UNIVERSITY OF
PLYMOUTH

DISCOVERING BIOMARKERS OF ALZHEIMER'S DISEASE BY STATISTICAL LEARNING APPROACHES

by

JINTAO LONG

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

DOCTOR OF PHILOSOPHY

Peninsula School of Medicine
March 2019

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 Plymouth University.

Publications:

Li, X., Long, J., He, T., Belshaw, R. and Scott, J., 2015. Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease. *Scientific reports*, 5.
Long, J., Pan, G., Ifeachor, E., Belshaw, R. and Li, X., 2016. Discovery of Novel Biomarkers for Alzheimer's Disease from Blood. *Disease markers*, 2016.

Grants:

£2500. Henry Lester Trust. (2016)

Word count of main body of thesis: 27358

t

Signed: _____

Date: _____

Abstract

In this work, statistical learning approaches are exploited to discover biomarkers for Alzheimer's disease (AD). The contributions has been made in the fields of both biomarker and software driven studies. Surprising discoveries were made in the field of blood-based biomarker search. With the inclusion of existing biological knowledge and a proposed novel feature selection method, several blood-based protein models were discovered to have promising ability to separate AD patients from healthy individuals. A new statistical pattern was discovered which can be potential new guideline for diagnosis methodology. In the field of brain-based biomarker, the positive contribution of covariates such as age, gender and APOE genotype to a AD classifier was verified, as well as the discovery of panel of highly informative biomarkers comprising 26 RNA transcripts. The classifier trained by the panetl of genes shows excellent capacity in discriminating patients from control. Apart from biomarker driven studies, the development of statistical packages or application were also involved. R package *metaUnion* was designed and developed to provide advanced meta-analytic approach applicable for microarray data. This package overcomes the defects appearing in previous meta-analytic packages – 1) the neglection of missing data, 2) the inflexibility of feature dimension 3) the lack of functions to support post-analysis summary. R package *metaUnion* has been applied in a published study as part of the integrated genomic approaches and resulted in significant findings. To provide benchmark references about significance of features for dementia researchers, a web-based platform AlzExpress was built to provide researchers with granular level of differential expression test and meta-analysis results. A combination of fashionable big data technologies and robust data mining algorithms make *AlzExpress* a flexible, scalable and comprehensive platform of valuable bioinformatics in dementia research.

Contents

List of Figures

List of Tables

Acknowledgement

It is with utmost sincerity that I express my gratitude to all of those who have supported me throughout my PhD research.

First and foremost, I would like to thank my supervisor Dr. Xinzhong Li and I really appreciate him for helping me through the ups and downs in my research life. I acquired too much inspiration and guidance from him throughout the past years.

I would also like to thank the other two supervisors Prof. Genhua Pan and Prof. Emmanuel Ifeachor for many discussions on my experiment, giving me precious opinions and suggestions and helping me accomplish my research project. This unique experience has greatly expanded my knowledge and skills and they are lifelong treasure.

Last but not the least, I would like to thank my colleagues Ms. Martha Paisi, Mr. Anastasios Plessas, Ms. Kathy Redfern, Mr. Safaa Mezban, Mr. Sebastian Stevens, Dr. Garry Famham, Mr. Robert O'Hara, Ms. Zoe Allen, Dr. Annegret Schneider, Ms. Lynsey Williams, Ms. Wajnat Tounsi, Dr. Majed Alghoribi, Ms. Marta Ksiazczyk and Mr Abdu Aldarhami in C507 Portland Square building, University of Plymouth. Also, I would like to thank my family, friends, and those who directly and indirectly involved in this project for their supports and encouragements all through my studies. To each of the above, I would like to extend my deepest appreciation.

Abbreviations

AA ..... Alzheimer's Association

ACC ..... Accuracy

AD ..... Alzheimer's Disease

ANN ..... Artificial Neural Networks

APOE ..... Apolipoprotein

AUC ..... Area Under Curve

BC ..... Breast Cancer

BPCA ..... Bayesian Principal Component Analysis

CE ..... Cerebellum

CNL ..... Control Group

CQT ..... Cassandra Query Language

CSF ..... Cerebrospinal Fluid

CT ..... Computed Tomography

CV ..... Cross Validation

DEM ..... Differentially Expressed Molecule

DEG ..... Differentially Expressed Gene

DTI ..... Diffusion Tensor Imaging

EC ..... Entorhinal Cortex

EOAD ..... Early-Onset Alzheimer's Disease

EOFAD ..... Early-Onset Familial Alzheimer's Disease

FAD ..... Familial Alzheimer's Disease

FDA ..... US Food and Drug Administration

FDR ..... False Discovery Rate

FFPE ..... Formalin-Fixed Paraffin-Embedded

FNR ..... False Negative Rate

fMRI ..... Functional Magnetic Resonance Imaging

FOR ..... False Omission Rate

FPR ..... False Postive Rate

GEO ..... Google Expression Omnibus

GFS ..... Google File System

GO ..... Gene Ontology

GP ..... Genetic Programming

GSH ..... Glutathione

GWAS ..... Genome-Wide Association Study

HD ..... Huntington's disease

HDFS ..... Hadoop Distributed File System

HIP ..... Hippocampus

IG ..... Information Gain

IGAP ..... International Genomics of Alzheimer's Project

IWG ..... International Working Group

KNN ..... K-Nearest Neighbour

KBFP ..... Knowledge-Based Feature Pool

LDA ..... Linear Discriminant Analysis

LOAD ..... Late Onset Alzheimer's Disease

LOOCV ..... Leave-One-Out Cross Validation

LOWESS ..... Locally Weighted Linear Regression

MCI ..... Mild Cognitive Impairment

MMSE ..... Mini-Mental State Exam

MRI ..... Magnetic Resonance Imaging

MTG ..... Medial Temporal Gyrus

NB ..... Naive Bayes

NDC ..... Non-Demented Control

NFL ..... Neurofilament Light Protein

NIA ..... National Institute on Aging

NGS ..... Next-Generation Sequencing

NFT ..... Neurofibrillary Tangle

NPV ..... Negative Predictive Value

OLGs ..... Oligodendroglia

OOB ..... Out-Of-Bag

PC ..... Posterior Cingulate

PCA ..... Principal Component Analysis

PD ..... Parkinson's Disease

PI ..... Performance Index

PPV ..... Positive Predictive Value

PET ..... Positron Emission Tomography

PFC ..... Prefrontal Cortex (also, Frontal Cortex, in this database)

POCG ..... Postcentral Gyrus

PPI ..... Protein-protein Interaction

PQTL ..... Protein Quantitative Trait Locus

PVC ..... Primary Visual Cortex

RF ..... Random Forest

RFBE ..... Random Forest Backward Elimination

RFE ..... Recursive Feature Elimination

RI ..... Ratio Intensity

SAM ..... Significant Analysis of Microarray

SFG ..... Superior Frontal Gyrus

SN ..... Sensitivity

SNP ..... Single Nucleotide Polymorphisms

SNR ..... Signal-to-Radio

SP ..... Specificity

SPECT ..... Single Photon Emission Computed Tomography

SVD ..... Singular Value Decomposition

SVM ..... Support Vector Machine

SVMFS ..... Support Vector Machine Forward Selection

SVMTFS ..... Support Vector Machine Top Forward Selection

TC ..... Temporal Cortex

TN ..... Training

TS ..... Testing

WM ..... White Matter

WRFP ..... Whole-Range Feature Pool

VC ..... Visual Cortex

Chapter 1

Introduction

1.1 Abstract

This chapter generally discusses the encyclopedic knowledge about Alzheimer's disease (AD). An overview of Alzheimer's disease (AD) is introduced with introudction of symptoms, prevalence, pathology, prognosis, diagnosis and most importantly biomarkers to date.

1.2 Alzheimer's disease (AD) introduction

Alzheimer's disease (AD) is a common type (60 ~80%) (`https://www.alzheimers.org.uk/about-us/news-and-media/facts-media`) of dementia, with common early symptoms such as short term memory loss which progressively advancing to confusion, irritability, aggression, mood swings, trouble with language, and long-term memory loss. 50 million people worldwide are living with dementia worldwide in 2018. This number is estimated to be more than triple to 152 million by 2050 [1]. There are 7.7 million new cases of dementia each year, implying that there is a new case of dementia somewhere in the world every four seconds. The total estimated worldwide cost of dementia was US$604 billion in 2010. About 70% of the costs occur in Western Europe and North America (`http://www.alz.co.uk/`). There are 850,000 people with dementia in the

UK, with numbers set to rise to over 1 million by 2025. This will soar to 2 million by 2051. 225,000 will develop dementia this year, that's one every three minutes. (`https://www.alzheimers.org.uk/about-us/news-and-media/facts-media`). 773,502 (94%) of these people with dementia were aged 65 years or over. Non-genetic risk factors include age, midlife high blood pressure, high BMI, diabetes, cerebrovascular disease, smoking, light-to-moderate alcohol consumption and antihypertensive therapy [2].

1.2.1 Symptoms, stages and types

Memory loss that disrupts daily life may be a symptom of AD or other dementia. To be more specific, Alzheimer's association (`https://www.alz.org`) suggests the following signs and symptoms as possible indicator of onset of AD:

- Memory loss that disrupts daily life
- Challenges in planning or solving problems
- Difficulty completing familiar tasks at home, at work or at leisure
- Confusion with time or place
- Trouble understanding visual images and spatial relationships
- New problems with words in speaking or writing
- Misplacing things and losing the ability to retrace steps
- Decreased or poor judgment
- Withdrawal from work or social activities
- Changes in mood and personality

1.2.1.1 Different stages of AD

Based on the intensity of the typical Alzheimer's symptoms, it can be classified into several subtypes and stages.

Mild AD (early stage) This includes the beginning of cognitive impairment that causes difficulties in remembering daily routine such as tasks at work, paying bills , and others. Because these symptoms are not very serious, the patients at this stage manage to remain functional with a certain amount of difficulty. They take longer to perform the same task which they used to do quicker before, and this becomes a pattern.

Moderate AD (middle stage) Because of a significant amount of neuronal damage, the symptoms of moderate Alzheimer's are more intense. The confusion becomes worse and due to the amount of memory loss, they become increasingly dependent on others. These individuals, even though physically agile, are not able to perform routine tasks as the delusions take over the sensory processing of their thoughts.

Severe AD (late stage) As the plaques and tangles spread, the brain cells start dying. This results in shrinkage of brain tissue. The patients with this condition are typically bedridden and are hardly able to communicate.

These subtypes are more like stages of the disease, and it often progresses from a milder to a more severe form. The sooner the patient is diagnosed with the condition, the better are the chances of treating and preventing its progression.

1.2.1.2 Mild cognitive impairment (MCI)

Mild cognitive impairment (MCI) is a condition where someone has minor or subtle problems with cognition - their mental capabilities such as memory or thinking. In MCI these difficulties are worse than would normally be expected for a healthy person of their age. However, the symptoms are not intense enough to interfere remarkably with daily life, and so are not defined as dementia.

It is estimated that between 5 and 20% of people aged over 65 have MCI (`https://www.alzheimers.org.uk/`). It is not regarded as a type of dementia, but a person identified with MCI is more likely to continue to develop dementia in the future.

Symptoms The term MCI describes a set of symptoms, rather than a specific disease. A person with MCI has mild problems with one or more of the following:

- Memory - for example, forgetting recent events or reiterating the same question
- Reasoning, planning or problem-solving - for example, struggling with thinking things through
- Attention - for example, being very easily distracted language - for example, taking much longer than usual to find the right word for something
- Visual depth perception - for example, struggling to interpret an object in three dimensions, judge distances or navigate stairs

These symptoms may have been noticed by the individual, or by those who know them. For a person with MCI, these changes may cause them to experience minor problems or need a little help with more demanding daily tasks (for example paying bills, managing medication, driving). However, MCI does not cause major problems with everyday living. If there is a significant impact on everyday activities, this may suggest dementia.

Most healthy people experience a gradual decline in mental abilities as part of ageing. In someone with MCI, however, the decline in mental abilities is greater than in normal ageing. For example, it's common in normal ageing to have to pause to remember directions or to forget words occasionally, but it's not normal to become lost in familiar places or to forget the names of close family members.

If the person with MCI has seen a doctor and taken tests of mental abilities, their problems will also be shown by a low test score or by falling test scores over time. This decline in mental abilities is often caused by an underlying illness.

Why is MCI important for AD study Neuropathological and neuroimaging evidence suggests that biological changes associated with dementia, and Alzheimer's disease (AD) in particular, occur long before, perhaps decades before, the onset of symptoms [3,

4]. Given this, it is therefore probable that there are indicators of incipient dementia occurring before the onset of the full dementia syndrome [5]. The syndrome of MCI appeared to have become widely accepted [6] as a general concept that is of subjective memory impairment in the context of cognitive impairment relative to age-matched controls and yet no loss of function and no dementia.

The firmly established concept of MCI have huge medical potentials, because it is very likely that a disease-modifying therapy will have only limited use in those with established disease because of extensive neuronal loss. Therefore one possible therapeutic approach would be to test new drugs and to use proven disease-modifying compounds in those at risk of developing dementia. Abundant evidence of high conversion from MCI to AD support the notion that MCI is the harbinger of dementia [4, 5, 7–13], which means people with this syndrome can be suitable cases for evaluation and treatment. Consequently, breakthrough of MCI pathology will definitely have massive positive impact on AD diagnosis and drug discovery, so MCI is very important for AD study.

1.2.1.3 Early onset AD (EOAD) and late onset AD (LOAD)

Despite AD mostly occurs in the elderly, a small proportion of AD occurred before the age of 65, which is called early onset AD (EOAD), and is considered to have a more aggressive course and shorter relative survival time [14]. About 5% of patients develop symptoms before age 65, where most of these patients have the sporadic form of the disease, but 10-15% have a genetic form that is generally inherited as an autosomal dominant fashion [15]. For the majority of AD cases, they are late onset AD (LOAD) that occurred after the age of 65.

1.2.1.4 Familial AD (FAD))

Familial Alzheimer's disease (FAD) is a form of Alzheimer's disease that doctors know for certain is linked to genes. In families that are affected, members of at least two generations have had the disease. FAD makes up less than 1% of all cases of Alzheimer's. Most people who have early onset Alzheimer's have FAD.

1.2.2 AD pathophysiology

The debate of AD pathophysiology can date back to the time in 1907 when Alzheimer observed the neuropathological traits of the disease i.e. amyloidal plaques and hyperphosphorylated neurofibrillary tangles (NFTs). Several hypotheses have been proposed on the basis of the diversed causative factors so as to explain this multifactorial disorder [16] such as the $A\beta$ hypothesis, cholinergic hypothesis, inflammation hypothesis and tau hypothesis [17]. Recently it has been implied that the most acknowledged $A\beta$ hypotheses, commonly used for the last two decades, is not responsibe for the complex pathophysiology of this incapacitating disease [18]. Recent studies have also remarked the role of $A\beta$ oligomers in synaptic impairment, indicating that these are basically the only one out of several other signals that disintegrate brain functions [18–21]. And formations of amyloid plaques that develop in the later age appear to be rather late event [20].

According to the amyloid cascade hypothesis, the APP is aberrantly processed by $\beta-$ and $\gamma-$secretases but normally cleaved by $\alpha-$secretase leading to an imbalance between production and clearance of $A\beta$ peptide [22]. Consequently, $A\beta$ peptides spontaneously aggregate into soluble oligomers and coalesce to assemble fibrils insoluble beta-sheet conformation and are eventually deposited in diffuse senile plaques [18]. Some recent studies has indicated that $A\beta42$ oligomers are formed by cooperative activities of both neurons and its associated astrocytes [20]. It was found that $A\beta42$ oligomers induce oxidative damage, promote tau hyperphosphorylation, results in toxic effects on synapses and mitochondria [7, 17]. But the role of $A\beta42$ senile plaques cannot be ignored as $A\beta42$ plaques that are supposed to be appear during late stage attract microglia [23]. Microglial activation attribute to the production and discharge of proinflammatory cytokines, including IL-1β, TNF-α, and IFN-γ. In turn, these cytokines stimulate the nearby astrocyte–neuron to produce further amounts of $A\beta42$ oligomers, thus activating more $A\beta42$ production and dispersal [20]. Oligodendroglia (OLGs) also has association with neurons–astrocyte complex; $A\beta$ oligomers also result in its decomposition [24].

$A\beta$ oligomers aggregates are regarded to contribute to the neuronal and vascular degeneration in AD brains [25]. It causes oxidative stress, a situation to which OLGs are particularly susceptible because their reduced glutathione (GSH) content is low and they have a high concentration of iron, thus presenting an impaired ability to scavenge oxygen radicals [21]. It has also been reported that $A\beta$ oligomers possesses an increased potential for damaging cholesterol rich membranes, such as those discovered in OLGs and myelin [25, 26].

Previous studies focusing on the receptors pharmacology of $A\beta$ have implied that $A\beta 42$ monomers activate the neuroprotective signaling of insulin-like growth factor-1 receptor (IGF-1R), whereas $A\beta 42$ oligomers target a host of neurons' and astrocytes' membrane receptors, such as the scavenger receptor for advanced glycation end products (RAGE), Frizzled receptor, insulin receptor, NMDAglutamate receptor, p75 neurotrophin receptor (p75NTR), $\alpha 7$ nicotinic ACh receptor (α7nAChR), ApoE receptors, formyl peptide receptor-like 1 (FPRL1/2), cellular prion protein (PrPc) acting as an Ab oligomer receptor, and the calcium-sensing receptor (CaSR) [27, 28]. Removal of $A\beta$ oligomers from the brain happens by several pathways including proteolytic degradation by the proteases neprilysin and insulin degrading enzyme (IDE), uptake by astrocytes and microglia, passive flow into the cerebrospinal fluid and sequestration into the vascular compartment by soluble form of the low-density lipoprotein receptor related protein 1 (LRP1) [29,30]. The effect of NO on IDE-mediated degradation of $A\beta$ has been investigated an The essential effect of the astrocyte–neuron interconnections is the astrocytes abilities to promote or lessen neurotransmitters release into the synapses they envelop with the Ca_{2+} they respectively let out or take up during their Ca_{2+} waves [24]. When neurons $A\beta 42$ production exceeds the safe limit, toxic $A\beta 42$ oligomers start spilling out of the neurons and onto their enveloping astrocytes both cell types being empowered with $A\beta 42$ oligomer-binding receptors besides accumulating or dispersing in the extracellular surrounding [20]. Due to the intimate physical and functional interdigitations in the neurons client group, the $A\beta 42$ oligomers releases by the neuron can directly connected with the a- 7nAChRs of its partner astrocytes [31]. The signals from these

receptors induce the astrocytes to exocytose the glutamate they have been taking up from the neuronal synapses [31]. The discharged glutamate activates the extrasynaptic NMDARs of the astrocytes' partner neurons [31, 32]. The resulting signals trigger Ca_{2+} surges evoking a cascade of events, including dysfunctional mitochondria pumping out ROS, which inflict an oxidative damage, caspase 3 activation, tau hyperphosphorylation, excess production of NO, ROS and VEG-F thereby destroying dendritic spines and neuronal synapses and severing communications within the astrocyte's neurons and beyond [31]. Armato and others have shown that CaSRs (present on the cell membranes of astrocytes and neurons on which Ab42 oligomers binds) selective allosteric antagonist (calcilytic) NPS 2143 specifically stops the excess release of endogenous $A\beta42$ from the $A\beta$25–35-exposed human astrocytes and neurons [28, 31].

d it has been shown that increased NO levels, which have been reported in AD, can weaken enzymatic function, potentially leading to increment $A\beta$ oligomers deposition in the brain and development of AD [32].

It has recently been reported that there is a 'contagion' like diffusion of $A\beta42$ oligomers and hyperphosphorylated tau oligomers via exocytosis (synapses) or exosomes to closely associated target cells (astrocytes and oligodendrocytes), which in turn become producer cells of $A\beta$ and tau oligomers [33]. Experimental evidence have implied that intracerebral (i.c.) administration of small amounts of brain extract containing misfolded $A\beta$ from patients with AD or from $A\beta$-APP transgenic (tg) mice induces cerebral β-amyloidosis and related pathologies in APP tg mice in a time- and concentration-dependent manner [33].

The essential effect of the astrocyte–neuron interconnections is the astrocytes abilities to promote or lessen neurotransmitters release into the synapses they envelop with the Ca_{2+} they respectively let out or take up during their Ca_{2+} waves [24]. When neurons $A\beta42$ production exceeds the safe limit, toxic $A\beta42$ oligomers start spilling out of the neurons and onto their enveloping astrocytes both cell types being empowered with $A\beta42$ oligomer-binding receptors besides accumulating or dispersing in the extracellular surrounding [20]. Due to the intimate physical and functional interdigitations

in the neurons client group, the $A\beta 42$ oligomers releases by the neuron can directly connected with the a- 7nAChRs of its partner astrocytes [31]. The signals from these receptors induce the astrocytes to exocytose the glutamate they have been taking up from the neuronal synapses [31]. The discharged glutamate activates the extrasynaptic NMDARs of the astrocytes' partner neurons [31, 32]. The resulting signals trigger Ca_{2+} surges evoking a cascade of events, including dysfunctional mitochondria pumping out ROS, which inflict an oxidative damage, caspase 3 activation, tau hyperphosphorylation, excess production of NO, ROS and VEG-F thereby destroying dendritic spines and neuronal synapses and severing communications within the astrocyte's neurons and beyond [31]. Armato and others have shown that CaSRs (present on the cell membranes of astrocytes and neurons on which Ab42 oligomers binds) selective allosteric antagonist (calcilytic) NPS 2143 specifically stops the excess release of endogenous $A\beta 42$ from the $A\beta$25–35-exposed human astrocytes and neurons [28, 31].

1.3 AD Biomarker

A biomarker is defined as a biological feature found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition. For any diseases including AD, the discovery of biomarkers will contribute to better understanding of the pathological process or cause of the onset of diseases, and could potentially leads to a successful treatment of the disease. The biomarkers can be of any different categories, from biochemical molecules, magnetic resonance imaging, to electrical signal. The diagnosis and treatment of many diseases such as cancer have benefit from the emergence of biomarkers in recent decades.

1.3.1 Cerebrospinal fluid biomarkers

For many decades, the definitive diagnosis of AD has depended on the postmortem detection of senile plaques (SPs) and neurofibrillary tangles (NFTs). The neuropathology

of AD is now better understood in relation to amyloid and tau pathology – consequently $A\beta$ and tau assays having been explored and examined during the last two decades to offer first 'core feasible' cerebrospinal fluid (CSF) biomarkers [34]. The biochemical composition of Cerebrospinal fluid (CSF) is also able to provide information on the brain chemistry. In the early 90s, the first publication documented elevated CSF amounts of t-tau in patients with AD dementia [35]. After that, augmented CSF concentrations of p-tau and reduced levels of $A\beta 1\text{–}42$ have been described. Replications of these findings were discovered by large numbers of studies. Compared to cognitively normal elderly subject, a level decrease of CSF $A\beta 1\text{–}42$ to about 50%, an increase in CSF t-tau to approximately 300% and a less evident growth in CSF p-tau to about 200% were reported recursively [36]. Combing $A\beta 1\text{–}42$ with tau provides excellent discriminative value for AD patients against age-matched healthy controls, with a sensitivity of 85% and a specificity of 86% (refer to equation 2.25 about sensitivity and specificity). Nonetheless, a deterioration of specificity is observed when these ratios are used to discriminate AD from other dementias [37]. Several other combinations of tau, $A\beta 1\text{–}38$, $A\beta 1\text{–}40$ and $A\beta 1\text{–}42$ were also trialed in many studies to execute the discriminations between AD, non-Alzheimer dementias (NAD) and healthy controls, among which the ratio $A\beta 1\text{–}42/A\beta 1\text{–}38$/p-tau robustly distinguished AD against NAD, reaching the requirements for an applicable screening and differential diagnostic AD biomarker [38, 39]. Except those core biomarkers with high certainty of pathological involvement, other molecules were also found to be of help for diagnosis. Notably, one study has reported the increase of concentration of eight amino acids in AD versus MCI [40], another larger examination has also managed to detect eight molecules with statistical significance [41].

1.3.2 Blood-based biomarkers

Blood is a biofluid which is much more easily reached and manageable than CSF, therefore, searching for consistent blood-based biomarkers is required. Yet, definite data regarding the association of plasma $A\beta 1\text{–}40$ and $A\beta 1\text{–}42$ concentrations with incipient

AD is presently lacking [34]. Some studies revealed that increased Aβ1–40 or Aβ1–42 levels be an indicator of development of AD [42, 43], while other analyses showed no associations [44, 45] or opposite [46] results. A low $A\beta 1\text{–}42/A\beta 1\text{–}40$ ratio is reported to foresee the onset of future AD [42, 47, 48], conversely, an increased [43, 49] or no obvious difference [44] between incipient AD subjects and healthy controls are also reported. It is described in a recent meta-analysis study that low $A\beta 1\text{–}42/A\beta 1\text{–}40$ ratio could forecast the progression of AD, yet no evidences of such link have been presented for single peptides [50]. In regard to tau, some studies have emphasized the discrepancy in the modulation of CSF tau levels comparing with blood. In case of hypoxic brain damage following cardiac arrest, tau is dramatically cleared after promptly released into blood [51], on the contrary, CSF tau levels remain high for several weeks after an acute neurological insult [52]. Moreover, remarkable increase in the tau level is observed in CSF of AD patients but less in commensurate plasma samples. In fact, the correlation between the tau levels in CSF and plasma compartment does not exist [53]. Recently a cross-sectional study has shown that the plasma concentration of tau appears to increase in AD samples compared to MCI and healthy controls. MCI-AD samples (i.e. MCI converters to AD) presents similar levels of tau as those detected in MCI-stable (i.e. MCI non-converters to AD) and healthy controls [54].

Last decade witnessed the development of mass spectrometry-based technologies that has elected proteomics as the chief platform to inspect the plasma/serum proteome for the discovery of next-generation biomarkers showing diagnostic, prognostic, or therapeutic efficacy [55]. The capability of synchronously quantifying large number of plasma analytes enables the discovery of more biomarker patterns which help distinguish AD patients from healthy controls [56, 57]. Although the existence of an inflammatory process in AD has been confidently suggested, based on the pathways those biomarkers are connected with, such panel of proteins are still difficult to reproduce in independent studies [58]. No findings of transcripts/proteins/metabolites in blood have been successfully replicated to be definitively approved as AD biomarkers [34]. In spite of the unsatisfactory repeatability of protein biomarkers, some encouraging findings with

promising diagnostic accuracy across cohorts still support the prospect of blood-based protein biomarkers as the future of AD diagnosis [57, 59, 60].

1.3.3 Neuroimaging biomarkers

Reduction of hippocampus volume, detected from structural MRI, is among the principal biomarkers of AD in the International Working Group (IWG) [61] and NIA-AA (the National Institute on Aging (NIA) at National Institutes of Health (NIH) and the Alzheimer's Association(AA)) criteria [62] (refer to Section 1.4.2). A reduced hippocampus volume has been found across multiple studies in AD and MCI subjects (for meta-analysis see [63]). However, it is not specific to AD and is found in other conditions, including fronto-temporal dementia [64], vascular dementia [65], Lewy-body dementia [66] and depression [67]. Another structural marker is whole brain volume as longitudinal marker of disease progression and treatment effects, and it has already been used as secondary endpoint in some clinical trials [68]. But it obtained less attention with the emergence of regionally more specific protocols, which are derived from local measures of grey matter concentration or cortical thickness. The foundation of the application of MRI is closely related to the assumption that regional brain volume can serve as in vivo surrogate of neuronal number, which some previous studies have provided evidences [69, 70]. However, more investigations are needed to address the associations between regional brain atrophy and regional markers of neuronal degeneration, because there is a study suggesting that cortical atrophy can reflect changes in neuronal and dendritic architecture, rather than regional neuron numbers and density [71].

Diffusion Tensor Imaging (DTI), a magnetic resonance (MR) technique, records the random thermal motion of water molecules, i.e. Brownian motion, within tissue [72]. It is a non-invasive way to obtain information of axonal organization of the brain. In the last decade, this technique has taken the lead to investigate white matter (WM) microarchitecture and integrity. It has been widely employed in AD and MCI [52, 73–75].

Functional MRI (fMRI) takes advantages of the blood-oxygen-level-dependent (BOLD)

contrasts in vascular capillary network surrounding the cerebral cortex to learn neuronal activity through non-invasive means during particular cognitive states. Several MRI studies have been able to recognize functional alternations before the onset of cognitive impairment or AD-related structural neurodegeneration [3, 76, 77].

Positron emission tomography (PET) and single-photon emission CT (SPECT) have long been extensively assessed as diagnostic instruments for dementia, and both techniques have given satisfying prognostic and diagnostic abilities. A few PET ligands targeting amyloid, tau, or metabolic activity have been investigated. An early radiotracer reported to bind both to amyloid plaques and NFT [78] is 2-(1-{6-[(2-[F-18]fluoroethyl)(methyl)amino]-2-naphthy- l}ethylidene)malononitrile (FDDNP). Cerebral metabolism, as measured by ^{18}F-fluorodeoxyglucose (FDG)-PET imaging is found to be decreased in AD. Several studies have reported the reductions in regional cerebral glucose metabolism in MCI and AD comparing with healthy controls [79, 80]. FDG-PET measures are strongly associated with cognitive deficits [80, 81]. Although PET is a reportingly promising approach for AD diagnosis, its high operational cost and restriction to specialist centres are the largest problem that researchers are facing forward.

1.4 AD prognosis and diagnosis

In 2011, clinical diagnostic criteria for Alzheimer's disease dementia were revised, and research guidelines for earlier stages of the disease were characterized to reflect a deeper understanding of the disorder. Development of the new guidelines was led by the National Institutes of Health and the Alzheimer's Association [82].

1.4.1 Early diagnosis

Research on new strategies for earlier diagnosis is among the most active areas in Alzheimer's science, with the hope that future treatments can target the disease in its earliest stages, before irreversible brain damage or mental decline has occurred. Several potential biomarkers are being studied for their ability to indicate early stages of

Alzheimer's disease. Examples being studied include beta-amyloid and tau levels in cerebrospinal fluid (CSF) and brain changes detectable by imaging. Recent research suggests that these indicators may change at different stages of the disease process.

Biomarker It is believed that biomarkers offer one of the most promising paths for an easy and accurate way to detect AD before devastating symptoms begin.

Blood and urine tests It is also investigated that whether AD causes consistent, measurable changes in blood or urine levels of tau, beta-amyloid or other biomarkers before symptoms appear. Moreover, whether early AD leads to detectable changes elsewhere in the body, such as the lens of the eye, has also been under investigation.

Brain imaging and neuroimaging Neuroimaging is among the most promising areas of research focused on early detection of Alzheimer's disease. To summarize, the following three types of neuroimaging technologies are used in AD research:

- **Structural imaging** provides information about the shape, position or volume of brain tissue. Structural techniques include magnetic resonance imaging (MRI) and computed tomography (CT).

- **Functional imaging** reveals how well cells in various brain regions are working by showing how actively the cells use sugar or oxygen. Functional techniques include positron emission tomography (PET) and functional MRI (fMRI).

- **Molecular imaging** uses highly targeted radiotracers to detect cellular or chemical changes linked to specific diseases. Molecular imaging technologies include PET, fMRI and single photon emission computed tomography (SPECT).

Genetic risk profiling Early-onset familial Alzheimer's disease (EOFAD) is caused by rare and highly penetrant mutations in three genes, namely: amyloid precursor protein (APP, located at chromosome region 21q21.2), presenilin 1 (PSEN1, located at

14q24.3), and presenilin 2 (PSEN2, located at 1q42.13). Presently, more than 220 distinct disease-causing mutations have been discovered across these genes [83]. In contrast to EOFAD, late-onset Alzheimer's disease (LOAD) exhibits a significantly more complex and intricate pattern of interplay between genetic and non-genetic factors. The earliest and by far best established genetic risk factor for LOAD is the presence of one or two copies of the e4 allele in the apolipoprotein E gene (APOE), located on chromosome 19q13.2 [84]. The second wave of genome-wide association study (GWAS) which have led to the identification of at least nine loci linked to mostly LOAD risk: BIN1, CLU, ABCA7, CR1, PICALM, MS4A, CD33, CD2AP, and EPHA1. Most recently, in the largest GWAS published to date which included 74,046 subjects and a large two-stage meta-analysis, 11 novel loci were discovered, the relevant genes are HLA-DRB5-HLA-DRB1, PTK2B, SORL1, SLC24A4-RIN3, NME8, ZCWPW1, FERMT2 [85].

Cerebrospinal fluid (CSF) proteins CSF is fluid that supports the brain and spinal cord. Adults have approximately one pint of CSF, which physicians can sample through a minimally invasive procedure called a lumbar puncture, or spinal tap. Research suggests that early stage AD may cause changes in CSF levels of tau and beta-amyloid, two proteins that generate abnormal brain deposits strongly linked to AD.

1.4.2 Clinical diagnosis

No single test has yet existed to prove an individual is AD patient or not. The physician may be able to make judgement about whether an individual has dementia, knowing the exact cause can be difficult. Diagnosing AD involves a complete assessment that considers all possible causes, as following describes:

Medical history To collect any current and past illnesses, medications, key medical conditions affecting other family members, including whether they may have had AD or other dementias.

Physical exam and diagnostic tests To collect information from a physical exam and laboratory tests, which help distinguish dementia from other diseases that have the same symptoms. Conditions other than AD that may cause confused thinking, trouble focusing or memory problems include anemia, infection, diabetes, kidney disease, liver disease, certain vitamin deficiencies, thyroid abnormalities, and problems with the heart, blood vessels and lungs.

Neurological exam During a neurological exam, the physician will closely assess the individual for problems that may signal brain disorders other than AD. Signs of small or large strokes, Parkinson's disease, brain tumors, fluid accumulation on the brain, and other illnesses that may impair memory or thinking, are all on the checklist.

Mental status tests Mental status testing evaluates memory, ability to solve simple problems and other thinking skills.

Mini-Mental State Exam (MMSE) During the MMSE, a health professional asks a patient a series of questions designed to assess a number of everyday mental skills. The maximum MMSE score is 30 points. A score of 20 to 24 suggests mild dementia, 13 to 20 suggests moderate dementia, and less than 12 indicates severe dementia. On average, the MMSE score of an AD patient declines about two to four points each year.

Mini-Cog test During the Mini-Cog, the individual is required to complete two tasks, first to remember and a few minutes later repeat the names of three common objects, and second to draw a face of a clock showing all twelve numbers in the right places and a time specified by the examiner.

Brain imaging A standard AD medical tests often includes structural imaging with magnetic resonance imaging (MRI) or computed tomography (CT). These tests are mainly employed to rule out other conditions that may render symptoms similar to AD but require different treatments. Structural imaging can reveal tumors, evidence of small

or large strokes, damage from severe head trauma, or a buildup of fluid in the brain.

1.5 Current treatment

Currently and/or 'only' approved treatments by US Food and Drug Administration (FDA), includes five drugs that are used to treat the cognitive manifestations of AD AChEIs–rivastigmine (Exelon), galantamine (Razadyne, Reminyl), tacrine (Cognex), and donepezil (Aricept) and NMDA receptor antagonist–memantine (Namenda) that target symptoms at its best [86]. Each drug acts in a different way to delay the breakdown of Ach (a chemical in the brain important for memory). AD is associated with inadequate levels of this important neurotransmitter. Tacrine (Cognex) is rarely prescribed due to its serious side effects (liver damage). In general, Reminyl, Exelon and Aricept are most effective when treatment is begun in the early stages. Memantine (Namenda) is the only drug shown to be effective for the later stages of the disease. They have all been shown to modestly slow the progression of cognitive symptoms and reduce problematic behaviors in some people, but at least half of the people who take these drugs do not respond to them. These present treatment strategies only delay the progression of symptoms associated with AD [87]. Much effort is being directed towards the discovery of disease-modifying therapies which can block the progression of the disease (i.e. clinical symptoms) and drugs targeting various molecular pathways. For development of disease modifying therapies complete knowledge about the various metabolic pathways is essential which includes production of $A\beta$ from APP, in vivo clearance and pathophysiological events that leads to fibril formation and deposition into plaques [17].

1.6 Aims of this thesis

This project aims to finding biomarkers by employing machine learning approaches to microarray transcriptomics data and proteomics data and DNA methylation data. The underlining hypothesis is that there exists multiple biomarkers in transcriptomics, pro-

teomics and epigenomics which show high relevance to the onset of AD, therefore can be used for diagnosis with promising performance.

The objectives are following:

- To go through a systematic study of machine learning approaches, mostly focus on SVM, random forests and their implementations
- To fulfil a detailed investigation of AD on its general mechanism and diagnosis methods currently applied, understand the advantage and disadvantage
- To select different kinds of significant biomarker by employing effective and robust feature selection methods in machine learning.
- To innovate novel methods to integrate genetics and pathway information, protein-protein interaction information in the machine learning modelling to enhance the performance in training and testing process.
- To develop robust application or software packages that can facilitate future research of AD or dementia either by providing statistical analysis or presenting benchmark results.

1.7 Synopsis of the thesis

This thesis is composed in the following order: Chapter 1 introduces the background of Alzheimer's disesae, its pathology, diagnosis and biomarker in all types. Chapter 2 introduces statistical learning approaches that is related or going to be used within the range of the thesis. Chapter 3 discusses some novel AD transcriptomic biomarker findings in brain, and chapter 4 discusses some interesting proteomic biomarker findings in blood, which was published as research article with the title of "Discovery of Novel Biomarkers for Alzheimer's Disease from Blood" in 2016. Chapter 5 describes applying meta-analysis in large scale studies. The R package metaUnion that overcame the

dimension persistency limit across all the studies in normal approaches, acted a crucial role in a comprehensive AD brain biomarkers investigation. The work was published with the title of "Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease" in 2015, and the R package metaUnion is open source in Github. A systematic database for archiving and streaming analysis is introduced in Chapter 6. Supported by powerful big data technology such as NoSQL, the system aims to automate and unify the procedure of bioinformatic data research, from data import and processing, to data analysis, presentation and visualization. Remarkable extensibility for diversified experiment type is one of the highlights. Chapter 7 describes the main contriution of the thesis and Chapter 8 is the summary and plan for future work.

Chapter 2

Methods

2.1 Abstract

This chapter generally discusses the main methodology used throughout the entire process of the project - statistical learning. State-of-the-art statistical learning algorithms and their application are also described, in addition to a brief introduction of big data technologies in the area of database models.

2.2 Data type and technology overview

Throughout the thesis, various types of data are employed to serve the research purpose. Microarray data extracted from RNA transcript and protein and DNA sequencing data are the main sources that were analysed. For future extension, novel technologies such as DNA methylation and RNA sequencing will also be potential entry points for new research ideas.

2.2.1 Omics technology

The complete seequencing of the human genome has ushered in a new era of system biology referred as **Omics technology**. The term "omics" refers to the comprehensive analysis of biological systems. Modern use of the term "omics" derive from the term

genome (hence genomics), a term derived invented by Hans Winkler in 1920, although the use of -ome is older, signifying the "collectivity" of a set of things. The word "genomics" is said to be appeared in the 1980s and became widely used in the 1990s. The first genome was completely sequenced by Sanger in Cambridge, UK, in the 1970s. Genome is the most fundamental part of many omics.

The word "Genomics" implies some hidden network among negetic elements. This network is regulated by many other omics such as proteomics, transcriptomics, metabolomics and physiomics. In this section, several omics technologies relating heavily to our research are introduced.

- Genomics, the study of genes and their funciton
- Proteomics, the study of proteins
- Metabonomics, the study of molecules involved in cellular metabolism
- Transcriptomics, the study of mRNA
- Glycomics, the study of cellular carbohydrates
- Lipomics, the study of cellular lipids

Omics technologies provide the tools needed to look at the differences in DNA, RNA, proteins, and other cellular molecules between species and among individuals of a species. These types of molecular profiles can vary with cell or tissue exposure to chemicals or drugs and thus have potential use in toxicological assessments. Omics experiments can often be conducted in high-throughput assays that produce tremendous amounts of data on the functional and/or structural alterations within the cell. "These new methods have already facilitated significant advances in our understanding of the molecular responses to cell and tissue damage, and of perturbations in functional cellular systems" [88].

2.2.2 Microarray technology

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene (Figure 2.1A). The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesised by the process of photolithography.

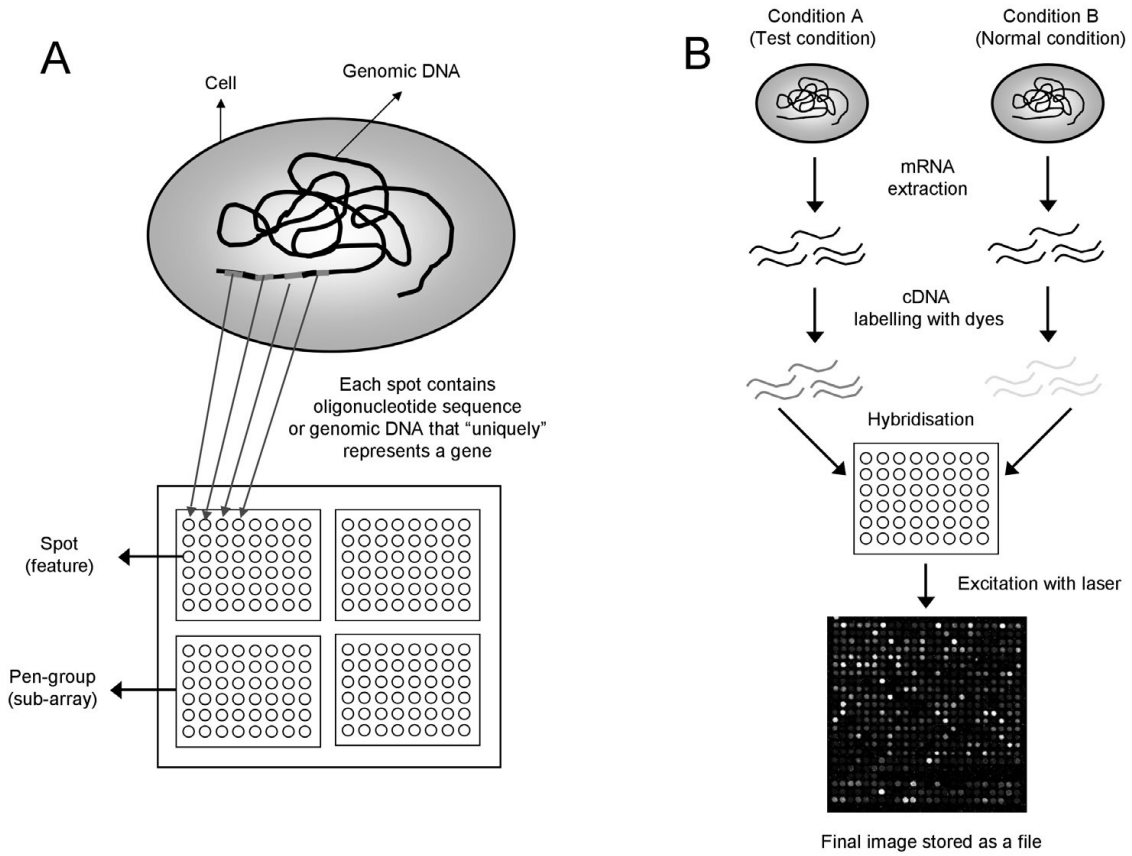


Figure 2.1: (A) A microarray may contain thousands of "spots". Each spot contains many copies of the same DNA sequence that uniquely represents a gene from an organism. Spots are arranged in an orderly fashion into Pengroups. (B) Schematic of the experimental protocol to study differential expression of genes. The organism is grown in two different conditions (a reference condition and a test condition). RNA is extracted from the two cells, and is labelled with different dyes (red and green) during the synthesis of cDNA by reverse transcriptase. Following this step, cDNA is hybridized onto the microarray slide, where each cDNA molecule representing a gene will bind to the spot containing its complementary DNA sequence. The microarray slide is then excited with a laser at suitable wavelengths to detect the red and green dyes. The final image is stored as a file for further analysis. Image source: `https://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray`

2.2.2.1 DNA microarray

Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Figure 2.1B gives a general picture

of the experimental steps involved. First, RNA is extracted from the cells. Next, RNA molecules in the extract are reverse transcribed into cDNA by using an enzyme reverse transcriptase and nucleotides labelled with different fluorescent dyes. For example, cDNA from cells grown in condition A may be labelled with a red dye and from cells grown in condition B with a green dye. Once the samples have been differentially labelled, they are allowed to hybridize onto the same glass slide. At this point, any cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples.

Following the hybridization step, the spots in the hybridized microarray are excited by a laser and scanned at suitable wavelengths to detect the red and green dyes. The amount of fluorescence emitted upon excitation corresponds to the amount of bound nucleic acid. For instance, if cDNA from condition A for a particular gene was in greater abundance than that from condition B, one would find the spot to be red. If it was the other way, the spot would be green. If the gene was expressed to the same extent in both conditions, one would find the spot to be yellow, and if the gene was not expressed in both conditions, the spot would be black. Thus, what is seen at the end of the experimental stage is an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene.

2.2.2.2 Protein microarray

Protein microarrays were developed due to the limitations of using DNA microarrays for determining gene expression levels in proteomics. They are the continuation of the DNA array approach [89]. However, owing to fundamental biochemical differences between DNAs and proteins, the chemical aspects of DNA microarrays cannot be simply applied to protein microarrays, which require a more sophisticated surface and immobilization chemistry [90]. DNAs are relatively simple polyanions and can be chemically modified and immobilized on solid surfaces based on electrostatic interaction or covalent coupling.

Proteins are chemically and structurally much more complex and heterogeneous. The activity and function is critically dependent on their delicate three-dimensional structure. Proteins tend to adsorb nonspecifically to most surfaces, which often results in the disruption of their structure. This strong tendency can be attributed to hydrophobic, ionic, and hydrogen bonding interactions with the solid surface.

2.2.3 Next-Generation Sequencing (NGS)

Next-Generation sequencing (NGS) technology [91] will be involved to enrich the data with sequencing results. It is dubbed because it is a radically different approach to decipher DNA sequences which transcends the barrier of scalability, throughput, speed and resolution compared to traditional DNA sequencing technology.

Massively parallel, this cutting-edge DNA sequencing technology has revolutionised genomic research. Using NGS an entire human genome can be sequenced within a single day. In contrast, the previous Sanger sequencing technology, used to decipher the human genome, required over a decade to deliver the final draft. In genome research NGS has mostly superseded conventional Sanger sequencing.

There are a number of different NGS platforms using different sequencing technologies. However, all NGS platforms perform sequencing of millions of small fragments of DNA in parallel. Bioinformatics analyses are used to piece together these fragments by mapping the individual reads to the human reference genome. Each of the three billion bases in the human genome is sequenced multiple times, providing high depth to deliver accurate data and an insight into unexpected DNA variation. NGS can be used to sequence entire genomes or constrained to specific areas of interest, including all 22000 coding genes (a whole exome) or small numbers of individual genes.

Though being overwhelmingly advantageous in genome research, NGS has not yet translated into routine clinical practice. This technology has huge potential to be used in clinical genetics, microbiology studies and oncology investigations.

2.2.3.1 Genome-Wide Association Study (GWAS)

Genome-wide association study (GWA study, or GWAS), is an examination of numerous common genetic variants in various individuals to see if any variant is related to a trait. GWAS typically targets on associations between single-nucleotide polymorphisms (SNPs) and traits like specific genotyping or major diseases.

Genome-wide association studies (GWAS) are a powerful hypothesis-free tool for the dissection of susceptibility to common heritable human diseases, including osteoporosis. To date, more than 2000 loci for common human diseases have been identified by GWAS. Success using the GWAS model depends on genetic risk being determined by shared stretches of DNA carried with different frequencies in cases and controls, inherited from ancient ancestors, termed the "common disease–common variant" hypothesis. Not all disease risk is caused by common variants, however, and thus GWAS will not detect all variants involved. Successful GWAS performance requires careful quality control, especially as the effect sizes under study are modest, and there are multiple potential sources of error. Conservative interpretation, use of stringent significance thresholds, and replication in independent cohorts are required to ensure results are robust. Despite these challenging parameters, much has been learnt from GWAS and, as the approach matures and is modified to identify a wider range of variants, significantly more will be learnt about the etiopathogenesis of common diseases such as osteoporosis.

2.2.4 DNA methylation

Recently, a trend of investigating with the alternation of DNA methylation, which is a biochemical process, has started. In this process, a methyl group is added to the cytosine or adenine DNA nucleotides. Pathologic DNA methylation may increase the likelihood of modified expression of genes in cells when cells make a division and differentiate from stem cells into specific tissues. Therefore it is recently employed as biomarker of diseases.

This stable and heritable covalent modification mostly affects cytosines in the context of a CpG dinucleotide in humans. It can be detected and quantified by a number of

technologies including genome-wide screening methods as well as locus- or gene-specific high-resolution analysis in different types of samples such as frozen tissues and FFPE samples, but also in body fluids such as urine, plasma, and serum obtained through non-invasive procedures. In some cases, DNA methylation based biomarkers have proven to be more specific and sensitive than commonly used protein biomarkers, which could clearly justify their use in clinics. However, very few of them are at the moment used in clinics and even less commercial tests are currently available.

2.3 Data preprocessing technique

The actual expression data needs to be extracted before analysing microarray data, which is then followed by the conduction of preprocessing for the data.

2.3.1 General data processing before experiment design

2.3.1.1 Image processing

The first step in the analysis of microarray data is to process this image. Most manufacturers of microarray scanners provide their own software.

Image processing includes the following procedures:

- Identification of the spots and distinguishing them from spurious signals.
- Determination of the spot area to be surveyed, determination of the local region to estimate background hybridization.
- Reporting summary statistics and assigning spot intensity after subtracting for background intensity.

We saw that the relative expression level for a gene can be measured as the amount of red or green light emitted after excitation. The most common metric used to relate this information is called expression ratio. It is denoted here as Tk and defifi ned as:

$$t_k = {}^{R_k}/_{G_k} \tag{2.1}$$

For each gene k on the array, where R_k represents the spot intensity metric for the test sample and G_k represents the spot intensity metric for the reference sample. As mentioned above, the spot intensity metric for each gene can be represented as a total intensity value or a background subtracted median value.

2.3.1.2 Transformation

To eliminate this inconsistency in the mapping interval which happens in the case of expresson ratio, one can perform two kinds of transformations of the expression ratio, namely, inverse transformation and logarithmic transformation.

Inverse or reciprocal transformation The inverse or reciprocal transformation converts the expression ratio into a fold-change, where for genes with an expression ratio of less than 1 the reciprocal of the expression ratio is multiplied by -1. If the expression ratio is larger than 1 then the fold change is equal to the expression ratio. The advantage of such a transformation is that one can represent upregulation and down-regulation with a similar mapping interval.

However, this method also has a problem in that the mapping space is discontinuous between –1 and $+1$ and hence becomes a problem in most mathematical analyses downstream of this step.

Logarithmic transformation A better transformation procedure is to take the logarithm base 2 value of the expression ratio (i.e. $log_2(expressionratio)$). This has the major advantage that it treats differential up-regulation and down-regulation equally, and also has a continuous mapping space. For example, if the expression ratio is 1, then $log_2(1)$ equals 0 represents no change in expression. If the expression ratio is 4, then $log_2(4)$ equals $+2$ and for expression ratio of $log_2(1/4)$ equals -2. Thus, in this transformation the mapping space is continuous and upregulation and down-regulation

are comparable.

Having explained the advantages of using expression ratios as a metric for gene expression, it should also be understood that there are disadvantages of using expression ratios or transformations of the ratios for data analysis. Even though expression ratios can reveal patterns inherent in the data, they remove all information about absolute expression levels of the genes. For example, genes that have R/G ratios of 400/100 and 4/1 will end up having the same expression ratio of 4, and associated problems will surface when one tries to reliably identify differentially regulated genes.

2.3.1.3 Normalization

When microarry is applied in biological studies, an ideal scenario which perfectly control all variables cannot be expected. This may be due to various reasons, for example, variation caused by differential labelling efficiency of the two fluorescent dyes or different amounts of starting mRNA material in the two samples. Thus, in the case of microarray experiments, as for any large-scale experiments, there are many sources of systematic variation that affect measurements of gene expression levels.

Normalization is a term that is used to describe the process of eliminating such variations to allow appropriate comparison of data obtained from the two samples.

The first step in a normalization procedure is to choose a gene-set (which consists of genes for which expression levels should not change under the conditions studied, that is the expression ratio for all genes in the gene-set is expected to be 1. From that set, a normalization factor, which is a number that accounts for the variability seen in the geneset, is calculated. It is then applied to the other genes in the microarray experiment.

Total intensity normalization The basic assumption in a total intensity normalization is that the total quantity of RNA for the two samples is the same. Also assuming that the same number of molecules of RNA from both samples hybridize to the microarray, the total hybridization intensities for the gene-sets should be equal. So, a normalization factor can be calculated as:

$$N_{total} = \frac{\sum_{k=1}^{N_{gene\_set}} R_k}{\sum_{k=1}^{N_{gene\_set}} G_k} \tag{2.2}$$

where R_k and G_k are expression intensity of red and green channel respectively for gene k, N_{gene_set} is the number of genes detected in the samples. The intensities are now rescaled such that $G'_k = G_k \times N_{total}$ and $R'_k = R_k$, where R'_k and G'_k are the normalised intensity of red and green channel respectively for gene k. The normalized expression ratio becomes:

$$T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{G_k \times N_{total}} = \frac{T_k}{N_{total}} \tag{2.3}$$

which is equivalent to:

$$log_2(T'_k) = log_2(T_k) - log_2(N_{total}) \tag{2.4}$$

This now adjusts the ratio such that the mean ratio for the gene set is equal to 1.

Mean log centering In this method, the basic assumption is that the mean $log2$ (expression ratio) should be equal to 0 for the gene-set. In this case, the normalization factor can be calculated as:

$$N_{mlc} = \frac{\sum_{k=1}^{N_{gene\_set}} log_2(\frac{R_k}{G_k})}{N_{gene\_set}} \tag{2.5}$$

where R_k and G_k are expression intensity of red and green channel respectively for gene k, N_{gene_set} is the number of genes detected in the samples. The intensities are now rescaled such that $G'k = G_k \times (2^{N_{mlc}})$ and $R'_k = R_k$, where R'_k and G'_k are the normalised intensity of red and green channel respectively for gene k. The normalized expression ratio becomes:

$$T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{G_k \times (2^{N_{mlc}})} = \frac{T_k}{2^{N_{mlc}}} \tag{2.6}$$

Which is equivalent to:

$$log_2(T'_k) = log_2(T_k) - log_2(2^{N_{mlc}}) = log_2(T_k) - N_{mlc} \quad (2.7)$$

This adjusts the ratio such that the mean log2 (expression ratio) for the gene-set is equal to 0.

Other normalization methods include: linear regression, Chen's ratio statistics and Lowess normalization. The next step following the normalization procedure is to filter low intensity data using specific threshold or relative threshold imposed according to the background intensity. If the experimental procedure included a replicate, averaging the values using the replicate data is the next step to be performed after data filtering. Finally, differentially expressed genes are identified.

2.3.2 Missing value imputation

Despite the wide use of microarray technology, microarray data quality is often plagued by missing value problem. It is estimated that up to 10% missing values can be contained in microarray data. And even in some data sets, one or more missing values can be found in up to 90% of genes [92]. The existence of missing values in microarray data can contribute from a variety of reasons including failures of hybridization, inclusion of artifacts, inadequate resolution, image noise and pollution, or spotting process derivatives [93]. In a worse case, the emergence of missing values can paralyse the usage of many analysis methods such as principal component analysis (PCA) and singular value decomposition (SVD). In a milder case, missing values adversely influence downstream analysis. The missing value issue has been reported to give unneglectable negative effect on some popular algorithms, such as hierarchical clustering and support vector machine (SVM) classifier.

To estimate the missing entries based on the incomplete gene expression data, a number of missing value imputation algorithms were proposed. The type of information used in the algorithm can be categorised into four different categories: 1) global approach 2) local approach 3) hybrid approach and 4) knowledge assisted approach.

2.3.2.1 Global approach

Algorithms in this category conduct missing value imputation according to global correlation information extracted from the entire data matrix. Widely used algorithms in this category include SVD imputation (SVDimpute) [94] and Bayesian principal component analysis (BPCA) [95]. In SVDimpute, a set of collaboratively orthogonal expression patterns (eigenvectors) are found via SVD, which can then be linearly combined to estimate the expression values of all genes within the data set. In BPCA, the expression of genes are similarly regarded as the combination of component vectors like the SVD approach (principal component in this context), i.e. $y = \sum_{i=1}^{K} w_l v_l + \epsilon$, where factor score w_l and residual error ϵ are assumed as normally distributed. An EM-like algorithm is then employed to approximate the posterior distributions of model parameters as well as missing values.

2.3.2.2 Local approach

Different from previous mentioned global approach, algorithms in this category use only local existing data with similar structure in the data set for imputation. The genes that are used to compute missing values is only those exhibits high correlation with the target gene with missing values. Algorithms like K nearest-neighbour imputation (KNNimpute) [94], least square imputation (LSimpute) [96], local least square imputation (LLSimpute) [97] belong to this category.

KNNimpute [94], is among the earliest and most well-known missing value imputation algorithms. KNNimpute exploits pairwise information between the target gene with missing value and the K reference genes which identified as nearest neighbours to impute the missing values. The missing value j in the target gene is approximated by the weighted average of the jth component of the K reference genes. The weights are defined proportional to the inverse of the Euclidean distance between the target and the reference genes. Several transformations to the original KNNimpute algorithm have been suggested [98, 99].

The concept of least square regression has also been widely adopted in many imputation algorithms. In least square imputation (LSimpute) [96], the K most correlated genes with the target gene are first selected by absolute Pearson correlation values. These reference genes x are assumed to be related with target gene y by a linear regression model $y = \alpha + \beta x + \epsilon$, by which a least square estimate of missing value is obtained. Lastly, the K estimates are combined linearly to produce the final estimate.

2.3.2.3 Hybrid approach

Depending on the nature of dataset to be either heterogeneous or homogenous, the favourable selection of imputation algorithm will be either a local approach or a global approach. Jornsten et al [100] proposes a hybrid approach named LinCmb that captures and utilizes both global and local correlation information in the data. The algorithm starts by obtaining estimates from five different imputation algorithms - SVDimpute, BPCA, KNNimpute, row average and GMCimpute – where the first two methods are global correlation based while the last three emphasize local correlations. The weighted average of the five estimates computed by the five algorithms are to be used as the final missing value estimate. To obtain the optimal weight set, LinCmb generates fake missing entries at the positions where true values are known and then uses the constituent methods to estimate the fake missing entries. The weights are then calculated by conducting a least square regression on the estimated fake missing entries.

2.3.2.4 Knowledge assisted approach

The idea of this category of imputation algorithms is the inclusion of domain knowledge or external information into the calculation process. The use of domain knowledge has the potential to remarkably enhance the imputation accuracy beyond what is achieved by purely data-driven approaches, particularly for data sets with small number of samples, much noise, or high missing rate. Algorithms in this category can exploit information such as the knowledge about the biological process in the microarray experiment [101], or the underlying biomolecular process as annotated in Gene Ontology (GO) [102] etc.

2.3.3 Normalization, filtering and averaging

Normalization is a type of transformation applied to the data that adjusts the individual hybridization intensities so that the data can align with a commonly agreed assumption as much as possible. Normalization is usually the first transformation the data undergoes. There are several reasons for the need to implement a normalization, including the unbalanced quantities of starting RNA, differences in labelling, detection efficiencies between different fluorescent dyes, and systematic biases.

2.3.3.1 Total intensity normalization

This method is based on the assumption that an approximately same number of labelled molecules from each sample should hybridize to the arrays, which therefore leads to the inference that the total sum of hybridization intensities in the arrays should be the same for each sample. Hence, a normalization factor is defined as

$$N_{total} = \frac{\sum_{i=1}^{N_{array}} R_i}{\sum_{i=1}^{N_{array}} G_i} \tag{2.8}$$

Where G_i and R_i are the measured intensities for gene i (for instance, the green and red intensities in a two-color microarray), and N_{array} is the total number of elements presented in the microarray. Using the normalization factor either one or both intensities can be scaled as follows

$$T_i = \frac{R_i}{G_i} = \frac{1}{N_{total}} \frac{R_i}{G_i} \tag{2.9}$$

Such that the mean ratio is equal to 1.

2.3.3.2 LOWESS normalization

Apart from a total intensity normalization mentioned above, a number of alternative methods are available, such as linear regression analysis [103], log centering, rank invariant methods [104], and Chen's ratio statistics [105]. However, none of these methods address the issue of systematic biases that the log2(ratio) values have a systematic dependence on intensity, which is reported in several studies [106, 107]. Locally weighted linear regression (LOWESS) [108] analysis has been proposed [106, 107] to remove such an effects in microarray data.

Model-based methods, such as neural networks and the mixture of Gaussians, use the data to build a parameterized model. After training, the model is used for predictions and the data are generally discarded. In contrast, 'memory-based' methods are non-parametric approaches that explicitly retain the training data, and use it each time a

prediction needs to be made. LOWESS, on the other hand, is a memory-based method that performs a regression around a point of interest using only training data that are 'local' to that point. In locally weighted regression, points are weighted by proximity to the current x in question using a kernel. A regression is then computed using the weighted points.

For clarification, a special ratio-intensity (RI) plot is introduced to expose and track the intensity-dependent effect, where the measured log2(Ri/Gi) for each element on the array as a function of the $log_{10}(R_i \times G_i)$ is plotted.

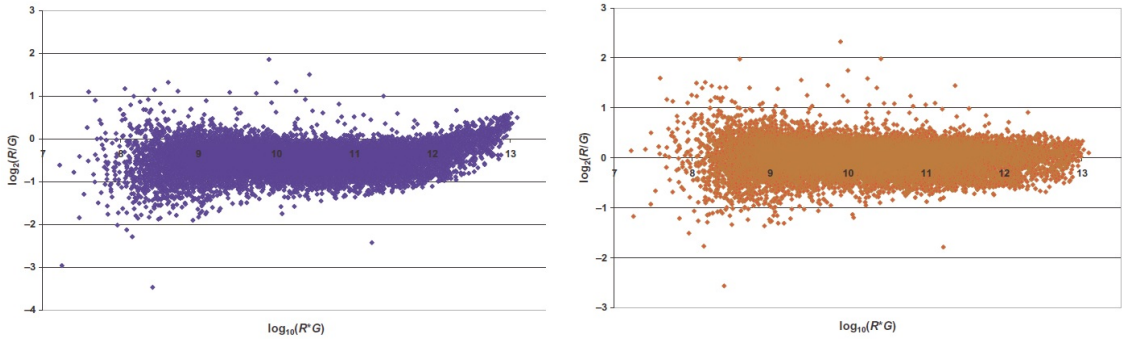


Image source: [109]

Figure 2.2: An R-I plot displays the $log_2(Ri/Gi)$ ratio for each element on the array as a function of the $log_{10}(Ri \times Gi)$ product intensities and can reveal systematic intensitydependent effects in the measured $log_2(ratio)$ values. Application of local lowess can correct for both systematic variation as a function of intensity and spatial variation on a DNA microarray. Figure on the left indicates the original data without normalization, figure on the right indicates the data normalized by LOWESS

2.3.3.3 Intensity-based filtering of array elements

Another kind of systematic bias appears when measured log2 ratio values increases as the measured hybridization intensity decreases. This is easy to understand because relative error increases at lower intensities where the signal approaches background. A widely used method to tackle this issue is to use only array elements with intensities that are remarkably different from background. Other methods include the use of absolute lower thresholds for acceptable array elements or percentage-based cut-offs where some fixed proportion of elements is abandoned.

2.3.3.4 Replicate filtering

Replication is necessary for finding and mitigating the variation in any experimental arrays, and there is no exception for microarrays. Biological replicates employs RNA independently obtained from different biological sources and carry out a measure to both of the natural biological variability in the system under study, or any random variation designed in sample preparation. Technical replicates offers information on the natural and systematic variability that appears in assay processing. A commonly used technical replication in two-color microarray array analysis is dye-reversal or flip-dye analysis [110], which comprises duplicating labelling and hybridization by swapping the fluorescent dyes used for each RNA sample.

In the ideal case, if a gene (presented in a pair of samples) does not have a preference over the dye they are attaching to, the following equation will be true

$$Log_2(T_{1i} \times T_{2i}) = log_2(\frac{A_{1i}}{B_{1i}} \times \frac{B_{2i}}{A_{2i}}) = 0 \tag{2.10}$$

Where A and B are two samples for gene i, and A_{1i} and A_{2i} is the expression of that gene in sample A before and after dye swap, likewise for B_{1i} and B_{2i}.

However in reality, experimental variation will make lead to a distribution of the log of the product ratios. So for this distribution, mean and standard distribution can be calculated, the results of which can help us make a decision on whether to keep the samples or not.

2.3.3.5 Averaging over replicates

To decrease the complexity of the data set, averaging over the replicate measures is a good choice. With the same replicate settings as mentioned before of two samples, A and B, we can adjust the value of both samples by

$$\overline{A} = \sqrt{A_{1i}A_{2i}}$$
$$\overline{B} = \sqrt{B_{1i}B_{2i}} \tag{2.11}$$

Where A and B are two samples for gene i, and A_{1i} and A_{2i} is the expression of that gene in sample A before and after dye swap, likewise for B_{1i} and B_{2i}.

2.4 Statistical learning

2.4.1 Statistical learning overview

Statistical learning, or specifically speaking, machine learning, is an adaptive process that enables computers to and learn by analogy, learn by example, learn from experience. In bioinformatics research, a number of machine learning approaches have been developed and applied to discover new meaningful knowledge from the biological databases, to analyse and predict diseases, to group similar genetic elements, and to find relationships or associations in biological data. Supervised learning is often used in the machine learning process in biomarker discovery. A general summary of machine learning in biomarker discovery is illustrated in the Figure 2.3. The workflow includes data acquisition, preprocessing of data, feature selection, classification and interpretation and validation of prediction model.

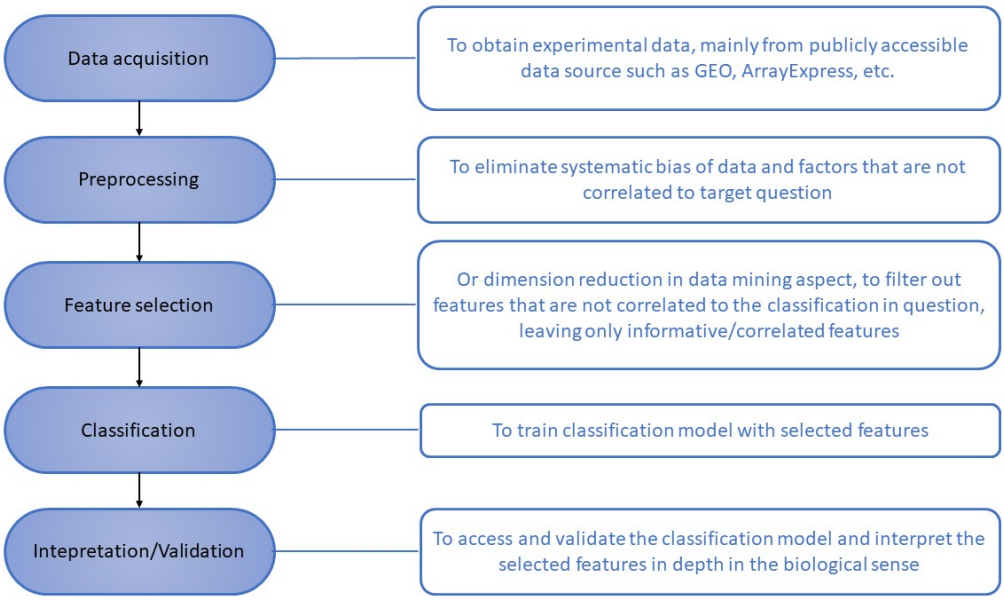


Figure 2.3: AD Biomarker discovery and validation project plan, which indicates the primary procedure for a full lifecyle of a study unit. It depicts the procedure from data acquisition, preprocessing of acquired data, feature selection with statistical approaches, to classification model training, assessing and result interpretation.

Generally, there are three sorts of machine learning algorithms; supervised learning, unsupervised learning and reinforcement learning. The difference between supervised learning and unsupervised learning is that the training samples of the former type of learning contains features values of items and associated "correct" class membership. For reinforcement learning, there will never be correct input or output presented, it is concerned with the problem that how a software agents could behave under a specific environment so as to gain the maximum predefined reward. Among these three types of learning, it is supervised learning that applied most frequently in the discovery of biomarkers, especially for the diagnosis of diseases such as cancer.

2.4.2 Supervised learning

Supervised learning is the machine learning task to obtain a solution (usually a function) from labelled training data. The training data is comprised of a set of training examples. In supervised learning, each example consists a pair of variables, one is the explanatory

variables and the other is signal variable. The set of explanatory variables is typically an input object vector and the signal variable is a desired output value that can be regarded as the label of a particular example. Through analysing the training data, a supervised learning algorithm produces an inferred function that can be used for mapping new sample, in other words, making prediction based on its input vector.

State-of-the-art supervised learning algorithms include artificial neural networks (ANN), k nearest neighbour (KNN), logistic regression, decision trees, linear or quadratic discriminant analysis (LDA/QDA), support vector machines (SVMs), kernel matching pursuit (KMP), logical analysis of data (LAD), stepwise discriminant analysis, partial least square projection, Naive Bayes, rule induction, and ensemble algorithms (e.g. boosting, bagging or random forest) combined with various base classifiers.

2.4.2.1 Decision tree

Decision trees are tree-like programs that classify instances by grouping them based on feature values. Each node in a decision tree is a feature to be classified, while each branch represents a value that is used as a measure to categorise an instance. Instances are sorted starting from the root node and classified according to their feature values. An example of a decision tree for the training set of Table 2.2 is shown in Figure 2.4.

Table 2.1: Comparisons between different supervised learning algorithms. Each column represents different supervised learning models, and each row represents a criteria to compare with between models.

	KNN[d]	Linear regression	Logistic regression	Decision trees	RF[e]	NN[f]	SVM[g]
Problem Type	Either	Regression	Classification	Either	Either	Either	Either
Intepretability[1]	Yes	Yes	Somewhat	Somewhat	A little	No	Somewhat
Explain difficulty[2]	Yes	Yes	Somewhat	Somewhat	No	No	No
Pred Acc[3]	Lower	Lower	Lower	Lower	Higher	Higher	Higher
Train Spd[4]	Fast	Fast	Fast	Fast	Slow	Slow	Slow
Pred Spd[5]	Depends on n	Fast	Fast	Fast	Moderate	Fast	Moderate
Param Amt[6]	Minimal	None[a]	None[a]	Some	Some	Lots	Some
SSS[7]	No	Yes	Yes	No	No	No	Yes
SNS[8]	No	No	No	No	Yes[c]	Yes	Yes
FI[9]	No	No	No	Yes	Yes	Yes	Yes
CPCM[10]	Yes	N/A	Yes	Possibly	Possibly	Possibly	Possibly
Parametric[11]	No	Yes	Yes	No	No	No	No
Scaling[12]	Yes	No[b]	No[b]	No	No	Yes	Yes
Algorithm	KNN[d]	Linear regression	Logistic regression	Decision trees	RF[e]	NN[f]	SVM[g]

[1] Is the results interpretable by you?
[2] Is the algorithm easy to be explained to others?
[3] Average predictive accuracy
[4] Training speed
[5] Prediction speed
[6] Amount of parameter tuning needed (excluding feature selection)
[7] Performs well with small number of observations?
[8] Handles lots of irrelevant features well (separates signal from noise)?
[9] Automatically learns feature interactions?
[10] Gives calibrated probabilities of class membership?
[11] Parametric?
[12] Features might need scaling?
[a] Excluding regularization
[b] Unless regularized
[c] Unless noise ratio is very high
[d] K-nearest neighbour
[e] Random forest
[f] Neural network
[g] Support vector machine

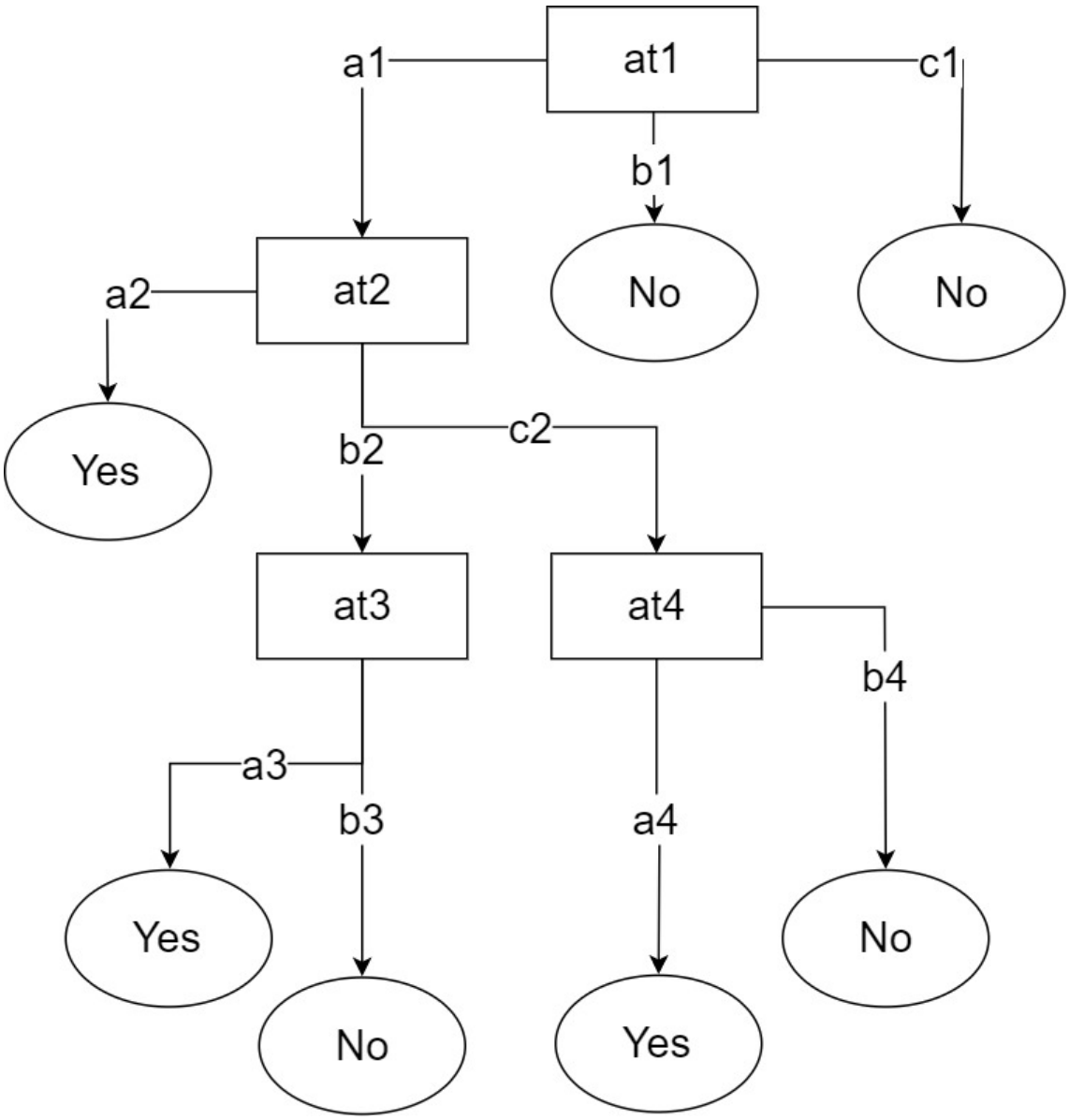


Figure 2.4: An example decision making procedure of a decision tree. For prediction, a new sample with unknown class label will go through the process of being classified into different nodes in the tree. Starting from node 'at1', where the attribute 'at1' of the sample's will determine which node is the next destination. A same iteration repeats until it reaches the 'leaves' which determine the prediction by 'Yes' or 'No' labels. Image source: [111]

For the example of the decision tree depicted in Figure 2.4, the instance $\langle$at1 = a1, at2 = b2, at3 = a3, at4 = b4$\rangle$ would sort to the nodes: at1, at2, and finally at3, which would classify the instance as being positive (represented by the values "Yes"). For the purpose of constructing optimal binary decision trees, theoreticians have endeavoured to

Table 2.2: An example of training set of decision tree. For each column, it represents an 'attribute' or 'feature', and the 'Class' column represents the class label which is necessary in all training data.

at1	at2	at3	at4	Class
a1	a2	a3	a4	Yes
a1	a2	a3	b4	Yes
a1	b2	a3	a4	Yes
a1	b2	b3	b4	No
a1	c2	ba3	a4	Yes
a1	c2	a3	b4	No
b1	b2	b3	b4	No
c1	b2	b3	b4	No

search for efficient heuristics of solving the problem.

The efficiency of decision tree requires that the root node of the tree should be the feature that best divides the training data set. Numerous methods were invented to the root node feature. Myopic measures such as information gain and gini index [112] estimate each attribute independently, whereas ReliefF algorithm [113] completes the estimation in the context of other attributes and their intercorrelations. Nonetheless, no single best method was found by a majority of studies to be universally robust [114]. Therefore, a selection of metric with comparison of individual methods is still an important step before constructing decision trees.

2.4.2.2 Random forest

Random forest is an algorithm for classification developed by Leo Breiman [115] that uses an ensemble of decision trees [111,112,116]. Each of the classification tress is built using a bootstrap sample of the data, and at each split the candidate set of variable is a random subset of variables. Thus, random forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners [111, 117], and random variable selection for tree building. Each tree is unpruned (grown fully), so as to obtain low-bias trees; at the same time, bagging and random variable selection result in low correlation of the individual tress. The algorithm yields an ensemble that can achieve both low bias and low variance (from averaging over a large ensemble of low-bias, high-variance but

low correlation trees).

2.4.2.3 Logistic regression and discriminant analysis

Discriminant analysis Fisher's linear discriminant analysis [118] is designed to find linear combinations, xw, of n-dimensional predictor variable values $x = (x_1, \ldots, x_n)$, to make large ratios of between-group to within-group sums of squares. For an $N \times (n+1)$ learning set data matrix, the ratio of between-group to within-group sums of squares is denoted by $w'Bw/w'Ww$, where B and W represent the $n \times n$ matrices of between-group and within-group sums of squares and cross products and w represent coefficients. The up-limit of $w'Bw/w'Ww$ can be inferred from performing eigendecomposition of $W^{-1}B$. Suppose matrix $W^{-1}B$ eigenvalues is denoted by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$, with corresponding linear independent eigenvectors $v_1, v_2, \ldots, v_s$. The discriminant variables are denoted as $\mu_l = xv_l, l = 1, \ldots, s$ on condition that $w = v_1$ maximizes $w'Bw/w'Ww$.

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are straight-forward methods in statistics or machine learning aiming to find the best linear combination of features to separate two or more classes of instances [119]. LDA specifies in working on observations with continuous quantites, while Discriminant Correspondence Analysis [120] is better dealing with categorical variables.

For a two-classes separating problem, LDA construct a hyperplane between the instances from two classes which can be described by a linear discriminant function $v_1x_1 + v_2x_2 + \ldots + v_nx_n + c$. The function output is equal to zero at the hyperplane. Meanwhile, two pre-conditions need to be fulfilled before training such a model:

- multivariate normal distribution in both datasets
- homogeneity of both covariance matrices

For discriminant analysis, the hyperplane is defined by the geometric means between the centroids of the two datasets. In normal practice, the variables are usually normalized first into standard means ($\mu = 0$) and variances ($\sigma^2 = 1$) and the Mahalanobis distance

(an ellipsoid distance determined from the covariance matrix of the dataset) is more favourable than the Euclidean distance [121].

Logistic regression The paradigm of logistic regression [122] is represented as

$$p(C = 1|x) = {}^{1}/_{[1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}]} \tag{2.12}$$

where x denotes an instance to be classified, and β_0, β_1, ..., β_n are the parameters of the model. These parameters should be estimated from the data so that a concrete model is acquired. The parameter estimation is completed in assist by the maximum likelihood estimation method. Since the system of $n+1$ equations and $n+1$ parameters to be solved does not have an analytic solution, the maximum likelihood estimations are performed iteratively. The Newton–Raphson procedure is a standard in this case. The modelling process is based on the Wald test and on the likelihood ratio test. The search in the space of models is usually carried out with forward, backward or stepwise approaches.

2.4.2.4 Neural networks and perceptrons

Neural networks as an artificial intelligence technique is inspired by the idea of mathematically modelling human intellectual abilities by means of biologically plausible engineering designs. The processing procedure from input to output is subdivided into several layers, in such a way that normally only units situating in two consecutive layers are binding to each other. By processing the input layer by layer, the output value which could be used as the criterion of prediction is calculated.

Single layered perceptrons The simplest neural network, called perceptron [123], deploys a single neutron to train classifier with a threshold activation function and separating two classes by a linear discrimination function.

A single layer perceptron can be described as follows:

Suppose x_1 through x_n are input feature values, and w_1 through w_n are connection

weights that are typically real numbers within the interval of -1 to 1. The perceptron first calculates the sum of weighted inputs $\sum_i x_i w_i$ and then the decision is made by comparing the sum with an adjustable threshold θ. If the sum is above θ then the output is 1, otherwise 0.

The most popular way that the perceptron algorithm is adopted to learn from a batch of training observations is to run the algorithm iteratively onto the training set until it manages to find a prediction vector which is fully correct on classifying all the training set. This prediction rule is then employed in the process of prediction of the test set.

WINNOW [124], whose name is derived from the fact that it had been designed for efficiency in separating relevant from irrelevant attributes, is among the first algorithm to be programmed based on perceptron idea that perform as a classifier model. It is designed to update the weights of each feature by either a *promotion parameter* α ($\alpha > 1$) or *demotion parameter* β ($0 < \beta < 1$). After the invention of WNNOW, a number of other algorithms have been developed such as those by Auer and Warmuth [125]. Later, a newer algorithm called *voted-perceptron* [126] improves the prediction performances on test set by including the information of prediction vectors after each mistake has been made.

Multilayered perceptrons Single layered perceptron can only classify linearly separable sets of instances, which otherwise will never classify all the instances properly if the data is not linearly separable. Multilayered perceptrons, also commonly known as Artificial Neural Network (ANN), have been developed to tackle the issue of linearly unseparable datasets [127].

A multi-layer neural network comprises numerous neurons concatenated together in a pattern of connections (Figure 2.5). Neurons are usually categorised into three types of layers: input layers, which receive information at the start of learning process; output layers, where the results of the processing are reported; and layers in between known as hidden layers. In particular, feed-forward ANNs (Figure 2.5) only allows data processing flow to travel one way only, from input to output.

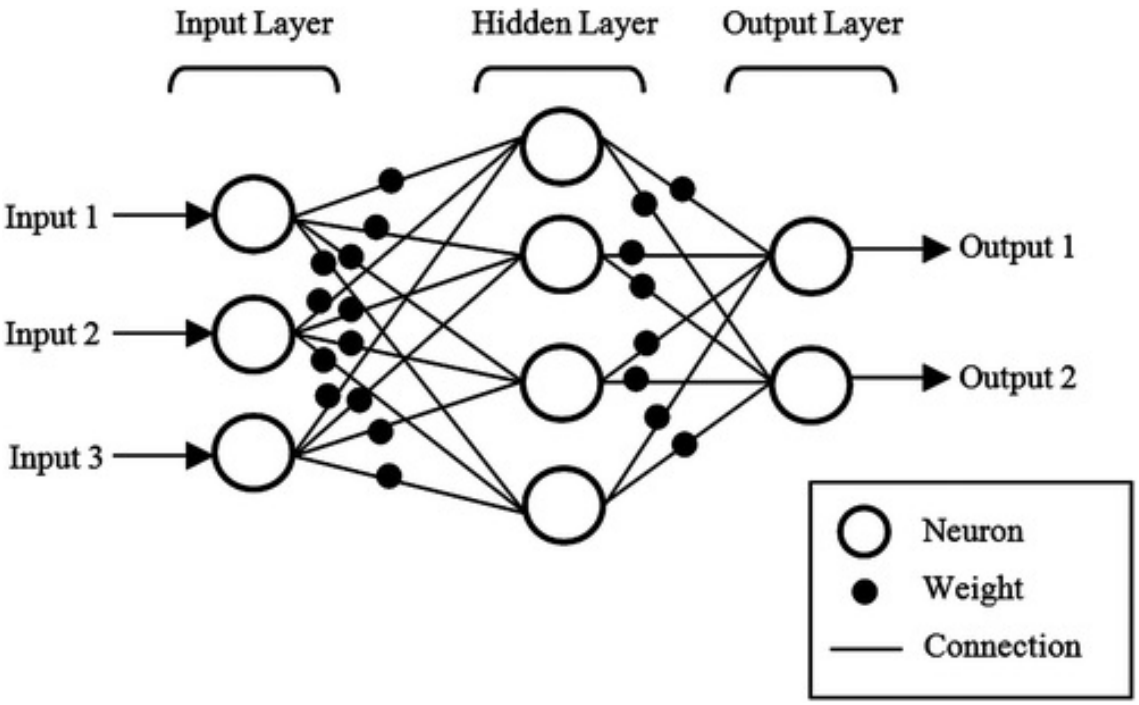


Figure 2.5: An artificial neural network is an interconnected group of nodes, similar to the vast network of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another. A weighted summation of input from top layers forms the output of the bottom layers, with the leftmost head as raw input and the rightmost bottom as output. Image source: [111]

During the process of classification, the signals propagates across the net from the input layer to the output layer. The activation values are initialized randomly in the input units which are then sent to each of the hidden units to which it is connected. Each of the activation values of these hidden units are computed and passed on to the output units. The activation values are all calculated according to a simple activation function. The function sums together the contributions from all connected units in the upstream, by multiplying the weight of the connection between the sending and receiving units (see Figure 2.5) and the sending unit's activation value.

The ANN model is trained by continually exposing the instances of the training set to the net. By comparing the desired output with the current output in the training process, all the weights in the net are modified slightly in the direction that would draw the output closer to the desired one. The network can be trained via several selection of

algorithms [128]. Nonetheless, the most widely used algorithm to compute and update the values of the weights is the Back Propagation (BP) algorithm.

In practice, the size of the hidden layer should be determined carefully, because undermining it can result in poor approximation and generalization ability of the model, while exaggerating it can lead to overfitting and complicate the search for the global optimum. A number of studies has excellent argument relating to this topic [129, 130].

A commonly known huge drawback of ANN is the speed of training. Since the BP algorithm requires a number of weight adjustments before acquiring a good weight configuration. One of the approaches for acceleration is to estimate optimal initial weights [131]. Another approach is weight-elimination algorithm that automatically derives the optimal topology with the benefit of avoiding overfitting [132]. Some mixture of other algorithms like genetic algorithms have also been used to train the weights [133] and to find the architecture [134] for the ANN model. Some researchers also attempt to train ANN with Bayesian methods [135].

2.4.2.5 Support Vector Machine (SVM)

The goal of support vector machines (SVM) [136] is to find the separating hyperplane with the largest margin (see Figure 2.6) which is defined by the positive distance between vector and the decision hyperplane. It originates from the reason that the larger the margin is, more confidence should be cast on the result of classification.

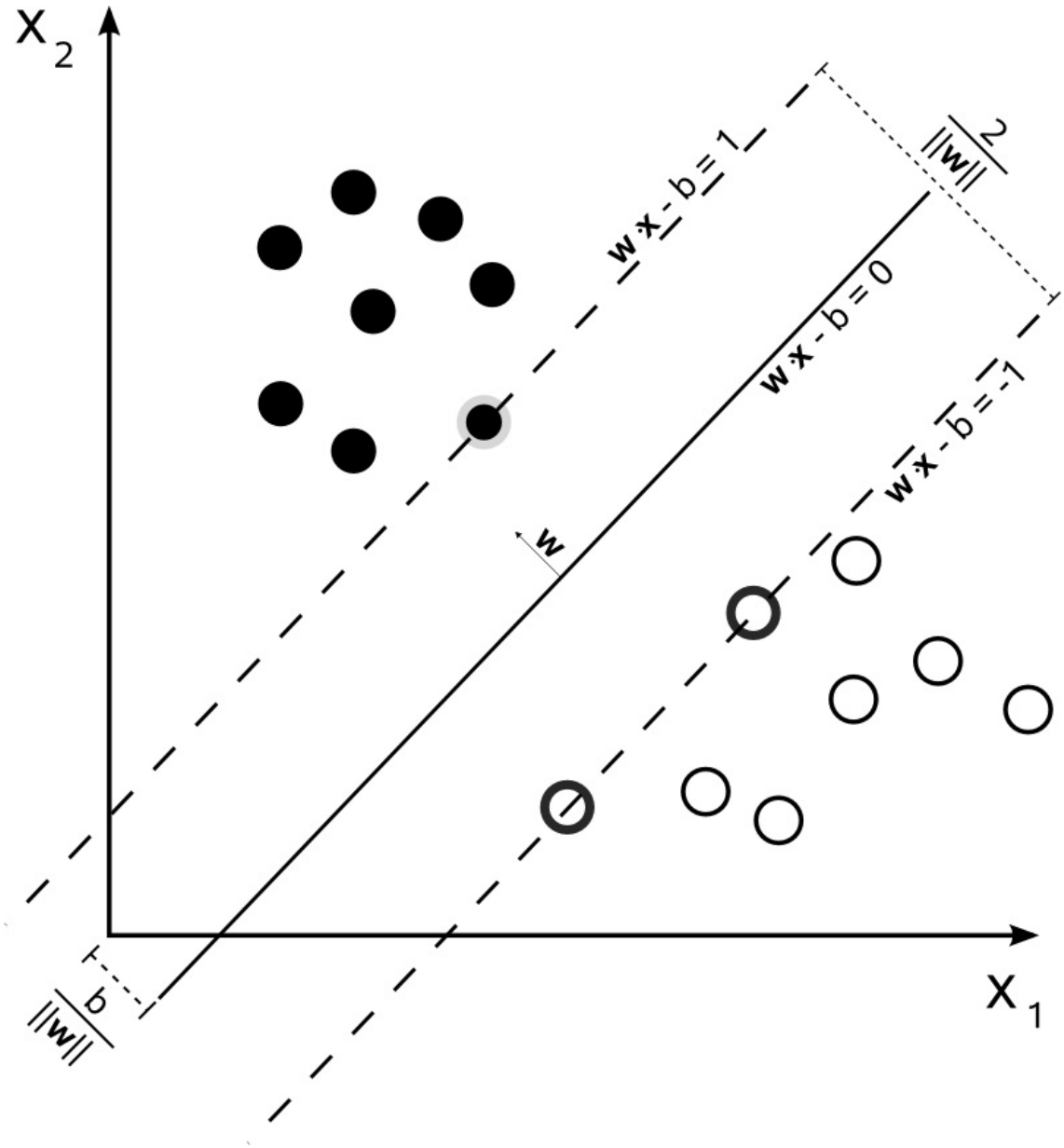


Figure 2.6: A demonstration of finding hyperplane with SVM model. In this example, two groups are separated by SVM, where the hyperplane (the diagonal line) separating two samples (black solid nodes and white unfilled nodes) with the maximum margin between the closest support vectors (highlighted with bolden edge of nodes which lie on dash lines parallel to the hyperplane). Image source: [111]

Suppose the training data is linearly separable and there exists a pair of (ω, b) such that

$$\begin{aligned} \omega^T x_i + b \geq 1, \quad \text{for all} \quad x_i \in P \\ \omega^T x_i + b \leq -1, \quad \text{for all} \quad x_i \in N \end{aligned} \tag{2.13}$$

So that the decision rule is defined as

$$f_{\omega,b}(x) = sgn(\omega^T + b) \tag{2.14}$$

Where ω is the weight vector, x_i is variables of sample i and b is the bias. Therefore, an optimum separating hyperlplane can be found by minimizing the squared norm of the separating hyperplane, which can be set up as a convex quadratic programming (QP) problem:

$$\begin{aligned} \text{Minimize} \quad \phi(\omega) = \frac{1}{2}\|\omega\|^2 \\ \text{Subject to} \quad y_i(\omega^T x_i + b) \geq 1, i = 1, \ldots, l \end{aligned} \tag{2.15}$$

Where y_i is the class label or dependent variables of sample i. For linearly separable data, the data points that lie on the margin of the optimum separating hyperplane is known as *support vectors*. The solution thus is represented as a linear combination of only these points.

However, for the cases that cannot be linearly separable, SVM may not be able to find any separating hyperplane at all. This issue can be addressed well by using a soft margin that accepts limited number of misclassifications on training instances [137]. The approach to realize it is to introduce positive slack variables ξ_i for each instance I = 1,. . . .,N in the constraints

$$\begin{aligned} \omega^T x_i + b \geq 1 - \xi, \quad \text{for} \quad y_i = +1 \quad \text{and} \quad \xi \geq 0 \\ \omega^T x_i + b \leq -1 + \xi, \quad \text{for} \quad y_i = -1 \quad \text{and} \quad \xi \geq 0 \end{aligned} \tag{2.16}$$

In this case, the Lagrangian is

$$L_p = \frac{1}{2}\|\omega\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\{y_i(x_i\omega - b) - 1 + \xi_i\} - \sum_i \mu_i\xi_i \tag{2.17}$$

Where the μ_i are the Lagrange multipliers introduced to enforce positivity of the ξ_i, and C is the parameter to reversely adjust margin size.

Another solution to address the case of inseparability of instances is to project the data onto a higher dimensional space where a new hyperplane is defined. With such a mapping to other spaces (possibly infinite), a linear separation in transformed feature space corresponds to a non-linear separation in the original input space. The application of *kernel function* is capable of achieving the goal. The *kernel function* is a special class of functions that allow inner products to be calculated directly in original input space, without actually performing the mapping or transformation into new features, the number of which is usually infinite [138].

It is very important to find such kernel functions that match the definition mentioned above. Genton [139] discussed several classes of kernels and their application in various domains, though not addressing the issue of suitability of functions to different type of problems. In the field of machine learning, commonly used kernels include the Gaussian and Laplace kernels, i.e.,

$$\begin{aligned} K(x, x') &= exp(-\frac{\|x - x'\|_2^2}{2\sigma_2}) \\ K(x, x') &= exp(-\frac{\|x - x'\|_1}{\sigma}) \end{aligned} \tag{2.18}$$

Recently, Muandet et al [140] provided a review of kernel mean embeddings of distributions and the new trend of research on kernel application extended to probability measures.

2.4.3 Unsupervised learning

Unsupervised learning targets on finding the hidden pattern or structure of the unlabelled data. Because the examples for the learner are without class labels or "responsive" variables, no error or reward signal is given to assess a potential solution. This is the major difference between supervised learning and unsupervised learning. Unsupervised learning is closely associated with the problem of density estimation in statistics [141]. Nonetheless, unsupervised learning also includes many other algorithms or techniques that endeavour to generate conclusions or explanations, or to explain principal features of the data.

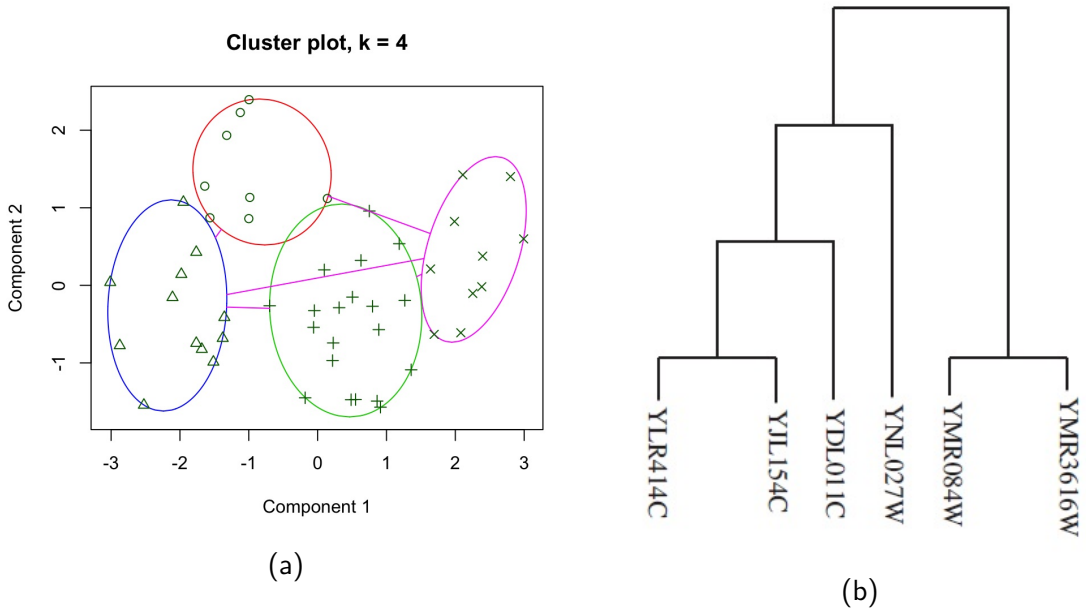
2.4.3.1 Clustering

Clustering is a technique partitioning a set of elements into subsets based on the differences between them. Especially, it is the process of grouping similar elements together. In clustering, no class information is provided and that is the main difference from supervised learning.

The clustering of genes in expression profiles is the most typical example of application of clustering in bioinformatics. As in microarray assays, the expression values for thousands of biomolecules in a limited number of samples are acquired. Extracted from these data, an interesting and intriguing piece of information is which molecules are coexpressed in the different samples. In practice, genes with similar expression level in all samples are amalgamated into one single cluster. Cluster analysis, also dubbed data segmentation, carries a variety of goals. By segmenting a collection of objects into clusters, those within each clusters are more closed associated to one another than those affiliated in other clusters. Sometimes the task of clustering is extended to arrange the clusters into a natural hierarchy, which involves iterative grouping of clusters to identify "clusters of clusters".

In all, the core idea of cluster analysis is to reveal the degree of similarity or dissimilarity between objects in interest.

Partition clustering Partition clustering specifies in obtaining a division of the data. Each point is classified into a unique cluster. Though commonly the number of clusters (see Figure 2.7 (a) k=4) is fixed, some algorithms are able to search for the optimal number of clusters while assigning objects to different clusters.



(a) (b)

Figure 2.7: Examples of clustering results from different approaches a) partition clustering - a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. b) hierarchical clustering - a set of nested clusters that are organized as a tree. Image source: [111]

One of the most popular iterative descent clustering methods is the *K-means algorithm* [142]. The objective of the algorithm is to group the data into K clusters such that the within-group sum of squares is minimized. A demonstration of the simplest form of K-means algorithm is by repeating the following two steps as depicted in Figure 2.8: 1) Assign objects to groups under the criterion that an object is assigned to the group with closest Euclidean distance 2) calculate new group means based on the previously updated assignments from step 1. The iteration ends when the class state of objects reach equilibrium.

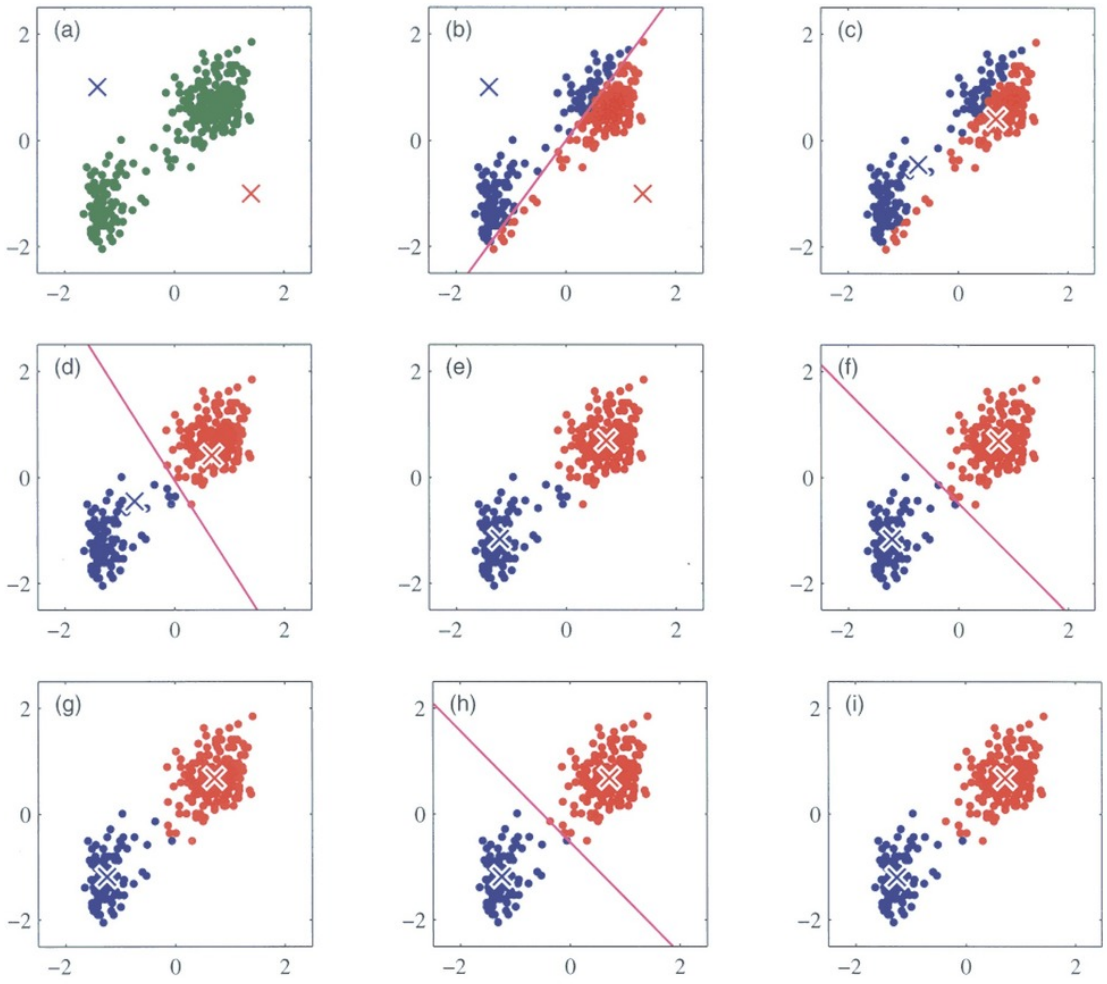


Figure 2.8: A visual procedure of applying k-means algorithm clustering approach. K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group. source: [111]

Hierarchical clustering Hierarchical clustering approaches [143] are the most popular methods to generalize data structures in the bioinformatics field. A hierarchical tree Figure 2.7 (b) is a nested set of partitions demonstrated by a tree diagram or dendrogram.

Strategies for hierarchical clustering are normally split into two types, "agglomerative" which is a "bottom up" approach that construct the tree from bottom to top by merging clusters along with higher hierarchy, or "divisive" which is a "top down" approach splitting the initial single cluster into groups. Generally, divisive algorithms are computationally inefficient. A measure of the distance or dissimilarity between two merged clusters determines which groups should be merged, or how the clusters should

be split. The matrix wrapping the dissimilarity information between pair of clusters is called dissimilarity matrix.

The following measures of distances between clusters are most commonly applied

- single-linkage: the distance between two groups is the distance between their closest members
- complete-linkage: the distance between the two farthest points
- Ward's hierarchical clustering method [144]: at each stage of the algorithm, the two groups that produce the smallest increase in the total within-group sum of squares are amalgamated
- centroid distance: defined as the distance between the cluster means or centroids
- median distance: distance between the medians of the clusters
- group average linkage: average of the dissimilarities between all pairs of individuals, one from each group

2.4.3.2 Principal component analysis (PCA)

Principal component analysis (PCA) is a multivariate technique that analyses a data matrix where instances are described by multiple inter-related quantitative dependent variables. It is likely to be the oldest multivariate technique. In fact, its origin can be traced back to over a century ago to Pearson [145] or even further [146]. Yet the term principal component was coined by Hotelling [147] who formalized its modern instantiation. Recently, Abdi and Williams [148] had a comprehensive review of the method.

The goals of PCA includes

- extracting the most significant and important information from the data matrix
- compressing the size of the data set by filtering out redundant information

- simplifying the description of the data set
- Analysing and decomposing the structure of the instances and the variables.

To realize these goals, new variables called principal components, which are formed by linear combinations of the original variables, are computed by PCA. It is required that the first principal component have the largest possible variance (i.e., inertia). As a result, this component will be able to explain or represent the most variance (the largest part of the inertia) of the data table. Under the constraint of 1) being orthogonal to the first component 2) to have the largest possible inertia, the second component is computed. The rest of components are computed likewise. Factor scores, which is the projections of the observations onto the principal components in a geometrical sense, are the values of these newly transformed variables for the observations.

Brief introduction of finding the components To find the components, singular value decomposition (SVD) of the data table X is executed.

$$X = P\Delta Q^T \tag{2.19}$$

where P and Q are the matrix of left and right singular vectors respectively, and Δ is the diagonal matrix of singular values. The matrix Q provides the coefficients of the linear combinations used to calculate the factors scores. Note that Δ is equal to the diagonal matrix of the (nonzero) eigenvalues of X^TX and XX^T. The factor scores denoted F is computed as

$$F = P\Delta = P\Delta Q^TQ = XQ \tag{2.20}$$

Interpreting PCA There are several aspects of information from the data table that can be extracted from the PCA results.

- The importance of a component can be obtained by evaluating its inertia, or more accurately, the proportion of the total inertia reflected or represented by this

component. The simplest way to get inertia is via the eigenvalue which is equal to the sum of the squared factor scores for a particular component.

- The contribution of an observation to a component can be obtained by the ratio of the squared factor score of this observation by the eigenvalue associated with that component.

- The *squared cosine* which indicates the importance of a component for a particular observation. It represents the contribution of a component to the squared distance of the observation to the origin.

- The correlation of a component and a variable is dubbed *loading* in PCA framework.

2.4.3.3 Genetic programming

Genetic programming (GP) is used as the feature selection algorithm. It performs better in classification with smaller amount of features compared to IG and REFS-F using NB, SVMs and J48. GP also outperforms NB and J48 as a classification method and has better performance in some cases, such as mining Mass Spectrometry data [149].

2.4.4 Probabilistic models

Different from logic-based algorithms like decision trees, perceptron-based algorithms like ANN or SVM, statistical learning approaches are characterized by mounting on an explicit probability model. These approaches delivers an output of probability that an instance belongs in each class, instead of simply a classification.

2.4.4.1 Bayesian machine learning

Bayesian learning is termed to represent the application of probability theory to learning from data [150]. The probability approach to modelling is conceptually very simple compared to other machine learning paradigms. First, probability distributions are used

to describe all the uncertain unobserved quantities within the model as well as their relationship to the data itself. Then the basic probability rules are applied to infer the unobserved data from the observed one. The behaviour of learning happens when the prior probability distribution (defined before observing the data) is transformed into posterior distributions (after observing the data). The Bayes rule can be generally summarised as

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y \in Y} P(x,y)} \tag{2.21}$$

Where $P(x)$ and $P(y)$ are the unconditional probability, $P(x|y)$ is the conditional probability of x given y, $P(x,y)$ is the probability that both x and y occur. Some replacement can be made to describe its application to machine learning

$$P(\theta|D,m) = \frac{P(D|\theta,m)P(\theta|m)}{P(D|m)} \tag{2.22}$$

Where D denotes the observed data, θ denotes the unknown parameters of a model, conditioning all terms on m which is the class of models in interest, $P(D|\theta,m)$ is the likelihood of parameters θ in model m, $P(\theta|m)$ is the prior probability and $P(\theta|D,m)$ is the posterior probability.

After the learning occurs and the prior knowledge is transformed into the posterior knowledge about the parameters, this posterior is now the prior to be used in the future data.

Naive Bayes classifier Naive Bayes (NB) is a very simple case of Bayesian networks which comprises directed acyclic graphs with only one single unobserved node and several observed nodes. A strong assumption is raised forward that the child nodes are strictly independent to one another in the context of the unobserved node [151]. Therefore, the model is characterized to estimate:

$$R = \frac{P(i|X)}{P(j|X)} = \frac{P(i)P(X|i)}{P(j)P(X|j)} = \frac{P(i)\prod P(X_r|i)}{P(j)\prod P(X_r|j)} \tag{2.23}$$

By comparing the two probabilities, the classifier makes a decision of predicting the

class label to be either i (if R $>$ 1) or j. Cestnik et al [152] were among the first to use BN in machine learning community. Further study [153] suggested using Laplace estimators or adding m-esimate to avoid the impact from a probability of 0 to the computation of probabilities $P(X, i)$.

While having the advantage of a short computational time, the biggest problem that the algorithm facing, is the strong assumption of independence among child nodes which is seldom the case. As a result, NB classifiers are usually less accurate than most other more sophisticated algorithms (such as ANNs). Many enhancement has been made on the original model in order to address the issue. Friedman et al [154] proposed some dependencies between features with the limitation that each feature can be related to only one other feature. Another attempt is the proposal of semi-naive Bayesian classifier [155] where attributes are further partitioned into independent groups.

2.4.4.2 Gaussian processes

Gaussian processes are a very flexible non-parametric model for unknown functions, which are commonly used in regression, classification, and other applications that require inference on functions [156].

2.4.5 Model evaluation and accessing methods

Once a classification model is trained and constructed, we need to assess its performance in regards to prediction accuracies. Some conventional measures are developed for this purpose.

2.4.5.1 Error rate and ROC curve

The error calculations are based on the confusion matrix (Table 2.3). Each cell in the matrix represents the number of samples fallen into that particular category. For example, the value of TP represents the number of instances whose true class is positive and predicted by the classification model to be positive. Therefore, the total number of

Table 2.3: Confusion matrix in a two classes problem. The column represents the prediction to be positive or negative, whereas the row represents the true label being positive or negative. The value in each cell represents the number of samples falling into the category, eg. TP means the number of positive samples that are predicted to be positive, FN means the number of positive samples that are predicted to be negative, etc.

		Predicted class	
		Positive	Negative
True class	Positive	TP: True positive	FN: False negative
	Negative	FP: False postive	TN: True negative

instances in the validation set is

$$N = TP + FP + TN + FN \tag{2.24}$$

Based on the confusion matrix, the following measures are defined:

$$\begin{aligned}
Sensitivity/Truepositiverate(TPR) &= \frac{TP}{TP+FN} \\
Specificity/Truenegativerate(TNR) &= \frac{TN}{TN+FP} \\
Precision/Positivepredictivevalue(PPV) &= \frac{TP}{TP+FP} \\
Negativepredictivevalue(NPV) &= \frac{TN}{TN+FN} \\
Missrate/Falsenegativerate(FNR) &= \frac{FN}{FN+TP} \\
Fall-out/Falsepositiverate(FPR) &= \frac{FP}{FP+TN} \\
Falsediscoveryrate(FDR) &= \frac{FP}{FP+TP} \\
Falseomissionrate(FOR) &= \frac{FN}{FN+TN} \\
Accuracy(ACC) &= \frac{TP+TN}{TP+TN+FP+FN}
\end{aligned} \tag{2.25}$$

To fine-tune a classifier, another method is to plot the receiver operating characteristics (ROCs) curve [157], which reveals hit rate versus false alarm rate, namely, $1 - specificity = FP/(FP+TN)$ versus $sensitivity = TP/(TP+FN)$, and has a form similar to Figure 2.9.

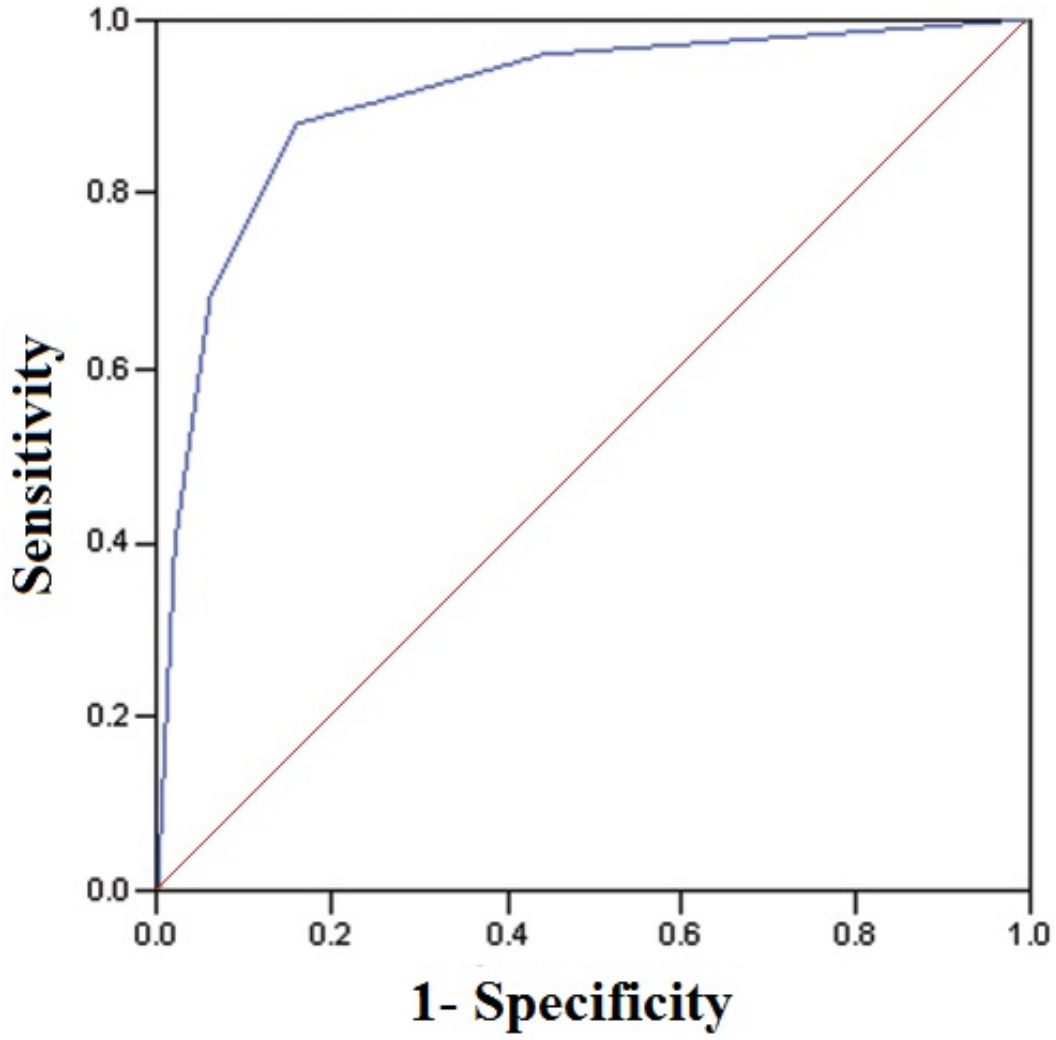


Figure 2.9: An example of ROC curve. The true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decisiont hreshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal). Image source: [111]

Each classification algorithm has a parameter, or in other words threshold of decision, which subject to changes so that classifying decisions can be made and consequently yielding different pairs of true positives versus false positives. Normally increasing the number of TP will also increase FP, and vice versa. As a result, a trade-off is often required to achieve the application aim to produce predict result that is within the level of error expectation, and the trade-off can be visualised as a point on this curve. The area

under the ROC curve is commonly employed as a measure of performance for machine learning models [158].

2.4.5.2 Estimating the classification error

An unneglectable issue relating to a designed classifier is how to calculate its error rate when it is applied to classify new instances that has not been used in the training process.

A popular method to estimate the error is by k-fold cross-validation [159]. To start with, all of the samples are partitioned into k folds where each fold is left out of the training process of the classifier and used as "stranger" dataset in the classifier testing process. The estimated error is the overall proportion of errors across all folds. A special case for this method occurs when k equals the total number of samples, which is dubbed as leave-one-out cross validation, where a single instance is left out each time as the testing dataset.

Another popular method for error estimation is the bootstrap methodology [160]. For this method, a resampling strategy is implemented based on the notion of "empirical distribution" where the bootstrapped samples are drawn with replacement by equal probability. The bootstrap zero estimator and the 0.632 bootstrap estimator [161] are used.

In bioinformatics, an overview was provided by Baldi et al. [162] of various methods to evaluate the accuracy of classification algorithms. The implementation of the previous error estimation methods has mostly focused on the analysis of supervised models specific for microarray data. In this sense, Michiels et al. [163] utilize multiple random sets for predicting cancer outcome by microarrays. Ambroise et al. [164] reported a 10-fold cross validation outperforms a leave-one-out cross validation, in a gene selection issue built upon microarray gene profiles. In regards to the bootstrap approach, the so-called 0.632 bootstrap error estimator proposed to tackle overfitted prediction is suggested. It has also been reported [165] that the 0.632 bootstrap estimator as a standout estimator in small-sample microarray classification. These same authors [166] have then developed a novel method called bolstered error estimation which bypass the performance of bootstrap in

feature ranking performance. Fu et al [167] proposed a method combining bootstrap and cross-validation which show very good results.

2.4.6 Meta-analysis

Different from machine supervised machine learning which focuses on prediction, meta-analysis comprises statistical methods for contrasting and combining results from different studies, in the hope of identifying patterns among different study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. The information extraction from accumulating mass of microarray data by meta-analysis can reconstruct cellular pathways or transcription factor network, leading to a more reliable understanding of the underlying biology. In the first application of meta-analysis to microarray, Rhodes et al [168] combined four data sets on prostate cancer to determine genes that are differentially expressed between benign prostate tissue and clinically localized prostate cancer. Current R packages to conduct meta-analysis for microarray studies include meta [169], metafor [170], metaMA [171], MAMA (depreciated), and scores of other packages, each of them is designed for specific type of input data and output data, with various objectives and algorithms in data processing.

2.5 Biomarker discovery by machine learning

2.5.1 Feature selection methods

Feature selection is extremely important for the classification that involves large numbers of features in the dataset. The need of feature selection is especially urgent for some cases, for example, microarray data based classification, where numerous noisy and redundant genes are presented in the dataset. Also, despite the small sample size retrieved from microarray data, it represents the state of a cell at a molecular level and plays gradually important roles in medical diagnosis. Therefore, many techniques are

developed and employed to work out how to select features efficiently and accurately. The technique for feature selection can be categorized into three main methods – filter methods, wrapper methods and embedded methods. Filter method selects only the highest ranking features based on general characteristics of the training data without influencing any learning algorithm. Moreover, it requires comparatively short computing time than other feature selection methods. Several kinds of filter methods are frequently used in microarray data to identify informative genes, such as Bayesian Network Information Gain (IG), Signal-to-Radio (SNR) and Euclidean Distance, among which IG has been reported to be the superior technique. Except for parametric techniques, some non-parametric techniques, for example, threshold number of misclassification, Pearson correlation coefficient and Significant Analysis of Microarray (SAM), are also applied for feature selection. Univariate filter method satisfies the need for validation of experimental results from biology and molecular domain experts and consumes less computational time. Nonetheless, the major drawback is the selected genes are most possibly redundant. Unlike the filter method which sort out meaningful genes independently, a wrapper method unifies the feature selection process with a classification algorithm. The learning algorithm which the model uses to train itself has great impact on the result of the feature selection process. Consequently, wrapper approach acquires more accuracy at the expense of computational cost. Genetic Algorithm (GA) is a randomized search algorithm that has been utilized for binary and multiclass cancer discrimination [172, 173]. The prime disadvantage of GA is that it has a higher risk of over-fitting than filter methods and is very computationally intensive. In terms of the specificity to learning algorithm, embedded methods are identical to wrapper methods. But a progressive advantage of this technique lies in the fact that it is more computationally simple than wrapper methods. Support Vector Machine (SVM) method of Recursive Feature Elimination (RFE) is one of those embedded methods that are used for gene selection.

2.5.2 Machine learning in practice

A common case for microarray data is that there exist a number of unannotated datasets that cannot contribute to the supervised classification using microarray data. Likewise, unsupervised learning algorithm cannot fully take advantage of the annotated datasets. It is reported that semi-supervised algorithms can address these situations. Blum and Mitchell [174] first introduced a co-training algorithm which enhanced the performance of classifier when there are both labelled and unlabelled samples. Another SSL algorithm is proposed [175] where the example data is comprised of DNA microarray expressions and phylogenic reconstructions. Lately, a Bayesian Semi-Supervised approach coined BGEN (Bayesian GENeralization) [176] was introduced whose predictions were proved to be more accurate than that from either K-means clustering or SVM classification. Harris and Ghaffari constructed a new SSL method classifying three cancer groups with both labelled and unlabelled data [177]. The new method shows improvement in the accuracy comparing to classifying solely with labelled data.

Extra-Trees Decision Trees is the learning algorithm which was used in a recent study [178] to define protein profiles of asthma-related inflammation and remodelling by ranking the potential biomarkers according to their importance. In that study, surface-enhanced laser desorption/ionization-time of flight-mass spectrometry (SELDI-TOF-MS) on lung samples is employed as input for machine learning and the discovery of several new key biomarkers that are likely to participate in the formation of asthmatic phenotype suggests that SELDI-TOF-MS is a useful tool to identify new potential therapeutic targets in inflammatory diseases.

Unlike standard approaches to estimate variable importance that contributed solely from the estimation of the conditional expectation of the outcome provided the biomarker and covariates, the approach of targeted maximum-likelihood estimation target the regression estimate specifically at the parameter of interest. In the case of a reported study about biomarker discovery of treatment of antiretroviral-resistant HIV infection [179], several ML algorithms are used to seek specific mutations which can reduce clinical

virologic response to antiretroviral regimens. It turns out that better performance if acquired by means of a targeted maximum-likelihood estimation compared to a standard approach.

It was proved that classifiers trained by combinatory algorithms have good performance in some studies. Wang et al [180] conducted a study which investigates the gene expression profiles in zebrafish treated with endocrine-disrupting chemicals of different modes of action. After testifying various combinations of gene feature selection/class prediction algorithms, it turns out that a classifier that is yielded by genetic algorithm and SVM is the best in terms of accuracy.

Among all of the developed supervised learning algorithms, support vector machine (SVM) and random forest (RF) exhibit promising classifying ability, stability in the expense of comparatively low computational cost, which make them two of the most popular algorithms applied in the studies of disease classification, prediction and biomarker discovery.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. After training, new examples are mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Random forest is an algorithm for classification is an ensemble learning method for classification (and regression) operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. There are many successful applications of SVM and random forest in disease diagnosis [181]. Guyon et al [182] utilized SVM as classifier training model to discover two genes that yield zero leave-one-out error in a leukemia cohort, and found out four genes with 98% accuracy in a colon cancer cohort. Furey et al [183] presented a method to analyse microarray expression data using SVM, with comparable performance in compared with another method which is based on perceptron algorithm in ovarian cancer datasets. Akay F. [184] built a SVM classifier for Wisconsin breast cancer dataset (WBCD) by only 5 features with as high as 99.51% classification accuracy. D-Uriarte R et al [185]

proposed a gene selection method based on random forest, with which a prediction model is trained with very small sets of genes while preserving predictive accuracy. Random forests are sets of ensemble learning methods for classification combining bagging idea and the random selection of features so as to build a collection of decision tress and outputting the class that is the mode of the classes output by individual trees. In Zhang's study [186] about identifying CSF biomarkers that distinguish Alzheimer (AD) and Parkinson Diseases (PD), random forest was used as the statistical analysis method to analyse the data, which validated a proteomics-discovered multianalyte profile (MAP) in CSF that is highly effective in classifying PD and moderately effective at identifying AD.

2.6 Big data technologies

We have entered an information driven era. The trend led by advancing hardware technologies and the rampage of internet in modern world has brought new scope of technology revolution in the broad aspect of information technology. The maturation of technology brings explosive boost of data volume. The data is being yielded at an exponential speed through various disparate potential resources and sensors, scientific instruments, and internet especially the social media, etc. The unprecedented outburst of information and virtual fortune challenged both the infrastructure and the paradigm of acquiring, storing and analysing data. Consequently, the large, complex, structured or unstructured, and unharmonized data has attracted significant attention.

Under the enormous focus, some modern data models that claim to process the extremely large data such as Big Data in a reliable and efficient way has come into horizon. NoSQL, largely being translated as "not only SQL", is a fashionable database "genre" that intends to be non-relational, highly scalable, efficiently distributed, and mostly open-source. They are often labelled as schema-free, straightforward API, simple replication, etc. Numerous potential data models are developed in the past few years but still the warehouse is growing to meet more challenges. The main mentality of

evolving NoSQL database capacity is very clear. The massive, complex, and fast-pace data production requires more efficient query (more queries) development that retrieves most accurate information (better results) with best optimal latency [187].

2.6.1 Data challenges

The mammoth growth of the modern data casts doubts on the computing capacity and efficiency of the existing database prototypes. In general, the concern that Big Data era is facing can be described to have four aspects of the data: volume, velocity, variety and veracity. All of the four concerns are briefly described here

- *Volume of data* The industries are inundated with vast volume of data of all types. The data is enlarging easily from terabytes to even petabytes. Tweets from the social media network alone consume up to 12 terabytes of data every day.

- *Velocity of data* The data collected at the enterprises is at an extremely high speed. A slight delay can make a huge impact on the analysed output. For the example of fraud detection, the 5 million trade transactions influx of data every day must be analysed instantly in order to identify the potential frauds.

- *Variety of the data* All of the possible types of data will be emerging and needed to be analysed – text, audio, video, log file, etc. A combinatory study of different types of data is also of high chance to report valuable information to the enterprises.

- *Veracity of the data* Data is not always trustworthy. It is especially crucial to have true output from input full of suspicion the decision making is based on such information.

2.6.2 NoSQL data models

The following section discusses some of the best database models that are considered to be reliable and efficient in the Big Data arena.

2.6.2.1 BigTable

BigTable, exclusively exploited by Google for managing various important projects, is designed as a distributed storage system to manage the petabytes size of the data, distributed across several thousand commodity servers and allows further horizontal scaling. The BigTable copes with latency and data size issues with high efficiency. Underpinned by the Google File System (GFS) [188] for log and data files storage and Google Sorted String Table (SSTable), it is a compressed, proprietary with high performance storage system.

The BigTables are intensively used by Google for its different applications such as Gmail [189], YouTube [190], Google Maps [191], etc.

2.6.2.2 Cassandra

Originally invented by Lakshman and Malik in 2011 [192], Cassandra is a distributed, open source data engine that deal with a gigantic amount of data of petabytes level, distributed across several commodity servers. The infrastructure selection of data clusters spanning multiple data centres guarantee the reliable support, high data availability and no single point of failure by Cassandra. The durable write function supported by memtable, which is a special memory space in the infrastructure of Cassandra acting as an intermediate station whenever a write operation occurs, help avoid the risk of losing data during hardware failure. Though being a member of NoSQL family, Cassandra does not support the joins and subqueries but is equipped with SQL-like languages called Cassandra Query Language (CQL).

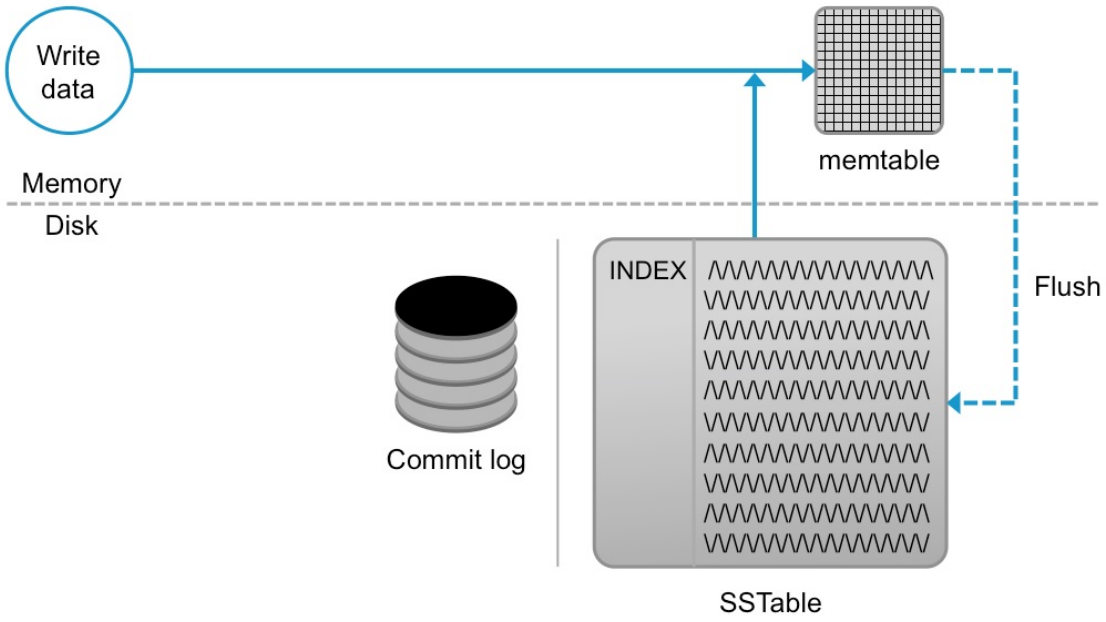


Figure 2.10: Writing path in Cassandra.Cassandra processes data at several stages on the write path, a) Logging data in the commit log b) Writing data to the memtable c) Flushing data from the memtable d) Storing data on disk in SSTables. Image source: [187]

2.6.2.3 HBase

HBase is a non-relational, column-oriented, open source, and distributed database management system [193] written in Java. HBase employs Hadoop Distributed File System (HDFS) underneath. It is envisioned to manage massive sparse datasets and follows the master-slave concept which is a model of communication between devices where one has unidirectional control over other devices. Though being a member of NoSQL family, HBase is a column-oriented data model that consists of the sets of tables, consisting of rows and columns. Like traditional databases, HBase assigns a primary key in each table that is used to access the data. HBase provides the function of grouping many columns (attributes) together into what are called column families that is open to join in anytime, with elements of a column family stored together.

2.6.2.4 MongoDB

MongoDB, etymologically derived from the word "humongous", is an open source, document-oriented NoSQL database that has lately gaining popularity in the data industry [194]. It is known as one of the most widely used NoSQL databases that inherit the master-slave replication. The responsibility of master is to implement the operations of reads and writes, whereas the role of slave is to duplicate the data fetched from master, to execute the read operation, and data backup. The slaves do not participate in write operations, but may select an alternate master in case of the current master failure. MongoDB uses binary encoded format of JSON documents called BSON behind the scenes. As a member of NoSQL family, no fixed schemas is required in contrast to the traditional relational databases. Numerous search refinement options like searching by fields, range queries, regular expression, or even the user-defined complex JavaScript functions are allowed in the query system of MongoDB.

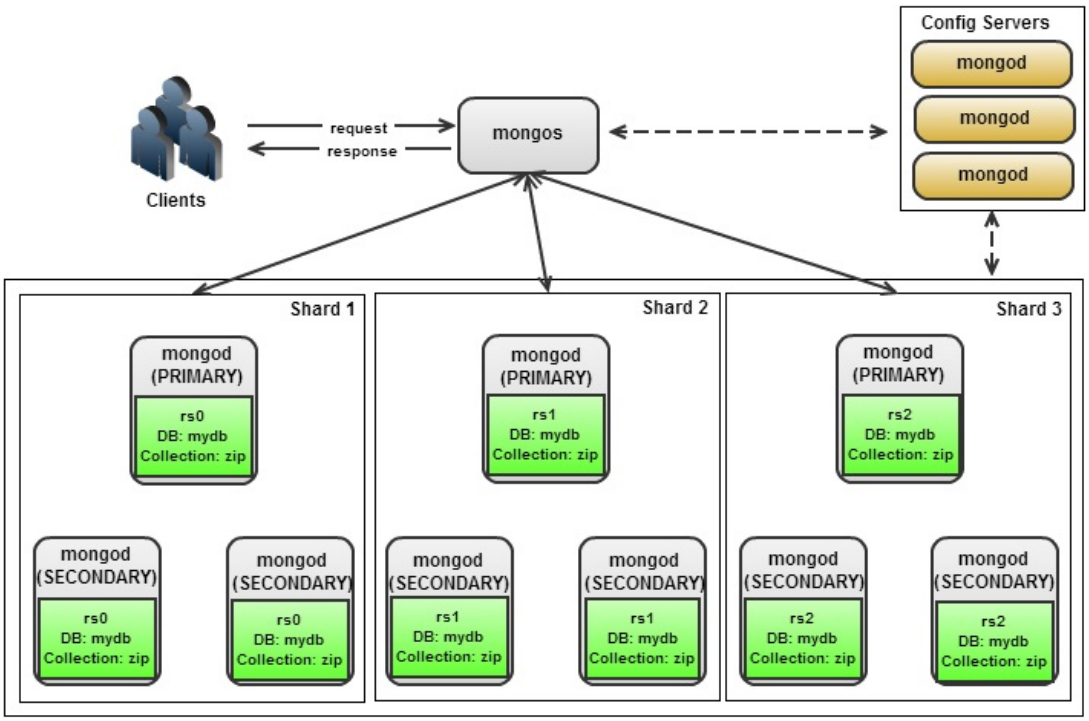


Figure 2.11: MongoDB architecture. MongoDB is a document database that provides high performance, high availability, and easy scalability. Some of its characteristic includes sharding, replication, the concept of collections and documents, etc. MongoDB has flexible schema, document structure, and atomicity of write operations. Image source: `https://chiruideas.weebly.com/mongo-db.html`

2.7 Methods in plan

2.7.1 Statistical approaches

In this thesis, the following statistical approaches are employed to achieve the research goals,

Random Forest Firstly, it is an ensembling-learning model which aggregate multiple machine learning models (decision tree in this case) to improve performance. Thus, a good combination of low variance and low bias can be expected from this model. Secondly, Random Forest is able to intepret classifying performance on a feature-by-feature basis, avoiding the mystification of feature performance explanation that appears in other models such as SVM.

Support Vector Machine Firstly, it has a regularisation parameter, which helps avoiding over-fitting to a manipulatable extent. Secondly, the fact that it uses kernel facilitates expert knowledge about the problem to be built in via engineering the kernel. Thirdly, an SVM is defined by a convex optimisation problem (no local minima) for which there are efficient methods. Lastly, it is an approximation to a bound on the test error rate.

Meta-analysis This method provides approaches to summarize and integrate results from multiple studies, and thus more general conclusions can be made by this analysis as it escape from the bottleneck of lack of coorperation between difference studies in terms of drawing conclusions.

2.7.2 Technology and software

And the following technology or software are used to serve for the research purposes,

R Entirely open source, R has the most comprehensive statistical analysis package available by far. There are large number of communities that strive to develop the various aspects of this tool. The graphical representation of R is extremely exemplary which makes it surpass other statistical and graphical packages.

Python is chosen among the rest of programming languages for the following reasons.

- Python is a programming language that is much easier to build prototypes compared to other languages, due to its simplicity and flexibility.
- Abundant Python libraries for data science such as Numpy, Scipy, Pandas, etc., and sufficient community support
- On developing extensive large scale applications, Python and Django in combination offer speed of development, flexibility, scalability and robust applications. The technology stack helps develop cost-effective application

MongoDB Like other NoSQL databases, MongoDB are more scalable and structurally schema-free compared to traditional relaional databases. Apart from these legacy advantages, MongoDB provides document oriented storage, replication, auto-sharding, and indexing on any attributes. These features are greatly beneficial for database performance in production, and more than adequate to serve a research application.

Ingenuity Pathway Analysis IPA is excellent in uncovering the significance of Omics data and identify new targets or candidate biomarkers within the context of biological systems. It has broadly been adopted by the life science research community and is cited in thousands of articles for the analysis, integration, and interpretation of data derived from Omics experiments, such as RNA-seq, small RNA-seq, microarrays including miRNA and SNP, metabolomics, proteomics, and small scale experiments.

Chapter 3

Prominent AD classifier enhanced by addition of covariate information

3.1 Abstract

In this chapter, transcriptomic AD biomarkers in brain are investigated. First, the impacts of covariate such as age, gender and APOE genotype information to the performances of AD classifiers were studied, with a result showing that the inclusion of age and APOE genotyping into the prediction model could improve classifiers performance. Then, a random forest based feature selection method was employed to select a panel of 26 RNA transcripts that has promising classification ability between AD patients and healthy controls.

3.2 Background

AD prognosis and diagnosis has always been hard. Like other neuro-degenerative diseases, the awareness of progression of the disease rely heavily on symptoms like long or short term memory loss, ambiguity, mood swing, etc.

In this study, the impacts of covariate to the performances of AD classifiers are evaluated, by comparing support vector machine (SVM) models trained with and without

covariates.

The interest of including age in the prediction model is obvious, since aging is closely linked with dementia. Cognitive impairment appears both in the process of normal aging and neurodegenerative diseases (especially AD). Major studies has been targeting on trying to distinguish cognitive impairment attributable to normal aging from those that suggest AD pathology [195, 196]. Age correction is a common approach to address the issue before analysing the experiment data. Recently studies show that AD first symptoms are age dependent [197], with disparity in the first cognitive and behavioural symptom experienced by AD patients. To understand the effect of age into the studies, several groups of researchers designed and examined different age correction methods for imaging data such as MRI [198, 199], sequencing data such as SNP [200].

The presence of an $\epsilon 4$ APOE allele was proved to be significantly associated with AD [201,202]. It is reported to be the strongest known susceptibility variant for AD [203–205] and this finding has been confirmed by genome-wide association studies [206,207]. There are three major isoforms of the APOE protein that are encoded by three alleles of the APOE gene ($\epsilon 2$, $\epsilon 3$ and $\epsilon 4$). As a result, APOE genotyping information has excellent potential to be one of the indicator variables for discriminating case and control samples.

Another factor to be investigated is gender. As per the report of US Institute of Medicine in 2012, "sex" refers to the classification of living things as male or female according to their reproductive organs and functions assigned by chromosomal complement, whereas "gender" refers to a person's self-representation as male or female or to how that person is responded to by social institutions on the basis of that presentation [208]. To avoid confusion, "gender" here refers to "sex" in the above definition. Though not proved to be a direct risk factor of AD, gender are reported to interact with other biological factors and gender-related factors. For example, majority of studies and a large meta-analysis [204] showed higher age-specific odds ratios of AD in women compared with men both for carriers of one $\epsilon 4$ allele and for carriers of two $\epsilon 4$ alleles. Other studies also showed that female $\epsilon 4$ allele carriers had greater hippocampal atrophy, more changes in the default mode connectivity, more cortical atrophy and worse memory

performance compared with men [209–211]. On the other hand, gender-related factors such as education, physical activity and occupation are also unneglectable. These all sum up to a plausible assumption that gender is also likely to be a risk factor, with more complicated interactions with other risk factors so that the effect might not be as obvious as a single direction. Therefore, more validation is needed in terms of gender effect on AD, and we were keen to include this factor in the validation process as well.

Despite the associations between the three covariates with the onset of the disease, none of the previous studies so far included both these covariates and the biomolecular expression profiling together as prediction/diagnosis model input. In this study, we evaluated the impact of covariates by parameterising them instead of eliminating their effects in the pre-processing. Age, gender and APOE genotyping information were combined to form different covariate variables groups, and their contribution to improve the predictive performances were evaluated respectively.

In this chapter, our main aims are the followings:

1. To investigate whether the inclusion of covariate information of individuals such as age, gender and APOE genotyping information will enhance the performance of prediction models

2. To apply feature selection methods to try to discover informative RNA that are AD related, and build a prediction model based on selected RNA

4. To investigate the biological linkage between selected RNA and AD

3.3 Materials and Methods

In this study, our main objective is to investigate and find out informative RNA biomarkers related to AD. In addition, covariates impact to the performance of prediction models are also investigated. Due to the limit of covariate data availability, we only included three covariates in this study: age, gender, APOE genotyping.

All simulations and analyses were carried out with R (R Development Core Team), using packages varselRF [185] for random forest backward elimination, e1071 [212] for SVM.

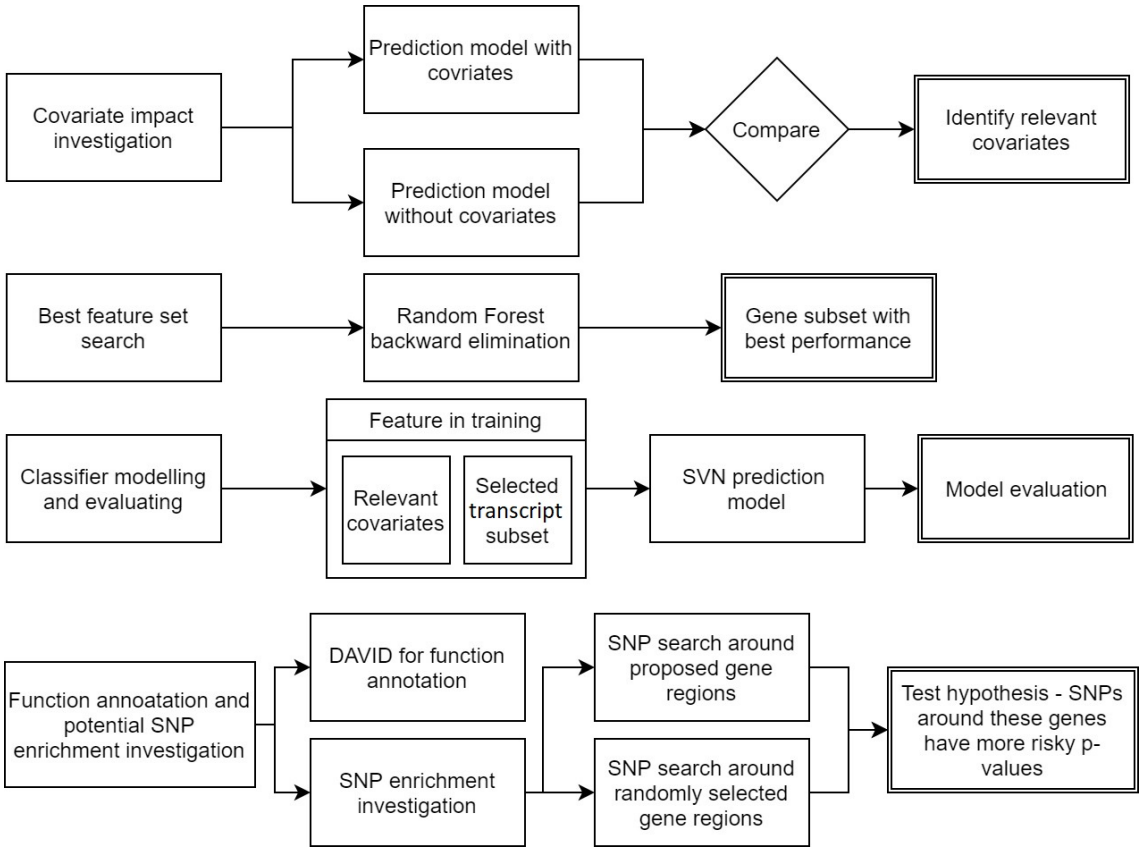


Figure 3.1: Work flow of study. The entire study contains covariate impact investigation, best feature set search, classifier modelling and evaluation, functional annotation and SNP enrichment investigation.

The work flow of the entire study was demonstrated in Figure 3.1.

3.3.1 Dataset overview

We tried to select the dataset that fit the purpose of our study. The dataset should match all the following requirements

- It is RNA expression profile extracted from human brain
- There are both AD and healthy control samples
- Apart from age and gender, APOE genotyping information should also be provided

Table 3.1: Sample summary of dataset GSE15222 in terms of experiment types, tissues, and covariate distribution.

Experiment type	RNA
Tissue	Brain
Sample count	187 health control, 176 LOAD samples
Age range	65-102
Gender	190male, 173 female
With APOE genotyping information	Yes

- Sample size should be as large as possible

With the application of the above requirements as dataset filters, we selected an RNA transcript expression profile for a total of 363 subjects summarized in the GEO (`http://www.ncbi.nlm.nih.gov/geo/`) series with accession number of **GSE15222**. This is the only dataset that matches all the requirements. The data set included 187 health control samples and 176 LOAD samples, with age (ranging from 65 to 102), gender (190 male and 173 female samples), and APOE genotyping information (category in detail: $\epsilon2/\epsilon2$, $\epsilon2/\epsilon3$, $\epsilon3/\epsilon3$, $\epsilon2/\epsilon4$, $\epsilon3/\epsilon4$, $\epsilon4/\epsilon4$) available (see Table 3.1).

A quick check about the distribution of samples between different conditions was also conducted, the result of which is shown in Table 3.2 and Figure 3.2

Table 3.2: Sample count of different disease states of dataset GSE15222 in different covariate groups.

		Disease status	
		AD	Control
Gender	Male	88	102
	Female	88	85
APOE	$\epsilon 2/\epsilon 2$	0	13
	$\epsilon 2/\epsilon 3$	3	15
	$\epsilon 3/\epsilon 3$	52	120
	$\epsilon 2/\epsilon 4$	9	2
	$\epsilon 3/\epsilon 4$	79	33
	$\epsilon 4/\epsilon 4$	33	4

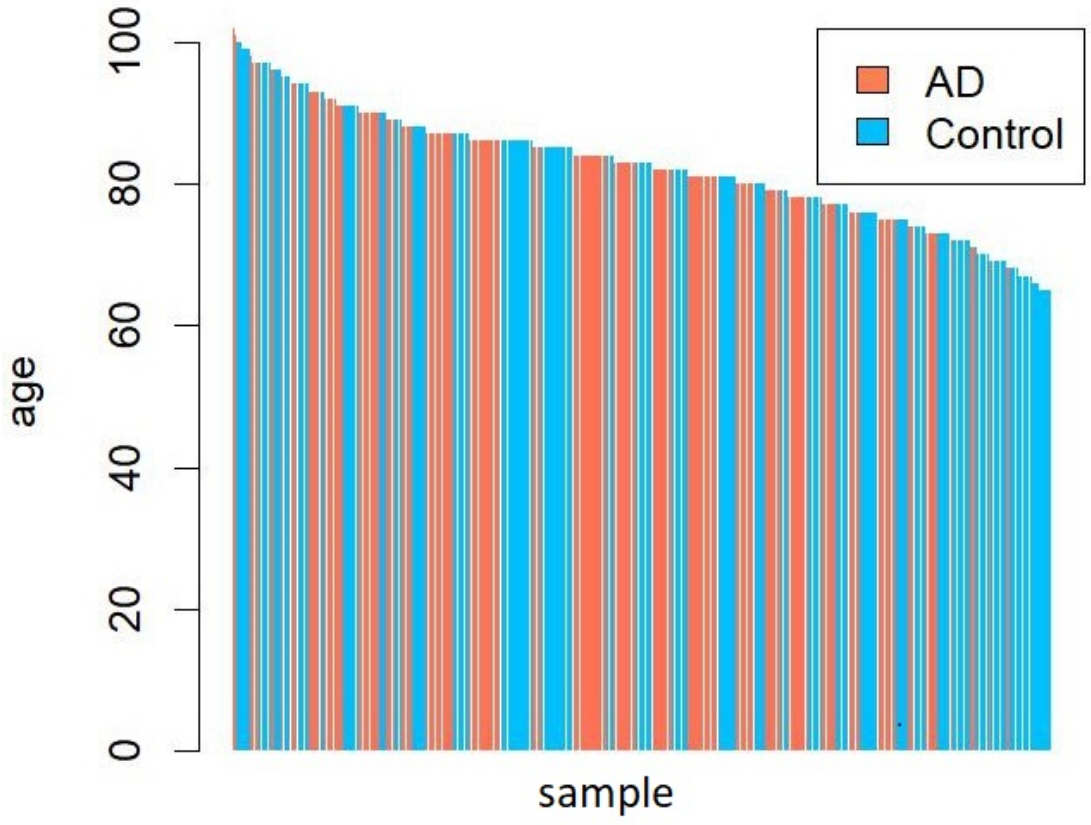


Figure 3.2: Sample age distribution across GSE15222 between AD and control. The horizontal axis indicates a sequence of samples sorted by age in a descending order. Red marks AD samples and blue marks healthy control samples.

We also employ PCA to identify the principle components and review if there are any outstanding features that represent a large portion of the data variance. The result is shown in Figure 3.3

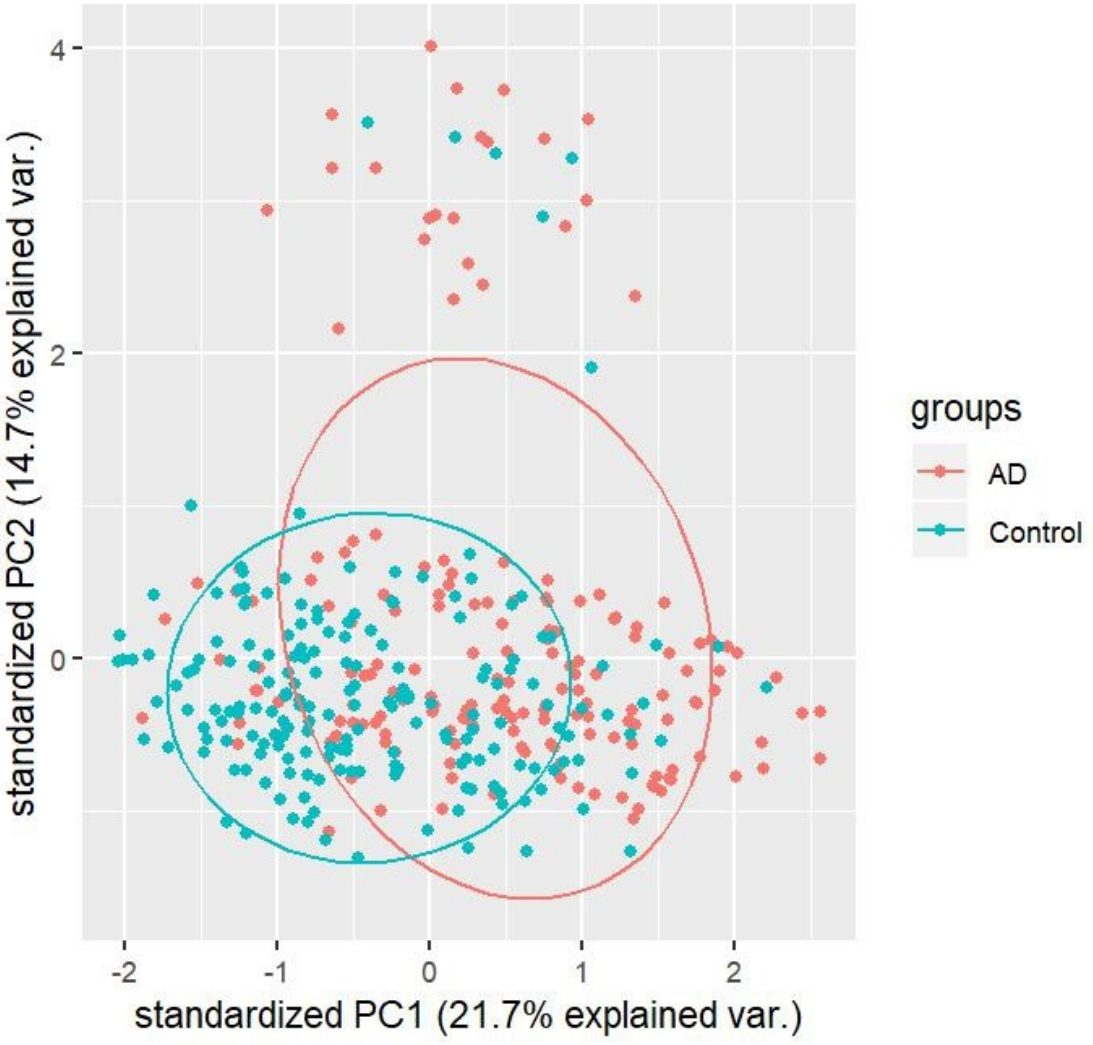


Figure 3.3: PCA analysis result. Top two PCs are shown and they account for below 40% of variation of data together.

The result of PCA indicates that to accounts for over 80% variance of the entire data, at least the top 22 components will need to be included. Therefore, there is no prominent compoenents (and thus feature) that can explain a significant portion of the data variance.

Table 3.3: Odd ratios for six allele combinations [204]

APOE allele combinations	Odd Ratio
$\epsilon 2/\epsilon 2$	0.6
$\epsilon 2/\epsilon 3$	0.6
$\epsilon 3/\epsilon 3$	1
$\epsilon 2/\epsilon 4$	2.6
$\epsilon 3/\epsilon 4$	3.2
$\epsilon 4/\epsilon 4$	14.9

3.3.2 Data collection and processing

We downloaded normalized data directly from Myers' lab (`http://labs.med.miami.edu/myers/LFuN/LFuN.html`), which includes 8650 genes were presented after the rank invariant normalized data was filtered by detection score and number of missing values [213]. This dataset has missing values because any intensity where the Illumina detection score (The detection score is a probability of expression beyond negative controls for each gene, calculated based on a normal distribution modelled using signals from 27 negative control probes) was less than 0.99 was coded as null, we applied k-nearest neighbour imputation (see section 2.3.2.2) with the following parameter settings (k = 10, rowmax = 0.5, colmax = 0.8, maxp = 1500, rng.seed=362436069) to treat missing values. The expression data were then log 10 transformed. The data collection and processing were carried out with R [214].

Covariate quantification The value of covariate gender were set to 0 and 1 representing male and female, whereas the APOE genotype consisting of six categories ($\epsilon 2/\epsilon 2$, $\epsilon 2/\epsilon 3$, $\epsilon 3/\epsilon 3$, $\epsilon 2/\epsilon 4$, $\epsilon 3/\epsilon 4$, $\epsilon 4/\epsilon 4$) were assigned with their respective odd ratios (OR) [204] (see Table 3.3).

3.3.3 Model catalog

The investigation covers covariate impact studies, performance assessment of proposed model consisting shortlisted RNA transcripts and the addition of covariates to those models. Besides, some benchmark models needs to be built. Therefore, the concept of a "random" model is developed, referring to models in which the features (including RNA probes in the microarray experiment and covariates) are selected randomly, instead of having a scheduled algorithm or plan behind. Finally, 24 models are built with each having different experiment contexts and different features selected for the model training and testing, which are listed as follows:

- 7 models with only covariates (age, gender, APOE, age+gender, age+APOE, gender+APOE, age+gender+APOE)
- 8 models, all containing proposed informative RNA transcript set (P-RNA-S) which is selected programatically by feature selection algoirthm, and differnt combination of covariates (P-RNA-S, P-RNA-S+age, P-RNA-S+gender, P-RNA-S+APOE, P-RNA-S+age+gender, P-RNA-S+age+APOE, P-RNA-S+gender+APOE, P-RNA-S+age+gender+APOE)
- 8 models, all containing randomly selected RNA transcript set (R-RNA-S), and differnt combination of covariates (R-RNA-S, R-RNA-S+age, R-RNA-S+gender, R-RNA-S+APOE, R-RNA-S+age+gender, R-RNA-S+age+APOE, R-RNA-S+gender+APOE, R-RNA-S+age+gender+APOE)
- 1 model trained in the validation from blood samples

3.3.4 RNA selection methods

There are two ways to select RNA in this study, a random selection or a Random Forest based selection.

3.3.4.1 Random selection

In contrary to Random Forest based selection, this method selects RNAs with a stochastic approach. Thus, the underlying idea of this method cannot be interpreted with conceivable logic or path.

3.3.4.2 Random Forest based selection

A feature selection method - Random Forest backward elimination (RFBE) [185] was conducted in the training set of each fold to select the best genes (mtryFactor=1, ntree=20, ntreeIterat=1000, vars.drop.frac=0.2).

Random forest backward elimination In this study, we used a random forests backward elimination method (RFBE) [185] to select significant genes which show strong relevance to the disease. This method employs the rank of variable importance of the tree with least out-of-bag (OOB) error among the forest to eliminate least important features in each iteration. Eventually the smallest feature subset upon which the tree is built with its error in a tolerable range is selected (see Algorithm 1).

input the full data matrix with all features present;
while *remaining feature size is larger than preset threshold (eg. 25)* **do**
    build a forest with remaining features;
    rank feature importance with OBB errors from each tree;
    select top features (eg. top 20%);
    drop unselected features;
end

Algorithm 1: Random forest backward elimination feature selection method

The foregoing workflow was repeated 1000 times to get an abundant group of potential feature sets, among which the best one will be used to train a prediction model. To unbiasedly assess the classifying performance of these feature sets, we trained a classifier with each of them by SVM model and assessed their training and testing accuracies. A performance index (PI) of this gene subset was then calculated as a weighted sum of its training and testing accuracies:

$$PI = \frac{n_{training}}{N} A_{training} + \frac{n_{testing}}{N} A_{testing} \tag{3.1}$$

where $n_{training}$ and $n_{testing}$ are the input sample size of training and testing, N is sample size of the entire dataset, $A_{training}$ and $A_{testing}$ are the training and testing accuracies. We then selected the transcript subset with highest performance index in the 1000 trials.

3.3.5 Model building and assessment

Once the a set of RNA set is shortlisted, cross validations are conducted and iterated to get assessment result for different models. These results are then used to intepret the questions we have suggested earlier in this chapter. To summarize:

- How are the proposed RNA features performing in classifying AD and control samples, comparing to other models
- Are the covariates contributing to the classifying performance? If so, which covariates are the most effective in improving the model performances?

3.3.5.1 The cross-validation cycle

The build-and-test cycle was conducted with every abovementioned model a) split samples into training and testing sets, b) filter data with selected features, c) a model was learned from the training set, d) assess the above generated model with the testing set.

Sample splitting Because we applied a five-fold cross validation (CV) approach (see definition in section 2.4.5.1), so for each of the five folds, four subsets are used for training and the remaining subset for testing.

Filter data to lower dimension As previously discussed in Section 3.3.3, we have 24 models in total to be trained and tested. For each model, there is different composition of selected features which are employed in the model training process. For instance,

the feature of model "proposed+age+apoe" consists of age, APOE genotyping, and an RNA set selected by a pre-designed algorithm discussed in Section 3.3.4.2, whereas that of model "random+gender" consists of gender and an RNA set selected randomly (discussed in Section 3.3.4.1).

Model training All the models are trained by Support Vector Machine (SVM), with a default *gamma* ($=1/featurenumber$) and a *cost* equal to 0.1.

Model assessment To assess the performances of all the models, metrics derived from the confusion matrix are calculated in each execution of the procedure, and an average and standard deviation were calculated. These include accuracy, SN, SP, PPV, NPV, AUC.

The procedure of the above descibed five-fold cross validation is repeated 200 times for each model to make more generalised conclusions.

3.3.5.2 Hypothesis testings

Part of the assessment in this study is about examing hypothesis of existence of improvement between models. Suppose we are comparing the performance between model A and model B and the latter is in favour in our assumption. The following hypothesis is definend

- H_0 There is no accuracy/AUC improvement in training or testing
- H_a The accuracy/AUC of model B is greater than that of model A

We then examined the hypothesis by one-way two-sample t-test to get the respective p-values.

3.3.6 IGAP (International Genomics of Alzheimer's Project)

International Genomics of Alzheimerś Project (IGAP) [85] is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ances-

try (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

We hypothesize that SNPs are more enriched in the regions of proposed genes than that of those selected randomly. Hence, we conducted a potential SNP enrichment investigation in order to examine the hypothesis that SNPs around these genes have more p-values in AD-associated genetic variants than SNPs around genes selected by chance. We employed the two-stage GWAS meta-analysis result (Lambert, Jean-Charles, et al) to test the hypothesis. We first extracted IGAP SNPs in the 200kbp and 500kbp up/down stream of the proposed genes. For each gene, a search for SNP records in the data source (200kbp/500kbp up/down) was carried out. After the candidate GWAS SNP group for the region of selected genes were found and congregated, we corrected the association p-values by the number of SNPs extracted or by the number of SNPs not in linkage disequilibrium (LD, $r^2 < 0.5$) calculated by SNAP online tool (http://www.broadinstitute.org/mpg/snap/). SNPs with corrected p-value < 0.05 and have predicted function will be the potential candidate SNPs. Afterwards, we test the above-mentioned hypothesis under the following procedure:

a) Randomly select equal number of genes from the genome, get their SNPs in region and calculate test statistics $ts = -log_{10}(p - values)$;

b) Get the median and average of ts;

c) Repeat step a) and b) 1000 times to get the distribution of median and average

of ts (imputated ts);

d) Calculate the average and median of ts of SNPs from proposed genes (observed ts);

e) Get empirical p-value by counting the number of cases when imputated ts is larger than observed ts;

In the final stage of this study, function annotation for the selected subset of genes was conducted. Because the 26 genes identified are not suitable for pathway enrichment analysis, DAVID (http://david.abcc.ncifcrf.gov/) was used for annotation and to learn the possible biological functions or pathways encrypted in these genes.

3.3.7 Validation in blood

The proposed RNA set was also validated in a blood dataset with GEO series number **GSE85426**. This dataset was refined by keywords "Alzheimer", "blood" and "RNA" in *ArrayExpress* (`https://www.ebi.ac.uk/arrayexpress/`). The RNA was extracted from peripheral blood cells, reverse-transcribed and labelled, then analysed for gene expression using GeneSpring GX12 (Agilent Technologies, USA). The dataset included 90 non-demented control and 90 AD samples, with age (ranging from 64 to 94), gender (92 male and 88 female) information provided. Since APOE information is not provided in this dataset, we chose not to include any covariate information in the validation.

The proposed RNA set, all identified with official gene symbol, were first mapped to Agilent probe IDs, and then the workflow described in section 3.3.5 was followed to obtain assessment result for the blood dataset.

3.4 Result

3.4.1 Classifier modelling and evaluating

We searched for the best feature set in the data set for classification. We discovered an SVM model parameterized with 26 RNAs which shows promising accuracies for both training and testing. In table 3.5, proposed model is trained with the 26 selected RNA.

Table 3.5 shows that on average, the proposed model outperforms random models (mean training/testing accuracy of proposed model: 0.888/0.859, of random model: 0.813/0.765, p-value=0 for both training and testing) without the addition of covariates. The proposed model with covaraites presents excellent classifying ability, with average training/testing accuracy of 0.914/0.876, SN 0.884/0.842, SP 0.943/0.907, PPV 0.936/0.898, NPV 0.897/0.862, AUC 0.961/0.942 for the entire dataset.

Table 3.4: Selected genes used in training the prediction model

Gene Name	Importance	Fold Change	P Value
ZNF264	4.94E-03	2.053	7.90E-19
ZDHHC23	2.32E-03	0.622	3.70E-28
SPOP	1.94E-03	1.405	3.20E-20
NRCAM	1.68E-03	0.727	5.00E-23
DSTYK	1.58E-03	1.382	2.20E-25
ABCC12	1.39E-03	0.645	1.80E-22
SRGAP1	1.34E-03	2.057	4.90E-19
NEUROD6	1.20E-03	0.485	2.20E-21
CHMP4B	1.18E-03	0.796	3.00E-23
C19ORF30	1.07E-03	0.597	2.30E-11
MRPL24	8.26E-04	0.527	8.00E-21
VEZF1	7.87E-04	1.629	2.40E-17
LETMD1	7.64E-04	0.817	8.00E-22
DDX23	6.64E-04	1.279	4.90E-19
PPEF1	6.34E-04	0.523	2.80E-28
KCTD13	6.25E-04	0.76	7.10E-16
ZHX3	6.02E-04	1.406	6.90E-21
FLJ37464	4.93E-04	0.744	9.80E-15
SGIP1	3.74E-04	0.645	1.80E-22
CASP8AP2	1.14E-04	1.98	1.10E-10
TCF3	1.06E-04	1.47	1.20E-20
LANCL2	0.00E+00	0.696	2.20E-26
KCNF1	0.00E+00	0.636	1.10E-20
RAB29	0.00E+00	1.775	5.50E-16
TUBB2A	0.00E+00	0.63	1.40E-18
STX1A	0.00E+00	0.739	5.10E-16

Fold change for each gene is calculated by the quotient between the means of expression values from both conditions (AD and health control, mean of AD as numerator) for each gene. (see Method). P-value is calculated as the p-value of the two-tail two-sample t-test for expression values from both conditions for each gene.

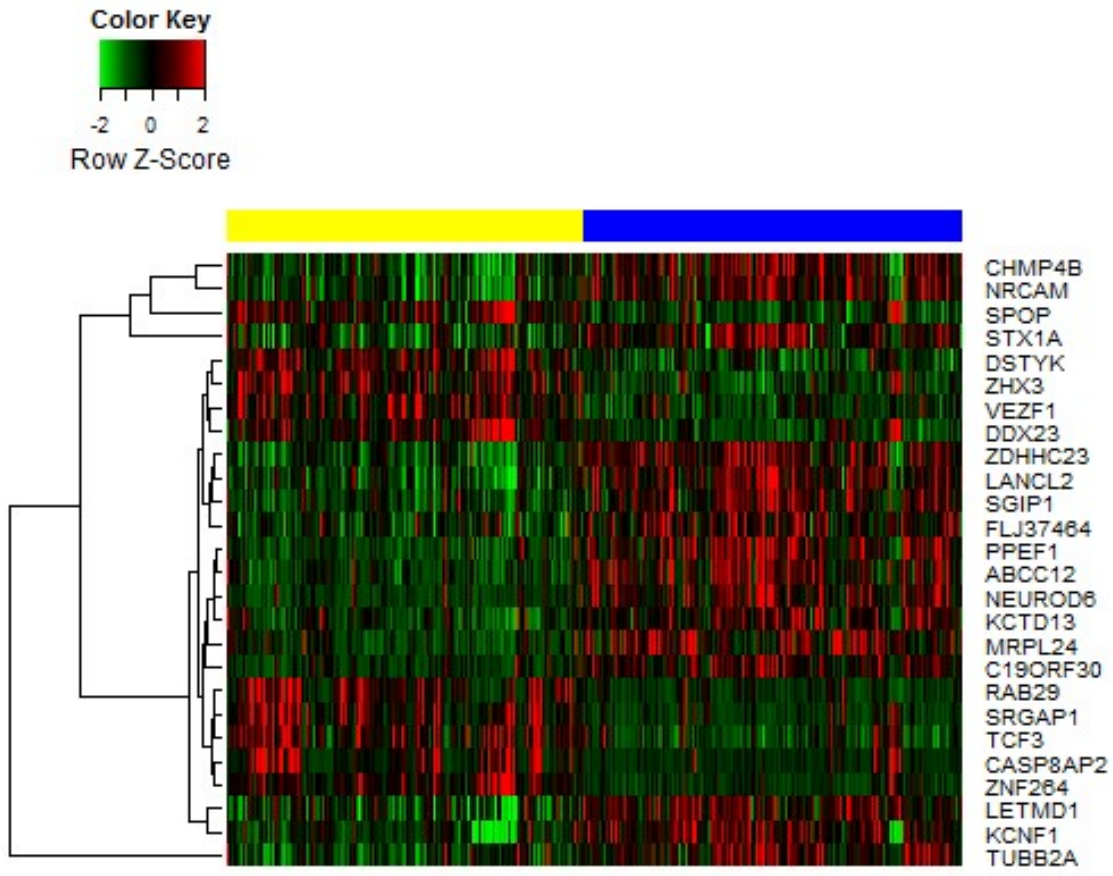


Figure 3.4: Heat map of expression level in data set for the 26 selected genes. The yellow bar on top of the heatmap represents AD sample block and blue bar represents normal health control sample block. Expression values are normalised by row.

3.4.2 Covariate impact investigation

The result of covariate impact investigation is shown in Table 3.5. Among the three covariates (age, gender, APOE), APOE has the best performance as a standalone classifier, with nearly 75% of both training and testing accuracy on average. Statistics also suggest that adding only age as a factor in the prediction model would improve the training accuracy of proposed model (p-value $= 1.73 \times 10^{-5}$), but not the testing accuracy (p-value$= 1$). A same difference in performance was observed in random model (p-value $= 2.31 \times 10^{-5}$ for training and 2.09×10^{-1} for testing). In terms of adding gender as an input variable in the model, there is a slight increase in training accuracy for both proposed and random model (p-value $= 7.77 \times 10^{-2}$ for random and 2.61×10^{-13} for

proposed), but the drop regarding testing accuracy demonstrates that the improvement hardly exists (p-value=8.77×10^{-1} for random and 1.65×10^{-1} for proposed model). It shows that gender as a factor is very unlikely to be relevant to the onset of AD and therefore lack of diagnostic value. The most remarkable enhancement is generated by including APOE as a factor in the classifier modelling process, with convincing support from statistics of both random and proposed model (random: training=0, testing=0; proposed: training=0, testing=0). It implies APOE is a significant risky factor for the onset of AD with distinctive classifying ability.

Table 3.5: Performances of all models

		Accuracy (mean +/- sd)	Sensitivity (mean +/- sd)	Specificity (mean +/- sd)
age	TN	0.558+/-0.017	0.660+/-0.047	0.462+/-0.024
	TS	0.558+/-0.048	0.660+/-0.080	0.462+/-0.088
apoe	TN	0.741+/-0.012	0.687+/-0.018	0.791+/-0.015
	TS	0.741+/-0.047	0.687+/-0.071	0.791+/-0.061
gender	TN	0.524+/-0.011	0.252+/-0.256	0.781+/-0.222
	TS	0.498+/-0.028	0.225+/-0.232	0.756+/-0.251
age + gender	TN	0.557+/-0.016	0.640+/-0.050	0.480+/-0.037
	TS	0.542+/-0.053	0.622+/-0.094	0.467+/-0.085
age + apoe	TN	0.741+/-0.012	0.687+/-0.018	0.791+/-0.015
	TS	0.741+/-0.047	0.687+/-0.071	0.791+/-0.061
apoe + gender	TN	0.741+/-0.012	0.687+/-0.018	0.791+/-0.015
	TS	0.741+/-0.047	0.687+/-0.071	0.791+/-0.061
age + gender + apoe	TN	0.741+/-0.012	0.688+/-0.018	0.791+/-0.015
	TS	0.741+/-0.047	0.687+/-0.071	0.791+/-0.061
proposed	TN	0.888+/-0.010	0.863+/-0.016	0.912+/-0.012

Continued on next page

Table 3.5 – *Continued from previous page*

		Accuracy (mean +/- sd)	Sensitivity (mean +/- sd)	Specificity (mean +/- sd)
proposed + age	TS	0.859+/-0.037	0.826+/-0.060	0.891+/-0.049
	TN	0.889+/-0.010	0.867+/-0.016	0.910+/-0.013
proposed + gender	TS	0.856+/-0.037	0.827+/-0.062	0.884+/-0.051
	TN	0.890+/-0.010	0.863+/-0.016	0.915+/-0.012
proposed + apoe	TS	0.860+/-0.036	0.824+/-0.060	0.893+/-0.049
	TN	0.908+/-0.009	0.882+/-0.016	0.933+/-0.011
proposed + age + apoe	TS	0.879+/-0.036	0.844+/-0.060	0.911+/-0.046
	TN	0.914+/-0.010	0.884+/-0.017	0.943+/-0.013
	TS	0.876+/-0.036	0.842+/-0.060	0.907+/-0.048
proposed + gender + apoe	TN	0.910+/-0.009	0.886+/-0.016	0.933+/-0.011
	TS	0.878+/-0.035	0.845+/-0.060	0.910+/-0.047
proposed + age + gender	TN	0.892+/-0.010	0.871+/-0.016	0.912+/-0.013
	TS	0.858+/-0.037	0.830+/-0.060	0.885+/-0.051
proposed + age + apoe + gender	TN	0.915+/-0.010	0.885+/-0.017	0.943+/-0.014
	TS	0.874+/-0.036	0.842+/-0.060	0.905+/-0.049
random	TN	0.813+/-0.025	0.783+/-0.035	0.840+/-0.032
	TS	0.765+/-0.051	0.736+/-0.079	0.793+/-0.071
random + age	TN	0.815+/-0.024	0.792+/-0.034	0.836+/-0.033
	TS	0.766+/-0.051	0.744+/-0.078	0.787+/-0.073
random + gender	TN	0.813+/-0.025	0.784+/-0.035	0.841+/-0.032
	TS	0.764+/-0.051	0.734+/-0.079	0.792+/-0.071
random	TN	0.848+/-0.022	0.824+/-0.028	0.870+/-0.029
+ apoe	TS	0.801+/-0.047	0.779+/-0.071	0.822+/-0.068
random + age + apoe	TN	0.853+/-0.021	0.834+/-0.027	0.872+/-0.029
	TS	0.804+/-0.046	0.787+/-0.070	0.821+/-0.068

Continued on next page

Table 3.5 – *Continued from previous page*

		Accuracy (mean +/- sd)	Sensitivity (mean +/- sd)	Specificity (mean +/- sd)
random + gender + apoe	TN	0.848+/-0.022	0.824+/-0.028	0.871+/-0.029
	TS	0.800+/-0.048	0.777+/-0.072	0.821+/-0.068
random + age + gender	TN	0.816+/-0.024	0.794+/-0.033	0.836+/-0.033
	TS	0.765+/-0.051	0.743+/-0.078	0.785+/-0.074
random + age + apoe + gender	TN	0.855+/-0.021	0.834+/-0.027	0.874+/-0.028
	TS	0.804+/-0.046	0.786+/-0.071	0.822+/-0.068
blood	TN	0.679+/-0.018	0.656+/-0.051	0.701+/-0.039
	TS	0.615+/-0.075	0.580+/-0.120	0.650+/-0.104

Abbreviations: TN - training TS - testing

Table 3.6: Performances of all models - other measurement

		PPV (mean +/- sd)	NPV (mean +/- sd)	AUC (mean +/- sd)
age	TN	0.535+/-0.014	0.592+/-0.026	0.579+/-0.015
	TS	0.537+/-0.043	0.592+/-0.063	0.579+/-0.060
apoe	TN	0.756+/-0.014	0.729+/-0.012	0.766+/-0.012
	TS	0.759+/-0.058	0.731+/-0.048	0.766+/-0.047
gender	TN	NaN+/-NA	0.531+/-0.017	0.520+/-0.016
	TS	NaN+/-NA	0.506+/-0.024	0.497+/-0.056
age + gender	TN	0.536+/-0.014	0.587+/-0.023	0.581+/-0.015
	TS	0.523+/-0.049	0.570+/-0.065	0.565+/-0.061
age +	TN	0.756+/-0.014	0.729+/-0.012	0.799+/-0.012

Continued on next page

Table 3.6 – *Continued from previous page*

		PPV (mean +/- sd)	NPV (mean +/- sd)	AUC (mean +/- sd)
apoe	TS	0.759+/-0.058	0.731+/-0.048	0.799+/-0.048
apoe + gender	TN	0.756+/-0.014	0.729+/-0.012	0.773+/-0.014
	TS	0.759+/-0.058	0.731+/-0.048	0.767+/-0.051
age + gender + apoe	TN	0.756+/-0.014	0.729+/-0.012	0.800+/-0.012
	TS	0.759+/-0.058	0.731+/-0.048	0.797+/-0.048
proposed	TN	0.903+/-0.012	0.876+/-0.013	0.934+/-0.007
	TS	0.879+/-0.048	0.847+/-0.045	0.911+/-0.032
proposed + age	TN	0.901+/-0.013	0.880+/-0.013	0.939+/-0.007
	TS	0.873+/-0.049	0.847+/-0.047	0.918+/-0.031
proposed + gender	TN	0.906+/-0.012	0.877+/-0.013	0.935+/-0.007
	TS	0.882+/-0.048	0.846+/-0.045	0.911+/-0.032
proposed + apoe	TN	0.926+/-0.011	0.894+/-0.013	0.954+/-0.006
	TS	0.902+/-0.046	0.864+/-0.046	0.932+/-0.027
proposed +	TN	0.936+/-0.013	0.897+/-0.014	0.961+/-0.005
age + apoe	TS	0.898+/-0.048	0.862+/-0.046	0.942+/-0.025
proposed + gender + apoe	TN	0.925+/-0.011	0.897+/-0.013	0.955+/-0.006
	TS	0.901+/-0.047	0.864+/-0.046	0.932+/-0.027
proposed + gender + age	TN	0.903+/-0.013	0.883+/-0.013	0.941+/-0.007
	TS	0.874+/-0.050	0.849+/-0.046	0.920+/-0.030
proposed + age + apoe + gender	TN	0.937+/-0.014	0.897+/-0.014	0.963+/-0.005
	TS	0.895+/-0.048	0.861+/-0.046	0.944+/-0.025
random	TN	0.823+/-0.031	0.805+/-0.027	0.879+/-0.022
	TS	0.773+/-0.063	0.764+/-0.056	0.837+/-0.050
random + age	TN	0.820+/-0.031	0.811+/-0.026	0.881+/-0.022
	TS	0.770+/-0.063	0.769+/-0.056	0.838+/-0.049

Continued on next page

Table 3.6 – *Continued from previous page*

		PPV (mean +/- sd)	NPV (mean +/- sd)	AUC (mean +/- sd)
random + gender	TN	0.824+/-0.031	0.806+/-0.027	0.879+/-0.022
	TS	0.772+/-0.063	0.763+/-0.056	0.836+/-0.050
random + apoe	TN	0.857+/-0.028	0.841+/-0.023	0.917+/-0.015
	TS	0.808+/-0.061	0.800+/-0.053	0.883+/-0.040
random + age + apoe	TN	0.860+/-0.028	0.848+/-0.022	0.922+/-0.014
	TS	0.809+/-0.060	0.806+/-0.053	0.888+/-0.038
random + gender + apoe	TN	0.858+/-0.029	0.841+/-0.023	0.917+/-0.015
	TS	0.807+/-0.061	0.799+/-0.054	0.881+/-0.040
random + age + gender	TN	0.821+/-0.031	0.812+/-0.026	0.882+/-0.022
	TS	0.768+/-0.063	0.767+/-0.056	0.838+/-0.049
random + age + apoe + gender	TN	0.863+/-0.028	0.849+/-0.022	0.923+/-0.014
	TS	0.810+/-0.060	0.806+/-0.053	0.887+/-0.039
blood	TN	0.688+/-0.020	0.672+/-0.026	0.765+/-0.013
	TS	0.626+/-0.083	0.612+/-0.077	0.700+/-0.077

Abbreviations: TN - training TS - testing

3.4.3 Validation in blood

The validation result in blood samples shows no significant classification performance, with an average training/testing accuracy 0.679/0.615 (See table 3.5).

3.4.4 Function annotation and pathway analysis for selected genes

DAVID for the 26 genes revealed a set of significant molecular biological functions and pathway information. The presence of KCTD13, KCNF1 and SPOP cast strong support

for the involvement of BTB/POZ (Broad-Complex, Tramtrack and Bric a brac, also known as poxvirus and zinc finger) domain (DAVID calculated p-value=3.0×10^{-2}). The nucleus activity function (DAVID calculated p-value=3.9×10^{-2}) is also in list because of the presence of DDX23, LANCL2, CASP3AP2, NEUROD6, KCTD13, SPOP, TCF3, VEZF1, ZNF264 and ZHX3.

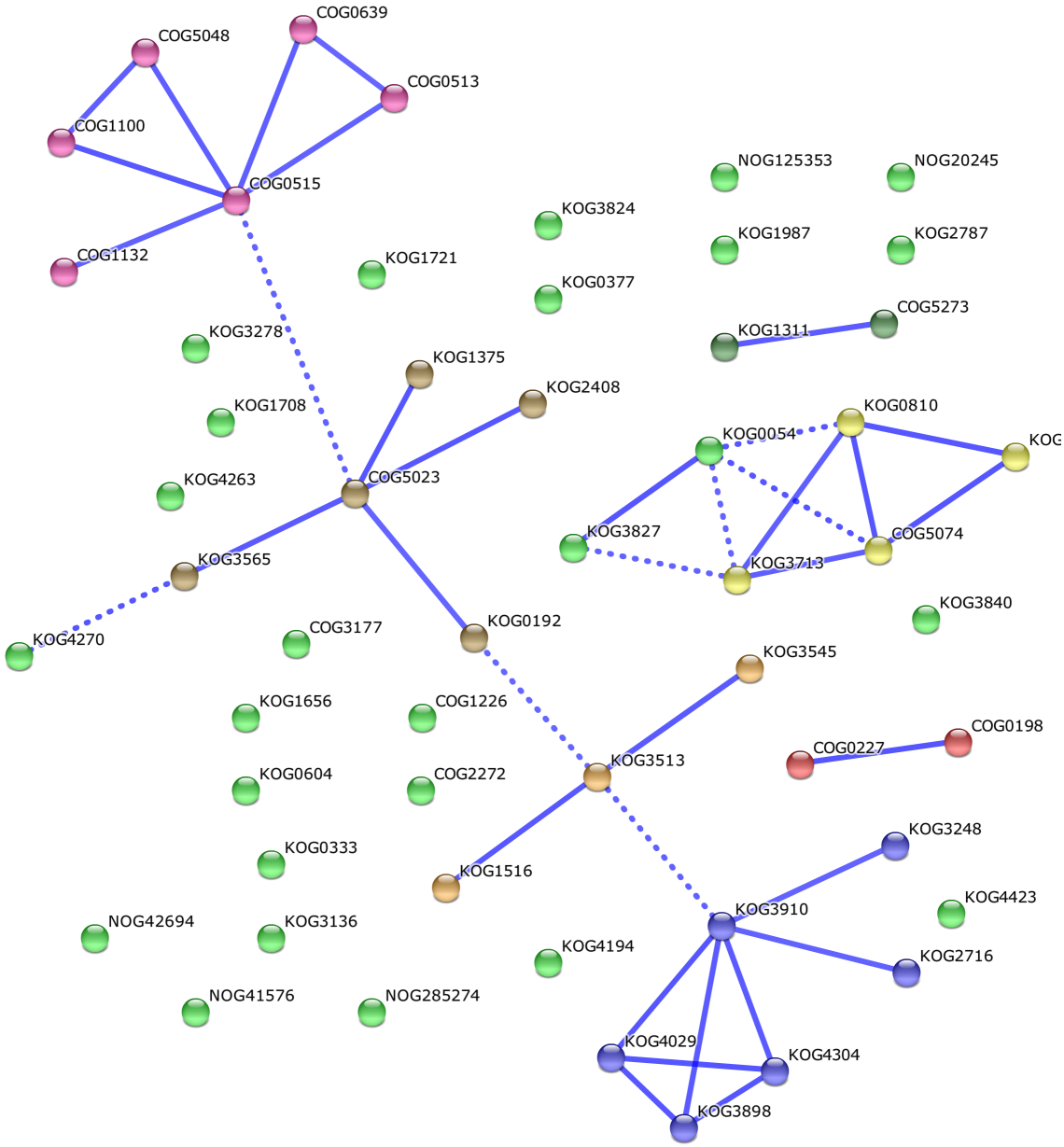


Figure 3.5: Confidence view of the protein-protein interaction network inferred from 26 genes. Stronger associations are represented by thicker lines. Different colours implicate different clustering which was calculated by a 10-means clustering method.

Table 3.7: Protein's annotation for clusters in the PPI network. This table corresponds to Figure 3.5 that displays only COG/KOG ID. The bolden genes are among the selected 26 genes.

COGs/KOGs	Gene Symbol	Protein Group Description	Cluter
COG1100	ARL6, DIRAS3, etc.	GTPase SAR1 and related small G proteins	1
COG1132	ABCC9, ABCC8 etc.	ABC-type multidrug transport system, ATPase and permease components	1
COG0515	MSSK1, SRPK1, SRPK2	Serine/threonine protein kinase	1
COG5048	ZNF511	FOG- Zn-finger	1
COG0513	DDX1, DDX50, etc.	Superfamily II DNA and RNA helicases	1
COG0639	**PPEF1**, PPEF2, etc.	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	1
COG5023	**TUBB2A**, TUBB6, etc.	Tubulin	2
KOG1375	TUBB4B, TUBB4A, etc.	Beta tubulin	2
KOG2408	PXDN, LPO, etc.	Peroxidase/oxygenase	2
KOG0192	MLKL, **DSTYK**, etc.	Tyrosine kinase specific for activated (GTP-bound)	2
KOG3565	SRGAP3, **SRGAP1**, SRGAP2	Cdc42-interacting protein CIP4	2
KOG1516	CEL, NLGN1, etc.	Carboxylesterase and related proteins	3

Continued on next page

Table 3.7 – *Continued from previous page*

COGs/KOGs	Gene Symbol	Protein Group Description	Cluter
KOG3513	**NRCAM**, CNTN1, etc.	Neural cell adhesion molecule L1	3
KOG3545	OLFML2A, LPHN1, etc.	Olfactomedin and related extracellular matrix glycoproteins	3
KOG3910	**TCF3**, TCF4, TCF12	Helix loop helix transcription factor	4
KOG4029	MSGN1, SCXB, MESP2, etc.	Transcription factor HAND2/Transcription factor TAL1/TAL2/LYL1	4
KOG3898	BHLHE22, NEUROG3, etc.	Transcription factor NeuroD and related HTH proteins	4
KOG4304	HES4, BHLHE41, etc.	Transcriptional repressors of the hairy/E(spl) family (contains HLH)	4
KOG2716	KCTD10, **KCTD13**, EDP1	Polymerase delta-interacting protein PDIP1 and related proteins, contain BTB/POZ domain	4
KOG3248	TCF7L2, TCF7L1, etc.	Transcription factor TCF-4	4
KOG0810	STX1B, STX3, etc.	SNARE protein Syntaxin 1 and related proteins	5
KOG1983	LLGL1, LLGL2	Tomosyn and related SNARE-interacting proteins	5
COG5074	STX1B, STX3, STX1C, etc.	t-SNARE complex subunit, syntaxin	5

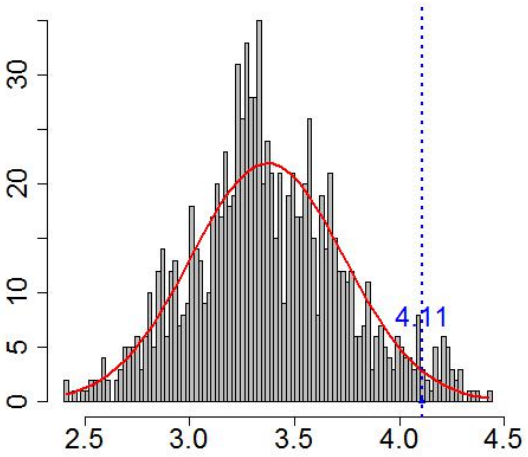
Continued on next page

Table 3.7 – *Continued from previous page*

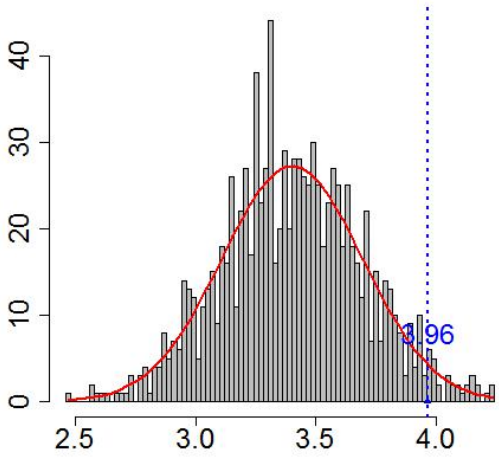
COGs/KOGs	Gene Symbol	Protein Group Description	Cluter
KOG3713	**KCNF1**, KCNS1, etc.	Voltage-gated K+ channel KCNB/KCNC	5
COG0227	MSRA	Ribosomal protein L28	6
COG0198	**MRPL24**, RPL26L1, RPL26	Ribosomal protein L24	6

3.4.5 SNP enrichment investigation

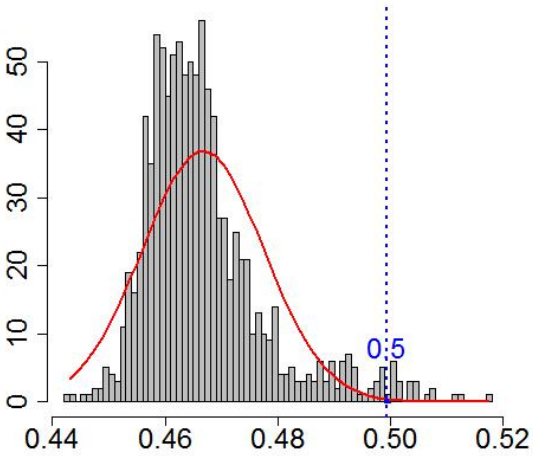
The result of SNP enrichment investigation managed to support the hypothesis that SNPs around these genes have more p-values in AD-associated genetic variants than SNPs around genes selected by chance. It suggests that the regions of the selected 26 genes prone to suffer more phenotyping mutations than that of normal genes. Figure 3.6 demonstrates the position where the observed average and median *ts* (marked with blue lines) locate among the distribution of average and median *ts* imputed by chance. The result reveals that the observed average *ts* are large enough(p-value=3.6×10^{-2}, 3×10^{-2}, 2.2×10^{-2}, 3.9×10^{-2} for 200k up/down search in stage 1 & 2 combined data, 500k up/down search in stage 1 & 2 combined data, 200k up/down search in stage 1 data, and 500k up/down search for stage 1 data only, corresponding to subgraph $a - d$ in Figure 3.6) to support the existence of the hypothesis. Comparatively, the statistics of median *ts* support the hypothesis as well (p-value=4.2×10^{-2}, 3.8×10^{-2}, 7×10^{-3}, 2.4×10^{-2} for 200k up/down search in stage 1 & 2 combined data and 500k up/down search in stage 1 & 2 combined data, 200k up/down search in stage 1 data, and 500k up/down search for stage 1 data only, corresponding to subgraph $e - h$ at the bottom in Figure 3.6.).



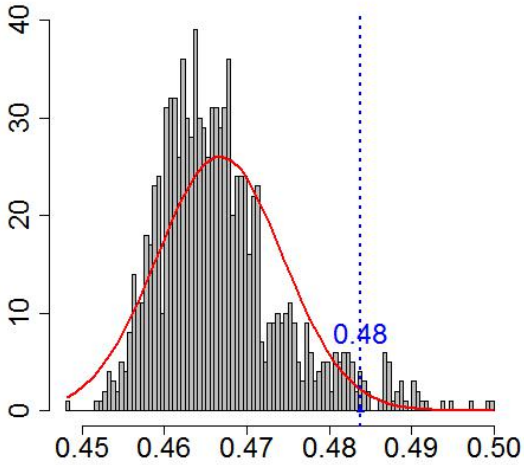
(a)



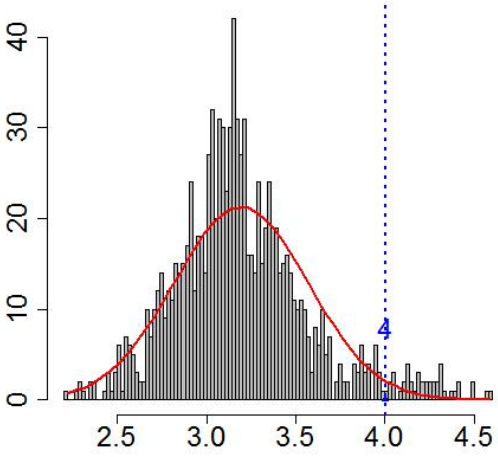
(b)



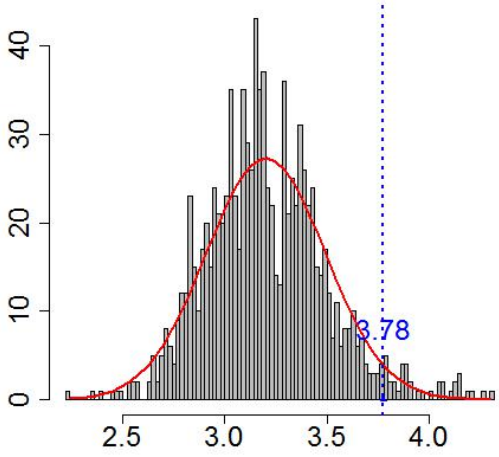
(c)



(d)

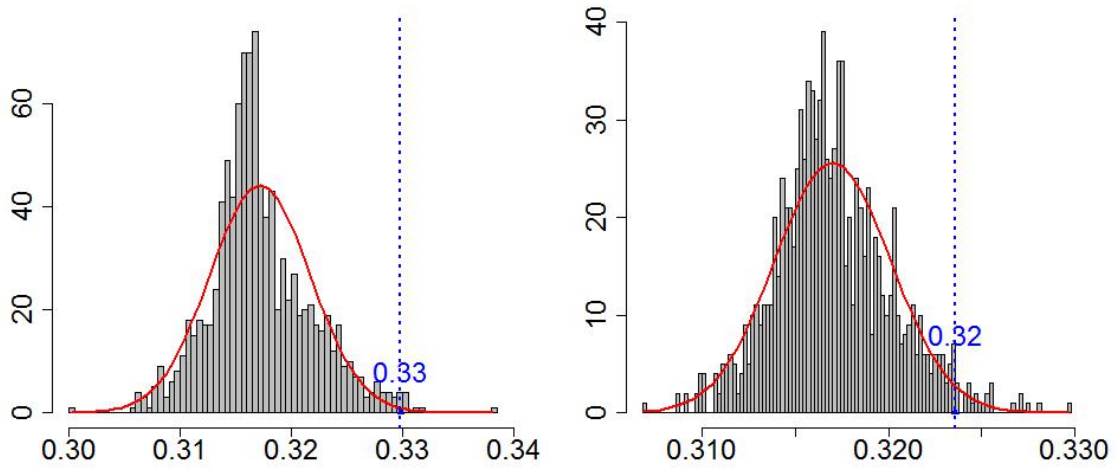


(e)



(f)

Figure 3.6



(g) (h)

Figure 3.6: Average and median imputed test statistics ($ts = -log_{10}pvalue$) distribution and observed test statistics from 26 selected transcripts. Vertical coordinate represents the occurrences of imputated average or median *ts*, horizontal coordinates represents average or median *ts*. The value of observed *ts* is marked with blue line. (a)Average, stage 1&2, 200kbp (b)Average, stage 1&2, 500kbp (c)Average, stage 1, 200kbp (d)Average, stage 1, 500kbp (e)Median, stage 1&2, 200kbp (f) Median, stage 1&2, 500kbp (g) Median, stage 1, 200kbp, (h)Median, stage 1, 500kbp

3.5 Discussion

We investigated the effect which three covariates (age, gender, APOE genotype) have made on the prediction model. We quantified the six categories of APOE genotype with their reported odd ratios. The idea orginates from the findings in previous studies [215–220] which suggest that $\epsilon 4$ presents a more than threefold increase in LOAD risk when compared to the common $\epsilon 3$, while carries of $\epsilon 2$ are exposed to lower risk of LOAD compared to non-carriers.

It is the first time that covariates such as age, gender and APOE genotype are to be parameterized in the process of modelling. We obtained better result when age and APOE genotype were included in our model. The enhancement the model acquired by including age as variable reveals the fact that there are still undiscovered biomolecule

markers which can fully express the age difference between individuals so as to make better predictions. Meanwhile, it is a novel attempt to employ genotyping information to achieve classification of disease status. APOE genotype information illustrated promising classifying ability. Thus further studies should be focusing not only on expression intensities but also exploiting the potential of genotyping information in terms of disease prediction.

We employed the method of RFBE to select significant genes before training the classifier. 26 out of 8650 genes were collected, with which we trained a prediction model with as high as 85% testing accuracy. It was a successful application of RFBE in terms of feature size and accuracy. Nevertheless, there are defects for the method itself. One main defect is that the variable importances are fixed. Variable importances are the measurement that determines what features are to be kept or eliminated. Despite their essence, they are not recalculated in each loop when some features are eliminated. Another defect of the method is the ignorance of link between features. According to the philosophy of Random forest, when calculating the variable importance, the interactive effect between variables will be ignored and variables are regarded independent, which is obviously against the actual situation. The interaction and co-effects of genes are definitely elements to be paid attention to in AD development. Some studies had dovated their attention this issue by various approaches. Tan et al [221] intergrated DEG analysis with Graqm-Schmidt process to identify genes that are both significant and highly informative for predicting tumour survival. A kernel-based multivariate feature selection method was proposed [222] to discover nonlinear correlations among features as well as between feature and target. Recently, a new hybrid method using Independent component analysis (ICA) and Artificial Bee Colony (ABC) was reported to have the ability to generate small subsets of genes and improve classification accuracy of classifier built upon Naive Bayes (NB) [223]. These above-mentioned methods all take into account the inter-correlation of genes when reducing the dimension, but the application of these models on dementia still need further investigations.

BTB/POZ domain is suggested to be a function cluster in the 26 genes with high

confidence, among which a gene member KCTD13 was reported to be the major driver for the neurodevelopmental phenotypes associated with the 16p11.2 copy number variant (CNV) [224]. It was also found to locate in a chromosomal rearrangement hotspot and dosage-sensitive loci, in which a rare duplication on chromosome 16p11.2 was identified in patients with psychosis in AD [225]. Another gene detected in the BTB/POZ domain, KCTF1, was listed in the genes whose expression is remarkably altered during senescence in both human and mouse cortex [226]. However, SPOP has not been reported to be associated with the pathology of any neurodegenerative diseases.

Apart from the likely appearance of BTB/POZ domain, functional annotation from DAVID also suggested that four genes encoding zinc finger proteins, vascular endothelial zinc finger 1 (VEZF1), zinc fingers and homeoboxes 3 (ZHX3), zinc finger protein 264 (ZNF264) and zinc finger DHHC-type containing 23 (ZDHHC23), were found among 26 genes.

We studied the protein-protein interaction network encoded by the 26 genes with STRING (https://tring-db.org). 54 proteins appeared to match the input of 26 genes in STRING database. The protein network is not enriched in interactions and 223 interactions were observed. The network was displayed with highest confidence (0.900) and zero interactor shown (See Figure 3.5.). There are basically 7 clusters found in the network (marked with different colors in Figure 3.5., nodes/proteins in every cluster annotated in Table 3.7.). In cluster 1 (nodes marked with pink in Figure 3.5), the polymorphisms of a homolog B of GTPase SAR1 (COG1100) was found recently to be associated with reduced risk of AD among APOE ϵ4 non-carriers [227]. FOG (Friend of GATA) is a zinc finger protein family with FOG-1 (friend of GATA-1) reported to form a complex during normal function with Retinoblastoma-associated proteins (RbAp48), whose deficiency or loss of function is related with AD [228].

Proteins of tubulin family appear in cluster 2 (nodes marked brown in Figure 3.5). The role of tubulin acetylation in AD has been investigated by several studies and it has been suggested that changes in the levels of this post-translational modification (PTM) are involved in the pathogenesis of AD [229–233]. However, whether this PTM has

detrimental or beneficial effects has not yet been clarified. Our study found out that tubulin β-2A encoder gene TUBB2A has been down-regulated in AD patients (see Figure 3.4), which supports the notion of a beneficial effect of tubulin β-2A to the pathology of AD. In addition, because the aggregation and hyper-phosphorylation of tau protein in degenerated neurons is a symbol of neurodegenerative disorder and tau protein is normally associated with microtubules (MTs), whether pathological changes of tubulin PTMs have disrupted the interactions of tau protein with MTs is a question worth to be addressed.

Of note, the basic helix-loop-helix (bHLH) family of transcription factors which has been shown to control critical aspects of development in many tissues emerge in cluster 4 (nodes marked blue in Figure 3.5). The reactivation of heart and neural crest derivatives expressed transcript 2 (HAND2) is reported to cause cardiac dilation and required for heart disease [234]. Transcription factors of NeuroD family is reported to be involved in glucotoxicity-induced beta cell dysfunction which is a critical factor in the development of type 2 diabetes [235]. NeuroD is also reported to increase the differentiation efficiency of mouse embryonic stem cells into insulin-producing cells [236], regulate neuronal migration [235] and regulate cell fate and neurite stratification in the developing retina [237]. The variety of tissue and functions with which these transcription factors are associated implies the possibility of links between diseases, for example, diabetes with AD, or heart failure with AD. More cross-disease investigations are needed to unveil the potential pathological relations.

Cluster 5 (nodes marked yellow in Figure 3.5) is a congregation of N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins which play crucial roles in synaptic vesicle trafficking and release, influencing the transportation of neurotransmitter. Syntaxin 1 (KOG0810 in Figure 3.5) is among a set of core neuronal SNARE molecules that directly mediate fusion of synaptic vesicles with the presynaptic membrane [238]. More evidences are revealing the connection between SNARE proteins and AD pathology. PICALM is a gene locus discovered by GWAS with affirming association with LOAD, appears to take part in directing the trafficking of SNARE protein

VAMP2 [207]. The t-SNARE complex deficit present in Lewy vody variant (LBV) is reported to be associated with the presence of LB-related pathology [239]. These findings cast light on the discovering new biomarker of AD in SNARE proteins.

In the process of testing the hypothesis that both training and testing performance of the model obtain improvements after including a combination of covariates as training and testing input, the p-values are calculated by a two-sample t-test. Yet, it is not a strict test in this case as we assumed without any further testing that the standard variances between the two samples were equalled.

The dataset GSE15222 itself cast some limitation on the study. Firstly, we only compared the difference between AD and control samples, however mild cognative impairment (MCI) was ignored. Secondly, some other quantitative measurement such as braak stage or amyloid level were not availble. Thirdly, no information about cell type was given in the meta information, so the possibility that all the observed difference was contributed from the variety of cell type cannot be eliminated. Lastly, the study only focus on biomarker in brain but the discovery of biomarkers in blood is the ultimate goal. Although it is pleasant to identify biomarkers in brain and validate in blood, it is also important to look into blood samples for discovery. As the validation of proposed biomarkers in blood suggests, changes in brain can be different from that in blood. Although that proves what happens in blood cannot mirror what is happening in the brain, they can still make a good biomarker.

Another limitation of the study is the lack of validation of the quantification method for APOE genotyping information. The numeric order of allele combinations assigned is solely determined by the previous conclusion drawn for single allele ($\epsilon 2$, $\epsilon 3$ or $\epsilon 4$). The effect of allele combinations is however not studied. In our study, more weight on AD risk was delegated to $\epsilon 4$. This partly contribute to the decision to assign greater numeric value for $\epsilon 2/\epsilon 4$ compared to $\epsilon 3/\epsilon 3$. But it is an assumption that has not been verified. Therefore, more validation on the AD risk of different allele combinations is needed.

3.6 Conclusion

Covariate addition of age, gender and APOE genotype are able to improve the performance of AD classifier. With a random forest based feature selection method, a small subset of genes are discovered to be highly informative and found to be enriched in SNP. The SVM model trained by selected genes has remarkable performance in distinguishing between case and control in the dataset. It is of high research value to continue validating the selected features in more AD datasets.

Chapter 4

Discovery of Novel Biomarkers for Alzheimer's Disease from Blood

4.1 Abstract

This chapter catalogues and discusses the investigation of blood-based AD biomarkers. It is based on the author's publication [240]. Following the construction of a knowledge feature pool under comprehensive knowledge collections, two novel SVM-based feature selection methods were proposed to select significant serum proteins with or without the constraint of feature pool, and a panel of only two and three proteins were identified to have good diagnostic ability. A surprising and novel statistical pattern was recognised.

4.2 Background

Although a number of genetic and cerebrospinal fluid (CSF) biomarkers for AD have been discovered in recent decades, few have been reported from the blood that have relevance to the disease [241]. There is thus a lack of robust and reliable blood-based biomarkers for AD diagnosis [34, 242]. With the expanding capacity of protein arrays and mass spectrometry-based detections, recent studies of blood profile biomarkers have attempted to address this problem. Ray and colleagues [56] were the first to use a profiling approach,

and they identified an 18-plasma protein profile that classified AD patients from healthy control subjects with high specificity. However, the result cannot be replicated. The same group later analyzed independent samples with different bioinformatics approaches and discovered that the majority of those 18 proteins were relevant to the levels of $A\beta$ or tau proteins in CSF [243]. Since these two studies, many profiling approaches have proposed protein panels with promising diagnostic ability, but the main issue has been reproducibility [58]. The problem of reproducibility has been addressed by Hu and colleagues [57] and Doecke and colleagues [59] using two well-characterized and large clinical cohorts to identify a series of inflammatory mediators associated with the onset of AD. Doecke and colleagues [59] and O'Bryant and colleagues [60] also reported diagnostic accuracy across cohorts (AUC=0.88, SN=0.75 and SP=0.91). In addition, researchers in plasma proteomics have used cross-validation across various cohorts to overcome the overfitting problem in high-dimensional studies. Molecules that have raised great hopes among these investigators include apolipoprotein E (APOE), NT-proBNP (N-terminal prohormone of Brain Natriuretic Peptide) and pancreatic polypeptide. It is been suggested that because AD is a mitochondrial dysfunction and immune system relevant disease [244, 245] focusing on genes involved in relevant pathways [246] may help in biomarker discovery [247]. However, few previous studies have used biological information in their modelling.

In the past decade, neurofilament light protein (NFL) is the only CSF biomarker for which the transition from CSF to blood has been relatively uncomplicated, but for tau and $A\beta$biomarkers, there is a signal also in blood, albeit with a smaller effect size than what can be obtained using the corresponding CSF measure.

To establish reliable blood biomarkers for $A\beta$ pathology in AD has always been difficult. $A\beta$ proteins can be measured in plasma yet historically its correlation with AD and/or cerebral β-amyloidosis has been weak or even absent [248]. Plasma $A\beta$ concentrations have been described as potentially affected by yield in platelets and other extra-cerebral tissues and the measurements have been complicated by matrix effects from plasma proteins [249]. Nonetheless, recent mass spectrometric studies indicates

that a ratio of a certain amyloid-β protein precursor (AβPP) fragment (AβPP669-711; an Aβ peptide that extends over the BACE1 cleavage site of AβPP with 3 amino acids), to Aβ42 or Aβ42/Aβ40 identifies individuals cerebral β-amyloidosis with excellent sensitivity and specificity [250,251]. Pilot data suggest associations of the concentrations of a number of plasma proteins (e.g., pancreatic polypeptide Y, IgM, chemokine ligand 13, interleukin 17, vascular cell adhesion protein 1, α2-macroglobulin, apolipoprotein A1, and complement proteins) with amyloid burden in the brain [252–254]. However, these data ought to be interpreted cautiously, because they are derived from multi-marker panels and as a mechanistic understanding of the associations is currently lacking.

CSF assays for T-tau and NFL were recently developed into ultrasensitive blood tests using Simoa technology [255]. Plasma or serum NFL concentration is associated with CSF and most CSF findings have been reproduced in blood [256]. Recent data suggest that serum NFL effectively signals onset of neurodegeneration in FAD [257] and Huntington's disease (HD) [258]. Plasma NFL concentration elevates in patients with Charcot-Marie-Tooth disease and correlates with disease severity, showing that peripheral nervesmay also release NFL [259].

For tau, the situation is promising but less clear. The correlation with the corresponding CSF concentration is absent [54] or weak [260]. Plasma T-tau concentration in AD is increased but the effect size is smaller than in CSF and there is no detectable increase in the MCI [**?**,54]. Altogether, the published studies on plasma T-tau as an AD biomarker so far acknowledged the feasibility of targeting on a predictive tau signal in blood. But the lack of correlation of plasma with CSF T-tau implies that investigation should shift to additional tau biomarkers in plasma.

In this chapter, existing biological knowledge of potential AD biomarkers is taken into consideration to construct a knowledge feature pool for a series of feature selection methods. We first established a feature pool comprising of numerous AD related biomarkers and then designed two novel SVM-based feature selection methods, which we used to select several panels of biomarkers. Finally, we validated the classifying performance of these panels with other serum and RNA expression cohorts. We found that

a panel of only two or three proteins gave us good diagnostic ability.

In this chapter, our main aims are the followings:

1. To develop feature selection methods that include the consideration of feature-feature interation within the dicision making

2. To search for proteins that have good classifying ability between healthy individuals and AD patients, and build prediction models based on those proteins

3. To investigate the biological linkage between selected features and AD

4.3 Materials and Methods

4.3.1 Data collection and pre-processing

We searched against Gene Expression Omnibus (GEO) (`http://www.ncbi.nlm.nih.gov/geo/`) for datasets with entries containing keywords "protein" and "blood" (search conducted in June 2015). Three datasets were chosen among the returned datasets: GSE29676, GSE39087 and GSE5281. GSE29676 consists of serum samples from 50 AD, 40 non-demented controls (NDC), 30 breast cancer (BC) and 29 Parkinson's disease (PD). The data were generated by a Invitrogen ProtoArray v5.0 protein platform including 9486 unique human protein antigens [261]. GSE39087 is also a human serum protein microarray dataset generated by the same platform as GSE29676 and contains 36 AD cases, 57 controls, 48 Parkinson disease, 18 breast cancers, and 7 multiple sclerosis [262]. GSE5281 is an RNA microarray dataset from brain tissues, with 87 AD cases and 74 controls. Each sample was collected from different brain regions comprising entorhinal cortex (EC), hippocampus (HIP), medial temporal gyrus MTG), posterior cingulate (PC), superior frontal gyrus (SFG), and primary visual cortex (PVC) [263].

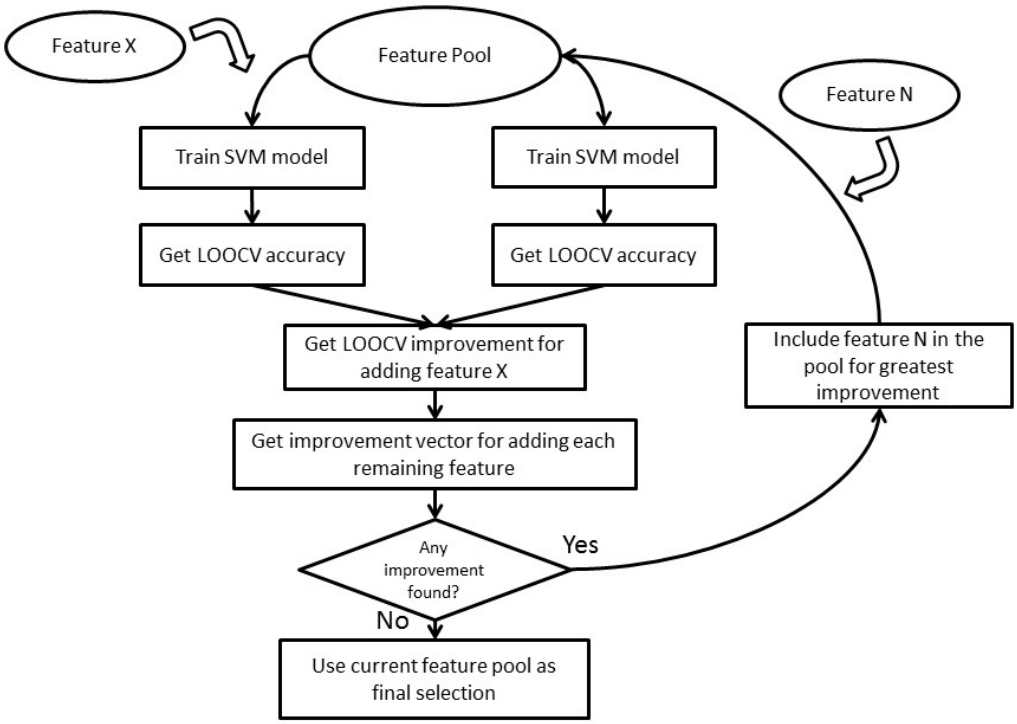


Figure 4.1: Workflow of Support Vector Machine Forward Selection (SVMFS). The selection starts with an empty aggregate of selected protein. In each iteration, one particular protein which yields the most improvement of the LOOCV accuracy of the model is selected. If no improvement is found in all the candidate proteins, the iteration breaks and no more proteins are added into the model, so the current model will be the final selection.

The normalized expression data of GSE29676 and GSE39087 were downloaded directly, then expression values smaller than 1 were set as 1 and 2-based logarithm transformation was conducted. To eliminate the potential bias caused by age and gender, the expression value was corrected using the following method. First, for each protein, a robust linear regression (rlm function in MASS R package [264]) was applied with the logarithm transformed expression value as the dependent variable and age and gender as the explanatory variables. Second, the sum of the intercept and residual was employed as the corrected expression value for that protein in each sample and used in subsequent analyses. For GSE5281, an age-gender-bias correction was also conducted on the normalized data before matching the probes with corresponding proteins. We used GSE29676

as the discovery dataset for biomarker identification, and GSE39087 and GSE5281 as the two validation datasets. Only AD and Control subjects were included in any subsequent analysis.

4.3.2 Feature candidate pool

In this study, we are going to investigate whether a feature selection under the context of existing biological knowledge will improve the resulta and yield better classifying models. Therefore, for comparison, we have two feature pools that the features can be selected fromL: **whole-range feature pool** and **knowledge-based feature pool**

4.3.2.1 Whole-range feature pool (WRFP)

As described in section 4.3.1, the dataset includes 9486 unique human protein antigens, and this is dubbed throughout the study as **whole-range feature pool** (WRFP).

4.3.2.2 Knowledge-based feature pool (KBFP)

In contrast with WRFP which includes all the available features within the dataset, a **knowledge-based feature pool** is a protein subset of WRFP with more biological meanings realting to AD. To collect all the AD-related genes or protein, we comprehensively searched the literature and online databases to construct a knowledge feature pool for the AD-related biomarkers. The text mining for AD biomarkers was conducted by searching publications on PubMed in December 2014 (`http://www.ncbi.nlm.nih.gov/pubmed`), producing a set of 611 genes. 172 genes were discovered from (a) large genome wide association study (GWAS) papers [83–85] and their first neighbors in a protein-protein interaction (PPI) network [265], and (b) AD-related genes and protein database in Alzforum (`http://www.alzforum.org/`). We collected 84 genes from a human AD real-time PCR array functional gene grouping (`http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-057Z.html`) and 876 genes in the Ingenuity Pathway Analysis (`http://www.ingenuity.com/`) tool filtered by keyword

"Alzheimer biomarker". From these searches a total of 1915 unique genes were placed in our knowledge-based gene pool.

4.3.3 Feature selection

We proposed a novel method – Support Vector Machine Forward Selection (SVMFS) – for selecting the best AD-related protein set for training our classification model (Figure 4.1) The framework of our method is built upon that of the Support Vector Machine (SVM) model. Throughout the study, we adopted the default settings of the SVM model in the e1071 R package [266] (gamma=1/feature number, cost=1, type=C-classification, kernel=radial). For a given protein set, an SVM model can be trained, whose leave-one-out cross validation (LOOCV) accuracy was then used as the evaluation score. The evaluation score improvement was calculated by comparing the evaluation scores of the previous protein set with the evaluation score of the updated protein set containing a selected additional protein. An alternative feature selection method was also used – SVM Top Forward Selection (SVMTFS). In this alternate method a ranking list for all the proteins based on the LOOCV accuracy of their respective single-protein SVM model was made.

The only difference between these two methods lies in the selection of the protein to be included in the next round. In SVMFS, the optimal protein among the rest is selected; in SVMTFS, the next protein in the ranking list based on the LOOCV is selected.

The abovementioned methods are applied to both WRFP and KBFP.

4.3.4 Classifier training and assessing

We conducted a cross-validation using the GSE29676 dataset on the protein sets discovered by our novel feature selection approach, and the 10 biomarkers discovered by Nagele and colleagues (named here as Nagele model) [261]. We trained classifiers with 60 samples (randomly selected, 30 each in AD and healthy samples) and then testing the classifiers with the remaining samples (10 AD and 20 healthy samples). The cross-

validation was repeated 5000 times for the calculation of average sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false discovery rate (FDR) and false omission rate (FOR) [267]. The ROC curve performance area under the curve (AUC) was plotted using the pROC R package [268].

4.3.5 Biomarker validation

We conducted both biomarker validation and classification model validation. Classifiers trained in discovery in the GSE29676 dataset were then tested for performance in the GSE39087 dataset. We also did cross-validation using GSE39087 of the features identified by GSE29676, i.e. randomly selecting 20 AD and 20 healthy in GSE39087 as training samples and the remainder as testing samples, and repeating 5000 times.

For GSE5281, target proteins identified in discovery dataset were matched with corresponding probes by their corresponding genes. An SVM classification model was built in the six different brain regions separately and LOOCV accuracy was used to assess the performance of the model in each region.

4.4 Result

Employing SVMFS and SVMTFS to select features in the KBFP and full feature pool respectively, we discovered three different protein sets that showed promising performance in discriminating AD patients from healthy individual as measured by LOOCV accuracy. Table 4.1 shows the LOOCV accuracy for each of the top 20 features (proteins) used in a single feature SVM model in discovery dataset. We found the following models (protein sets) whose classifying performance is shown in Table 4.2.

- A two-feature model selected in WRFP by SVMFS had 98.8% SVM-LOOCV accuracy and consisted of ECH1+NHLRC2 (enoyl coenzyme A hydratase 1 peroxisomal plus NHL repeat containing 2).
- A three-feature model selected in KBFP by SVMFS had 96.5% SVM-LOOCV ac-

Table 4.1: Top 20 proteins with largest LOOCV accuracy

NCBI Accession ID	Protein Name	LOOCV Accuracy
BC011792.1	ECH1	96.5%
NM_004502.2	HOXB7	96.5%
NM_177924.1	ASAH1	96.5%
BC030814.1	IGKV1-5	95.4%
BC034142.1	IGKV1-5	95.4%
BC034146.1	IGKV1-5	95.4%
BC034937.1	C10orf64	95.4%
NM_176884.1	TAS2R43	95.4%
PV3366	ERBB2	94.2%
NM_201278.1	MTMR2	94.2%
BC038406.1	C3orf20	94.2%
NM_152776.1	MGC40579	94.2%
NM_014110.3	PPP1R8	93.0%
XM_294794.1	LOC339065	93.0%
NM_019891.1	ERO1LB	93.0%
BC068078.1	NPM2	93.0%
NM_002613.3	PDPK1	93.0%
NM_031268.3	PDPK1	93.0%
BC032101.1	JAGN1	93.0%
NM_000963.1	PTGS2	93.0%

curacy and consisted of ERBB2+FN1+SLC6A13 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian) transcript variant 2, fibronectin 1, plus solute carrier family 6 (neurotransmitter transporter, GABA), member 13).

- A two-feature model selected in WRFP by SVMTFS had 97.7% SVM-LOOCV accuracy and consisted of ECH1+HOXB7 (homeobox B7).

Evaluation by cross-validation in the same dataset showed a good performance of these models (Table 4.2). The average sensitivity and specificity of models ECH1+NHLRC2, ECH1+HOXB7, and ERBB2+FN1+SLC6A13 all reached at least 88%. Among the selected proteins, an interesting statistical pattern for the expression level was discovered in ECH1, HOXB7 and ERBB2 (Figures 4.5 and 4.6). In each of these three proteins, the normal expression range has two thresholds (one upper limit and one lower limit). To the best of our knowledge, such biomarkers with banded distributions between healthy and AD samples have not previously been reported. Typically there is a binary separation between AD and healthy samples with only one threshold.

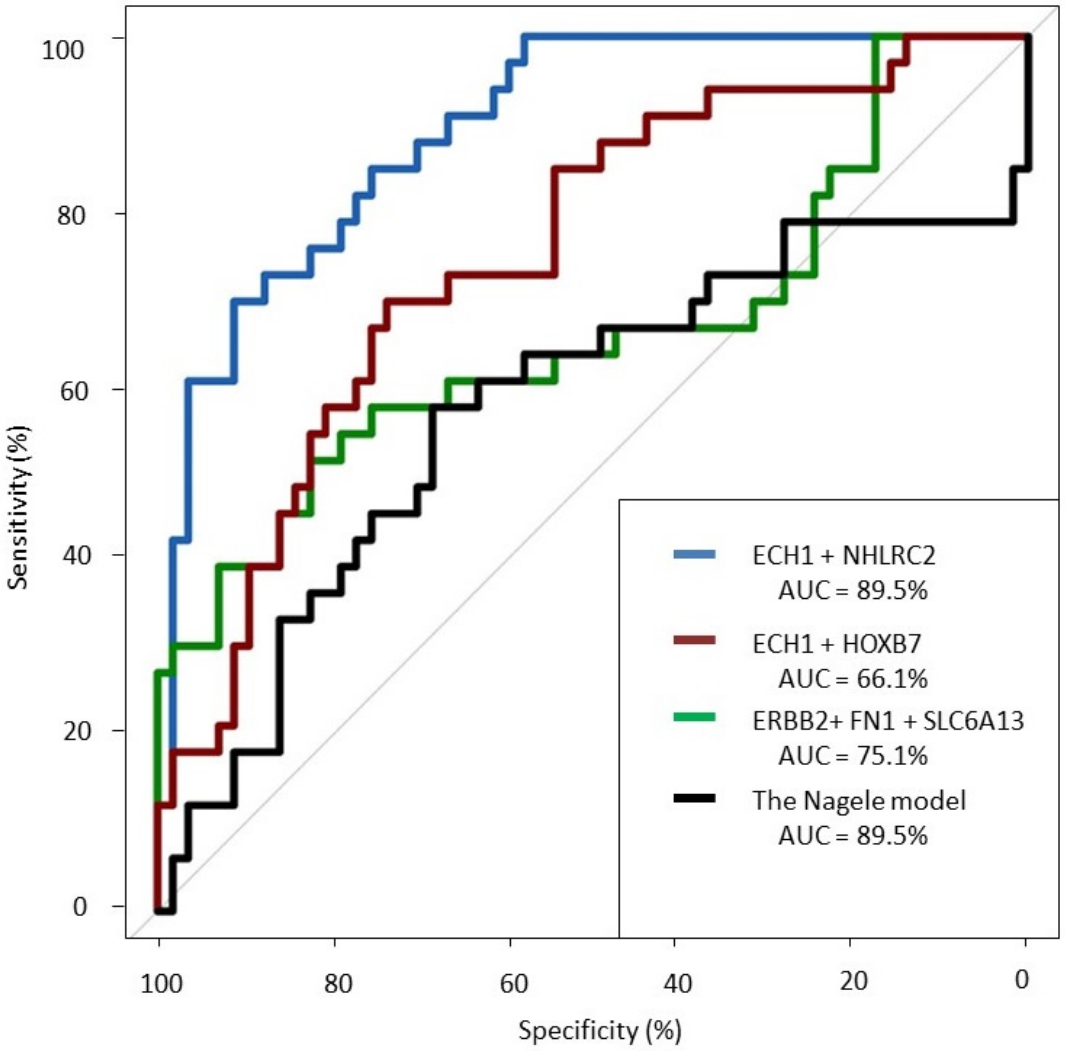


Figure 4.2: ROC curves of the three proposed models in the cross-cohort validation using GSE39087. The curves were plotted with the best performed classifiers across all models during the validation.

4.4.1 Cross-cohort validation

The cross-validation using cohort GSE39087, which is also a serum protein microarray data, showed that the three models still maintained good classification ability, with SVM-LOOCV accuracies of 88.9% (ECH1+NHLRC2), 97.8% (ECH1+HOXB7), and 74.4% (ERBB2+FN1+SLC6A13). Model ECH1+HOXB7 outperformed the others in this process of validation, with over 95% in sensitivity and specificity (Table

4.2). Model ECH1+NHLRC2 also exhibited good predictive performance except for a decreased sensitivity, which could result from the relatively small training sample size. Despite the seemingly good result in cross-validation using GSE39087, the performances of models deteriorated when they were trained and tested by different cohorts (AUC: ECH1+NHLRC2: 89.5%, ECH1+HOXB7: 66.1%, ERBB2+FN1+SLC6A13: 75.1%, see Figure 4.2). This could be an indication of over-training in those models, especially for model ECH1+HOXB7. The reason for this could be different experimental environments between the two cohorts.

We also investigated the distribution pattern for all the proteins using dataset GSE39087 and found that ECH1 still maintained its banded distribution, while in ERBB2 and HOXB7 more spread of control is found (Figure 4.7 & Figure 4.8). The disparity may be caused by the different data processing methods employed by GSE29676 and GSE39087; the former dataset were characterized into disease and control groups and then linearly normalized while the latter dataset were normalized via the compare-function embedded in Invitrogen's Prospector [269].

We conducted LOOCV separately for our three proposed protein sets in the six different brain regions of dataset GSE5281 (thus 18 models were evaluated in total). The result shows that our three models maintain excellent overall classification ability in EC but are poorer in the others (Table 4.3).

Table 4.2: Performances of three proposed models in dataset GSE29676 and GSE39087.

	Average LOOCV accuracy	Validation Accuracy	Sensitivity	Specificity	NPV	PPV	FDR	FOR
Cross-validation in GSE29676								
ECH1 + NHLRC2	98.8%	95.4%	94.0%	97.1%	93.0%	97.7%	2.3%	7.0%
ECH1 + HOXB7	97.7%	95.6%	95.0%	96.3%	94.1%	97.1%	2.9%	5.9%
ERBB2 + FN1 + SLC6A13	96.5%	89.6%	87.9%	91.8%	86.4%	93.4%	6.6%	13.6%
Nagele model	64.0%	56.8%	58.9%	54.3%	51.8%	62.2%	37.8%	48.2%
Cross-validation in GSE39087								
ECH1 + NHLRC2	88.9%	87.4%	79.7%	90.8%	91.2%	80.7%	19.4%	8.8%
ECH1 + HOXB7	97.8%	96.9%	96.8%	96.9%	98.6%	93.8%	6.2%	1.4%
ERBB2 + FN1 + SLC6A13	74.4%	69.8%	80.9%	65.0%	89.4%	51.4%	48.6%	10.6%
Nagele model	70.0%	69.4%	80.2%	64.7%	89.0%	50.6%	49.4%	11.0%

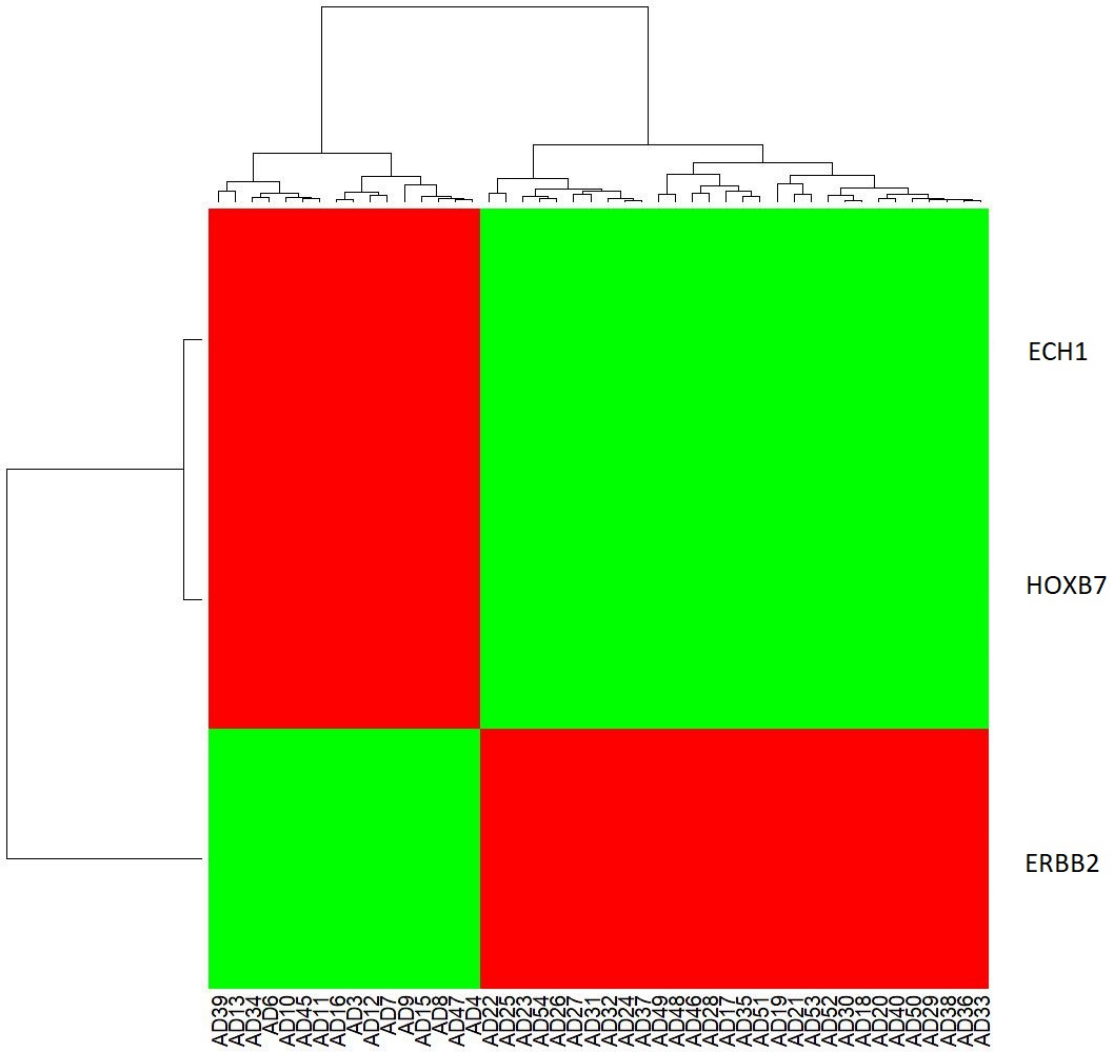


Figure 4.3: Up/down regulation for ECH1, HOXB7 and ERBB2 in GSE29676. Horizontal coordinate represent AD samples and vertical cooridinate represents three proteins. Expression values are scaled by rows for each protein, and red/green indicate relatively up/down regulated expression.

Table 4.3: Accuracy performances of our three proposed models in dataset GSE5281 (see Methods for full name of brain regions)

	EC	HIP	MTG	PC	SFG	VCX
ECH1 + NHLRC2	95.5%	78.3%	60.0%	57.1%	68.0%	50.0%
ECH1 + HOXB7	86.4%	87.0%	80.0%	85.7%	68.0%	40.0%
ERBB2 + FN1 + SLC6A13	90.9%	56.5%	88.0%	81.0%	60.0%	43.3%

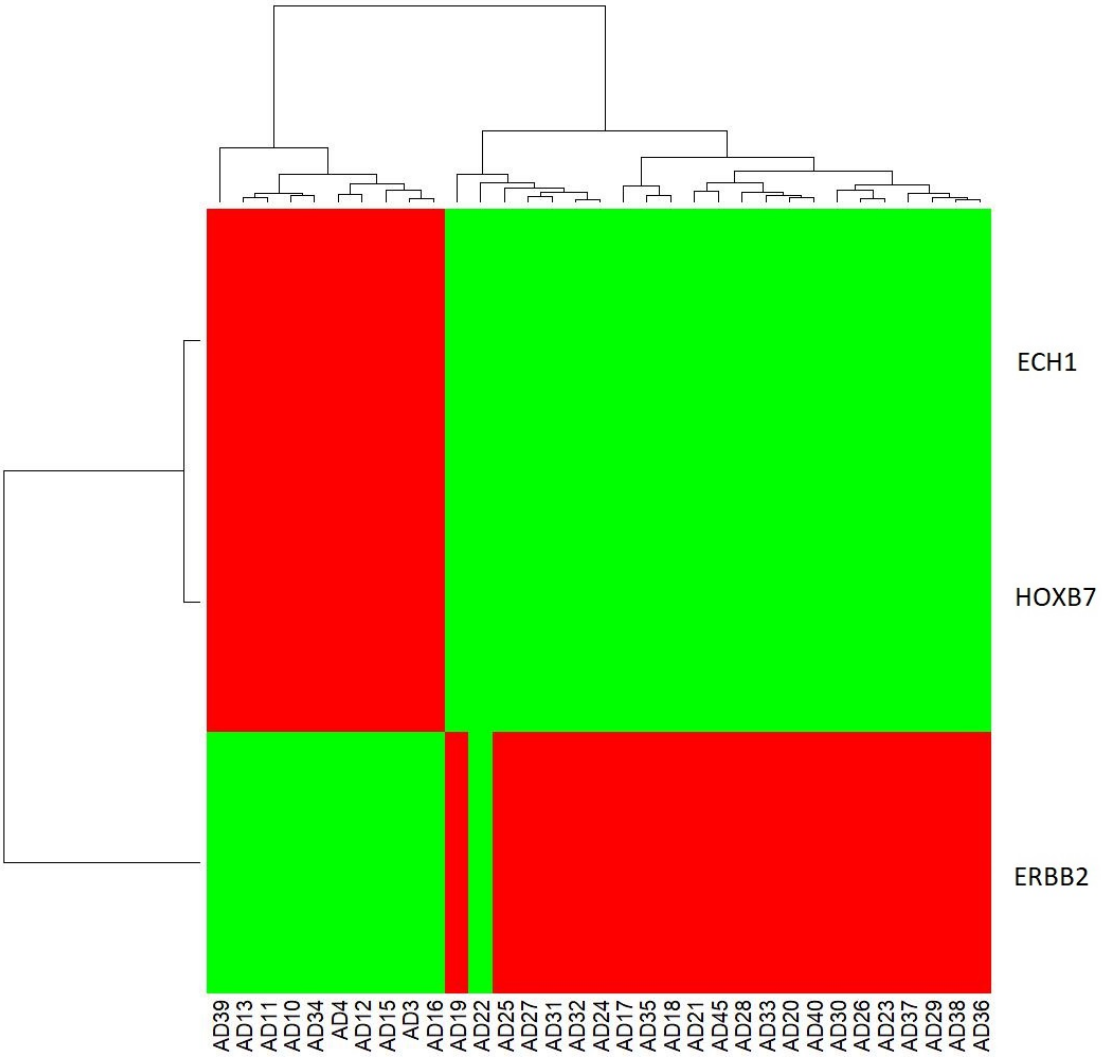


Figure 4.4: Up/down regulation for ECH1, HOXB7 and ERBB2 in GSE39087. Horizontal coordinate represent AD samples and vertical cooridinate represents three proteins. Expression values are scaled by rows for each protein, and red/green indicate relatively up/down regulated expression.

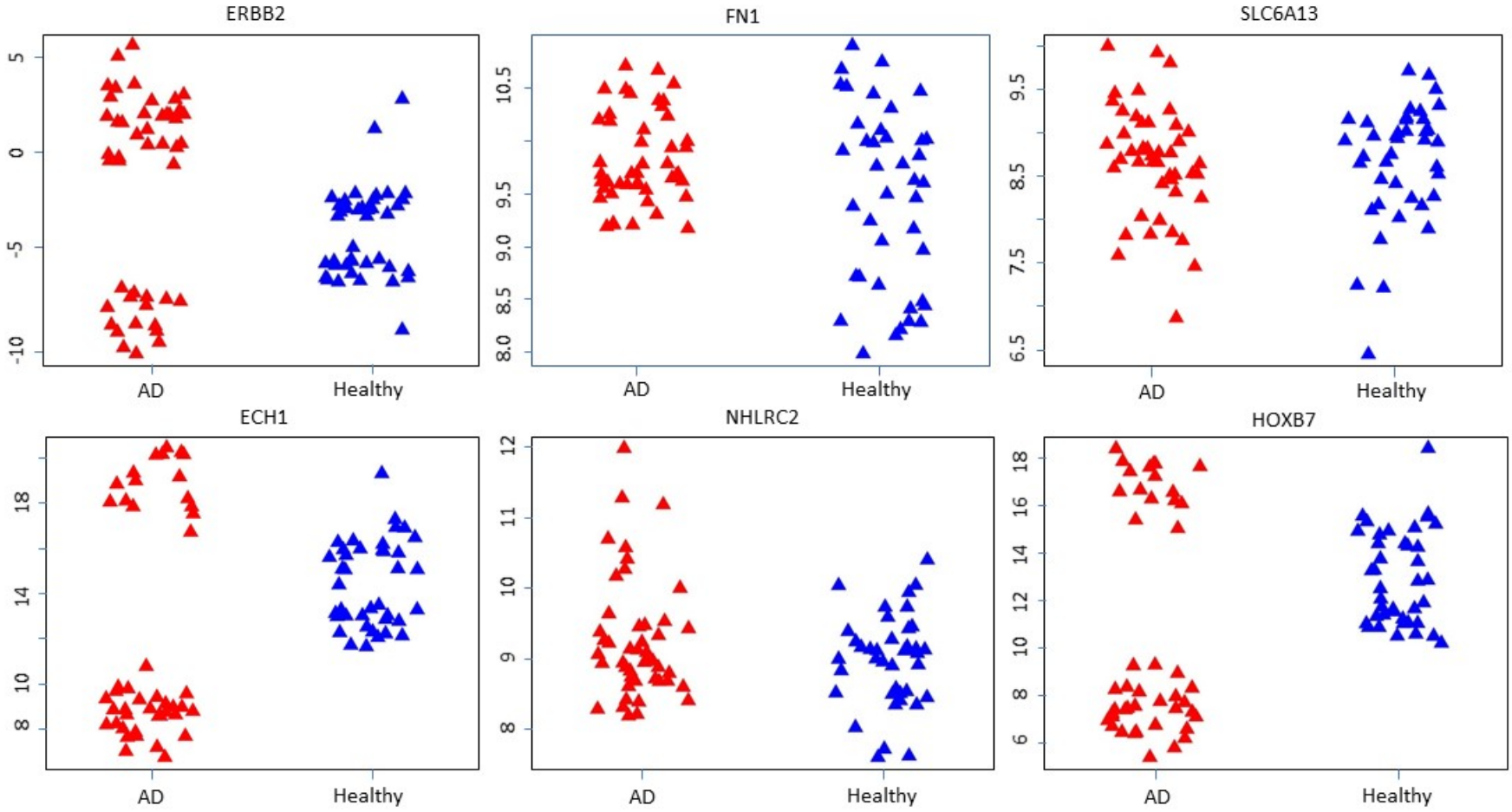


Figure 4.5: Expression level of six proteins in three proposed models under different conditions (red for AD samples, blue for healthy controls) in dataset GSE29676. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

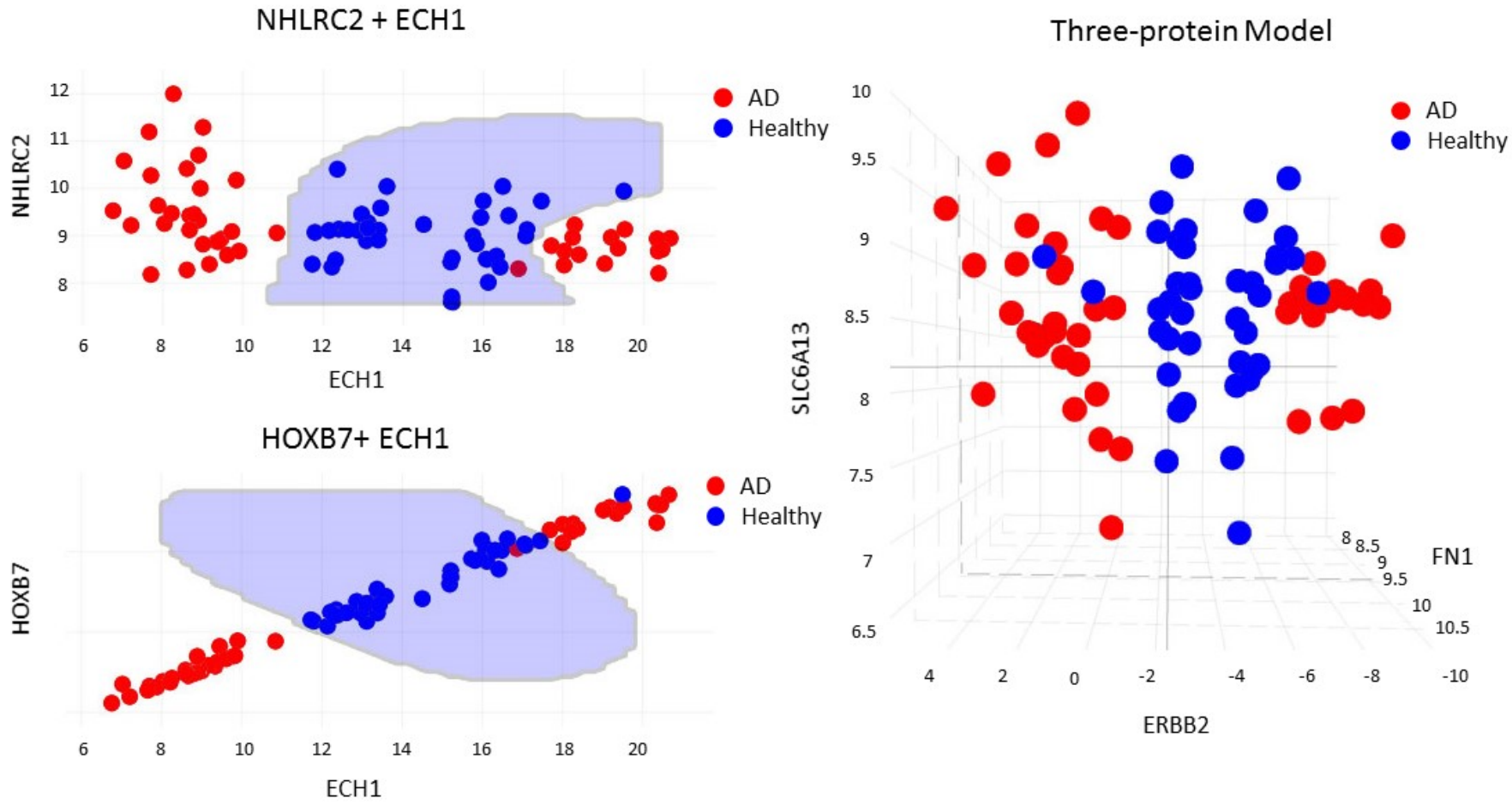


Figure 4.6: Three proposed models in dataset GSE29676. Blue shaded area indicates where a sample will be classified as healthy by the prediction model. Coordinates in each plot represents the processed expression value. Red represents AD samples and blue represents healthy control samples

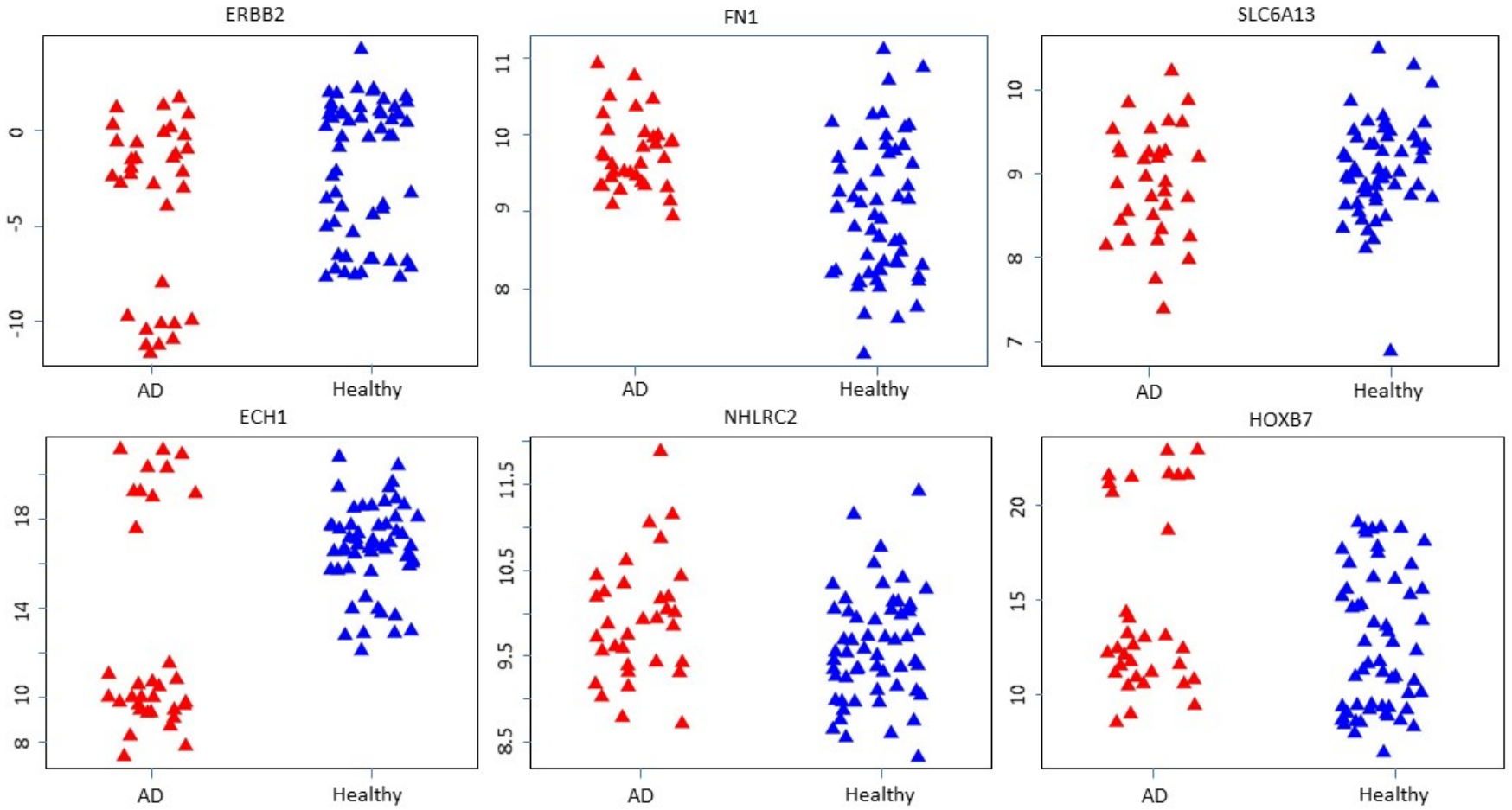


Figure 4.7: Expression level of six proteins in three proposed models under different conditions (red for AD samples, blue for healthy controls) in dataset GSE39087

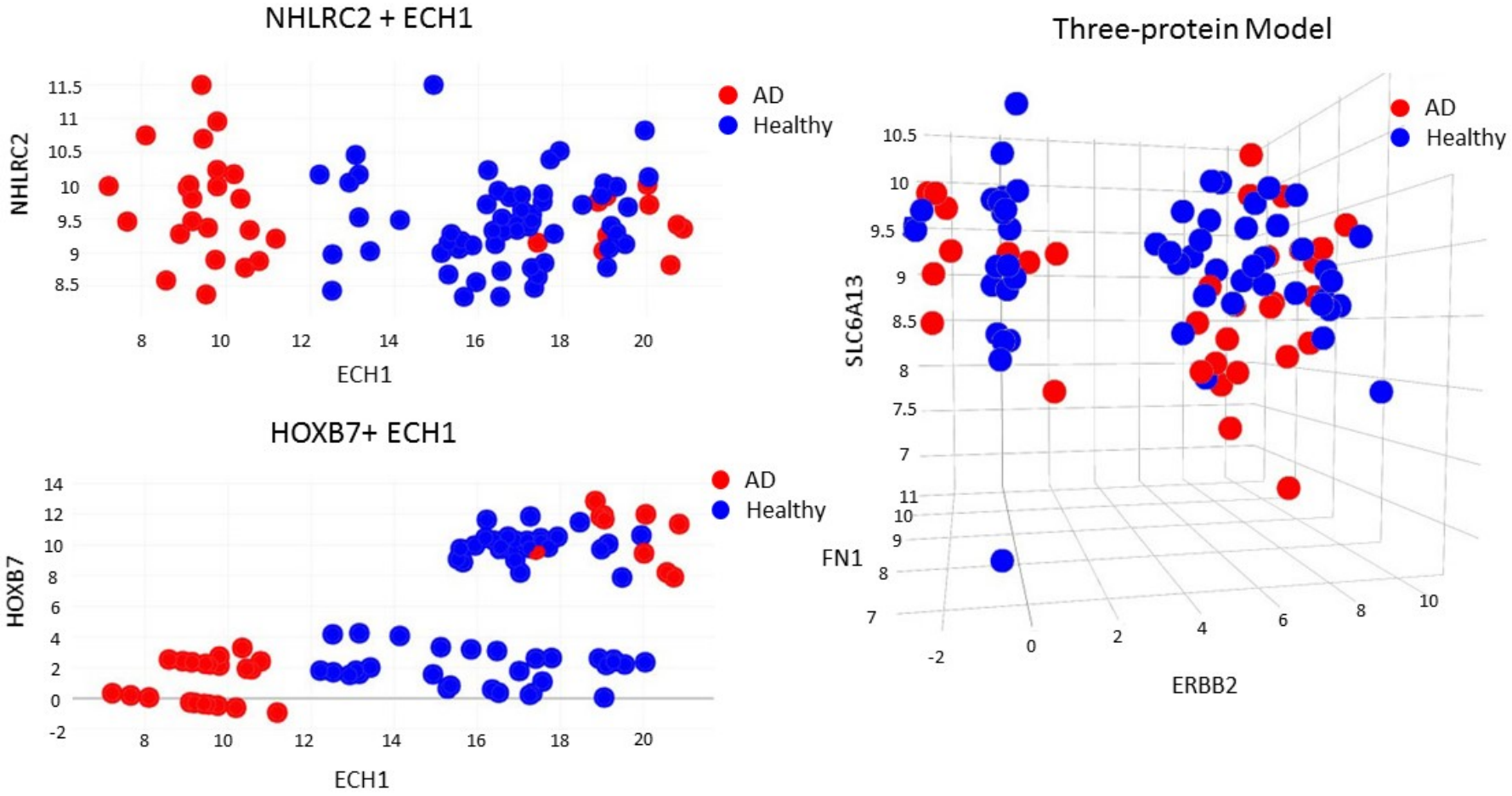


Figure 4.8: Three proposed models in dataset GSE39087

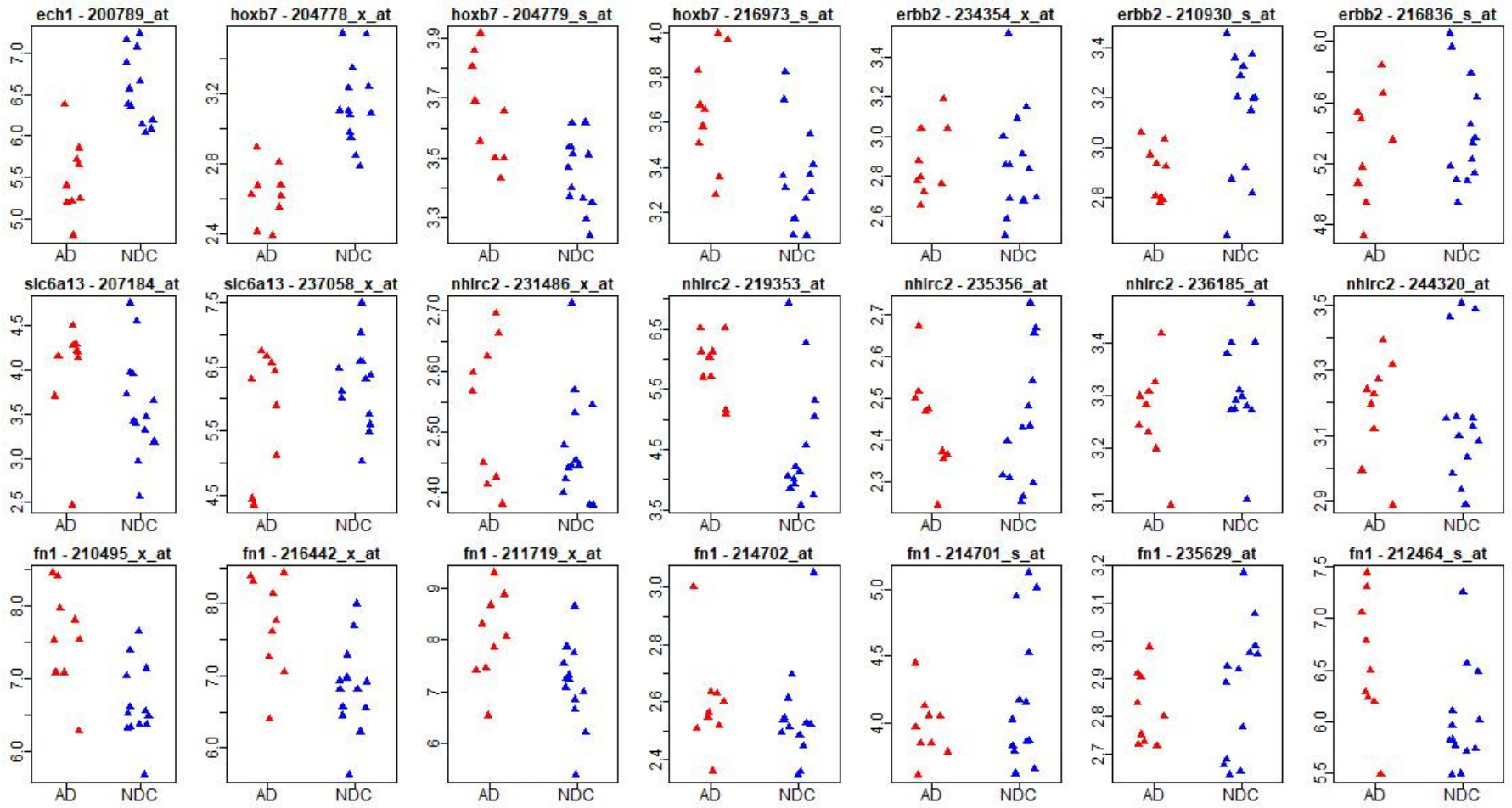


Figure 4.9: Expression level of gene probes mapping with six proteins from brain region EC in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

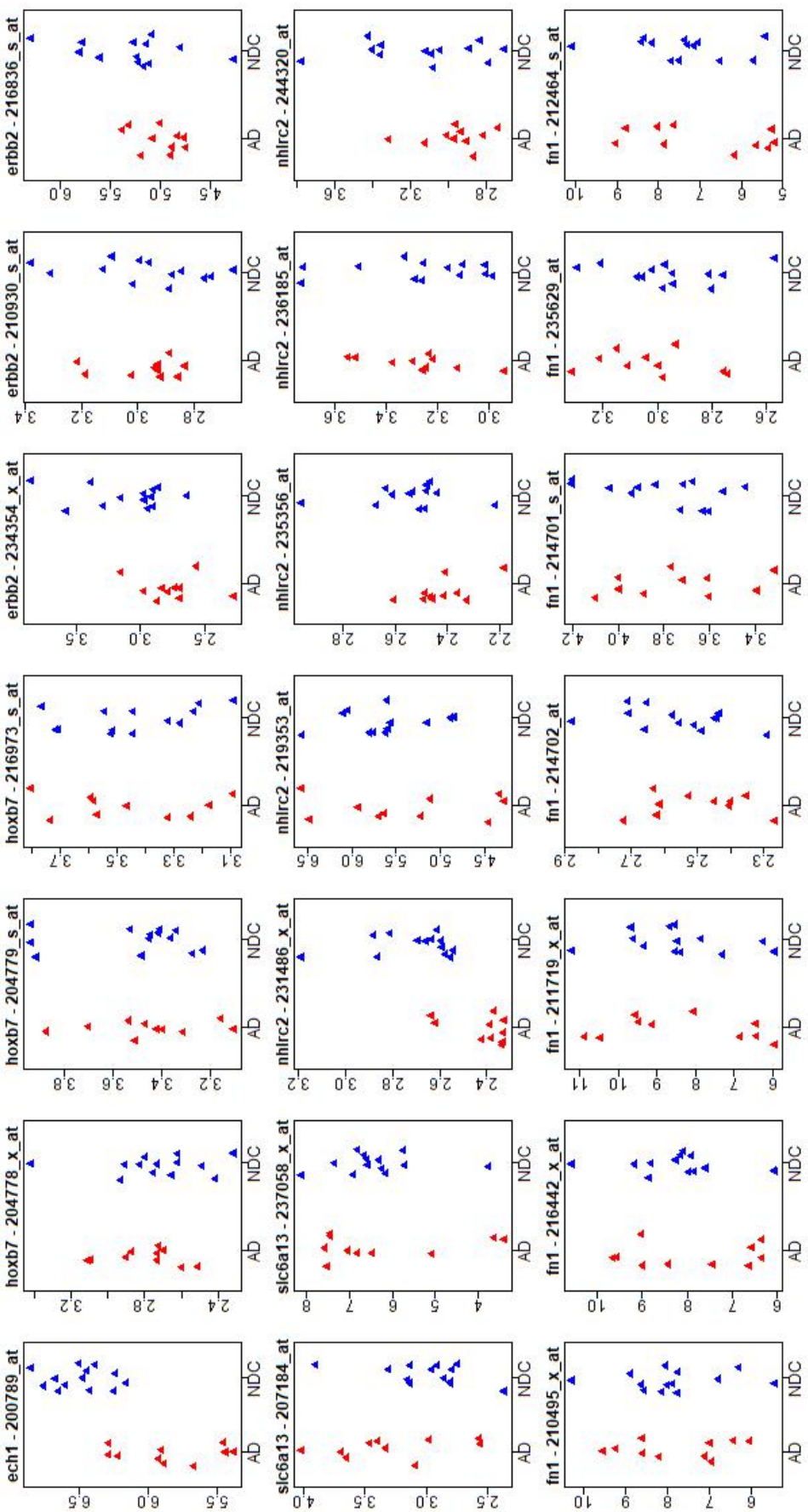


Figure 4.10: Expression level of gene probes mapping with six proteins from brain region HI in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

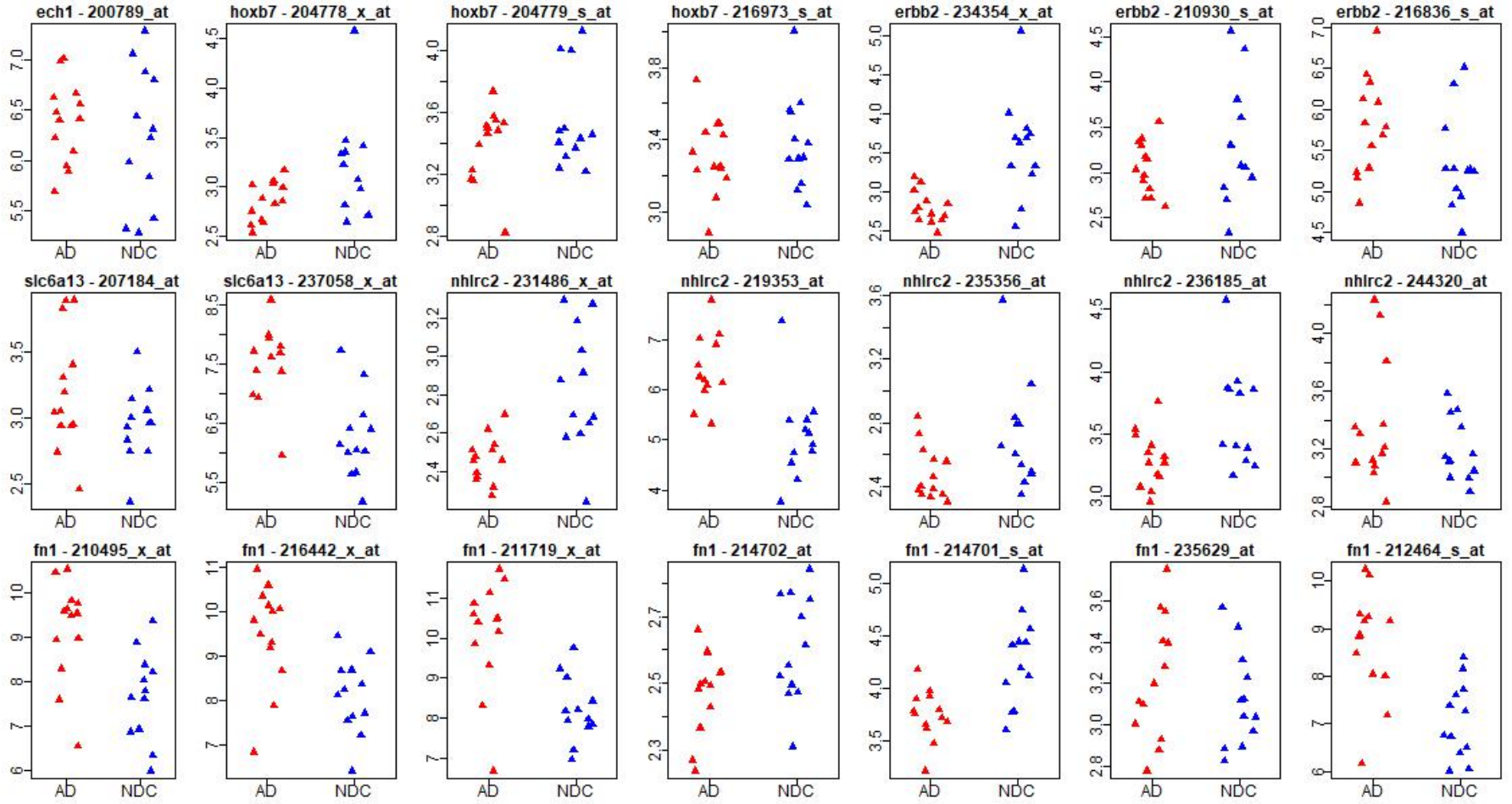


Figure 4.11: Expression level of gene probes mapping with six proteins from brain region MTG in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

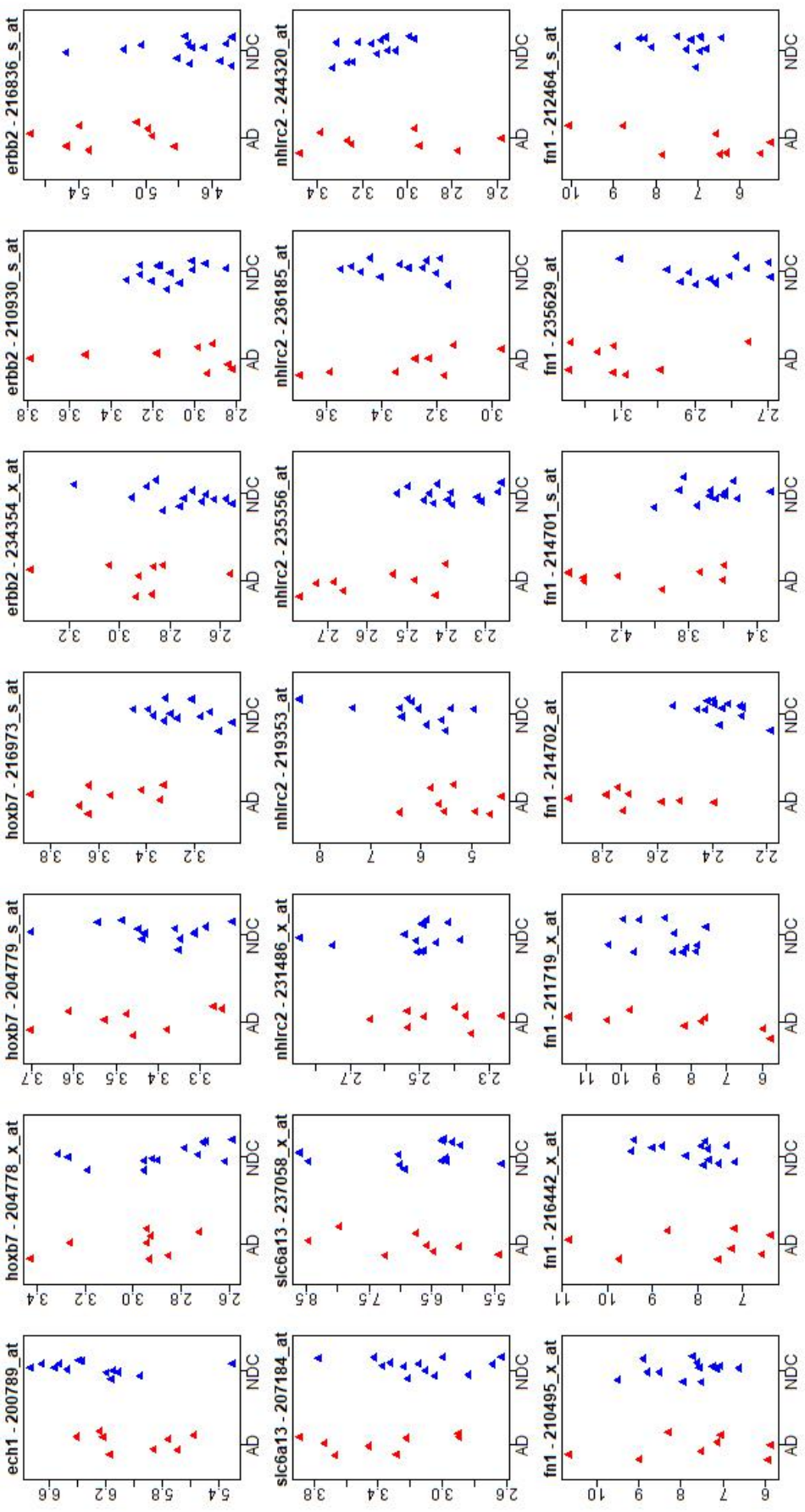


Figure 4.12: Expression level of gene probes mapping with six proteins from brain region PS in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

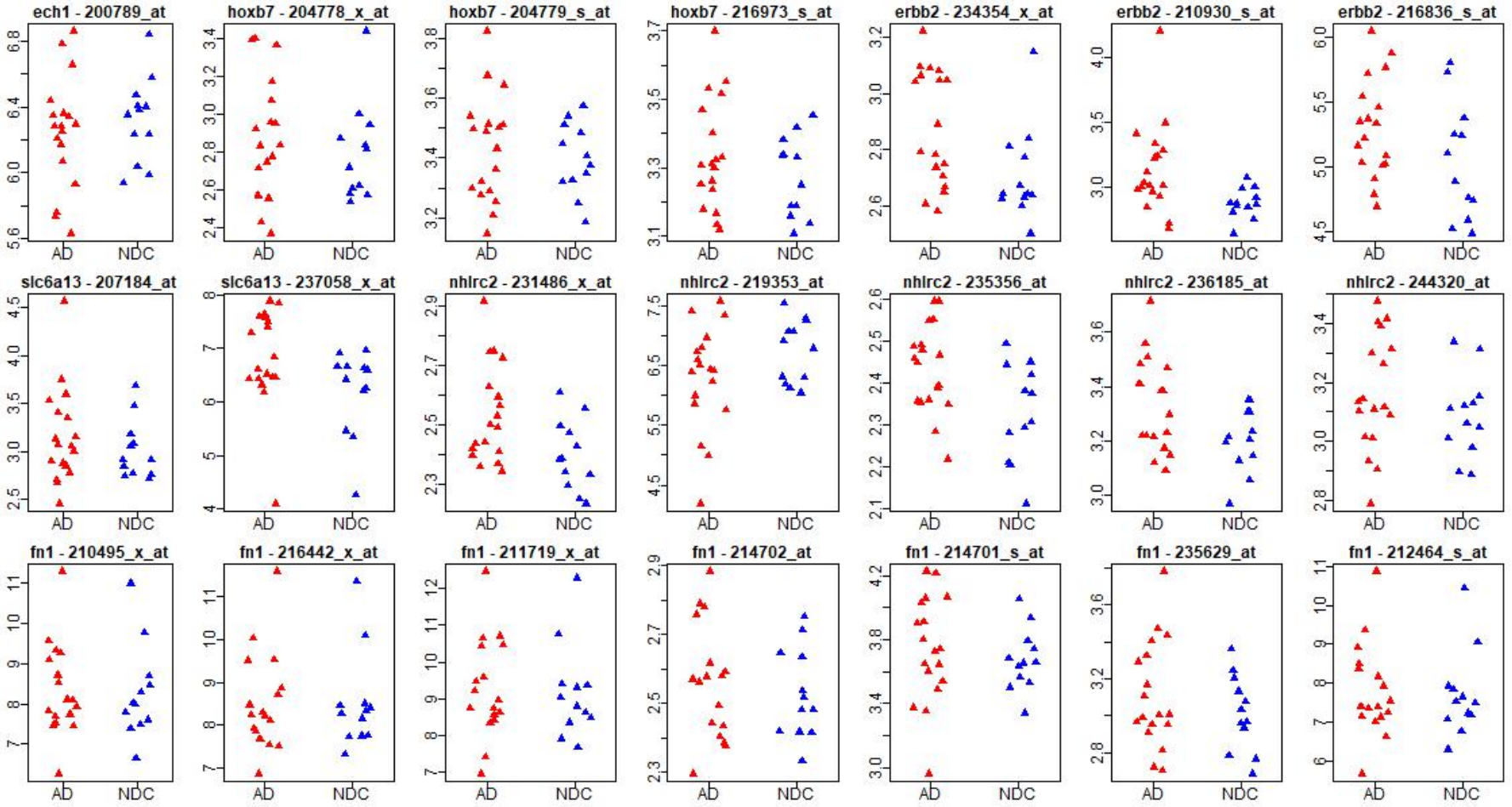


Figure 4.13: Expression level of gene probes mapping with six proteins from brain region PVC in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

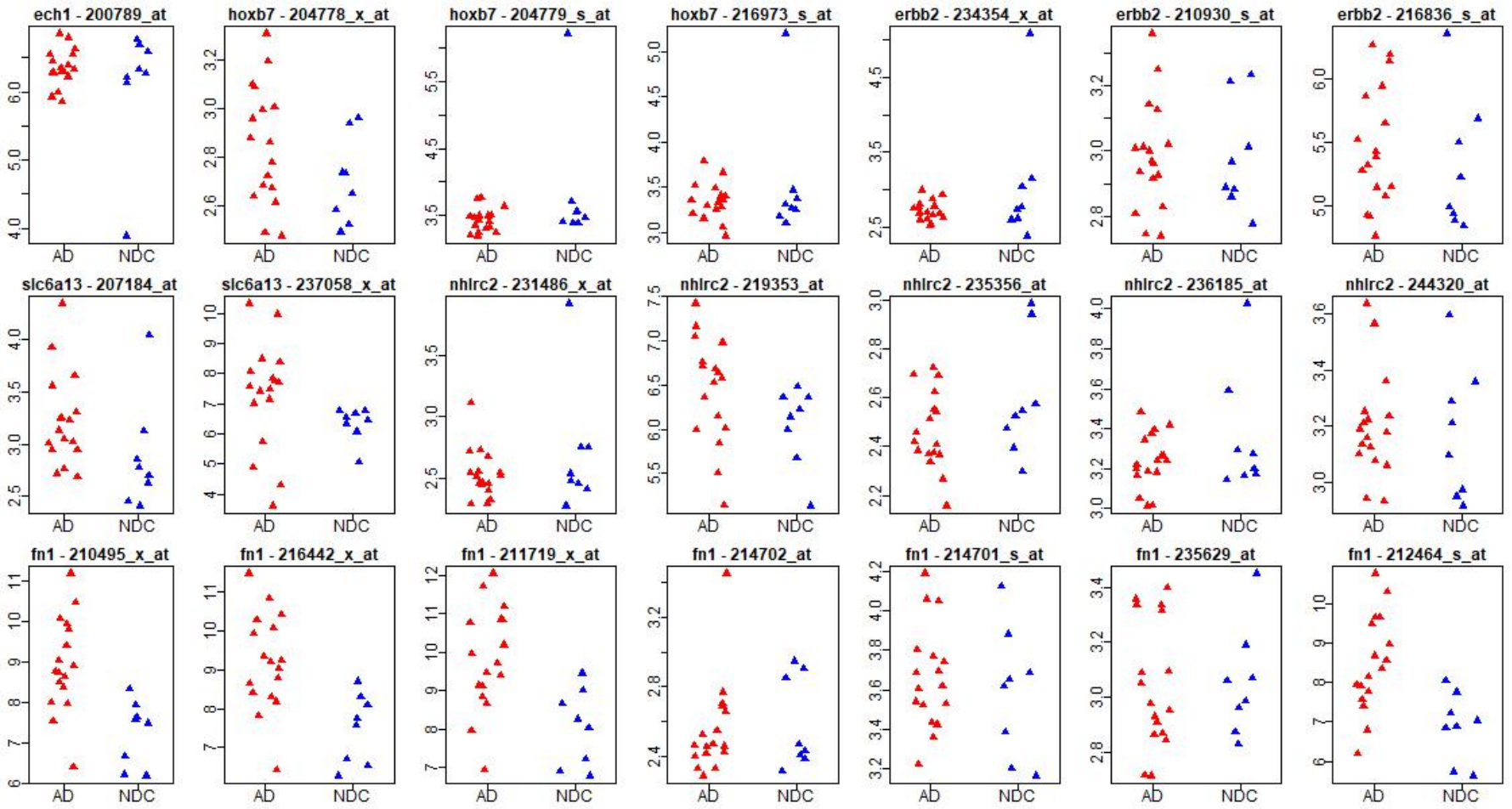


Figure 4.14: Expression level of gene probes mapping with six proteins from brain region SFG in GSE5281. For each subchart, the first part of the title separated by hyphen suggests the gene name, whereas the later suggest the gene probe ID. Red indicates AD samples and blue for healthy controls. The vertical coordinate of each plot represents the processed expression value, the horizontal coordinate represents different sample categories

4.5 Discussion

The original study of dataset GSE29676 reported 10 autoantibodies as diagnostic AD biomarkers [261]. The authors constructed a descending ranked list sorted by the difference in prevalence between AD and healthy groups using *Predictive Analysis for Microarrays (PAM)*, and then the top 10 features were selected. This method of feature selection did not take the combinatory effect of feature sets into consideration, as each autoantibody was selected exclusively according to their own discriminant ability between groups. To overcome the weaknesses in feature selection, we used a SVM radial kernel embedded feature selection method, which not only compensates for the ignorance of combinatory effect of significant differentiator feature sets, but also adds the ability to discover complex patterns in the data. More importantly, in the original study the predictive models were trained and validated in samples that were randomly selected, only once, which may lead to uncertain results. In contrast, our study has cross-validation by repeating the sampling for 5000 times to compensate for any uncertainty in bootstrapping. Also, the impact of age and gender on the prediction models were ignored in Nagele's study. In a later re-examination by the same researchers, those two factors (age and gender) were identified to strongly influence the number of autoantibodies detected using protein microarrays [262]. We eliminated such effects by simulating a robust linear regression model between age, gender and the expression value. The expression value was then corrected by summing the intercept and the residue.

We also see a potentially novel pattern of expression in AD and healthy samples with two boundaries. An assumption can be made that there is a normal level of protein expression in healthy individuals. The LOOCV accuracies of those proteins with this particular pattern suggest that any subject with an abnormal expression level, either being up or down regulated, can be diagnosed as having AD with high confidence.

The existence of upper and lower bound of normal expression in these proteins can have multiple possible explainations. Firstly, it can imply the potential to subdivide AD into two or more categories, where the samples with increased expression level affiliate

with one category while those with less expression level affiliate with the other.

A second and more likely explaination to the banded distribution of AD samples, is that there are regions of potential protein quantitative trait locus (pQTL) that varies in degree, attributing to the pathology of AD.

The sample expression distribution of the corresponding genes to the six proteins are also examined, the results of which are shown in Figure **??**. The figures show that the banded pattern disappear in probes mapped to ECH1, HOXB7 and ERBB2. It excludes the possibility of a genetic cause for a banded distribution pattern found in blood samples. Furthermore, in the cross validation in brain, a better overall performance of three models in region EC is observed (see Table 4.3). A likely explanation about the discrepancy of performances between difference regions, is that EC is the first area of the brain to be affected in Alzheimer's disease [270], so it is reasonable that the alternation of genes happen more frequently in this region than in other regions.

Furthermore, we find a correlation between the expression levels of the proteins with two boundaries in our study. For instance, in the dataset GSE29676, the AD sample group with a down-regulated expression level of protein ECH1 also have down-regulated HOXB7 and up-regulated ERBB2 (Pearson correlation r=0.99 for ECH1 & HOXB7; r=-0.95 for ECH1 & ERBB2; and r=-0.94 for ERBB2 & HOXB7, see Figure 4.3). The same situation was observed in the dataset GSE39087 (see Figure 4.4). These observations suggest that there is an underlying linkage between the upstream activities of these proteins. We predict that further investigations will reveal co-expression, regulation or antagonistic relation between the precursor molecules of those proteins, at the level either of transcription or translation.

Considering the proteins in our panels, ECH1 is a gene encoding a member of the hydratase/isomerase superfamily. The gene product shows high sequence similarity with the enoyl-coenzyme A (CoA) hydratases of several species, especially within a conserved domain that is characteristic of these proteins. The encoded protein contains a C-terminal peroxisomal targeting sequence that localizes to the peroxisome. Its rat ortholog is a delta3,5-delta2,4-dienoyl-CoA isomerase that functions in the auxiliary step

of the fatty acid beta-oxidation pathway. This transcript was reported to be significantly up-regulated in response to neuronal silencing in the rat [271] but no linkage to AD or dementia has been reported previously. HOXB7 is a member of the Antp homeobox family and encodes a protein with a homeobox DNA-binding domain. It is included in a cluster of homeobox B genes located on chromosome 17. The encoded nuclear protein functions as a sequence-specific transcription factor that is involved in cell proliferation and differentiation. HOXB7 is age-repressed in mesenchymal stromal cells and conversely age-induced in hematopoietic progenitor cells [272]. ERBB2 is a member of a family of single-transmembrane receptor tyrosine kinases called ERBB and plays the main role in mediating Neuregulin-1 (NRG1) function [273, 274]. NRG1 participates in numerous neurodevelopmental processes, and is implicated in nerve cell differentiation and synapse formation [275, 276], radial glia formation and neuronal migration [277, 278], oligodendrocyte development and axon myelination [279, 280], axon navigation [281], and neurite outgrowth [282, 283].

Our findings suggest that the combined expression levels of ECH1, HOXB7 and ERBB2 have good potential to be an indicator of AD pathology. ECH1 and HOXB7 are expressed in almost all tissues and are enriched in the central nervous system, while ERBB2 is absent from many tissues and is not detected in the central nervous system (`http://www.proteinatlas.org/`). Whether these proteins can pass the blood brain barrier is yet to be investigated.

We note that our approach is different from the recursive feature elimination (RFE) method, which searches features starting from the sorted full feature space and eliminates features by a certain number or proportion in each iteration [261]. In contrast, our approach searches features by including important and informative features in each iteration. Such methods are greedy and may achieve global solutions, but are computationally expensive. To overcome this, we restricted our method to include just one feature in each iteration and terminated the searching when the improvement of prediction model caused by including a new feature was less than a predefined threshold (zero in the study).

4.5.1 Limitation of the study

There are some limitations throughout the study. A major obstacle in cross validating the models between GSE29676 and GSE39087 is that no raw data was provided, thus, the disparity or bias caused by difference in data pre-processing cannot be eliminated. This could have also contributed to the discovery of the wider spread of control samples found in protein ERBB2 and HOXB7 in GSE39087.

A second limitation of the study is the absence of braak stage data that can indicate the extent of AD pathology, which can possibly provide more possible explanations on different performances among different brain regions in the validation, since AD onset is a process that render sequential effect of different brain regions.

Another limitation lies on the fact that no blood cell type information is provided in any datasets analysed in the study. Thus, the effect of abundance of particular blood cell type cannot be eliminated.

We predefined a knowledge-based feature pool (KBFP) based on existing knowledge about AD pathology before feature selection was conducted. However, this feature pool is imperfect in the following perspectives. Many of the features collected are only brain specific and thus alternations of biomolecules oriented from blood may not be fully included. As some genes or proteins are altered in blood in AD but in a peripheral response to AD pathology, they do not mirror what has happened or going to happen in brain, but they are still equally good biomarkers. In other words, biomarkers fallen into this category are missing in KBFP. We also consider to include the 26 genes discovered in brain RNA biomarker study described in Chapter 3 into the pool to further improve its diversity.

4.6 Conclusion

The inclusion of existing biological knowledge and use of a novel feature selection method has allowed us to find several protein models that have a promising ability to distinguish

AD patients from healthy individuals. We also find a new statistical pattern involving both upper and lower bounds to expression in our model. The reproducibility of these findings needs now to be tested in larger cohorts.

Chapter 5

MetaUnion R package development

5.1 Abstract

This chapter describes the development of R package *metaUnion* - an advanced meta-analytic approach applicable for microarray data. This package is designed to overcome the defects appear in other similar meta-analytic packages, such as the neglection of missing data, the inflexibility of feature dimension, and the lack of functions to support post-analysis summary. *metaUnion* has been applied in a study to identify differentially expressed genes as part of the integrated genomic approaches. Genes like NEUROD6, ZCCHC17, PPEF1 and MANBAL were identified to be potentially implicated in LOAD. The result of thefuncz study was published in 2015 [265]. In addition, a part of functions in *metaUnion* is also applied and programmatically embedded in the data analysis process of biomarker database *Alzexpress* that will be discussed in Chapter 6. The package is now only publicly available on Github (`https://github.com/chingtoe365/metaUnion`).

5.2 Introduction

5.2.1 Meta-analysis overview

Meta-analysis is a statistical method that allows an analyst to combine effect sifunczzes across different studies into one meaningful estimate. In contrast to results from a single primary study, results from a meta-analysis deliver better generalizability, greater precision, and the ability to explore heterogeneity across studies [284, 285].

The conventional meta-analytic model calculates an average effect size by weighting each effect with the inverse of an effect size's variance (i.e., the squared standard error). The idea is that studies with a smaller variance have a greater impact on the average effect size relative to studies with a larger variance. This is referred to as a fixed-effect estimate [284]. An alternative and widely-used approach takes into account the variance between studies to estimate a random-effects meta-analytic model. Compared to a fixed-effect meta-analytic estimate, the random-effects meta-analytic estimate will have a greater confidence interval and the weights of each study will be similar [286].

Recent advances in meta-analytic methods have brought positive evolvement to the models. For instance, in tackling the issue of multiple effect sizes for one study, robust variance estimation can be used to incorporate all effect sizes into one model [287], whereas traditional methods only use priori decision rules to choose one effect size or simply average all the effect size in one study. Other multivariate meta-analytic approaches was also suggested [288], which require that the user know the covariance structure of effect sizes within each study.

Similar progresses have been made for a number of common application issues. To implement structural equation modelling with meta-analytic correlation matrices, a two-step modelling approach was designed by Cheung [289]. The first step is the synthesis of correlation matrix, and the second step is to conduct the usual structural equation models such as factor analysis or latent variable modelling, with the correlation matrix prepared by the first step. The framework was extended by Polanin et al [290] extended this

framework to accommodate complicated datasets with manifold levels of dependency. R package *metaSEM* developed by Cheung et al [289] allows users to conduct a meta-analytic structural equation model. Another advance is the application of network model by means of network meta-analysis, where the objective is to concurrently contrast multiple interventions. Albeit not as fashionable in the social or educational sciences, network meta-analyses are now becoming popular in medicine [291].

5.2.2 Application in genomics and microarray

With the advent of novel high throughput technology and chronic accumulation of microarray based biological studies, meta-analysis is becoming more important in biomarker discovery.

Microarray analysts are always facing a challenging task in terms of how to extract, compare and consolidate information from a magnificent amount of data. However, this is tampered by the complicated experimental protocols and designs embedded in microarray data. The main reason is the lack of normalised standards for microarray experiments for which heterogeneous datasets are generated. Under that condition, direct comparison is impossible. To solve the conflict of research interest and experiment reality, an approach combining results from different microarray datasets is quested.

Numerous R packages are developed for the purpose of conducting meta-analysis on genomic and microarray data. *MADAM* [292] provides the ability to calculate effect sizes, run fixed- or random-effects models, and also the functions to create some plots. The *metaARRAY* package [293] is specifically for large-scale meta-analysis of microarray data with a Bayesian framework. An author-written demonstration of the package capabilities using publicly available data is also included. The *metaDE* package [294] enables users to conduct 12 types of meta-analysis to discover differential expressed genes. Users are also provided with the option to choose which test statistic to calculated according to the outcome variable and choose corrections for one-sided tests. To date, the *metaDE* package emerges as the only microarray meta-analysis R package that compensate the

Table 5.1: Function overview of meta-analysis R package for genome. Source: [286]

Package	Effect Size	Dimension Compatibility	Missing Data	Dependent Effects	Fixed Effect	Random Effects	Moderator Analyses	Publication Bias	Sensitivity Analysis	Creates Plots
metaMA	✓				✓	✓			✓	
MAMA	✓					✓	✓			✓
metaDE	✓		✓		✓	✓			✓	✓
metaARRAY						✓				
MADAM										✓
metaRNAseq	✓				✓				✓	
metaPCA			✓	✓						✓
metaQC			✓	✓					✓	✓
metaSKAT	✓		✓		✓	✓				
MultiMETA						✓	✓			✓
Eas✓Strata										
gap	✓			✓	✓	✓				✓
metaUnion	✓	✓	✓							✓

omission of missing data. The imputation of missing data is designated for any gene with less than 30% missing data. In *metaUnion*, instead of imputation, the effect is reflected in the result by linking the missing data with the weighting coefficients during the combination. The *metaMA* package [171] searches for differentially expressed genes via merging either p values or t-test statistics in both paired and unpaired data. A demonstration vignette is also embedded using publicly available data for this package. Package *metaRNAseq* [295] shares similar functionality yet targeting on applying to RNA sequencing experiments. The *MAMA* package [296] relies on several R packages GeneMeta [297], *metaMA*, and *metaARRAY*.

Methodology wise, Choi et al [298] and Rhodes et al [168] were among the first authors to explore the practise of meta-analysis in the context of microarray data to find differentially expressed (DE) genes. Part of Choi et al [298] methods are applied in the Bioconductor package GeneMeta [297]. These approaches utilize either the effect size calculation [298] by inter-study variation modelling or the combination of p-values [168]. Conlon et al [299] and Scharp et al [300] also suggested to employ Bayesian methods to combine microarray data.

5.2.3 Defects in existing packages

Nonetheless, many defects appear when current meta-analysis R packages are used in practice. First, current R packages cannot handle data from different studies with different dimensions. Second, the presence of null values in the input data is ignored by directly omitting cases with missing value in these packages, which bias the true result of meta-analysis. Third, the output of result is not user-friendly and intuitive due to the unobvious illustration of p-values, z-scores and regulation direction for each feature in each study. Finally, a union calculation of targeted features selected by different effect size calculation methods is not provided. To improve the usability and practicability of meta-analysis for bioinformatics research, a new meta-analysis package *metaUnion* was developed to fix the existing deficiencies of current packages. It will be able, first, to

handle data from different studies with different dimensions; second, to rectify the number according to the number of missing value in the input; third, to provide the result of p-values, z-scores and effect sign for each feature; last, to provide a union calculation of targeted features selected by three different algorithms.

5.3 Statistic

The basic methodology of meta-analysis has been introduced in section 2.4.6.

The analysis of gene expression array data has always been centring on a fundamental question, which is whether the level of expression is significantly different in samples from two different conditions. Statistical hypothesis testing is a powerful approach to answer this question.

In *metaUnion*, two types of hypothesis tests are chosen to conduct differentially expressed gene analysis. Users are provided two options - student's t-test and a moderated t-test - where they can choose either approach to conduct DEG analysis.

5.3.1 Student's t-test

A t-test is most commonly employed when test statistics follows a Student's t-distribution under null hypothesis. In a t-test, the well-known t-statistics is measured as the extent of difference between two populations, in the form of

$$t = {(m_c - m_t)}/{\sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}} \tag{5.1}$$

where, c and t represent different experiment conditions to be compared with each other, and for each $m = \sum_i x_i/n$ and $s^2 = \sum_i (x-m)^2/n-1$ are the estimates for the mean and standard deviation. Knowingly that t follows approximately a Student distribution, with

$$f = \frac{\left(s_c^2/n_c + s_t^2/n_t\right)^2}{\frac{(s_c^2/n_c)^2}{n_c - 1} + \frac{(s_t^2/n_t)^2}{n_t - 1}} \tag{5.2}$$

degress of freedom. The advantage of using t-test comparing to the above-mentioned fold change approach, is that the means between different populations are adjusted by empirical standard deviations, therefore the fixed fold-threshold issue can be addressed. [301] However, limited by the high expense and tediousness of experiments, the sample numbers n_c and n_t is often very small. Thus the variance is often significantly underestimated.

5.3.2 Moderated t-test in R package limma

Another approach to conduct DEG analysis is the moderated t-test implemented in R package *limma*. This [302] is an R/Bioconductor software package that integrate linear modelling with complex experimental design analysis, remarkably developed with information borrowing to tackle the problem of small sample size in differential expression in array or sequencing studies.

The highly parallel nature of gene expression experiments entitled this package to borrow information between genes in a method commonly known as empirical Bayes. The coefficients inferred by linear models fitted to each gene provide additional information needed in Bayesian method to moderate residual variances.

This function fits multiple linear models by weighted or generalized least squares. It accepts data from a experiment involving a series of microarrays with the same set of probes. A linear model is fitted to the expression data for each probe. The expression data should be log-ratios for two-color array platforms or log-expression values for one-channel platforms

$$E(y_g) = X\alpha_g \tag{5.3}$$

where $E(y_g)$ is the logarithm transformed expression levels, X is a design matrix of full column rank and α is a coefficient vector. Certain contrasts of the coefficients are

assumed to be of biological interest and their coefficients are extracted by

$$\beta_g = C^T \alpha_g \tag{5.4}$$

where C is the contrast matrix. The target hypothesis to be tested is whether individual contrast values β_g are equal to zero. Some distributional assumptions are customised for the convenience of hypothesis testing.

a) The residual variances s_g^2 are assumed to follow approximately a scaled chisquare distribution.

$$s_g^2 | \sigma_2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2 \tag{5.5}$$

Where d_g is the residual degrees of freedom for the linear model for gene g.

b) It is assumed that the variances of linear coefficient estimators α_g and that of the contrast estimators β_g comply the followings

$$var(\hat{\alpha}_g) = V_g s_g^2 \tag{5.6}$$

$$var(\hat{\beta}_g) = C^T V_g C s_g^2 \tag{5.7}$$

where V_g is a positive definite matrix not depending on s_g^2. Let v_{gj} be the jth diagonal element of $C^T V_g C$. The contrast estimators are assumed to be approximately normal with mean β_g and covariance matrix $C^T V_g C \sigma_g^2$

$$\hat{\beta}_{gj} | \beta_{gj}, \sigma_g^2 \sim N(\beta, v_{gj} \sigma_g^2) \tag{5.8}$$

Under the above-mentioned assumptions, the ordinary t-statistic

$$t_{gj} = \frac{\hat{\beta_{gj}}}{s_g \sqrt{v_{gj}}} \tag{5.9}$$

follows an an approximate t-distribution on d_g degrees of freedom.

To describe how the unknown coefficients β_{gj} and unknown variances σ_g^2 vary across genes, prior distributions for these sets of parameters are defined. To describe how the variances are expected to vary across genes, prior information is assumed on σ_g^2 equivalent to a prior estimator s_0^2 with d_0 degrees of freedom, i.e., where 0 stands for the initial estimator,

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2 \tag{5.10}$$

In regards to the prior information of β_{gj}, it is assumed it complies with a normal distribution with mean equal to zero and unscaled variance v_{0j}, i.e.,

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{gj} \sigma_g^2) \tag{5.11}$$

Swapping the related parametrizations from those proposed by Lonnstedt and Speed [303], where $d_g = f, v_g = {}^1/_n, d_0 = 2v, s_0^2 = {}^a/_{(d_0 v_g)}$ and $v_0 = c$, the posterior mean of σ_g^2 given s_g^2 under the above hierarchical model is

$$\tilde{s}_g^2 = E(\sigma_g^2 | s_g^2) = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \tag{5.12}$$

The posterior values shrinks the observed variances towards the prior values with the extent of shrinkage in relation to the relative sizes of prior and observed degrees of

freedom. Define moderated t-statistics by

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \tag{5.13}$$

Which is proved to follow a t-distribution under the null hypothesis $H_0 : \beta_{gj} = 0$ with degrees of freedom $d_g + d_0$. Different from the fully Bayesian approach which allows users to choose the hyperparameters such as d_0, s_0^2, v_{0j}, the empirical Bayes approach estimate them from the data.

5.3.3 Statistics in metaUnion

Meta-analysis has been widely used in medicine and health policy to interpret contradictory results from various studies. And microarray experiments are a typical example of small sample size designs. The package *metaUnion* aims to provide users a robust tool to conduct meta-analysis with the option of choosing the methods desired to calculate the statistics. Compared with other meta-analysis packages such as *metaMA*, *metaUnion* improves the compatibility with microarray experiments by unlimiting the dimension of input expression value matrix. This improvement allows results from different microarray platform to be able to analysed together, which magnificently enlarge the sample sizes that the results are based on. Large samples consequently reduces the skewness of result caused by extreme samples and increases the liability of the conclusions. The package also takes the absence of expression values in each study into consideration, by adjusting the weights that are used during the intercourse of parameters estimation.

5.3.3.1 Effect size calculation

The following assumption is established for the calculation of effect size. Let Y_{sigr} and Y_{sjgr} be the expression levels for gene g in conditions i and j for study s and replicate r. The data are assumed to be normally distributed as $Y_{sigr} \sim N(\mu_{sig}, \sigma_{sg}^2)$ and $Y_{sjgr} \sim N(\mu_{sjg}, \sigma_{sg}^2)$. Then the effect size is

$$\delta_{sg} = \frac{(\mu_{sig} - \mu_{sjg})}{\sigma_{sg}} \tag{5.14}$$

Therefore, the effect sizes can be linked to the test statistics calculated in student's t-test or moderated t-test in *limma* by

$$t = d\sqrt{\tilde{n}} \tag{5.15}$$

with $\tilde{n} = {n_i n_j}/{(n_i + n_j)}$ where ni (respectively, nj) is the number of replicates in condition i (respectively j). The unbiased estimators of effect size d can be defined as

$$d' = c(m)d \tag{5.16}$$

with

$$c(m) = \frac{\gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\gamma(\frac{m-1}{2})} \tag{5.17}$$

where m is the number of degrees of freedom. For *limma* [302], m equals to the sum of prior degrees of freedom and residual degrees of freedom for the linear model for gene g. For student's t-test, degrees of freedom are $m = n_i + n_j – 2$.

Variances of effect sizes are also needed to apply meta-analysis procedure. As described by Marot et al [171], the exact form of the variances for effect sizes is inferred using the distribution of effect sizes provided by Hedges et al [304]

$$var(d) = \frac{m}{(m-2)\tilde{n}}[1 + \tilde{n}\delta^2] - \frac{\delta^2}{[c(m)]^2} \tag{5.18}$$

5.3.3.2 Effect size combination

The hierarchical model described by Choi et al [298] is adopted in order to combine effect size and obtain a test statistics for differential expression

$$\begin{aligned} d_{sg} &= \theta_{sg} + e_{sg}, e_{sg} \sim N(0, \omega_{sg}^2) \\ \theta_{sg} &= \mu_{sg} + v_{sg}, v_{sg} \sim N(0, \tau_g^2) \end{aligned} \tag{5.19}$$

where d_{sg} is the estimation of the effect size for study s and gene g, τ_g^2 represents between-study variance while ω_{sg}^2 are the within-study variances. The within-study variance has been estimated in the same stage as the estimation of the effect size. An estimation of the between-study variances τ_g^2 can be obtained using the method of moments as suggested by Choi et al [298]. Parameter $\mu - g$ is estimated as in the generalized least squares method

$$\hat{\mu_g}(\tau_g^2) = \sum(\omega_{sg}^2 + \tau_g^2)^{-1} d_{sg} / \sum(\omega_{sg}^2 + \tau_g^2)^{-1} \tag{5.20}$$

with

$$var(\hat{\mu_g}(\tau_g^2)) = {}^1/\sum(\omega_{sg}^2 + \tau_g^2)^{-1} \tag{5.21}$$

where $\omega_{sg}^2 + \tau_g^2$ is equivalent to the $var(d)$ in equation 5.18 and d_{sg} is estimated in equation 5.16. A z-score to test for differentially expressed genes is then constructed as follows

$$z_g = \frac{\hat{\mu_g}(\tau_g^2)}{\sqrt{var(\hat{\mu_g}(\tau_g^2))}} \tag{5.22}$$

The z-score is assumed to follow a normal distribution after an investigation by a q-q plot.

5.3.3.3 P-value combination

To combine p-values many authors such as Rhodes et al [168] and Hu et al [305] utilized Fisher's combined probability test across studies. However, the main drawback of this approach is that over- and under-expressed genes are treated separately. Marot et al [171] hence suggested to use the inverse normal method that is symmetric in the perspective that P-values near zero are accumulated in the same way as P-values near unity [306]. The inverse normal method refers to the averaging of transformed individual p-values to normal scores. This procedure was first introduced by Stouffer et al (1949) [307] and Liptak [308]. An alternative method to implement the inverse normal method is proposed by Marot et al [171] as

$$S_g = \sum_{s=1}^{N_s} \omega_s \phi^{-1}(1 - \tilde{p}_g(s)) \tag{5.23}$$

with

$$\omega_s = \sqrt{\frac{n(s)}{\sum_{i=1}^{N_s} n(i)}} \tag{5.24}$$

Where $n(s)$ is the number of replicates without null expression values for gene g in study s, and $\tilde{p}_g(s)$ is the one-sided p-values for each study, for the purpose of avoiding directional conflicts. Under the null hypothesis, S_g follows a standard normal distribution [171], thus an overall two-sided p-value can be inferred by

$$p_g = 2(1 - \phi(|S_g|)) \tag{5.25}$$

5.3.3.4 Special one-study case

In the case that expression values exist in only one study, the test statistic is assumed to follow a normal distribution. The p-value for the only-existing study is assigned as the p-value for meta-analysis, and the meta-analysis z-score can be inferred by

$$S_g = sign(\tilde{t}) * \phi^{-1}(1 - \frac{\tilde{p}_g(s)}{2}) \tag{5.26}$$

Where $\tilde{t}$ is the test statistic, and $\tilde{p}_g(s)$ is the two-sided p-value, both calculated from that single study. $sign()$ is a function to persist the sign of test-statistics.

5.4 Functions in the package

5.4.1 inputExpCheck

inputDataCheck is a pre-check function for input datasets. Auto-run in function *meta-Analysis()*. Checks if the information in the input data list matches, and filters out probes without annotations or with multiple Entrez ID annotations. We use Entrez ID as the gene ID.

Usage inputExpCheck(inputList)

Arguments

inputList	A list object containing lists for respective studies. Each list of a particular study contains an expression matrix and a sample information vector. The expression matrix is constructed with each row representing a gene (probeset) and each column representing a sample. The order of the columns should be [case samples][control samples][two additional columns]. The two additional columns appended at the end indicate the gene Entrez ID (this column must be named as "Entrez.Gene") and then the gene symbol. The sample information vector contains two elements, the first one represents the number of case samples and the second one represents the number of control samples.

Value A filtered data list with complete gene annotations (Entrez gene ID) and correct format for ongoing meta-analysis. The following probesets are excluded from the list: those which are expressed in less than 2 samples in each group; those which are unannotated or annotated in the wrong format.

5.4.2 metaAnalysis

Apply *limma* or student t-test to each probe and calculate statistics to identify differentially expressed genes in each study, followed by meta-analysis combining effect-size or p-value from different studies. Detailed procedure pipeline is shown in Figure 5.1

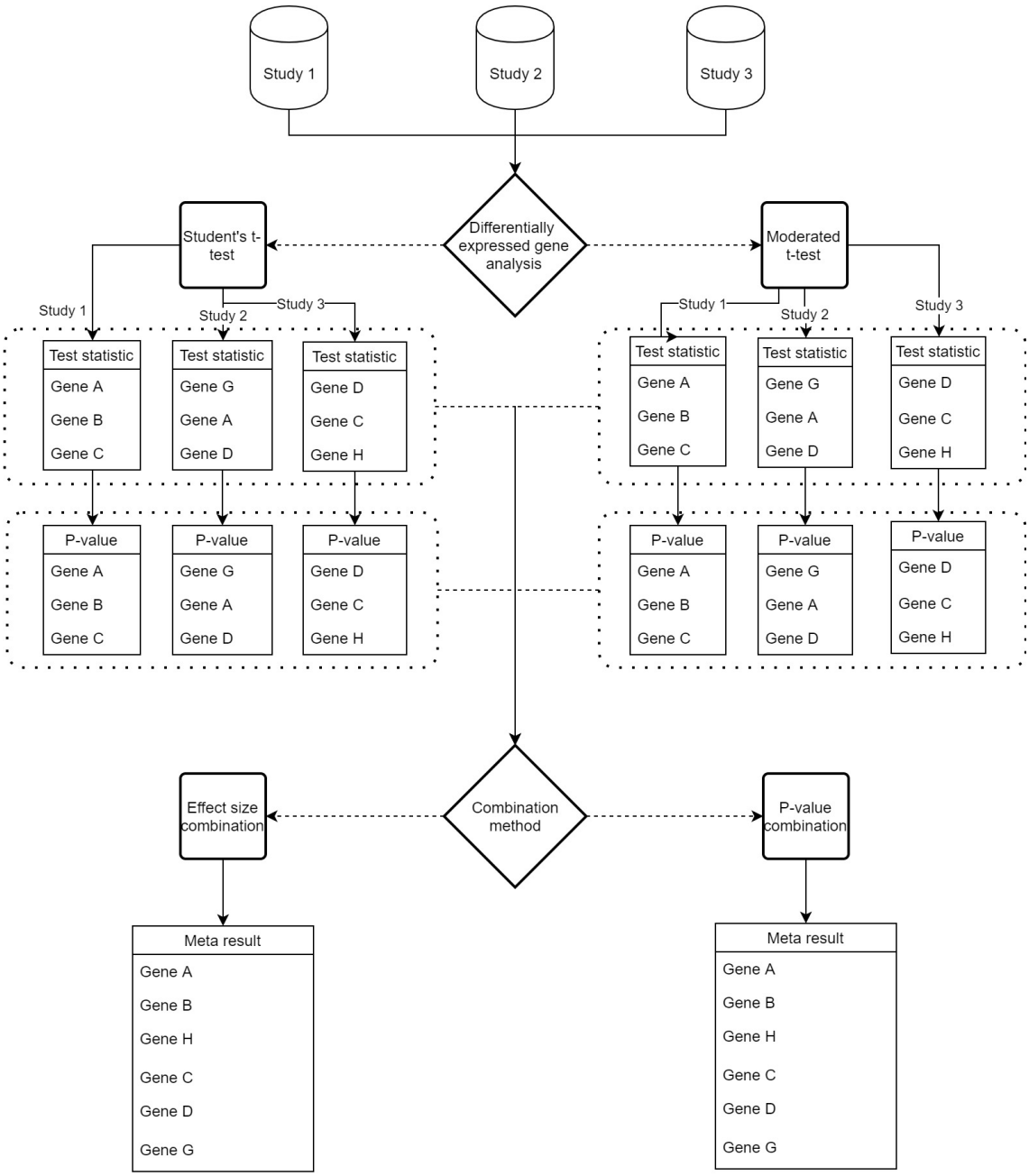


Figure 5.1: Meta-analysis pipeline of *metaUnion*. A three-study example is used for illustration. As depicted, two options are subjected to users' changes - 1) the method to conduct differentially expressed genes analysis with either student's t-test or moderated t-test; 2) the method to combine outcomes from every study with either effect size combination or p-value combination

Usage metaAnalysis(dataList,uniqGeneSelMethod="dprime",calWithLimma=TRUE, combinedPval=FALSE,filterData=TRUE)

Arguments

dataList	A list object containing lists for respective studies. Each list of a particular study contains an expression matrix and a sample information vector. The expression values should have been log 2 transformed. The expression matrix is constructed with each row representing a gene (probeset) and each column representing a sample. The order of the columns should be [case samples][control samples][two additional columns]. The two additional columns appended at the end indicate the gene Entrez ID (this column must be named as "Entrez.Gene") and then the gene symbol (this column must be named as "Symbol"). The sample information vector contains two elements, the first one represents the number of case samples and the second one represents the number of control samples.
uniqGeneSelMethod	Specifying the method used to select a gene if multiple entries correspond to one single gene, with "dprime" as default. Methods are as follows: logFC: select entry with largest absolute fold change; t: select entry with largest absolute t score; pval: select entry with smallest p value; n: select entry with largest number of expressed samples; df: select entry with largest absolute degree of freedom; d: select entry with largest absolute delta; vard: select entry with largest absolute vard; dprime: select entry with largest absolute dprime; vardprime: select entry with largest absolute vardprime. If two entriesy share the same value, then the one with the smaller row index in the expression matrix will be selected.

Continued on next page

Table 5.2 - - *Continued from previous page*

calWithLimma	Specifying which method (student's t-test or moderated t-test) will be used to calculate the statistics for each study, default TRUE using *limma* (moderated) t-test. If using t-test for calculation, the statistics of probesets whose expressed sample number in either case or control group is less than 2 will all set to NA and these probesets will be excluded if expCheckInput() function is triggered beforehand.
combinedPval	Choosing whether to combine p-value or effect size to execute meta-analysis, default FALSE combining effect size.
filterData	Choosing whether to filter input data or not by using function inputExpCheck(), default TRUE (do filtering).

Value A matrix summarizing the following information for each probe: EntrezID, gene official symbol, FoldChange in each study, effect-size in each study, test two-tailed p-value in each study, z-score after meta-analysis, p-value after meta-analysis, effect, significance, Bonferroni corrected p-value after meta-analysis. Probes expressed in less than 2 case/control samples will be ignored. Here "effect" indicates the direction of gene regulation, e.g. symbol "+", "-" and "?" represent a gene being up-regulated, down-regulated or missing in that study. In "significance" tab, symbol "!", "#" and "?" represent a gene being significant, not significant and missing in that study, where genes with p-value $<= 0.05$ are considered significant.

5.4.3 metaInfoStat

metaInfoStat is to summarize the input data list and calculate the number of genes found in all studies.

Usage metaInfoStat(inputList, metaResult, filterData=TRUE)

Arguments

inputList	A list object containing lists for respective studies. Each list of a particular study contains an expression matrix and a sample information vector. The expression values should have been log 2 transformed. The expression matrix is constructed with each row represents a gene (feature) and each column represents a sample. The order of the column should be [case samples][control samples][two additional columns]. The two additional columns appended at the end indicate the gene Entrez ID (this column must be named as "Entrez.Gene") and then the gene symbol (this column must be named as "Symbol"). The sample information vector contains two elements, the first one represents the number of case samples and the second one represents the number of control samples.
metaResult	A meta-analysis result list generated by function metaAnalysis().
filterData	Specifying whether to filter unannotated or multi-annotated probes or not, default TRUE (do filtering).

Value A list containing two matrices. One matrix summarizes the general information about the input data list, including number of genes, and number of case and control samples in each study. The other matrix displays the numbers of found in all studies (n) and to how many are found in n-1, n-2 studies etc.

5.4.4 metaZscoreQQplot

inputDataCheck is to generates a Q-Q plot figure for the z scores from the meta-analysis.

Usage metaZscoreQQplot(metaResult,mainTitle,savefile,width=500,height=500)

Arguments

metaResult	A meta-analysis result list generated by function metaAnalysis()
mainTitle	Main title of the Q-Q plot figure
savefile	File name
width	Width of the figure, default 500
height	Height of the figure, default 500

Value A Q-Q plot figure of meta-analysis z scores.

5.4.5 sortMetaStat

inputDataCheck is to extracts top ranking genes according to different measurements.

Usage sortMetaStat(metaResult, displayNumber, decreasing=FALSE)

Arguments

metaResult	A meta-analysis result list generated by function metaAnalysis()
displayNumber	Displays gene number
decreasing	Ranking by increasing or decreasing order, default FALSE (increasing order)

Value A short-listed metaResult with specified number of genes ordered by Bonferroni corrected p-values, and the -abs(metaZscore) as the second measurement.

5.5 Demo data Alzdata

A dataset for easy demonstration of all the usages in *metaUnion* is attached. The dataset is a subset of a larger aggregate of datasets investigated by Li et al [265].

Usage data(Alzdata)

Format List of 6 studies: gse5281, gse48350, gse15222, gse33000, gse44772, gse36980, which are a subset of microarray datasets downloaded from GEO (`http://www.ncbi.nlm.nih.gov/geo/`).

Each list of studies contains two data frames as the following:

expression.mat A data.frame, each object/row represents a probeset and each variable/column represents a sample. The two additional columns appended at the end indicate the gene Entrez ID (this column must be named as "Entrez.Gene") and then the gene symbol.

sample.info A data.frame with 1 object of 2 variables, numbers of case and control sample.

Source Datasets are trimmed from original studies, including only probesets for the top 30 differential expressed genes (DEGs) identified in our study [265] and the 50-150 randomly selected probesets from the original dataset.

5.6 Application in biomarker mining

One of the aims for *metaUnion* is to maximize the number of samples that are included in meta-analysis, such that those DE genes which are not commonly shared by all studies will not be omitted in the course of analysis. This advantage of *metaUnion* has been exploited in a study conducted by Li et al [265], where integrated genomic approaches pioneered by meta-analysis identified major pathways and upstream regulators in LOAD.

5.6.1 Method and Approach

5.6.1.1 Dataset selection and probeset pre-filtering

Before the study, datasets to be investigated were narrowed down by the following criteria,

- Late-onset Alzheimer's disease (LOAD) related mRNA expression studies.
- Data extracted from human post mortem brain tissues in either brain region super frontal gyrus (SFG) or prefrontal cortex (PFC), since both of which are part of the brain frontal lobe.

With the above criteria, six profile datasets with GEO accession number of GSE15222, GSE36980, GSE44770, GSE5281, GSE33000 and GSE48350 were selected.

Both GSE48350 and GSE5281 datasets were generated from the Affymatrix HG-U133-plus2 platform (GEO platform ID: GPL570) which contains 54675 probesets, containing 252 and 161 profiles respectively. Raw data were extracted from CEL files and then normalized using RMA (Robust Multi-array Average) method programmed in the *affy2* R package (`http://www.bioconductor.org`). A pre-filtering process was carried out to preclude probesets in each studies by consecutively applying the following filters:

- Probes with more than 10% absent calls (present/absent call by affy mas5 algorithm) across samples.
- Probesets with average expression level (after log2 transformation) lower than three.

The latest version of probesets annotation file was downloaded from Affymetrix NetAffx Analysis Centre (`http://www.affymetrix.com/analysis/index.affx`).

GSE36980 was generated from the Affymatrix Human Gene 1.0 ST platform (GEO platform ID: GPL6244) containing 79 profiles. Raw data were extracted from CEL files followed by an RMA normalization on the transcript level. A same pre-filtering

process described above in the pre-processing of GSE48350 and GSE5281 is implemented. Microarray annotation database package hugene10sttranscriptcluster.db [309] in R was used for annotating this dataset.

The common platform to generate GSE44770 and GSE33000 datasets is the Rosetta/Merck Human 44k microarray platform (GEO platform ID: GPL4372), containing 39302 probeset. These two datasets contain 230 and 623 case-control profiles respectively. Processed data are used without any pre-filtering.

GSE15222, containing 363 profiles, was generated from the Illumina microarray platform (GEO platform ID: GPL2700) including 24354 probesets. A sub-dataset downloaded from Dr Myer's lab (`http://labs.med.miami.edu/myers`) was used in the study. In the sub-dataset, probesets are defined to have null values if their relevant detection scores were less than 0.99. The sub-dataset therefore contains 8650 probesets resulting from a pre-filtering schema - probesets with more than 30% null values were excluded.

5.6.1.2 Sample filtering

In this study, the samples to be investigated are filtered by the following rules:

- Only samples that are collected from brain region PFC are retained to be investigated because maximum number of gene expression profiles can be obtained for this region. Samples collected from other regions are abandoned.

- Only samples with age information are included.

- Only samples aged between 65 and 95 years old are included to reduce the analysis bias because Alzheimer's disease is age-related.

After the above region and age control, a total number of 212 controls and 450 cases were left for the remaining analysis.

Table 5.3: Meta-analysis dataset meta information

Platform	Study	EntrezGene	Case	Control	Sum
GPL570	GSE5281	13424	23	9	32
GPL570	GSE48350	14903	21	23	44
GPL2700	GSE15222	3995	30	37	67
GPL4372	GSE33000	21576	249	88	337
GPL4372	GSE44770	21576	116	39	155
GPL6244	GSE36980	18922	11	16	27
Total	6	23530	450	212	662

5.6.1.3 Annotating probesets

Throughout the study, each probeset in every dataset is mapped to Entrez Gene IDs (as gene IDs). Those probesets without corresponding gene IDs were discarded. Furthermore, if a gene has more than one relevant probesets, the probeset having the largest absolute estimated effect size would be kept and the rest were discarded.

Table 5.3 lists the numbers of case and control samples, and the number of unique genes in each study used for the meta-analysis.

5.6.1.4 Conduct meta-analysis

With pre-processed datasets described above, both the effect-size based and p-value based meta-analysis method programmed in *metaUnion* were applied. Relevant statistical p-values were obtained by assuming a normal distribution with Bonferroni correction for multiple testing to identify combined-study DEGs.

5.6.2 Meta-analysis Result

Meta-analysis was implemented on previously selected six studies containing 450 AD and 212 healthy human brain tissue samples from the frontal cortex, summing up to include 23530 unique genes shown in Table 5.3 . After Bonferroni correction, 3124 differentially expressed genes (DEGs) were identified (with 1358 up-regulated and 1766

down-regulated). Only 3838 out of the 23530 genes (16.3%) appear in all six studies (the common genes), among which only 918 (23.9%) were found to be DEGs. More composition information of the 3124 DEGs is captured - 1582 (50.6%) genes were found in five studies; 242 (7.7%) in four; 213 (6.8%) in three and 169 (5.4%) in two studies. Obviously, if only common genes were to be investigated, most of the DEGs (the top 30 of 3124 DEGs can be viewed in Table 5.4 would be neglected (See Figure 5.2 (a)). The results of an alternate p-value based meta-analysis approach with Bonferroni correction also suggested similar results (see Table 5.5). 3315 DEGs were identified by this approach, with 3123 overlapping with the effect-size approaches (representing 99.9% and 94.2% of the DEGs identified separately). Also, high level of overlapping between DEGs identified by the effect-size approach and the p-value approach is shown in Figure 5.2 (b). This indicates a good homogeneity between two approaches.

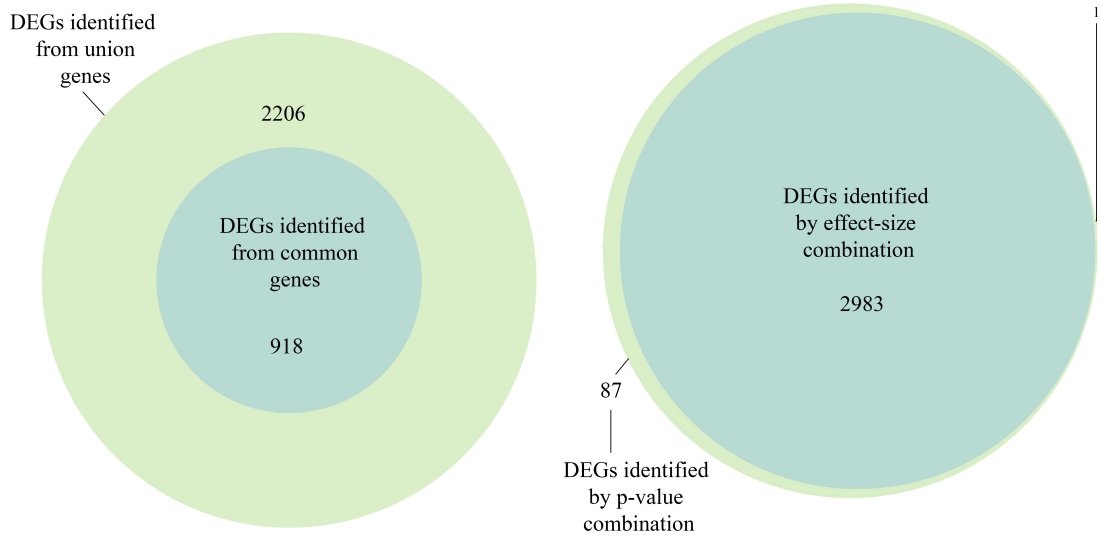


Figure 5.2: DEGs distribution comparison from different sources and methods. (a) DEGs identified from all and common genes (b) DEGs identified by effect-size and p-value combination

Once the meta-analysis was conducted, ensuing result presentation can be done by other functions in *metaUnion*. For instance, function *sortMetaStat* can sort the meta-analysis result by specific metric assigned by the user, and then a table similar to what

is presented in Table 5.4 can be generated where the top 30 DEGs are displayed. And meta z-score QQ plot can be generated by function *metaZscoreQQplot* (See Figure 5.3)

Table 5.4: Top 30 DEGs identified by effect-size based approach

Symbol	**metaZscore**	**metaPval**	**effect**	**significance**	**metaPvalBonf**
NEUROD6	-9.93	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ZCCHC17	-8.91	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
PPEF1	-8.90	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
C1QA	8.77	<2.22E-16	?+++++	?#!!!#	<2.22E-16
MANBAL	-8.76	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
BDNF	-8.74	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
CRH	-8.73	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
ITPKB	8.69	<2.22E-16	++++++	!#!!!#	<2.22E-16
FAM211A	-8.68	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
FKBP5	8.59	<2.22E-16	++++++	##!!!#	<2.22E-16
PCYOX1L	-8.58	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
MS4A6A	8.57	<2.22E-16	++++++	##!!!#	<2.22E-16
DUSP4	-8.51	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
TM7SF2	-8.49	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
SEMA3F	8.47	<2.22E-16	++++++	##!!!#	<2.22E-16
TRIP10	8.42	<2.22E-16	++++++	!#!!!#	<2.22E-16
LHFPL2	8.40	<2.22E-16	++++++	##!!!#	<2.22E-16
NREP	-8.40	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ARF5	-8.35	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ST6GAL2	-8.33	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
GPCPD1	-8.31	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
DOK3	8.30	<2.22E-16	++?+++	##?!!#	<2.22E-16

Continued on next page

Table 5.4 - - *Continued from previous page*

Symbol	**metaZscore**	**metaPval**	**effect**	**significance**	**metaPvalBonf**
KCNF1	-8.30	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
NFKBIA	8.28	2.22E-16	++?+++	!#?!!#	5.22E-12
ST8SIA5	-8.26	2.22E-16	- - ?- - -	!#?!!#	5.22E-12
NUPR1	8.19	2.22E-16	++++++	!#!!!#	5.22E-12
DPH2	-8.15	4.44E-16	- - ?- - -	##?!!#	1.04E-11
CCKBR	-8.14	4.44E-16	- - - - - -	!#!!!#	1.04E-11
LATS2	8.13	4.44E-16	++?+++	!#?!!#	1.04E-11
HIST1H2BD	8.11	4.44E-16	++++++	##!!!#	1.04E-11

Note: '-', '+' and '?' in **effect** column represent down-regulated, up-regulated, and missing in data. '!', '#' and '?' in **significance** represent gene being significantly differentially expressed, insignificantly differently expressed, and missing in data. The smallest value displayed is 2.2E-16 and those smaller than the value are marked as <2.2E-16.

Table 5.5: Top 30 DEGs identified by p-value based approach

Symbol	**metaZscore**	**metaPval**	**effect**	**significance**	**metaPvalBonf**
NEUROD6	10.15	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ZCCHC17	8.99	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
PPEF1	8.97	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ITPKB	-8.95	<2.22E-16	++++++	!#!!!#	<2.22E-16
BDNF	8.85	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
MANBAL	8.85	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
C1QA	-8.85	<2.22E-16	?+++++	?#!!!#	<2.22E-16

Continued on next page

Table 5.5 - - *Continued from previous page*

Symbol	**metaZscore**	**metaPval**	**effect**	**significance**	**metaPvalBonf**
CRH	8.82	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
FAM211A	8.79	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
FKBP5	-8.69	<2.22E-16	++++++	##!!!#	<2.22E-16
PCYOX1L	8.68	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
MS4A6A	-8.67	<2.22E-16	++++++	##!!!#	<2.22E-16
TM7SF2	8.66	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
DUSP4	8.62	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
SEMA3F	-8.59	<2.22E-16	++++++	##!!!#	<2.22E-16
ARF5	8.54	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
LHFPL2	-8.49	<2.22E-16	++++++	##!!!#	<2.22E-16
TRIP10	-8.48	<2.22E-16	++++++	!#!!!#	<2.22E-16
NREP	8.47	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
GPCPD1	8.44	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
ST6GAL2	8.42	<2.22E-16	- - ?- - -	##?!!#	<2.22E-16
NFKBIA	-8.42	<2.22E-16	++?+++	!#?!!#	<2.22E-16
DOK3	-8.40	<2.22E-16	++?+++	##?!!#	<2.22E-16
NUPR1	-8.39	<2.22E-16	++++++	!#!!!#	<2.22E-16
KCNF1	8.39	<2.22E-16	- - - - - -	!#!!!#	<2.22E-16
ST8SIA5	8.37	<2.22E-16	- - ?- - -	!#?!!#	<2.22E-16
LATS2	-8.33	<2.22E-16	++?+++	!#?!!#	<2.22E-16
GSG1	8.23	2.22E-16	???- - -	???!!#	5.22E-12
DPH2	8.23	2.22E-16	- - ?- - -	##?!!#	5.22E-12
HSPB3	8.20	2.22E-16	-?- - - -	!?!!!#	5.22E-12

Note: '-', '+' and '?' in **effect** column represent down-regulated, up-regulated, and missing in data. '!', '#' and '?' in **significance** represent gene being significantly differentially expressed, insignificantly differently expressed, and missing in data. The smallest value displayed is 2.2E-16 and those smaller than the value are marked as <2.2E-16.

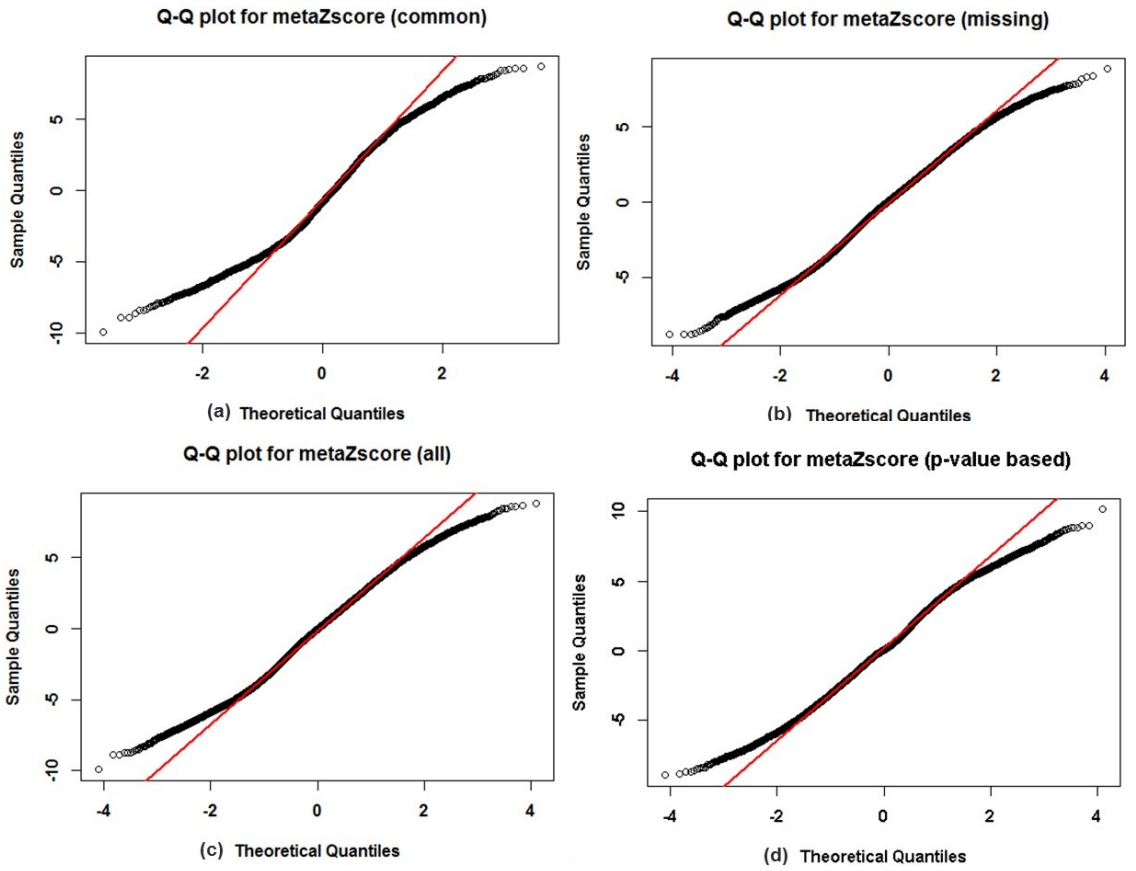


Figure 5.3: Normal Q-Q plot for the meta-analysis. Source: [265] Q-Q plot of the metaZscore calculated for those 3838 common genes across six studies (a); and those missing in at least one study (b); all 23530 genes (c). The distributions are approximated to normal distribution with two fat tails. Clearly the normality of the overall metaZscore is improved by merging data from those incompleted genes (missed in at least one study) to common genes. metaZscore of sample-size weighted P-value combined meta-analysis has similar normality (d).

5.6.3 Discussion

Both effect size and p-value combination aim to generate conclusions that summarize results from different studies. They differ from each other by how new hypothesis is constructed. Effect size combination reconstruct the null hypothesis by establishing a model which contains withiin-study and between-study variance, while p-value combination estimate a t-statistic by reverting p-values from each study and then aggregating

them. No logical impropriety exists in any of the two approach, and their linkage to the biological reality still needs to be investigated and clarified.

There are some limitation of *metaUnion*. No moderator analysis, publication bias and sensitivity analysis provided, as well as the lack of consideration in different types of effect sizes such as dependent effects, fixed effect and random effects. Also, there is no consideration about how to eliminate duplication of gene records, which at this stage has to be treated by the user before the processing by the main meta-anlaysis function.

5.7 Summary

In this chapter, the design, development and validation of novel R meta-analytic package *metaUnion* is described. Specifically aiming to compensate the defects found in similar predecessor packages in meta-analysis, *metaUnion* addresses the several issues which had been neglected in aspects such as missing data treatment and the fixed feature dimension requirement. *metaUnion* also provides additional functions for users to conduct post-calculation analysis easily. The validation of *metaUnion* reveals that it is a powerful R package with notable advantages can be summarised as follows:

- *metaUnion* can retain a maximum number of genes to be included in the analysis, thus a large proportion of DEGs that do not commonly probed across all the studies are not neglected - more than 70% in the validation test.
- *metaUnion* provides useful functions for downstream analysis once the major meta-analysis is completed. It provides the functions to sort the result table by specific measurement and to plot Q-Q plot.

In conclusion, *metaUnion* has advanced features compared to other existing meta-analytic R packages and is likely to gain popularity in conducting meta-analysis in microarray studies in the future.

Chapter 6

AlzExpress: an advanced biomarker database with meta-analysis for dementia

6.1 Abstract

This chapter describes the development of *AlzExpress* - a database to provide neurodegenerative disease association references for molecules in the interest of the researchers and users. *AlzExpress* offers the functionality of query on biomolecule, such as RNA, protein, and provides reference for researchers to identify differentially expressed molecules (DEM) in different categories of samples, with high statistically significance. It also provides meta-analysis to combine statistics from separate studies in order to generalize the conclusions made the and improve the reliability. It is the first system with a web-based interface to integrate DEM analysis and meta-analysis, in additional to some dataset meta-information and overlapping DEMs between different categories of studies. By virtue of the fashionable NoSQL technology for data management, the system shows high scalability and compatibility to include various types of molecular data and different experiment designs.

6.2 Background

The advent of high throughput microarray data has provided an effective method to search targets for drug discovery and diagnostics, and also explosively increases the volume of information for researchers in life science. Dementia and neurodegenerative disease studies are one of the beneficiaries of it, especially Alzheimer's disease (AD) studies, as it is one of the most challenging modern medical topics. The prosperity of data allows researchers to integrate the experimental data, analyse according to different study purposes and eventually develop into data-driven web servers/software founded on the basis of systematically designed databases.

6.2.1 Existing databases relating to AD research

With the booming wealth of information about putative AD susceptibility genes, it was increasingly difficult for researchers to follow and interpret the updates of genetic association studies about the disease. In 2007, Bertram et al [310] created a publicly available and persistently updated database *AlzGene* (`http://www.alzgene.org`), for the purpose of cataloguing all the genetic association studies in the field of AD. A comprehensive meta-analyses was conducted for polymorphisms existing in at least three case-control samples. Apart from the well-established association between $\epsilon 4$ allele of APOE, a dozen of potential AD susceptible genes were identified (ACE, CHRNB2, CST3, ESR1, GAPDHS, IDE, MTHFR, NCSTN, PRNP, PSEN1, TF, TFAM and TNF).

In 2014, Bai et al [311] initialized a database *AlzBase* (`http://alz.big.ac.cn/alzBase`) integrating several categories of information about AD from multiple sources, where elucidating the evolution of molecular pathway in AD is the main focus. Plenty of information is enclosed

- Gene dysregulation in AD and closely associated processes and diseases such as aging and neurological disorders
- Correlation between gene dysregulation and AD progression

- Plentiful annotations on the functional and regulatory information
- Gene-wise network relationship. A comprehensive summary for the top ranked genes is also provide in *AlzBase*

AlzBase allows researchers to prioritize genes from their own study and propose novel hypothesis in terms of the molecular mechanism of AD.

6.2.2 Specialties of AlzExpress

6.2.2.1 NoSQL database

The selection of NoSQL database as data storage and retrieval model has granted large scalability to the platform. Among the NoSQL family databases, MongoDB is selected as the central database underneath, because it owns good quality like being open-source, document-oriented and schema-free. The indexing and load balancing capability of MongoDB enables the platform to run fast and efficiently, meanwhile allowing ad hoc queries makes the development and structure design easy and flexible. Different from tabular relational databases, NoSQL databases utilize a weaker concurrency model than most relational (SQL) database systems. The ability to dynamically add new attributes to data records empower great flexibility to NoSQL database. In a practical sense, if a new attribute, for example education, is to be added in some of the samples in a dataset, the update will be a simple operation which otherwise will cause huge change of schemas in relational SQL. This characteristic makes *AlzExpress* very adaptive to future changes such as attribute addition of samples, new linkage between datasets, etc.

6.2.2.2 Meta-analysis on expression profiles

To the best of our knowledge, previous databases relating to AD do not cover statistical support of meta-analysis from the array profiling perspective. As described in the section 6.2, *AlzGene* specifies in revealing genetic association of AD via polymorphisms; *AlzBase* addresses on the pathways and functional annotations, meanwhile including differential

expression analysis but no meta-analysis is applied. *AlzExpress* include not only differential expression analysis but also meta-analytic results across all studies collected. The process of meta-analysis is conducted with the intrinsic method of R package *metaUnion* described in chapter 5. The inclusion of meta-analysis generalises the analysis output from various studies and thus generating more unbiased and reliable conclusions.

6.2.2.3 Resilience in comparison selection and cross-talking

Another huge scalability of *AlzExpress* stems from the capacity to include DEM/meta-analytic results from numerous comparisons. The meaning of diversity of comparisons is multi-fold. First, though it is completely AD-focused at this stage, more datasets can be easily imported into the database and ensuing analysis can be executed. Datasets of diseases like Parkinson's disease, other types of dementia or neurodegenerative diseases are able to be enclosed. Second, different variables of samples like brain regions or MMSE scores are included, so the comparisons can be made between different states of these variables. These variances result in a diversified comparisons that *AlzExpress* can provide. Given such an advantage, cross-talking of DEMs identified by different groups/diseases can be conducted and very meaningful results can be obtained, such as the DEM overlapping analysis.

6.2.3 Application summary

AlzExpress provides the functionality of query on biomolecule, such as mRNA, protein, and provides reference for researchers in this field to identify differentially expressed molecules in different categories of samples, with high statistically significance. It also provides meta-analysis for separate results across different studies, which help take varying statistical power from different studies into consideration. Meta-analysis serves to generalize and improve statistical reliability for each queried feature. AlzExpress adopts R package *metaUnion* (`https://github.com/chingtoe365/metaUnion`) which was developed in 2015 to conduct meta-analysis. This package has been imployed to identify

several major pathways and upstream regulators in AD [265].

6.3 System implementation

6.3.1 Data structure

The database side of *AlzExpress* is structured in a NoSQL convention and programmed in MongoDB. It is composed of four database for different use – sample database, annotation database, individual group test statistics database and meta-analysis statistics database. The detailed design of the database cluster is demonstrated in Figure 6.1

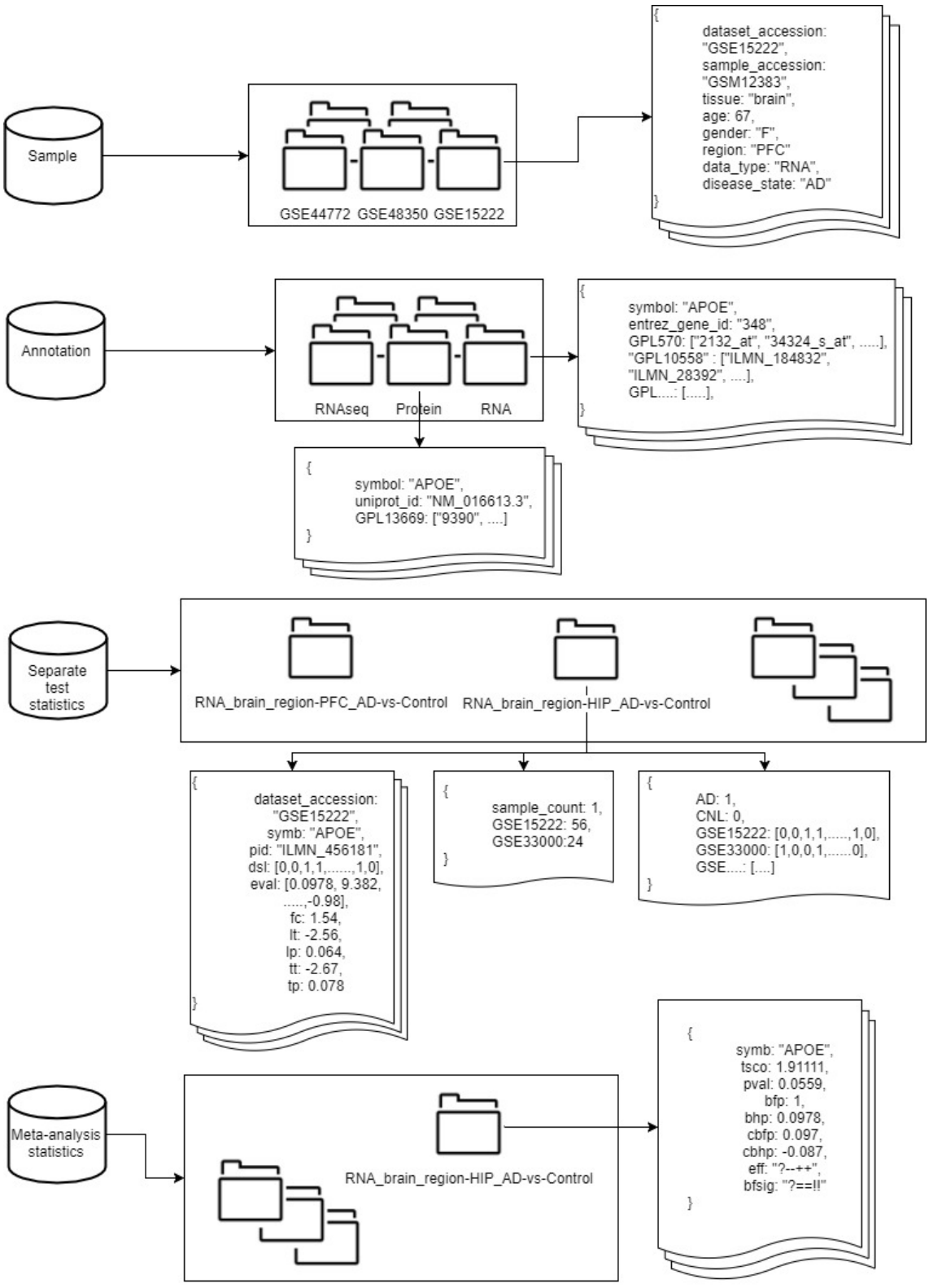


Figure 6.1: Data structure of *AlzExpress*. *AlzExpress* comprises four main databases depending of functionality: sample, annotation, separate test statistics and meta-analysis statistics.

The sample database consists of a number of collections with each representing a dataset for a single study. Each dataset then consists of a number of documents, each created for a sample with sample information such as age, gender, disease state, etc.

The annotation database consists of two collections which are separately for mRNA and protein. The mRNA collection is composed of thousands of documents with each representing an mRNA with unique Entrez Gene ID. Every document contains the official gene symbols and probe IDs which the particular mRNA corresponds to across different platforms. The protein collection is composed in a similar way except Uniprot ID is adopted as the identifier of each protein.

The individual group test statistics database and meta-analysis statistics database locates the group comparison statistics for each probe between different disease states. It includes fold change, hypothesis test z-score and p-value from either individual or multiple studies, adjusted p-value, and regulation status across different studies (for meta-analysis database). To facilitate the mediated data processing, two documents are specially designed to be the parameter check-book for each collection in test statistics database. One is designed for easy acquisition of sample count, and the other is designed for easy acquisition of disease state of each sample (see Figure 6.1).

6.3.2 Data processing and analytical methods

Figure 6.2 shows the detailed data processing and analysis pipeline. All of the sample data are stored in the sample database and are downloaded from GEO by GEOquery package in R [312], except a study with GEO series GSE15222 which are downloaded from the website of the experiment organisation – Myer's lab. Meta-data of each sample including age, gender, disease state (case or control), brain region where the sample locates (only for samples collected from brain) are extracted. We follow pre-designed prototype (see Table 6.1) to abbreviate the brain regions. Assay data are also extracted to an array of expression value for a specific sample, the order of which corresponds to another array for probe IDs.

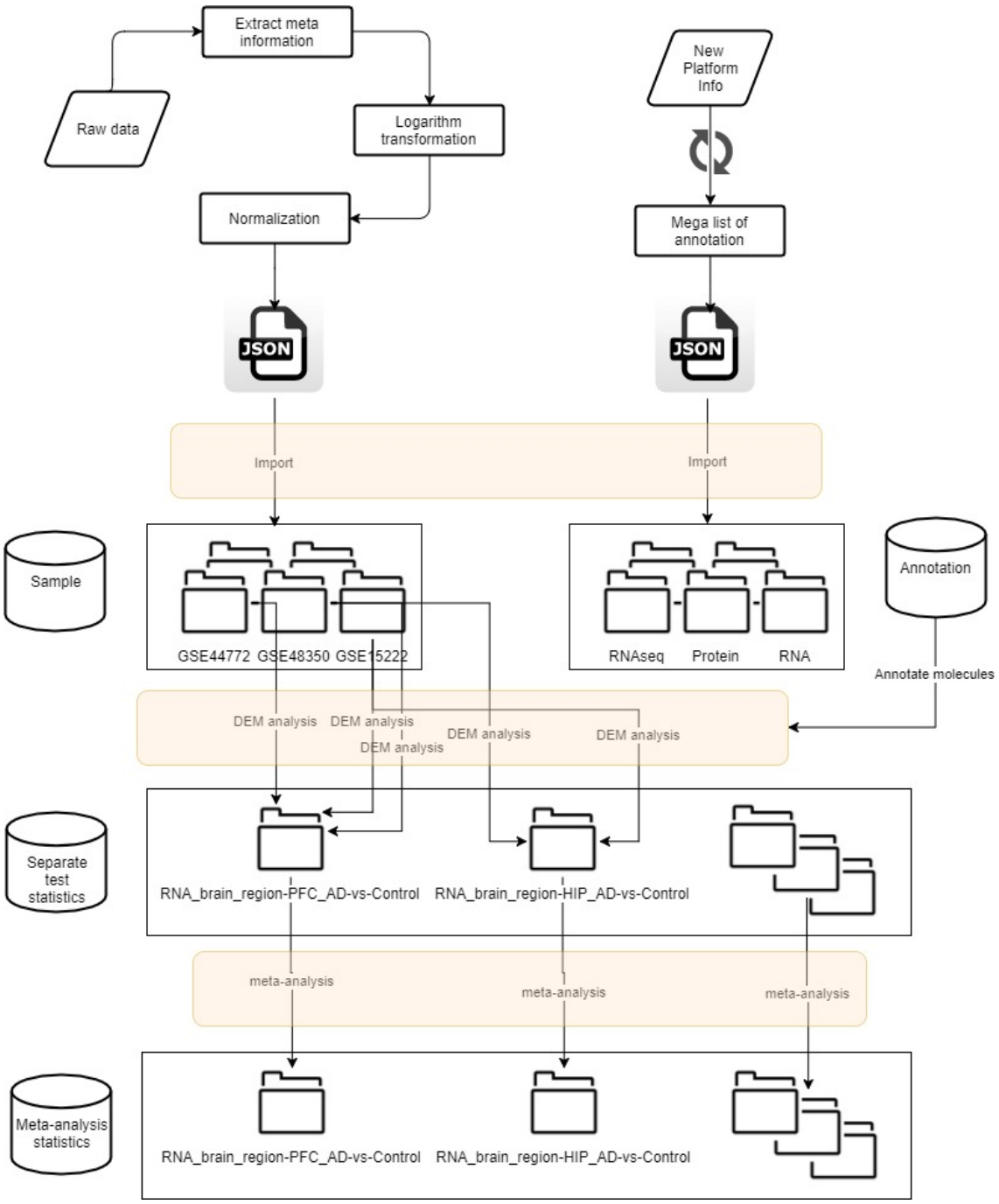


Figure 6.2: Data preprocessing and analysis pipeline of *AlzExpress*. The procedure can be summarized as a) raw sample and annotation data preprocessed and imported in JSON format b) separate tests are conducted for different datasets, with annotation assigning UIDs (entrez gene ID) across datasets c) conduct meta-analysis based on UIDs for specific groups

We adopted the expression data that the previous researchers had processed with certain kinds of pre-processing methods, but we made sure that the expression values

across all studies are logarithm transformed by two. For example, the values of GSE36980 are RMA normalized, wherein a logarithm transformation by two is performed, therefore we adopted the original data and import into our database. The values of GSE44772 are logarithm transformed by 10, so we transform it by exponentiation based by 10, after which a logarithm transformation by two is then performed, details in Table 6.1.

Individual group DEG tests are conducted across different categorical groups for each study. In each study, only samples from a particular category (eg. In region PFC) are selected. Then both *limma* moderated t-test and student t-test are conducted to identify DEMs. Their respective z-scores and p-values are recorded in the individual group test statistics database. Meta-analysis is conducted across various categorical groups. In each group (eg. region PFC), R package *metaUnion* is employed to perform meta-analysis by combining p-values for each feature from different studies. And p-values are adjusted by both BH and Bonforroni correction, with both common and union features numbers.

6.3.3 User interface

6.3.3.1 Functionality summary

Query function for specific genes are offered by *AlzExpress*, and individual test statistics and meta-analysis results for those genes of interest are displayed. Genes are annotated by NCBI and linked with STRING database. *AlzExpress* also provides investigation of common differentially expressed biomoleculars across different categories. Also *AlzExpress* provides various visualization plots for researchers, such as volcano plots, heatmap and bar-charts. All tables can be downloaded as CSV files.

6.3.3.2 Sample information overview

The summary section (shown in figure 6.3) provides an overview of meta data of datasets collected in the database. A summary of datasets offers information for users to decide what comparisons are able to be carried out from the platform. At this stage, the data types of datasets include RNA, protein and RNAseq, and the tissues from which the

Table 6.1: Summary of loaded datasets in *AlzExpress*

Dataset accession	Tissue	Data type	Note
GSE36980	brain	RNA	
GSE44772	brain	RNA	
GSE5281	brain	RNA	
GSE33000	brain	RNA	Frontal cortex to "PFC"; 10 to expression value and then log two transformed
GSE48350	brain	RNA	Postcentral gyrus to "POCG"; 10 to expression value and then log two transformed
GSE15222	brain	RNA	parietal lobe to "Postcentral gyrus"; both expression and covariate data from myers lab
GSE13214	brain	RNA	pathologic & defenitive Alzheimer are both classed as AD; use replicate 2, remove probes with all null values (234 probes)
GSE84890	brain	RNA	
GSE29378	brain	RNA	GPL6947 used but GPL10558 is the latest version so use latter
GSE63063	blood	RNA	GPL6947 used but GPL10558 is the latest version so use latter
GSE39087	blood	protein	NCBI reference sequence are used as unique identifiers in protein studies
GSE53697	brain	RNAseq	No probe ids so use symbols directly; use log 2 rpkm value as expression value; filter duplicate feature with largest foldchange

data is extracted include brain and blood.

6.3.3.3 Dataset meta information and analysis summary

The entire section is to provide information as per dataset basis.

Meta information

This section presents meta information breakdown for each dataset. In this section, pie charts are plotted to provide these information (see figure 6.4). Moreover, a volcano plot is also provided for a quick overview of the DEM analysis result for a particular group that the user is interested in. It helps identify molecules with high likelihood to be DEMs. If more than one regions exist in a dataset, a region input is required to bootstrap the volcano plot.

Volcano plot and DEM analysis result collection

A volcano plot binds the view of distribution of p-values and fold change together into one graph, such that those molecules that are likely to be DEMs can be identified easily. Figure 6.5 is an example of this section, red points in the plot represent those molecules with,

- Absolute fold change larger than 1.5
- limma p-value less than 0.05

From this statistic point of view, these molecules are assumed to be highly associated with the onset of the disease, since the fact is supported by both fold change and *limma* p-value.

At the bottom, a csv file download handler is also provided for the user to download the statistics into a csv file format for all the molecules. This is for the purpose of further analysis carried out based on the statistics calculated.

6.3.3.4 Molecule based and sample filtered query

In this section, *Alzexpress* provides querying functions for the users on the molecule they are interested in (shown in Figure 6.6).

The sample filter is very important before conducting queries to answer questions or various hypothesis proposed by the users. The purpose of the filter is to control variables so that only samples complying the required criteria remains. After applying the filter, only the value of two variables - disease status and molecules - varies across all selected samples. Therefore, the relationship between the expression of molecules and the disease status can be revealed by unbias results obtained from *Alzexpress* database.

There are four main variables that are needed to be constrained

- **Data type** where there are three options at this stage - RNA, protein and RNAseq. RNA stands for the expression profiling array data of transcripts, protein stands for protein profiling by protein array, RNAseq stands for expression profiling of transcripts by high throughput sequencing. This field is extendable and more data type is to be included such as DNA methylation, microRNA, etc.

- **Tissue** from which the data is extracted is required specified. There are two options at this stage - brain and blood.

- The third variable defines more detailed filters for the samples, where this filter can further refine the samples by only selecting those meeting the category. At this stage, only **region** as a category type is available and it will only be active with "brain" as "tissue" being selected, because samples extracted from blood will not have region as a property obviously. And for the **region** category type, there are 11 values available - PFC, HIP, etc. (listed in Table 6.2). This field is extendable and more category will be included such as MMSE score, ethnicity, etc.

- **Comparison** is the last variable required to be specified. At this stage, only the comparison of AD versus Control is available. This field is extendable and more disease status comparisons can be included in the future.

Table 6.2: Required fields for the query form and their available options. Check Chapter for abbreviations

Fields	Options
Data Type	RNA, protein, RNAseq
Tissue	Brain, blood
Category	Region
Category Value of region	PFC, HIP, EC, CE,,SFG, PC, VI, TE, MTG, PVC, TC, ALL
Comparison	AD-vs-Control
Duplicate removal method	Largest fold change

Table 6.2 shows detailed options for each field.

After required fields are filled, query can be submitted and an example result is shown in figure 6.7. The tabs above the table display all the datasets where there are at least one queried molecule found. In the example, there are four datasets found. For each dataset, a table is appended, displaying the analysis results for each molecule. Useful statistics including fold change, z-score and p-value calculated with normal student t-test or *limma* moderated t-test respectively.

Additional annotation for each molecule is also provided, with the entries to external sources of databases such as NCBI and STRING. It also provides an entry to the meta-analysis result section which will be discussed in the next section.

6.3.3.5 Meta-analysis in AlzExpress

The meta-analysis result is calculated by combining the results from previous section. The table on the top (see Figure 6.8) lists the following statistics

- **meta z-score** the z-score re-calculated by meta-analysis
- **meta p-value** the p-value re-calculated by meta-analysis
- BH adjusted p-values for union molecules
- BH adjusted p-values for commonly shared molecules respectively

- Bonferroni adjusted p-values for union molecules
- Bonferroni adjusted p-values for commonly shared molecules
- **Significance** describes whether the particular molecule is significantly differentially expressed across all the datasets, with "**!**", "**#**" and "**?**" representing being significant, not significant and missing
- **Effect** indicate the direction of molecule regulation, with "**+**", "**-**" and "**?**" representing up-regulated, down-regulated and missing

A table displaying the top 10 molecules with least meta p-values is also generated. Two tables are specially separated to address the different dimensional contexts in which the p-values were adjusted. One table is displaying the statistics calculated in the context of union molecules, where molecules appearing at least in one dataset will be included in the p-value adjustment process. In compared to this, another table is in the context of overlapping molecules, where molecules appearing in all datasets will be included in the p-value adjustment process.

The bottom table describes meta information for that particular meta-analysis. Sample count, AD sample count, control sample count and feature for each dataset are all in list. **Union feature count**is the number of molecules appearing at least in one dataset, whereas **Intersect feature count** is the number of molecules appearing in all datasets.

6.3.3.6 Common DEMs study

Alzexpress also provide a section for users to identify overlapping DEMs identified from different sample groups/categories. In this section, users are able to choose whatever sample groups they are interested in.

Suppose both "Protein-blood" and "RNA-PFC" are selected (see figure 6.9), the result will be the overlapping DEMs identified by the two groups. The statistics displayed in the table can have two sources. If the sample group selected only contains one dataset, the statistics display for that group is *limma* p-value calculated with that single dataset.

Otherwise, if the sample group contains more than one dataset, it is the Bonferroni corrected meta p-value. Molecules are identified as DEMs if this value is less than 0.05.

In addition, there are cases when the selected groups do not share the same data type, for example, one is in type "protein" and the other two are in type "RNA". In those situations, gene name will always act as the bridge between two different types. For example, protein "Apolipoprotein E" will be interchangeable to RNA with gene symbol "APOE", so that the overlapping calculation can proceed.

The graph at the bottom indicates the accountability of DEM identified and the unique molecules in each sample group.

6.4 Limitation

AlzExpress is designed as a reference warehouse for AD reserchers, focusing on converting past experiment results into conclusive statistics. There are a couple of limitations:

- Some visualization targets were not achieved, such as the idea of a heatmap indicating the full range of probes involved in a particular study, due to the built-in limitation of hardware performance from the perspective of users. An over-complicated figure should overload most of the browers on normal computers.
- A feasible prototype for data pre-processing has not yet been developed for all the other data types apart from microarray data, the data type which has been majorly accommodated by the platform. Therefore, the automation of pre-processing is still in very low level and the population of datasets needs to be supervised.

6.5 Summary

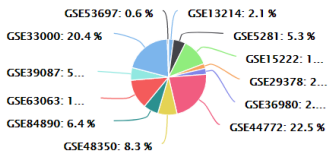
The invention of *AlzExpress* provides researchers with a powerful tool to obtain statistical support to test and verify hypothesis in interest. The result of DEM analysis and meta-analysis from different categorical groups for molecules in interest can be acquired easily via *AlzExpress*. It also provides useful graphs such as volcano plots and pie charts for

visualization of meta data, as well as convenient export of results via CSV files. Apart from the microarray data being focused on at current stage, the scalability enables more different types of measurements such as DNA methylation and sequencing data to be imported, processed and analysed in the future. *AlzExpress* is an ideal warehouse for high-throughput biostatistics in dementia research.

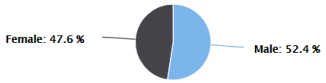
The future plans include importing and analysing samples with additional conditions such as mild cognitive impairment and Parkinson's disease in the future, and also aim to analyse micro-RNA and DNA methylation microarray data to reveal a more complete scope for molecular expression regulations in dementia patients.

Sample distribution

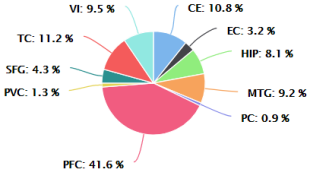
Sample proportion for different datasets



Sample proportion for different gender



Sample proportion for different regions



Dataset data type and located tissue

Dataset	Data Type	Tissue
GSE13214	RNA	brain
GSE5281	RNA	brain
GSE15222	RNA	brain
GSE29378	RNA	brain
GSE36980	RNA	brain
GSE44772	RNA	brain
GSE48350	RNA	brain
GSE84890	RNA	brain
GSE63063	RNA	blood
GSE39087	protein	blood
GSE33000	RNA	brain
GSE53697	RNAseq	brain

Figure 6.3: Sample overview page

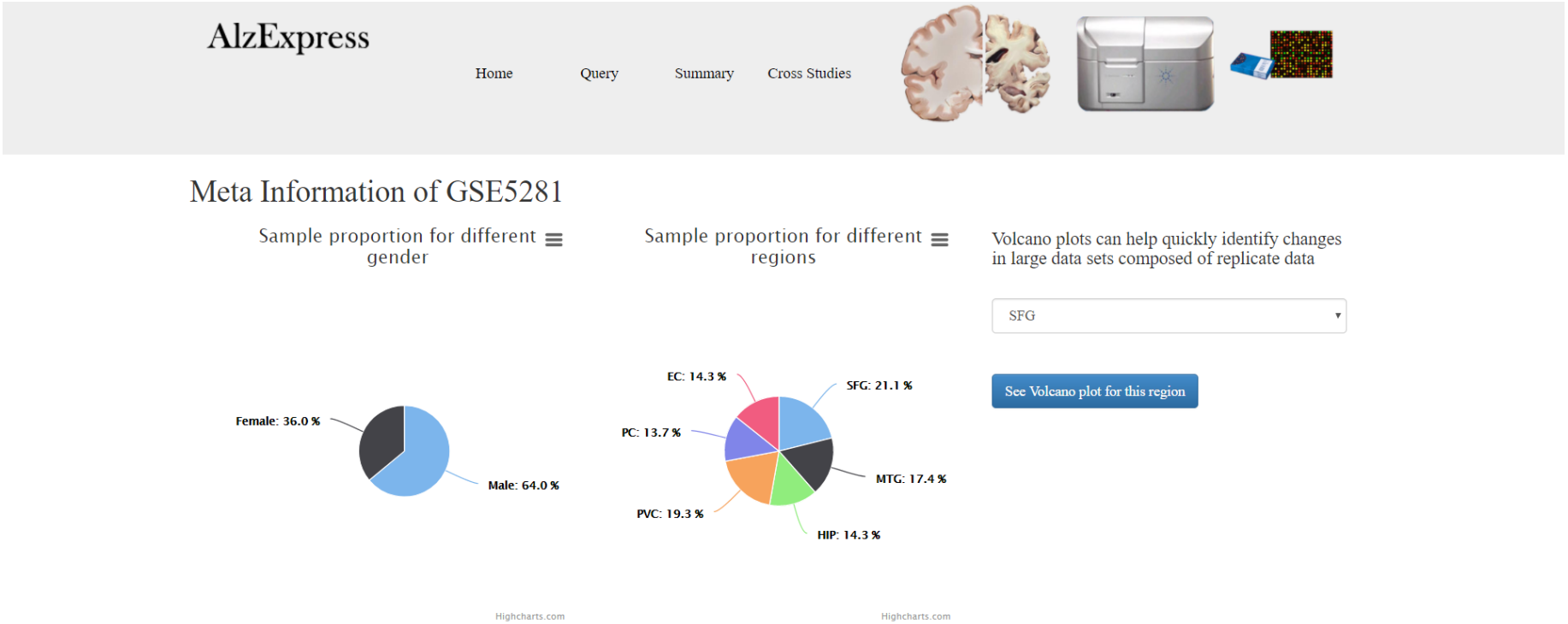


Figure 6.4: Meta information page in user interface - GSE5281

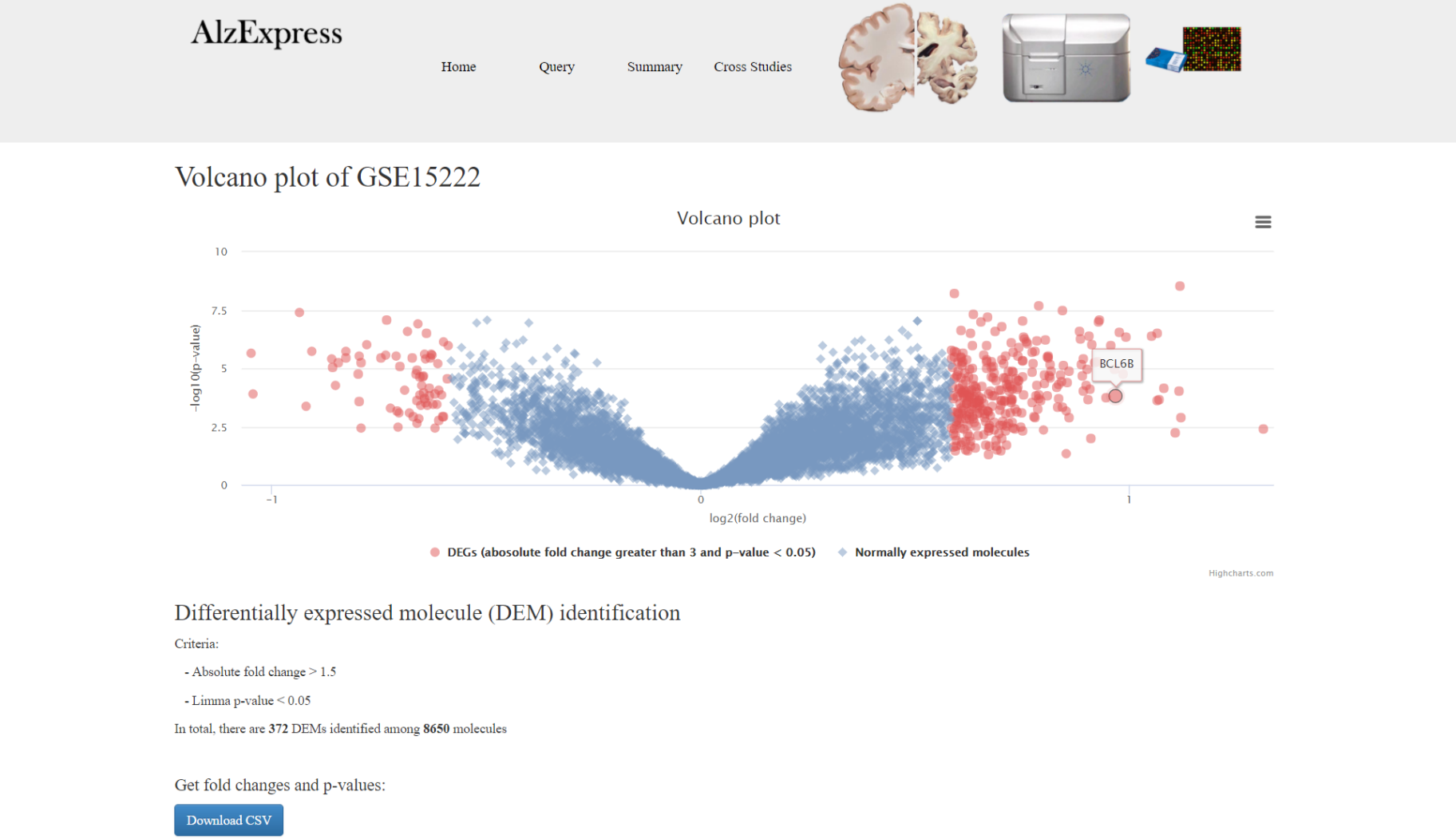


Figure 6.5: Volcano plot page in user interface – GSE15222

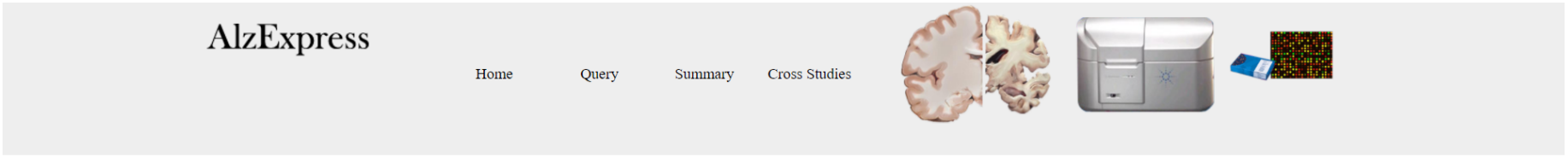


Figure 6.6: Biomolecule query page

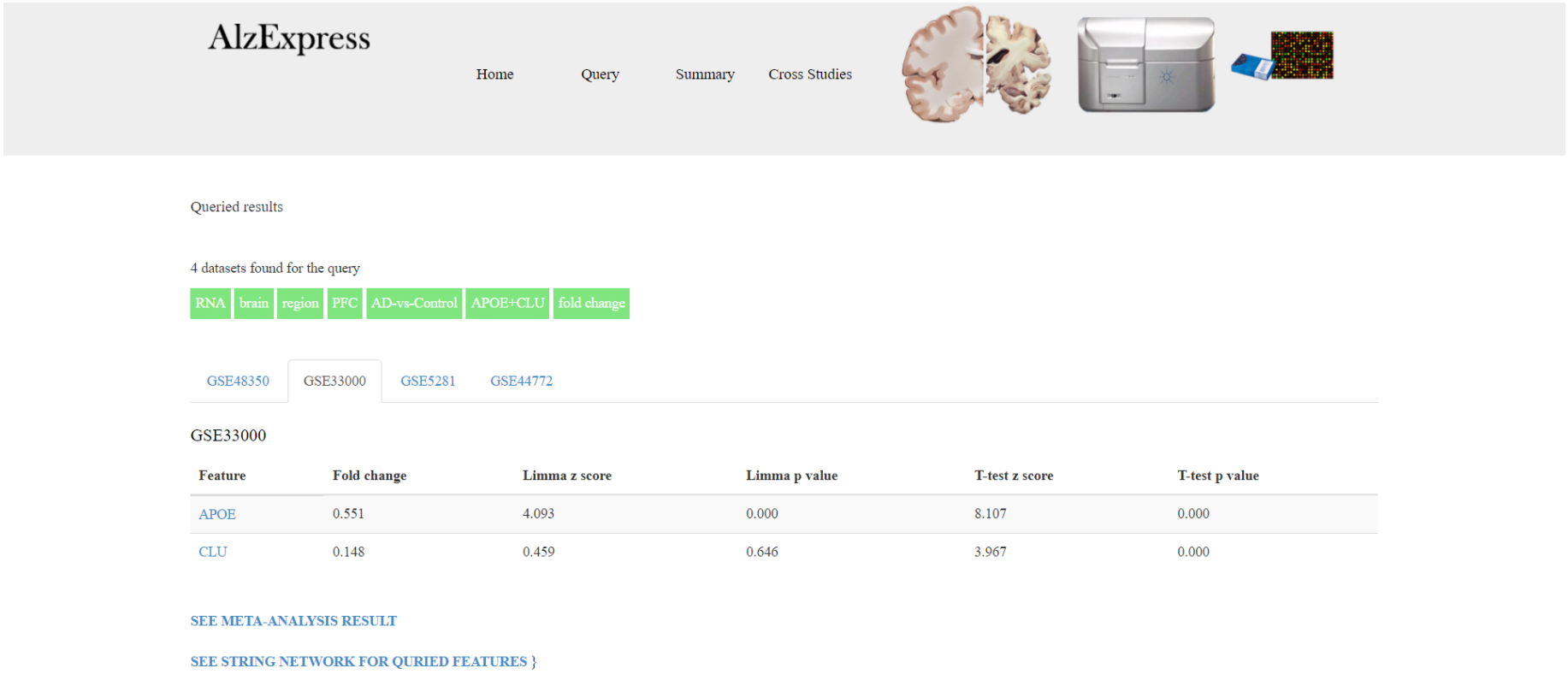
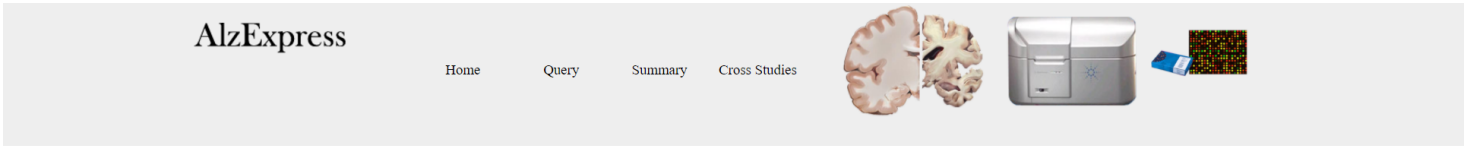


Figure 6.7: Query result of DEM analysis. The constraints are, data type – RNA, tissue – brain, region – PFC, comparison – AD vs. control, molecule in interest – APOE, CLU, method to select probe – largest fold change.



Meta-analysis result for this group

Download CSV

Feature	Meta t-score	Meta p-value	Adjusted p-value (BH)	Adjusted p-value (Bonferroni)	Adjusted p-value (BH, common features)	Adjusted p-value (Bonferroni, common features)	Significance	Effect
APOE	10.222	0	0	0	nan	nan	??!-!--	??++++++
CLU	inf	0	0	0	nan	nan	??!-!--	??-----

Top 10 features by Meta p-value

Filtered by Union features

Download CSV

Feature	Meta t-score	Meta p-value	Adjusted p-value (BH)	Adjusted p-value (Bonferroni)	Adjusted p-value (BH, common features)	Adjusted p-value (Bonferroni, common features)	Significance	Effect
tAKR	inf	0	0	0	nan	nan	??!-!??	??+-+??
EPHA8	inf	0	0	0	nan	nan	??!-!--	??++++-
NEFL	inf	0	0	0	nan	nan	??!-!--	??+-++-
EPHA6	inf	0	0	0	nan	nan	??!-!--	??-----
NEIL1	inf	0	0	0	nan	nan	?-!?!??	?++?+??
EPHA4	inf	0	0	0	nan	nan	-?!-!--	-?-----
NEIL2	inf	0	0	0	nan	nan	?-!-!--	?++++++
EPHA10	inf	0	0	0	nan	nan	??!-!--	??-+---
EPHA1	inf	0	0	0	nan	nan	-?!-!--	+?-----

Meta Information

Dataset	Sample Count	AD Sample Count	Control Sample Count	Feature Count	Union Feature Count	Intersect Feature Count
GSE5281	34	23	11	54675	24107	876
GSE13214	23	13	10	4566		
GSE36980	33	15	18	33297		
GSE48350	69	21	48	54675		
GSE44772	230	129	101	39280		
GSE33000	467	310	157	39280		
GSE15222	71	31	40	8650		

Figure 6.8: Meta-analysis result page

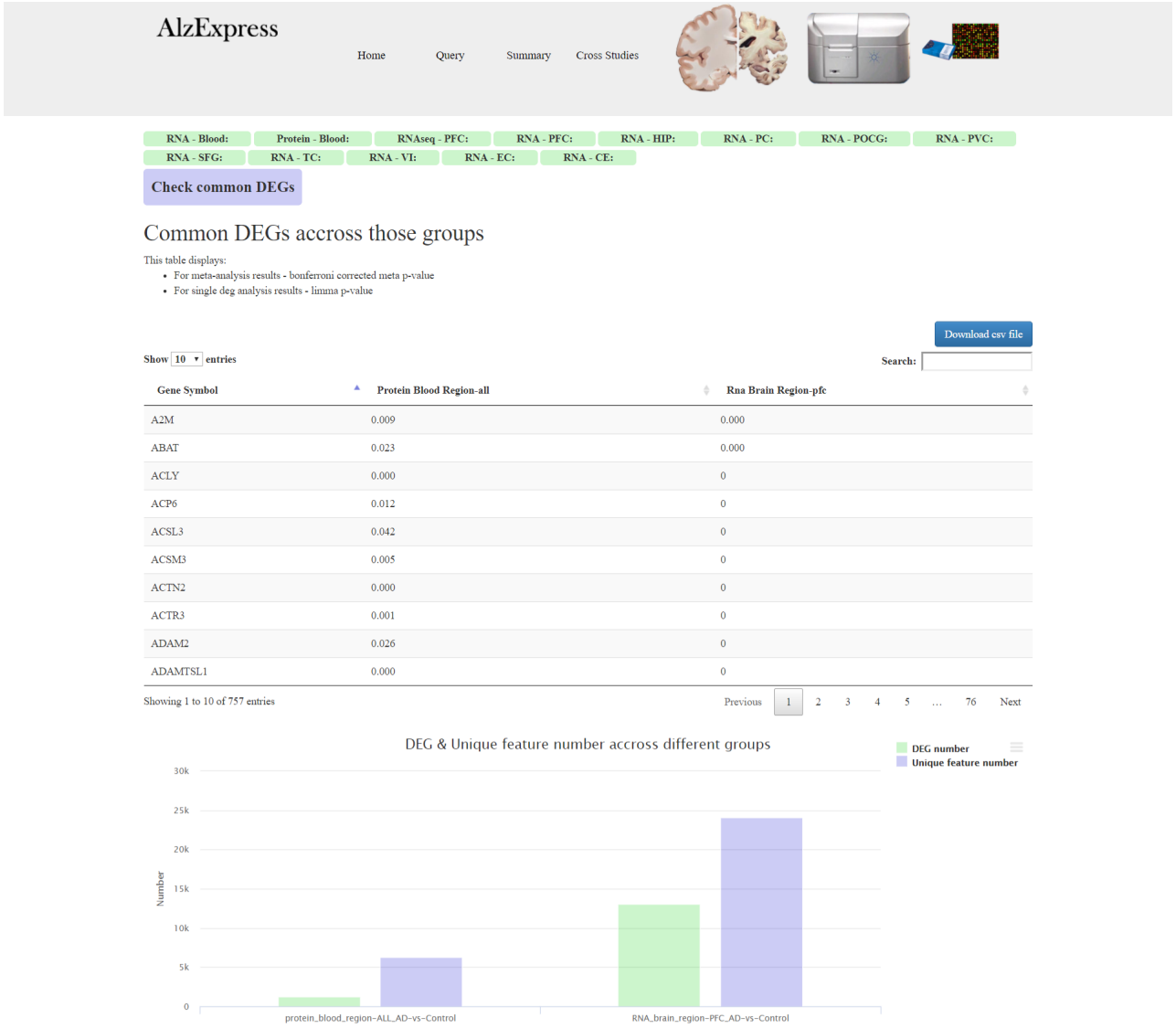


Figure 6.9: DEM overlapping query page

Chapter 7

Contributions

The main contributions can be summarized in the following list

1. We discovered that the inclusion of age and APOE genotyping into the prediction model can improve classifiers performance. A panel of 26 RNA transcripts was discovered that has promising classification ability between AD patients and healthy controls.

2. Three models (ECH1+NHLRC2, ECH1+HOXB7, ERBB2+FN1+SLC6A13) were discovered in blood which were identified to have good diagnostic ability. A surprising and novel statistical pattern was recognised which shed light on the existence of new pQTL for AD. Two novel SVM-based feature selection methods were proposed to select significant serum proteins with or without the constraint of feature pool. The result of the study was published in 2016 [240]

3. We developed an R package *metaUnion*, an advanced meta-analytic approach applicable for microarray data. This package is designed to overcome the defects appear in other similar meta-analytic packages, such as the neglection of missing data, the inflexibility of feature dimension, and the lack of functions to support post-analysis summary. *metaUnion* has been applied in a study to identify differentially expressed genes as part of the integrated genomic approaches. Genes like NEUROD6, ZCCHC17, PPEF1 and MANBAL were identified to be potentially

implicated in LOAD. The result of the study was published in 2015 [265].

4. A database to provide neuro-degenerative disease association references for molecules - *AlzExpress* - was developed. *AlzExpress* offers the functionality of query on biomolecule, such as RNA, protein, and provides reference for researchers to identify differentially expressed molecules (DEM) in different categories of samples, with high statistically significance. It also provides meta-analysis to combine statistics from separate studies in order to generalize the conclusions made the and improve the reliability. It is the first system with a web-based interface to integrate DEM analysis and meta-analysis, in additional to some dataset meta-information and overlapping DEMs between different categories of studies. By virtue of the fashionable NoSQL technology for data management, the system shows high scalability and compatibility to include various types of molecular data and different experiment designs.

Chapter 8

Conclusions and future work

8.1 Summary

The possibility to accurately and convincingly diagnose Alzheimer's disease (AD) in its earliest stages has gained public health attention and priority resulting from the growing prevalence of AD. The ability to diagnose early stage of AD is instantly important. Over the past decade, many machine learning and pattern classification methods have been used for early diagnosis of AD and MCI based on different modalities of biomarkers. With the advent of high throughput biomolecule detection technologies, such as next generation sequencing and microarray, and neuroimaging technique like MRI and PET, diverse means of diagnosis have been provided for the potential biomarker discovery.

My investigation of AD biomarkers started with looking into transcriptomic biomarker in brain and also studying the impacts of covariates, such as age, to the performances of AD classifiers. The result shows that the inclusion of age and APOE genotyping into the prediction model could enhance the performance of classifiers by around 8% in both training and testing. I have also discovered a panel of 26 RNA transcripts that has promising classification ability between AD patients and healthy controls, with the performance of over 90% training accuracy and over 85% testing accuracy. A random forest backward elimination is adopted as the major feature selection method, with the assist of a recursive cross validation process and weighted accuracy of training and

testing. The panel of transcripts turns out to be enriched in SNP, underpinned by more risky p-values compared to SNPs around randomly selected genes. Though brain biomarker may not have great potential in clinic diagnostic pratice, it offers great clues for understanding AD pathology, which if well exploited, can also be a good starting point for the discovery of new biomarkers in other tissue, such as blood.

The aim to discover non-invasive biomarker drove my interest into the blood-based AD biomarkers. A comprehensive literature review is carried out to gain existing biological knowledge of potential AD biomarkers with which a knowledge feature pool is constructed. A feature pool comprising of numerous AD related biomarkers was established which we used to select several panels of biomarkers. Two novel SVM-based feature selection methods were designed and deployed to select panels of serum features. With the selected panels of serum, several classifiers were trained by SVM from training dataset and then their performances were evaluated. After validation, we found that a panel of only two or three proteins gave us good diagnostic ability, with two of the three models estimated to have over 90% sensitivity and specificity. Beside, a novel statistical pattern was also recognised, separating case and normal with upper and lower boundaries. However, the reproducibility of these findings needs to be further validated in larger cohorts.

With the advent of novel high throughput technology and chronic accumulation of microarray based biological studies, meta-analysis is becoming more important in biomarker discovery. Nonetheless, after assessing the current meta-analytic R packages, many defects are discovered when being used in practice. First, current R packages cannot handle data from different studies with different dimensions which reduces the power of meta-analysis because of reduced sample size. Second, the presence of null values in the input data is ignored by directly omitting cases with missing values in these packages, which bias the true result of meta-analysis. Third, a lack of functions for users to conduct post-calculation analysis were commonly found in previous packages. To improve the usability and practicability of meta-analysis for bioinformatics research, a new meta-analysis package *metaUnion* was developed to fix the existing deficiencies

of current packages. It is able, first, to handle data from different studies with different dimensions; second, to rectify the number according to the number of missing value in the input; last, to provide the result of p-values, z-scores and effect sign for each gene. The validation of *metaUnion* reveals that it is a powerful R package with notable advantages. *metaUnion* can retain a maximum number of genes to be included in the analysis, thus a large proportion of DEGs that do not commonly probed across all the studies are not neglected - more than 70% in the validation test. *metaUnion* provides useful functions for downstream analysis once the major meta-analysis process is completed. It provides the functions to sort the result table by specific measurement and to plot Q-Q plot. In conclusion, *metaUnion* has advanced features compared to other existing meta-analytic R packages and is likely to gain popularity in conducting meta-analysis in microarray studies in the future.

As the project proceed further, a growing need of managing and integrating experiment data arises. Warehousing the data to make an efficient and fluent translation of data into solid results become increasingly a big concern. With the outstanding urge of this, the idea of developing a database to provide neuro-degenerative disease association references for molecules emerged. And *AlzExpress*, a NoSQL backed web-based platform is the answer. *AlzExpress* allows users to make queries on a biomolecule level, such as RNA, protein, and delivers reference with high statistically confidence for researchers to identify DEM in different categories of samples. Meta-analysis is embedded to combine statistics from separate studies so that the conclusions can be generalised. As the first system with a web-based interface to integrate DEM analysis and meta-analysis, it is also equipped with some dataset meta-information and functionalities like overlapping DEMs checking between different categories of studies. In support from the popular NoSQL database MongoDB, the system shows high scalability and excellent capability to extend its usage in the future.

8.2 Limitations

In general, the limitations of our studies includes the following,

- The investigation and search of biomarkers did not cnsider the cellular heterogeneity in AD. Therefore, the possibility that the different cellular types are contributing to the classifying performances of all models proposed cannot be eliminated.

- Only AD samples are included throughout all studies, but samples of other types of dementia are not included, such as MCI or Parkinson's disease (PD) samples. Since some individuals can have mixed dementia, so the complexity has not been excluded and properly explained. Also, this will obviously prevent the study from progressing further onto investigations on other types of dementia, and the potential of cross referencing between biomarkers from different types of disease in dementia will be wasted.

- In general, the model assessment logic in every study weighs more on evaluating the classifying ability of features themselves, while less on evaluating that of the predictive models. So the performances metrics reported are actually indicators of how good the selected features are on average, which in return will mislead the system to choose a good model but not the best one.

- Different datasets are used in different studies, so there is a lack of continuity and adherence for the results across studies.

8.3 Future work

Correlation based feature selection algorithms Although there are various kinds of feature selection methods that show promising performance in specific microarray datasets, almost all of the previous methods used in my past work have only considered the classifying or separating ability for a single feature, which leads to the fact that these

feature selection methods would only select features that are performing well on its own. However, human being is an entity in which genes or proteins are cooperating to maintain the biological function. Different types of relationships between genes, such as co-expression, regulation (either positive or negative), antagonism. . . , can indicate different physical condition an organism is in and thus can imply certain disease status. Therefore, the information from specific feature combination should be taken into account as an important signal during the process of feature selection. Unfortunately, most of the current feature selection methods did not include this aspect into their methodology. As a result, a feature selection method that can identify the feature set in combination with best classifying ability will be of great interest to researchers in biomarker discovery from a statistical learning perspective. Instead of identifying only those features that perform well separately, these algorithms will be comparatively more robust and more generalized findings that can be explained in a biological sense.

DNA methylation biomarkers Recently, the link between methylomic variation and AD has been studied. A number of genes were identified as potentially associated with LOAD [313], and some methylated regions in particular genes, such as ANK1 [314], also manifest association with neuropathology. In the light of exploring more biological truth in this aspect, we plan to investigate the probability of DNA methylomic variation as potential biomarkers for AD diagnosis. We are going to employ robust and effective feature selection and machine learning methods to current DNA methylation datasets, in the hope of discovering some loci which are meaningful with classifying ability and can be further used as diagnostic biomarker for AD.

Extension of AlzExpress *AlzExpress* provides a scalable web-based platform for the integrative gene-specific analysis for the investigation of dementia biomarkers from an expression level perspective. However, the potential of this platform is not yet fully exploited. For the next stage of work, Scaling up the database with more experiment data of more diseases from various sources will be of great research value. The high

scalability and schema-free nature of NoSQL database empower the platform with the potential to include other studies with different objectives diseases, such as Parkinson's disease or other types of dementia. Moreover, additional measurements or data type from a variety of scientific instruments can be encompassed into the platform, such as data of RNA sequencing, DNA methylation, microRNA expression, etc. The participation of these subjects will not only enrich the diversity of database, but also envision a good future of cross referencing and comparison of findings between different diseases and different types of biomolecules.

Graphene biosensors for AD diagnosis From the perspective of developing practical AD diagnostic instruments, new material will come on stage with the aim of bringing excitement. Graphene biosensors exploit the promising physical, chemical and electronic properties of the atomic-thick 2-dimentional material to develop a new generation of label-free biosensors and sensor arrays with unprecedented sensitivities [315], which can be used for the detection of the abnormal proteins/biomolecules of AD noninvasively and at lower levels of concentrations than existing methods such as ELISA. In the past several years, numerous improvement and breakthrough were made in graphene biosensors research, and the engineering group from project BBDiag (Blood Biomarker-based Diagnostic Tools for Early-stage Alzheimer's Disease) in Plymouth University has been largely contributing to the remarks. Outstanding breakthrough was made in progressing graphene label-free biosensors from this group [316–318]. In the light of these biosensors with high sensitivity, an AD diagnostic system combining this sensor with multiple significant biomarkers from transcript array, protein array and methylation array data is pragmatically more possible. The application of the potential multi-biomarkers to design biosensor matrix integrating with other diagnostic methods such as MRI image and EEG signal of objects, will further accelerate the realization of AD prognosis and diagnosis.

Academic institutions, the pharmaceutical industry and regulatory organizations all agree that biomarkers have an important role in the drug development process. Throughout my PhD the discovery of AD biomarkers by specifically statistical learning approaches

has been the main focus. Biomarkers have several potential uses in clinical trials. These include their use as diagnostic aids to enrich the patient sample with cases of AD; as tools to identify and monitor the biochemical effect of the drug candidate; and as safety markers to detect potential side effects of the drug. We followed strictly two directions during the course of the entire research - biomarker driven studies and software driven studies. Therefore, on one hand, novel findings were discovered in both RNA and blood as tissue. On the other hand, fashionable software packages and platform were developed to suit the need of future researchers in this field. We believe by the collaborative work from different generations, the mystery of AD will finally be unveiled and a practical medical solution will appear in the end.

Bibliography

[1] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] C. Qiu, M. Kivipelto, and E. von Strauss, "Epidemiology of alzheimer's disease: occurrence, determinants, and strategies toward intervention," *Dialogues in clinical neuroscience*, vol. 11, no. 2, p. 111, 2009.

[3] S. Y. Bookheimer, M. H. Strojwas, M. S. Cohen, A. M. Saunders, M. A. Pericak-Vance, J. C. Mazziotta, and G. W. Small, "Patterns of brain activation in people at risk for alzheimer's disease," *New England journal of medicine*, vol. 343, no. 7, pp. 450–456, 2000.

[4] H. Braak and E. Braak, "Evolution of neuronal changes in the course of alzheimer's disease," in *Ageing and dementia*. Springer, 1998, pp. 127–140.

[5] M. Bruscoli and S. Lovestone, "Is mci really just early dementia? a systematic review of conversion studies," *International Psychogeriatrics*, vol. 16, no. 2, pp. 129–140, 2004.

[6] C. Flicker, S. H. Ferris, and B. Reisberg, "Mild cognitive impairment in the elderly predictors of dementia," *Neurology*, vol. 41, no. 7, pp. 1006–1006, 1991.

[7] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta neuropathologica*, vol. 82, no. 4, pp. 239–259, 1991.

[8] ——, "Staging of alzheimer's disease-related neurofibrillary changes," *Neurobiology of aging*, vol. 16, no. 3, pp. 271–278, 1995.

[9] H. Braak, E. Braak, and J. Bohl, "Staging of alzheimer-related cortical destruction," *European neurology*, vol. 33, no. 6, pp. 403–408, 1993.

[10] T. Gómez-Isla, J. L. Price, D. W. McKeel Jr, J. C. Morris, J. H. Growdon, and B. T. Hyman, "Profound loss of layer ii entorhinal cortex neurons occurs in very mild alzheimer's disease," *Journal of Neuroscience*, vol. 16, no. 14, pp. 4491–4500, 1996.

[11] B. T. Hyman, G. W. Van Hoesen, A. R. Damasio, and C. L. Barnes, "Alzheimer's disease: cell-specific pathology isolates the hippocampal formation," *Science*, vol. 225, no. 4667, pp. 1168–1170, 1984.

[12] J. H. Kordower, Y. Chu, G. T. Stebbins, S. T. DeKosky, E. J. Cochran, D. Bennett, and E. J. Mufson, "Loss and atrophy of layer ii entorhinal cortex neurons in elderly people with mild cognitive impairment," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 49, no. 2, pp. 202–213, 2001.

[13] E. Mufson, E. Chen, E. Cochran, L. Beckett, D. Bennett, and J. Kordower, "Entorhinal cortex β-amyloid load in individuals with mild cognitive impairment," *Experimental neurology*, vol. 158, no. 2, pp. 469–490, 1999.

[14] B. Seltzer and I. Sherwin, "A comparison of clinical features in early-and late-onset primary degenerative dementia: one entity or two?" *Archives of neurology*, vol. 40, no. 3, pp. 143–146, 1983.

[15] A. Awada, "Early and late-onset alzheimer's disease: What are the differences?" *Journal of neurosciences in rural practice*, vol. 6, no. 3, p. 455, 2015.

[16] A. Kumar and S. Dogra, "Neuropathology and therapeutic management of alzheimer's disease–an update," *Drugs of the Future*, vol. 33, no. 5, pp. 433–446, 2008.

[17] A. Kurz and R. Perneczky, "Novel insights for the treatment of alzheimer's disease," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 373–379, 2011.

[18] J. Hardy, "The amyloid hypothesis for alzheimer's disease: a critical reappraisal," *Journal of neurochemistry*, vol. 110, no. 4, pp. 1129–1134, 2009.

[19] R. Anand, K. D. Gill, and A. A. Mahdi, "Therapeutics of alzheimer's disease: Past, present and future," *Neuropharmacology*, vol. 76, pp. 27–50, 2014.

[20] I. Dal Prà, A. Chiarini, L. Gui, B. Chakravarthy, R. Pacchiana, E. Gardenal, J. F. Whitfield, and U. Armato, "Do astrocytes collaborate with neurons in spreading the "infectious" aβ and tau drivers of alzheimer's disease?" *The Neuroscientist*, vol. 21, no. 1, pp. 9–29, 2015.

[21] D. Galimberti, L. Ghezzi, and E. Scarpini, "Immunotherapy against amyloid pathology in alzheimer's disease," *Journal of the neurological sciences*, vol. 333, no. 1-2, pp. 50–54, 2013.

[22] S. Salomone, F. Caraci, G. M. Leggio, J. Fedotova, and F. Drago, "New pharmacological strategies for treatment of alzheimer's disease: focus on disease modifying drugs," *British journal of clinical pharmacology*, vol. 73, no. 4, pp. 504–517, 2012.

[23] H. Rosenmann, "Immunotherapy for targeting tau pathology in alzheimer's disease and tauopathies," *Current Alzheimer Research*, vol. 10, no. 3, pp. 217–228, 2013.

[24] D. S. Antanitus, "A theory of cortical neuron-astrocyte interaction," *The Neuroscientist*, vol. 4, no. 3, pp. 154–159, 1998.

[25] A. D. Roth, G. Ramírez, R. Alarcón, and R. Von Bernhardi, "Oligodendrocytes damage in alzheimer's disease: beta amyloid toxicity and inflammation," *Biological research*, vol. 38, no. 4, pp. 381–387, 2005.

[26] S. Subasinghe, S. Unabia, C. J. Barrow, S. S. Mok, M.-I. Aguilar, and D. H. Small, "Cholesterol is necessary both for the toxic effect of aβ peptides on vascular smooth muscle cells and for aβ binding to vascular smooth muscle cell membranes," *Journal of neurochemistry*, vol. 84, no. 3, pp. 471–479, 2003.

[27] Y. Xu, J. Yan, P. Zhou, J. Li, H. Gao, Y. Xia, and Q. Wang, "Neurotransmitter receptors and cognitive dysfunction in alzheimer's disease and parkinson's disease," *Progress in neurobiology*, vol. 97, no. 1, pp. 1–13, 2012.

[28] U. Armato, B. Chakravarthy, R. Pacchiana, and J. F. Whitfield, "Alzheimer's disease: An update of the roles of receptors, astrocytes and primary cilia," *International journal of molecular medicine*, vol. 31, no. 1, pp. 3–10, 2013.

[29] K. Yasojima, E. McGeer, and P. McGeer, "Relationship between beta amyloid peptide generating molecules and neprilysin in alzheimer disease and normal brain," *Brain research*, vol. 919, no. 1, pp. 115–121, 2001.

[30] H. Braak and K. Del Tredici, "Alzheimer's pathogenesis: is there neuron-to-neuron propagation?" *Acta neuropathologica*, vol. 121, no. 5, pp. 589–595, 2011.

[31] M. Talantova, S. Sanz-Blasco, X. Zhang, P. Xia, M. W. Akhtar, S.-i. Okamoto, G. Dziewczapolski, T. Nakamura, G. Cao, A. E. Pratt *et al.*, "Aβ induces astrocytic glutamate release, extrasynaptic nmda receptor activation, and synaptic loss," *Proceedings of the National Academy of Sciences*, vol. 110, no. 27, pp. E2518–E2527, 2013.

[32] E. E. Tuppo and H. R. Arias, "The role of inflammation in alzheimer's disease," *The international journal of biochemistry & cell biology*, vol. 37, no. 2, pp. 289–305, 2005.

[33] Y. S. Eisele, U. Obermüller, G. Heilbronner, F. Baumann, S. A. Kaeser, H. Wolburg, L. C. Walker, M. Staufenbiel, M. Heikenwalder, and M. Jucker, "Peripherally applied aβ-containing inoculates induce cerebral β-amyloidosis," *Science*, p. 1194516, 2010.

[34] H. Hampel, S. Lista, S. J. Teipel, F. Garaci, R. Nisticò, K. Blennow, H. Zetterberg, L. Bertram, C. Duyckaerts, H. Bakardjian *et al.*, "Perspective on future role of biological markers in clinical therapy trials of alzheimer's disease: a long-range point of view beyond 2020," *Biochemical pharmacology*, vol. 88, no. 4, pp. 426–449, 2014.

[35] M. Vandermeeren, M. Mercken, E. Vanmechelen, J. Six, A. Voorde, J.-J. Martin, and P. Cras, "Detection of proteins in normal and alzheimer's disease cerebrospinal fluid with a sensitive sandwich enzyme-linked immunosorbent assay," *Journal of neurochemistry*, vol. 61, no. 5, pp. 1828–1834, 1993.

[36] K. Blennow, "Cerebrospinal fluid protein biomarkers for alzheimer's disease," *NeuroRx*, vol. 1, no. 2, pp. 213–225, 2004.

[37] R. Tarawneh and D. M. Holtzman, "Biomarkers in translational research of alzheimer's disease," *Neuropharmacology*, vol. 59, no. 4, pp. 310–322, 2010.

[38] M. Shoji, E. Matsubara, M. Kanai, M. Watanabe, T. Nakamura, Y. Tomidokoro, M. Shizuka, K. Wakabayashi, Y. Igeta, Y. Ikeda *et al.*, "Combination assay of csf tau, aβ1-40 and aβ1-42 (43) as a biochemical marker of alzheimer's disease," *Journal of the neurological sciences*, vol. 158, no. 2, pp. 134–140, 1998.

[39] V. Welge, O. Fiege, P. Lewczuk, B. Mollenhauer, H. Esselmann, H.-W. Klafki, S. Wolf, C. Trenkwalder, M. Otto, J. Kornhuber *et al.*, "Combined csf tau, p-tau181 and amyloid-β 38/40/42 for diagnosing alzheimer's disease," *Journal of neural transmission*, vol. 116, no. 2, pp. 203–212, 2009.

[40] E. Kaiser, P. Schoenknecht, S. Kassner, W. Hildebrandt, R. Kinscherf, and J. Schroeder, "Cerebrospinal fluid concentrations of functionally important amino acids and metabolic compounds in patients with mild cognitive impairment and alzheimer's disease," *Neurodegenerative Diseases*, vol. 7, no. 4, pp. 251–259, 2010.

[41] C. Czech, P. Berndt, K. Busch, O. Schmitz, J. Wiemer, V. Most, H. Hampel, J. Kastler, and H. Senn, "Metabolite profiling of alzheimer's disease cerebrospinal fluid," *PLoS one*, vol. 7, no. 2, p. e31501, 2012.

[42] M. van Oijen, A. Hofman, H. D. Soares, P. J. Koudstaal, and M. M. Breteler, "Plasma aβ 1–40 and aβ 1–42 and the risk of dementia: a prospective case-cohort study," *The Lancet Neurology*, vol. 5, no. 8, pp. 655–660, 2006.

[43] R. Mayeux, L. Honig, M.-X. Tang, J. Manly, Y. Stern, N. Schupf, and P. Mehta, "Plasma aβ40 and aβ42 and alzheimer's disease relation to age, mortality, and risk," *Neurology*, vol. 61, no. 9, pp. 1185–1190, 2003.

[44] O. Hansson, H. Zetterberg, E. Vanmechelen, H. Vanderstichele, U. Andreasson, E. Londos, A. Wallin, L. Minthon, and K. Blennow, "Evaluation of plasma aβ 40 and aβ 42 as predictors of conversion to alzheimer's disease in patients with mild cognitive impairment," *Neurobiology of aging*, vol. 31, no. 3, pp. 357–367, 2010.

[45] O. Lopez, L. Kuller, P. Mehta, J. Becker, H. Gach, R. Sweet, Y. Chang, R. Tracy, and S. DeKosky, "Plasma amyloid levels and the risk of ad in normal subjects in the cardiovascular health study," *Neurology*, vol. 70, no. 19, pp. 1664–1671, 2008.

[46] J. Sundelöf, V. Giedraitis, M. C. Irizarry, J. Sundström, E. Ingelsson, E. Rönnemaa, J. Ärnlöv, M. D. Gunnarsson, B. T. Hyman, H. Basun *et al.*, "Plasma β amyloid and the risk of alzheimer disease and dementia in elderly men: a prospective, population-based cohort study," *Archives of Neurology*, vol. 65, no. 2, pp. 256–263, 2008.

[47] N. R. Graff-Radford, J. E. Crook, J. Lucas, B. F. Boeve, D. S. Knopman, R. J. Ivnik, G. E. Smith, L. H. Younkin, R. C. Petersen, and S. G. Younkin, "Association of low plasma aβ42/aβ40 ratios with increased imminent risk for mild cognitive impairment and alzheimer disease," *Archives of neurology*, vol. 64, no. 3, pp. 354–362, 2007.

[48] K. Yaffe, A. Weston, N. R. Graff-Radford, S. Satterfield, E. M. Simonsick, S. G. Younkin, L. H. Younkin, L. Kuller, H. N. Ayonayon, J. Ding *et al.*, "Association of plasma β-amyloid level and cognitive reserve with subsequent cognitive decline," *Jama*, vol. 305, no. 3, pp. 261–266, 2011.

[49] R. Mayeux, M.-X. Tang, D. M. Jacobs, J. Manly, K. Bell, C. Merchant, S. A. Small, Y. Stern, H. M. Wisniewski, and P. D. Mehta, "Plasma amyloid β-peptide 1–42 and incipient alzheimer's disease," *Annals of neurology*, vol. 46, no. 3, pp. 412–416, 1999.

[50] A. Koyama, O. I. Okereke, T. Yang, D. Blacker, D. J. Selkoe, and F. Grodstein, "Plasma amyloid-β as a predictor of dementia and cognitive decline: a systematic review and meta-analysis," *Archives of neurology*, vol. 69, no. 7, pp. 824–831, 2012.

[51] J. Randall, E. Mörtberg, G. K. Provuncher, D. R. Fournier, D. C. Duffy, S. Rubertsson, K. Blennow, H. Zetterberg, and D. H. Wilson, "Tau proteins in serum predict neurological outcome after hypoxic brain injury from cardiac arrest: results of a pilot study," *Resuscitation*, vol. 84, no. 3, pp. 351–356, 2013.

[52] C. Hesse, L. Rosengren, N. Andreasen, P. Davidsson, H. Vanderstichele, E. Vanmechelen, and K. Blennow, "Transient increase in total tau but not phospho-tau in human cerebrospinal fluid after acute stroke," *Neuroscience letters*, vol. 297, no. 3, pp. 187–190, 2001.

[53] C. Rosén, O. Hansson, K. Blennow, and H. Zetterberg, "Fluid biomarkers in alzheimer's disease–current concepts," *Molecular neurodegeneration*, vol. 8, no. 1, p. 20, 2013.

[54] H. Zetterberg, D. Wilson, U. Andreasson, L. Minthon, K. Blennow, J. Randall, and O. Hansson, "Plasma tau levels in alzheimer's disease," *Alzheimer's research & therapy*, vol. 5, no. 2, p. 9, 2013.

[55] S. Ray, P. J. Reddy, R. Jain, K. Gollapalli, A. Moiyadi, and S. Srivastava, "Proteomic technologies for the identification of disease biomarkers in serum: advances and challenges ahead," *Proteomics*, vol. 11, no. 11, pp. 2139–2161, 2011.

[56] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, D. R. Galasko, M. Jutel, A. Karydas *et al.*, "Classification and prediction of clinical alzheimer's diagnosis based on plasma signaling proteins." *Nature medicine*, vol. 13, no. 11, 2007.

[57] W. T. Hu, D. M. Holtzman, A. M. Fagan, L. M. Shaw, R. Perrin, S. E. Arnold, M. Grossman, C. Xiong, R. Craig-Schapiro, C. M. Clark *et al.*, "Plasma multianalyte profiling in mild cognitive impairment and alzheimer disease," *Neurology*, vol. 79, no. 9, pp. 897–905, 2012.

[58] M. Björkqvist, M. Ohlsson, L. Minthon, and O. Hansson, "Evaluation of a previously suggested plasma biomarker panel to identify alzheimer's disease," *PloS one*, vol. 7, no. 1, p. e29868, 2012.

[59] J. D. Doecke, S. M. Laws, N. G. Faux, W. Wilson, S. C. Burnham, C.-P. Lam, A. Mondal, J. Bedo, A. I. Bush, B. Brown *et al.*, "Blood-based protein biomarkers for diagnosis of alzheimer disease," *Archives of neurology*, vol. 69, no. 10, pp. 1318–1325, 2012.

[60] S. E. O'Bryant, G. Xiao, R. Barber, R. Huebinger, K. Wilhelmsen, M. Edwards, N. Graff-Radford, R. Doody, R. Diaz-Arrastia, T. A. R. . C. Consortium *et al.*, "A

blood-based screening tool for alzheimer's disease that spans serum and plasma: findings from tarc and adni," *PloS one*, vol. 6, no. 12, p. e28092, 2011.

[61] B. Dubois, H. H. Feldman, C. Jacova, J. L. Cummings, S. T. DeKosky, P. Barberger-Gateau, A. Delacourte, G. Frisoni, N. C. Fox, D. Galasko *et al.*, "Revising the definition of alzheimer's disease: a new lexicon," *The Lancet Neurology*, vol. 9, no. 11, pp. 1118–1127, 2010.

[62] C. R. Jack, M. S. Albert, D. S. Knopman, G. M. McKhann, R. A. Sperling, M. C. Carrillo, B. Thies, and C. H. Phelps, "Introduction to the recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 257–262, 2011.

[63] L. Clerx, P. J. Visser, F. Verhey, and P. Aalten, "New mri markers for alzheimer's disease: a meta-analysis of diffusion tensor imaging and a comparison with medial temporal lobe measurements," *Journal of Alzheimer's Disease*, vol. 29, no. 2, pp. 405–429, 2012.

[64] 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*, vol. 36, no. 1, pp. 57–66, 2013.

[65] M. Laakso, K. Partanen, P. Riekkinen, M. Lehtovirta, E.-L. Helkala, M. Hallikainen, T. Hanninen, P. Vainio, and H. Soininen, "Hippocampal volumes in alzheimer's disease, parkinson's disease with and without dementia, and in vascular dementia an mri study," *Neurology*, vol. 46, no. 3, pp. 678–681, 1996.

[66] 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*, vol. 51, no. 2, pp. 357–362, 1998.

[67] P. Videbech and B. Ravnkilde, "Hippocampal volume and depression: a meta-analysis of mri studies," *American Journal of Psychiatry*, vol. 161, no. 11, pp. 1957–1966, 2004.

[68] N. Fox, R. Black, S. Gilman, M. Rossor, S. Griffith, L. Jenkins, M. Koller *et al.*, "Effects of aβ immunization (an1792) on mri measures of cerebral volume in alzheimer disease," *Neurology*, vol. 64, no. 9, pp. 1563–1572, 2005.

[69] C. Zarow, H. V. Vinters, W. G. Ellis, M. W. Weiner, D. Mungas, L. White, and H. C. Chui, "Correlates of hippocampal neuron number in alzheimer's disease and ischemic vascular dementia," *Annals of neurology*, vol. 57, no. 6, pp. 896–903, 2005.

[70] M. Bobinski, M. De Leon, J. Wegiel, S. Desanti, A. Convit, L. Saint Louis, H. Rusinek, and H. Wisniewski, "The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in alzheimer's disease," *Neuroscience*, vol. 95, no. 3, pp. 721–725, 1999.

[71] S. H. Freeman, R. Kandel, L. Cruz, A. Rozkalne, K. Newell, M. P. Frosch, E. T. Hedley-Whyte, J. J. Locascio, L. A. Lipsitz, and B. T. Hyman, "Preservation of neuronal number despite age-related cortical brain atrophy in elderly subjects without alzheimer disease," *Journal of Neuropathology & Experimental Neurology*, vol. 67, no. 12, pp. 1205–1212, 2008.

[72] P. J. Basser and D. K. Jones, "Diffusion-tensor mri: theory, experimental design and data analysis–a technical review," *NMR in Biomedicine*, vol. 15, no. 7-8, pp. 456–467, 2002.

[73] L. O'Dwyer, F. Lamberton, A. L. Bokde, M. Ewers, Y. O. Faluyi, C. Tanner, B. Mazoyer, D. O'Neill, M. Bartley, D. R. Collins *et al.*, "Multiple indices of diffusion identifies white matter damage in mild cognitive impairment and alzheimer's disease," *PloS one*, vol. 6, no. 6, p. e21745, 2011.

[74] T. C. Chua, W. Wen, M. J. Slavin, and P. S. Sachdev, "Diffusion tensor imaging in mild cognitive impairment and alzheimer's disease: a review," *Current opinion in neurology*, vol. 21, no. 1, pp. 83–92, 2008.

[75] M. Bozzali and A. Cherubini, "Diffusion tensor mri to investigate dementias: a brief review," *Magnetic resonance imaging*, vol. 25, no. 6, pp. 969–977, 2007.

[76] P. R. Borghesani, L. C. Johnson, A. L. Shelton, E. R. Peskind, E. H. Aylward, G. D. Schellenberg, and M. M. Cherrier, "Altered medial temporal lobe responses during visuospatial encoding in healthy apoe* 4 carriers," *Neurobiology of aging*, vol. 29, no. 7, pp. 981–991, 2008.

[77] N. Filippini, B. J. MacIntosh, M. G. Hough, G. M. Goodwin, G. B. Frisoni, S. M. Smith, P. M. Matthews, C. F. Beckmann, and C. E. Mackay, "Distinct patterns of brain activity in young carriers of the apoe-ε4 allele," *Proceedings of the National Academy of Sciences*, vol. 106, no. 17, pp. 7209–7214, 2009.

[78] K. Shoghi-Jadid, G. W. Small, E. D. Agdeppa, V. Kepe, L. M. Ercoli, P. Siddarth, S. Read, N. Satyamurthy, A. Petric, S.-C. Huang *et al.*, "Localization of neurofibrillary tangles and beta-amyloid plaques in the brains of living patients with alzheimer disease," *The American Journal of Geriatric Psychiatry*, vol. 10, no. 1, pp. 24–35, 2002.

[79] S. M. Landau, D. Harvey, C. M. Madison, R. A. Koeppe, E. M. Reiman, N. L. Foster, M. W. Weiner, W. J. Jagust, A. D. N. Initiative *et al.*, "Associations between cognitive, functional, and fdg-pet measures of decline in ad and mci," *Neurobiology of aging*, vol. 32, no. 7, pp. 1207–1218, 2011.

[80] J. B. Langbaum, K. Chen, W. Lee, C. Reschke, D. Bandy, A. S. Fleisher, G. E. Alexander, N. L. Foster, M. W. Weiner, R. A. Koeppe *et al.*, "Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the alzheimer's disease neuroimaging initiative (adni)," *Neuroimage*, vol. 45, no. 4, pp. 1107–1116, 2009.

[81] W. J. Jagust, S. Landau, L. Shaw, J. Trojanowski, R. Koeppe, E. Reiman, N. Foster, R. C. Petersen, M. Weiner, J. Price *et al.*, "Relationships between biomarkers in aging and dementia," *Neurology*, vol. 73, no. 15, pp. 1193–1199, 2009.

[82] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen *et al.*, "The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[83] M. Cruts, J. Theuns, and C. Van Broeckhoven, "Locus-specific mutation databases for neurodegenerative brain diseases," *Human mutation*, vol. 33, no. 9, pp. 1340–1344, 2012.

[84] W. J. Strittmatter, A. M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G. S. Salvesen, and A. D. Roses, "Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease." *Proceedings of the National Academy of Sciences*, vol. 90, no. 5, pp. 1977–1981, 1993.

[85] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, vol. 45, no. 12, pp. 1452–1458, 2013.

[86] D. S. Auld, T. J. Kornecook, S. Bastianetto, and R. Quirion, "Alzheimer's disease and the basal forebrain cholinergic system: relations to β-amyloid peptides, cognition, and treatment strategies," *Progress in neurobiology*, vol. 68, no. 3, pp. 209–245, 2002.

[87] M. R. Farlow, M. L. Miller, and V. Pejovic, "Treatment options in alzheimer's disease: maximizing benefit, managing expectations," *Dementia and geriatric cognitive disorders*, vol. 25, no. 5, pp. 408–422, 2008.

[88] M. J. Aardema and J. T. MacGregor, "Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies," in *Toxicogenomics*. Springer, 2003, pp. 171–193.

[89] J. Piehler, "New methodologies for measuring protein interactions in vivo and in vitro," *Current opinion in structural biology*, vol. 15, no. 1, pp. 4–14, 2005.

[90] W. Kusnezow and J. D. Hoheisel, "Solid supports for microarray immunoassays," *Journal of Molecular Recognition*, vol. 16, no. 4, pp. 165–176, 2003.

[91] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[92] A. G. De Brevern, S. Hazout, and A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering," *BMC bioinformatics*, vol. 5, no. 1, p. 114, 2004.

[93] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparison of methods for image analysis on cdna microarray data," *Journal of computational and graphical statistics*, vol. 11, no. 1, pp. 108–136, 2002.

[94] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[95] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[96] T. H. Bø, B. Dysvik, and I. Jonassen, "Lsimpute: accurate estimation of missing values in microarray data with least squares methods," *Nucleic acids research*, vol. 32, no. 3, pp. e34–e34, 2004.

[97] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for dna microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2004.

[98] K.-Y. Kim, B.-J. Kim, and G.-S. Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," *BMC bioinformatics*, vol. 5, no. 1, p. 160, 2004.

[99] L. P. Brás and J. C. Menezes, "Improving cluster-based missing value estimation of dna microarray data," *Biomolecular engineering*, vol. 24, no. 2, pp. 273–282, 2007.

[100] R. Jörnsten, H.-Y. Wang, W. J. Welsh, and M. Ouyang, "Dna microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, no. 22, pp. 4155–4161, 2005.

[101] X. Gan, A. W.-C. Liew, and H. Yan, "Microarray missing data imputation based on a set theoretic framework and biological knowledge," *Nucleic Acids Research*, vol. 34, no. 5, pp. 1608–1619, 2006.

[102] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in microarray data with gene ontology," *Bioinformatics*, vol. 22, no. 5, pp. 566–572, 2005.

[103] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[104] G. Tseng, M. Oh, L. Rohlin, J. Liao, and W. Wong, "Issues in cdna microarray analysis: quality filtering," 2001.

[105] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cdna microarray images," *Journal of Biomedical optics*, vol. 2, no. 4, pp. 364–374, 1997.

[106] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic acids research*, vol. 30, no. 4, pp. e15–e15, 2002.

[107] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li *et al.*, "Within the fold: assessing differential expression measures and reproducibility in microarray assays," *Genome biology*, vol. 3, no. 11, pp. research0062–1, 2002.

[108] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[109] J. Quackenbush, "Microarray data normalization and transformation," *Nature genetics*, vol. 32, pp. 496–501, 2002.

[110] P. R. Bevington, D. K. Robinson, J. M. Blair, A. J. Mallinckrodt, S. McKay *et al.*, "Data reduction and error analysis for the physical sciences," *Computers in Physics*, vol. 7, no. 4, pp. 415–416, 1993.

[111] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[112] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[113] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European conference on machine learning*. Springer, 1994, pp. 171–182.

[114] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data mining and knowledge discovery*, vol. 2, no. 4, pp. 345–389, 1998.

[115] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[116] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 2007.

[117] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[118] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[119] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[120] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.* IEEE, 1999, pp. 41–48.

[121] I. Kononenko, "Inductive and bayesian learning in medical diagnosis," *Applied Artificial Intelligence an International Journal*, vol. 7, no. 4, pp. 317–337, 1993.

[122] D. G. Kleinbaum, L. L. Kupper, and L. E. Chambless, "Logistic regression analysis of epidemiologic data: theory and practice," *Communications in Statistics-Theory and Methods*, vol. 11, no. 5, pp. 485–547, 1982.

[123] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," CORNELL AERONAUTICAL LAB INC BUFFALO NY, Tech. Rep., 1961.

[124] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and computation*, vol. 108, no. 2, pp. 212–261, 1994.

[125] P. Auer and M. K. Warmuth, "Tracking the best disjunction," *Machine Learning*, vol. 32, no. 2, pp. 127–150, 1998.

[126] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.

[127] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[128] C. Neocleous and C. Schizas, "Artificial neural network learning: A comparative review," *Methods and Applications of Artificial Intelligence*, pp. 750–750, 2002.

[129] L. Camargo and T. Yoneyama, "Specification of training sets and the number of hidden neurons for multilayer perceptrons," *Neural computation*, vol. 13, no. 12, pp. 2673–2680, 2001.

[130] M. A. Kon and L. Plaskota, "Information complexity of neural networks," *Neural Networks*, vol. 13, no. 3, pp. 365–375, 2000.

[131] J. Y. Yam and T. W. Chow, "Feedforward networks training speed enhancement by optimal initialization of the synaptic coefficients," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 430–434, 2001.

[132] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in *Advances in neural information processing systems*, 1991, pp. 875–882.

[133] M. Siddique and M. Tokhi, "Training neural networks: backpropagation vs. genetic algorithms," in *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, vol. 4. IEEE, 2001, pp. 2673–2678.

[134] G. G. Yen and H. Lu, "Hierarchical genetic algorithm based neural network design," in *Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on*. IEEE, 2000, pp. 168–175.

[135] F. Vivarelli and C. K. Williams, "Comparing bayesian neural network algorithms for classifying segmented outdoor images," *Neural Networks*, vol. 14, no. 4, pp. 427–437, 2001.

[136] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[137] K. Veropoulos, C. Campbell, N. Cristianini *et al.*, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, 1999, pp. 55–60.

[138] B. Schölkopf, C. J. Burges, and A. J. Smola, *Advances in kernel methods: support vector learning*. MIT press, 1999.

[139] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," *Journal of machine learning research*, vol. 2, no. Dec, pp. 299–312, 2001.

[140] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[141] A. B. Tucker, *Computer science handbook*. CRC press, 2004.

[142] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[143] N. Jardine and R. Sibson, "Mathematical taxonomy," *London etc.: John Wiley*, 1971.

[144] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[145] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[146] A. Cauchy, "Sur l'equation á l'aide de laquelle on détermine les inégalites séculaires des mouvements des planétes [on the equation by which the inequalities for the secular movement of planets is determined]," *Oeuvres Completes (II*, 1829.

[147] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[148] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[149] S. Ahmed, M. Zhang, and L. Peng, "Feature selection and classification of high dimensional mass spectrometry data: A genetic programming approach." in *EvoBIO*. Springer, 2013, pp. 43–55.

[150] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[151] I. J. Good, "Probability and the weighing of evidence," 1950.

[152] B. Cestnik, I. Kononenko, and I. Bratko, "Assistant 86: A knowledge-elicitation tool for sophisticated users," in *Proceedings of the 2nd European Conference on European Working Session on Learning*. Sigma Press, 1987, pp. 31–45.

[153] B. Cestnik *et al.*, "Estimating probabilities: a crucial task in machine learning." in *ECAI*, vol. 90, 1990, pp. 147–149.

[154] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[155] I. Kononenko, "Semi-naive bayesian classifier," in *Machine Learning—EWSL-91*. Springer, 1991, pp. 206–219.

[156] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.

[157] S. GreenAM, "Signaldetectiontheoryandpsychophysics," 1966.

[158] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[159] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.

[160] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[161] ——, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[162] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[163] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[164] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the national academy of sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.

[165] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[166] C. Sima, U. Braga-Neto, and E. R. Dougherty, "Superior feature-set ranking for small samples using bolstered error estimation," *Bioinformatics*, vol. 21, no. 7, pp. 1046–1054, 2004.

[167] W. J. Fu, R. J. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 21, no. 9, pp. 1979–1986, 2005.

[168] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays," *Cancer research*, vol. 62, no. 15, pp. 4427–4433, 2002.

[169] G. Schwarzer, "meta: An R package for meta-analysis," *R News*, vol. 7, no. 3, pp. 40–45, 2007.

[170] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *Journal of Statistical Software*, vol. 36, no. 3, pp. 1–48, 2010. [Online]. Available: http://www.jstatsoft.org/v36/i03/

[171] G. Marot, J.-L. Foulley, C.-D. Mayer, and F. Jaffrézic, "Moderated effect size and p-value combinations for microarray meta-analyses," *Bioinformatics*, vol. 25, no. 20, pp. 2692–2699, 2009.

[172] J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling, "Multiclass cancer classification and biomarker discovery using ga-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[173] L. Li, T. A. Darden, C. Weingberg, A. Levine, and L. G. Pedersen, "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method," *Combinatorial chemistry & high throughput screening*, vol. 4, no. 8, pp. 727–739, 2001.

[174] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[175] T. Li, S. Zhu, Q. Li, and M. Ogihara, "Gene functional classification by semi-supervised learning from heterogeneous data," in *Proceedings of the 2003 ACM symposium on Applied computing*. ACM, 2003, pp. 78–82.

[176] Y. Qi, P. E. Missiuro, A. Kapoor, C. P. Hunter, T. S. Jaakkola, D. K. Gifford, and H. Ge, "Semi-supervised analysis of gene expression profiles for lineage-specific development in the caenorhabditis elegans embryo," *Bioinformatics*, vol. 22, no. 14, pp. e417–e423, 2006.

[177] C. Harris and N. Ghaffari, "Biomarker discovery across annotated and unannotated microarray datasets using semi-supervised learning," *BMC genomics*, vol. 9, no. 2, p. S7, 2008.

[178] F. Q. Calvo, M. Fillet, D. De Seny, M.-A. Meuwis, R. Marée, C. Crahay, G. Paulissen, N. Rocks, M. Gueders, L. Wehenkel *et al.*, "Biomarker discovery in asthma-related inflammation and remodeling," *Proteomics*, vol. 9, no. 8, pp. 2163–2170, 2009.

[179] O. Bembom, M. L. Petersen, S.-Y. Rhee, W. J. Fessel, S. E. Sinisi, R. W. Shafer, and M. J. van der Laan, "Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection," *Statistics in medicine*, vol. 28, no. 1, pp. 152–172, 2009.

[180] R.-L. Wang, D. Bencic, A. Biales, D. Lattier, M. Kostich, D. Villeneuve, G. T. Ankley, J. Lazorchak, and G. Toth, "Dna microarray-based ecotoxicological biomarker discovery in a small fish model species," *Environmental toxicology and chemistry*, vol. 27, no. 3, pp. 664–675, 2008.

[181] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical systems and signal processing*, vol. 21, no. 6, pp. 2560–2574, 2007.

[182] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[183] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[184] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[185] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[186] J. Zhang, I. Sokal, E. R. Peskind, J. F. Quinn, J. Jankovic, C. Kenney, K. A. Chung, S. P. Millard, J. G. Nutt, and T. J. Montine, "Csf multianalyte profile distinguishes alzheimer and parkinson diseases," *American journal of clinical pathology*, vol. 129, no. 4, pp. 526–529, 2008.

[187] S. Sharma, R. Shandilya, S. Patnaik, and A. Mahapatra, "Leading nosql models for handling big data: a brief review," *International Journal of Business Information Systems*, vol. 22, no. 1, pp. 1–25, 2016.

[188] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS operating systems review*, vol. 37, no. 5. ACM, 2003, pp. 29–43.

[189] J. Copper, "How entities and indexes are stored," *Google App Engine, Google Code*, 2009.

[190] K. Cordes, "Youtube scalability (talk), their new solution for thumbnails is to use google's bigtable, which provides high performance for a large number of rows, fault tolerance, caching, etc," *This is a nice (and rare?) example of actual synergy in an acquisition*, 2007.

[191] A. Hitchcock, "Google's bigtable," *There are currently around 100 cells for services such as Print, Search History, Maps, and Orkut*, 2005.

[192] A. Lakshman and P. Malik, "The apache cassandra project," *Retrieved from the World Wide Web October*, vol. 31, p. 2014, 2011.

[193] N. Dimiduk, A. Khurana, M. H. Ryan, and M. Stack, *HBase in action*. Manning Shelter Island, 2013.

[194] K. Chodorow, *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*. " O'Reilly Media, Inc.", 2013.

[195] K. Ghosh, G. Haggerty, and P. Agarwal, "Alzheimer/s disease - not an exaggeration of healthy aging," *Indian Journal of Psychological Medicine*, vol. 33, no. 2, p. 106, 2011. [Online]. Available: https://doi.org/10.4103/0253-7176.92047

[196] D. T. Jones, M. M. Machulda, P. Vemuri, E. M. McDade, G. Zeng, M. L. Senjem, J. L. Gunter, S. A. Przybelski, R. T. Avula, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Age-related changes in the default mode network are more advanced in alzheimer disease," *Neurology*, vol. 77, no. 16, pp. 1524–1531, oct 2011. [Online]. Available: https://doi.org/10.1212/wnl.0b013e318233b33d

[197] J. Barnes, B. C. Dickerson, C. Frost, L. C. Jiskoot, D. Wolk, and W. M. van der Flier, "Alzheimer's disease first symptoms are age dependent: Evidence from the NACC dataset," *Alzheimer's & Dementia*, vol. 11, no. 11, pp. 1349–1357, nov 2015. [Online]. Available: https://doi.org/10.1016/j.jalz.2014.12.007

[198] J. Dukart, M. L. Schroeter, and K. M. and, "Age correction in dementia – matching to a healthy brain," *PLoS ONE*, vol. 6, no. 7, p. e22193, jul 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0022193

[199] F. Falahati, , D. Ferreira, H. Soininen, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, S. Lovestone, M. Eriksdotter, L.-O. Wahlund, A. Simmons, and E. Westman, "The effect of age correction on multivariate classification in alzheimer's disease, with a focus on the characteristics of incorrectly and correctly classified subjects," *Brain Topography*, vol. 29, no. 2, pp. 296–307, oct 2015. [Online]. Available: https://doi.org/10.1007/s10548-015-0455-1

[200] B. P. Printy, N. Verma, M. C. Cowperthwaite, and M. K. Markey, "Effects of genetic variation on the dynamics of neurodegeneration in alzheimer's disease," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, aug 2014. [Online]. Available: https://doi.org/10.1109/embc.2014.6944121

[201] M. A. Williams, G. J. McKay, R. Carson, D. Craig, G. Silvestri, and P. Passmore, "Age-related macular degeneration–associated genes in alzheimer disease," *The American Journal of Geriatric Psychiatry*, vol. 23, no. 12, pp. 1290–1296, dec 2015. [Online]. Available: https://doi.org/10.1016/j.jagp.2015.06.005

[202] C. A. Hostage, K. R. Choudhury, P. M. Doraiswamy, and J. R. P. and, "Mapping the effect of the apolipoprotein e genotype on 4-year atrophy rates in an alzheimer disease–related brain network," *Radiology*, vol. 271, no. 1, pp. 211–219, apr 2014. [Online]. Available: https://doi.org/10.1148/radiol.13131041

[203] M. M. Mielke, P. Vemuri, and W. A. Rocca, "Clinical epidemiology of alzheimer's disease: assessing sex and gender differences," *Clinical epidemiology*, vol. 6, p. 37, 2014.

[204] L. A. Farrer, "Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease," *JAMA*, vol. 278, no. 16,

p. 1349, oct 1997. [Online]. Available: https://doi.org/10.1001/jama.1997.03550160069041

[205] A. F. Jorm, K. A. Mather, P. Butterworth, K. J. Anstey, H. Christensen, and S. Easteal, "APOE genotype and cognitive functioning in a large age-stratified population sample." *Neuropsychology*, vol. 21, no. 1, pp. 1–8, 2007. [Online]. Available: https://doi.org/10.1037/0894-4105.21.1.1

[206] J.-C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, M. Hiltunen, O. Combarros, D. Zelenika, M. J. Bullido, B. Tavernier, L. Letenneur, K. Bettens, C. Berr, F. Pasquier, N. Fiévet, P. Barberger-Gateau, S. Engelborghs, P. D. Deyn, I. Mateo, A. Franck, S. Helisalmi, E. Porcellini, O. Hanon, M. M. de Pancorbo, C. Lendon, C. Dufouil, C. Jaillard, T. Leveillard, V. Alvarez, P. Bosco, M. Mancuso, F. Panza, B. Nacmias, P. Bossù, P. Piccardi, G. Annoni, D. Seripa, D. Galimberti, D. Hannequin, F. Licastro, H. Soininen, K. Ritchie, H. Blanché, J.-F. Dartigues, C. Tzourio, I. Gut, C. V. Broeckhoven, A. Alpérovitch, M. Lathrop, and P. Amouyel, "Genome-wide association study identifies variants at CLU and CR1 associated with alzheimer's disease," *Nature Genetics*, vol. 41, no. 10, pp. 1094–1099, sep 2009. [Online]. Available: https://doi.org/10.1038/ng.439

[207] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. S. Pahwa, V. Moskvina, K. Dowzell, A. Williams *et al.*, "Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease," *Nature genetics*, vol. 41, no. 10, pp. 1088–1093, 2009.

[208] NaN, *Women's Health Research*. National Academies Press, oct 2010. [Online]. Available: https://doi.org/10.17226/12908

[209] Y. Liu, T. Paajanen, E. Westman, L.-O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Proitsi, J. Powell, P. Mecocci, and et al., "Effect of apoe ϵ4 allele on cortical thicknesses and volumes: The addneuromed study," *Journal of*

Alzheimer's Disease, vol. 21, no. 3, p. 947–966, Sep 2010. [Online]. Available: http://doi.org/10.3233/JAD-2010-100201

[210] J. S. Damoiseaux, W. W. Seeley, J. Zhou, W. R. Shirer, G. Coppola, A. Karydas, H. J. Rosen, B. L. Miller, J. H. Kramer, and M. D. G. and, "Gender modulates the APOE 4 effect in healthy older adults: Convergent evidence from functional brain connectivity and spinal fluid tau levels," *Journal of Neuroscience*, vol. 32, no. 24, pp. 8254–8262, jun 2012. [Online]. Available: https://doi.org/10.1523/jneurosci.0305-12.2012

[211] A. Fleisher, "Sex, apolipoprotein e ϵ4 status, and hippocampal volume in mild cognitive impairment," *Archives of Neurology*, vol. 62, no. 6, p. 953, jun 2005. [Online]. Available: https://doi.org/10.1001/archneur.62.6.953

[212] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "Misc functions of the department of statistics (e1071), tu wien," *R package*, vol. 1, pp. 5–24, 2008.

[213] J. A. Webster, J. R. Gibbs, J. Clarke, M. Ray, W. Zhang, P. Holmans, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem *et al.*, "Genetic control of human brain transcript expression in alzheimer disease," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 445–458, 2009.

[214] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[215] J. M. Castellano, J. Kim, F. R. Stewart, H. Jiang, R. B. DeMattos, B. W. Patterson, A. M. Fagan, J. C. Morris, K. G. Mawuenyega, C. Cruchaga *et al.*, "Human apoe isoforms differentially regulate brain amyloid-β peptide clearance," *Science translational medicine*, vol. 3, no. 89, pp. 89ra57–89ra57, 2011.

[216] M. Koistinaho, S. Lin, X. Wu, M. Esterman, D. Koger, J. Hanson, R. Higgs, F. Liu, S. Malkani, K. R. Bales *et al.*, "Apolipoprotein e promotes astrocyte colocalization and degradation of deposited amyloid-[beta] peptides," *Nature medicine*, vol. 10, no. 7, p. 719, 2004.

[217] R. W. Mahley and S. C. Rall Jr, "Apolipoprotein e: far more than a lipid transport protein," *Annual review of genomics and human genetics*, vol. 1, no. 1, pp. 507–537, 2000.

[218] X. He, K. Cooley, C. H. Chung, N. Dashti, and J. Tang, "Apolipoprotein receptor 2 and x11α/β mediate apolipoprotein e-induced endocytosis of amyloid-β precursor protein and β-secretase, leading to amyloid-β production," *Journal of Neuroscience*, vol. 27, no. 15, pp. 4052–4060, 2007.

[219] S. Ye, Y. Huang, K. Müllendorff, L. Dong, G. Giedt, E. C. Meng, F. E. Cohen, I. D. Kuntz, K. H. Weisgraber, and R. W. Mahley, "Apolipoprotein (apo) e4 enhances amyloid β peptide production in cultured neuronal cells: Apoe structure as a potential therapeutic target," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 700–18 705, 2005.

[220] Y. Huang, "Apolipoprotein e and alzheimer disease," *Neurology*, vol. 66, no. 1 suppl 1, pp. S79–S85, 2006.

[221] Q. Tan, M. Thomassen, K. M. Jochumsen, O. Mogensen, K. Christensen, and T. A. Kruse, "A combinatory approach for selecting prognostic genes in microarray studies of tumour survivals," *Advances in Bioinformatics*, vol. 2009, pp. 1–7, 2009. [Online]. Available: https://doi.org/10.1155/2009/480486

[222] S. Sun, Q. Peng, and A. Shakoor, "A kernel-based multivariate feature selection method for microarray data classification," *PLoS ONE*, vol. 9, no. 7, p. e102541, jul 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0102541

[223] R. Aziz, C. Verma, and N. Srivastava, "A novel approach for dimension reduction of microarray," *Computational Biology and Chemistry*, vol. 71, pp. 161–169, dec 2017. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2017.10.009

[224] C. Golzio, J. Willer, M. E. Talkowski, E. C. Oh, Y. Taniguchi, S. Jacquemont, A. Reymond, M. Sun, A. Sawa, J. F. Gusella *et al.*, "Kctd13 is a major driver of mirrored neuroanatomical phenotypes associated with the 16p11. 2 cnv," *Nature*, vol. 485, no. 7398, p. 363, 2012.

[225] X. Zheng, F. Y. Demirci, M. M. Barmada, G. A. Richardson, O. L. Lopez, R. A. Sweet, M. I. Kamboh, and E. Feingold, "A rare duplication on chromosome 16p11. 2 is identified in patients with psychosis in alzheimer's disease," *PloS one*, vol. 9, no. 11, p. e111462, 2014.

[226] E. H. Sharman, K. G. Sharman, and S. C. Bondy, "Parallel changes in gene expression in aged human and mouse cortex," *Neuroscience letters*, vol. 390, no. 1, pp. 4–8, 2005.

[227] J.-H. Chen, Y.-L. Huang, C.-J. Hsieh, T.-F. Chen, Y. Sun, L.-L. Wen, P.-K. Yip, Y.-M. Chu, and Y.-C. Chen, "Genetic polymorphisms of lipid metabolism gene sar1 homolog b and the risk of alzheimer's disease and vascular dementia," *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 10, no. 4, p. P500, 2014.

[228] E. Pavlopoulos, S. Jones, S. Kosmidis, M. Close, C. Kim, O. Kovalerchik, S. A. Small, and E. R. Kandel, "Molecular mechanism for age-related memory loss: the histone-binding protein rbap48," *Science translational medicine*, vol. 5, no. 200, pp. 200ra115–200ra115, 2013.

[229] J. Chen, Y. Zhou, S. Mueller-Steiner, L.-F. Chen, H. Kwon, S. Yi, L. Mucke, and L. Gan, "Sirt1 protects against microglia-dependent amyloid-β toxicity through inhibiting nf-κb signaling," *Journal of Biological Chemistry*, vol. 280, no. 48, pp. 40 364–40 374, 2005.

[230] M. Perez, I. Santa-Maria, D. Barreda, E. Gomez, X. Zhu, R. Cuadros, J. R. Cabrero, F. Sanchez-Madrid, H. N. Dawson, M. P. Vitek *et al.*, "Tau–an inhibitor of deacetylase hdac6 function," *Journal of neurochemistry*, vol. 109, no. 6, pp. 1756–1766, 2009.

[231] D. Kim, M. D. Nguyen, M. M. Dobbin, A. Fischer, F. Sananbenesi, J. T. Rodgers, I. Delalle, J. A. Baur, G. Sui, S. M. Armour *et al.*, "Sirt1 deacetylase protects against neurodegeneration in models for alzheimer9s disease and amyotrophic lateral sclerosis," *The EMBO journal*, vol. 26, no. 13, pp. 3169–3179, 2007.

[232] K. N. Green, J. S. Steffan, H. Martinez-Coria, X. Sun, S. S. Schreiber, L. M. Thompson, and F. M. LaFerla, "Nicotinamide restores cognition in alzheimer's disease transgenic mice via a mechanism involving sirtuin inhibition and selective reduction of thr231-phosphotau," *Journal of Neuroscience*, vol. 28, no. 45, pp. 11 500–11 510, 2008.

[233] B. Hempen and J.-P. Brion, "Reduction of acetylated α-tubulin immunoreactivity in neurofibrillary tangle-bearing neurons in alzheimer's disease," *Journal of Neuropathology & Experimental Neurology*, vol. 55, no. 9, pp. 964–972, 1996.

[234] E. Dirkx, M. M. Gladka, L. E. Philippen, A.-S. Armand, V. Kinet, S. Leptidis, H. El Azzouzi, K. Salic, M. Bourajjaj, G. J. Da Silva *et al.*, "Nfat and mir-25 cooperate to reactivate the transcription factor hand2 in heart failure," *Nature cell biology*, vol. 15, no. 11, p. 1282, 2013.

[235] J.-W. Kim, Y.-H. You, S. Jung, H. Suh-Kim, I.-K. Lee, J.-H. Cho, and K.-H. Yoon, "mirna-30a-5p-mediated silencing of beta2/neurod expression is an important initial event of glucotoxicity-induced beta cell dysfunction in rodent models," *Diabetologia*, vol. 56, no. 4, pp. 847–855, 2013.

[236] H. Xu, K. S. Tsang, J. C. Chan, P. Yuan, R. Fan, H. Kaneto, and G. Xu, "The combined expression of pdx1 and mafa with either ngn3 or neurod improves the dif-

ferentiation efficiency of mouse embryonic stem cells into insulin-producing cells," *Cell transplantation*, vol. 22, no. 1, pp. 147–158, 2013.

[237] T. J. Cherry, S. Wang, I. Bormuth, M. Schwab, J. Olson, and C. L. Cepko, "Neurod factors regulate cell fate and neurite stratification in the developing retina," *Journal of Neuroscience*, vol. 31, no. 20, pp. 7365–7379, 2011.

[238] T. Söllner, M. K. Bennett, S. W. Whiteheart, R. H. Scheller, and J. E. Rothman, "A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion," *Cell*, vol. 75, no. 3, pp. 409–418, 1993.

[239] E. B. Mukaetova-Ladinska, J. H. Xuereb, F. Garcia-Sierra, J. Hurt, H.-J. Gertz, R. Hills, C. Brayne, F. A. Huppert, E. S. Paykel, M. A. McGee *et al.*, "Lewy body variant of alzheimer's disease: selective neocortical loss of t-snare proteins and loss of map2 and α-synuclein in medial temporal lobe," *The Scientific World Journal*, vol. 9, pp. 1463–1475, 2009.

[240] J. Long, G. Pan, E. Ifeachor, R. Belshaw, and X. Li, "Discovery of novel biomarkers for alzheimer's disease from blood," *Disease markers*, vol. 2016, 2016.

[241] B. Schreitmüller, T. Leyhe, E. Stransky, N. Köhler, and C. Laske, "Elevated angiopoietin-1 serum levels in patients with alzheimer's disease," *International journal of Alzheimer's disease*, vol. 2012, 2012.

[242] S. Patel, R. J. Shah, P. Coleman, and M. Sabbagh, "Potential peripheral biomarkers for the diagnosis of alzheimer's disease," *International Journal of Alzheimer's Disease*, vol. 2011, 2011.

[243] M. Britschgi, K. Rufibach, S. L. B. Huang, C. M. Clark, J. A. Kaye, G. Li, E. R. Peskind, J. F. Quinn, D. R. Galasko, and T. Wyss-Coray, "Modeling of pathological traits in alzheimer's disease based on systemic extracellular signaling proteome," *Molecular & Cellular Proteomics*, vol. 10, no. 10, pp. M111–008 862, 2011.

[244] P. I. Moreira, C. Carvalho, X. Zhu, M. A. Smith, and G. Perry, "Mitochondrial dysfunction is a trigger of alzheimer's disease pathophysiology," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1802, no. 1, pp. 2–10, 2010.

[245] J.-C. Lambert, B. Grenier-Boley, V. Chouraki, S. Heath, D. Zelenika, N. Fievet, D. Hannequin, F. Pasquier, O. Hanon, A. Brice *et al.*, "Implication of the immune system in alzheimer's disease: evidence from genome-wide pathway analysis," *Journal of Alzheimer's disease*, vol. 20, no. 4, pp. 1107–1118, 2010.

[246] M. P. Mattson, "Pathways towards and away from alzheimer's disease," *Nature*, vol. 430, no. 7000, pp. 631–639, 2004.

[247] H. Ruffner, A. Bauer, and T. Bouwmeester, "Human protein–protein interaction networks and the value for drug discovery," *Drug discovery today*, vol. 12, no. 17, pp. 709–716, 2007.

[248] B. Olsson, R. Lautner, U. Andreasson, A. Öhrfelt, E. Portelius, M. Bjerke, M. Hölttä, C. Rosén, C. Olsson, G. Strobel *et al.*, "Csf and blood biomarkers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis," *The Lancet Neurology*, vol. 15, no. 7, pp. 673–684, 2016.

[249] H. Zetterberg, "Plasma amyloid β—quo vadis?" *Neurobiology of aging*, vol. 36, no. 10, pp. 2671–2673, 2015.

[250] V. Ovod, K. N. Ramsey, K. G. Mawuenyega, J. G. Bollinger, T. Hicks, T. Schneider, M. Sullivan, K. Paumier, D. M. Holtzman, J. C. Morris *et al.*, "Amyloid β concentrations and stable isotope labeling kinetics of human plasma specific to central nervous system amyloidosis," *Alzheimer's & Dementia*, vol. 13, no. 8, pp. 841–849, 2017.

[251] N. Kaneko, A. Nakamura, Y. Washimi, T. Kato, T. Sakurai, Y. Arahata, M. Bundo, A. Takeda, S. Niida, K. Ito *et al.*, "Novel plasma biomarker surro-

gating cerebral amyloid deposition," *Proceedings of the Japan Academy, Series B*, vol. 90, no. 9, pp. 353–364, 2014.

[252] S. Westwood, E. Leoni, A. Hye, S. Lynham, M. R. Khondoker, N. J. Ashton, S. J. Kiddle, A. L. Baird, R. Sainz-Fuertes, R. Leung *et al.*, "Blood-based biomarker candidates of cerebral amyloid using pib pet in non-demented elderly," *Journal of Alzheimer's Disease*, vol. 52, no. 2, pp. 561–572, 2016.

[253] S. C. Burnham, C. C. Rowe, D. Baker, A. I. Bush, J. D. Doecke, N. G. Faux, S. M. Laws, R. N. Martins, P. Maruff, S. L. Macaulay *et al.*, "Predicting alzheimer disease from a blood-based biomarker profile a 54-month follow-up," *Neurology*, vol. 87, no. 11, pp. 1093–1101, 2016.

[254] N. Voyle, D. Baker, S. C. Burnham, A. Covin, Z. Zhang, D. P. Sangurdekar, C. A. Tan Hehir, C. Bazenet, S. Lovestone, S. Kiddle *et al.*, "Blood protein markers of neocortical amyloid-β burden: a candidate study using somascan technology," *Journal of Alzheimer's Disease*, vol. 46, no. 4, pp. 947–961, 2015.

[255] U. Andreasson, K. Blennow, and H. Zetterberg, "Update on ultrasensitive technologies to facilitate research on blood biomarkers for central nervous system disorders," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 3, pp. 98–102, 2016.

[256] H. Zetterberg, "Neurofilament light: a dynamic cross-disease fluid biomarker for neurodegeneration," *Neuron*, vol. 91, no. 1, pp. 1–3, 2016.

[257] P. S. Weston, T. Poole, N. S. Ryan, A. Nair, Y. Liang, K. Macpherson, R. Druyeh, I. B. Malone, R. L. Ahsan, H. Pemberton *et al.*, "Serum neurofilament light in familial alzheimer disease: A marker of early neurodegeneration," *Neurology*, vol. 89, no. 21, pp. 2167–2175, 2017.

[258] L. M. Byrne, F. B. Rodrigues, K. Blennow, A. Durr, B. R. Leavitt, R. A. Roos, R. I. Scahill, S. J. Tabrizi, H. Zetterberg, D. Langbehn *et al.*, "Neurofilament light

protein in blood as a potential biomarker of neurodegeneration in huntington's disease: a retrospective cohort analysis," *The Lancet Neurology*, vol. 16, no. 8, pp. 601–609, 2017.

[259] Å. Sandelius, H. Zetterberg, K. Blennow, R. Adiutori, A. Malaspina, M. Laura, M. M. Reilly, and A. M. Rossor, "Plasma neurofilament light chain concentration in the inherited peripheral neuropathies," *Neurology*, vol. 90, no. 6, pp. e518–e524, 2018.

[260] N. Mattsson, H. Zetterberg, S. Janelidze, P. S. Insel, U. Andreasson, E. Stomrud, S. Palmqvist, D. Baker, C. A. T. Hehir, A. Jeromin *et al.*, "Plasma tau in alzheimer disease," *Neurology*, vol. 87, no. 17, pp. 1827–1835, 2016.

[261] E. Nagele, M. Han, C. DeMarshall, B. Belinka, and R. Nagele, "Diagnosis of alzheimer's disease based on disease-specific autoantibody profiles in human sera," *PloS one*, vol. 6, no. 8, p. e23112, 2011.

[262] E. P. Nagele, M. Han, N. K. Acharya, C. DeMarshall, M. C. Kosciuk, and R. G. Nagele, "Natural igg autoantibodies are abundant and ubiquitous in human sera, and their number is influenced by age, gender, and disease," *PLoS One*, vol. 8, no. 4, p. e60726, 2013.

[263] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, T. G. Beach, A. Grover, T. L. Niedzielko, L. E. Schneider, D. Mastroeni, R. Caselli *et al.*, "Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons," *Proceedings of the National Academy of Sciences*, vol. 105, no. 11, pp. 4441–4446, 2008.

[264] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[265] 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*, vol. 5, 2015.

[266] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017, r package version 1.6-8. [Online]. Available: https://CRAN.R-project.org/package=e1071

[267] R. H. Fletcher, S. W. Fletcher, and G. S. Fletcher, *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2012.

[268] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[269] T. F. S. Inc, *ProtoArray Prospector v5.2.3*, 2015.

[270] U. A. Khan, L. Liu, F. A. Provenzano, D. E. Berman, C. P. Profaci, R. Sloan, R. Mayeux, K. E. Duff, and S. A. Small, "Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease," *Nature neuroscience*, vol. 17, no. 2, p. 304, 2014.

[271] M. Gleichmann, Y. Zhang, W. H. Wood, K. G. Becker, M. R. Mughal, M. J. Pazin, H. van Praag, T. Kobilo, A. B. Zonderman, J. C. Troncoso *et al.*, "Molecular changes in brain aging and alzheimer's disease are mirrored in experimentally silenced cortical neuron networks," *Neurobiology of aging*, vol. 33, no. 1, pp. 205–e1, 2012.

[272] W. Wagner, S. Bork, P. Horn, D. Krunic, T. Walenda, A. Diehlmann, V. Benes, J. Blake, F. Huber, V. Eckstein, P. Boukamp, and D. Ho, "Aging and replicative senescence have related effects on human stem and progenitor cells," *PLoS ONE*, vol. 4, no. 6, 2009.

[273] Y. Yarden and M. Sliwkowski, "Untangling the erbb signalling network," *Nat Rev Mol Cell Biol*, vol. 2, pp. 127–137, 2001.

[274] D. Falls, "Neuregulins: functions, forms, and signaling strategies," *Exp Cell Res*, vol. 284, pp. 14–30, 2003.

[275] A. Buonanno and G. Fischbach, "Neuregulin and erbb receptor signaling pathways in the nervous system," *Curr Opin Neurobiol*, vol. 11, pp. 287–296, 2001.

[276] G. Corfas, K. Roy, and J. Buxbaum, "Neuregulin 1-erbb signaling and the molecular/cellular basis of schizophrenia," *Nat Neurosci*, vol. 7, pp. 575–580, 2004.

[277] E. Anton, M. Marchionni, K. Lee, and P. Rakic, "Role of ggf/neuregulin signaling in interactions between migrating neurons and radial glia in the developing cerebral cortex," *Development*, vol. 124, pp. 3501–3510, 1997.

[278] C. Rio, H. Rieff, P. Qi, T. Khurana, and G. Corfas, "Neuregulin and erbb receptors play a critical role in neuronal migration," *Neuron*, vol. 19, pp. 39–50, 1997.

[279] P. Fernandez, D. Tang, L. Cheng, A. Prochiantz, A. Mudge, and M. Raff, "Evidence that axon-derived neuregulin promotes oligodendrocyte survival in the developing rat optic nerve," *Neuron*, vol. 28, pp. 81–90, 2000.

[280] V. Calaora, B. Rogister, K. Bismuth, K. Murray, H. Brandt, P. Leprince, M. Marchionni, and M. Dubois-Dalcq, "Neuregulin signaling regulates neural precursor growth and the generation of oligodendrocytes in vitro." *J Neurosci*, vol. 21, pp. 4740–4751, 2001.

[281] G. López-Bendito, A. Cautinat, J. Sánchez, F. Bielle, N. Flames, A. Garratt, D. Talmage, L. Role, P. Charnay, O. Marín, and S. Garel, "Tangential neuronal migration controls axon guidance: a role for neuregulin-1 in thalamocortical axon navigation." *Cell*, vol. 125, pp. 127–142, 2006.

[282] O. Bermingham-McDonogh, K. McCabe, and T. Reh, "Effects of ggf/neuregulins on neuronal survival and neurite outgrowth correlate with erbb2/neu expression in developing rat retina." *Development*, vol. 122, pp. 1427–1438, 1996.

[283] K. Gerecke, J. Wyss, and S. Carroll, "Neuregulin-1beta induces neurite extension and arborization in cultured hippocampal neurons." *Mol Cell Neurosci*, vol. 27, pp. 379–393, 2004.

[284] M. Borenstein, L. V. Hedges, J. Higgins, and H. R. Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," *Research synthesis methods*, vol. 1, no. 2, pp. 97–111, 2010.

[285] T. Pigott, *Advances in meta-analysis*. Springer Science & Business Media, 2012.

[286] J. R. Polanin, E. A. Hennessy, and E. E. Tanner-Smith, "A review of meta-analysis packages in r," *Journal of Educational and Behavioral Statistics*, vol. 42, no. 2, pp. 206–242, 2017.

[287] L. V. Hedges, E. Tipton, and M. C. Johnson, "Robust variance estimation in meta-regression with dependent effect size estimates," *Research synthesis methods*, vol. 1, no. 1, pp. 39–65, 2010.

[288] A. Gasparrini, "Mvmeta: multivariate and univariate meta-analysis and meta-regression," 2015.

[289] M. W.-L. Cheung, "metasem: An r package for meta-analysis using structural equation modeling," *Frontiers in Psychology*, vol. 5, p. 1521, 2015.

[290] J. R. Polanin and S. J. Wilson, "Meta-analyzing a complex correlational dataset: A case study using correlations that measure the relationship between parental involvement and academic achievement." *Society for Research on Educational Effectiveness*, 2014.

[291] B. Neupane, D. Richer, A. J. Bonner, T. Kibret, and J. Beyene, "Network meta-analysis using r: a review of currently available automated packages," *PLoS One*, vol. 9, no. 12, p. e115065, 2014.

[292] K. G. Kugler, L. A. Mueller, and A. Graber, "Madam-an open source meta-analysis toolbox for r and bioconductor," *Source code for biology and medicine*, vol. 5, no. 1, p. 3, 2010.

[293] D. Ghosh and H. Choi, "metaarray package for meta-analysis of microarray data," 2010.

[294] X. Wang, D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, S. G. Liao, Z. Huo, S. Tang, Y. Ding *et al.*, "An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection," *Bioinformatics*, vol. 28, no. 19, pp. 2534–2536, 2012.

[295] G. Marot, F. Jaffrézic, and A. Rau, "metarnaseq: Differential meta-analysis of rna-seq data," *dim (param)*, vol. 1, no. 26408, p. 3.

[296] I. Ihnatova, "Mama: an r package for meta-analysis of microarray," *R package version*, vol. 2, no. 1, 2013.

[297] L. Lusa, R. Gentleman, and M. Ruschhaupt, "Genemeta: metaanalysis for high throughput experiments," *R package version*, vol. 1, no. 1, 2013.

[298] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics*, vol. 19, no. suppl_1, pp. i84–i90, 2003.

[299] E. M. Conlon, J. J. Song, and A. Liu, "Bayesian meta-analysis models for microarray data: a comparative study," *BMC bioinformatics*, vol. 8, no. 1, p. 80, 2007.

[300] R. B. Scharpf, H. Tjelmeland, G. Parmigiani, and A. B. Nobel, "A bayesian model for cross-study differential gene expression," *Journal of the American*

Statistical Association, vol. 104, no. 488, pp. 1295–1310, dec 2009. [Online]. Available: https://doi.org/10.1198/jasa.2009.ap07611

[301] P. Baldi and A. D. Long, "A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[302] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–25, 2004.

[303] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica sinica*, pp. 31–46, 2002.

[304] L. V. Hedges, "Distribution theory for glass's estimator of effect size and related estimators," *Journal of educational statistics*, vol. 6, no. 2, pp. 107–128, 1981.

[305] P. Hu, C. M. Greenwood, and J. Beyene, "Statistical methods for meta-analysis of microarray data: a comparative study," *Information Systems Frontiers*, vol. 8, no. 1, pp. 9–20, 2006.

[306] L. Hedges and I. Olkin, "Statistical methods for meta-analysis. academic press, orlando (flor.)," 1985.

[307] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr, "The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1," 1949.

[308] T. Liptak, "On the combination of independent tests," *Magyar Tud Akad Mat Kutato Int Kozl*, vol. 3, pp. 171–197, 1958.

[309] J. MacDonald, "hugene10sttranscriptcluster. db: Affymetrix hugene10 annotation data (chip hugene10sttranscriptcluster)," *R package version*, vol. 8, no. 1.

[310] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database." *Nature genetics*, vol. 39, no. 1, 2007.

[311] Z. Bai, G. Han, B. Xie, J. Wang, F. Song, X. Peng, and H. Lei, "Alzbase: an integrative database for gene dysregulation in alzheimer's disease," *Molecular neurobiology*, vol. 53, no. 1, pp. 310–319, 2016.

[312] S. Davis and P. Meltzer, "Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 14, pp. 1846–1847, 2007.

[313] K. M. Bakulski, D. C. Dolinoy, M. A. Sartor, H. L. Paulson, J. R. Konen, A. P. Lieberman, R. L. Albin, H. Hu, and L. S. Rozek, "Genome-wide dna methylation differences between late-onset alzheimer's disease and cognitively normal controls in human frontal cortex," *Journal of Alzheimer's Disease*, vol. 29, no. 3, pp. 571–588, 2012.

[314] K. Lunnon, R. Smith, E. Hannon, P. L. De Jager, G. Srivastava, M. Volta, C. Troakes, S. Al-Sarraj, J. Burrage, R. Macdonald *et al.*, "Methylomic profiling implicates cortical deregulation of ank1 in alzheimer's disease," *Nature neuroscience*, vol. 17, no. 9, pp. 1164–1170, 2014.

[315] N. Mohanty and V. Berry, "Graphene-based single-bacterium resolution biodevice and dna transistor: interfacing graphene derivatives with nanoscale and microscale biocomponents," *Nano letters*, vol. 8, no. 12, pp. 4469–4476, 2008.

[316] B. Li, G. Pan, N. D. Avent, R. B. Lowry, T. E. Madgett, and P. L. Waines, "Graphene electrode modified with electrochemically reduced graphene oxide for label-free dna detection," *Biosensors and Bioelectronics*, vol. 72, pp. 313–319, 2015.

[317] K. Islam, A. Suhail, and G. Pan, "A label-free and ultrasensitive immunosensor for detection of human chorionic gonadotrophin based on graphene fets," *Biosensors*, vol. 7, no. 3, p. 27, 2017.

[318] B. Li, G. Pan, A. Suhail, K. Islam, N. Avent, and P. Davey, "Deep uv hardening of photoresist for shaping of graphene and lift-off fabrication of back-gated field effect biosensors by ion-milling and sputter deposition," *Carbon*, vol. 118, pp. 43–49, 2017.