD and 104 controls, all with 19147 features. Information on number of samples in training and test datasets for AD and PD datasets is shown in table 6.2.

Local structures in the data and outlier samples were identified by reducing dimensionality using t-SNE. The t-SNE plots with perplexity of 30 are shown in figure 6.3 indicating no outliers in the data. The perplexity of 30 was chosen as it gave the clearest visualisation of the data. Plots of PD and AD data show no clear distinction between disease and control samples suggesting that simple classification may be difficult.

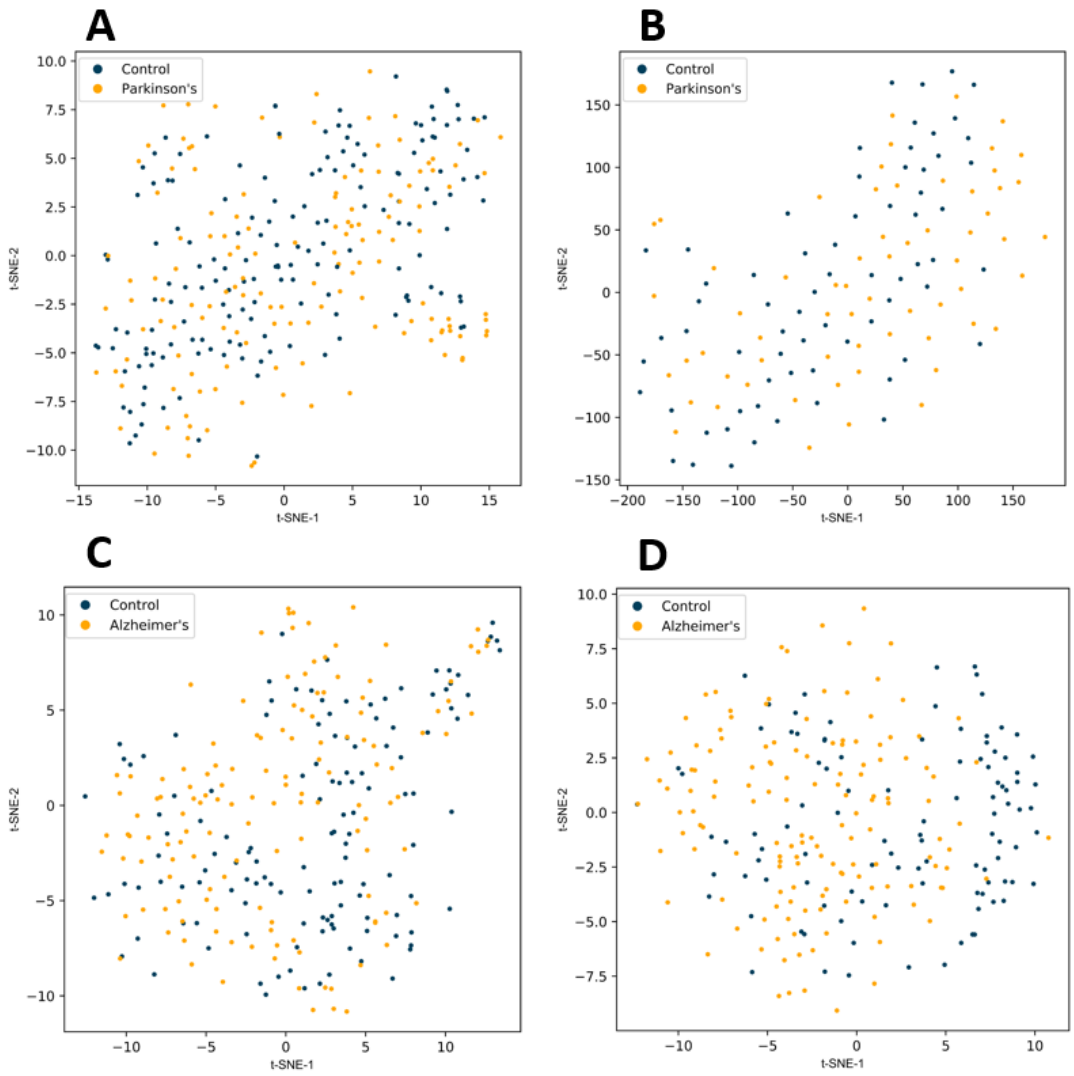


Figure 6.3: **t-SNE plots for training and test data of PD and AD.** PD training (A) and test (B) datasets and AD training (C) and test (D) datasets show no outliers and no clear distinction between disease and control samples.

6.4.2 Feature selection

Six approaches to feature selection were used on the AD and PD training datasets to identify the best panel of genes to use in classification algorithms. The python code including optimising the hyperparameters of models used in feature selection approaches is available at `https://doi.org/10.5281/zenodo.4483751`

6.4.2.1 Knowledge-based feature selection

To see if they yield better classification models, a set of features based on existing biological knowledge was used. In the PD dataset, a combination of the 2500 knowledge-based features and 3000 highest MAD features returned 4981 unique features in the dataset. In the AD dataset, a combination of the 5953 knowledge-based features and 3000 highest MAD features returned 7520 unique features in the dataset.

6.4.2.2 Recursive feature elimination with variable step size

Feature weights in VSSRFE are found using LR. On AD training data VSSRFE identified a panel of 159 genes which gave a prAUC, ROC-AUC and accuracy of 1.00 (shown in figure 6.4A). On PD training data, VSSRFE identified a panel of 5 genes (*DGKK, PTGDS, LSP1, PDLIM7* and *KIR2DL3*) that gave the maximum prAUC of 0.686, ROC-AUC of 0.704 and accuracy of 0.690 (shown in figure 6.4B).

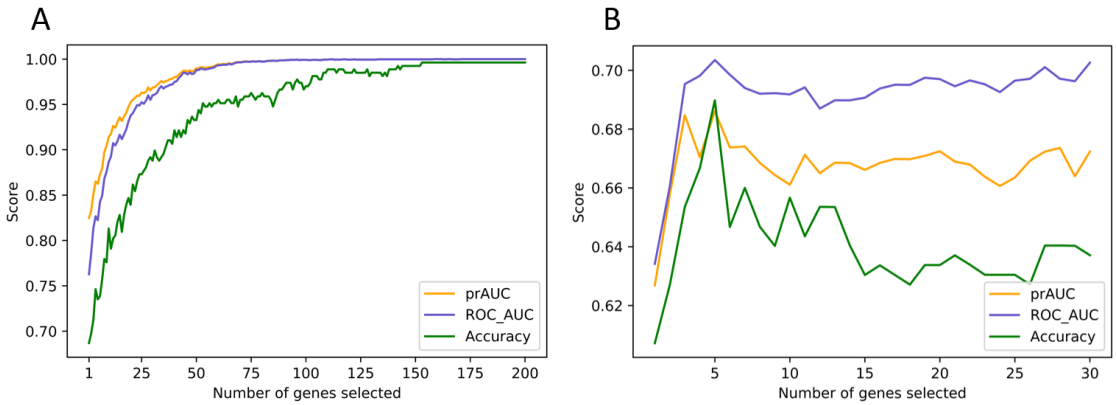


Figure 6.4: **Evaluation scores for different numbers of genes selected using VSSRFE**. VSSRFE identified a panel of 159 genes on AD data (A) and a panel of 5 genes on PD data (B) that gave the best prAUC, ROC-AUC and accuracy scores.

6.4.2.3 Feature reduction using LASSO

The number of features were reduced by regularization using LASSO. LASSO identified a gene set of 2 genes (*NDUFS5, RPL36AL*) that gave the best model (prAUC = 0.8191) using the AD dataset. Using the PD dataset, LASSO identified a gene set of 19 genes that gave the best but still poor model (prAUC = 0.5861).

6.4.2.4 Variational autoencoder (VAE)

Dimensionality of the data was reduced using a VAE. Multiple VAE architectures based on VAE proposed by Zhang *et al.* [175] were tested to identify the best model for the data by using the in-built softmax classifier to see which architecture was most effective at classifying the data using 5 fold CV.

On the AD training data, the basic VAE architecture performed the best (accuracy of 0.623) over VAE with batch normalisation (accuracy of 0.537) and dropout layers (accuracy of 0.560). The number of features was reduced to 128 by a VAE using this architecture. The learning rate for the VAE had to be reduced to 0.00001 as the model did not converge at 0.001. On the PD training data, the VAE with batch normalisation and dropout layers gave the greatest accuracy (0.554), though not much more than either VAEs without dropout, which both had an accuracy of 0.548.

6.4.3 Machine learning for classification

Optimisation of all classification algorithms was performed using bayesian optimisation with 5-fold cross validation. Python code of tuned models are available at `https://doi.org/10.5281/zenodo.4483751`. Five classification algorithms (LR, SVM with radial kernel, RF, XGBoost, MLP) were optimised and ran using each of the gene sets identified using the feature selection approaches. Additionally, a VAE and a CNN were optimised and used to reduce dimensionality and classify data.

For the PD dataset, the evaluation scores of each classification algorithm is shown in table 6.3. The ROC curves for each classification algorithm is shown in figure 6.5. All models except one (MLP using VAE feature selection) had an accuracy higher than the proportion of the largest observed class (non-information rate) of the test data (0.519). The RF model trained using all genes gave the best accuracy (0.702), ROC AUC (0.743) and prAUC (0.762), however had a much lower sensitivity (0.571) than specificity (0.824). The confusion matrix summarising the performance of this best model is shown in table 6.4. This may be advantageous for biomarkers as a false negative diagnosis is much preferred to a false positive. The CNN performed well with consistently high scores

across all evaluation approaches.

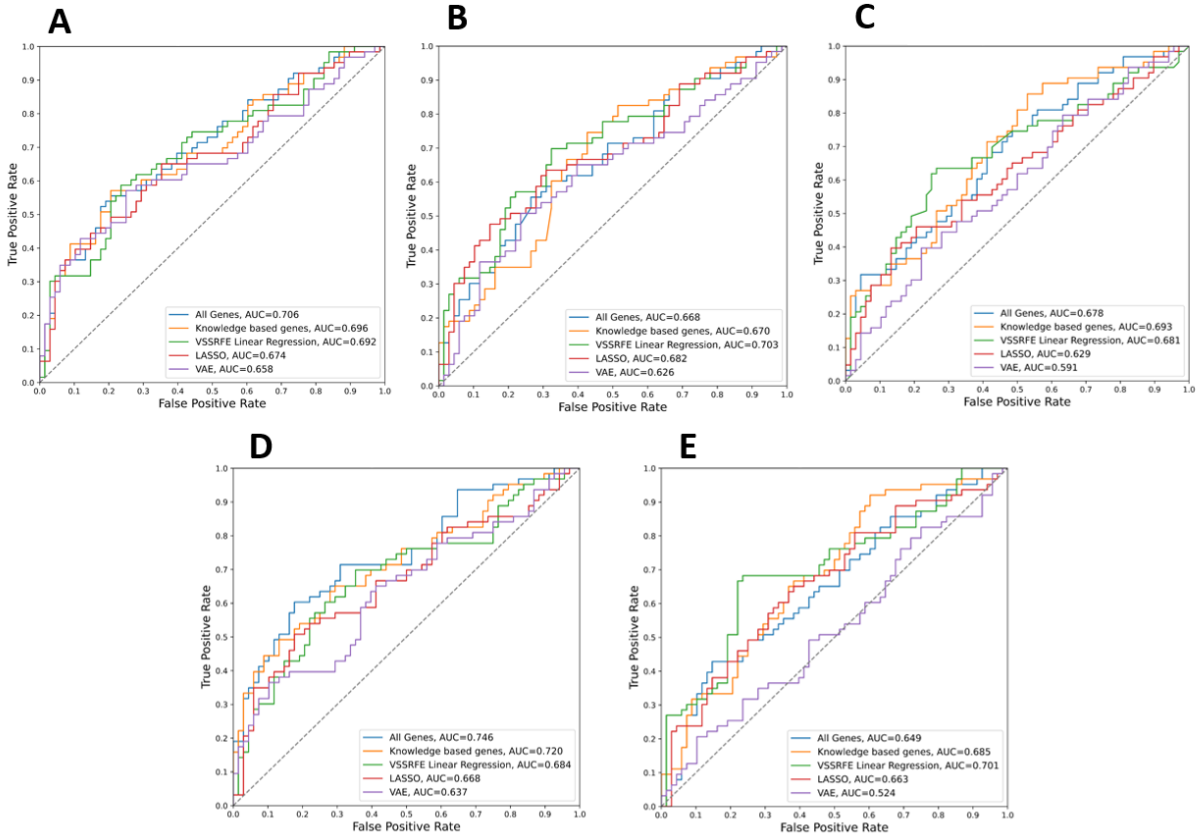


Figure 6.5: **ROC curves for each classification algorithms on PD data**. The classification algorithms used are logistic regression (A), SVM (B), random forest (C), XGBoost (D), and MLP (E).

The evaluation scores of each classification algorithm for the AD dataset is shown in table 6.5. The ROC curves for each classification algorithm is shown in figure 6.6. All models except two (MLP and SVM using VAE feature selection) had an accuracy higher than the non-information rate of the test data (0.579). The RF model trained using the 159 feature set identified using VSSRFE gave the best accuracy (0.810), ROC AUC (0.889) and prAUC (0.919). The confusion matrix summarising the performance of this best model is shown in table 6.6.

Table 6.3: **Evaluation of classification algorithms on PD data.** The results for each classification approach using all feature sets identified in feature selection is shown. There were 20183 genes in the PD dataset and 4981 in the knowledge genes feature set. VSSRFE feature selection selected 5 features(*DGKK*, *PTGDS*, *LSP1*, *PDLIM7* and *KIR2DL3*), LASSO selected 19 features and VAE reduced all features to a representative 128 features. CNN and VAE classifiers inherently reduce feature dimensions so do not require feature selection.

Classification algorithm	Feature selection	Accuracy	Sensitivity	Specificity	ROC-AUC	prAUC
LR	All genes	0.664	0.571	0.750	0.706	0.714
	Knowledge genes	0.672	0.540	0.794	0.696	0.710
	VSSRFE	0.672	0.587	0.750	0.692	0.681
	LASSO	0.641	0.571	0.706	0.674	0.682
	VAE	0.656	0.587	0.721	0.658	0.694
SVM	All genes	0.626	0.476	0.765	0.668	0.668
	Knowledge genes	0.565	0.397	0.721	0.670	0.665
	VSSRFE	0.641	0.460	0.809	0.703	0.696
	LASSO	0.641	0.556	0.721	0.682	0.689
	VAE	0.618	0.540	0.691	0.625	0.592
XGBoost	All genes	0.588	0.556	0.618	0.678	0.679
	Knowledge genes	0.618	0.524	0.706	0.693	0.698
	VSSRFE	0.679	0.603	0.750	0.681	0.675
	LASSO	0.588	0.556	0.618	0.629	0.642
	VAE	0.550	0.540	0.559	0.591	0.567
RF	All genes	0.702	0.571	0.824	0.743	0.762
	Knowledge genes	0.672	0.476	0.853	0.716	0.737
	VSSRFE	0.672	0.603	0.735	0.684	0.682
	LASSO	0.641	0.556	0.721	0.668	0.671
	VAE	0.573	0.508	0.632	0.637	0.666
MLP	All genes	0.603	0.540	0.662	0.649	0.609
	Knowledge genes	0.618	0.540	0.691	0.685	0.663
	VSSRFE	0.672	0.556	0.779	0.701	0.684
	LASSO	0.626	0.556	0.691	0.663	0.626
	VAE	0.511	0.444	0.574	0.517	0.504
CNN		0.695	0.667	0.721	0.715	0.710
VAE		0.672	0.556	0.779	0.713	0.712

Table 6.4: **Confusion matrix summarising the the performance of a best classification model on PD data.** The RF model trained using all 20183 genes in the dataset gave the best evaluation scores (accuracy = 0.702, ROC AUC = 0.743, prAUC = 0.762)

	True Positive	True Negative	Total
Predicted Positive	56	12	68
Predicted Negative	27	36	63
Total	83	48	131

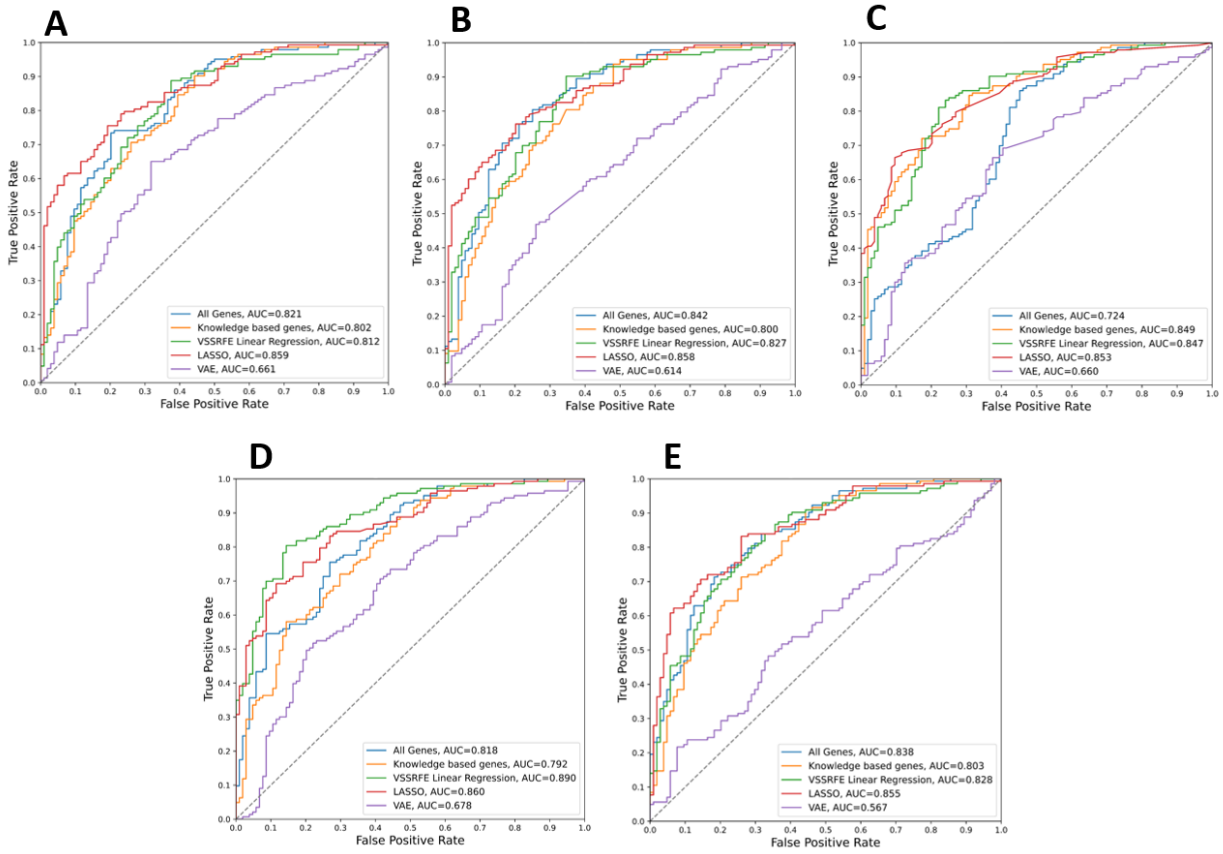


Figure 6.6: **ROC curves for each classification algorithms on AD data**. The classification algorithms used are logistic regression (A), SVM (B), random forest (C), XGBoost (D), and MLP (E).

6.5 Discussion

Diagnosis of AD and PD are still challenging in the clinic, partly due to a lack of accessible and accurate blood biomarkers. Here, classification algorithms are applied to tran-

Table 6.5: **Evaluation of classification algorithms on AD data.** The results for each classification approach using all feature sets identified in feature selection is shown. There were 19147 genes in the AD dataset and 7520 in the knowledge genes feature set. VSSRFE feature selection selected 159 features, LASSO selected 2 features and VAE reduced all features to a representative 128 features. CNN and VAE classifiers inherently reduce feature dimensions so do not require feature selection.

Classification algorithm	Feature selection	Accuracy	Sensitivity	Specificity	ROC-AUC	prAUC
LR	All genes	0.737	0.811	0.635	0.821	0.848
	Knowledge genes	0.713	0.783	0.615	0.802	0.830
	VSSRFE	0.733	0.755	0.702	0.812	0.842
	LASSO	0.777	0.790	0.760	0.859	0.899
	VAE	0.648	0.657	0.635	0.661	0.692
SVM	All genes	0.769	0.853	0.654	0.842	0.860
	Knowledge genes	0.737	0.797	0.654	0.800	0.822
	VSSRFE	0.745	0.769	0.712	0.827	0.858
	LASSO	0.769	0.797	0.731	0.858	0.898
	VAE	0.579	0.497	0.692	0.615	0.661
XGBoost	All genes	0.599	0.559	0.654	0.724	0.764
	Knowledge genes	0.741	0.713	0.779	0.841	0.875
	VSSRFE	0.794	0.853	0.712	0.847	0.883
	LASSO	0.725	0.587	0.913	0.858	0.902
	VAE	0.628	0.839	0.337	0.660	0.709
RF	All genes	0.741	0.748	0.731	0.820	0.855
	Knowledge genes	0.700	0.720	0.673	0.792	0.820
	VSSRFE	0.810	0.818	0.798	0.889	0.919
	LASSO	0.717	0.573	0.913	0.860	0.903
	VAE	0.656	0.790	0.471	0.678	0.684
MLP	All genes	0.761	0.839	0.654	0.838	0.873
	Knowledge genes	0.721	0.790	0.625	0.803	0.829
	VSSRFE	0.757	0.804	0.692	0.828	0.863
	LASSO	0.765	0.720	0.827	0.855	0.890
	VAE	0.514	0.378	0.702	0.567	0.659
CNN		0.765	0.895	0.587	0.810	0.845
VAE		0.757	0.923	0.529	0.798	0.816

Table 6.6: **Confusion matrix summarising the the performance of a best classification model on AD data.** The RF model trained using the 159 features identified using VSSRFE gave the best evaluation scores (accuracy = 0.810, ROC AUC = 0.889, prAUC = 0.919)

	True Positive	True Negative	Total
Predicted Positive	83	21	104
Predicted Negative	26	117	143
Total	109	138	247

scriptomics data to identify a panel of genes and model that has a potential as a biomarker for ND.

Using a diverse variety of feature selection and machine learning approaches the best performing models on AD and PD blood data are identified. On the PD data, the best performing model was RF on all genes (accuracy = 0.702, ROC AUC = 0.743, prAUC = 0.762). The best AD model performed better than this, using a RF model trained on a 159 gene set identified using VSSRFE (accuracy = 0.810, ROC AUC = 0.889, prAUC = 0.919). This demonstrates the viability of biomarkers using gene expression data.

Many previous AD studies using machine learning to identify biomarkers in AD have been done on small datasets [357]. Long *et al*. [5] used a novel feature selection approach of SVM forward selection followed by classification using SVM. They identified a panel of two proteins (*ECH1* and *ERBB2*) which returned a ROC AUC of 0.895. This model, however, was trained on a small dataset, comprised of only 30 AD and 30 control samples. They addressed this in a later paper which used a much larger sample size of 143 AD patients and 104 controls [52]. Here, they used LASSO feature selection to identify a panel of four probes mapping to 3 genes (*NDUFA1, MRPL51* and *RPL36AL*) which can classify AD from control patients with a ROC AUC of 0.87 using a SVM classifier. The three genes they identified as the optimum gene panel were included in the 159 gene set used in the best AD model in this study. Although the AUC was lower, this studies result has a greater power due to the much larger sample size.

A recent study by Lee and Lee [194] tested various feature selection and classification

approaches to three AD datasets. The highest ROC AUC of 0.874 was identified by using deep neural network with DEGs with a high convergent functional genomics score [365]. However, they also validated models between datasets and no combination of feature selection and classification algorithms they used achieved a ROC AUC above 0.580 using the datasets used in this study.

Many of the models built outperformed previous studies. The random forest trained on the feature set identified by VSSRFE with logistic regression gave very promising results with the best accuracy (0.810), ROC AUC (0.889) and prAUC (0.919). This model had a relatively balanced sensitivity and specificity (0.818 and 0.798 respectively). Better specificity was found in other models, with multiple models having a specificity of 0.913, however this came at the cost of sensitivity. This set of 159 features has the potential as a diagnosis panel of AD if validated in the future.

Various previous studies have worked to identify blood-based gene expression variations and signatures associated with PD. Jiang *et al.* [366] performed feature selection on PD blood trancriptomics data by identifying DEGs, reducing dimensions using LASSO and then performing recursive feature addition with a SVM on the remaining features. This identified a panel of 9 genes (*PTGDS, GPX3, SLC25A20, CACNA1D, LRRN3, POLR1D, ARHGAP26, TNFSF14* and *VPS11*) which were used with SVM, random forest and decision tree model classifiers. *PTGDS* and *LRRN3* were the only genes from their feature set that were in any of the feature selection method used in this study, with the former being identified in all feature selection approaches and the latter only present in genes based on previous knowledge. They identified the best classification approach to be random forest with a ROC AUC of 0.777, however this study has many limitations. Their limited approach to feature selection that involved only using DEGs likely removed many key features early before LASSO could be applied. The largest limitation of their study is the small size of the test dataset, which can introduce bias that result in performance estimations that do not reflect the true quality of the model. Work by Shamir *et al.* [355], who achieved a ROC AUC greater than those found in this study using the same dataset, also had this limitation.

Falchetti *et al.* [367] used much larger test datasets by performing a meta-analysis of

four PD blood datasets. For feature selection, they selected the top 100 DEGs by absolute effect size and used RFE to identify a 59 gene set that was used with 9 classification algorithms. Using an 80% training, 20% test split of the data they had a more balanced split of data than previous studies. The best model they identified was a SVM with radial kernel which achieved a ROC AUC of 0.791, although many of their models outperformed those created in this study. Datasets used by Falchetti *et al*. were combined by merging after re-scaling each gene in each dataset which, although made the sample size much greater, may have introduced covariates to the data, especially with high levels of technical noise present in microarrays.

The large sample size of the train and test data that have come from the same datasets used in this study avoids many of the limitations that these previous studies have had. Despite this, the PD models in this study underperform compared to some previous results. Although the models perform worse, the larger sample size increase the likelihood that the results are reproducible, which is extremely important for diagnostic study. The best PD model identified had a low sensitivity but high specificity. A high specificity is important in biomarkers, as it ensures that patients who do not have the disease are not misdiagnosed or overdiagnosed.

The results show that VAE feature selection performs relatively poorly at capturing a representation of microarray data that can be used for classification. Previous work has also shown that VAE approaches lose important information in ND microarray data [194]. This is likely due to the complex nature of gene expression in blood for NDs. When used in tissue that is more impacted by disease, VAE has been an effective way of reducing feature dimensionality while retaining feature information [368]. This is possible in disease in which the impacted tissue can be directly biopsied such as cancer, however is not practical in NDs. CNNs, on the other hand, performed well for classification. Traditionally applied to imaging data, CNNs work well with many layers making them suited to reducing data dimensionality and classification [195]. Previously, they have been shown to work well in classifying various cancer types [196]. On AD data, the CNN had high sensitivity and low specificity. This makes them good for detecting actual cases of disease, however they have a high rate of false positives. This can be advantageous for

biomarkers as those predicted to have ND can undergo further testing and diagnosis, and healthy people can have ND ruled out quickly by healthcare professionals. CNNs can be built to be extremely complex, and improving CNN for classification in NDs would be an interesting approach for potential future work.

These results describe potential future diagnostic biomarkers for NDs, however there are some limitations throughout the study. Studies to identify gene expression biomarkers require very large sample sizes to identify a reliable diagnostic signature [369]. The datasets used in this study are the largest that are publicly available and so should give comprehensive results, however would likely require validation in thousands of samples [369].

Additionally, information on other variables that can impact the data would allow for the impact of these on results to be reduced. Phenotypic information, such as age, gender, smoking status, BMI and other factors that impact disease progression would be useful, as would information that can affect data collection and processing, such as abundance of blood cell types in samples. Further information on patients disease history and symptoms would make it possible to investigate the effect of ND as the disease progresses and develop prognostic biomarkers. Additionally, it would be possible to create biomarkers that predict risk of certain symptoms developing.

Misdiagnosis rates in ND are very high, for example, misdiagnosis rates of AD ranging from 12% to 23% in pathologically confirmed studies [7]. Diagnosis of ND are generally based on clinical examination and ruling out other potential causes of symptoms using PET scans and blood tests. As a result of this, there is also the potential that patients in the cohorts with which the model is being trained and tested on are misdiagnosed. If misdiagnosed patients are present in the initial cohort to identify the model it is likely the models will continue to misdiagnose patients with similar conditions. In the datasets used in this study, diagnosis criteria are more strict than the minimal required for diagnosis in clinical settings, which should reduce the impact of this on the results of this work.

6.6 Conclusion

This study aimed to identify blood-based gene expression biomarkers for AD and PD. Additionally, it aimed to assess whether feature selection under the context of existing biological knowledge can improve classification performance. The approach to feature selection and classification used in this study is the most thorough to date of ND data. The models shown in this chapter have successfully classified AD and PD patients from controls with very good evaluation metrics and show promise as biomarkers for PD and AD. The potential of deep learning, particularly CNNs, is also demonstrated in use for biomarker detection using transcriptomics data, which can be improved and refined through future work. The use of knowledge-based feature pools did not improve classification performances suggesting that the pathophysiological importance in ND brain tissue does not directly translate to biomarker potential in blood tissue, however there is still potential for more data-driven approaches including feature weighting that would likely improve feature selection.

These promising findings now need to be investigated in larger cohorts before their clinical viability can be determined. Like most biomarkers, gene expression biomarkers would work best when used as a tool in combination with other diagnosis approaches.

Chapter 7

Conclusion

7.0.1 Final discussion

Gene expression data is important in the investigation of underlying processes of NDs. One of the biggest risk factors for many NDs is age, and as number of over 65s in the world increases the prevalence of NDs is growing [1]. Understanding NDs and elucidate processes that are involved in these diseases can uncover new therapeutic approaches and areas of investigation. Additionally, being able to accurately diagnose NDs is becoming increasingly important. Misdiagnosis of NDs is high and so biomarkers are an invaluable tool [7]. Many approaches to identifying ND biomarkers have been used in the past including blood and CSF protein levels and neuroimaging techniques like PET and MRI, with very little success. Gene expression data is a good target to identify biomarkers due to accessibility and how they are related to disease.

Applications of differential expression analysis to RNA-seq data was used to identify novel HD associated genes. Differential gene expression is the most common approach to gene expression data and so the importance of using up to date techniques to identify DEGs is demonstrated by identifying dyregulation in known important pathways of HD, such as Immune response and inflammatory pathways, and novel therapeutic targets within these pathways. The involvement of astrocytes and microglia in inflammatory pathways that take place in HD brains is demonstrated, and *NFE2L2* and *PITX1* are identified as potential therapeutic targets of inflammation in HD brain.

DSP was highlighted as an important gene in multiple facets of investigation, with literature supporting its role in protecting against telomere DNA damage and cell apoptosis and a potential protective role in HD. IHC was used to confirm this dysregulation of DSP protein expression in the prefrontal cortex of HD patients. There is potential for desmoplakin to be investigated as a CSF or blood plasma biomarker for HD based on previous literature. Desmoplakin levels in plasma have previously been shown to be a potential early biomarker for statin responsiveness after ischemic stroke [370], and CSF levels have been investigated as a biomarker to rule out false positive rates of 14-3-3 proteins in CJD [371]. One of the advantages of RNA-seq is that old data can be still used with improved quality control and alignment software. In this thesis an approach to RNA-seq data for identification of DEGs is discussed that identifies more known disease pathways and DEGs than previous approaches, as well as uncovers novel results.

As gene profiling technology becomes more accessible and cheaper, the number of available gene expression studies increases and meta-analysis becomes important in identification of important genes associated with NDs. A combined effect size approach has been successful in meta-analysis of AD microarray data previously [162], which gives the advantage of using all combined gene sets from all the studies included in the meta-analysis rather than the genes that are common between all datasets as other approaches have done [163, 164]. As this approach was successful on AD data, it was applied to largest PD microarray cohort to date to identify underlying processes in disease and investigate the cross-talk between AD and PD gene expression.

YWHAZ and other genes coding 14–3-3 proteins are highlighted as important DEGs in signaling pathways and in PPINs. 14-3-3 proteins have been shown to interact with α-synuclein, Parkin and LRRK2 proteins [283] however are a novel target in investigation of PD. The modulation of *YWHAZ* has been proposed for therapeutic approaches to cancer [372], although no functional domains that are druggable have been identified in the structure of 14–3-3ζ [373]. However, a recent study has shown a novel small molecule protosappanin A that can perform allosteric regulation of 14–3-3ζ [373] that has future potential in therapeutic trials for diseases associated with *YWHAZ* dysregulation. Although there is little research on using individual 14-3-3 proteins as biomarkers, there

has been much on using 14-3-3 proteins as blood and CSF biomarkers for ischemic CNS damage and NDs, especially CJD [374]. This opens up the possibility of research using 14-3-3 proteins as biomarkers of PD in blood and CSF samples. Additionally, 14-3-3 proteins are not found in the CSF of people with other ND diseases, suggesting that 14-3-3 presence is not just as a result of brain cell death [374].

Perturbed pathways also include oxidative stress and mitochondrial dysfunction. Interestingly, of the 1046 DEGs identified a significant majority of 71% were downregulated. Other NDs, including AD, do not have a significant difference in up and downregulated genes, suggesting that PD is driven by loss of gene function. Additionally, investigating the cross-link between PD and AD elucidated the common pathological and physiological links between the diseases. By doing this, known targets in AD can be identified in PD, allowing for an application of research and future approaches to help in investigating both diseases. For example, the sirtuin signalling pathway and the upstream regulator *REST* have both been researched and investigated in AD, however are novel in PD. Targeting sirtuins and *REST* in PD presents a strong therapeutic potential that had been recognised in AD but is novel in PD.

As AD and PD cross-talk is shown to be important further research was required to further understand the similarities, and differences, between the diseases. There is increasing evidence in the literature [257] and it has been shown here that molecular pathways, including mitochondrial function, oxidative stress and inflammation underlie the pathogenesis of both AD and PD. Investigating the similar pathogenic mechanisms of both diseases allows for better understating of how biological changes lead to the similar symptoms of NDs and how they develop as individual diseases. Additionally, it would allow for the repurposing of drugs that target biological mechanisms that are shared between diseases. The availability of blood gene expression data for both PD and AD allowed for similar methodology to be applied to both and diseases. Using blood gene expression data, network analysis identified a large proportion of disease miRNA associated SNPs are shared between PD and AD, suggesting similarities in genetic risk factors between the diseases.

One of the key findings of the PD network was the insulin resistance module in PD patients not being preserved in controls networks. Insulin resistance is known to be im-

portant in PD [327] and a potentially very important therapeutic target. A recent approach to PD and AD treatment has emerged where brain insulin function are enhanced via intranasal delivery of insulin [375]. This approach bypasses the blood brain barrier and ensure quick delivery to the brain and does not impact blood insulin level. This approach requires much more work to see clinical applications, as optimal dosage to reduce any side effects and improve disease symptoms has yet to be found [375].

Up until now, identifying hub genes in gene co-expression networks has been limited to using single approaches with no indication of statistical significance. A novel hub detection permutation test was developed which uses multiple approaches to identify the key hubs within important disease subnetworks. Identifying hub genes increases understanding of these subnetworks and the biological processes associated with them and offers treatment targets in the future. Using this approach, *HDAC6* was identified as a hub gene in insulin resistance pathways dysregulated in PD. HDAC6 has been shown to promote the formation of α-synuclein toxic oligomer inclusions [328] and the PD risk gene Parkin has been shown to degrade impaired mitochondria via HDAC6 pathways [376], highlighting its importance. HDAC6 inhibitors have been proposed as a treatment of AD, and have been shown to re-establish memory and recognition in AD models [377]. However, within treatment in humans, the safety profile of HDAC6 inhibitors are currently being improved for use in treatment of cancer and AD [378]. This research into HDAC6 inhibitors can be applied to PD and potentially other NDs in the future as a therapeutic approach.

A lipolysis subnetwork was present in AD patients but not in MCI and control patients. Lipolysis is promoted by insulin resistance and in turn impairs insulin signalling. *TRPC5* and *BRAP* are identified as hubs in lipolysis subnetwork and have both been previously associated with insulin signalling [338, 379]. With this identification of novel pathways and hub genes that present new areas of mechanistic research and important targets in both diseases, the potential of blood gene expression for identification of biomarkers became clear.

In the future, network approaches can be expanded upon by inferring causality between genes. Undirected networks give insight into if there is a relationship between genes and how these change in disease, however, finding the direction of edges can give

further insight into potential therapeutic targets and disease pathways. As more genetic data become available and methodologies are developed for inferring causality in gene networks, these approaches will become increasingly important.

Blood tissue is very accessible and blood gene expression biomarkers have been shown to have promise for diagnosis of disease [5, 52]. As transcriptomics data measures the expression of thousands of genes, and approaches become more accessible, a panel of genes where the expression can be objectively measured to classify control and disease is potentially more obtainable than measurement of individual proteins or molecules. Although I show that DEGs in PD brain and blood samples were not significantly overlapped, my network analysis of blood PD data demonstrated that many of the biological processes identified were similar between the two and so blood gene expression can reflect changes that are taking place in the brain. When looking for biomarkers, it is more of interest if changes can be used for diagnosis than if the individual genes used play an important role in disease processes in the brain. The large datasets that are used in this work allow for a better opportunity to develop more accurate blood biomarkers.

Using the large gene expression datasets used in network analysis, a large selection of feature selection and statistical learning classifiers were trained to diagnose ND. These approaches aimed to identify a novel panel of gene biomarkers that can effectively classify disease from control patients. One of the key strengths of this approach is that multiple genes are used as biomarkers. Single biomarkers often forgo specificity for sensitivity or vice versa and so are not as effective as using multiple biomarkers [380]. This is especially true for gene expression data as a result of the large amount of noise present in microarray data. The models built and novel gene sets identified in this thesis demonstrated a great ability to classify disease from control patients. The best model distinguishing AD from control patients was shown with a high performance of 0.919 prAUC and the best classifier for PD and control shown to have a 0.762 prAUC. Random forest was identified to be the optimum classifier in both diseases, trained on a 159 gene set identified using VSSRFE in the AD data and all genes in the PD data.

This work also demonstrates the importance and value of revisiting and reanalysing publicly available. With the development of novel and superior computational and statis-

tical approaches, patterns and information in previously analysed data can be identified. Approaches such as meta-analysis can combine the information from a number of smaller old studies to identify novel disease knowledge. Additionally, as new approaches to gene expression profiling like RNA-seq become available and accessible it does not mean older approaches like microarray analysis become redundant and in fact can still offer a fantastic approach to gene expression analysis with large sample sizes and advanced analysis approaches.

7.0.2 Future work

The aim of the work undertaken in this thesis was to investigate approaches to gene expression data to elucidate underlying processes of NDs and use statistical learning to identify potential biomarkers. This was achieved, identifying many novel processes and important genes that regulate ND and identifying models that can accurately classify ND patients from healthy people. These results open up many areas of future research.

This work elucidates the underlying common characteristics between AD and PD. Particularly, REST was identified as a regulator in both diseases. Research has been done investigating the role of REST in AD [293] which has the potential to be applicable to PD after further research. Future work on particular common pathways or key genes that are shared between AD and PD may find a target for therapeutic intervention in both diseases, or allow for established knowledge or treatments approaches in one disease to be applied to the other.

The promising gene expression biomarkers detected in this study would need to be investigated in larger cohorts to determine their clinical viability. Additionally, new technologies are being refined that can speed up the detection of a gene expression biomarker panel. Graphene-based biosensors have been used in cancer diagnosis for quantitative detection of DNA, miRNA and proteins [381]. Although they are very early in research they have been shown to have great potential and could be used with gene expression panels and models found in this work. Additionally, protein and methylation array data could be used in tangent with biosensors to have high sensitivity of ND diagnosis.

This work also shows CNNs worked well in reducing dimensionality and classifying data in both AD and PD, demonstrating their future potential in biomarker detection for ND research. As CNNs reduces dimensionality from all features it can include information that other feature selection approaches do not. CNNs can be built with very custom architectures for the problem they are addressing. They have been successfully used with neuroimaging data for many years and are still improving [382] and so it stands to reason that with more work they could be greatly improved with their use on gene expression data as well [196]. Even with a limited number of architectures tried in this work CNNs performed well, achieving a ROC-AUC of 0.810 on AD data.

Bibliography

[1] United Nations Department of Economic and Social Affairs. *World Population Ageing 2019*. Tech. rep. Economic and Social Affairs, Population Division, 2019.

[2] R. Cacace, K. Sleegers, and C. V. Broeckhoven. "Molecular genetics of early-onset Alzheimer disease revisited". *Alzheimer's & Dementia* 12.6 (2016), pp. 733–748.

[3] Alzheimer's Association. *2018 Alzheimer's Disease Facts and Figures*. Tech. rep. 2018.

[4] L. Shi, A. L. Baird, S. Westwood, A. Hye, R. Dobson, M. Thambisetty, and S. Lovestone. "A Decade of Blood Biomarkers for Alzheimer's Disease Research: An Evolving Field, Improving Study Designs, and the Challenge of Replication". *Journal of Alzheimer's Disease* 62.3 (2018), pp. 1181–1198.

[5] J. Long, G. Pan, E. Ifeachor, R. Belshaw, and X. Li. "Discovery of Novel Biomarkers for Alzheimer's Disease from Blood". *Disease Markers* 2016 (2016), Article ID 4250480.

[6] G. M. McKhanna et al. "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging- Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". *Alzheimers Dementia* 7.3 (2012), pp. 263–269.

[7] C. A. Hunter, N. Y. Kirson, U. Desai, A. K. G. Cummings, D. E. Faries, and H. G. Birnbaum. "Medical costs of Alzheimer's disease misdiagnosis among US Medicare beneficiaries". *Alzheimer's and Dementia* 11.8 (2015), pp. 887–895.

[8] J. Cummings, G. Tong, and C. Ballard. “Treatment Combinations for Alzheimer’s Disease: Current and Future Pharmacotherapy Options”. *Journal of Alzheimers Disease* 67.3 (2019), pp. 779–794.

[9] S. M. McCurry, C. F. Reynolds, S. Ancoli-Israel, L. Teri, and M. V. Vitiello. “Treatment of sleep disturbance in Alzheimer’s disease”. *Sleep Medicine Reviews* 4.6 (2000), pp. 603–628.

[10] J. Weller and A. Budson. “Current understanding of Alzheimer’s disease diagnosis and treatment”. *F1000Research* 7.F1000 Faculty Rev (2018), p. 1161.

[11] M. Saxena and R. Dubey. “Target Enzyme in Alzheimer’s Disease: Acetylcholinesterase Inhibitors”. *Current Topics in Medicinal Chemistry* 19.4 (2019), pp. 264–275.

[12] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings. “Alzheimer’s disease”. *Nature Reviews Disease Primers* 1 (2015), Article number: 15056.

[13] A. Habib, D. Sawmiller, and J. Tan. “Restoring sAPP functions as a potential treatment for Alzheimer’s disease”. *Journal of Neuroscience Research* 95.4 (2017), pp. 973–991.

[14] R. Shu et al. “APP intracellular domain acts as a transcriptional regulator of miR-663 suppressing neuronal differentiation”. *Cell Death Disease* 6 (2015), e1651.

[15] H. Nhan, K. Chiang, and E. Koo. “The multifaceted nature of amyloid precursor protein and its proteolytic fragments: friends and foes”. *Acta Neuropathologica* 129 (2015), pp. 1–19.

[16] G. K. Gouras, T. T. Olsson, and O. Hansson. “β-amyloid Peptides and Amyloid Plaques in Alzheimer’s Disease”. *Neurotherapeutics* 12.1 (2015), pp. 3–11.

[17] K. Wildsmith, M. Holley, J. Savage, R. Skerrett, and G. Landreth. “Evidence for impaired amyloid β clearance in Alzheimer’s disease”. *Alzheimer’s Research Therapy* 5 (2013), p. 4.

[18] S. Sadigh-Eteghad, B. Sabermarouf, A. Majdi, M. Talebi, M. Farhoudi, and J. Mahmoudi. “Amyloid-Beta: A Crucial Factor in Alzheimer’s Disease”. *Medical Principles and Practice* 24 (2015), p. 1.

[19] V. Lattanzi, K. Bernfur, E. Sparr, U. Olsson, and S. Linse. “Solubility of Aβ40 peptide”. *JCIS Open* 4 (2021), p. 100024.

[20] J. Zhao, Y. Deng, Z. Jiang, and H. Qing. “G protein-coupled receptors (GPCRs) in Alzheimer’s disease: A focus on BACE1 related GPCRs”. *Frontiers in Aging Neuroscience* 8 (2016), p. 58.

[21] Y. Gao, L. Tan, J.-T. Yu, and L. Tan. “Tau in Alzheimer’s Disease: Mechanisms and Therapeutic Strategies”. *Current Alzheimer Research* 15.3 (2018), pp. 283–300.

[22] C. Gong and K. Iqbal. “Hyperphosphorylation of Microtubule-Associated Protein Tau: A Promising Therapeutic Target for Alzheimer Disease”. *Curr Med Chem.* 15.23 (2008), pp. 2321–2328.

[23] L. M. Bekris, C.-E. Yu, T. D. Bird, and D. W. Tsuang. “Genetics of Alzheimer disease”. *Journal of Geriatric Psychiatry and Neurology* 23.4 (2010), pp. 213–227.

[24] A. Pilotto, A. Padovani, and B. Borroni. “Clinical, biological, and imaging features of monogenic Alzheimer’s disease”. *BioMed Research International* 2013 (2013), Article ID: 689591.

[25] C. Liu, C. Liu, T. Kanekiyo, H. Xu, and G. Bu. “Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy”. *Nature Reviews Neurology* 9.2 (2013), pp. 106–118.

[26] Q. Jiang et al. “ApoE Promotes the Proteolytic Degradation of Aβ”. *Neuron* 58.5 (2008), pp. 681–693.

[27] K. Lee, Y. M. Lee, J. M. Park, B. D. Lee, E. Moon, H. J. Jeong, S. Y. Kim, Y. I. Chung, and J. H. Kim. “Right hippocampus atrophy is independently associated with Alzheimer’s disease with psychosis”. *Psychogeriatrics* 19.2 (2019), pp. 105–110.

[28] L. C. De Souza, M. Chupin, M. Bertoux, S. Lehéricy, B. Dubois, F. Lamari, I. Le Ber, M. Bottlaender, O. Colliot, and M. Sarazin. “Is hippocampal volume a

good marker to differentiate alzheimer's disease from frontotemporal dementia?" *Journal of Alzheimer's Disease* 36 (2013), pp. 57–66.

[29] M. Hashimoto, H. Kitagaki, T. Imamura, N. Hirono, T. Shimomura, H. Kazui, S. Tanimukai, T. Hanihara, and E. Mori. "Medial temporal and whole-brain atrophy in dementia with Lewy bodies: A volumetric MRI study". *Neurology* 51.2 (1998), pp. 357–362.

[30] C. H. Andrade-Moraes et al. "Cell number changes in Alzheimer's disease relate to dementia, not to plaques and tangles." *Brain : a journal of neurology* 136.12 (2013), pp. 3738–3752.

[31] M. Wirth, C. M. Madison, G. D. Rabinovici, H. Oh, S. M. Landau, and W. J. Jagust. "Alzheimer's disease neurodegenerative biomarkers are associated with decreased cognitive function but not-amyloid in cognitively normal older individuals". *Journal of Neuroscience* 33.13 (2013), pp. 5553–5563.

[32] R. A. Sperling, M. C. Donohue, R. Raman, C. K. Sun, R. Yaari, K. Holdridge, E. Siemers, K. A. Johnson, and P. S. Aisen. "Association of Factors with Elevated Amyloid Burden in Clinically Normal Older Individuals". *JAMA Neurology* 77.6 (2020), pp. 735–745.

[33] L. Wang et al. "Evaluation of Tau Imaging in Staging Alzheimer Disease and Revealing Interactions Between β-Amyloid and Tauopathy". *JAMA Neurology* 73.9 (2016), pp. 1070–1077.

[34] R. L. Joie et al. "Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET". *Science Translational Medicine* 12.524 (2020), eaau5732.

[35] O. Hansson, S. Lehmann, M. Otto, H. Zetterberg, and P. Lewczuk. "Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease". *Alzheimer's Research and Therapy* 11 (2019), Article number: 34.

[36] K. Blennow and H. Zetterberg. “Biomarkers for Alzheimer’s disease: current status and prospects for the future”. *Journal of Internal Medicine* 284.6 (2018), pp. 643–663.

[37] K. Blennow, N. Mattsson, M. Schöll, O. Hansson, and H. Zetterberg. “Amyloid biomarkers in Alzheimer’s disease”. *Trends in Pharmacological Sciences* 36.5 (2015), pp. 297–309.

[38] K. Blennow and H. Zetterberg. “Cerebrospinal Fluid Biomarkers for Alzheimer’s Disease”. *Journal of Alzheimer’s Disease* 18.2 (2009), pp. 413–417.

[39] E. Kapaki, G. P. Paraskevas, I. Zalonis, and C. Zournas. “CSF tau protein and β-amyloid (1-42) in Alzheimer’s disease diagnosis: Discrimination from normal ageing and other dementias in the Greek population”. *European Journal of Neurology* 10.2 (2003), pp. 119–128.

[40] K. Blennowa, B. Duboisb, A. M. Faganc, P. Lewczukd, M. J. de Leone, and H. Hampel. “Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer’s disease”. *Alzheimer’s & Dementia* 11.1 (2015), pp. 58–69.

[41] J. Toombs and H. Zetterberg. “In the blood: biomarkers for amyloid pathology and neurodegeneration in Alzheimer’s disease”. *Brain Communications* 2.1 (2020), fcaa054.

[42] A. Nakamura et al. “High performance plasma amyloid-β biomarkers for Alzheimer’s disease”. *Nature* 554.7691 (2018), pp. 249–254.

[43] E. H. Thijssen et al. “Diagnostic value of plasma phosphorylated tau181 in Alzheimer’s disease and frontotemporal lobar degeneration”. 26.3 (2020), pp. 387–397.

[44] S. Janelidze et al. “Plasma P-tau181 in Alzheimer’s disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer’s dementia”. *Nature Medicine* 26.3 (2020), pp. 379–386.

[45] L. Gaetani, K. Blennow, P. Calabresi, M. D. Filippo, L. Parnetti, and H. Zetterberg. “Neurofilament light chain as a biomarker in neurological disorders”. *Journal of Neurology, Neurosurgery, and Psychiatry* 90.8 (2019), pp. 870–881.

[46] M. M. Mielke et al. “Plasma and CSF neurofilament light: Relation to longitudinal neuroimaging and cognitive measures”. *Neurology* 93.3 (2019), e252–e260.

[47] N. Mattsson et al. “Plasma tau in Alzheimer disease”. *Neurology* 87.17 (2016), pp. 1827–1835.

[48] T. K. Karikari et al. “Blood phosphorylated tau 181 as a biomarker for Alzheimer’s disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts”. *The Lancet Neurology* 19.5 (2020), pp. 422–433.

[49] D. Pellegrino-Coppola, A. Claringbould, M. Stutvoet, B. Consortium, D. I. Boomsma, M. A. Ikram, P. E. Slagboom, H.-J. Westra, and L. Franke. “Correction for both common and rare cell types in blood is important to identify genes that correlate with age”. *BMC Genomics* 22 (2021), Article number: 184.

[50] C. Wang, L. Chen, Y. Yang, M. Zhang, and G. Wong. “Identification of potential blood biomarkers for Parkinson’s disease by gene expression and DNA methylation data integration analysis”. *Clinical Epigenetics* 11 (2019), Article number: 24.

[51] S. Ahmed, M. Kabir, Z. Ali, M. Arif, F. Ali, and D.-J. Yu. “An Integrated Feature Selection Algorithm for Cancer Classification using Gene Expression Data”. *Combinatorial Chemistry & High Throughput Screening* 21.9 (2018), pp. 631–645.

[52] X. Li et al. “Systematic Analysis and Biomarker Study for Alzheimer’s Disease”. *Scientific Reports* 8 (2018), p. 17394.

[53] Parkinson’s UK. *The prevalence and incidence of Parkinson’s in the UK*. Tech. rep. London, 2017.

[54] C. Marras et al. “Prevalence of Parkinson’s disease across North America”. *NPJ Parkinson’s Disease* 4 (2018), Article number: 21.

[55] M. T. Hayes. “Parkinson’s Disease and Parkinsonism”. *American Journal of Medicine* 132.7 (2019), pp. 802–807.

[56] K. Kalinderi, S. Bostantjopoulou, and L. Findani. “The genetic background of Parkinson’s disease: current progress and future prospects”. *Acta Neurologica Scandinavica* 134.5 (2016), pp. 314–326.

[57] A. H. Schapira, K. R. Chaudhuri, and P. Jenner. “Non-motor features of Parkinson disease”. *Nature Reviews Neuroscience* 18.7 (2017), pp. 435–450.

[58] H. A. Hanagasi, Z. Tufekcioglu, and M. Emre. “Dementia in Parkinson’s disease”. *Journal of the Neurological Sciences* 374 (2017), pp. 26–31.

[59] S. G. Reich and J. M. Savitt. “Parkinson’s Disease”. *Medical Clinics of North America* 103.2 (2019), pp. 337–350.

[60] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino. “Accuracy of clinical diagnosis of Parkinson disease”. *Neurology* 86.6 (2016), pp. 566–576.

[61] J. E. Hardebo and C. Owman. “Barrier mechanisms for neurotransmitter monoamines and their precursors at the blood-brain interface”. *Ann Neurol.* 8.1 (1980), pp. 1–31.

[62] D. B. Calne, J. L. Reid, S. D. Vakil, S. Rao, A. Petrie, C. A. Pallis, J. Gawler, P. K. Thomas, and A. Hilson. “Idiopathic Parkinsonism Treated with an Extracerebral Decarboxylase Inhibitor in Combination with Levodopa”. *Br Med J.* 3.5777 (1971), pp. 729–732.

[63] D. A. Figge, K. L. E. Jaunarajs, and D. G. Standaert. “Dynamic DNA Methylation Regulates Levodopa-Induced Dyskinesia”. *J Neurosci.* 36.24 (2016), pp. 6514–6524.

[64] G. B. Baker, R. T. Coutts, K. F. McKenna, and R. L. Sherry-McKenna. “Insights into the mechanisms of action of the MAO inhibitors phenelzine and tranylcypromine: a review”. *J Psychiatry Neurosci.* 17.5 (1992), pp. 206–214.

[65] P. Riederer and G. Laux. “MAO-inhibitors in Parkinson’s Disease”. *Exp Neurobiol.* 20.1 (2011), pp. 1–17.

[66] C. V. Verschuur et al. “Randomized Delayed-Start Trial of Levodopa in Parkinson’s Disease”. *New England Journal of Medicine* 380.4 (2019), pp. 315–324.

[67] V. Ghiglieri, V. Calabrese, and P. Calabresi. “Alpha-Synuclein: From Early Synaptic Dysfunction to Neurodegeneration”. *Front Neurol.* 9 (2018), p. 295.

[68] H. A. Lashuel, C. R. Overk, A. Oueslati, and E. Masliah. “The many faces of -synuclein: from structure and toxicity to therapeutic target”. *Nat Rev Neurosci.* 14.1 (2013), pp. 38–48.

[69] F. Samuel, W. P. Flavin, S. Iqbal, C. Pacelli, S. D. S. Renganathan, L.-E. Trudeau, E. M. Campbell, P. E. Fraser, and A. Tandon. “Effects of Serine 129 Phosphorylation on -Synuclein Aggregation, Membrane Association, and Internalization”. *Nat Rev Neurosci.* 291.9 (2016), pp. 4374–4385.

[70] K. Wakabayashi, K. Tanji, S. Odagiri, Y. Miki, F. Mori, and H. Takahashi. “The Lewy body in Parkinson’s disease and related neurodegenerative disorders.” *Molecular neurobiology* 47.2 (2013), pp. 495–508.

[71] J. H. T. Power, O. L. Barnes, and F. Chegini. “Lewy Bodies and the Mechanisms of Neuronal Cell Death in Parkinson’s Disease and Dementia with Lewy Bodies”. *Brain Pathology* 27.1 (2017), pp. 3–12.

[72] S. Selvaraj and S. Piramanayagam. “Impact of gene mutation in the development of Parkinson’s disease”. *Genes and Diseases* 6.2 (2019), pp. 120–128.

[73] D. Bazazeh, R. M. Shubair, and W. Q. Malik. “Biomarker Discovery and Validation for Parkinson’s Disease: A Machine Learning Approach”. In: *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*. 2016.

[74] U. Saeed, J. Compagnone, R. I. Aviv, A. P. Strafella, S. E. Black, A. E. Lang, and M. Masellis. “Imaging biomarkers in Parkinson’s disease and Parkinsonian syndromes: Current and emerging concepts”. *Translational Neurodegeneration* 6 (2017), Article number: 8.

[75] J. L. Cummings, C. Henchcliffe, S. Schaier, T. Simuni, A. Waxman, and P. Kemp. “The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration”. *Brain* 134.11 (2011), pp. 3146–3166.

[76] F. Sampedro, J. Marín-Lahoz, S. Martínez-Horta, V. Camacho, D.-A. Lopez-Mora, J. Pagonabarraga, and J. Kulisevsky. “Extra-striatal SPECT-DAT uptake correlates with clinical and biological features of de novo Parkinson’s disease”. *Neurobiology of Aging* 97 (2020), pp. 120–128.

[77] H. Oikawa, M. Sasaki, Y. Tamakawa, S. Ehara, and K. Tohyama. “The substantia nigra in Parkinson disease: Proton density-weighted spin-echo and fast short inversion time inversion-recovery MR findings”. *American Journal of Neuroradiology* 23.10 (2002), pp. 1747–1756.

[78] P. Péran et al. “Magnetic resonance imaging markers of Parkinson’s disease nigrostriatal signature”. *Brain* 133.11 (2010), pp. 3423–3433.

[79] L. Minati, M. Grisoli, F. Carella, T. De Simone, M. G. Bruzzone, and M. Savoiardo. “Imaging degeneration of the substantia Nigra in Parkinson disease with inversion-recovery MR imaging”. *American Journal of Neuroradiology* 28.2 (2007), pp. 309–313.

[80] D. H. Kwon, J. M. Kim, S. H. Oh, H. J. Jeong, S. Y. Park, E. S. Oh, J. G. Chi, Y. B. Kim, B. S. Jeon, and Z. H. Cho. “Seven-tesla magnetic resonance images of the substantia nigra in Parkinson disease”. *Annals of Neurology* 71.2 (2012), pp. 267–277.

[81] A. Atik, T. Stewart, and J. Zhang. “Alpha-Synuclein as a Biomarker for Parkinson’s Disease”. *Brain Pathology* 26.3 (2018), pp. 410–418.

[82] M. Waragai, J. Wei, M. Fujita, M. Nakai, G. J. Ho, E. Masliah, H. Akatsu, T. Yamada, and M. Hashimoto. “Increased level of DJ-1 in the cerebrospinal fluids of sporadic Parkinson’s disease”. *Biochemical and Biophysical Research Communications* 345.3 (2006), pp. 967–972.

[83] P. T. Kotzbauer, N. J. Cairns, M. C. Campbell, A. W. Willis, B. A. Racette, S. D. Tabbal, and J. S. Perlmutter. “Pathologic accumulation of -synuclein and Aβ in Parkinson disease patients with dementia”. *Arch Neurol.* 69.10 (2012), pp. 1326–1331.

[84] X. Hu, Y. Yang, and D. Gong. “Changes of cerebrospinal fluid Aβ42, t-tau, and p-tau in Parkinson’s disease patients with cognitive impairment relative to those with normal cognition: a meta-analysis”. *Neurol Sci.* 38.11 (2017), pp. 1953–1961.

[85] T. J. Montine et al. “CSF Aβ42 and tau in Parkinson’s disease with cognitive impairment”. *Movement Disorders* 25.15 (2010), pp. 2682–2685.

[86] A. M. Fagan, M. A. Mintun, A. R. Shah, P. Aldea, C. M. Roe, R. H. Mach, D. Marcus, J. C. Morris, and D. M. Holtzman. “Cerebrospinal fluid tau and ptau181 increase with cortical amyloid deposition in cognitively normal individuals: Implications for future clinical trials of Alzheimer’s disease”. *EMBO Molecular Medicine* 1.8-9 (2009), pp. 371–380.

[87] A. Gorostidi et al. “α-Synuclein Levels in Blood Plasma from LRRK2 Mutation Carriers”. *PLoS ONE* 7.12 (2012), e52312.

[88] D. Besong-Agbo et al. “Naturally occurring α-synuclein autoantibody levels are lower in patients with Parkinson disease”. *Neurology* 80.2 (2013), pp. 169–175.

[89] R. Duran, F. J. Barrero, B. Morales, J. D. Luna, M. Ramirez, and F. Vives. “Plasma a-Synuclein in Patients with Parkinson’s Disease With and Without Treatment”. *Movement Disorders* 25.4 (2010), pp. 489–493.

[90] P. H. Lee, G. Lee, H. J. Park, O. Y. Bang, I. S. Joo, and K. Huh. “The plasma alpha-synuclein levels in patients with Parkinson’s disease and multiple system atrophy”. *Journal of Neural Transmission* 113.10 (2006), pp. 1435–1439.

[91] L. M. Smith, M. C. Schiess, M. P. Coffey, A. C. Klaver, and D. A. Loeffler. “α-Synuclein and Anti-α-Synuclein Antibodies in Parkinson’s Disease, Atypical Parkinson Syndromes, REM Sleep Behavior Disorder, and Healthy Controls”. *PLoS ONE* 7.12 (2012), e52285.

[92] P. G. Foulds, J. D. Mitchell, A. Parker, R. Turner, G. Green, P. Diggle, M. Hasegawa, M. Taylor, D. Mann, and D. Allsop. “Phosphorylated α-synuclein can be detected in blood plasma and is potentially a useful biomarker for Parkinson’s disease”. *The FASEB Journal* 25.12 (2011), pp. 4127–4137.

[93] M. J. Park, S. M. Cheon, H. R. Bae, S. H. Kim, and J. W. Kim. “Elevated levels of α-synuclein oligomer in the cerebrospinal fluid of drug-naïve patients with Parkinson’s disease”. *Journal of Clinical Neurology (Korea)* 7.4 (2011), pp. 215–222.

[94] C. Ballard, I. Ziabreva, R. Perry, J. P. Larsen, J. O’Brien, I. McKeith, E. Perry, and D. Aarsland. “Differences in neuropathologic characteristics across the Lewy body dementia spectrum”. *Neurology* 67.11 (2006), pp. 1931–1934.

[95] M. Waragai, M. Nakai, J. Wei, M. Fujita, H. Mizuno, G. Ho, E. Masliah, H. Akatsu, F. Yokochi, and M. Hashimoto. “Plasma levels of DJ-1 as a possible marker for progression of sporadic Parkinson’s disease”. *Neuroscience Letters* 425.1 (2007), pp. 18–22.

[96] C. Maita, S. Tsuji, I. Yabe, S. Hamada, A. Ogata, H. Maita, S. M. Iguchi-Ariga, H. Sasaki, and H. Ariga. “Secretion of DJ-1 into the serum of patients with Parkinson’s disease”. *Neuroscience Letters* 431.1 (2008), pp. 86–89.

[97] C. An, X. Pu, W. Xiao, and H. Zhang. “Expression of the DJ-1 protein in the serum of Chinese patients with Parkinson’s disease”. *Neuroscience Letters* 665 (2018), pp. 236–239.

[98] M. Shi et al. “Significance and confounders of peripheral DJ-1 and alpha-synuclein in Parkinson’s disease”. *Neuroscience Letters* 480.1 (2010), pp. 78–82.

[99] E. Angelopoulou, Y. N. Paudel, and C. Piperi. “miR-124 and Parkinson’s disease: A biomarker with therapeutic potential”. *Pharmacological Research* 150 (2019), p. 104515.

[100] C. A. Ross and S. J. Tabrizi. “Huntington’s disease: From molecular pathogenesis to clinical treatment”. *The Lancet Neurology* 10.1 (2011), pp. 83–98.

[101] A. D. Ha and V. S. Fung. “Huntington’s disease”. *Current Opinion in Neurology* 25.4 (2012), pp. 491–498.

[102] D. R. Langbehn et al. “CAG-repeat length and the age of onset in Huntington Disease (HD): A review and validation study of statistical approaches”. *American*

Journal of Medical Genetics, Part B: Neuropsychiatric Genetics 153B.2 (2010), pp. 397–408.

[103] P. McColgan and S. J. Tabrizi. "Huntington's disease: a clinical review". *European Journal of Neurology* 25.1 (2018), pp. 24–34.

[104] C. A. Ross et al. "Huntington disease: Natural history, biomarkers and prospects for therapeutics". *Nature Reviews Neurology* 10.4 (2014), pp. 204–216.

[105] E. J. Wild and S. Tabrizi. "Therapies targeting DNA and RNA in Huntington's disease". *Lancet Neurol.* 16.10 (2017), pp. 837–847.

[106] H. B. Kordasiewicz et al. "Sustained therapeutic reversal of Huntington's disease by transient repression of huntingtin synthesis". *Neuron* 74.6 (2013), pp. 1031–1044.

[107] M. DiFiglia et al. "Therapeutic silencing of mutant huntingtin with siRNA attenuates striatal and cortical neuropathology and behavioral deficits". *Proceedings of the National Academy of Sciences of the United States of America* 104.43 (2007), pp. 17204–17209.

[108] S. Yang et al. "CRISPR/Cas9-mediated gene editing ameliorates neurotoxicity in mouse model of Huntington's disease". *The Journal of Clinical Investigation* 127.7 (2016), pp. 2719–2724.

[109] J. P. Vonsattel, R. H. Myers, T. J. Stevens, R. J. Ferrante, E. D. Bird, and E. P. Richardson. "Neuropathological classification of Huntington's disease". *Journal of Neuropathology and Experimental Neurology* 44.6 (1985), pp. 559–577.

[110] C. Yapijakis. "Huntington Disease: Genetics, Prevention, and Therapy Approaches". *Advances in Experimental Medicine and Biology* 987 (2017), pp. 55–65.

[111] Y. Jiang, S. R. Chadwick, and P. Lajoie. "Endoplasmic reticulum stress: The cause and solution to Huntington's disease?" *Brain Research* 1648.Part B (2016), pp. 650–657.

[112] C. Zuccato, M. Valenza, and E. Cattaneo. "Molecular mechanisms and potential therapeutical targets in Huntington's disease". *Physiological Reviews* 90.3 (2010), pp. 905–981.

[113] D. Weir, A. Sturrock, and B. Leavitt. “Development of biomarkers for Huntington’s disease”. *The Lancet Neurology* 10.6 (2011), pp. 573–590.

[114] P. Fazio, M. Paucar, P. Svenningsson, and A. Varrone. “Novel Imaging Biomarkers for Huntington’s Disease and Other Hereditary Choreas”. *Current Neurology and Neuroscience Reports* 18.12 (2018), p. 85.

[115] J. S. Paulsen et al. “Detection of Huntington’s disease decades before diagnosis: The Predict-HD study”. *Journal of Neurology, Neurosurgery and Psychiatry* 79.8 (2008), pp. 874–880.

[116] S. J. Tabrizi et al. “Biological and clinical manifestations of Huntington’s disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data”. *The Lancet Neurology* 8.9 (2009), pp. 791–801.

[117] G. J. Harris, A. M. Codori, R. F. Lewis, E. Schmidt, A. Bedi, and J. Brandt. “Reduced basal ganglia volume associated with the gene for Huntington’s disease in asymptomatic at-risk persons”. *Brain* 122.9 (1999), pp. 1667–1678.

[118] S. J. Tabrizi et al. “Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington’s disease in the TRACK-HD study: Analysis of 36-month observational data”. *The Lancet Neurology* 12.7 (2013), pp. 637–649.

[119] E. H. Aylward, P. C. Nopoulos, C. A. Ross, D. R. Langbehn, R. K. Pierson, J. A. Mills, H. J. Johnson, V. A. Magnotta, A. R. Juhl, and J. S. Paulsen. “Longitudinal change in regional brain volumes in prodromal Huntington disease”. *Journal of Neurology, Neurosurgery and Psychiatry* 82.4 (2011), pp. 405–410.

[120] D. A. Majid, A. R. Aron, W. Thompson, S. Sheldon, S. Hamza, D. Stoffers, D. Holland, J. Goldstein, J. Corey-Bloom, and A. M. Dale. “Basal ganglia atrophy in prodromal Huntington’s disease is detectable over one year using automated segmentation”. *Movement Disorders* 26.14 (2011), pp. 2544–2551.

[121] A. Feigin, C. Tang, Y. Ma, P. Mattis, D. Zgaljardic, M. Guttman, J. S. Paulsen, V. Dhawan, and D. Eidelberg. “Thalamic metabolism and symptom onset in preclinical Huntington’s disease”. *Brain* 130.11 (2007), pp. 2858–2867.

[122] O. F. Odish, K. Johnsen, P. van Someren, R. A. Roos, and J. G. van Dijk. "EEG may serve as a biomarker in Huntington's disease using machine learning automatic classification". *Scientific Reports* 8.1 (2018), p. 16090.

[123] F. Niccolini et al. "Altered PDE10A expression detectable early before symptomatic onset in Huntington's disease". *Brain* 138.10 (2015), pp. 3016–3029.

[124] D. S. Russell et al. "The phosphodiesterase 10 positron emission tomography tracer, [18F]MNI-659, as a novel biomarker for early huntington disease". *JAMA Neurology* 71.12 (2014), pp. 1520–1528.

[125] D. S. Russell et al. "Change in PDE10 across early Huntington disease assessed by [18 F]MNI-659 and PET imaging". *Neurology* 86.8 (2016), pp. 748–754.

[126] L. M. Byrne, F. B. Rodrigues, E. B. Johnson, E. De Vita, K. Blennow, R. Scahill, H. Zetterberg, A. Heslegrave, and E. J. Wild. "Cerebrospinal fluid neurogranin and TREM2 in Huntington's disease". *Scientific Reports* 8 (2018), Article number: 4260.

[127] E. J. Wild et al. "Quantification of mutant huntingtin protein in cerebrospinal fluid from Huntington's disease patients". *Journal of Clinical Investigation* 125.5 (2015), pp. 1979–1986.

[128] Z. Tan et al. "Huntington's disease cerebrospinal fluid seeds aggregation of mutant huntingtin". *Molecular Psychiatry* 20.11 (2015), pp. 1286–1293.

[129] A. Weiss et al. "Mutant huntingtin fragmentation in immune cells tracks Huntington's disease progression". *Journal of Clinical Investigation* 122.10 (2012), pp. 3731–3736.

[130] L. M. Byrne et al. "Neurofilament light protein in blood as a potential biomarker of neurodegeneration in Huntington's disease: a retrospective cohort analysis". *The Lancet Neurology* 16.8 (2017), pp. 601–609.

[131] G. Ellrichmann, C. Reick, C. Saft, and R. a. Linker. "The Role of the Immune System in Huntington's Disease". *Clinical and Developmental Immunology* 2013 (2013), Article ID: 541259.

[132] J. A. Bouwens, E. Van Duijn, C. M. Cobbaert, R. A. Roos, R. C. Van Der Mast, and E. J. Giltay. "Plasma Cytokine Levels in Relation to Neuropsychiatric Symptoms and Cognitive Dysfunction in Huntington's disease". *Journal of Huntington's Disease* 5.4 (2016), pp. 369–377.

[133] F. Leblhuber, J. Walli, K. Jellinger, G. P. Tilz, B. Widner, F. Laccone, and D. Fuchs. "Activated immune system in patients with Huntington's disease". *Clinical Chemistry and Laboratory Medicine* 36.10 (1998), pp. 747–750.

[134] D. A. Lashkari, J. L. Derisi, J. H. Mccusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. "Yeast microarrays for genome wide parallel genetic and gene expression analysis". *Proceedings of the National Academy of Sciences of the United States of America* 94.24 (1997), pp. 13057–13062.

[135] R. Bumgarner. "Overview of DNA Microarrays: Types, Applications, and Their Future". *Current Protocols in Molecular Biology* 101.1 (2013), pp. 22.1.1–22.1.11.

[136] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. "Overview of DNA Microarrays: Types, Applications, and Their Future". *Current Protocols in Molecular Biology* 122.1 (2018), e59.

[137] R. Stark, M. Grzelak, and J. Hadfield. "RNA sequencing: the teenage years". *Nature Reviews Genetics* (2019).

[138] M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". *Molecular Systems Biology* 15.6 (2019), e8746.

[139] A. Conesa et al. "A survey of best practices for RNA-seq data analysis". *Genome Biology* 17 (2016), Article number: 13.

[140] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. "affy—analysis of Affymetrix GeneChip data at the probe level". *Bioinformatics* 20.3 (2004), pp. 307–315.

[141] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. "Summaries of Affymetrix GeneChip probe level data". *Nucleic acids research* 31.4 (2003), e15.

[142] J. H. Do and D. K. Choi. "Normalization of microarray data: Single-labeled and dual-labeled arrays". *Molecules and Cells* 22.3 (2006), pp. 254–261.

[143] C. R. Williams, A. Baccarella, J. Z. Parrish, and C. C. Kim. “Trimming of sequence reads alters RNA-Seq gene expression estimates”. *BMC Bioinformatics* 17 (2016), p. 103.

[144] P. Pérez-Rubio, C. Lottaz, and J. C. Engelmann. “FastqPuri: High-performance preprocessing of RNA-seq data”. *BMC Bioinformatics* 20 (2019), p. 226.

[145] A. Dobin, T. R. Gingeras, C. Spring, R. Flores, J. Sampson, R. Knight, N. Chia, and H.-t. S. Technologies. “Mapping RNA-seq with STAR”. *Curr Protoc Bioinformatics* 51.4 (2016), pp. 586–597.

[146] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. *Genome Biology* 14 (2013), R36.

[147] S. Anders, P. T. Pyl, and W. Huber. “HTSeq-A Python framework to work with high-throughput sequencing data”. *Bioinformatics* 31.2 (2015), pp. 166–169.

[148] Y. Liao, G. K. Smyth, and W. Shi. “FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features”. *Bioinformatics* 30.7 (2014), pp. 923–930.

[149] R. Patro, S. M. Mount, and C. Kingsford. “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms”. *Nature Biotechnology* 32.5 (2014), pp. 462–464.

[150] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. “Near-optimal probabilistic RNA-seq quantification”. *Nature Biotechnology* 34.5 (2016), pp. 525–527.

[151] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. “Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference”. *Nature Methods* 14.4 (2017), pp. 417–419.

[152] Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou. “Advanced applications of RNA sequencing and challenges”. *Bioinformatics and Biology Insights* 9.Suppl 1 (2015), pp. 29–46.

[153] S. Anders and W. Huber. “Differential expression analysis for sequence count data”. *Genome Biology* 11.10 (2010), R106.

[154] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology* 15 (2014), Article number: 550.

[155] N. J. Schurch et al. “Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment”. *arXiv e-prints*, arXiv:1505.02017 (May 2015), arXiv:1505.02017.

[156] N. J. Schurch et al. “How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?” *RNA* 22.6 (2016), pp. 839–851.

[157] J. Costa-Silva, D. Domingues, and F. M. Lopes. “RNA-Seq differential expression analysis: An extended review and a software tool”. *PLoS ONE* 12.12 (2017), e0190152.

[158] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology* 15 (12 2014), p. 550.

[159] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. “Combining multiple microarray studies and modeling interstudy variation”. *Bioinformatics* 19.Suppl 1 (2003), pp. 84–90.

[160] D. Toro-Domínguez, J. A. Villatoro-García, J. Martorell-Marugán, Y. Román-Montoya, M. E. Alarcón-Riquelme, and P. Carmona-Sáez. “A survey of gene expression meta-analysis: methods and applications”. *Briefings in Bioinformatics* (2020), bbaa019.

[161] G. Borrageiro, W. Haylett, S. Seedat, H. Kuivaniemi, and S. Bardien. “A review of genome-wide transcriptomics studies in Parkinson’s disease”. *European Journal of Neuroscience* 47.1 (2018), pp. 1–16.

[162] X. Li, J. Long, T. He, R. Belshaw, and J. Scott. “Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer’s disease”. *Scientific Reports* 5 (2015), p. 12393.

[163] Y. Feng and X. Wang. “Systematic analysis of microarray datasets to identify Parkinson’s disease-associated pathways and genes”. *Molecular Medicine Reports* 15.3 (2017), pp. 1252–1262.

[164] M. Cruz-Monteagudo, F. Borges, C. Paz-Y-Miño, M. N. D. Cordeiro, I. Rebelo, Y. Perez-Castillo, A. M. Helguera, A. Sánchez-Rodríguez, and E. Tejera. “Efficient and biologically relevant consensus strategy for Parkinson’s disease gene prioritization”. *BMC Medical Genomics* 9 (2016), p. 12.

[165] J. A. Miller, M. C. Oldham, and D. H. Geschwind. “A Systems Level Analysis of Transcriptional Changes in Alzheimer’s Disease and Normal Aging”. *The Journal of Neuroscience* 28.6 (2008), pp. 1410–1420.

[166] P. Langfelder, B. Zhang, and S. Horvath. “Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R”. *Bioinformatics* 24.5 (2008), pp. 719–720.

[167] D. J. Allocco, I. S. Kohane, and A. J. Butte. “Quantifying the relationship between co-expression, co-regulation and gene function”. *BMC Bioinformatics* 5 (2004), Article number: 18.

[168] M. T. Weirauch. “Gene Coexpression Networks for the Analysis of DNA Microarray Data”. In: *Applied Statistics for Network Biology*. John Wiley & Sons, Ltd, 2011. Chap. 11, pp. 215–250.

[169] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang. “Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types”. *Nature Communications* 5 (2014), Article number: 3231.

[170] K. Lunnon et al. “Mitochondrial dysfunction and immune activation are detectable in early Alzheimer’s disease blood”. *Journal of Alzheimer’s Disease* 30.3 (2012), pp. 685–710.

[171] S. Chen, D. Yang, C. Lei, Y. Li, X. Sun, M. Chen, X. Wu, and Y. Zheng. “Identification of crucial genes in abdominal aortic aneurysm by WGCNA”. *PeerJ* 7 (2019), e7873.

[172] J. A. Botía, J. Vandrovcova, P. Forabosco, S. Guelfi, K. D'Sa, The United Kingdom Brain Expression Consortium, J. Hardy, C. M. Lewis, M. Ryten, and M. E. Weale. "An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks". *BMC Systems Biology* 11 (2017), p. 47.

[173] K. Shastry and H. Sanjay. "Machine Learning for Bioinformatics". In: *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*. 2020.

[174] M. A. Myszczynska, P. N. Ojamies, A. M. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo. "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases". *Nature Reviews Neurology* 16.8 (2020), pp. 440–456.

[175] X. Zhang, J. Zhang, K. Sun, X. Yang, C. Dai, and Y. Guo. "Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, pp. 765–769.

[176] R. Liu, C. A. Mancuso, A. Yannakopoulos, K. A. Johnson, and A. Krishnan. "Supervised learning is an accurate method for network-based gene classification". *Bioinformatics* 36.11 (Apr. 2020), pp. 3457–3465.

[177] T. G. Nick and K. M. Campbell. "Logistic Regression". *Methods in Molecular Biology* 404 (2007), pp. 273–301.

[178] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[179] K. Veropoulos, C. Campbell, and N. Cristianini. "Controlling the Sensitivity of Support Vector Machines". *Proceedings of International Joint Conference Artificial Intelligence* (June 1999).

[180] W. A. Belson. "Matching and Prediction on the Principle of Biological Classification". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 8.2 (1959), pp. 65–75.

[181] L. Breiman. "Random forests". *Machine Learning* 45 (2001), pp. 5–32.

[182] R. Toth et al. “Random forest-based modelling to detect biomarkers for prostate cancer progression”. *Clinical Epigenetics* 11 (2019), p. 148.

[183] J. H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. *The Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[184] T. Chen and C. Guestrin. “XGBoost”. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016).

[185] C. Van Der Malsburg. “Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms”. In: *Brain Theory*. Ed. by G. Palm and A. Aertsen. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 245–248.

[186] D. Marques, A. Barradas Filho, A. Romariz, I. Viegas, D. Luz, A. K. Barros Filho, S. Labidi, and A. Ferraudo. “Recent Developments on Statistical and Neural Network Tools Focusing on Biodiesel Quality”. *International Journal of Computer Science and Application* 3 (Jan. 2014), p. 97.

[187] P. Ramachandran, B. Zoph, and Q. V. Le. *Searching for Activation Functions*. 2017.

[188] M. Kon and L. Plaskota. “Information complexity of neural networks”. *Neural Networks* 13.3 (2000), pp. 365–375.

[189] L. Camargo and T. Yoneyama. “Specification of Training Sets and the Number of Hidden Neurons for Multilayer Perceptrons”. *Neural Computation* 13 (Dec. 2001), pp. 2673–2680.

[190] J. Y. F. Yam and T. W. S. Chow. “Feedforward networks training speed enhancement by optimal initialization of the synaptic coefficients”. *IEEE Transactions on Neural Networks* 12.2 (2001), pp. 430–434.

[191] D. Zimmerer, S. A. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein. *Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection*. 2018.

[192] H. Xu et al. “Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications”. In: *Proceedings of the 2018 World Wide Web Conference*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 187–196.

[193] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto. “A Convolutional Autoencoder Approach for Feature Extraction in Virtual Metrology”. *Procedia Manufacturing* 17 (Jan. 2018), pp. 126–133.

[194] T. Lee and H. Lee. “Prediction of Alzheimer’s disease using blood gene expression data”. *Scientific Reports* 10 (2020), Article number: 3485.

[195] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe. “Deep learning with convolutional neural network in radiology”. *Japanese Journal of Radiology* 36.4 (2018), pp. 257–272.

[196] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen. “Convolutional neural network models for cancer type prediction based on gene expression”. *BMC Medical Genomics* 13 (2020), Article number: 44.

[197] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang. “Deep Learning for Health Informatics”. *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21.

[198] R. Singh, J. Lanchantin, G. Robins, and Y. Qi. “DeepChrome: deep-learning for predicting gene expression from histone modifications”. *Bioinformatics* 32.17 (Aug. 2016), pp. i639–i648.

[199] V. Kiselev, T. Andrews, and M. Hemberg. “Challenges in unsupervised clustering of single-cell RNA-seq data.” *Nature Reviews. Genetics* 20.5 (2019-01-01 00:00:00.002), pp. 273–282.

[200] X. Yu, G. Yu, and J. Wang. “Clustering cancer gene expression data by projective clustering ensemble”. *PLOS ONE* 12.2 (Feb. 2017), pp. 1–21.

[201] K. Daniels and C. Giraud-Carrier. “Learning the Threshold in Hierarchical Agglomerative Clustering”. In: *2006 5th International Conference on Machine Learning and Applications (ICMLA’06)*. 2006, pp. 270–278.

[202] I. T. Jollife and J. Cadima. “Principal component analysis: A review and recent developments”. *Philosophical Transactions of the Royal Society A* 374 (2016), p. 20150202.

[203] D. Kobak and P. Berens. “The art of using t-SNE for single-cell transcriptomics”. *Nature Communications* 10 (2019), Article number: 5416.

[204] L. van der Maaten and G. Hinton. “Visualizing Data using t-SNE”. *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[205] J. Leevy, T. Khoshgoftaar, R. Bauder, and N. Seliya. “A survey on addressing high-class imbalance in big data”. *Journal of Big Data* 5 (Nov. 2018), Article number: 42.

[206] T. Fawcett. “An introduction to ROC analysis”. *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874.

[207] J. Davis and M. Goadrich. “The Relationship between Precision-Recall and ROC Curves”. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML ’06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233–240.

[208] M. Stone. “Cross-Validatory Choice and Assessment of Statistical Predictions”. *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 111–147.

[209] X. Zhang, T. Zhou, L. Zhang, K. Y. Fung, and K. M. Ng. “Food Product Design: A Hybrid Machine Learning and Mechanistic Modeling Approach”. *Industrial & Engineering Chemistry Research* 58.36 (2019), pp. 16743–16752.

[210] T. T. Joy, S. Rana, S. Gupta, and S. Venkatesh. “Hyperparameter tuning for big data using Bayesian optimisation”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016, pp. 2574–2579.

[211] H. Shaziya and R. Zaheer. “Impact of Hyperparameters on Model Development in Deep Learning”. In: *Proceedings of International Conference on Computational Intelligence and Data Engineering*. Ed. by N. Chaki, J. Pejas, N. Devarakonda, and R. M. Rao Kovvur. Singapore: Springer Singapore, 2021, pp. 57–67.

[212] B. Kumari and T. Swarnkar. “Filter versus wrapper feature subset selection in large dimensionality micro array: A review”. *International Journal of Computer Science and Information Technologies* 2.3 (2011), pp. 1048–1053.

[213] Z. Li, W. Xie, and T. Liu. “Efficient feature selection and classification for microarray data”. *PLoS ONE* 13.8 (2018), e0202167.

[214] R. Muthukrishnan and R. Rohini. “LASSO: A feature selection technique in predictive modeling for machine learning”. In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. 2016, pp. 18–20.

[215] X. Feng, S. Luo, and B. Lu. “Conformation Polymorphism of Polyglutamine Proteins”. *Trends in Biochemical Sciences* 43.6 (2018), pp. 424–435.

[216] S. J. Tabrizi et al. “Targeting Huntingtin Expression in Patients with Huntington’s Disease”. *New England Journal of Medicine* 380.24 (2019), pp. 2307–2316.

[217] E. Valionyte, Y. Yang, S. L. Roberts, J. Kelly, B. Lu, and S. Luo. “Lowering Mutant Huntingtin Levels and Toxicity: Autophagy-Endolysosome Pathways in Huntington’s Disease”. *Journal of Molecular Biology* 432.8 (2020). Autophagy in Neurodegenerative Diseases, pp. 2673–2691.

[218] A. Neueder and G. P. Bates. “A common gene expression signature in Huntington’s disease patient brain regions”. *BMC Medical Genomics* 7 (2014), p. 60.

[219] E. Mina, W. Van Roon-Mom, K. Hettne, E. Van Zwet, J. Goeman, C. Neri, P. A. Hoen, B. Mons, and M. Roos. “Common disease signatures from gene expression analysis in Huntington’s disease human blood and brain”. *Orphanet Journal of Rare Diseases* 11 (2016), p. 97.

[220] A. Labadorf et al. “RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression”. *PLoS ONE* 10.12 (2015), e0143563.

[221] C. Soneson, M. I. Love, and M. D. Robinson. “Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences”. *F1000Research* 4 (2015), p. 1521.

[222] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. “Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt”. *Nature Protocols* 4.8 (2009), pp. 1184–1191.

[223] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. “BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis”. *Bioinformatics* 21.16 (2005), pp. 3439–3440.

[224] X. Wang, Y. Lin, C. Song, E. Sibille, and G. C. Tseng. “Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder”. *BMC Bioinformatics* 13 (2012), p. 52.

[225] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. “Data-driven hypothesis weighting increases detection power in genome-scale multiple testing”. *Nature Methods* 13.7 (2016), pp. 577–580.

[226] P. DI Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. “Nextflow enables reproducible computational workflows”. *Nature Biotechnology* 35.4 (2017), pp. 316–319.

[227] S. Mubeen, C. T. Hoyt, A. Gemünd, M. Hofmann-Apitius, H. Fröhlich, and D. Domingo-Fernández. “The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling”. *Frontiers in Genetics* 10 (2019), p. 1203.

[228] M. V. Kuleshov et al. “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”. *Nucleic Acids Research* 44 (2016), W90–W97.

[229] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma’ayan. “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. *BMC Bioinformatics* 14 (2013), Article Number: 128.

[230] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, and D. Telikicherla. “Human Protein Reference Database–2009 update”. *Nucleic Acids Res.* 37 (2009), pp. D767–772.

[231] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks”. *Genome Research* 13.11 (2003), pp. 2498–2504.

[232] J. Schindelin et al. “Fiji: An open-source platform for biological-image analysis”. *Nature Methods* 9.7 (2012), pp. 676–682.

[233] C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri. “ImageJ2: ImageJ for the next generation of scientific image data”. *BMC Bioinformatics* 18 (2017), p. 529.

[234] Y. Zhang et al. “Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse”. *Neuron* 89.1 (2016), pp. 37–53.

[235] J. Qu, L. Zhu, Z. Zhou, P. Chen, S. Liu, M. L. Locy, V. J. Thannickal, and Y. Zhou. “Reversing mechanoinductive DSP expression by CRISPR/dCas9–mediated epigenome editing”. *American Journal of Respiratory and Critical Care Medicine* 198.5 (2018), pp. 599–609.

[236] D. Scarabino, L. Veneziano, M. Peconi, M. Frontali, E. Mantuano, and R. M. Corbo. “Leukocyte telomere shortening in Huntington’s disease”. *Journal of the Neurological Sciences* 396 (2019), pp. 25–29.

[237] P. Li, Y. Meng, Y. Wang, J. Li, M. Lam, L. Wang, and L.-J. Di. “Nuclear localization of desmoplakin and its involvement in telomere maintenance”. *International Journal of Biological Sciences* 15.11 (2019), pp. 2350–2362.

[238] H. Wang, M. Wu, Y. Lu, K. He, X. Cai, X. Yu, J. Lu, and L. Teng. “LncRNA MIR4435-2HG targets desmoplakin and promotes growth and metastasis of gastric cancer by activating Wnt/β-catenin signaling”. *Aging* 11.17 (2019), pp. 6657–6673.

[239] C. Smith-Geater et al. “Aberrant Development Corrected in Adult-Onset Huntington’s Disease iPSC-Derived Neuronal Cultures via WNT Signaling Modulation”. *Stem Cell Reports* 14.3 (2020), pp. 406–419.

[240] Q. Li et al. “Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing”. *Neuron* 101.2 (2019), 207–223.e10.

[241] H. M. Yang, S. Yang, S. S. Huang, B. S. Tang, and J. F. Guo. “Microglial activation in the pathogenesis of Huntington’s Disease”. *Frontiers in Aging Neuroscience* 9 (2017), p. 193.

[242] W. Kim, M. H. Cho, P. Sakornsakolpat, D. A. Lynch, H. O. Coxson, R. Tal-Singer, E. K. Silverman, and T. H. Beaty. “DSP variants may be associated with longitudinal change in quantitative emphysema”. *Respiratory Research* 20 (2019), p. 160.

[243] P. W. Holland. “Evolution of homeobox genes”. *Wiley Interdisciplinary Reviews: Developmental Biology* 2.1 (2013), pp. 31–45.

[244] R. A. Alharbi, R. Pettengell, H. S. Pandha, and R. Morgan. “The role of HOX genes in normal hematopoiesis and acute leukemia”. *Leukemia* 27.5 (2013), pp. 1000–1008.

[245] A. G. Hoss et al. “MicroRNAs Located in the Hox Gene Clusters Are Implicated in Huntington’s Disease Pathogenesis”. *PLoS Genetics* 10.2 (2014), e1004188.

[246] A. Labadorf, S. H. Choi, and R. H. Myers. “Evidence for a pan-neurodegenerative disease response in Huntington’s and Parkinson’s disease expression profiles”. *Frontiers in Molecular Neuroscience* 10 (2018), p. 430.

[247] J. S. Byun et al. “The transcription factor PITX1 drives astrocyte differentiation by regulating the SOX9 gene”. *The Journal of biological chemistry* 295.39 (2020), pp. 13677–13690.

[248] M. Oksanen et al. “NF-E2-related factor 2 activation boosts antioxidant defenses and ameliorates inflammatory and amyloid properties in human Presenilin-1 mutated Alzheimer’s disease astrocytes”. *Glia* 68.3 (2020), pp. 589–599.

[249] J. R. Liddell. “Are astrocytes the predominant cell type for activation of Nrf2 in aging and neurodegeneration?” *Antioxidants* 6.3 (2017), p. 65.

[250] J. Ma, T. Jiang, L. Tan, and J. T. Yu. “TYROBP in Alzheimer’s Disease”. *Molecular Neurobiology* 51.2 (2015), pp. 820–826.

[251] T. H. Palpagama, H. J. Waldvogel, R. L. Faull, and A. Kwakowsky. “The Role of Microglia and Astrocytes in Huntington’s Disease”. *Frontiers in Molecular Neuroscience* 12 (2019), p. 258.

[252] T. Liu, L. Zhang, D. Joo, and S. C. Sun. “NF-κB signaling in inflammation”. *Signal Transduction and Targeted Therapy* 2 (2017), e17023.

[253] E. Marcora and M. B. Kennedy. “The Huntington’s disease mutation impairs Huntingtin’s role in the transport of NF-κB from the synapse to the nucleus”. *Human Molecular Genetics* 19.22 (2010), pp. 4373–4384.

[254] P. A. C. Valadão, K. B. S. Santos, T. H. Ferreira e Vieira, T. Macedo e Cordeiro, A. L. Teixeira, C. Guatimosim, and A. S. de Miranda. “Inflammation in Huntington’s disease: A few new twists on an old tale”. *Journal of Neuroimmunology* 348 (2020), p. 577380.

[255] R. C. Wolf, N. Vasic, C. Schönfeldt-Lecuona, G. B. Landwehrmeyer, and D. Ecker. “Dorsolateral prefrontal cortex dysfunction in presymptomatic Huntington’s disease: Evidence from event-related fMRI”. *Brain* 130.11 (2007), pp. 2845–2857.

[256] J. Kelly, R. Moyeed, C. Carroll, D. Albani, and X. Li. “Gene expression meta-analysis of Parkinson’s disease and its relationship with Alzheimer’s disease”. *Molecular Brain* 12 (2019), Article Number: 16.

[257] A. Xie, J. Gao, L. Xu, and D. Meng. “Shared Mechanisms of Neurodegeneration in Alzheimer’s Disease and Parkinson’s Disease”. *BioMed Research International* 2014 (2014), Article ID: 648740.

[258] J. B. Anang, T. Nomura, S. R. Romenets, K. Nakashima, J.-f. Gagnon, and R. B. Postuma. “Dementia Predictors in Parkinson Disease: A Validation Study”. *Journal of Parkinson’s Disease* 7.1 (2017), pp. 159–162.

[259] A. Kaźmierczak, G. A. Czapski, A. Adamczyk, B. Gajkowska, and J. B. Strosznajder. “A novel mechanism of non-Aβ component of Alzheimer’s disease amyloid (NAC) neurotoxicity. Interplay between p53 protein and cyclin-dependent kinase 5 (Cdk5)”. *Neurochemistry International* 58.2 (2011), pp. 206–214.

[260] O. D. Kwon. “Is There Any Relationship between Apolipoprotein E Polymorphism and Idiopathic Parkinson’s Disease?” *Journal of Alzheimer’s Disease & Parkinsonism* 7 (2017), p. 292.

[261] X. Li, S. James, and P. Lei. “Interactions Between α-Synuclein and Tau Protein: Implications to Neurodegenerative Disorders”. *Journal of Molecular Neuroscience* 60.3 (2016), pp. 298–304.

[262] B. E. Riley et al. “Systems-based analyses of brain regions functionally impacted in Parkinson’s disease reveals underlying causal mechanisms”. *PLoS ONE* 9.8 (2014), e102909.

[263] A. Dumitriu, J. Golji, A. T. Labadorf, B. Gao, T. G. Beach, R. H. Myers, K. A. Longo, and J. C. Latourelle. “Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease”. *BMC Medical Genomics* 9 (2016), p. 5.

[264] H. Braak, K. Del Tredici, U. Rüb, R. A. De Vos, E. N. Jansen Steur, and E. Braak. “Staging of brain pathology related to sporadic Parkinson’s disease”. *Neurobiology of Aging* 24.2 (2003), pp. 197–211.

[265] Y. Zhang, M. James, F. A. Middleton, and R. L. Davis. “Transcriptional analysis of multiple brain regions in Parkinson’s disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms”. *American Journal of Medical Genetics - Neuropsychiatric Genetics* 137B.1 (2005), pp. 5–16.

[266] E. Oerton and A. Bender. “Concordance analysis of microarray studies identifies representative gene expression changes in Parkinson’s disease: a comparison of 33 human and animal studies”. *BMC neurology* 17.1 (2017), p. 58.

[267] B. Zheng et al. “PGC-1α, A Potential Therapeutic Target for Early Intervention in Parkinson’s Disease”. *Science Translational Medicine* 2.52 (2011), 52ra73.

[268] L. B. Moran, D. C. Duke, M. Deprez, D. T. Dexter, R. K. Pearce, and M. B. Graeber. “Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson’s disease”. *Neurogenetics* 7 (2006), pp. 1–11.

[269] E. Mariani, F. Frabetti, A. Tarozzi, M. C. Pelleri, F. Pizzetti, and R. Casadei. "Meta-analysis of Parkinson's disease transcriptome data using TRAM software: Whole substantia nigra tissue and single dopamine neuron differential gene expression". *PLoS ONE* 11.9 (2016), e0161567.

[270] J. Chi, Q. Xie, J. Jia, X. Liu, J. Sun, Y. Deng, and L. Yi. "Integrated analysis and identification of novel biomarkers in Parkinson's disease". *Frontiers in Aging Neuroscience* 10 (2018), p. 178.

[271] E. Mariani, L. Lombardini, F. Facchin, F. Pizzetti, F. Frabetti, A. Tarozzi, and R. Casadei. "Sex-specific transcriptome differences in substantia Nigra tissue: A meta-analysis of parkinson's disease data". *Genes* 9 (2018), p. 275.

[272] R Core Team. *R: A language and environment for statistical computing*. 2017.

[273] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. "Affy - Analysis of Affymetrix GeneChip data at the probe level". *Bioinformatics* 20.3 (2004), pp. 307–315.

[274] J. Gross and U. Ligges. *nortest: Tests for Normality*. 2015.

[275] L. J. Carithers et al. "A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project". *Biopreservation and Biobanking* 13.5 (2015), pp. 311–319.

[276] W. E. Johnson, C. Li, and A. Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* 8.1 (2007), pp. 118–127.

[277] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, and Y. Zhang. *sva: Surrogate Variable Analysis*. 2019.

[278] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* 43.7 (2015), e47.

[279] A. Krämer, J. Green, J. Pollard, and S. Tugendreich. "Causal analysis approaches in ingenuity pathway analysis". *Bioinformatics* 30.4 (2014), pp. 523–530.

[280] D. Chang et al. "A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci". *Nature Genetics* 49.10 (2017), pp. 1511–1516.

[281] V. Carelli et al. “Syndromic parkinsonism and dementia associated with OPA1 missense mutations”. *Annals of Neurology* 78.1 (2015), pp. 21–38.

[282] K. Pennington, T. Chan, M. Torres, and J. Andersen. “The dynamic and stress-adaptive signaling hub of 14-3-3: emerging mechanisms of regulation and context-dependent protein–protein interactions”. *Oncogene* 37 (2018), pp. 5587–5604.

[283] M. Foote and Y. Zhou. “14-3-3 Proteins in Neurological Disorders”. *International Journal of Biochemistry and Molecular Biology* 3.2 (2012), pp. 152–164.

[284] Y. Cau, D. Valensin, M. Mori, S. Draghi, and M. Botta. “Structure, function, involvement in diseases and targeting of 14-3-3 proteins: an update”. *Current Medicinal Chemistry* 25.1 (2018), pp. 5–21.

[285] S. R. Slone, N. Lavalley, M. Mcferrin, B. Wang, and T. Alene. “Increased 14-3-3 phosphorylation observed in Parkinson’s disease reduces neuroprotective potential of 14-3-3 proteins”. 79 (2015), pp. 1–13.

[286] Q. Wang, Y. Liu, and J. Zhou. “Neuroinflammation in Parkinson’s disease and its potential as therapeutic target”. *Translational Neurodegeneration* 4 (2015), p. 19.

[287] J. Meiser, D. Weindl, and K. Hiller. “Complexity of dopamine metabolism”. *Cell Communication and Signaling* 11 (2013), Article number: 34.

[288] H. R. Griffith, J. A. den Hollander, O. C. Okonkwo, T. O’Brien, R. L. Watts, and D. C. Marson. “Brain metabolism differs in Alzheimer diesease and Parkinson disease dementia”. *Alzheimers Dementia* 4.6 (2008), pp. 421–427.

[289] X. Zhang, F. Gao, D. Wang, C. Li, Y. Fu, W. He, and J. Zhang. “Tau pathology in Parkinson’s disease”. *Frontiers in Neurology* 9 (2018), p. 809.

[290] J. T. Fukasaw, R. W. de Labio, L. T. Rasmussen, L. C. de Oliveira, E. Chen, J. Villares, G. Tureck, M. d. A. C. Smith, and S. L. M. Payao. “CDK5 and MAPT Gene Expression in Alzheimer’s Disease Brain Samples”. *Current Alzheimer Research* 15.2 (2018), pp. 182–186.

[291] T. Rittman, M. Rubinov, P. E. Vértes, A. X. Patel, C. E. Ginestet, B. C. Ghosh, R. A. Barker, M. G. Spillantini, E. T. Bullmore, and J. B. Rowe. "Regional expression of the MAPT gene is associated with loss of hubs in brain networks and cognitive impairment in Parkinson disease and progressive supranuclear palsy". *Neurobiology of Aging* 48 (2016), pp. 153–160.

[292] J.-Y. Hwang and R. S. Zukin. "REST, a master transcriptional regulator in neurodegenerative disease". *Curr Opin Neurobiol.* 48 (2018), pp. 193–200.

[293] T. Lu et al. "REST and Stress Resistance in Aging and Alzheimer's Disease". *Nature* 507.7493 (2014), pp. 448–454.

[294] M. Yu, L. Cai, M. Liang, Y. Huang, H. Gao, S. Lu, J. Fei, and F. Huang. "Alteration of NRSF expression exacerbating 1-methyl-4-phenyl-pyridinium ion-induced cell death of SH-SY5Y cells". *Neuroscience Research* 65.3 (2009), pp. 236–244.

[295] N. Sugeno, S. Jäckel, A. Voigt, Z. Wassouf, J. Schulze-Hentrich, and P. J. Kahlea. "α-Synuclein enhances histone H3 lysine-9 dimethylation and H3K9me2-dependent transcriptional responses". *Scientific reports* 6 (2016), p. 36328.

[296] B. L. Tang. "Sirtuins as modifiers of Parkinson's disease pathology". *Journal of Neuroscience Research* 95.4 (2017), pp. 930–942.

[297] R. S. Turner et al. "A randomized, double-blind, placebo-controlled trial of resveratrol for Alzheimer disease". *Neurology* 85.16 (2015). Ed. by et al., pp. 1383–1391.

[298] J. H. Malone and B. Oliver. "Microarrays, deep sequencing and the true measure of the transcriptome". *BMC Biology* 9 (2011), Article number: 34.

[299] J. Kelly, R. Moyeed, C. Carroll, S. Luo, and X. Li. "Genetic networks in Parkinson's and Alzheimer's disease". *AGING* 12.6 (2020), pp. 5221–5243.

[300] P. Chatterjee, D. Roy, M. Bhattacharyya, and S. Bandyopadhyay. "Biological networks in Parkinson's disease: An insight into the epigenetic mechanisms associated with this disease". *BMC Genomics* 18 (2017), p. 721.

[301] R. Pinho et al. "Gene expression differences in peripheral blood of Parkinson's disease patients with distinct progression profiles". *PLoS ONE* 11.6 (2016), e0157852.

[302] R. H. Swerdlow. “Mitochondria and Mitochondrial Cascades in Alzheimer’s Disease”. *Journal of Alzheimer’s Disease* 62.3 (2018), pp. 1403–1416.

[303] M. Manczak, B. S. Park, Y. Jung, and P. H. Reddy. “Differential Expression of Oxidative Phosphorylation Genes in Patients with Alzheimer’s Disease: Implications for Early Mitochondrial Dysfunction and Oxidative Damage”. *NeuroMolecular Medicine* 5.2 (2004), pp. 147–162.

[304] B. Zhang et al. “Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease”. *Cell* 153.3 (2013), pp. 707–720.

[305] S. Jevtic, A. S. Sengar, M. W. Salter, and J. A. McLaurin. “The role of the immune system in Alzheimer disease: Etiology and treatment”. *Ageing Research Reviews* 40 (2017), pp. 84–94.

[306] V. Schmidt et al. “Quantitative modelling of amyloidogenic processing and its influence by SORLA in Alzheimer’s disease”. *EMBO Journal* 31.1 (2012), pp. 187–200.

[307] S. C. Ritchie, S. Watts, L. G. Fearnley, K. E. Holt, G. Abraham, and M. Inouye. “A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets”. *Cell Systems* 3.1 (2016), pp. 71–82.

[308] E. R. Gamazon, W. Zhang, A. Konkashbaev, S. Duan, E. O. Kistner, D. L. Nicolae, M. E. Dolan, and N. J. Cox. “SCAN: SNP and copy number annotation”. *Bioinformatics* 26.2 (2010), pp. 259–262.

[309] S. Buckberry, S. J. Bent, T. Bianco-Miotto, and C. T. Roberts. “MassiR: A method for predicting the sex of samples in gene expression microarray datasets”. *Bioinformatics* 30.14 (2014), pp. 2084–2085.

[310] S. Sood et al. “A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status”. *Genome Biology* 16 (2015), p. 185.

[311] J. Taminau et al. “Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages”. *BMC Bioinformatics* 13 (2012), p. 335.

[312] L. Yuan, L. Chen, K. Qian, G. Qian, C. L. Wu, X. Wang, and Y. Xiao. "Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC)". *Genomics Data* 14 (2017), pp. 132–140.

[313] U. Brandes. "A faster algorithm for betweenness centrality". *The Journal of Mathematical Sociology* 25.2 (2001), pp. 163–177.

[314] M. J. Newman. "A measure of betweenness centrality based on random walks". *Social Networks* 27.1 (2005), pp. 39–54.

[315] J. M. Kleinberg. "Authoritative sources in a hyperlinked environment". *Journal of the ACM* 46.5 (1999), pp. 604–632.

[316] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine". *Computer Networks and ISDN Systems* 30.1-7 (1998), pp. 107–117.

[317] G. Csardi and T. Nepusz. "The igraph software package for complex network research". *InterJournal* Complex Sy (2006), p. 1695.

[318] I. E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". *Nature Genetics* 51 (2019), pp. 404–413.

[319] M. Martins et al. "Convergence of mirna expression profiling, α-synuclein interacton and GWAS in Parkinson's disease". *PLoS ONE* 6.10 (2011), e25443.

[320] A. Zovoilis et al. "MicroRNA-34c is a novel target to treat dementias". *EMBO Journal* 30.20 (2011), pp. 4299–4308.

[321] Z. Guo, H. Wang, Y. Li, B. Li, C. Li, and C. Ding. "A microRNA-related single nucleotide polymorphism of the XPO5 gene is associated with survival of small cell lung cancer patients". *Biomedical Reports* 1.4 (2013), pp. 545–548.

[322] C. Liu, F. Zhang, T. Li, M. Lu, L. Wang, W. Yue, and D. Zhang. "MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs". *BMC Genomics* 13 (2012), p. 661.

[323] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui. “HMDD v3.0: A database for experimentally supported human microRNA-disease associations”. *Nucleic Acids Research* 47.D1 (2019), pp. D1013–D1017.

[324] R. Albert. “Scale-free networks in cell biology”. *Journal of Cell Science* 118.21 (2005), pp. 4947–4957.

[325] C. Richter-Landsberg and J. Leyk. “Inclusion body formation, macroautophagy, and the role of HDAC6 in neurodegeneration”. *Acta Neuropathologica* 126.6 (2013), pp. 793–807.

[326] M. Bastian, S. Heymann, and M. Jacomy. “Gephi: An open source software for exploring and manipulating networks”. In: *International AAAI Conference on Weblogs and Social Media*. 2009.

[327] G. Vidal-Martinez, B. Yang, J. Vargas-Medrano, and R. G. Perez. “Could α-synuclein modulation of insulin and dopamine identify a novel link between parkinson’s disease and diabetes as well as potential therapies?” *Frontiers in Molecular Neuroscience* 11 (2018), p. 465.

[328] C. Simões-Pires, V. Zwick, A. Nurisso, E. Schenker, P. A. Carrupt, and M. Cuendet. “HDAC6 as a target for neurodegenerative diseases: What makes it different from the other HDACs?” *Molecular Neurodegeneration* 8 (2013), p. 7.

[329] X. Qu, C. Huang, H. Qu, B. Jia, Q. Cui, C. Sun, and Y. Chu. “Histone deacetylase 6 promotes insulin resistance via deacetylating phosphatase and tensin homolog (PTEN) in ovarian OVCAR-3 cells”. *International Journal of Clinical and Experimental Pathology* 9.7 (2016), pp. 7105–7113.

[330] S. Sekar and C. Taghibiglou. “Elevated nuclear Phosphatase and tensin homolog (PTEN) and Altered Insulin Signaling in Substantia Nigral Region of Patients with Parkinson’s Disease”. *Neuroscience Letters* 666 (2018), pp. 139–143.

[331] J. Shirakawa, M. Fernandez, T. Takatani, A. E. Ouaamari, P. Jungtrakoon, E. R. Okawa, W. Zhang, P. Yi, A. Doria, and R. N. Kulkarni. “Insulin signaling regulates the FoxM1/PLK1/CENP-A pathway to promote adaptive pancreatic β-cell proliferation”. *Cell Metabolism* 25.4 (2017), pp. 868–882.

[332] Š. Lehtonen, T. M. Sonninen, S. Wojciechowski, G. Goldsteins, and J. Koistinaho. “Dysfunction of cellular proteostasis in Parkinson’s disease”. *Frontiers in Neuroscience* 13 (2019), p. 457.

[333] J. A. Santiago and J. A. Potashkin. “Blood transcriptomic meta-analysis identifies dysregulation of hemoglobin and iron metabolism in Parkinson’ disease”. *Frontiers in Aging Neuroscience* 9 (2017), p. 73.

[334] I. Diner et al. “Aggregation properties of the small nuclear ribonucleoprotein U1-70K in Alzheimer disease”. *Journal of Biological Chemistry* 289.51 (2014), pp. 35296–35313.

[335] X. Liu, J. Chen, T. Guan, H. Yao, W. Zhang, Z. Guan, and Y. Wang. “MiRNAs and target genes in the blood as biomarkers for the early diagnosis of Parkinson’s disease”. *BMC Systems Biology* 13 (2019), Article Number: 10.

[336] Z. Wan, D. Mah, S. Simtchouk, A. Kluftinger, and J. P. Little. “Role of amyloid b in the induction of lipolysis and secretion of adipokines from human adipose tissue”. *Adipocyte* 4.3 (2015), pp. 212–216.

[337] L. S. S. Ferreira, C. S. Fernandes, M. N. N. Vieira, and F. G. De Felice. “Insulin Resistance in Alzheimer’s Disease”. *Frontiers in Neuroscience* 12 (2018), p. 830.

[338] T. Imaizumi et al. “Effect of dietary energy and polymorphisms in BRAP and GHRL on obesity and metabolic traits”. *Obesity Research & Clinical Practice* 12.1 (2016), pp. 39–48.

[339] M. W. Albers et al. “At the interface of sensory and motor dysfunctions and Alzheimer’s disease”. *Alzheimer’s and Dementia* 11.1 (2015), pp. 70–98.

[340] C. Murphy. “Olfactory and other sensory impairments in Alzheimer disease”. *Nature Reviews Neurology* 15 (2019), pp. 11–24.

[341] B. Malnic, P. A. Godfrey, and L. B. Buck. “The human olfactory receptor gene family”. *Proceedings of the National Academy of Sciences of the United States of America* 101.8 (2004), pp. 2584–2589.

[342] E. Casadei, L. Tacchi, C. R. Lickwar, S. T. Espenschied, J. M. Davison, P. Muñoz, J. F. Rawls, and I. Salinas. "Commensal Bacteria Regulate Gene Expression and Differentiation in Vertebrate Olfactory Systems Through Transcription Factor REST". *Chemical Senses* 44.8 (2019), pp. 615–630.

[343] N. T. Seyfried et al. "A Multi-Network Approach Identifies Protein-specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease". *Cell Systems* 4.1 (2017), pp. 60–72.

[344] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A. E. Schrag, and A. E. Lang. "Parkinson disease". *Nature Reviews Disease Primers* 3 (2017), p. 17013.

[345] M. T. Gray and J. M. Woulfe. "Striatal blood-brain barrier permeability in Parkinson's disease". *Journal of Cerebral Blood Flow and Metabolism* 35.5 (2015), pp. 747–750.

[346] L. F. Lue, A. Guerra, and D. G. Walker. "Amyloid Beta and Tau as Alzheimer's Disease Blood Biomarkers: Promise From New Technologies". *Neurology and Therapy* 6.s1 (2017), pp. 25–36.

[347] J. Chojdak-Łukasiewicz, M. Małodobra-Mazur, A. Zimny, L. Noga, and B. Paradowski. "Plasma tau protein and Aβ42 level as markers of cognitive impairment in patients with Parkinson's disease". *Advances in Clinical and Experimental Medicine* 29.1 (2020), pp. 115–121.

[348] W. C. Lin et al. "Peripheral leukocyte apoptosis in patients with parkinsonism: Correlation with clinical characteristics and neuroimaging findings". *BioMed Research International* 2014 (2014), p. 635923.

[349] S. J. Annesley et al. "Immortalized Parkinson's disease lymphocytes have enhanced mitochondrial respiratory activity". *DDM Disease Models and Mechanisms* 9.11 (2016), pp. 1295–1305.

[350] D. Bäckström, G. Granåsen, M. E. Domellöf, J. Linder, S. J. Mo, K. Riklund, H. Zetterberg, K. Blennow, and L. Forsgren. "Early predictors of mortality in

parkinsonism and Parkinson disease A population-based study". *Neurology* 91.22 (2018), e2045–e2056.

[351] K. Rezai-Zadeh, D. Gate, C. A. Szekely, and T. Town. "Can peripheral leukocytes be used as Alzheimer's disease biomarkers?" *Expert Review of Neurotherapeutics* 9.11 (2009), pp. 1623–1633.

[352] H. Li, G. Hong, M. Lin, Y. Shi, L. Wang, F. Jiang, F. Zhang, Y. Wang, and Z. Guo. "Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease". *Scientific Reports* 7 (2017), p. 14027.

[353] Z. Cai, P. F. Qiao, C. Q. Wan, M. Cai, N. K. Zhou, and Q. Li. "Role of Blood-Brain Barrier in Alzheimer's Disease". *Journal of Alzheimer's Disease* 63.4 (2018), pp. 1223–1234.

[354] P. T. Nelson et al. "Limbic-predominant age-related TDP-43 encephalopathy (LATE): Consensus working group report". *Brain* 142.6 (2019), pp. 1503–1527.

[355] R. Shamir et al. "Analysis of blood-based gene expression in idiopathic Parkinson disease". *Neurology* 89.16 (2017), pp. 1676–1683.

[356] C. Chai and K.-L. Lim. "Genetic Insights into Sporadic Parkinson's Disease Pathogenesis". *Current Genomics* 14.8 (2013), pp. 486–501.

[357] M. Karaglani, K. Gourlia, I. Tsamardinos, and E. Chatzaki. "Accurate Blood-Based Diagnostic Biosignatures for Alzheimer's Disease via Automated Machine Learning". *Journal of Clinical Medicine* 9.9 (2020), p. 3016.

[358] S. Cui, Q. Wu, J. West, and J. Bai. "Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease". *PLoS Computational Biology* 15.8 (2019), e1007264.

[359] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[360] R. Bedre. *reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit*. 2020.

[361] M. Kanehisa and S. Goto. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. *Nucleic Acids Research* 28.1 (2000), pp. 27–30.

[362] M. Kanehisa. “Toward understanding the origin and evolution of cellular organisms”. *Protein Science* 28.11 (2019), pp. 1947–1951.

[363] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe. “KEGG: integrating viruses and cellular organisms”. *Nucleic Acids Research* 49 (2020), pp. D545–D551.

[364] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794.

[365] A. B. Niculescu and H. Le-Niculescu. “Convergent functional genomics: What we have learned and can learn about genes, pathways, and mechanisms”. *Neuropsychopharmacology* 35 (2010), pp. 355–356.

[366] F. Jiang, Q. Wu, S. Sun, G. Bi, and L. Guo. “Identification of potential diagnostic biomarkers for Parkinson’s disease”. *FEBS Open Bio* 9.8 (2019), pp. 1460–1468.

[367] M. Falchetti, R. D. Prediger, and A. Zanotto-Filho. “Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic Parkinson’s disease”. *Computers in Biology and Medicine* 124 (2020), p. 103925.

[368] D. Ai, Y. Wang, X. Li, and H. Pan. “Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder”. *Biomolecules* 10.9 (2020), p. 1207.

[369] R. Shamir et al. “Analysis of blood-based gene expression in idiopathic Parkinson disease”. *Neurology* 89.16 (2017), pp. 1676–1683.

[370] A. J. López-Farré, J. J. Zamorano-León, A. Segura, P. J. Mateos-Cáceres, J. Modrego, P. Rodríguez-Sierra, L. Calatrava, J. Tamargo, and C. Macaya. “Plasma desmoplakin I biomarker of vascular recurrence after ischemic stroke”. *Journal of Neurochemistry* 121.2 (2012), pp. 314–325.

[371] J. Gawinecka, B. Ciesielczyk, P. Sanchez-Juan, M. Schmitz, U. Heinemann, and I. Zerr. “Desmoplakin as a potential candidate for cerebrospinal fluid marker to rule out 14-3-3 false positive rates in sporadic Creutzfeldt-Jakob disease differential diagnosis”. *Neurodegener Dis.* 9.3 (2012), pp. 139–144.

[372] J. Zhao, C. L. Meyerkord, Y. Du, F. R. Khuri, and H. Fu. “14-3-3 proteins as potential therapeutic targets”. *Semin Cell Dev Biol.* 22.7 (2011), pp. 705–712.

[373] 1. Yan-Jun Wan et al. “Allosteric regulation of protein 14-3-3ζ scaffold by small-molecule editing modulates histone H3 post-translational modifications”. *Theranostics* 10.2 (2020), pp. 797–815.

[374] E. Cho and J.-Y. Park. “Emerging roles of 14-3-3γ in the brain disorder”. *BMB reports* 53.10 (2020), pp. 500–511.

[375] C. Benedict and C. A. Grillo. “Insulin Resistance as a Therapeutic Target in the Treatment of Alzheimer’s Disease: A State-of-the-Art Review”. *Theranostics* 12 (2018), p. 215.

[376] J.-Y. Lee, Y. Nagano, J. P. Taylor, K. L. Lim, and T.-P. Yao. “Disease-causing mutations in parkin impair mitochondrial ubiquitination, aggregation, and HDAC6-dependent mitophagy”. *J Cell Biol.* 189.4 (2010), pp. 671–679.

[377] Y. Li, S. Sang, W. Ren, Y. Pei, Y. Bian, Y. Chen, and H. Sun. “Inhibition of Histone Deacetylase 6 (HDAC6) as a therapeutic strategy for Alzheimer’s disease: A review (2010–2020)”. *European Journal of Medicinal Chemistry* 226 (2021), p. 113874.

[378] M. Runfola, S. Sestito, S. Gul, G. Chiellini, and S. Rapposelli. “Collecting data through high throughput in vitro early toxicity and off-target liability assays to rapidly identify limitations of novel thyromimetics”. *Data Brief* 29 (2020), p. 105206.

[379] J. Qiu, E. J. Wagner, O. K. Rønnekleiv, and M. J. Kelly. “Insulin and leptin excite anorexigenic pro-opiomelanocortin neurones via activation of TRPC5 channels”. *Journal of Neuroendocrinology* 30.2 (2018), e12501.

[380] J. E. McDermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland. “Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data”. *Expert Opin Med Diagn.* 7.1 (2013), pp. 37–51.

[381] Y. Bai, T. Xu, and X. Zhang. “Graphene-based biosensors for detection of biomarkers”. *Micromachines* 11.1 (2020), p. 60.

[382] J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. De Bruijne, M. W. Vernooij, H. H. Adams, M. A. Ikram, W. J. Niessen, and G. V. Roshchupkin. “Gray matter age prediction as a biomarker for risk of dementia”. *Proceedings of the National Academy of Sciences of the United States of America* 116.42 (2019), pp. 21213–21218.

Appendix A

Meta-analysis of gene expression for Parkinson's disease and the crosstalk between Parkinson's and Alzheimer's diseases

A.1 Table of significant pathways identified using PD DEGs

Table A.1: IPA canonical pathway analysis for significant pathways identified using all PD DEGs, included with the information for pathways shared with those identified as significant using all AD DEGs.

Ingenuity Canonical Pathways	**DEGs**	**Genes in Pathway**	**Ratio**	**adjPval**	**AD DEGs**	**AD Ratio**	**AD adjPval**
Breast Cancer Regulation by Stathmin1	33	204	0.162	2.40e-06	47	0.23	8.51e-03
Sirtuin Signaling Pathway	40	284	0.141	2.40e-06	70	0.247	3.39e-04
14-3-3-mediated Signaling	24	130	0.185	9.55e-06	34	0.262	5.13e-03
Phagosome Maturation	23	138	0.167	9.12e-05	34	0.246	1.15e-02
Remodeling of Epithelial Adherens Junctions	15	67	0.227	1.12e-04			
Mitochondrial Dysfunction	24	166	0.145	4.27e-04	41	0.248	4.79e-03
Oxidative Phosphorylation	18	105	0.173	4.68e-04			
CDK5 Signaling	17	99	0.173	7.24e-04	27	0.276	7.08e-03
Huntington's Disease Signaling	30	248	0.121	7.59e-04			
B Cell Receptor Signaling	25	188	0.133	7.59e-04	50	0.266	6.92e-04
Epithelial Adherens Junction Signaling	21	143	0.147	7.59e-04			
Gap Junction Signaling	24	191	0.126	2.51e-03			
Germ Cell-Sertoli Cell Junction Signaling	22	170	0.130	2.69e-03	41	0.243	7.08e-03
Axonal Guidance Signaling	43	447	0.096	2.95e-03	94	0.211	2.09e-03
Sertoli Cell-Sertoli Cell Junction Signaling	22	174	0.127	3.31e-03	38	0.22	3.72e-02
Rac Signaling	17	116	0.147	3.31e-03	29	0.25	1.70e-02
Synaptic Long Term Potentiation	17	119	0.143	4.17e-03	29	0.244	2.34e-02
AMPK Signaling	25	216	0.116	4.17e-03			
Glycolysis I	7	24	0.292	4.27e-03			

PI3K/AKT Signaling	17	124	0.138	5.50e-03	34	0.276	2.09e-03
ERK/MAPK Signaling	23	199	0.116	7.08e-03	51	0.256	9.55e-04
Iron homeostasis signaling pathway	17	127	0.134	7.41e-03			
Clathrin-mediated Endocytosis Signaling	23	206	0.112	1.05e-02			
Gαi Signaling	16	121	0.133	1.05e-02			
Signaling by Rho Family GTPases	26	250	0.104	1.38e-02	61	0.243	9.55e-04
Neuropathic Pain Signaling In Dorsal Horn Neurons	15	114	0.132	1.62e-02			
G Protein Signaling Mediated by Tubby	7	31	0.226	1.62e-02			
Role of NFAT in Cardiac Hypertrophy	23	217	0.106	1.62e-02	47	0.218	2.19e-02
Reelin Signaling in Neurons	13	93	0.141	1.62e-02			
Amyloid Processing	9	50	0.180	1.62e-02			
Cardiac β-adrenergic Signaling	17	141	0.121	1.62e-02			
TCA Cycle II (Eukaryotic)	6	24	0.250	1.82e-02			
p70S6K Signaling	16	132	0.122	1.95e-02			
PTEN Signaling	15	120	0.126	1.95e-02	31	0.261	7.94e-03
Sumoylation Pathway	13	97	0.135	2.00e-02	30	0.312	9.33e-04
Protein Ubiquitination Pathway	26	264	0.099	2.00e-02			
Gluconeogenesis I	6	25	0.240	2.00e-02			
STAT3 Pathway	13	98	0.134	2.04e-02	20	0.27	2.63e-02
HIPPO signaling	12	86	0.140	2.09e-02	32	0.372	4.17e-05
fMLP Signaling in Neutrophils	15	122	0.123	2.14e-02	33	0.27	3.47e-03
Dopamine-DARPP32 Feedback in cAMP Signaling	18	161	0.112	2.14e-02			
D-myo-inositol (1,4,5)-trisphosphate Degradation	5	18	0.278	2.14e-02			

Dopamine Receptor Signaling	11	76	0.145	2.14e-02			
GNRH Signaling	18	163	0.111	2.29e-02	38	0.235	1.45e-02
Insulin Receptor Signaling	16	137	0.117	2.34e-02			
Cardiac Hypertrophy Signaling	23	233	0.099	2.63e-02	51	0.219	1.45e-02
Cyclins and Cell Cycle Regulation	11	80	0.138	3.02e-02	20	0.253	4.79e-02
IGF-1 Signaling	13	106	0.123	3.55e-02			
Protein Kinase A Signaling	33	386	0.086	3.72e-02	82	0.214	3.24e-03
Gαq Signaling	17	159	0.107	3.89e-02	42	0.264	1.58e-03
Aspartate Degradation II	3	7	0.429	3.98e-02	5	0.714	9.55e-03
Renin-Angiotensin Signaling	14	121	0.116	3.98e-02			
ATM Signaling	12	97	0.124	4.27e-02			
BMP signaling pathway	10	75	0.135	4.47e-02			

A.2 Table of significant pathways identified using down-regulated PD DEGs

Table A.2: IPA canonical pathway analysis for significant pathways identified using down-regulated PD substantia nigra DEGs.

Ingenuity Canonical Pathways	DEGs	Genes in Pathway	Ratio	Adj. Pval
Breast Cancer Regulation by Stathmin1	32	204	0.157	1.86e-09
Phagosome Maturation	23	138	0.167	2.19e-07
Sirtuin Signaling Pathway	34	284	0.120	2.19e-07
Mitochondrial Dysfunction	24	166	0.145	9.55e-07
14-3-3-mediated Signaling	21	130	0.162	9.55e-07
Remodeling of Epithelial Adherens Junctions	15	67	0.227	9.55e-07
Oxidative Phosphorylation	18	105	0.173	2.82e-06
Axonal Guidance Signaling	39	447	0.087	3.39e-05
Gap Junction Signaling	22	192	0.115	1.17e-04
CDK5 Signaling	15	99	0.153	1.32e-04
Huntington's Disease Signaling	25	248	0.101	2.14e-04
Sertoli Cell-Sertoli Cell Junction Signaling	20	173	0.116	2.34e-04
Germ Cell-Sertoli Cell Junction Signaling	19	170	0.112	5.13e-04
Cardiac β-adrenergic Signaling	17	141	0.121	5.13e-04
Glycolysis I	7	24	0.292	5.37e-04
Epithelial Adherens Junction Signaling	17	143	0.119	6.03e-04
Synaptic Long Term Potentiation	15	120	0.126	8.51e-04
PI3K/AKT Signaling	15	123	0.122	1.17e-03
Iron homeostasis signaling pathway	15	128	0.118	1.62e-03
Neuropathic Pain Signaling In Dorsal Horn Neurons	14	114	0.123	1.74e-03
Dopamine-DARPP32 Feedback in cAMP Signaling	17	161	0.106	1.91e-03
Rac Signaling	14	116	0.121	1.95e-03
Dopamine Receptor Signaling	11	76	0.145	1.99e-03
Role of NFAT in Cardiac Hypertrophy	20	216	0.093	2.88e-03
Signaling by Rho Family GTPases	22	252	0.088	2.95e-03
TCA Cycle II (Eukaryotic)	6	24	0.250	3.24e-03
Protein Kinase A Signaling	29	386	0.075	3.80e-03
Gluconeogenesis I	6	25	0.240	3.80e-03
HIPPO signaling	11	86	0.128	4.79e-03
GNRH Signaling	16	162	0.099	5.01e-03

p70S6K Signaling	14	131	0.107	5.01e-03
AMPK Signaling	19	215	0.088	5.62e-03
ERK/MAPK Signaling	18	199	0.091	5.89e-03
Amyloid Processing	8	50	0.160	5.89e-03
Gαi Signaling	13	121	0.108	6.17e-03
Opioid Signaling Pathway	20	237	0.084	6.46e-03
fMLP Signaling in Neutrophils	13	122	0.107	6.92e-03
Clathrin-mediated Endocytosis Signaling	18	206	0.087	7.59e-03
CREB Signaling in Neurons	18	211	0.086	8.91e-03
Gαq Signaling	15	160	0.094	8.91e-03
Protein Ubiquitination Pathway	21	265	0.080	8.91e-03
RhoGDI Signaling	16	177	0.091	8.91e-03
Melatonin Signaling	9	70	0.129	1.10e-02
Calcium Signaling	17	198	0.086	1.12e-02
α-Adrenergic Signaling	10	85	0.118	1.12e-02
Synaptic Long Term Depression	15	168	0.089	1.38e-02
Actin Cytoskeleton Signaling	18	222	0.081	1.38e-02
BMP signaling pathway	9	74	0.122	1.41e-02
Aspartate Degradation II	3	7	0.429	1.41e-02
B Cell Receptor Signaling	16	189	0.085	1.48e-02
Insulin Receptor Signaling	13	137	0.095	1.48e-02
Regulation of eIF4 and p70S6K Signaling	14	155	0.091	1.48e-02
IGF-1 Signaling	11	106	0.104	1.55e-02
Reelin Signaling in Neurons	10	92	0.109	1.70e-02
Fcγ Receptor-mediated Phagocytosis in Macrophages and Monocytes	10	93	0.108	1.82e-02
Parkinson's Signaling	4	16	0.250	1.86e-02
Pyridoxal 5'-phosphate Salvage Pathway	8	64	0.125	1.86e-02
CCR3 Signaling in Eosinophils	12	127	0.095	1.86e-02
Cardiac Hypertrophy Signaling	18	232	0.078	1.86e-02
Phototransduction Pathway	7	52	0.137	1.91e-02
Salvage Pathways of Pyrimidine Ribonucleotides	10	96	0.105	1.91e-02
D-myo-inositol (1,4,5)-Trisphosphate Biosynthesis	5	28	0.185	2.00e-02
Cdc42 Signaling	12	130	0.093	2.08e-02
Tight Junction Signaling	14	167	0.084	2.40e-02
Chemokine Signaling	8	68	0.118	2.40e-02
Aldosterone Signaling in Epithelial Cells	14	168	0.084	2.45e-02
D-myo-inositol (1,4,5)-trisphosphate Degradation	4	19	0.222	2.45e-02

P2Y Purigenic Receptor Signaling Pathway	12	134	0.090	2.45e-02
PAK Signaling	10	100	0.100	2.45e-02
Role of CHK Proteins in Cell Cycle Checkpoint Control	7	57	0.123	3.09e-02
G Protein Signaling Mediated by Tubby	5	32	0.161	3.16e-02
Renin-Angiotensin Signaling	11	122	0.091	3.16e-02
Inhibition of Angiogenesis by TSP1	5	33	0.156	3.63e-02
Xenobiotic Metabolism Signaling	19	273	0.070	3.80e-02
G-Protein Coupled Receptor Signaling	19	275	0.069	3.98e-02
GDNF Family Ligand-Receptor Interactions	8	77	0.105	3.98e-02
IL-1 Signaling	9	93	0.098	3.98e-02
Ceramide Signaling	9	93	0.097	4.17e-02
Arsenate Detoxification I (Glutaredoxin)	2	4	0.500	4.37e-02
CXCR4 Signaling	13	164	0.079	4.37e-02
Mevalonate Pathway I	3	12	0.250	4.68e-02

Appendix B

Network analysis to identify key dysregulated processes and hub genes in neurodegenerative diseases

B.1 Table of significant hubs identified in non-preserved modules between PD and healthy controls using network analysis

Table B.1: Significant hubs identified in non-preserved modules between PD and healthy controls using network analysis.

Module	Gene	Hub detection method	Score	P-value
PD modules not preserved in HC				
Darkseagreen4	GINS2	Betweenness	3826	0.005
	S1PR5	Kleinberg's centrality; PageRank; MM	0.30751; 0.02637; 0.90234	0.006; 0.006; 0.007
	AGBL2	Closeness	10.00256	0.007
	NKG7	PageRank	0.02512	0.007
Navajowhite2	SNRNP70	PageRank; Kleinberg's centrality	0.02359; 0.27933	0.003; 0.007
	POPDC2	Closeness	18.03573	0.008
	CHKB	Kleinberg's centrality	0.28034	0.009
	MIR142	MM	0.85297	0.009
Salmon	TYSND1	PageRank; MM; Kleinberg's centrality	0.00978; 0.84787; 0.17499	0.002; 0.002; 0.008
	C17orf97	Closeness	4.4882	0.002
	HDAC6	Kleinberg's centrality; MM; PageRank	0.17867; 0.83636; 0.00958	0.003; 0.006; 0.007
	FAM114A1	Betweenness	12901	0.004
	ZNF804A	Betweenness; Closeness	12956; 4.27567	0.005; 0.007
	ABCD1	PageRank; MM	0.00904; 0.83955	0.006; 0.006
	ZNF526	PageRank	0.00908	0.006
	TMEM147-AS1	Betweenness	12566	0.008
	RENBP	PageRank	0.00823	0.009

HC modules not preserved in PD				
	FAM110C	Closeness; Betweenness	0.72585; 33683	0.000; 0.002
	TXLNGY	Betweenness	40661	0
	PAK4	Kleinberg's centrality; Pagerank; MM	0.12262; 0.00467; 0.83401	0.001; 0.002; 0.003
	GIGYF1	Kleinberg's centrality; PageRank; MM	0.12332; 0.00473; 0.85428	0.002; 0.002; 0.002
	WDTC1	Kleinberg's centrality; MM; PageRank	0.11337; 0.82836; 0.00441	0.002; 0.004; 0.008
	NEB	Closeness; Betweenness	0.70015; 21395	0.003; 0.004
	SH3BGR	Closeness; Betweenness	0.63727; 19636	0.004; 0.005
	FCGBP	Betweenness	16988	0.005
Purple	INO80B	PageRank; Kleinberg's centrality; MM	0.00417; 0.10391; 0.82766	0.005; 0.007; 0.007
	ZNF582-AS1	Closeness; Betweenness	0.59408; 0.06978	0.006; 0.008
	PLA2G4C	Betweenness	20491	0.007
	TBC1D25	PageRank; MM	0.00401; 0.81547	0.007; 0.007
	MFSD12	Kleinberg's centrality; PageRank; MM	0.10808; 0.00411; 0.80996	0.007; 0.009; 0.009
	MCM2	Closeness	0.57973	0.008
	SPATA6	Closeness	0.65087	0.009
	RPS6KA4	MM	0.80597	0.009
	FIZ1	MM	0.81009	0.009

B.2 Table of significant hubs identified in non-preserved modules between AD, MCI and healthy controls using network analysis

Table B.2: Significant hubs identified in non-preserved modules between AD, MCI and healthy controls using network analysis.

Module	Gene	Hub detection method	Score	P-value
AD modules				
	TARM1	Betweenness; Closeness	8087; 0.01321	0.000; 0.001
	GPRACR	Kleinberg's centrality; PageRank; MM	0.04680; 0.00133; 0.70116	0.000; 0.003; 0.000
	LINC01122	Closeness; Betweenness	0.01357; 9575	0.000; 0.003
	DNAJB4	Betweenness; Closeness	6040; 0.01296	0.001; 0.008
	MDFI	PageRank; Kleinberg's centrality; MM	0.00133; 0.04662; 0.69908	0.001; 0.001; 0.001
	TRPC5	PageRank; Kleinberg's centrality; MM	0.00133; 0.04646; 0.69609	0.001; 0.002; 0.004
	UXS1	MM; Kleinberg's centrality; PageRank	0.66258; 0.04457; 0.00128	0.001; 0.005; 0.006
	GRIP2	Betweenness; Closeness	11154; 0.01304	0.002; 0.009
	BLNK	Closeness; Betweenness	0.01333; 6775	0.002; 0.007
	TACR2	Betweenness; Closeness	7150; 0.01312	0.003; 0.003
	LHFPL1	MM; PageRank; Kleinberg's centrality	0.65754; 0.00128; 0.04443	0.003; 0.006; 0.007
	USP18	Betweenness	6055	0.004
	MRAP2	Closeness; Betweenness	0.01294; 4184	0.004; 0.007
	LY6G6C	Kleinberg's centrality; PageRank; MM	0.04587; 0.00131; 0.68542	0.004; 0.004; 0.005
Blue	*LSR*	Kleinberg's centrality; MM	0.04445; 0.65775	0.004; 0.009
	WDR90	PageRank; MM; Kleinberg's centrality	0.00129; 0.66619; 0.04486	0.004; 0.006; 0.007
	LINC01547	PageRank; Kleinberg's centrality	0.00127; 0.04422	0.004; 0.007
	TYMSOS	Closeness	0.01288	0.005
	BRAP	Kleinberg's centrality	0.04389	0.005
	B4GALT2	MM; Kleinberg's centrality; PageRank	0.66218; 0.04463; 0.00128	0.005; 0.007; 0.008

	CCER1	Betweenness; Closeness	4412; 0.01296	0.006; 0.007
	ST8SIA5	Betweenness; Closeness	4355; 0.01289	0.007; 0.007
	HERC2P2	Betweenness	4157	0.007
	PTPN5	Kleinberg's centrality	0.04384	0.007
	TIMM13	PageRank	0.00126	0.007
	ZNF829	Closeness	0.0131	0.008
	ICAM1	Kleinberg's centrality; MM	0.04369; 0.64437	0.008; 0.009
	FOXH1	Betweenness	3872	0.009
	MYMK	Closeness	0.01271	0.009
MCI modules				
	OR10A7	Betweenness; Closeness	2938; 0.01844	0.000; 0.001
	UXS1	MM; Kleinberg's centrality; PageRank	0.58467; 0.04861; 0.00168	0.000; 0.001; 0.001
	UGT2B11	Closeness; Betweenness	0.01834; 2433	0.000; 0.003
	ADPRH	PageRank; MM; Kleinberg's centrality	0.00166; 0.57500; 0.04805	0.001; 0.003; 0.005
	RBMS3	Betweenness; Closeness	2780; 0.01796	0.001; 0.007
	COL12A1	PageRank; Kleinberg's centrality; MM	0.00166; 0.04801; 0.57814	0.002; 0.004; 0.006
	TRPC5	PageRank; MM	0.00164; 0.56563	0.003; 0.007
	OR4A16	Betweenness	2210	0.004
	LRRC59	Closeness; Betweenness	0.01808; 1851	0.004; 0.009
	MIR99AHG	Betweenness	1873	0.005
Sienna3	*MUC16*	Closeness; Betweenness	0.01778; 1858	0.005; 0.006
	TXN2	PageRank; Kleinberg's centrality	0.00165; 0.04788	0.005; 0.007
	ANKRD35	PageRank; Kleinberg's centrality	0.00162; 0.04706	0.006; 0.008

	CBARP	PageRank; MM	0.00164; 0.56376	0.006; 0.009
	ERLIN1	Kleinberg's centrality	0.04537	0.007
	MDFI	Kleinberg's centrality; PageRank	0.00162; 0.04673	0.007; 0.008
	CFAP47	Kleinberg's centrality	0.04686	0.008
	ZNF624	PageRank	0.00162	0.008
	CYP26B1	Closeness	0.0179	0.008
	PPP1R14D	Betweenness	1851	0.009
	HEPACAM	Betweenness	1856	0.009
HC modules				
	LOC150051	Kleinberg's centrality; PageRank; MM	0.06961; 0.00266; 0.63584	0.000; 0.000; 0.000
	TUBA3E	Betweenness; Closeness	3142; 0.08073	0.001; 0.006
	SYN1	Betweenness; Closeness	4572; 0.08297	0.001; 0.006
	BRMS1L	Kleinberg's centrality; MM; PageRank	0.06801; 0.62877; 0.00262	0.003; 0.003; 0.006
	DEFB123	Kleinberg's centrality; PageRank; MM	0.06542; 0.00255; 0.60318	0.003; 0.006; 0.008
	FAM81B	Closeness; Betweenness	0.08489; 2573	0.003; 0.009
Darkolivegreen	*C6*	Closeness; Betweenness	0.08390; 2995	0.004; 0.005
	VN1R2	Kleinberg's centrality; PageRank; MM	0.06545; 0.00253; 0.60514	0.004; 0.005; 0.009
	ENKUR	Kleinberg's centralityv; MM	0.06562; 0.59780	0.004; 0.007
	EFCAB1	Closeness	0.08297	0.006
	TCTEX1D4	Closeness	0.08121	0.006
	SCG3	Betweenness	2657	0.007
	OR5AS1	PageRank	0.00253	0.007
	YIPF5	Closeness	0.08046	0.009

Darkorange2	*AP5Z1*	PageRank; MM; Kleinberg's centrality	0.00635; 0.67903; 0.10642	0.001; 0.003; 0.009
	MED18	Closeness; Betweenness	0.17321; 1833	0.003; 0.004
	NHLRC4	Closeness	0.16515	0.004
	NEK8	PageRank	0.00592	0.005
	ZNF585A	PageRank; Kleinberg's centrality	0.00608; 0.10271	0.006; 0.009
	CHEK1	Betweenness	1277	0.009
Skyblue	*IL1RL1*	Betweenness	1595	0.004
	ADRA1A	Closeness; Betweenness	0.24625; 1276	0.005; 0.007
	RBM5	MM; PageRank	0.87355; 0.01192	0.006; 0.007
Red	*CHST11*	Betweenness; Closeness	5964; 0.06598	0.000; 0.002
	PLK3	PageRank; MM; Kleinberg's centrality	0.00225; 0.65033; 0.06427	0.001; 0.001; 0.002
	CPEB2	MM; Kleinberg's centrality; PageRank	0.63587; 0.06243; 0.00219	0.001; 0.002; 0.002
	PNMA8A	Betweenness; Closeness	5555; 0.06434	0.001; 0.008
	ZFY	Closeness	0.06577	0.002
	IGFBPL1	MM; PageRank; Kleinberg's centrality	0.62685; 0.00216; 0.06143	0.002; 0.003; 0.004
	LINC01623	Closeness; Betweenness	0.06630; 3297	0.002; 0.004
	PLXND1	Kleinberg's centrality; MM; PageRank	0.06357; 0.64174; 0.00222	0.003; 0.004; 0.006
	WFDC13	Betweenness	3644	0.005
	CD200R1L	Closeness	0.06528	0.005
	VGLL2	Closeness; Betweenness	0.06555; 3580	0.005; 0.008

SLC28A1	PageRank; MM	0.00214; 0.61949	0.006; 0.008
SKA1	Betweenness	3161	0.007
ABCA9	Betweenness	4142	0.007
KREMEN2	Closeness	0.06402	0.008
SNF8	MM; Kleinberg's centrality	0.61736; 0.06063	0.008; 0.009
TXN2	PageRank	0.00214	0.009
FBXO42	PageRank	0.00209	0.009
KIF2C	Closeness	0.06184	0.009