IDENTIFICATION OF ALZHEIMER DISEASE-ASSOCIATED PATHWAYS AND NETWORK USING TRANSCRIPTOME ANALYSIS

LAU CHING YEE

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

University of Malaya

IDENTIFICATION OF ALZHEIMER DISEASE-ASSOCIATED PATHWAYS AND NETWORK USING TRANSCRIPTOME ANALYSIS

LAU CHING YEE

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

University of Malaya

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Lau Ching Yee

Matric No: SGR130019

Name of Degree: Master of Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Identification of Alzheimer Disease-Associated Pathways and Network Using Transcriptome Analysis

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

IDENTIFICATION OF ALZHEIMER DISEASE-ASSOCIATED PATHWAYS AND NETWORK USING TRANSCRIPTOME ANALYSIS

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disease and the most common form of dementia. The disease mainly affects people aged 65 and older. The mechanisms underlying AD aetiology is still not clearly understood due to its complex nature. In this study, we integratively reanalyzed the publicly available transcriptome data sets of AD studies using human post-mortem brains. By using this method, the capability of detecting weak signals could be improved and novel biological insights which could not be obtained from the individual studies could be gained. In order to get reliable biological inference from the data, we compared and evaluated existing bioinformatic methods in transcriptomic analysis and selected the superior ones to be included in our data analysis pipeline. Since complex diseases like AD can be better understood from the perspective of network biology than at the individual gene level, we used NetDecoder, a state-of-the-art network-based transcriptomic analysis algorithm to capture genes that are associated with the differentially expressed genes in a network context. The networks established based on protein-protein interactions included key genes such as UBC, ABL1, YWHAZ, APP, TP53 and CTNNB1, which have also been reported by other AD studies. The networks potentially provide mechanistic insights to better understand how these genes interact and drive AD pathogenesis. Thus, the present study provides a workflow for mining promising target genes for further confirmatory experiments which can lead to more effective treatment of AD or better diagnostics for AD.

Keywords: Alzheimer's disease, integrative analysis, microarray, RNA-seq

University of Malaya

PENGENALPASTIAN LALUAN DAN RANGKAIAN BERKAITAN DENGAN PENYAKIT ALZHEIMER MENGGUNAKAN ANALISIS TRANSKRIPTOM

ABSTRAK

Penyakit Alzheimer (AD) suatu penyakit kemerosotan saraf yang progresif dan sejenis demensia yang biasa. Penyakit ini khasnya memberi kesan kepada orang yang berumur 65 ke atas. Mekanisme asas etiologi AD tidak difahami jelas kerana AD sejenis penyakit kompleks. Dalam kajian ini, kami menganalisa data transkriptom awam daripada kajian-kajian AD yang menggunakan tisu bedah siasat otak manusia. Dengan ini, keupayaan untuk mengesan isyarat yang lemah ditambahbaik dan pandangan biologi baru yang tidak dapat diperolehi daripada kajian individu terhasil. Bagi mendapatkan inferens biologi yang dipercayai daripada data, kita membanding dan menilai kaedah-kaedah bioinformatik sedia ada dalam analisis transkriptom dan memilih kaedah yang unggul untuk dimasukkan dalam perancangan analisis data. Oleh sebab penyakit kompleks seperti AD lebih senang difahami dari perspektif biologi rangkaian berbanding pada peringkat gen individu, kami menggunakan NetDecoder, suatu algoritma rangkaian baru untuk mengesan gen yang berkait dengan gen terekspres terbeza dalam konteks rangkaian. Rangkaian yang dikenalpasti berdasarkan interaksi protein termasuk gen-gen utama seperti UBC, ABL1, YWHAZ, APP, TP53 dan CTNNB1, yang telah dilaporkan kajian-kajian AD terdahulu. Rangkaian-rangkaian ini memberi pandangan mekanistik terhadap cara gen-gen berinteraksi dan menyumbang kepada patogenesis AD. Dengan demikian, kajian ini menyediakan satu aliran kerja perlombongan gen-gen sasaran untuk eksperimen pengesahan selanjutnya yang mampu menjana rawatan AD yang lebih berkesan mahupun diagnosis AD yang lebih baik.

Kata kunci: penyakit Alzheimer, analisis integratif, mikroarray, RNA-seq

University of Malaya

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Khang Tsung Fei, my research supervisor, for his valuable suggestions and guidance during the planning and development of this research study. I would also like to thank Dr. Lim Yat Yuen, my second research supervisor, for his constructive advices on thesis writing.

Finally, I wish to thank my family and friends for their support and encouragement throughout my research study.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

LIST OF TABLES

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

Aβ	:	Beta-amyloid
AD	:	Alzheimer's disease
APP	:	Beta-amyloid precursor protein
bp	:	Base pair
CASP1	:	Caspase-1
ComBat	:	Empirical Bayes method
CRYAB	:	αB-crystallin
DAVID	:	The Database for Annotation, Visualization and Integration Discovery
DBI	:	Diazepam binding inhibitor
DEG	:	Differentially expressed gene
EBV	:	Epstein-Barr virus
ECM	:	Extra-cellular matrix
edgeR	:	Empirical analysis of digital gene expression in R
FC	:	Fold change
fRMA	:	Frozen robust multiarray analysis
$GABA_A$	:	γ-aminobutyric acid type A
GC	:	Guanine and cytosine
GEO	:	Gene Expression Omnibus
GO	:	Gene Ontology
GSEA	:	Gene Set Enrichment Analysis
GSK-3	:	Glycogen synthase kinase-3
IL-1β	:	Interleukin 1 beta
IP	:	Impact score
iRefIndex	:	Interaction reference index

University of Malaya

kDa	:	Kilo Dalton
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
lincRNA		Long intergenic non-coding ribonucleic acid
log	:	Logarithm
MAP	:	Maximum *a posteriori*
MDS	:	Multidimensional scaling
mRNA	:	Messenger ribonucleic acid
n	:	Sample size
NAD+	:	Nicotinamide adenine dinucleotide
NGS	:	Next-generation sequencing
NPM1	:	Nucleophosmin
NSF	:	N-ethylmaleimide-sensitive factor
OSA	:	Omicsoft sequence aligner
PAK1	:	p21-activated kinase
PCC	:	Pearson correlation coefficient
PHB	:	Prohibitin
PPI	:	Protein-protein interaction
PPV	:	Positive predictive value
PPV^*	:	Expected positive predictive value
qPCR	:	Quantitative polymerase chain reaction
RLE	:	Relative log expression
RMA	:	Robust multiarray analysis
RNA	:	Ribonucleic acid
RNA-seq	:	RNA-sequencing
SAM	:	Significance analysis of microarrays
SNARE	:	Soluble N-ethylmaleimide-sensitive factor attachment protein receptor

University of Malaya

SNCA	:	Alpha-synuclein
SRA	:	Sequence Read Archive
TMM	:	Trimmed mean of M values
UBC	:	Polyubiquitin-C
WGCNA	:	Weighted correlation network analysis

University of Malaya

LIST OF APPENDICES

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Transcriptome

The transcriptome is the full range of transcripts (RNAs expressed from the genome) and their quantity in a particular cell or population of cells at a particular time. The percentage of the human genome that is transcribed into RNAs was estimated to be less than 5% (Frith et al., 2005). According to Vogel and Marcotte (2012),

> In general, in both bacteria and eukaryotes, the cellular concentrations of proteins correlate with the abundances of their corresponding mRNAs, but not strongly. They often show a squared Pearson correlation coefficient of ~0.40, which implies that ~40% of the variation in protein concentration can be explained by knowing mRNA abundances. Higher correlations have also been observed. To explain the remaining ~60% of the variation, some combination of post-transcriptional regulation and measurement noise needs to be invoked. (p. 228)

Transcriptomics is the study of the transcriptome. Studying the transcriptome is an important part of understanding a cell's entire story or perturbed systems — connecting the gap between the genetic code and the functional proteins. According to Wang et al. (2009),

> The key aims of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions. (p. 57)

University of Malaya

1.2 Microarray and RNA-sequencing for transcriptomics

During the past decades the development and widespread use of microarray technology have encouraged the study of the transcriptome. Recently, the application of next-generation sequencing (NGS) technology to sequence steady-state RNA in a sample, termed RNA-sequencing (RNA-seq), has emerged as a powerful alternative approach for transcriptomics (Oshlack et al., 2010; Wang et al., 2009). In contrast to microarray, expression levels of genes are based on direct counts of transcripts rather than probe intensities.

Microarray has several limitations, such as high background levels caused by cross-hybridization, bias introduced by the variation of probe binding efficiency and dependence on existing genome sequence for probe design (Wang et al., 2009). Moreover, microarray lacks sensitivity for transcripts with low and very high expression levels owing to background and saturation of signals respectively. Consequently, it has a limited dynamic range of expression levels.

Unlike microarray technology, RNA-seq is sequencing-based approach that has very low background levels and large dynamic range of detection with no upper limit, thus it is more predictive of true expression levels (Wang et al., 2009). RNA-seq allows the entire transcriptome to be surveyed and therefore it is capable of discovering novel transcripts and isoforms, capturing alternative splicing comprehensively, distinguishing allelic expression and RNA editing (Oshlack et al., 2010; Wang et al., 2009). Besides that, RNA-seq also has lower technical variation and higher resolution on expression levels compared to microarray.

Nevertheless, RNA-seq technology faces several challenges. For instance, specific sequencing protocols create biases in the outcome (Oshlack et al., 2010). Furthermore, nucleotide sequence bias, transcript length bias and GC content bias have been observed

University of Malaya

in RNA-seq data (Rung & Brazma, 2013; Zheng et al., 2011). Besides that, the presence of splicing events, paralogous sequences or repetitive sequences make mapping reads or alignment more difficult (Oshlack et al., 2010; Wang et al., 2009).

1.3 Preprocessing of data

Before apparent patterns of variation in microarray or RNA-seq data can be attributed to biological variation, it is crucial to first eliminate unwanted non-biological variations that are present in raw data. Failure to do so would introduce errors and these errors are retained throughout the ensuing analyses, which can consequently affect the results and conclusions of a study. Thus, preprocessing of raw data is pivotal step in the analysis of microarray or RNA-seq data.

1.3.1 Preprocessing of microarray data

Typically, preprocessing of microarray data aims to handle background noise, processing effects, between array variation and summarization of probes (McCall & Almudevar, 2012). The preprocessing steps: image-processing and data normalization are used to remove systematic variation (Allison et al., 2006). Other potential preprocessing steps consist of data transformation, data filtering and, for two-colour arrays, background subtraction. A lot of methods have been developed for preprocessing of microarray data. For example, RMA (Irizarry et al., 2003) and fRMA (McCall et al., 2010) are preprocessing algorithms for high density oligonucleotide microarrays which involve 3 steps: background correction, normalization and summarization.

University of Malaya

1.3.2 Preprocessing of RNA-seq data

Generally, the major steps in preprocessing of RNA-seq data include removal of technical sequences, quality analysis, *de novo* assembly of short reads (without reference) or mapping short reads to a reference genome and/or a transcriptome, feature counting and normalization of raw counts. Many methods or tools have been developed for preprocessing of RNA-seq data. For instances, Trimmamatic (Bolger et al., 2014) is a tool for removal of technical sequences and quality filtering; Trinity (Haas et al., 2013) is a platform for *de novo* transcript sequence reconstruction; TopHat (Trapnell et al., 2009) and OSA (Hu et al., 2012) are read mapping algorithms, htseq-count (Anders et al., 2015) is a feature counting tool; TMM (Robinson & Oshlack, 2010) is a normalization method.

1.4 Differential expression analysis

A differentially expressed gene (DEG) is a gene that shows statistically significant difference in mean expression level between two phenotype classes. This involves the application of statistical methodology such as test for equality of two means for each gene. There are two possible approaches to generate a list of candidate DEG. The conceptually and practically simplest one is the single-gene approach, where we treat each gene as being independent of one another, apply a statistical test on the equality of mean between the two phenotype classes, and then control for false discovery rates (Tusher et al., 2001). An estimate of the biological significance in the form of fold-change, and also the *p*-value (for measuring statistical significance), is returned. Using information from *p*-value (sometimes, jointly with fold-change), a final list of selected gene candidates is produced. Alternatively, a method that takes advantage of prior knowledge about the associations of genes (gene sets) available in curated databases potentially returns candidates that are more biologically relevant. The most popular

University of Malaya

implementation is the Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005) for microarray data. The method is based on score that is a modification of the Kolmogorov-Smirnov statistic. Computationally it is more expensive because it needs to simulate the null distribution. Furthermore, GSEA has a strong focus on cancer-related gene sets, and therefore may not be suitable as a general-purpose method for differential expression analysis. For these reasons, we will focus on the single-gene approach for differential expression analysis.

In general, although naïve methods such as the t-test can be applied for microarray and RNA-seq data, superior tests based on specific models of gene expression (e.g. linear model for microarray data; negative binomial distribution of gene counts of RNA-seq data) can provide additional statistical power for detecting DEG, if the model approximates the underlying data distribution well, and the statistic is robust to deviations from model assumptions.

Many methods have been proposed for differential expression analysis of transcriptome data. The examples of these methods for microarray data are linear models and empirical Bayes Methods (Smyth, 2004) implemented in the limma software package (Ritchie et al., 2015), Significance Analysis of Microarrays (SAM; Tusher et al., 2001), VarMixt (Delmar et al., 2005), whereas the examples for RNA-seq data are DESeq (Anders & Huber, 2010), DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010).

While the *p*-value can be used to determine which genes pass the selection criteria as DEGs, a more balanced approach is to consider statistical significance jointly with biological significance, as estimated using the fold change. Such a joint filtering criteria was proposed by Li (2012). Specifically, let p denote the *p*-value of a statistical test, and let FC denote the fold change. According to the method of Xiao et al. (2014), a reciprocal function decision boundary for filtering DEGs using statistical and biological significance

University of Malaya

can be obtained in the following way. Suppose we require $p < 0.01$ and FC ≥ 2 to call for up-regulated genes, and $p < 0.01$ and FC $\leq 1/2$ to call for down-regulated genes. The product of $-\log_{10}p > 2$ and $|\log_2 FC| \geq 1$ yields the inequality $-\log_{10}p > 2/|\log_2 FC|$. Thus, genes that fall in the region defined by $-\log_{10}p > 2/\log_2 FC$ are differentially up-regulated; those in the region of $-\log_{10}p > -2/\log_2 FC$ are differentially down-regulated. The union of the sets of differentially up and down-regulated genes constituted the set of DEGs. The *p* and FC cut offs for selection of DEGs can be different with $p < 0.01$ and FC ≥ 2. Thus, the final form of the decision boundary would be more general in the form of $-\log_{10}p > c/|\log_2 FC| + k$, where c is a constant for controlling the stretching ($c > 1$) or shrinking ($0 < c < 1$) of the reciprocal function, and k is a vertical translation.

1.5 Integrative analysis

With the exceptional growth of transcriptome data sets in public repositories, new possibilities lie in integrative analysis of multiple data sets to increase the statistical power of transcriptome analysis (Taminau et al., 2012). By using this method, the capability of detecting weak signals could be improved and novel biological insights which could not be obtained from the individual studies could be captured (Rung & Brazma, 2013). According to Jiang and Liu (2015),

> There are also studies integrating expression data sets from GEO to make new discoveries. For example, expression compendia integration identified the conditional activity of expression modules in cancer, expression outlier analysis predicted the frequent fusion of the *TMPRSS2* and *ETS* transcription factor genes in prostate cancer and mutual information has been used to infer post-translational modulators of transcription factor activity. The current study by Fehrmann *et al*. represents a fresh angle for big data integration and novel discovery. (p. 103)

University of Malaya

Meta-analysis and data merging are the two different strategies of integrative analysis (Lazar et al., 2013). Meta-analysis consists in analyzing each data set independently and the results or summary-level data such as *p*-values are then combined (Lazar et al., 2013; Rung & Brazma, 2013). In contrast, data merging involves merging different data sets into a bigger data set and the subsequent analysis being performed using the new integrated data set (Lazar et al., 2013).

Meta-analysis is prone to high false-negative rates upon the statistical hypothesis test when data sets contain only a few samples but it is often a better option for better control of between-laboratory heterogeneity (Lazar et al., 2013; Rung & Brazma, 2013). The main advantage of data merging over meta-analysis is more robust inference can be made owing to the higher statistical relevance of the results. However, removing generic sources of unwanted variation is the main challenge for data merging. Batch effect is the main source that obscures meaningful biological information with non-biological perturbations. Methods that remove or adjust batch variation, which enable the merging of multiple microarray data sets into a bigger data set, have been developed.

1.6 Network analysis

Network biology refers to the network that characterizes biological system. Here we define the most basic network nomenclature with Figure 1.6.1. Undirected network (Figure 1.6.1 (a)) has edges that connect nodes without direction whereas directed network (Figure 1.6.1 (b)) has edges that connect nodes with direction. Optionally, nodes can be in different colour, shape or size to add more information with appropriate annotation. Similarly, edges can be in different colour or width to add more information with appropriate annotation.

University of Malaya

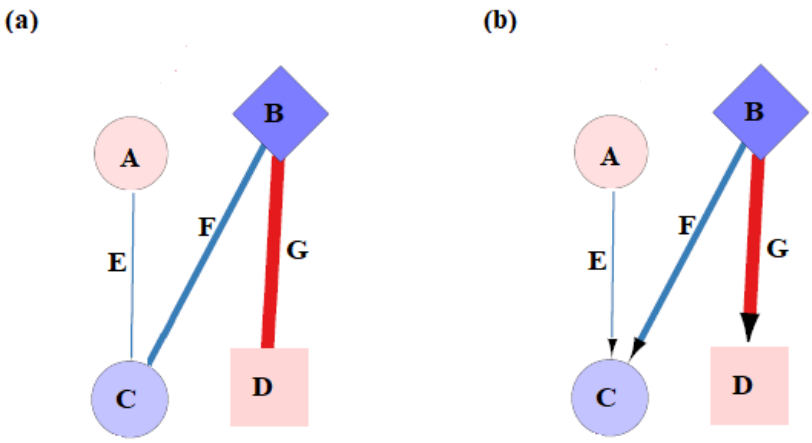


Figure 1.6.1: Network nomenclature. (a) Undirected network and (b) directed network with nodes A, B, C, D and edges E, F, G.

According to Barabási et al. (2011),

> Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances in this direction are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases. (p. 56)

These advances thus may lead to potential therapeutic targets or better biomarkers with predictive and diagnostic values.

University of Malaya

1.6.1 NetDecoder

Recently, a network biology platform called NetDecoder (da Rocha et al., 2016) for finding a context-specific biological network from a protein-protein interaction (PPI) network, given the gene expression profiles and associated genes (such as DEGs), has been developed. Proof-of-concept of the utility of NetDecoder as a network analysis tool was shown in the form of its success in deriving meaningful biological results from three case studies involving microarray data from breast cancer, dyslipidemia, and Alzheimer's disease (AD). There, NetDecoder successfully recovered subnetworks whose gene members had functions well-known to be associated with the disease phenotype.

Here, we briefly look at core computational biology ideas that underly NetDecoder. NetDecoder uses the iRefIndex version 14.0 to construct a PPI network containing 15 608 proteins and 180 044 interactions. Starting from a set of source genes (usually DEG) and transcriptome data (microarray or RNA-seq) defined by the user, NetDecoder runs a process-guided flow algorithm to identify interaction paths that connect the source genes to the target genes (transcriptional regulators as the default). In doing so, it finds a subnetwork consisting of the source genes, target genes, and importantly, intermediary genes that modulate context-specific information flows between source and target genes. These intermediary genes are usually not differentially expressed, but function as important determinants of information flow paths that result in particular biological phenotypes (Jacunski & Tatonetti, 2013). The contextual nature of the subnetwork thus discovered implies that genes in disease-specific subnetwork would be enriched in disease-related signaling pathways, thus highly interpretable from a biological perspective.

A schematic overview of NetDecoder workflow is given in Figure 1.6.1.1. Given the gene expression profile consisting of two phenotypes (such as control and disease)

University of Malaya

(Figure 1.6.1.1 a), gene-wise Pearson correlation coefficients (PCCs) are calculated across samples of the same phenotype for all possible pairs of genes (Figure 1.6.1.1 b and c). In the PPI network, the absolute value of the PCC is used to set the edge weight of each interaction and the negative of its logarithm (-logPCC) is used to set the cost of the edge. Red colour of edges represents positive PCC and blue colour of edges represents negative PCC. Thus, an edge-weighted PPI network is obtained for each of the two classes (Figure 1.6.1.1 d). Then, NetDecoder uses the process-guided flow algorithm to find paths through the PPI network to obtain a sparse subnetwork beginning from source genes to target genes with minimum cost (minimum-cost flow optimization) (Figure 1.6.1.1 e). The two subnetworks specific to the control and the disease phenotype classes are then compared. NetDecoder uses a novel scoring scheme to identify key players (Figure 1.6.1.1 f) that contribute to a disease phenotype: (i) Network routers are "intermediary proteins that have influence over many genes and show high flow differences between two phenotypes" (da Rocha et al., 2016, p. 4). (ii) Key targets are "the target/sink nodes (genes related to transcriptional regulation) with high flow differences" (da Rocha et al., 2016, p. 4). (iii) High impact genes are "genes that experience a significant change in regulation between control and disease conditions including flow differences, establishment of new inflows and change of directionality of gene expression correlations (i.e. from positive correlation to negative correlation or vice versa) between two phenotypes" (da Rocha et al., 2016, p. 4). NetDecoder "developed a novel scoring scheme, termed impact score (IP), to rank and assess genes based on their importance in mediating differences in information flow profiles between two given phenotypes" and "define genes with high IP scores as high impact genes" (da Rocha et al., 2016, p. 4). In the disease-specific subnetwork, NetDecoder further selects "paths enriched with at least two types of key genes and termed these paths as prioritized subnetworks" (da Rocha et al., 2016, p. 9) (Figure 1.6.1.1 g).

University of Malaya

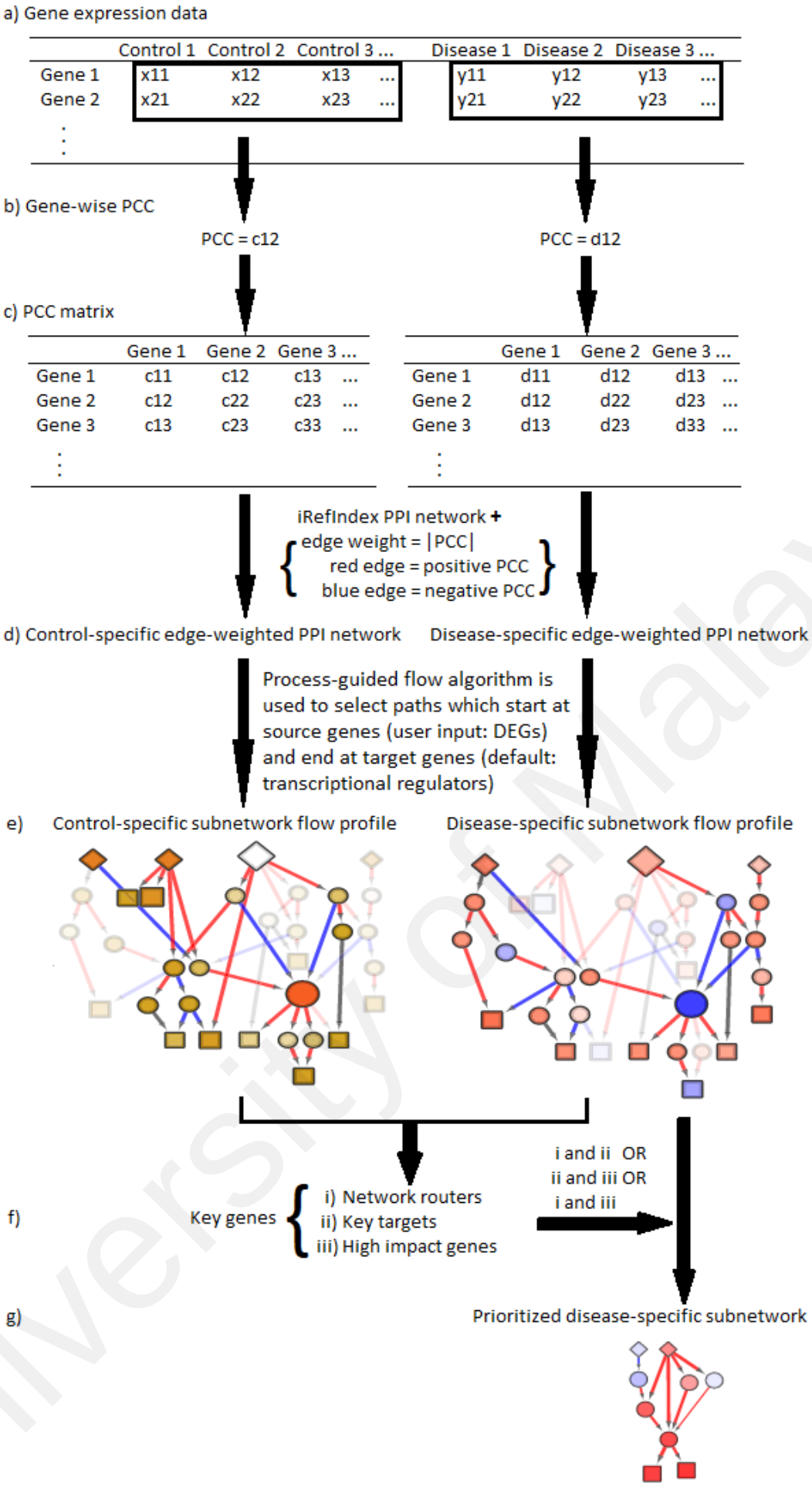


Figure 1.6.1.1: Schematic overview of NetDecoder workflow. (a) Given a gene expression data consisting of control and disease biological replicates, (b) PCCs are calculated for all possible pairs of genes across samples of the same phenotype (control and disease) and (c) thus a pairwise PCC matrix is obtained for each of the two phenotypes (control and disease). (d) Integrating the PPI network with the absolute values of the PCC (edge weight) and the PCC directionality (red edge for positive PCC and blue edge for negative PCC) produces edge-weighted PPI network for each of the two phenotypes (control and disease). (e) Source genes such as DEGs or any other gene list of interest is required as input to NetDecoder. By default, target genes (sinks) are genes involved in transcriptional regulation. Target genes can also be defined by the user based on study goal. The process-guided flow algorithm is used to select paths, which start at source genes (diamond shape of nodes), passing through intermediary genes (circle shape of nodes) and end at target genes (square shape of nodes), along the edge-weighted PPI networks. In the resulting subnetworks, edge width represents the amount of flow through an edge. Red edge represents positive PCC and blue edge represents negative PCC. Node size represents the total flow (the in and out flows) at a node. Nodes are coloured according to the node flow difference – red represents high flow in disease but low flow in control; conversely, blue represents low flow in disease but high flow in control. (f) The two subnetworks specific to the control and the disease phenotypes are then compared. The following key genes are identified: (i) network routers (ii) key targets (iii) high impact genes. (g) In the disease-specific subnetwork, NetDecoder further selects paths that include at least two types of key genes and the resulting subnetwork is termed as prioritized disease-specific subnetwork. Network images in (e) and (g) are from da Rocha et al. (2016).

1.7 Enrichment analysis

In general, a gene seldom affects the phenotype of interest on its own, but through interaction of its product with the products of other genes. Thus, for a phenotype of interest, we can imagine that it is associated with a network of genes whose products interact with each other to bring about the manifestation of the phenotype. The inverse problem is that, given a gene list (such as DEG) obtained from microarray or RNA-seq experiments, how can one know what biological process is most likely associated with members of the gene list?

To do so, we first note that a gene can be associated with various biological terms such as Gene Ontology (GO), KEGG pathways. The task of a functional enrichment algorithm is to determine which process is over-represented or under-represented in the gene list, thus providing the critical biological context for gene lists discovered from analyses of microarray or RNA-seq data. By far, the DAVID (Huang et al., 2009) Gene Functional Classification Tool remains the method of choice in the bioinformatic community for performing functional enrichment analysis (~15,000 citations as of January 2018). DAVID is based on an agglomerative algorithm that clusters members in a gene list into classes of related biology (biological modules). By doing so, it groups functionally-related genes and terms into manageable number of biological modules, so that the network and biological context of a gene list can be inferred.

1.8 Alzheimer's disease

Alzheimer's disease (AD) is a progressive neurodegenerative disease and the most common form of dementia. The disease mainly affects people aged 65 and older (Evans et al., 1989). Globally, an estimated 35.6 million people are affected by dementia (of which AD forms a significant subset) in 2010. This figure is expected to double to 65.7

University of Malaya

million in 2030, and 115.4 million in 2050 (Prince et al., 2013). From a public health perspective, AD presents a challenging problem to manage by all countries in the future as the world population ages.

AD is characterized clinically by cognitive impairment and often accompanied by non-cognitive symptoms (Twine et al., 2011). Physically, the presence of neuropathologies in the form of neurofibrillary tangles, senile plaques, neuropil threads, specific neuron loss, and synapse loss (Terry, 1994) is found in the brain of AD patients (Murphy & LeVine, 2010). The mechanisms underlying AD aetiology is still not clearly understood due to its complex nature (Kavanagh et al., 2013).

The pathogenesis of AD involves the dysregulated production and deposition of the β-amyloid peptide (Aβ). Indeed, diagnosis of AD requires two hallmark pathologies: presence of extracellular plaque deposits of the Aβ in the brain, and the flame-shaped neurofibrillary tangles of the microtubule binding protein tau (Murphy & LeVine, 2010). In heritable early onset AD cases, mutations are observed in either the β-amyloid precursor protein (APP) for Aβ, or in presenilin-1 (PS1) or presenilin-2 (PS2), which function as the catalytic subunit of γ-secretase. The latter enzyme is the final endoprotease in the biochemical pathway that produces the Aβ peptide. The Aβ peptide is a 4 kDa molecule which is derived from the larger APP molecule. It was first discovered as the main component of amyloid deposits in the brain and cerebrovasculature of patients with AD and Down's Syndrome (Glenner & Wong, 1984a; Masters, Multhaup et al., 1985; Masters, Simms et al., 1985).

The pathogenesis, diagnosis and therapy of AD remain challenging to research efforts (Twine et al., 2011). Over the years, our understanding of the key players involved in AD pathogenesis has been gradually improving. For example, Miller et al. (2008) applied the weighted gene coexpression network analysis method (WGCNA) (Horvath et al., 2006;

University of Malaya

Oldham et al., 2006; Zhang & Horvath, 2005) on microarray data, and successfully identified the signaling molecule YWHAZ for the first time as hub in both aging and AD patients, in addition to recovering presenilin 1 (PSEN1) which is known to be a catalytic subunit of γ-secretase. Blair et al. (2013) found age-associated increases in FKBP51 which interacts with Hsp90 to promote neurotoxic tau accumulation in AD brains. Using RNA-seq data, Twine et al. (2011) identified APOE splice variants in AD brains which are postulated to be associated with neurodegeneration progression. More recently, Mills et al. (2013) reported upregulation of the diazepam-binding inhibitor (DBI) in the parietal cortex of AD brains, and Satoh et al. (2014) reported downregulation of NeuroD6 as a biomarker in AD brains.

1.9 Transcriptome data mining for Alzheimer's disease

Global research efforts have been dedicated to understand biology of AD by comparing the gene expression profiles between healthy individuals and individuals diagnosed with AD using microarray and RNA-seq technology, contributing to an ever-growing of disparate transcriptome data sets that are deposited in public archives. In this issue, the publicly available transcriptome data sets from AD studies using human post-mortem brains can be collected and reanalyzed. Microarray data merging can be done to discover new insights on AD. This big data mining can also be used to draw conclusions about the general properties of gene expression in large sample groups, such as broad transcriptional patterns of AD regardless of the brain region studied and the stage of disease. Differential expression analyses of the merged microarray data set and RNA-seq data sets can be performed to identify differentially expressed genes. Complex diseases like AD can be better understood from the perspective of network biology than at the individual gene level. Network analyses can be carried out to decode AD-specific networks. Finally, enrichment analyses can be done to identify enriched biological

University of Malaya

pathways. A lot of computational methods or bioinformatic tools are available for transcriptome analysis. In order to extract the maximum biological information and get reliable biological inference from the data, the existing methods or tools can be studied, examined, compared or evaluated and finally the superior one be included in the data analysis pipeline.

1.10 Research aims and objectives

The aim of this thesis is to develop a coherent bioinformatic workflow for converting integrative analyzed public transcriptome AD data from microarray and RNA-seq platforms into new insights of biological processes involved in AD pathogenesis. Towards this aim, several objectives are set. Firstly, the relevant AD microarray and RNA-seq data sets are identified from the literature. Secondly, comparing of the relevant existing bioinformatic methods or tools is done and the optimal one is selected to build a robust data analysis pipeline. Thirdly, data merging of microarray data sets is done with batch effects removal and the suitability of the resulting merged data set for downstream analysis is determined. Fourthly, differential expression analyses between normal and AD conditions are carried out to identify differentially expressed genes (DEGs). Fifthly, network analyses are done to identify AD-associated networks. Finally, enrichment analyses are done to extract the biological context of the DEG lists and the AD-associated networks.

University of Malaya

CHAPTER 2: METHODOLOGY

2.1 Data collection

Twelve raw Affymetrix microarray data sets from AD studies using human post-mortem brains were obtained from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo). Table 2.1.1 shows the metadata of these data sets. In total, the 12 microarray data sets contain 670 samples (326 AD samples and 344 control samples).

Two raw Illumina RNA-seq data sets from AD studies using human post-mortem brains were obtained from the Sequence Read Archive (SRA; www.ncbi.nlm.nih.gov/sra; www.ebi.ac.uk/ena; trace.ddbj.nig.ac.jp). Table 2.1.2 shows the metadata of these data sets. In total, the 2 RNA-seq data sets contain 21 samples (11 AD samples and 10 control samples).

University of Malaya

Table 2.1.1: Metadata of the 12 microarray data sets used in the current study.

GEO accession	Source	Sample size		Platform	Reference
		AD	Control		
GSE1297	Hippocampal CA1	22	9	Affymetrix Human Genome U133A Array	(Blalock et al., 2004)
	Total	22	9		
GSE4757	Entorhinal cortex	10	10	Affymetrix Human Genome U133 Plus 2.0 Array	(Dunckley et al., 2006)
	Total	10	10		
GSE5281	Entorhinal cortex	10	13	Affymetrix Human Genome U133 Plus 2.0 Array	(Liang, Dunckley et al., 2008; Liang et al., 2007; Liang, Reiman et al., 2008)
	Hippocampus	10	13		
	Medial temporal gyrus	16	12		
	Posterior cingulate	9	13		
	Superior frontal gyrus	23	11		
	Primary visual cortex	19	12		
	Total	87	74		

University of Malaya

Table 2.1.1, continued.

GEO accession	Source	Sample size		Platform	Reference
		AD	Control		
GSE9770	Entorhinal cortex	6	0	Affymetrix Human Genome U133 Plus 2.0 Array	(Liang et al., 2010)
	Hippocampus	6	0		
	Middle temporal gyrus	6	0		
	Posterior cingulate cortex	5	0		
	Superior frontal gyrus	6	0		
	Primary visual cortex	5	0		
	Total	34	0		
GSE12685	Frontal cortex synaptoneurosome	6	8	Affymetrix Human Genome U133A Array	(Williams et al., 2009)
	Total	6	8		
GSE16759	Parietal lobe cortex	4	4	Affymetrix Human Genome U133 Plus 2.0 Array	(Nunez-Iglesias et al., 2010)
	Total	4	4		

University of Malaya

Table 2.1.1, continued.

GEO accession	Source	Sample size		Platform	Reference
		AD	Control		
GSE26972	Entorhinal cortex	3	3	Affymetrix Human Exon 1.0 ST Array	(Berson et al., 2012)
	Total	3	3		
GSE28146	Hippocampus	22	8	Affymetrix Human Genome U133 Plus 2.0 Array	(Blalock et al., 2011)
	Total	22	8		
GSE29652	Temporal cortex astrocytes	18	0	Affymetrix Human Genome U133 Plus 2.0 Array	(Simpson et al., 2011)
	Total	18	0		
GSE36980	Frontal cortex	15	18	Affymetrix Human Gene 1.0 ST Array	(Hokama et al., 2014)
	Temporal cortex	10	19		
	Hippocampus	7	10		
	Total	32	47		
GSE37263	Temporal cortex	8	8	Affymetrix Human Exon 1.0 ST Array	(Tan et al., 2010)
	Total	8	8		

University of Malaya

Table 2.1.1, continued.

GEO accession	Source	Sample size		Platform	Reference
		AD	Control		
GSE48350	Entorhinal cortex	15	39	Affymetrix Human Genome U133 Plus 2.0 Array	(Blair et al., 2013)
	Hippocampus	19	43		
	Post-central gyrus	25	43		
	Superior frontal gyrus	21	48		
	Total	80	173		

University of Malaya

Table 2.1.2: Metadata of the 2 RNA-seq data sets used in the current study.

SRA accession	Source	Sample size		Platform	Library layout	Read length	Strand-specific assay	Reference
		AD	Control					
SRP004879	Total brain	1	1	Illumina Genome Analyzer (GAII)	Single end	35 bp / 36 bp	No	(Twine et al., 2011)
	Frontal lobe	1	1					
	Temporal lobe	1	1					
	Total	3	3					
SRP056863	Frontal cortex	8	7	Illumina HiSeq 2000	Paired end	90 bp / 101 bp	No	(Bai et al., 2013)
	Total	8	7					

University of Malaya

2.2 Determination of the methods/tools to be used in the current study

A lot of computational methods or bioinformatic tools are available for transcriptome analysis. In order to extract the maximum biological information and get reliable biological inference from the data, existing methods or tools were studied, examined, compared or evaluated and finally the superior one was included in the data analysis pipeline of this study.

2.2.1 Data analysis pipeline for microarray

2.2.1.1 Preprocessing of data

The effect of microarray preprocessing methods on certain multivariate analyses was examined by a study (McCall & Almudevar, 2012). According to the study, typical trade-off between bias and precision was observed in nine preprocessing methods with an exception, where frozen Robust Multiarray Analysis (fRMA) was found to have better accuracy, given its precision. For this reason, fRMA was selected to participate in the present study.

2.2.1.2 Integrative analysis

Taminau et al. (2014), in their validation analysis using microarray data from six lung cancer studies, showed that data merging identified all DEGs that were also identified through meta-analysis, in addition to many more genes missed by the latter approach. Literature review further showed that the DEGs identified using data merging and also meta-analysis were corroborated by other studies with respect to their involvement in lung cancer development.

Certain microarray data sets used in the current study have low sample size, which puts meta-analysis at a disadvantage since meta-analysis is prone to high false-negative

University of Malaya

rates in this situation. Furthermore, certain microarray data sets have no control samples and thus these data sets are unable to be included in meta-analysis. For all the reasons above, meta-analysis was not considered and data merging was performed in the present study.

Removing batch effects is essential in data merging. A comparative study of six batch effect removal methods using multiple measures of precision, accuracy and overall performance reported that ComBat (empirical Bayes method) outperformed the other five candidates (Chen et al., 2011). ComBat had satisfactorily performance on all measures whereas each of the others had at least one major drawback. Besides that, only ComBat was found to be robust when adjusting small batches. For these reasons, ComBat was chosen to be the batch effect removal method in the present study.

2.2.1.3 Differential expression analysis

A comparative study of differential expression analysis methods for microarray data reported that the empirical Bayes statistic implemented in limma was the most robust method across all sample sizes (Jeffery et al., 2006). Another comparative study of eight statistical tests with variance modeling strategies reported that limma and VarMixt offered significant improvement when compared to the t-test (Jeanmougin et al., 2010). Additionally, limma shows several practical advantages. For these reasons, limma was selected to participate in differential expression analysis of microarray data in the current study.

University of Malaya

2.2.2 Data analysis pipeline for RNA-seq

2.2.2.1 Preprocessing of data

A comparative analysis of several adapter and quality filtering tools showed that Trimmomatic had the best performance, particularly in Maximum Information mode (Bolger et al., 2014). Thus, Trimmomatic with Maximum Information mode was selected to participate in adapter and quality filtering process.

A comparative study of fifty gene profiling pipelines (the combinations of alignment and quantification tools) showed that the pipeline with OSA and htseq-ine (htseq-count with intersection-nonempty mode) is in the top of the overall rankings based on two metrics (relative error and Spearman correlation) (Fonseca et al., 2014). Therefore, OSA and htseq-ine were selected as the alignment and quantification tools respectively in the current study.

2.2.2.2 Differential expression analysis

The two RNA-seq data sets used in the current study have small sample size. Since there is no comparative study of differential expression analysis methods using small data sets in the literature, such a comparative study was done to find out which method should be used in the current study.

As of 22 September 2015, a survey of the methods for performing differential expression analysis using RNA-seq data showed that there were 22 methods available (Khang & Lau, 2015). These methods vary in their effectiveness for calling DEG when sample size is small, which is the situation for the two RNA-seq data sets in the current study. Because of this, it is unclear as to which was the best one to choose. Besides sample size, an additional consideration is the ability of the differential expression analysis to detect DEG is also a function of the biological effect size between the phenotype classes

University of Malaya

under consideration. In the context of AD versus control phenotype classes, the biological effect can be assumed to be strong.

In order to select the most appropriate method, an empirical assessment of method that received the most attention from the scientific community (i.e. high citations per year) such as edgeR, DESeq and its new version, DESeq2, was carried out. These methods are parametric, since they explicitly model the distribution of gene counts using the negative binomial distribution. To be balanced, we included the well-known non-parametric NOISeq for additional contrast. Furthermore, Zhang et al. (2014) showed that edgeR had slightly superior performance in the receiver operating characteristic curve compared to DESeq and Cuffdiff2. Thus, edgeR was included in this comparison. Two methods with high citations per year: Cuffdiff2 and DEGSeq were not included, based on conclusions from recent method comparative analyses. For example, Cuffdiff2 was found to have very low precision when replicate size increased in the analysis of two large RNA-seq data sets from mouse and human (Seyednasrollah et al., 2015). Another comparative study involving DESeq, DEGseq, edgeR, NBPSeq, TSPM and baySeq showed that DEGseq had the largest false positive rate among them (Guo et al., 2013).

(a) ***Benchmarking***

The Recount database (Frazee et al., 2011) contains raw RNA-seq count data sets from 18 major studies which have been assembled from raw reads using the Myrna (Langmead et al., 2010) pipeline. A search through Recount database identified the Bottomly data set (Bottomly et al., 2011) as a suitable benchmarking data set. This data set contains gene expression data (22 million Illumina reads per sample, read length of ~30 bases) obtained from the brain striatum tissues of two mice strains: C57BL/6J (n = 10) and DBA/2J (n = 11). These two strains of mice are known to show large, strain-specific variation in

University of Malaya

neurological response (Grice et al., 2007; Korostynski et al., 2006; Korostynski et al., 2007), and thus mimic the strong biological effect between control and AD phenotypes in the present study.

(b) ***Constructing a reference DEG set***

To construct a reference DEG set for which the result of differential expression analysis from edgeR, DESeq, DESeq2, and NOISeq could be compared, voom (Law et al., 2014; Ritchie et al., 2015) was used. Whereas the differential expression analysis methods considered either model the mean–variance relationships in the count data nonparametrically, or parametrically using the Poisson/negative binomial distributions, voom log-transforms count data into a microarray-like data type. This transformed data can then be handled using the robust limma algorithm (Ritchie et al., 2015; Smyth, 2004) developed for microarray analysis. Since voom is based on a different algorithmic architecture, using it to set the reference DEG set can avoid the issue of calling similar DEG due to algorithmic similarities. Nonetheless, the validity of using voom to set the reference DEG sets requires empirical justification. One way to do this is to compare its performance with other DEG call methods on some RNA-seq data set where qPCR validation results are available for sufficiently large numbers of genes. Such type of data set is scarce in the literature, and only one data set was found to be suitable. The Rajkumar data set (Rajkumar et al., 2015) consists of gene expression count data (26,119 genes; minimum of 10 million Illumina reads per sample, read length of ~50 bases) from the amygdala tissues of C57BL/6NTac strain mice. There are two phenotype classes: wild type (n = 8), and heterozygotes for the Brd1 gene deletion (n = 8). A total of 115 genes were selected for qPCR validation (additional Table 5 in (Rajkumar et al., 2015)), and 60 of them were found to have differential expression. The differential expression analysis methods considered were voom, edgeR, DESeq, DESeq2, and NOISeq.

University of Malaya

A simple requirement for a differential expression method that is reasonable for setting the reference DEG set is that it should not return extreme results (too few (tens) or too many (thousands)). When this requirement is satisfied, the differential expression method that shows relatively higher positive predictive value (PPV; the complement of the false discovery rate) is preferable. Let NTP be the number of observed true positives, and NFP the number of observed false positives, and Ng the number of DEG called by a differential expression method. Since only 115 out of 26,119 genes had qPCR validation, it is not possible to estimate PPV, but only its expectation.

Technically, the actual number of true positives consists of an observed part (N_{TP}), and an unobserved part (N^*_{TP}). The number of DEG (i.e. total predicted positives) that lacks validation is $U = N_g - N_{FP} - N_{TP}$. The expected number of unobserved true positives can be computed as $N^*_{TP} = [N_{TP} / (N_{TP} + N_{FP})] \times U$, and the expected PPV is then given by $PPV^* = (N_{TP} + N^*_{TP}) / N_g$.

(c) ***Simulation and performance evaluation***

To simulate low sample size scenarios (100 instances), 3 and 6 individuals within each phenotype class were randomly sampled (without replacement) respectively. The performance of the DEG call methods: edgeR, DESeq, DESeq2, NOISeq was assessed using sensitivity and positive predictive value (PPV). For each DEG call method, sensitivity was computed as the proportion of reference DEG called. PPV was computed as the proportion of DEG called belonging the reference DEG set. The mean and the standard deviation of these metrics were given. For ranking the methods, the F-score was used. This score combines information from both PPV and sensitivity, in the form of by F-score=2(PPV x sensitivity) / (PPV + sensitivity). The F-score lies between 0 and 1. A high F-score is desirable because it indicates good balance of PPV and sensitivity.

University of Malaya

2.3 Microarray data analysis

Each raw data set was preprocessed using frozen robust multiarray analysis (fRMA). During this process, background correction, normalization and summarization of the probes in each probe set were done. A single gene may be detected by multiple probe sets referred to as set of probe sets. For genes that are detected by more than one probe set, the probe set with the highest mean expression in each set of probe sets was selected to represent its corresponding gene.

The twelve data sets were merged into a single dataset and only genes that are present in all the data sets were retained. The newly merged data set, which consists of expression levels of 8673 genes in 326 AD samples and 344 control samples, was adjusted for batch effects using empirical Bayes method (ComBat). In order to assess the effectiveness of ComBat, multidimensional scaling (MDS) plots and relative log expression (RLE) plots were constructed from the merged data set before and after applying ComBat.

Differential expression analysis of the merged data set was performed with linear models and empirical Bayes methods. Genes with $-\log_{10}p \geq 16/\left|\log_2 FC\right| - 17$ were identified as differentially expressed genes, where p is the p-value from linear models and empirical Bayes methods, $\log_2 FC$ is the estimate of $\log_2$ fold change from linear models and empirical Bayes methods.

Table 2.3.1 shows the implementation of the above methods or algorithms.

University of Malaya

Table 2.3.1: Implementation of the methods/algorithms used in the microarray data analysis.

Analysis step	Method/Algorithm	R implementation		Reference
		R function	R package	
Preprocessing	fRMA	frma	frma	(McCall et al., 2010)
Collapsing multiple probe sets for a single gene	MaxMean	collapseRows	WGCNA	(Miller et al., 2011)
Adjusting batch effects	ComBat	merge	inSilicoMerging	(Johnson et al., 2007; Taminau et al., 2012)
Differential expression analysis	Linear models and empirical Bayes	lmFit, contrasts.fit, eBayes, topTable	limma	(Ritchie et al., 2015; Smyth, 2004)

University of Malaya

2.4 RNA-seq data analysis

For each raw data set, removal of technical sequences and quality filtering were applied to each read/read pair using Trimmomatic (http://www.usadellab.org/cms/index.php?page=trimmomatic). Subsequently, read mapping was performed using Omicsoft sequence aligner (OSA; http://omicsoft.com/osa) with Omicsoft-provided genome (Human.B38) and gene model (Ensembl.R83). Next, htseq-count (http://www-huber.embl.de/HTSeq) with intersection-nonempty mode and Homo_sapiens.GRCh38.83.gtf file (ftp://ftp.ensembl.org/pub/release-83/gtf/homo_sapiens/Homo_sapiens.GRCh38.83.gtf.gz) was used to count for each gene how many aligned reads overlap its exons.

Differential expression analyses of the per-gene counts were performed with DESeq2 (http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html). For the SRP004879 data set, genes with $-\log_{10}p \geq 38/|\log_2 FC|-16$ were identified as differentially expressed genes; for the SRP056863 data set, genes with $-\log_{10}p \geq 15/|\log_2 FC|-5$ were identified as differentially expressed genes; where p is the Wald test p-value from DESeq2, $\log_2 FC$ is the $\log_2$ fold change (MAP, maximum *a posteriori*) from DESeq2. Subsequently, the genes with differences in median normalized counts between AD and control below 20 were omitted from the list of differentially expressed genes.

University of Malaya

2.5 Network analysis

Network analyses were performed using NetDecoder (http://www.NetDecoder.org). Each of the analyses was given the DEG set and its corresponding gene expression profile. NetDecoder parameters were set as -corThreshold 0.5 -ratioThreshold 5 -top 10 -g none and -overlap was not provided.

2.6 Enrichment analysis

Enrichment analyses were performed using the Database for Annotation, Visualization and Integration Discovery (DAVID) Functional Annotation Tool (http://david.abcc.ncifcrf.gov/summary.jsp). Each of the analyses was given the DEG list or the list of genes contained in prioritized AD-specific subnetwork and its corresponding gene population background with KEGG_PATHWAY annotation category to identify the most relevant biological pathways associated with the given gene list.

University of Malaya

CHAPTER 3: RESULTS

3.1 Comparisons of RNA-seq differential expression analysis methods

For the analysis of the Rajkumar data set, the DEG set size and expected PPV of each method are given in Table 3.1.1. Only voom and edgeR produced gene sets with sizes that had reasonable order of magnitude, whereas the rest either returned too few (DESeq2, NOISeq) or too many (DESeq). However, voom had relatively higher expected PPV over edgeR. In addition, the DEG set size called using voom had standard error (SE) that was about 4 times smaller than that of edgeR's (1000 iterations of bootstrap sampling with replacement of biological replicates). For these reasons, voom was considered to be the better choice for constructing the reference DEG set, and was therefore used to set the reference DEG set for the Bottomly data set.

An interesting observation in Table 3.1.1 relates to the fact that DESeq2 called substantially less DEGs compared to DESeq. It is possible that the implementation of a shrinkage estimation of dispersion parameter and fold change to improve the performance of DESeq for DESeq2 can lead to over-correction that yields too few DEGs.

Table 3.1.1: DEG set sizes and PPVs of the five differential expression analysis methods considered with respect to the Rajkumar data set. The values given are means ± standard errors.

Method	DEG set size	PPV (%)
voom	287 ± 43	88.9 ± 4.1
edgeR	564 ± 694	72.6 ± 15.0
DESeq	3384	Not relevant
NOISeq	31	Not relevant
DESeq2	10	Not relevant

University of Malaya

Table 3.1.2 shows the behaviour of the four differential expression analysis methods considered with respect to the Bottomly data set (see also Figure 3.1.1). For n=3 and n=6, DESeq2 was found to have the best PPV and sensitivity balance as indicated by the top F-score of DESeq2. The second rank of methods was edgeR, followed by NOISeq, and finally DESeq. Comparing DESeq2 with its closest competitor, DESeq2 returned DEG sets with sizes that were reasonably large (270 ± 128 for n=3; 390 ± 84 for n=6), whereas DEG sets returned by edgeR were too large in size (780 ± 199 for n=3; 854 ± 118 for n=6). Note also that doubling the sample size from 3 to 6 for each method led to ~49% increase in F-score for DESeq2, but only ~21% for edgeR, ~26% for NOISeq, and ~33% for DESeq.

Table 3.1.2: PPVs, sensitivities, DEG set sizes and F-scores of the four differential expression analysis methods considered with respect to the Bottomly data set and the two simulation scenarios (n=3 and n=6). The values given are means ± standard deviations.

	DESeq2	edgeR	NOISeq	DESeq
n = 3				
PPV (%)	52.5 ± 10.8	28.7 ± 4.1	40.2 ± 13.8	10.9 ± 0.7
Sensitivity (%)	36.0 ± 5.7	59.8 ± 5.4	20.1 ± 9.9	48.7 ± 4.9
DEG set size	270 ± 128	780 ± 199	268 ± 297	1619 ± 97
F-score	0.43	0.39	0.27	0.18
n = 6				
PPV (%)	62.1 ± 7.7	33.9 ± 3.0	50.4 ± 8.6	14.7 ± 0.5
Sensitivity (%)	65.1 ± 4.5	79.0 ± 4.6	26.2 ± 6.8	67.8 ± 3.5
DEG set size	390 ± 84	854 ± 118	208 ± 140	1671 ± 66
F-score	0.64	0.47	0.34	0.24

University of Malaya

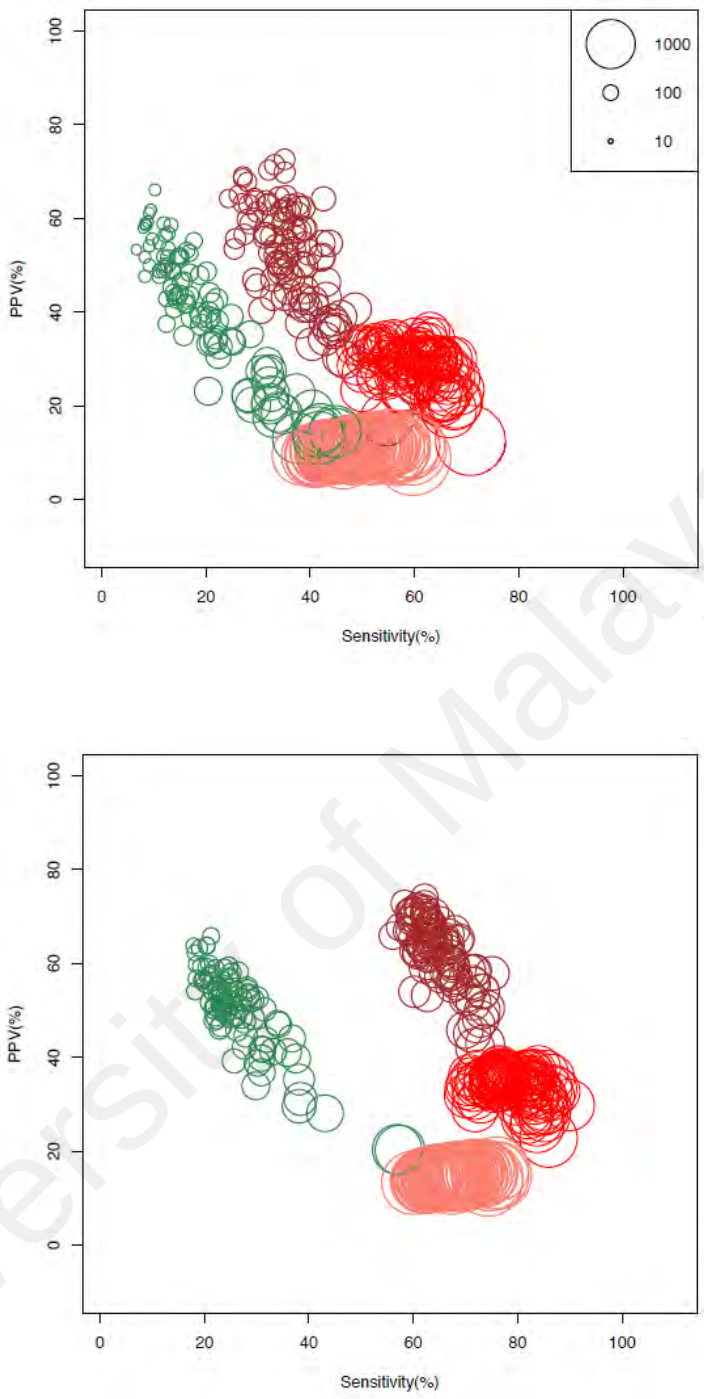


Figure 3.1.1: Scatter plots of PPV against sensitivity for the four differential expression analysis methods considered with respect to the Bottomly data set and the two simulation scenarios (n=3 and n=6). Scatter plot for the n = 3 scenario is given in upper panel and scatter plot for the n=6 scenario is given in lower panel. The diameters of circles are proportional to the DEG set sizes. The brown colour represents DESeq2, the red colour represents edgeR, the green colour represents NOISeq and the pink colour represents DESeq.

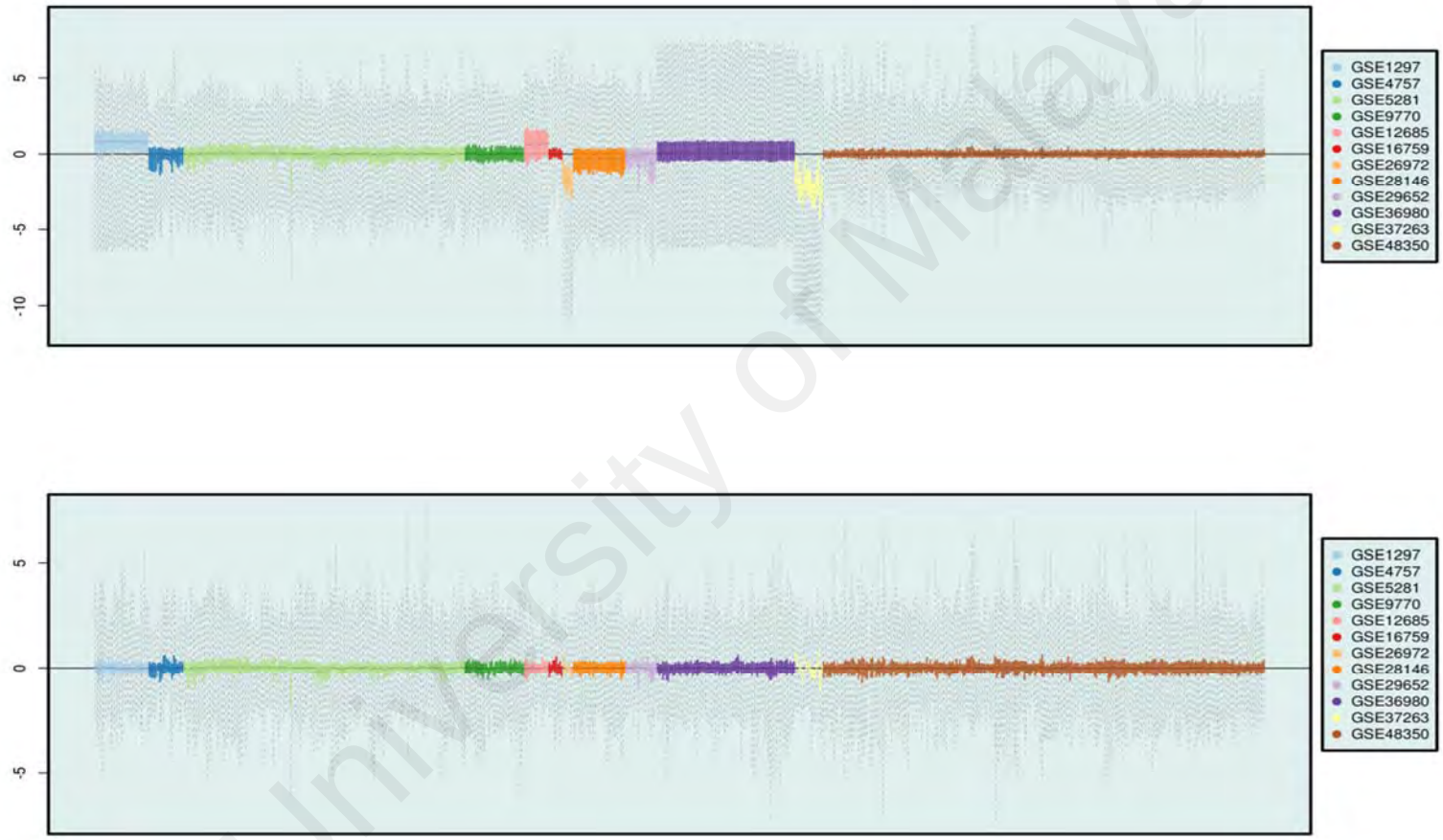
University of Malaya

To summarize, DESeq2 is the clear method of choice for differential expression analysis when sample size is small. Hence, for the two RNA-seq data sets (SRP004879: 3 AD and 3 control samples; SRP056863: 8 AD and 7 control samples), DESeq2 was used to identify DEGs.

3.2 Assessing the effectiveness of ComBat

Figure 3.2.1 shows the RLE plots for assessing the effectiveness of ComBat. The top panel shows the presence of batch effects before applying ComBat, since the median of the samples were not all sampled at 0, and the variances were not approximately constant as indicated by the large variation in the length of the box plot whiskers. The bottom panel shows that most samples have median centered at 0 after applying ComBat. The box plot whisker lengths were also relatively similar compared to before application of ComBat. Based on the comparison, the batch effect removal by ComBat appeared to be effective. Figure 3.2.2 further supports this conclusion. The left panel shows that the clustering of samples was by experiment before applying ComBat. The right panel shows the presence of two approximate clusters defined by disease condition after applying ComBat. The overlap between the two clusters is expected, since distances between samples were computed using the entire vector of expression level, which is affected by other sources of variation such as age, sex, ethnicity, comorbidity, and brain tissue source, which are unrelated to disease condition.

University of Malaya



University of Malaya

Figure 3.2.1: Relative log expression (RLE) plots for the merged microarray data set before (top) and after (bottom) applying ComBat.

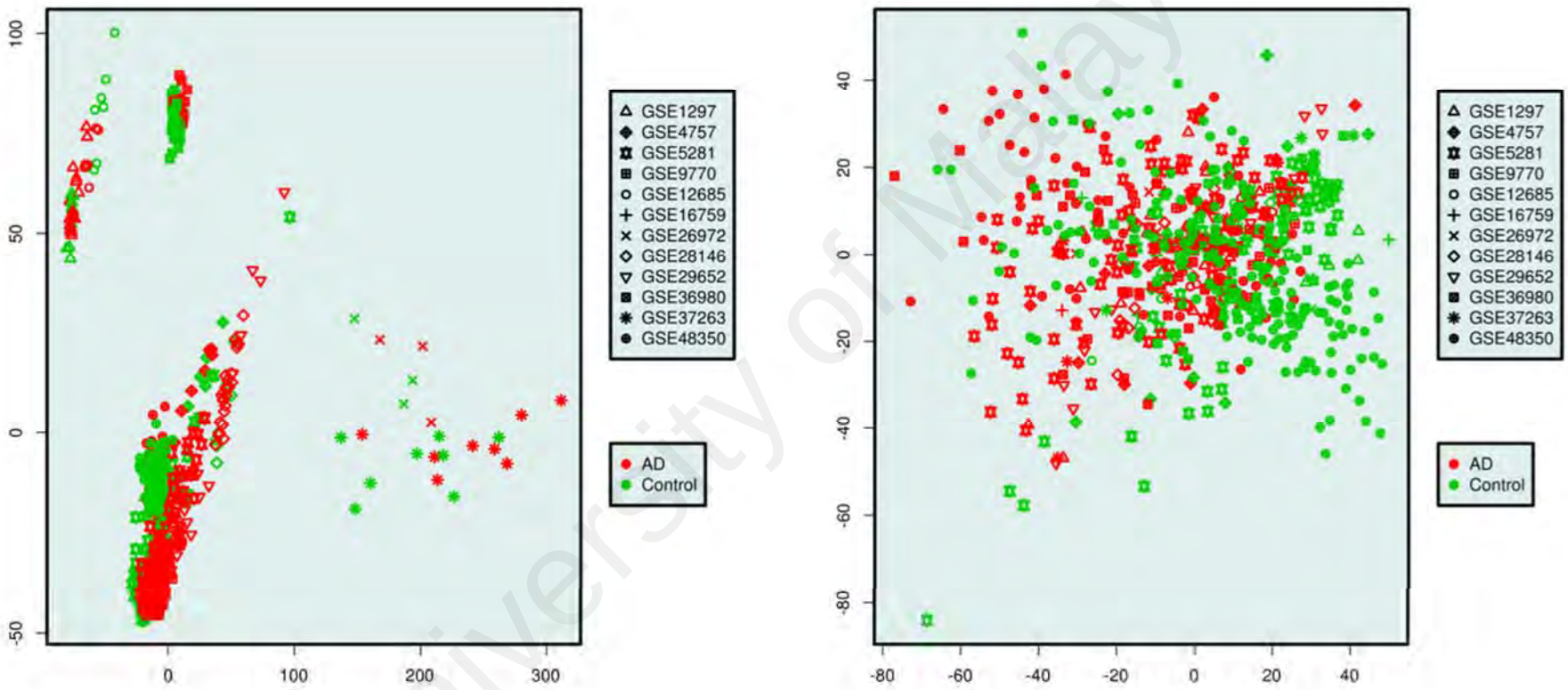


Figure 3.2.2: Multidimensional scaling (MDS) plots for the merged microarray data set before (left) and after (right) applying ComBat.

University of Malaya

3.3 Differential expression analyses

The volcano plots show the joint distribution of $\log_2$FC and $-\log_{10}p$ of each gene are given in Figure 3.3.1 (for the merged microarray data set), Figure 3.3.2 (for SRP004879 data set) and Figure 3.3.3 (for SRP056863 data set). Table 3.3.1 shows the acceptance region used to select DEG. For the merged microarray data set, the genes at the acceptance region were identified as DEG. For SRP004879 and SRP056863 data sets, the genes at the acceptance region were further filtered as described in Chapter 2.4 and the remainder of genes were identified as DEG. In general, the number of DEG ranged from 100 to 300. For the merged microarray data set and SRP056863 data set, there are about 4 times and 2 times more downregulated genes, respectively. In contrast, for SRP004879 data set, the number of upregulated genes was about 50% more than downregulated genes.

Table 3.3.1: DEG acceptance region and the number of DEG (upregulated and downregulated genes).

Data set	DEG acceptance region	Number of DEG	Upregulated in AD	Downregulated in AD
Merged microarray	$-\log_{10}p \geq 16/\lvert \log_2\text{FC} \rvert - 17$	269	52	217
SRP004879	$-\log_{10}p \geq 38/\lvert \log_2\text{FC} \rvert - 16$	226	140	86
SRP056863	$-\log_{10}p \geq 15/\lvert \log_2\text{FC} \rvert - 5$	112	37	75

University of Malaya

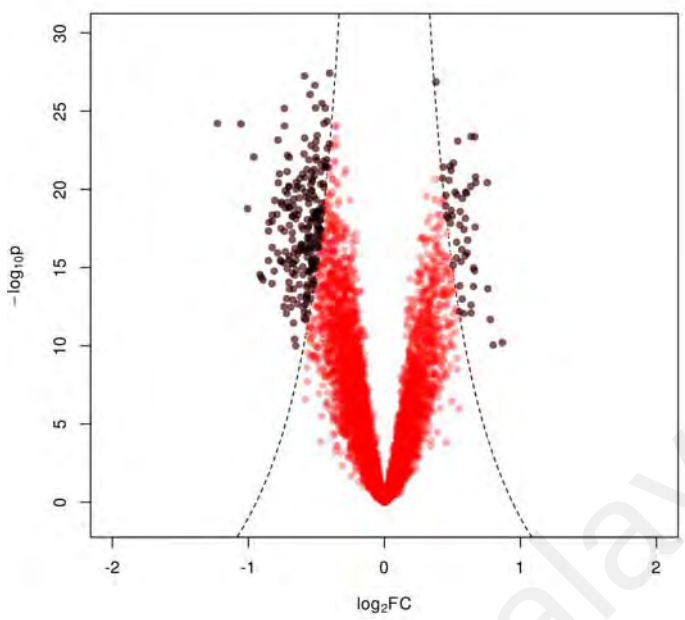


Figure 3.3.1: Volcano plot for the merged microarray data set. Dash line represents -$\log_{10}p$ = 16/ | $\log_2$FC | -17. Brown dots represent genes with -$\log_{10}p \geq$ 16/ | $\log_2$FC | -17 and red dots represent genes with -$\log_{10}p <$ 16/ | $\log_2$FC | -17.

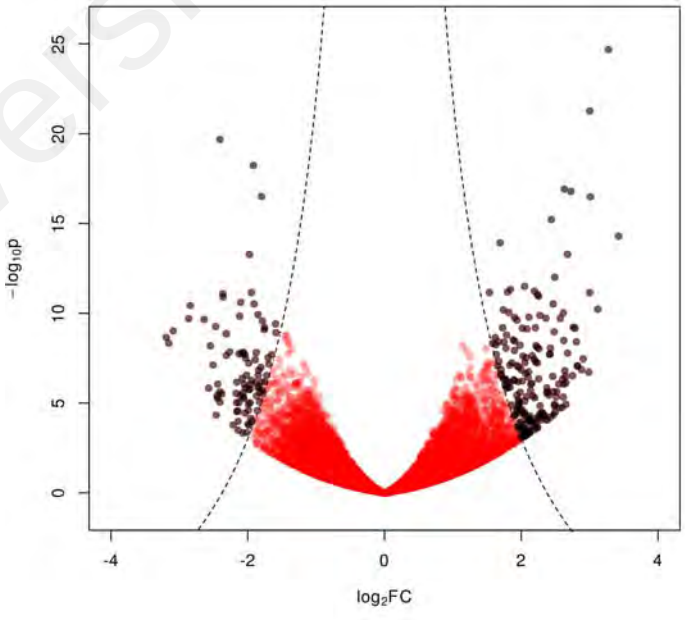


Figure 3.3.2: Volcano plot for SRP004879 data set. Dash line represents -$\log_{10}p$ = 38/ | $\log_2$FC | -16. Brown dots represent genes with -$\log_{10}p \geq$ 38/ | $\log_2$FC | -16 and red dots represent genes with -$\log_{10}p <$ 38/ | $\log_2$FC | -16.

University of Malaya

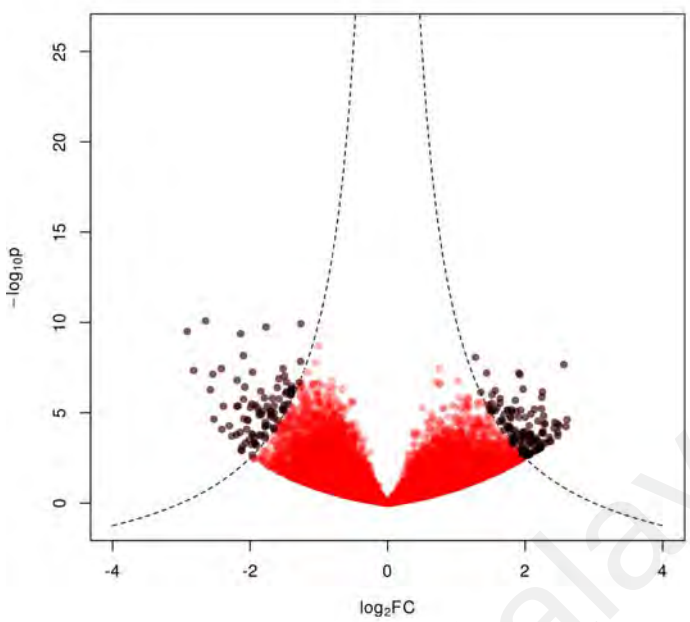


Figure 3.3.3: Volcano plot for SRP056863 data set. Dash line represents $-\log_{10}p = 15/|\log_2 FC| - 5$. Brown dots represent genes with $-\log_{10}p \geq 15/|\log_2 FC| - 5$ and red dots represent genes with $-\log_{10}p < 15/|\log_2 FC| - 5$.

For the merged microarray data set, the DEG signature profile of the control group appeared qualitatively to have sufficient dissimilarity with that of the AD group (Figure 3.3.4). For the two RNA-seq data sets, the DEG set selected using the present methodology allowed unambiguous association of DEG signature with disease status (Figure 3.3.5 and Figure 3.3.6).

For complete lists of DEGs, see Appendix A for the merged microarray data set, Appendix B for SRP004879 data set and Appendix C for SRP056863 data set.

University of Malaya

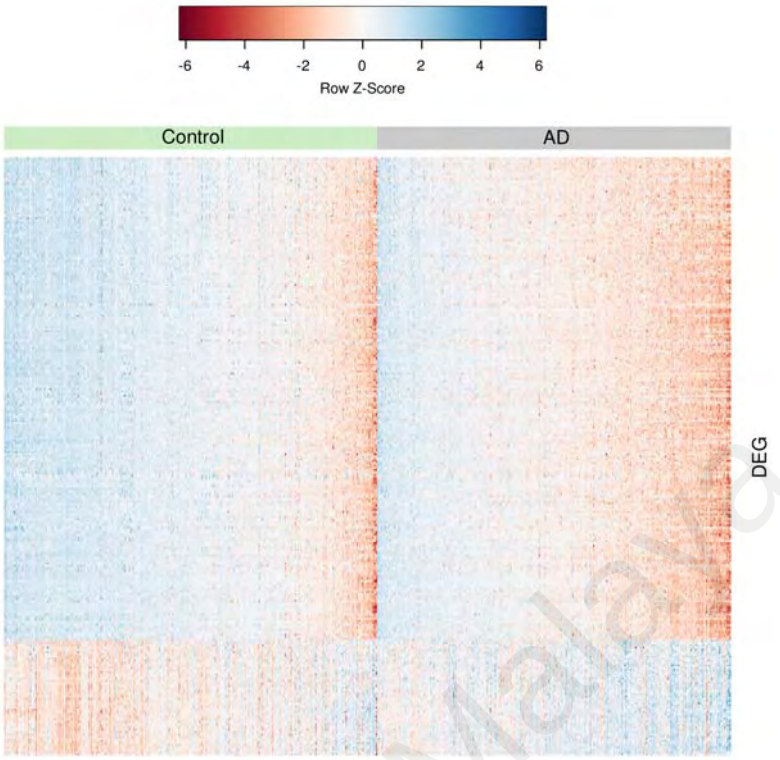


Figure 3.3.4: Heat map of DEG expression profile for the merged microarray data set.

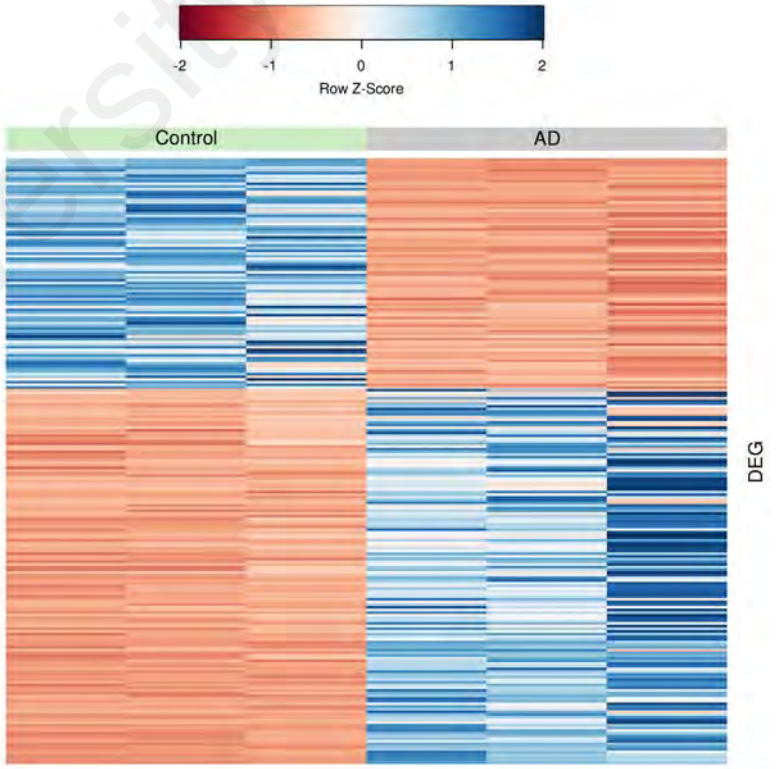


Figure 3.3.5: Heat map of DEG expression profile for SRP004879 data set.

University of Malawi

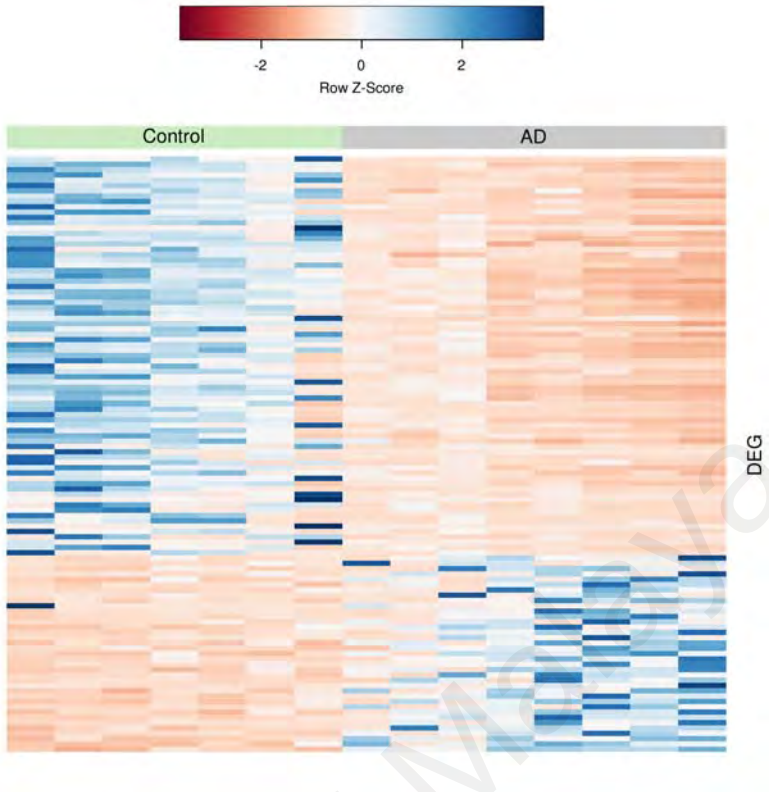


Figure 3.3.6: Heat map of DEG expression profile for SRP056863 data set.

3.4 Network analyses using NetDecoder

The prioritized AD-specific subnetworks returned by NetDecoder provide clues for the aetiology of AD. They are shown in Figure 3.4.1 (for the merged microarray data set), Figure 3.4.2 (for SRP004879 data set) and Figure 3.4.3 (for SRP056863 data set). The identities of the source, intermediary and target genes in the prioritized AD-specific subnetworks returned by NetDecoder are shown in Table 3.4.1. Key genes returned by NetDecoder are shown in Figure 3.4.4 (high impact genes) and Figure 3.4.5 (network routers and key targets). In the literature, a number of genes have been reported to be associated with AD (Table 3.4.2).

University of Malaya

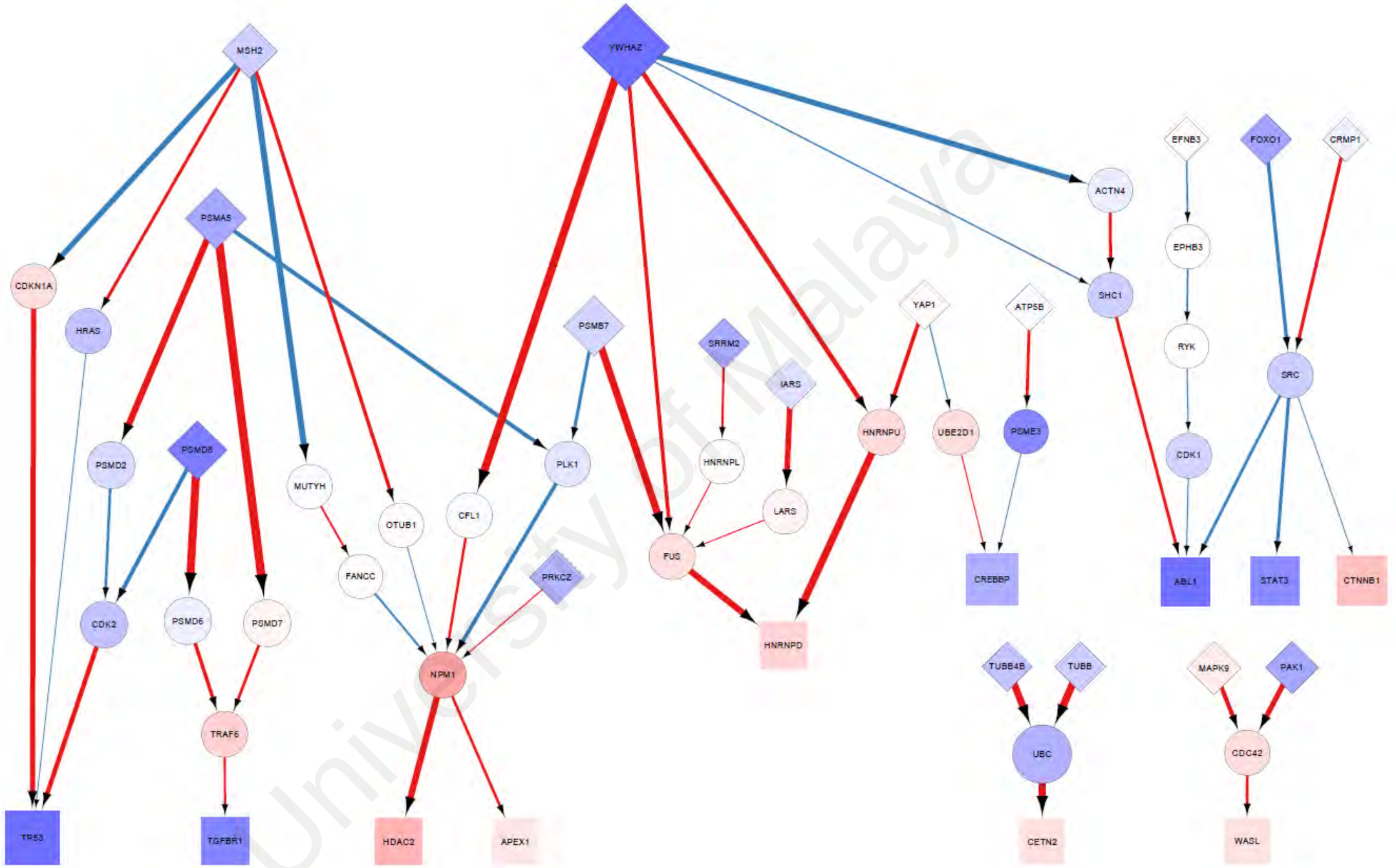


Figure 3.4.1: Prioritized AD-specific subnetwork for the merged microarray data set. These paths start at source genes (diamond shape of nodes), passing through intermediary genes (circle shape of nodes) and end at target genes (square shape of nodes). Edge width represents the amount of flow through an edge. Red edge represents positive PCC and blue edge represents negative PCC. Node size represents the total flow (the in and out flows) at a node. Nodes are coloured according to the node flow difference – red represents high flow in AD but low flow in control; conversely, blue represents low flow in AD but high flow in control.

University of Malaya

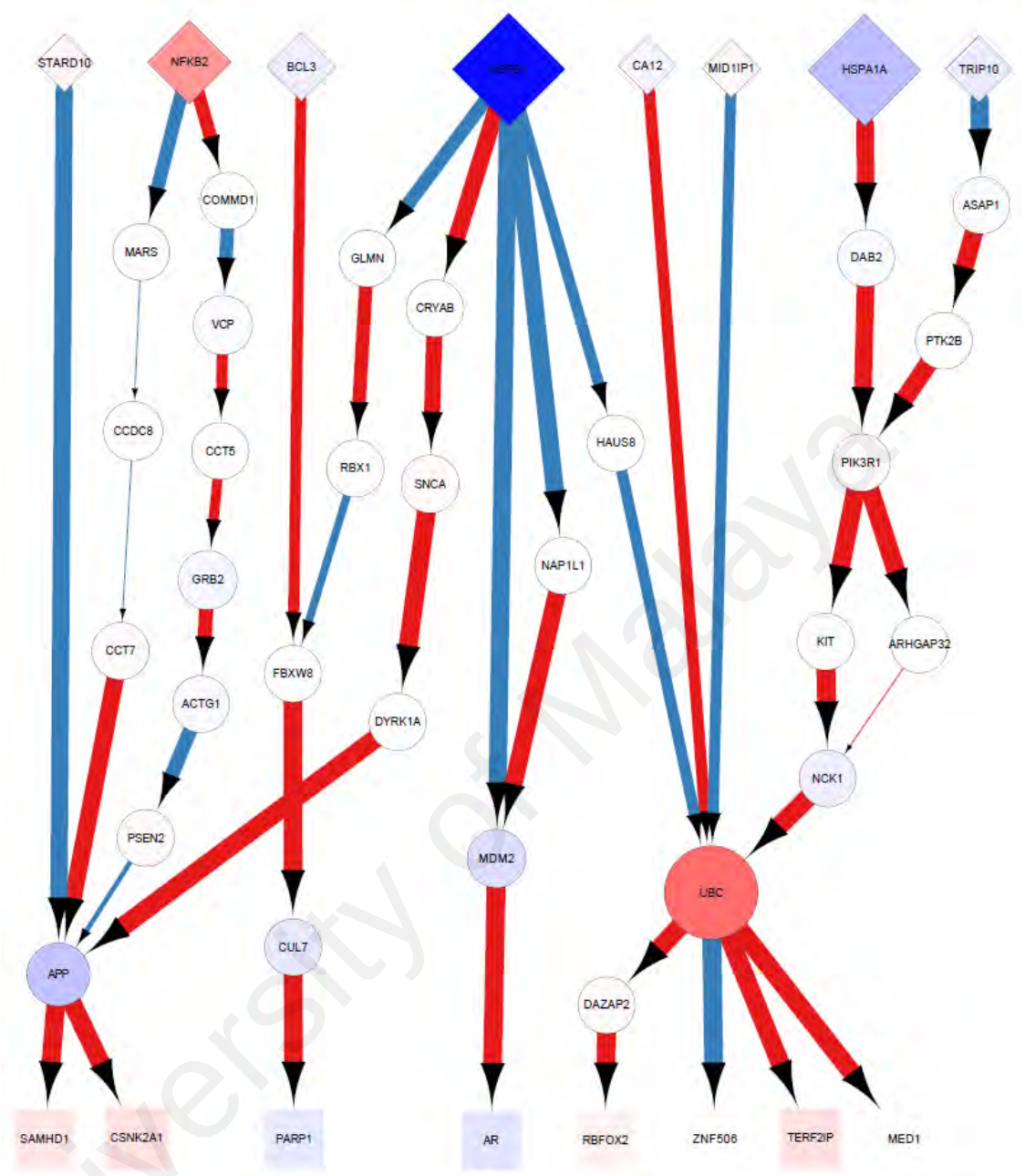


Figure 3.4.2: Prioritized AD-specific subnetwork for SRP004879 data set. These paths start at source genes (diamond shape of nodes), passing through intermediary genes (circle shape of nodes) and end at target genes (square shape of nodes). Edge width represents the amount of flow through an edge. Red edge represents positive PCC and blue edge represents negative PCC. Node size represents the total flow (the in and out flows) at a node. Nodes are coloured according to the node flow difference – red represents high flow in AD but low flow in control; conversely, blue represents low flow in AD but high flow in control.

University of Malaya

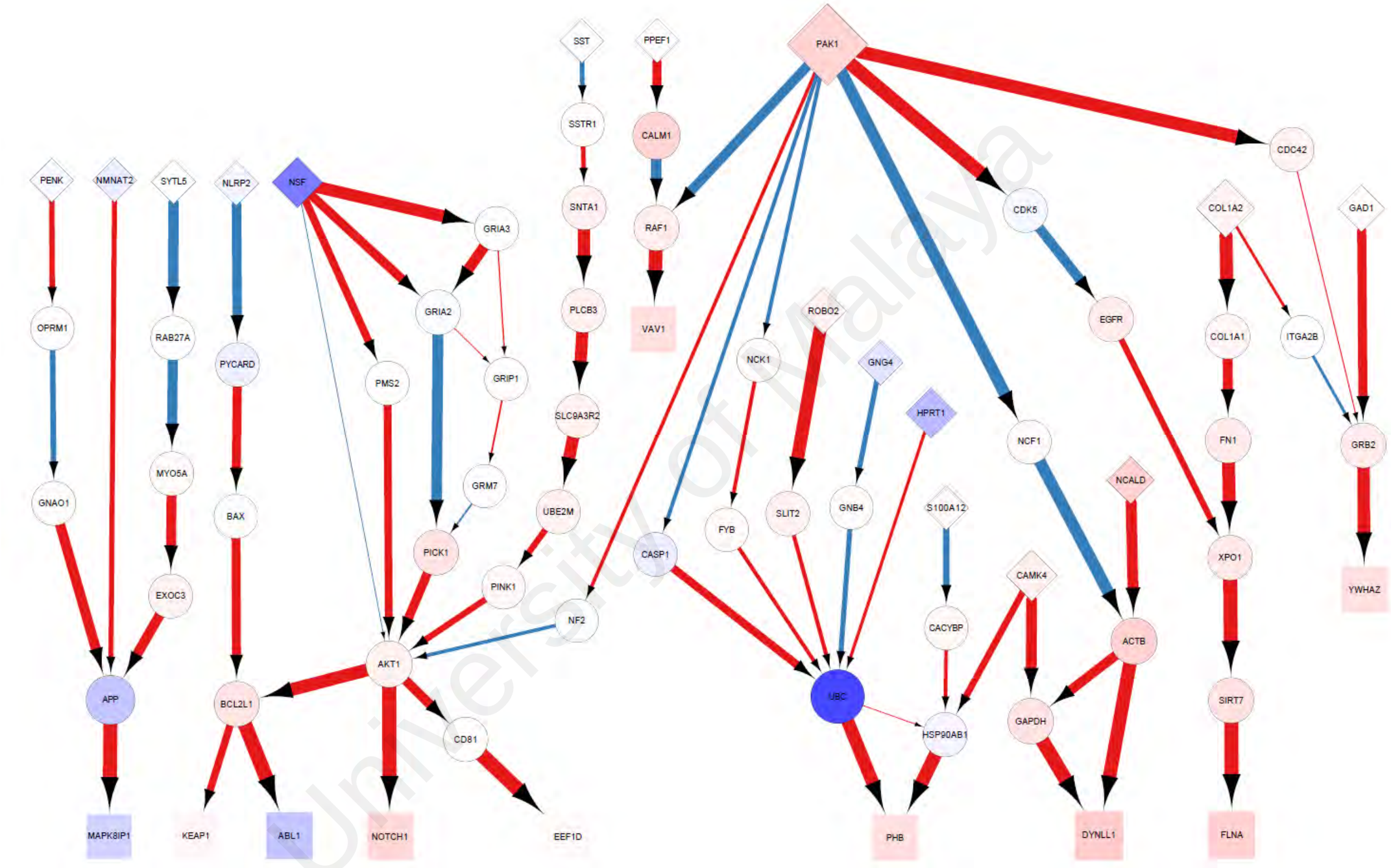


Figure 3.4.3: Prioritized AD-specific subnetwork for SRP056863 data set. These paths start at source genes (diamond shape of nodes), passing through intermediary genes (circle shape of nodes) and end at target genes (square shape of nodes). Edge width represents the amount of flow through an edge. Red edge represents positive PCC and blue edge represents negative PCC. Node size represents the total flow (the in and out flows) at a node. Nodes are coloured according to the node flow difference – red represents high flow in AD but low flow in control; conversely, blue represents low flow in AD but high flow in control.

University of Malaya

Table 3.4.1: Source, intermediary, and target genes in the prioritized AD-specific subnetworks.

Data set	Gene		
	Source	Intermediary	Target
Merged microarray	ATP5B	ACTN4	ABL1
	CRMP1	CDC42	APEX1
	EFNB3	CDK1	CETN2
	FOXO1	CDK2	CREBBP
	IARS	CDKN1A	CTNNB1
	MAPK9	CFL1	HDAC2
	MSH2	EPHB3	HNRNPD
	PAK1	FANCC	STAT3
	PRKCZ	FUS	TGFBR1
	PSMA5	HNRNPL	TP53
	PSMB7	HNRNPU	WASL
	PSMD8	HRAS	
	SRRM2	LARS	
	TUBB	MUTYH	
	TUBB4B	NPM1	
	YAP1	OTUB1	
	YWHAZ	PLK1	
		PSMD2	
		PSMD6	
		PSMD7	
		PSME3	
		RYK	
		SHC1	
		SRC	
		TRAF6	
		UBC	
		UBE2D1	
Total	17	27	11

University of Malaya

Table 3.4.1, continued.

Data set	Gene		
	Source	Intermediary	Target
SRP004879	BCL3	ACTG1	AR
	CA12	APP	CSNK2A1
	HSPA1A	ARHGAP32	MED1
	HSPB1	ASAP1	PARP1
	MID1IP1	CCDC8	RBFOX2
	NFKB2	CCT5	SAMHD1
	STARD10	CCT7	TERF2IP
	TRIP10	COMMD1	ZNF506
		CRYAB	
		CUL7	
		DAB2	
		DAZAP2	
		DYRK1A	
		FBXW8	
		GLMN	
		GRB2	
		HAUS8	
		KIT	
		MARS	
		MDM2	
		NAP1L1	
		NCK1	
		PIK3R1	
		PSEN2	
		PTK2B	
		RBX1	
		SNCA	
		UBC	
		VCP	
Total	8	29	8

University of Malaya

Table 3.4.1, continued.

Data set	Gene		
	Source	Intermediary	Target
SRP056863	CAMK4	ACTB	ABL1
	COL1A2	AKT1	DYNLL1
	GAD1	APP	EEF1D
	GNG4	BAX	FLNA
	HPRT1	BCL2L1	KEAP1
	NCALD	CACYBP	MAPK8IP1
	NLRP2	CALM1	NOTCH1
	NMNAT2	CASP1	PHB
	NSF	CD81	VAV1
	PAK1	CDC42	YWHAZ
	PENK	CDK5	
	PPEF1	COL1A1	
	ROBO2	EGFR	
	S100A12	EXOC3	
	SST	FN1	
	SYTL5	FYB	
		GAPDH	
		GNAO1	
		GNB4	
		GRB2	
		GRIA2	
		GRIA3	
		GRIP1	
		GRM7	
		HSP90AB1	
		ITGA2B	
		MYO5A	
		NCF1	
		NCK1	
		NF2	
		OPRM1	
		PICK1	
		PINK1	
		PLCB3	
		PMS2	
		PYCARD	
		RAB27A	
		RAF1	
		SIRT7	
		SLC9A3R2	
		SLIT2	
		SNTA1	
		SSTR1	
		UBC	
		UBE2M	
		XPO1	
Total	16	46	10

University of Malaya

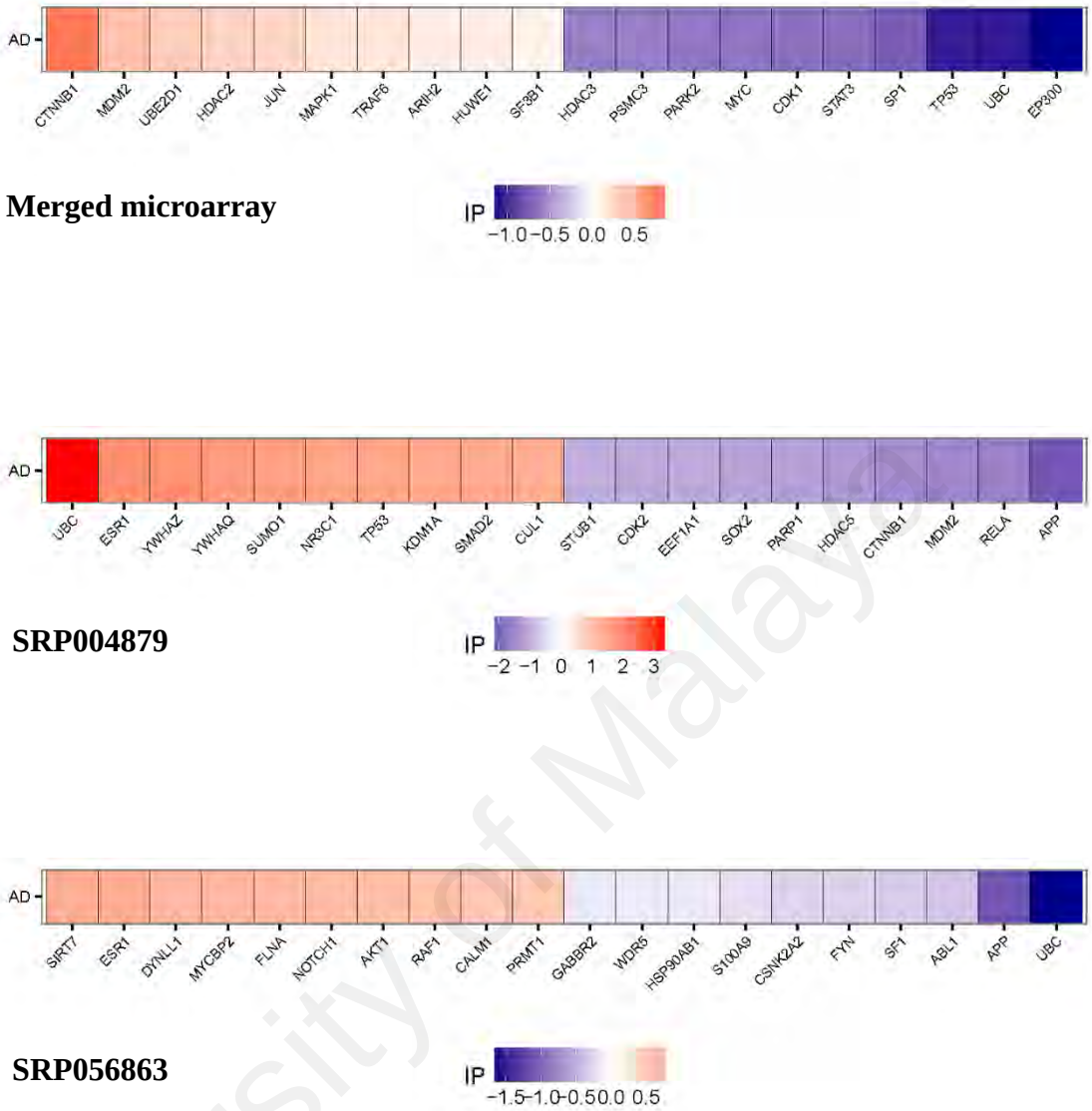


Figure 3.4.4: High impact genes for the merged microarray, SRP004879 and SRP056863 data sets. Impact score (IP) returned by NetDecoder ranks genes based on their importance in mediating differences in flow profiles between control and AD states. Genes with high IP scores (positive or negative) are defined as high impact genes. Genes with larger magnitudes of IP scores are more likely to be involved in AD aetiology. Heat maps for top 10 genes with high positive IP scores and top 10 genes with high negative IP scores are shown.

University of Malaya

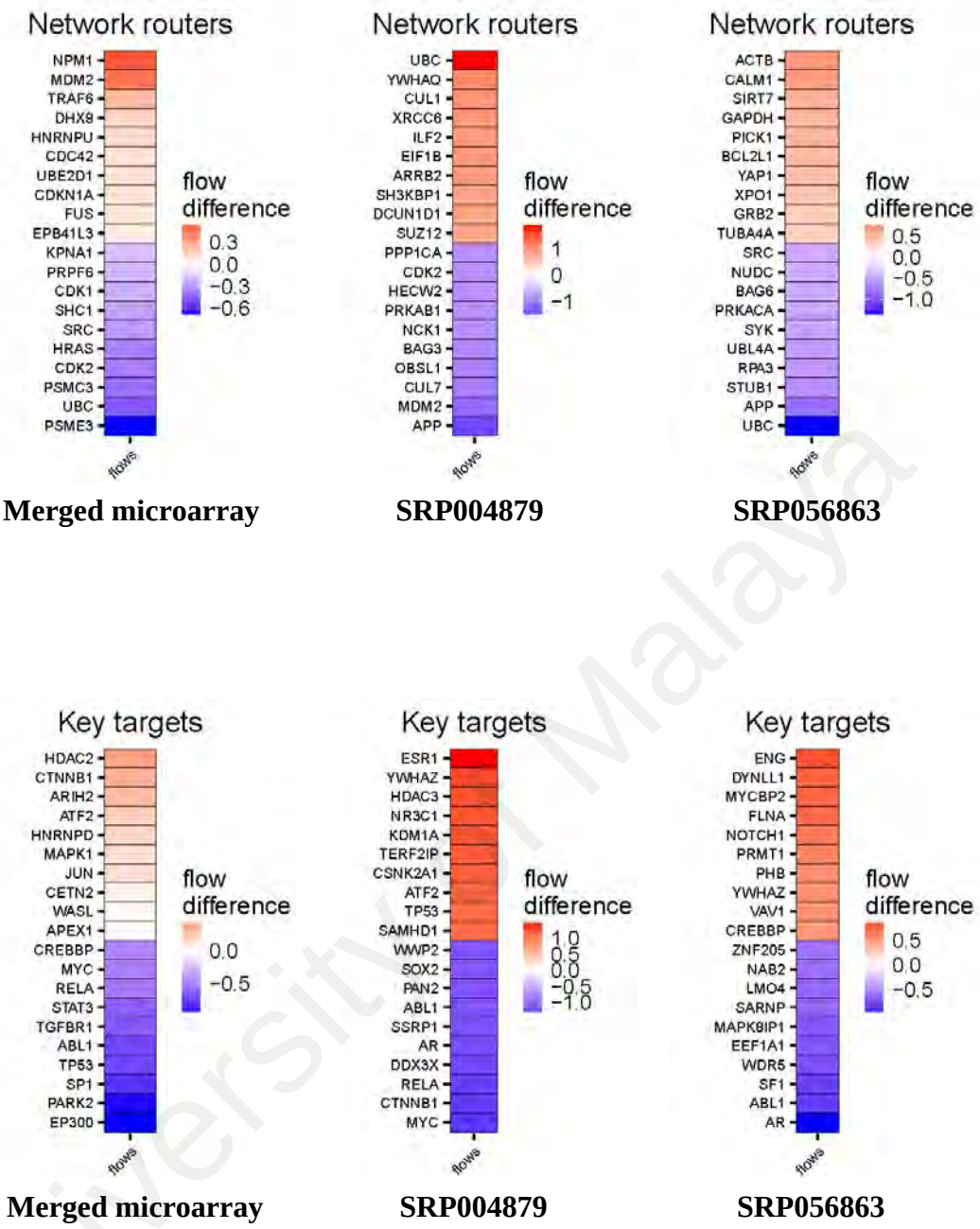


Figure 3.4.5: Network routers and key targets for the merged microarray, SRP004879 and SRP056863 data sets. Network routers are intermediary genes that show high flow differences (positive or negative) between control and AD states. Key targets are target (or sink) genes (transcriptional regulators) that show high flow differences (positive or negative) between control and AD states. Heat maps for top 10 genes with high positive flow differences and top 10 genes with high negative flow differences are shown.

University of Malaya

Table 3.4.2: Fifteen genes known to be associated with AD. Genes that are included in the prioritized AD-specific subnetworks are represented by ticks. Key gene types (high impact gene, network router, key target) are shown in parentheses.

Gene	Merged microarray	SRP004879	SRP056863
ABL1	√ (key target)	(key target)	√ (high impact gene, key target)
AKT1			√ (high impact gene)
APP		√ (high impact gene, network router)	√ (high impact gene, network router)
CSNK2A1		√ (key target)	
CTNNB1	√ (high impact gene, key target)	(high impact gene, key target)	
DYNLL1			√ (high impact gene, key target)
HSPB1		√	
NPM1	√ (network router)		
NSF			√
PAK1	√		√
PSME3	√ (network router)		
SIRT7			√ (high impact gene, network router)
TP53	√ (high impact gene, key target)	(high impact gene, key target)	
UBC	√ (high impact gene, network router)	√ (high impact gene, network router)	√ (high impact gene, network router)
YWHAZ	√	(high impact gene, key target)	√ (key target)

University of Malaya

In the prioritized AD-specific subnetworks of all three data sets, UBC (polyubiquitin-C) is present as high impact gene and network router (Table 3.4.2). UBC is involved in the ubiquitin-proteasome complex which targets proteins for degradation (Vilchez et al., 2014). Protein aggregation is a hallmark of AD brains (Ross & Poirier, 2004) and occurs when defects in UBC prevents clearance of misfolded proteins. The resulting protein plaques lead to cellular cytoskeletal pathologies in AD (Bamburg & Bloom, 2009).

APP recovery (as high impact gene and network router) in the prioritized AD-specific subnetworks of the RNA-seq data sets (Table 3.4.2) adds confidence that the subnetworks are biologically meaningful, since APP is the known precursor molecule of the Aβ peptide, which makes up the neuritic deposits found in AD brain.

YWHAZ is a ubiquitous signaling protein in numerous essential cellular processes (Aitken, 2006) and Miller et al. (2008) showed that this protein was correlated with AD and aging and recommended its continued study.

NPM1 (nucleophosmin) is a nucleolar protein with histone-binding property and is involved in chromatin organization (Tamada et al., 2006). Importantly, altered NPM1 gene expression was observed in the CA1 region of the hippocampus during early stage AD, suggesting of nucleolar stress (Hernández-Ortega et al., 2016). Impaired nucleolar activity can contribute to the pathogenesis of neurodegenerative diseases (Erickson & Bazan, 2013).

HSPB1 has been reported to be protective against the neurotoxic effect of Aβ peptides, as well as involved in modulating cellular APP levels, though the exact mechanism is unclear (Conway et al., 2014).

NSF (N-ethylmaleimide-sensitive factor) is associated with the SNARE (Soluble N-ethylmaleimide-sensitive factor attachment protein receptor) proteins which are essential

University of Malaya

components that regulate neurotransmitter exocytosis at the presynaptic site (Söllner et al., 1993). Presynaptic dysfunction produces cognitive alterations in AD (Terry et al., 1991).

AKT1 is a serine/threonine kinase that regulates the activity of glycogen synthase kinase-3 (GSK-3), which is involved in hyperphosphorylation of tau proteins that form neurofibrillary tangles, and Aβ peptide production and deposition (Bhat & Budd, 2002). Increased activities were found to be significantly increased in the soluble fractions of the mid-temporal cortex of AD brains, compared to non-AD controls (Rickle et al., 2004).

PAK1 (p21-activated kinase) is an important regulator of actin cytoskeleton, and influences dendritic spine morphogenesis. Loss of PAK1 is associated with cognitive defects in AD patients (Zhao et al., 2006).

CTNNB1 encodes the well-known beta-catenin protein, which is a member of the Wnt signaling pathway (Logan & Nusse, 2004), and interacts with presenilin-1, a known subunit of the γ-secretase which cleaves the APP protein (Haass & De Strooper, 1999). Ghanevati and Miller (2005) observed that phospho-beta-catenin accumulation in AD is a consequence of impaired proteosome function.

CSNK2A1 encodes casein kinase 2 alpha 1, which is a serine/threonine protein kinase involved in the phosphorylation of acidic proteins such as casein. Pigino et al. (2009) found that intraneuronal soluble intracellular oligomeric Aβ causes abnormal activation of CSNK2A1 in the axons, leading to excessive phosphorylation of kinesin-1 which removes the anterograde motor from vesicles. Consequently, fast axonal transport is inhibited. Dysregulated fast axonal transport has been suggested as a pathological mechanism in AD (Morfini et al., 2002; Pigino et al., 2003).

University of Malaya

TP53 encodes the well-known tumor protein p53, which is involved in inducing cell cycle arrest and apoptosis (Levine et al., 1991). Its association with AD was reported by Hooper et al. (2007), who found that p53 could indirectly phosphorylate the tau proteins. Hyperphosphorylated tau proteins lose their ability to regulate axonal transport which then causes neurofibrillary tangles and toxic soluble tau species to accumulate, leading to neurodegeneration (Grundke-Iqbal et al., 1986; Iqbal et al., 2016; Iqbal et al., 2010).

SIRT7 is a member of a class of enzymes called sirtuins, which function as deacetylases that depend on nicotinamide adenine dinucleotide (NAD+) for activity. Sirtuin modulation has been found influence the progression of neurodegenerative disorders such as AD, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and spinal and bulbar muscular atrophy, through its action on modulating transcription factor activity, as well as direct action on deacetylating proteotoxic species (Herskovits & Guarente, 2013).

ABL1 is one of the most well-studied tyrosine kinase in AD. The ABL1 protein is known to phosphorylate tau proteins. In adult mouse neurons, overexpression of ABL1 results in neurodegeneration and neuroinflammation (Schlatterer et al., 2011).

DYNLL1 encodes the dynein light chain 1 which is involved in axonal transport in neurons. Its homologue in *Drosophila melanogaster* - DDLC1, has been shown to be necessary for protein clearance by autophagy which prevents neurodegeneration (Batlevi et al., 2010).

PSME3 encodes the proteasome activator subunit 3 (PA28γ), which functions as a regulator of the 20S proteasome in the cytoplasm to regulate oxidative stress (i.e. enhance the degradation of oxidized protein). As Aβ proteins accumulate, they inhibit 20S

University of Malaya

Proteasome activity, eventually leading to neuron cell death due to the homeostatic breakdown in the cell (Gillardon et al., 2007).

While listing down the known biological functions of the individual genes with respect to AD pathogenesis can provides useful biological context, the resulting information may nonetheless be fragmentary. Instead of examining the biological functions of individual genes in Table 3.4.2, focusing on sets of genes linked in a path (Table 3.4.3) and gene pairs with biologically important PCCs (Table 3.4.4) in the prioritized AD-specific subnetworks may yields additional biological insights.

Table 3.4.3: Selected paths in the prioritized AD-specific subnetworks.

Data set	Selected path
Merged microarray	PSMA5-PLK1-NPM1-HDAC2
SRP004879	HSPB1-CRYAB-SNCA-DYRK1A-APP
SRP056863	PAK1-CASP1-UBC-PHB

Table 3.4.4: Gene pairs with notable PCCs in the prioritized AD-specific subnetworks.

Data set	Gene pair with notable PCC	PCC directionality
Merged microarray	CDK2-TP53	Positive
SRP004879	DYRK1A-APP	Positive
	PSEN2-APP	Negative
SRP056863	UBC-PHB	Positive

University of Malaya

We now relate the selected paths in Table 3.4.3 and gene pairs with notable PCCs in Table 3.4.4 to known association with AD in the literature. If a gene in Table 3.4.3 or Table 3.4.4 has appeared also in Table 3.4.2, its association with AD in the literature will not be repeated.

(i) PSMA5-PLK1-NPM1-HDAC2

PSMA5 encodes the Proteasome subunit alpha type-5 protein which forms one of the subunits of the 20S proteasome complex. A dysfunctional ubiquitin-proteasome system leads to defective protein clearance, resulting in aberrant protein accumulation in neurodegenerative diseases such as AD (Oddo, 2008). Song et al. (2011) showed that inhibition of Plk1 kinase activity or depletion of Plk1 using RNA interference in the hippocampal tissues of AD patients reduces Aβ-induced neuronal cell death, suggesting that Plk1 might be a potential therapeutic target for AD treatment. Feng et al. (2001) found that Plk1 phosphorylates the proteasome subunits, and enhance the proteolytic activity of proteasomes (of which PSMA5 is a member). HDAC2 has recently been shown to cooperate with the transcription factor Sp3 to regulate synaptic gene function. Since overexpression of HDAC2 results in the turning off of genes that are important for memory creation, hence inhibition of the HDAC2-Sp3 complex may be useful for ameliorating cognitive impairment in AD (Yamakawa et al., 2017).

(ii) CDK2-TP53

The positive correlation between CDK2 and TP53 is supported by the findings of Yu et al. (2005), who found that when neuroblastoma cells suffer DNA-damage, they induce a p53-mediated inhibition of cell cycle progression, and induction of cdk2-cyclin E which eventually leads to cell death.

University of Malaya

(iii) HSPB1-CRYAB-SNCA-DYRK1A-APP

This path suggests that the intermediate molecules between HSPB1 and APP may be useful therapeutic targets in AD treatment. It was recently shown that the protein product of CRYAB (αB-crystallin) when attached to the endoplasmic-reticulum, prevents protein aggregate formation (Yamamoto et al., 2014). SNCA (alpha-synuclein) has been reported to have pathogenic interaction with Aβ peptide (Suh & Checler, 2002). Finally, DYRK1A inhibition was recently proposed as a potential therapeutic strategy for treatment of AD (Stotani et al., 2016). Overexpressed DYRK1A contributes to neurofibrillary degeneration via enhanced phosphorylation of APP, resulting in aggregation of Aβ plaques in brain tissues which leads to early-onset neurodegeneration, neuronal loss, and dementia in patients with Down's Syndrome (Wegiel et al., 2011). Patients who have Down's Syndrome suffer from mental retardation, and can show similar neuropathology as AD (Glenner & Wong, 1984a, 1984b). The positive correlation between DYRK1A and APP is therefore consistent with what is known in the literature.

(iv) PSEN2-APP

PSEN2 encodes the major component of the y-secretase enzyme. This enzyme functions in the sequential proteolytic cleavages of APP, and the subsequent formation of Aβ peptides. It was found to be significantly downregulated in the auditory cortex of AD patients relative to controls (Delabio et al., 2014). The negative correlation between PSEN2 and APP is then consistent with elevated APP, which is a hallmark of AD.

University of Malaya

(v) PAK1-CASP1-UBC-PHB

CASP1 encodes caspase-1. Caspases belong to a family of endoproteases that play key links in cellular apoptosis and inflammation (McIlwain et al., 2013). Using a mouse model, Heneka et al. (2013) showed that when the NLRP3 inflammosome detects inflammatory Aβ aggregates, it responds by secreting caspase-1, which activates the cytokine IL-1β (interleukin 1 beta). As a result of the inflammatory environment surrounding the Aβ plaque, APP degradation is downregulated, and destruction of Aβ plaques by microglia is decreased. The association of PAK1 with CASP1 is interesting here, because it has been shown that PAK-1 induced phosphorylation is crucial for CASP1 activation, which in turns activates IL-1β (Basak et al., 2005). Prohibitins (PHB), which consist of the PHB1 and PHB2 subunits, function as membrane scaffolds. They are localized at the inner membrane of the mitochondria and mitochondrial cristae (folds of the inner membranes) are known to be sites of Aβ accumulation (Hansson Petersen et al., 2008). It is known that mitochondrial dysfunction characterizes AD disease pathology (Bonet-Costa et al., 2016; Castellani et al., 2002). Merkwirth et al. (2012) found that inactivation of Phb2 in the mouse forebrain induces early onset tau hyperphosphorylation and formation of filaments in the hippocampus, which result in behavioral and cognitive impairments. Thus, the positive correlation between UBC and PHB implies that downregulation of UBC, which occurs in AD, would reduce PHB expression and eventually lead to AD symptoms.

University of Malaya

3.5 Enrichment analyses using DAVID

Table 3.5.1 shows the most enriched KEGG pathways extracted by DAVID. Enrichment analyses of genes in the prioritized AD-specific subnetworks yielded interesting pathways that have been discussed in the AD literature. For both the merged microarray data set and SRP004879 data set, the most enriched KEGG pathway is Epstein-Barr virus (EBV) infection. For SRP056863 data set, focal adhesion is the most enriched KEGG pathway.

In contrast, enrichment analysis of DEG set for the merged microarray data set recovered two enriched KEGG pathways that are synapse-related, with reported relevance to AD (Sheng et al., 2012). The most enriched pathway is synaptic vesicle cycle. This pathway is implicated in Parkinson's Disease, but is noted to be also associated with AD (Abeliovich & Gitler, 2016). No hits were returned for SRP004879 data set. For SRP056863 data set, GABAergic synapse is the most enriched pathway. Limon et al. (2012) reported the loss of functional $GABA_A$ receptors in the brains of AD patients. The implications of GABAergic neurotransmission in AD were recently discussed by Li et al. (2016).

University of Malaya

Table 3.5.1: The most enriched KEGG pathways extracted by DAVID.

Data set	Input of DAVID	KEGG pathway	Genes in pathway
Merged microarray	Genes in prioritized AD-specific subnetwork	Epstein-Barr virus infection	CDK1, YWHAZ, CREBBP, TP53, CDK2, STAT3, CDKN1A, HDAC2, PSMD2, MAPK9, PSMD6, TRAF6, PSMD7, PSMD8
	DEGs	Synaptic vesicle cycle	SYT1, STX1A, ATP6V1H, ATP6V1B2, ATP6V1D, ATP6V0B, SLC17A7, ATP6V1C1, ATP6V1A, ATP6V1E1, VAMP2, ATP6V0D1, SNAP25, NSF, AP2M1
SRP004879	Genes in prioritized AD-specific subnetwork	Epstein-Barr virus infection	CSNK2A1, MDM2, HSPB1, HSPA1, NFKB2, PIK3R1
	DEGs	-	-
SRP056863	Genes in prioritized AD-specific subnetwork	Focal adhesion	ACTB, EGFR, GRB2, RAF1, VAV1, FLNA, AKT1, CDC42, COL1A2, PAK1, COL1A1, ITGA2B, FN1
	DEGs	GABAergic synapse	SLC32A1, GAD2, GABRA4, GNG4, GAD1, NSF

University of Malaya

CHAPTER 4: DISCUSSION AND CONCLUSION

4.1 Pipeline robustness

Even supposing that ComBat provided effective cross-platform normalization to remove batch effects due to different study, there remains considerable heterogeneity introduced into the gene expression levels after data merging. These sources of variation, such as sex, age, ethnicity, comorbidity, and source of brain tissue, can potentially reduce the statistical power of detecting differentially expressed genes under a marginal gene candidate selection approach, which is used in the present work. On the other hand, the increased sample size may offset this loss of statistical power. Here, the DEG detected from the merged microarray data set produced an interesting hypothesis (together with SRP004879) about the involvement of EBV infection (see section 4.3) after NetDecoder and DAVID enrichment analysis. Since EBV infection has only been speculated recently (Carbone et al., 2014; Licastro & Porcellini, 2016; Mawanda & Wallace, 2013) as a possible process involved in AD pathogenesis, the agreement between the outcome of the *in silico* analysis and the literature appears to support the robustness of the proposed bioinformatic pipeline.

4.2 Biological interpretation of prioritized AD-specific subnetworks

It is important to remember that, while the subnetworks produced in NetDecoder are directed, the directedness does not imply any causative relationship between two interacting genes, and merely serves as a convenience for obtaining the minimum cost path from the source genes to the target genes through intermediary genes. If evidence from the literature is available, directedness of edges between two genes in a path of interest can be inferred, and this would allow the integration of numerous disjointed results in the AD literature into more coherent, testable network-oriented hypotheses.

University of Malaya

Interestingly, in the prioritized AD-specific subnetwork of SRP004879, there is a path from the well-known proinflammation nuclear factor NFKB2 to APP, through a path with highly correlated genes: NFKB2-COMMD1-VCP-CCT5-GRB2-ACTG1-PSEN2-APP. While currently the linkage between inflammation and AD is unclear, researchers have already considered the possible connection (Granic et al., 2009). Interestingly, copper dyshomeostasis is found in AD patients (Lovell et al., 1998; Squitti & Polimanti, 2013), and COMMD1 is the gene involved in copper metabolism. The usefulness of such paths that connect well-known key genes in AD pathogenesis is that it allows the subject matter expert to make use of biochemistry knowledge to infer causality, or design experiments that could allow such conclusions to be made.

To summarise, genes that are known to be associated with AD pathogenesis in the literature were successfully recovered among the intermediary genes in the prioritized AD-specific subnetworks. Furthermore, paths connecting well-known genes in AD can be recovered from the prioritized AD-specific subnetworks, and they provide a basis for integrating known experimental results in the literature. Finally, correlations between genes known to be involved in AD are also consistent with results in the literature. The rich results obtained support NetDecoder as a useful tool for harvesting biologically meaningful subnetworks.

4.3 Enrichment analyses of gene sets

Enrichment analyses of gene sets that define the prioritized AD-specific subnetworks revealed rich results, recovering pathways that have been discussed in the AD literature. Two major pathways were identified from Chapter 3.5, namely focal adhesion (SRP056863) and EBV infection (merged microarray; SRP004879). Focal adhesions (Chen et al., 2003) are macromolecular structures containing integrins that function as

University of Malaya

mechanical links connecting the cellular cytoskeleton to the extra-cellular matrix (ECM). Integrins are transmembrane receptors that mediate cell-cell and cell-ECM adhesions (Howe et al., 1998). Focal adhesions are known to regulate beta-amyloid peptide signaling and cell death in AD (Caltagarone et al., 2007). A viral association with AD is more puzzling, but not totally implausible (Mawanda & Wallace, 2013). While an early study did not detect EBV in peripheral blood cells and post-mortem brain tissues from AD patients and normal controls (Kittur et al., 1992), the most recent evidence suggest EBV infection could be an environmental risk factor for AD progression in elderly patients (Carbone et al., 2014; Licastro & Porcellini, 2016). Thus, it is interesting that the current *in-silico* analysis points to a line of investigation that has until now received very little attention from experimental scientists. It is noteworthy that EBV is also reported to be associated with multiple sclerosis (Zivadinov et al., 2009), another neurodegenerative disease.

In contrast, although enrichment analyses of DEG sets also revealed some pathways associated with AD, the pathways seemed to be more peripheral. For example, in the merged microarray data set, the most enriched pathway is synaptic vesicle cycle, which has been reported to be associated with AD (Sheng et al., 2012). While associated with AD, this pathway is primarily implicated in the pathogenesis of Parkinson's Disease (Abeliovich & Gitler, 2016). For SRP056863, GABAergic synapse is the most enriched pathway. Limon et al. (2012) reported the loss of functional $GABA_A$ receptors in the brains of AD patients. The implications of GABAergic neurotransmission in AD were recently discussed by Li et al. (2016). No pathways were found significantly enriched for SRP004879.

University of Malaya

4.4 Limitations and future study

Although the data merging process appeared to indicate the success of ComBat (see Figure 3.2.1, Figure 3.2.2) in removing batch effects from the 12 independent microarray data sets, there is not yet any systematic analysis of how successful data merging may be in general when applied to other experiments. One possible future work is to focus on empirical assessment of data merging as a feasible integrative analysis approach for gene expression profiles. The two RNA-seq data sets were analyzed separately, since currently there are no algorithms that could reliably remove batch effects arising from the merging of multiple RNA-seq data sets.

There might be some concerns about the stability of results from NetDecoder for SRP004879, which has 3 controls and 3 AD. The reason is because the gene-wise PCCs can only be estimated using 3 data points for each phenotype class. Consequently, the correlation estimates have large standard errors. Because of this, it is reasonable to expect this situation to affect the identification of appropriate subnetworks through a minimum-cost optimization process, since the edge weights and edge costs are functions of the correlation estimate. It is therefore reasonable to be less optimistic about the NetDecoder result for this data set. Surprisingly, the result of enrichment analysis of genes in the prioritized AD-specific subnetwork for SRP004879 (involvement of EBV) is similar to that from the one obtained from the merged microarray, in which the correlation estimates had much lower standard error due to the large sample size (326 AD; 344 control). A possible future work is to perform enrichment analysis of the genes from the subnetwork obtained from ESSNet (Lim et al., 2015), a recent tool for handling analysis of small sample gene expression data sets. If EBV is recovered too, then this additional support suggests a surprising new direction for AD research.

University of Malaya

Assuming that the association of AD with EBV infection is not a false positive, we expect to detect EBV gene expression in the transcriptome data of (age-matched) AD subjects, but not in the control subjects. Unfortunately, we were not able to test this expectation in the only RNA-seq data set SRP004879 where genes of the AD-specific prioritized subnetwork were found to be enriched in EBV infection pathway, because the control (n=3, average about 30 years old) and the AD subjects (n=3, all above 80 years old) were not age-matched. It was also not possible to test this expectation using the merged microarray data, as the platforms were all optimized to be specific to human genes.

4.5 Conclusion

In this thesis, I aimed to identify pathways and network associated with Alzheimer's Disease via a bioinformatic analysis of publicly available transcriptomic data. The integrative analysis of these data has the potential to generate new hypotheses regarding the molecular aspects of AD pathogenesis that is not possible from previous analyses of multiple single transcriptomic data sets. I showed that multiple microarray data sets could be successfully (i.e. with batch effects removed) merged into a super data set (326 AD and 344 control samples) using the ComBat algorithm, thus greatly increasing the statistical power to detect DEGs (using limma) in AD condition that are not possible in the analysis of single data sets. Furthermore, I showed DESeq2 to be optimal for calling DEGs in RNA-seq data sets. Using the DEGs obtained from the merged microarray data set and RNA-seq data sets as input genes in the recently proposed NetDecoder algorithm. I found that the AD-specific prioritized subnetworks contain genes that have been validated in the AD literature, such as UBC, ABL1, YWHAZ, APP, TP53 and CTNNB1. The recovery of known key genes in AD is an important validation of the interpretability of the data mining work flow in producing biologically meaningful results. Additionally,

University of Malaya

the novel paths recovered such as PSMA6-PLK1-NPM1-HDAC2, HSPB1-CRYAB-SNCA-DYRK1A-APP, and PAK1-CASP1-UBC-PHB potentially generate practical hypotheses for inferring previously unsuspected linkages in the systems biology of AD pathogenesis, as all members in the paths have been reported in the AD literature. Importantly, functional enrichment analysis of the genes in the prioritized subnetworks supports the involvement of EBV viral infection with AD progression. A viral hypothesis for AD has only been considered recently (Carbone et al., 2014; Licastro & Porcellini, 2016; Mawanda & Wallace, 2013), so the concordance between the present *in silico* result with opinions in the literature may not be a coincidence.

Since the subnetwork search step of the NetDecoder algorithm relies on edge weights and costs that are a function of the Pearson correlation, transcriptome data sets with small sample size may have unreliably estimated correlation values (e.g. SRP004879 – 3 AD and 3 control samples), which could affect what subnetworks are recovered. Additionally, it should be remembered that causality cannot be inferred from the prioritized subnetworks, but must be inferred from evidence in the literature, or experimentally tested. Despite these concerns, biologically meaningful pathways and functions have been inferred from the present analysis, much more than what is possible traditionally using only DEGs.

To summarize, I believe the integrative analysis of transcriptome data proposed in this thesis has produced biologically meaningful candidate genes for AD research, together with hypotheses about connections between these genes that biologists may find meaningful to explore further. It seems that the work flow can be feasibly extended to similar two-phenotype class problems (e.g. in the study of various cancers, other neurodegenerative diseases, etc.) where publicly available transcriptome data is abundant.

University of Malaya

REFERENCES

Abeliovich, A., & Gitler, A. D. (2016). Defects in trafficking bridge Parkinson's disease pathology and genetics. *Nature, 539*(7628), 207-216.

Aitken, A. (2006). 14-3-3 proteins: A historic overview. *Seminars in Cancer Biology, 16*(3), 162-172.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics, 7*(1), 55-65.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*(10), R106.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166-169.

Bai, B., Hales, C. M., Chen, P. C., Gozal, Y., Dammer, E. B., Fritz, J. J., . . . Peng, J. (2013). U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America, 110*(41), 16562-16567.

Bamburg, J. R., & Bloom, G. S. (2009). Cytoskeletal pathologies of Alzheimer disease. *Cell Motility and the Cytoskeleton, 66*(8), 635-649.

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics, 12*(1), 56-68.

Basak, C., Pathak, S. K., Bhattacharyya, A., Mandal, D., Pathak, S., & Kundu, M. (2005). NF-kappaB- and C/EBPbeta-driven interleukin-1beta gene expression and PAK1-mediated caspase-1 activation play essential roles in interleukin-1beta release from Helicobacter pylori lipopolysaccharide-stimulated macrophages. *Journal of Biological Chemistry, 280*(6), 4279-4288.

Batlevi, Y., Martin, D. N., Pandey, U. B., Simon, C. R., Powers, C. M., Taylor, J. P., & Baehrecke, E. H. (2010). Dynein light chain 1 is required for autophagy, protein clearance, and cell death in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America, 107*(2), 742-747.

Berson, A., Barbash, S., Shaltiel, G., Goll, Y., Hanin, G., Greenberg, D. S., . . . Soreq, H. (2012). Cholinergic-associated loss of hnRNP-A/B in Alzheimer's disease impairs cortical splicing and cognitive function in mice. *EMBO Molecular Medicine, 4*(8), 730-742.

Bhat, R. V., & Budd, S. L. (2002). GSK3beta signalling: Casting a wide net in Alzheimer's disease. *Neurosignals, 11*(5), 251-261.

University of Malaya

Blair, L. J., Nordhues, B. A., Hill, S. E., Scaglione, K. M., O'Leary, J. C., Fontaine, S. N., . . . Dickey, C. A. (2013). Accelerated neurodegeneration through chaperone-mediated oligomerization of tau. *Journal of Clinical Investigation, 123*(10), 4158-4169.

Blalock, E. M., Buechel, H. M., Popovic, J., Geddes, J. W., & Landfield, P. W. (2011). Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *Journal of Chemical Neuroanatomy, 42*(2), 118-126.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., & Landfield, P. W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America, 101*(7), 2173-2178.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bonet-Costa, V., Pomatto, L. C., & Davies, K. J. (2016). The proteasome and oxidative stress in Alzheimer's disease. *Antioxidants & Redox Signaling, 25*(16), 886-901.

Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., . . . Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE, 6*(3), e17820.

Caltagarone, J., Jing, Z., & Bowser, R. (2007). Focal adhesions regulate Abeta signaling and cell death in Alzheimer's disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1772*(4), 438-445.

Carbone, I., Lazzarotto, T., Ianni, M., Porcellini, E., Forti, P., Masliah, E., . . . Licastro, F. (2014). Herpes virus in Alzheimer's disease: Relation to progression of the disease. *Neurobiology of Aging, 35*(1), 122-129.

Castellani, R., Hirai, K., Aliev, G., Drew, K. L., Nunomura, A., Takeda, A., . . . Smith, M. A. (2002). Role of mitochondrial dysfunction in Alzheimer's disease. *Journal of Neuroscience Research, 70*(3), 357-360.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE, 6*(2), e17238.

Chen, C. S., Alonso, J. L., Ostuni, E., Whitesides, G. M., & Ingber, D. E. (2003). Cell shape provides global control of focal adhesion assembly. *Biochemical and Biophysical Research Communications, 307*(2), 355-361.

Conway, M., Nafar, F., Straka, T., & Mearow, K. (2014). Modulation of amyloid-β protein precursor expression by HspB1. *Journal of Alzheimer's Disease, 42*(2), 435-450.

University of Malaya

da Rocha, E. L., Ung, C. Y., McGehee, C. D., Correia, C., & Li, H. (2016). NetDecoder: A network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Research, 44*(10), e100.

Delabio, R., Rasmussen, L., Mizumoto, I., Viani, G. A., Chen, E., Villares, J., . . . Payão, S. L. (2014). PSEN1 and PSEN2 gene expression in Alzheimer's disease brain: A new approach. *Journal of Alzheimer's Disease, 42*(3), 757-760.

Delmar, P., Robin, S., & Daudin, J. J. (2005). VarMixt: Efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics, 21*(4), 502-508.

Dunckley, T., Beach, T. G., Ramsey, K. E., Grover, A., Mastroeni, D., Walker, D. G., . . . Stephan, D. A. (2006). Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. *Neurobiology of Aging, 27*(10), 1359-1371.

Erickson, J. D., & Bazan, N. G. (2013). The nucleolus fine-tunes the orchestration of an early neuroprotection response in neurodegeneration. *Cell Death & Differentiation, 20*(11), 1435-1437.

Evans, D. A., Funkenstein, H. H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., . . . Taylor, J. O. (1989). Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported. *JAMA, 262*(18), 2551-2556.

Feng, Y., Longo, D. L., & Ferris, D. K. (2001). Polo-like kinase interacts with proteasomes and regulates their activity. *Cell Growth & Differentiation, 12*(1), 29-37.

Fonseca, N. A., Marioni, J., & Brazma, A. (2014). RNA-Seq gene profiling--a systematic empirical comparison. *PLoS ONE, 9*(9), e107026.

Frazee, A. C., Langmead, B., & Leek, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics, 12*, 449.

Frith, M. C., Pheasant, M., & Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *European Journal of Human Genetics, 13*(8), 894-897.

Ghanevati, M., & Miller, C. A. (2005). Phospho-beta-catenin accumulation in Alzheimer's disease and in aggresomes attributable to proteasome dysfunction. *Journal of Molecular Neuroscience, 25*(1), 79-94.

Gillardon, F., Kloss, A., Berg, M., Neumann, M., Mechtler, K., Hengerer, B., & Dahlmann, B. (2007). The 20S proteasome isolated from Alzheimer's disease brain shows post-translational modifications but unchanged proteolytic activity. *Journal of Neurochemistry, 101*(6), 1483-1490.

Glenner, G. G., & Wong, C. W. (1984a). Alzheimer's disease and Down's syndrome: Sharing of a unique cerebrovascular amyloid fibril protein. *Biochemical and Biophysical Research Communications, 122*(3), 1131-1135.

University of Malaya

Glenner, G. G., & Wong, C. W. (1984b). Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications, 120*(3), 885-890.

Granic, I., Dolga, A. M., Nijholt, I. M., van Dijk, G., & Eisel, U. L. (2009). Inflammation and NF-kappaB in Alzheimer's disease and diabetes. *Journal of Alzheimer's Disease, 16*(4), 809-821.

Grice, D. E., Reenilä, I., Männistö, P. T., Brooks, A. I., Smith, G. G., Golden, G. T., . . . Berrettini, W. H. (2007). Transcriptional profiling of C57 and DBA strains of mice in the absence and presence of morphine. *BMC Genomics, 8*, 76.

Grundke-Iqbal, I., Iqbal, K., Tung, Y. C., Quinlan, M., Wisniewski, H. M., & Binder, L. I. (1986). Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences of the United States of America, 83*(13), 4913-4917.

Guo, Y., Li, C. I., Ye, F., & Shyr, Y. (2013). Evaluation of read count based RNAseq analysis methods. *BMC Genomics, 14 Suppl 8*, S2.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., . . . Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*(8), 1494-1512.

Haass, C., & De Strooper, B. (1999). The presenilins in Alzheimer's disease--proteolysis holds the key. *Science, 286*(5441), 916-919.

Hansson Petersen, C. A., Alikhani, N., Behbahani, H., Wiehager, B., Pavlov, P. F., Alafuzoff, I., . . . Ankarcrona, M. (2008). The amyloid beta-peptide is imported into mitochondria via the TOM import machinery and localized to mitochondrial cristae. *Proceedings of the National Academy of Sciences of the United States of America, 105*(35), 13145-13150.

Heneka, M. T., Kummer, M. P., Stutz, A., Delekate, A., Schwartz, S., Vieira-Saecker, A., . . . Golenbock, D. T. (2013). NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice. *Nature, 493*(7434), 674-678.

Hernández-Ortega, K., Garcia-Esparcia, P., Gil, L., Lucas, J. J., & Ferrer, I. (2016). Altered machinery of protein synthesis in Alzheimer's: From the nucleolus to the ribosome. *Brain Pathology, 26*(5), 593-605.

Herskovits, A. Z., & Guarente, L. (2013). Sirtuin deacetylases in neurodegenerative diseases of aging. *Cell Research, 23*(6), 746-758.

Hokama, M., Oka, S., Leon, J., Ninomiya, T., Honda, H., Sasaki, K., . . . Nakabeppu, Y. (2014). Altered expression of diabetes-related genes in Alzheimer's disease brains: The Hisayama study. *Cerebral Cortex, 24*(9), 2476-2488.

Hooper, C., Meimaridou, E., Tavassoli, M., Melino, G., Lovestone, S., & Killick, R. (2007). p53 is upregulated in Alzheimer's disease and induces tau phosphorylation in HEK293a cells. *Neuroscience Letters, 418*(1), 34-37.

University of Malaya

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., . . . Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proceedings of the National Academy of Sciences of the United States of America, 103*(46), 17402-17407.

Howe, A., Aplin, A. E., Alahari, S. K., & Juliano, R. L. (1998). Integrin signaling and cell growth control. *Current Opinion in Cell Biology, 10*(2), 220-231.

Hu, J., Ge, H., Newman, M., & Liu, K. (2012). OSA: A fast and accurate alignment tool for RNA-Seq. *Bioinformatics, 28*(14), 1933-1934.

Huang, d. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57.

Iqbal, K., Liu, F., & Gong, C. X. (2016). Tau and neurodegenerative disease: The story so far. *Nature Reviews Neurology, 12*(1), 15-27.

Iqbal, K., Liu, F., Gong, C. X., & Grundke-Iqbal, I. (2010). Tau in Alzheimer disease and related tauopathies. *Current Alzheimer Research, 7*(8), 656-664.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4*(2), 249-264.

Jacunski, A., & Tatonetti, N. P. (2013). Connecting the dots: Applications of network medicine in pharmacology and disease. *Clinical Pharmacology & Therapeutics, 94*(6), 659-669.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., & Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE, 5*(9), e12336.

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics, 7*, 359.

Jiang, P., & Liu, X. S. (2015). Big data mining yields novel insights on cancer. *Nature Genetics, 47*(2), 103-104.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127.

Kavanagh, T., Mills, J. D., Kim, W. S., Halliday, G. M., & Janitz, M. (2013). Pathway analysis of the human brain transcriptome in disease. *Journal of Molecular Neuroscience, 51*(1), 28-36.

Khang, T. F., & Lau, C. Y. (2015). Getting the most out of RNA-seq data analysis. *PeerJ, 3*, e1360.

University of Malaya

Kittur, S. D., Hoh, J. H., Kawas, C. H., Hayward, G. S., Endo, H., & Adler, W. H. (1992). A molecular hybridization study for the presence of Herpes simplex, cytomegalovirus and Epstein-Barr virus in brain and blood of Alzheimer's disease patients. *Archives of Gerontology and Geriatrics, 15*(1), 35-41.

Korostynski, M., Kaminska-Chowaniec, D., Piechota, M., & Przewlocki, R. (2006). Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. *BMC Genomics, 7*, 146.

Korostynski, M., Piechota, M., Kaminska, D., Solecki, W., & Przewlocki, R. (2007). Morphine effects on striatal transcriptome in mice. *Genome Biology, 8*(6), R128.

Langmead, B., Hansen, K. D., & Leek, J. T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology, 11*(8), R83.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*(2), R29.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., . . . Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: A survey. *Briefings in Bioinformatics, 14*(4), 469-490.

Levine, A. J., Momand, J., & Finlay, C. A. (1991). The p53 tumour suppressor gene. *Nature, 351*(6326), 453-456.

Li, W. (2012). Volcano plots in analyzing differential expressions with mRNA microarrays. *Journal of Bioinformatics and Computational Biology, 10*(6), 1231003.

Li, Y., Sun, H., Chen, Z., Xu, H., Bu, G., & Zheng, H. (2016). Implications of GABAergic neurotransmission in Alzheimer's disease. *Frontiers in Aging Neuroscience, 8*, 31.

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Ramsey, K., . . . Stephan, D. A. (2008). Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: A reference data set. *Physiological Genomics, 33*(2), 240-256.

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Ramsey, K., . . . Stephan, D. A. (2010). Neuronal gene expression in non-demented individuals with intermediate Alzheimer's disease neuropathology. *Neurobiology of Aging, 31*(4), 549-566.

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., . . . Stephan, D. A. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiological Genomics, 28*(3), 311-322.

University of Malaya

Liang, W. S., Reiman, E. M., Valla, J., Dunckley, T., Beach, T. G., Grover, A., . . . Stephan, D. A. (2008). Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences of the United States of America, 105*(11), 4441-4446.

Licastro, F., & Porcellini, E. (2016). Persistent infections, immune-senescence and Alzheimer's disease. *Oncoscience, 3*(5-6), 135-142.

Lim, K., Li, Z., Choi, K. P., & Wong, L. (2015). A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. *Journal of Bioinformatics and Computational Biology, 13*(4), 1550018.

Limon, A., Reyes-Ruiz, J. M., & Miledi, R. (2012). Loss of functional GABA(A) receptors in the Alzheimer diseased brain. *Proceedings of the National Academy of Sciences of the United States of America, 109*(25), 10071-10076.

Logan, C. Y., & Nusse, R. (2004). The Wnt signaling pathway in development and disease. *Annual Review of Cell and Developmental Biology, 20*, 781-810.

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

Lovell, M. A., Robertson, J. D., Teesdale, W. J., Campbell, J. L., & Markesbery, W. R. (1998). Copper, iron and zinc in Alzheimer's disease senile plaques. *Journal of the Neurological Sciences, 158*(1), 47-52.

Masters, C. L., Multhaup, G., Simms, G., Pottgiesser, J., Martins, R. N., & Beyreuther, K. (1985). Neuronal origin of a cerebral amyloid: Neurofibrillary tangles of Alzheimer's disease contain the same protein as the amyloid of plaque cores and blood vessels. *The EMBO Journal, 4*(11), 2757-2763.

Masters, C. L., Simms, G., Weinman, N. A., Multhaup, G., McDonald, B. L., & Beyreuther, K. (1985). Amyloid plaque core protein in Alzheimer disease and Down syndrome. *Proceedings of the National Academy of Sciences of the United States of America, 82*(12), 4245-4249.

Mawanda, F., & Wallace, R. (2013). Can infections cause Alzheimer's disease? *Epidemiologic Reviews, 35*, 161-180.

McCall, M. N., & Almudevar, A. (2012). Affymetrix GeneChip microarray preprocessing for multivariate analyses. *Briefings in Bioinformatics, 13*(5), 536-546.

McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics, 11*(2), 242-253.

McIlwain, D. R., Berger, T., & Mak, T. W. (2013). Caspase functions in cell death and disease. *Cold Spring Harbor Perspectives in Biology, 5*, a008656.

University of Malaya

Merkwirth, C., Martinelli, P., Korwitz, A., Morbin, M., Brönneke, H. S., Jordan, S. D., . . . Langer, T. (2012). Loss of prohibitin membrane scaffolds impairs mitochondrial architecture and leads to tau hyperphosphorylation and neurodegeneration. *PLoS Genetics, 8*(11), e1003021.

Miller, J. A., Cai, C., Langfelder, P., Geschwind, D. H., Kurian, S. M., Salomon, D. R., & Horvath, S. (2011). Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics, 12*, 322.

Miller, J. A., Oldham, M. C., & Geschwind, D. H. (2008). A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *Journal of Neuroscience, 28*(6), 1410-1420.

Mills, J. D., Nalpathamkalam, T., Jacobs, H. I., Janitz, C., Merico, D., Hu, P., & Janitz, M. (2013). RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. *Neuroscience Letters, 536*, 90-95.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., . . . Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics, 34*(3), 267-273.

Morfini, G., Szebenyi, G., Elluru, R., Ratner, N., & Brady, S. T. (2002). Glycogen synthase kinase 3 phosphorylates kinesin light chains and negatively regulates kinesin-based motility. *The EMBO Journal, 21*(3), 281-293.

Murphy, M. P., & LeVine, H. (2010). Alzheimer's disease and the amyloid-beta peptide. *Journal of Alzheimer's Disease, 19*(1), 311-323.

Nunez-Iglesias, J., Liu, C. C., Morgan, T. E., Finch, C. E., & Zhou, X. J. (2010). Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PLoS ONE, 5*(2), e8898.

Oddo, S. (2008). The ubiquitin-proteasome system in Alzheimer's disease. *Journal of Cellular and Molecular Medicine, 12*(2), 363-373.

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences of the United States of America, 103*(47), 17973-17978.

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology, 11*(12), 220.

Pigino, G., Morfini, G., Atagi, Y., Deshpande, A., Yu, C., Jungbauer, L., . . . Brady, S. (2009). Disruption of fast axonal transport is a pathogenic mechanism for intraneuronal amyloid beta. *Proceedings of the National Academy of Sciences of the United States of America, 106*(14), 5907-5912.

University of Malaya

Pigino, G., Morfini, G., Pelsman, A., Mattson, M. P., Brady, S. T., & Busciglio, J. (2003). Alzheimer's presenilin 1 mutations impair kinesin-based axonal transport. *Journal of Neuroscience, 23*(11), 4499-4508.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia, 9*(1), 63-75.e62.

Rajkumar, A. P., Qvist, P., Lazarus, R., Lescai, F., Ju, J., Nyegaard, M., . . . Christensen, J. H. (2015). Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genomics, 16*, 548.

Rickle, A., Bogdanovic, N., Volkman, I., Winblad, B., Ravid, R., & Cowburn, R. F. (2004). Akt activity in Alzheimer's disease and other neurodegenerative disorders. *NeuroReport, 15*(6), 955-959.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology, 11*(3), R25.

Ross, C. A., & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine, 10*, S10-S17.

Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics, 14*(2), 89-99.

Satoh, J., Yamamoto, Y., Asahina, N., Kitano, S., & Kino, Y. (2014). RNA-Seq data mining: Downregulation of NeuroD6 serves as a possible biomarker for Alzheimer's disease brains. *Disease Markers, 2014*, 123165.

Schlatterer, S. D., Acker, C. M., & Davies, P. (2011). c-Abl in neurodegenerative disease. *Journal of Molecular Neuroscience, 45*(3), 445-452.

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics, 16*(1), 59-70.

Sheng, M., Sabatini, B. L., & Südhof, T. C. (2012). Synapses and Alzheimer's disease. *Cold Spring Harbor Perspectives in Biology, 4*(5), a005777.

Simpson, J. E., Ince, P. G., Shaw, P. J., Heath, P. R., Raman, R., Garwood, C. J., . . . Group, M. C. F. a. A. N. S. (2011). Microarray analysis of the astrocyte transcriptome in the aging brain: Relationship to Alzheimer's pathology and APOE genotype. *Neurobiology of Aging, 32*(10), 1795-1807.

University of Malaya

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3,* Article3.

Song, B., Davis, K., Liu, X. S., Lee, H. G., Smith, M., & Liu, X. (2011). Inhibition of Polo-like kinase 1 reduces beta-amyloid-induced neuronal cell death in Alzheimer's disease. *Aging (Albany NY), 3*(9), 846-851.

Squitti, R., & Polimanti, R. (2013). Copper phenotype in Alzheimer's disease: Dissecting the pathway. *American Journal of Neurodegenerative Disease, 2*(2), 46-56.

Stotani, S., Giordanetto, F., & Medda, F. (2016). DYRK1A inhibition as potential treatment for Alzheimer's disease. *Future Medicinal Chemistry, 8*(6), 681-696.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550.

Suh, Y. H., & Checler, F. (2002). Amyloid precursor protein, presenilins, and alpha-synuclein: Molecular pathogenesis and pharmacological applications in Alzheimer's disease. *Pharmacological Reviews, 54*(3), 469-525.

Söllner, T., Whiteheart, S. W., Brunner, M., Erdjument-Bromage, H., Geromanos, S., Tempst, P., & Rothman, J. E. (1993). SNAP receptors implicated in vesicle targeting and fusion. *Nature, 362*(6418), 318-324.

Tamada, H., Van Thuan, N., Reed, P., Nelson, D., Katoku-Kikyo, N., Wudel, J., . . . Kikyo, N. (2006). Chromatin decondensation and nuclear reprogramming by nucleoplasmin. *Molecular and Cellular Biology, 26*(4), 1259-1271.

Taminau, J., Lazar, C., Meganck, S., & Nowé, A. (2014). Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinformatics, 2014,* 345106.

Taminau, J., Meganck, S., Lazar, C., Steenhoff, D., Coletta, A., Molter, C., . . . Nowé, A. (2012). Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC Bioinformatics, 13,* 335.

Tan, M. G., Chua, W. T., Esiri, M. M., Smith, A. D., Vinters, H. V., & Lai, M. K. (2010). Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *Journal of Neuroscience Research, 88*(6), 1157-1169.

Terry, R. D. (1994). Neuropathological changes in Alzheimer disease. *Progress in Brain Research, 101,* 383-390.

Terry, R. D., Masliah, E., Salmon, D. P., Butters, N., DeTeresa, R., Hill, R., . . . Katzman, R. (1991). Physical basis of cognitive alterations in Alzheimer's disease: Synapse loss is the major correlate of cognitive impairment. *Annals of Neurology, 30*(4), 572-580.

University of Malaya

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America, 98*(9), 5116-5121.

Twine, N. A., Janitz, K., Wilkins, M. R., & Janitz, M. (2011). Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS ONE, 6*(1), e16266.

Vilchez, D., Saez, I., & Dillin, A. (2014). The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nature Communications, 5*, 5659.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics, 13*(4), 227-232.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57-63.

Wegiel, J., Gong, C. X., & Hwang, Y. W. (2011). The role of DYRK1A in neurodegenerative diseases. *The FEBS Journal, 278*(2), 236-245.

Williams, C., Mehrian Shai, R., Wu, Y., Hsu, Y. H., Sitzer, T., Spann, B., . . . Miller, C. A. (2009). Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease. *PLoS ONE, 4*(3), e4936.

Xiao, Y., Hsiao, T. H., Suresh, U., Chen, H. I., Wu, X., Wolf, S. E., & Chen, Y. (2014). A novel significance score for gene selection and ranking. *Bioinformatics, 30*(6), 801-807.

Yamakawa, H., Cheng, J., Penney, J., Gao, F., Rueda, R., Wang, J., . . . Tsai, L. H. (2017). The transcription factor Sp3 cooperates with HDAC2 to regulate synaptic function and plasticity in neurons. *Cell Reports, 20*(6), 1319-1334.

Yamamoto, S., Yamashita, A., Arakaki, N., Nemoto, H., & Yamazaki, T. (2014). Prevention of aberrant protein aggregation by anchoring the molecular chaperone αB-crystallin to the endoplasmic reticulum. *Biochemical and Biophysical Research Communications, 455*(3-4), 241-245.

Yu, X., Caltagarone, J., Smith, M. A., & Bowser, R. (2005). DNA damage induces cdk2 protein levels and histone H2B phosphorylation in SH-SY5Y neuroblastoma cells. *Journal of Alzheimer's Disease, 8*(1), 7-21.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology, 4*, Article17.

University of Malaya

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., . . . Zhao, Q. Y. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE, 9*(8), e103207.

Zhao, L., Ma, Q. L., Calon, F., Harris-White, M. E., Yang, F., Lim, G. P., . . . Cole, G. M. (2006). Role of p21-activated kinase pathway defects in the cognitive deficits of Alzheimer disease. *Nature Neuroscience, 9*(2), 234-242.

Zheng, W., Chung, L. M., & Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics, 12*, 290.

Zivadinov, R., Zorzon, M., Weinstock-Guttman, B., Serafin, M., Bosco, A., Bratina, A., . . . Ramanathan, M. (2009). Epstein-Barr virus is associated with grey matter atrophy in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry, 80*(6), 620-625.

University of Malaya

LIST OF PUBLICATIONS AND PAPERS PRESENTED

1. Khang, T. F., & **Lau, C. Y.** (2015). Getting the most out of RNA-seq data analysis. *PeerJ*, *3*, e1360.

University of Malaya