UNIVERSITY OF CALIFORNIA SAN DIEGO

Transcriptomic Profiling of Sporadic Alzheimer's Disease Patients

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Bioengineering

by

Balaji Ganesh Anantharaman

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Marcos Intaglietta
Professor Steven Wagner

2020

The thesis of Balaji Ganesh Anantharaman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To

My Family

and

Friends

EPIGRAPH

What is rational is actual and what is actual is rational.

G.W.F Hegel

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to thank Dr Shankar Subramaniam for his support as the chair of my committee. I greatly appreciate him helping me put together different parts of my research into a cohesive thesis. I'd also like to thank him for giving me the opportunity and access to resources to conduct research in his lab.

I'd also like to thank Dr Andrew Caldwell for his unrelenting support from day one. His mentorship has been invaluable and has helped me mature as a Bioinformatics and Systems Biology researcher over the past year and a half.

I'd also like to thank Dr Srinivasan Ramachandran, Milenka Mitic, and other people from the Subramaniam Lab for their help.

I'm also grateful for the support from my family back in India. I'd also like to thank my friends from Graduate School - Aashish, Raghav, Manas, Raja, Krithika and many others for their constant support. Special thanks to my roommates, Swetha and Shradha, for putting up with me.

Chapter 2 is coauthored with Subramaniam, Shankar; Caldwell, Andrew; Wagner, Steven; and Anantharaman, Balaji Ganesh. The thesis author was the primary author of this chapter.

ABSTRACT OF THE THESIS

Transcriptomic Profiling of Sporadic Alzheimer's Disease Patients

by

Balaji Ganesh Anantharaman

Master of Science in Bioengineering

University of California San Diego, 2020

Professor Shankar Subramaniam, Chair

Alzheimer's Disease (AD) is a neurodegenerative disorder characterized physically by dementia and physiologically by senile plaques and neurofibrillary tangles in the brain. Mutations to the genes PSEN1, PSEN2 and APP result in the manifestation of the dominantly inherited form of AD, Familial AD. Though a number of risk factors, including genetic mutations, environmental factors, and aging, have been attributed to the sporadic form of AD, the underlying mechanistic basis of the disease is yet to be unearthed. Analysing sporadic AD RNA-seq samples together with non-demented controls allows us to uncover these molecular mechanisms and to this end, we have analysed a 50-sample RNA-Seq dataset, with 40 AD samples and 10 controls, and identified disease-associated endotypes that arise from gene expression changes between the AD cases and

the controls. We have also described an adapted framework for analysing low-quality RNA-seq samples (RIN $> 1, < 3$), and applying this framework to our data results in the categorization of the samples into two groups which show different degrees of differential expression with respect to the controls. Endotypes such as Dedifferentiation and Synaptic Signalling are preferentially enriched in samples from one group, although a small subset of samples from this group exhibits a significantly higher enrichment of said endotypes compared to other group members. We hypothesize that these differences in endotype signatures manifest due to varying severity among the samples, and scrutinizing the similarities and dissimilarities among the groups can provide insights into the etiology of Sporadic AD.

Chapter 1

Introduction

Neurodegenerative disease (ND) is a blanket term used to describe a group of neurological disorders with diverse clinical and pathological implications that result in a steady loss of functioning neurons [59]. Alzheimer's Disease (AD) is a progressive ND characterized by gradual memory impairment and loss of cognitive abilities, especially those related to learning, behaviour, speech, visuospatial cognition, and the motor system. It is also the most common form of dementia and was first established as a neuropathological phenotype by the Bavarian psychiatrist, Alois Alzheimer, in 1906. [21].

1.1 Aging

Susceptibility to AD is unequivocally linked to aging, and the percentage of people with AD increases with age - 3% of people age 65-74, 17% of people age 75-84, and 32% of people age 85 or older have AD [38]. An estimated 5.8 million Americans aged 65 or older suffer from AD in 2020, and as the population of Americans aged 65 or older continues to increase, so will the number of individuals suffering from AD [2]. These numbers, estimated on the basis of symptoms such as memory loss and cognitive impairment, could be significantly different if a biomarker-based method is to calculate the prevalence [43].

1.2 Classification

AD cases are classified on the basis of family history [32] and age of onset [9] for clinical purposes. Table 1 [32] summarizes this classification.

Table 1.1: AD classification based on 1) ancestry and 2) aging

Factor	Categories	Frequency	Description
Family History	Autosomal Dominant	< 5%	Occurs in more than **3** individuals in at least **2** generations, with two of the individuals being first-degree relatives of the third.
	Familial	~ 15 - 25%	Occurs in more than **2** individuals and at least **2** of the affected individuals are third-degree relatives or closer.
	Sporadic	~ 75%	Occurs in more than **2** individuals and at least **2** of the affected individuals are third-degree relatives or closer.
Age of Onset	Early Onset	~ 5%	Onset **before** the age of **65**
	Late Onset	~ 95%	Onset **on or after** the age of **65**

1.3 Pathology

AD is characterized by the abnormal accumulation of two aggregates [57] - Aβ, an extracellular deposit, and Neurofibrillary Tangles (NFT), which are intracellular deposits. A series of proteolysis events, catalysed by a family of proteases called secretases, acting upon an intramembrane protein, APP, results in the production of Aβ fragments. Intracellular NFT aggregates are formed when Tau, a protein that usually binds to microtubules and localizes to axons, is phosphorylated at several sites. This causes the protein to release from the microtubules and start binding to each other, forming axonal NFTs [36].

Familial AD is an autosomal dominant disease characterized by mutations in genes encoding for three proteins - APP, PSEN1,and PSEN2. The last two are homologous proteins that can mimic the catalytic activity of γ-secretase. The formation of Aβ fragments from APP is initiated by the cleavage of APP's extracellular N-terminus by secretase [BACE] [17], followed by γ-secretase/PSEN1/PSEN2 cleaving APP , twice, inside the cell, resulting in the formation of an intracellular residue and extracellular Aβ. Two major forms of Aβ can be produced, Aβ40 and Aβ42 [31] based on the cleavage site of γ-secretase/PSEN1/PSEN2. All three mutations result in a higher Aβ42/Aβ40 ratio [49][45]. The DIAN study [64] confirmed that the mutation type along with the associated Aβ42/Aβ40 ratio can predict the mean age of onset of dementia. These Aβ42 fragments aggregate to form highly insoluble fibrils that eventually deposit as plaques [51]. However, there is evidence that soluble Aβ oligomers are also neurotoxic as they damage nearby neurons and stimulate the formation of NFTs [29]. Called the amyloid cascade, this theory ties together the two major AD pathologies and is backed up by experimental evidence [37].

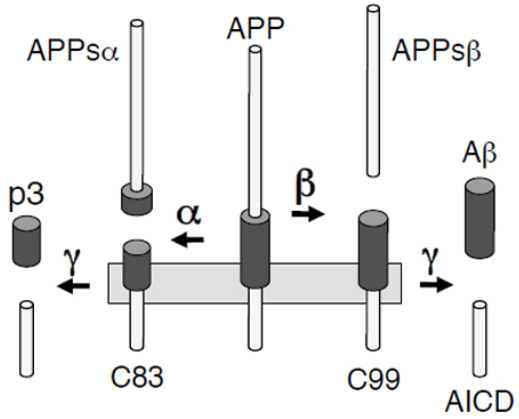

Figure 1.1: Two APP processing pathways [23] - Sequential cleaving of APP by β-secretase followed by γ-secretase results in amyloid formation; Sequential cleaving by α-secretase followed by γ-secretase results in the formation of P3 peptide, a hydrophobic protein that doesn't form oligomers [22].

1.4 APOE

The Apolipoprotein (APOE) gene on chromosome 19 is a strong risk factor for developing AD. The main carrier of cholesterol in the CNS, APOE has been found to play an important role in Aβ metabolism, aggregation and deposition [40]. Humans possess three common APOE alleles: APOE2, APOE3, and APOE4 [56]. Population studies have found that APOE4 increases the risk of developing AD (odds ratio ranging from 3 to 10) [24] and is also associated with an earlier onset of the disease. APOE2, on the other hand, decreases the risk of getting AD [18]. Increased plaque deposition has been observed in patients with the APOE4 [40] allele and it has been shown that the variant protein's inefficiency in Aβ clearance is the reason behind this phenomenon [72]. There is no conclusive evidence for the involvement of APOE4 in tau phosphorylation.

1.5 Amyloid Cascade and Sporadic AD

The amyloid cascade hypothesis is largely based on data from familial AD patients, and hence, it's relevance for patients with sporadic AD is questionable [16]. The accumulation of Aβ has been suggested to be a host response to an underlying neurological condition or a stressor such as a brain injury, and hence might even be protective in nature [41]. Studies, conducted in cognitively healthy people, analysing the relationship between Aβ accumulation and atrophy in AD-related brain regions have been largely incongruous in their findings [43], [54]. However, biomarker studies show an increased risk for cognitive impairment in healthy elderly people with signs of cerebral amyloidosis [34]. Moreover, the APOE4 allele, a familial AD risk factor, has been confirmed to be a major risk factor for sporadic AD as well [18].

Genome-wide association studies have identified genes, that contribute to phenotypic pathways such as innate immunity and cholesterol metabolism, as minor risk factors for sporadic AD [46]. Non-genetic and modifiable factors such as alcohol intake and education have also been

considered as AD risk factors [28] [60]. Clinical and pathological heterogeneity among patients with sporadic AD makes it tougher to pinpoint associated risk factors.

1.6 Transcriptomic Profiling

A tissue's transcriptome can give an accurate snapshot of its cellular activity at a given point in time [69]. The process of finding an association between a genetic variant and AD remains challenging, and alternatively, integrating gene expression analysis into this process can help determine the effect of these risk factors at the transcriptomic level in a specific tissue or a specific cell-type or at a particular point in time [70]. Moreover, it is this paradigm shift in the gene regulation and expression patterns, caused by underlying genetic risks, that results in the expression of various disease states [19]. Transcriptomic profiling is commonly performed using either Microarray Hybridization or Next Generation RNA Sequencing, and both methods allow for the investigation of changes in gene expression and mRNA splicing patterns between different conditions [70]. Profiling followed by systems level analysis of these changes can help identify misfiring regulatory mechanisms that engender disease states.

Microarray Hybridization involves the binding of cDNA libraries, reverse transcribed from RNA samples, to a probe DNA that is immobilized to a solid surface. Despite this technology improving leaps and bounds in recent years, it still possesses a number of limitations – microarrays cannot identify novel transcripts as probe design is based on known genome sequences [65]; their reliance on non-specific Hybridization reactions can make quantification results unreliable; due to differences in probe design and number across different platforms, inter- and intra-platform consistency is uncertain. RNA-Seq is a NGS assay that starts off with the reverse-transcription of RNAs to give cDNAs. Adapters are subsequently ligated to these strands which are then sequenced, unidirectionally or bidirectionally, in a massively parallel fashion, in flow cells on the sequencing instrument. Enrichment of a specific subset of RNA, such as mRNA or miRNA, is

done before sequencing. This usually involves the removal of highly abundant ribosomal RNAs either by isolation and degradation using heat and/or chemicals [39], or in the case of mRNA sequencing, preferential isolation of mRNAs using their polyA tails. Compared to microarray-based techniques, RNA-Seq allows for the discovery of new variants - reads from a unique mRNA will map to a different part of the reference genome. Moreover, RNA-Seq has a wider dynamic range as compared to the microarray technology, thereby allowing for the detection of more differentially expressed genes between two conditions [73].

For the study of neurodegenerative diseases, the main source of RNA is post-mortem brain tissues, though this tissue is difficult to obtain and the susceptibility of post-mortem RNA to degrade is high [53] [8]. In this study, we've compared the gene expression data of 40 AD samples – 19 early onset and 21 late onset, against 10 Non-Demented Control (NDC) samples. The patients were chosen using a rigorous vetting criteria so as to maintain clinical uniformity, and measures were taken to account for RNA-degradation while processing the expression data.

Chapter 2

Data Collection

Bulk RNA sequencing was performed on RNA extracted from patient brain samples frozen and preserved at Alzheimer's Disease Research Center (ADRC), UCSD. A total of 50 samples were sequenced - 10 controls and 40 AD cases. All samples had a RIN score between 1 and 3. All patients were followed clinically, and the 40 samples were split into two groups based on the age of onset of AD - 19 Early Onset AD samples (Table 2.2) , with an age of onset $<$ 60 years, and 21 Late-Onset AD samples (Table 2.3), with an age of onset between 70 and 80. Three scores - BIMC (Blessed Memory-Information-Concentration) [12], MMSE (Mini-Mental State Examination) [26] and Mattis' DRS (Dementia Rating Scale) [52] were used to classify the selected patients as AD/ NDC cases. All controls had a BIMC score (on a scale of 0 - 35; Higher score == more extreme dementia) <= 4, MMSE (on a scale of 0 - 30; Lower score == more extreme dementia) score between 26 and 30, and an aggregate DRS (Maximum of 144 points - Split among Initiation/Preservation(37), Attention(37), Construction(6), Conceptualization(39) and Memory(25); Lower score == more extreme dementia) score between 127 and 140. Each brain sample was also staged on the basis of the concentration of Neurofibrillary Tangles(NFTs) in different brain regions, using a modified version of the staging scheme introduced by Braak and Braak. In this scheme, the concentration of NFTs in the hippocampal and entorhinal cortex (EC),

two regions which are believed to be the starting point for tau pathology [20], is used as a proxy to estimate the progression of the disease. Braak stages I and II reflect mild NFT concentration in EC and hippocampal regions; Stages III and IV reflect the moderate concentrations of NFT in EC and hippocampal regions; Stages V and VI are characterized by the EC and hippocampal regions being severely entrenched by the tau pathology, and its spread to the isocortical regions. All AD samples were at BRAAK stage 6 while the controls were at BRAAK stage 1 or BRAAK stage 2. The APOE status of the AD samples was also determined - all samples either had the APOE3/3 or the APOE3/4 genotype.

2.1 Metadata

The following metadata were also collected for each AD sample - sex, age (at death), age of onset, and the concentration of neuritic plaques and tangles in Mid Frontal Cortex (MF), Inferior Parietal Cortex (IP), Superior Temporal Cortex (ST) and the hippocampus. Since AD was ascertained to be the cause of death of all patients from this study, Disease Specific Survival(DSS) time was estimated by subtracting the age of onset from age at death.

2.2 Reads to Counts

Each sample is split among 4 lanes in the sequencer. The lane fastq files were first combined together using the 'cat' command line function. Trim Galore v0.6.5, a wrapper script for the command-line tools Cutadapt [1] v2.9 and FastQC, was used to trim adapter sequences, isolate all pair-end reads with a Phred score (Q) of 20 or more, and subsequently, estimate a few metrics for quality control. The trimmed reads were then mapped to the GRCh38.p12 human transcriptome using Kallisto [13] v 0.46.1 run with the following options -bias –rf-stranded. MultiQC was used to collate and summarize quality metrics from the previous steps.

Table 2.1: Metadata for controls - samples with less than 25% mapped reads are highlighted in green.

Metdata Table - Control samples

			Time (in years)	
SampleID	Sex	BRAAK1	Age at Death	Percentage of Mapped Reads
COL				
COL1	2	1.0	63	42.5
COL2	1	0.0	83	37.5
COL3	1	0.0	76	39.2
COL4	1	3.0	91	15.7
COL5	2	1.0	102	15.2
COL6	2	1.0	97	41.6
COL7	1	1.1	87	37.4
COL8	2	1.0	93	49.4
COL9	2	1.0	80	33.9
COL10	1	2.0	94	39.3

Sex - 1 = Male, 2 = Female

2.3 Counts to Differentially Expressed Genes (DEGs)

The R package tximport [66] v1.16.1 was used to summarize the Kallisto transcript abundance counts to the gene level and import the resulting count data into the R programming environment. A DGEList object was then created from the read counts using the DGEList() function from edgeR [62] v3.30.3, and the sample phenotype data. Only genes which had 10 counts or more in at least 5 samples were considered for further analysis. The data was

Table 2.2: Metadata for Early Onset Cases - samples with less than 25% mapped reads are highlighted in green; samples with missing age on onset/age at death are highlighted in yellow.

Metdata Table - Early Onset AD samples

				Time (in years)											
SampleID	Sex	APOE	BRAAK1	Age of Onset	Age at Death	DSS	MF Tangles	IP Tangles	ST Tanlges	HP Tangles	MF Plaques	IP Plaques	ST Plaques	HP Plaques	Percentage of Mapped Reads
EOL															
EOL1	1	34	6	60	67	7	7	6	12	10	50	50	50	25	40.6
EOL2	1	33	6	58	69	11	8	6	11	24	50	50	50	25	22.6
EOL3	1	34	6	52	65	13	11	8	10	41	50	50	50	24	35.3
EOL4	2	33	6	60	74	14	3	11	12	34	50	50	50	33	37.8
EOL5	1	33	6	58	68	10	4	NA	5	35	50	NA	50	24	31.0
EOL6	2	33	6	57	69	12	4	6	6	3	50	50	27	6	11.5
EOL7	1	33	6	40	46	6	6	9	3	3	50	50	50	50	18.1
EOL8	2	33	6	35	46	11	11	11	12	26	50	39	26	47	24.8
EOL9	2	34	6	51	60	9	16	15	13	26	50	50	50	17	39.0
EOL10	1	34	6	53	60	7	7	11	4	9	50	50	43	15	39.0
EOL11	1	33	6	44	52	8	12	15	21	42	50	50	27	18	40.4
EOL12	1	33	6	35	41	6	19	19	7	22	43	50	44	46	40.8
EOL13	1	34	6	57	72	15	13	9	17	39	50	50	50	24	38.5
EOL14	2	34	6	49	61	12	6	6	10	5	50	45	27	20	33.8
EOL15	1	33	6	37	46	9	23	12	19	53	50	50	50	50	26.1
EOL16	2	33	6	57	71	14	3	3	2	18	49	37	35	29	37.5
EOL17	1	33	6	54	65	11	8	10	12	8	50	50	27	10	34.2
EOL18	1	34	6	58	69	11	24	19	21	20	50	50	50	22	40.5
EOL19	2	33	6	58	68	10	8	NA	10	22	50	NA	50	33	39.5

Sex - 1 = Male, 2 = Female; DSS - Disease Specific Survival, MF - Mid Frontal Cortex, IP - inferiro Pareital Cortex, ST - Superior Temporal Cortex, HP - Hippocampus

then normalized using the TMM (Trimmed Mean of M-values) normalization method. The Voom function from Limma [61] v3.44.1 was used to model the mean-variance trend and get gene-specific weights, which were subsequently used to fit a linear model to the count data.

Table 2.3: Metadata for Late Onset Cases - samples with less than 25% mapped reads are highlighted in green; samples with missing age of onset/age at death are highlighted in yellow.

Metdata Table - Late Onset AD samples

				Time (in years)											
SampleID	Sex	APOE	BRAAK1	Age of Onset	Age at Death	DSS	MF Tangles	IP Tangles	ST Tanlges	HP Tangles	MF Plaques	IP Plaques	ST Plaques	HP Plaques	Percentage of Mapped Reads
LOL															
LOL1	2	34	6	71	82	11	7	8	NA	66	50	50	50	15	31.80
LOL2	1	33	6	75	86	11	5	4	8	26	50	50	43	9	30.00
LOL3	2	34	6	73	NA	NA	6	6	10	17	50	50	50	45	37.30
LOL4	1	34	6	72	NA	NA	7	4	4	18	50	50	50	16	37.60
LOL5	2	34	6	72	78	6	5	6	5	34	50	50	50	26	39.10
LOL6	1	34	6	75	79	4	3	6	6	24	50	50	47	17	40.60
LOL7	1	34	6	72	83	11	6	7	16	29	38	50	33	9	39.80
LOL8	1	34	6	73	83	10	3	9	6	20	35	35	47	17	50.17
LOL9	2	33	6	78	85	7	4	7	11	6	40	32	50	15	48.60
LOL10	1	34	6	72	83	11	3	11	6	26	50	50	50	32	32.60
LOL11	2	33	6	76	83	7	4	4	6	17	24	21	35	8	39.50
LOL12	1	33	6	71	85	14	8	12	20	34	50	50	50	17	30.70
LOL13	2	34	6	74	85	11	6	5	5	6	50	50	48	12	27.50
LOL14	2	33	6	73	81	8	4	10	10	8	41	50	46	10	44.30
LOL15	2	34	6	73	80	7	5	8	8	17	50	50	34	16	32.50
LOL16	1	33	6	72	79	7	5	4	4	18	48	43	35	13	33.30
LOL17	1	33	6	76	85	9	6	3	3	17	46	47	35	21	34.00
LOL19	1	33	6	72	84	12	4	5	NA	26	41	43	NA	9	37.00
LOL20	1	34	6	76	87	11	2	3	5	24	50	50	31	10	40.30
LOL21	2	33	6	78	83	5	2	6	9	8	50	50	43	25	37.30

Sex - 1 = Male, 2 = Female; DSS - Disease Specific Survival, MF - Mid Frontal Cortex, IP - inferiro Pareital Cortex, ST - Superior Temporal Cortex, HP - Hippocampus

A contrast matrix was used to compare gene expression between the AD subtypes and NDC, and empirical Bayesian statistics for the differential expression analysis was estimated using the eBayes function from Limma. Genes with an FDR-adjusted p-value of less than 0.05 were

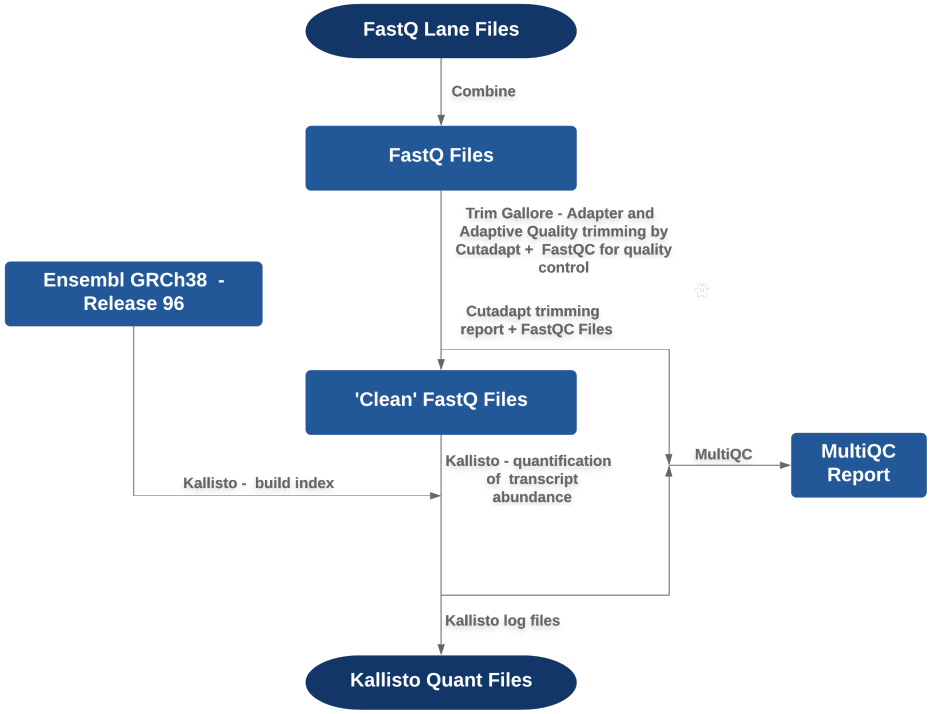

Figure 2.1: RNA-Seq analysis pipeline flowchart - command line tools.

deemed as being differentially expressed between the conditions. Fig 2.1 and 2.2 summarize the pipeline that was adopted to analyze the data.

2.4 Results

The two controls with low-mapping percentages were removed from the analysis. The early onset samples have 2290 upregulated DEGs (logFC > 0, qval < 0.05) and 2165 downregulated DEGs (logFC < 0, qval < 0.05), while the late onset samples have 7 upregulated DEGs (logFC > 0, qval < 0.05) and 13 downregulated DEGs (logFC < 0, qval < 0.05) (Fig. 2.3, 2.4).

The low number of differentially expressed genes in the case of the Late Onset samples seemed unusual. We theorized that inherent heterogeneity in the samples was confounding any differential expression compared to the controls. We decided to explore the data further to figure out if this was indeed the case and if yes, identify the cause(s) of this heterogeneity.

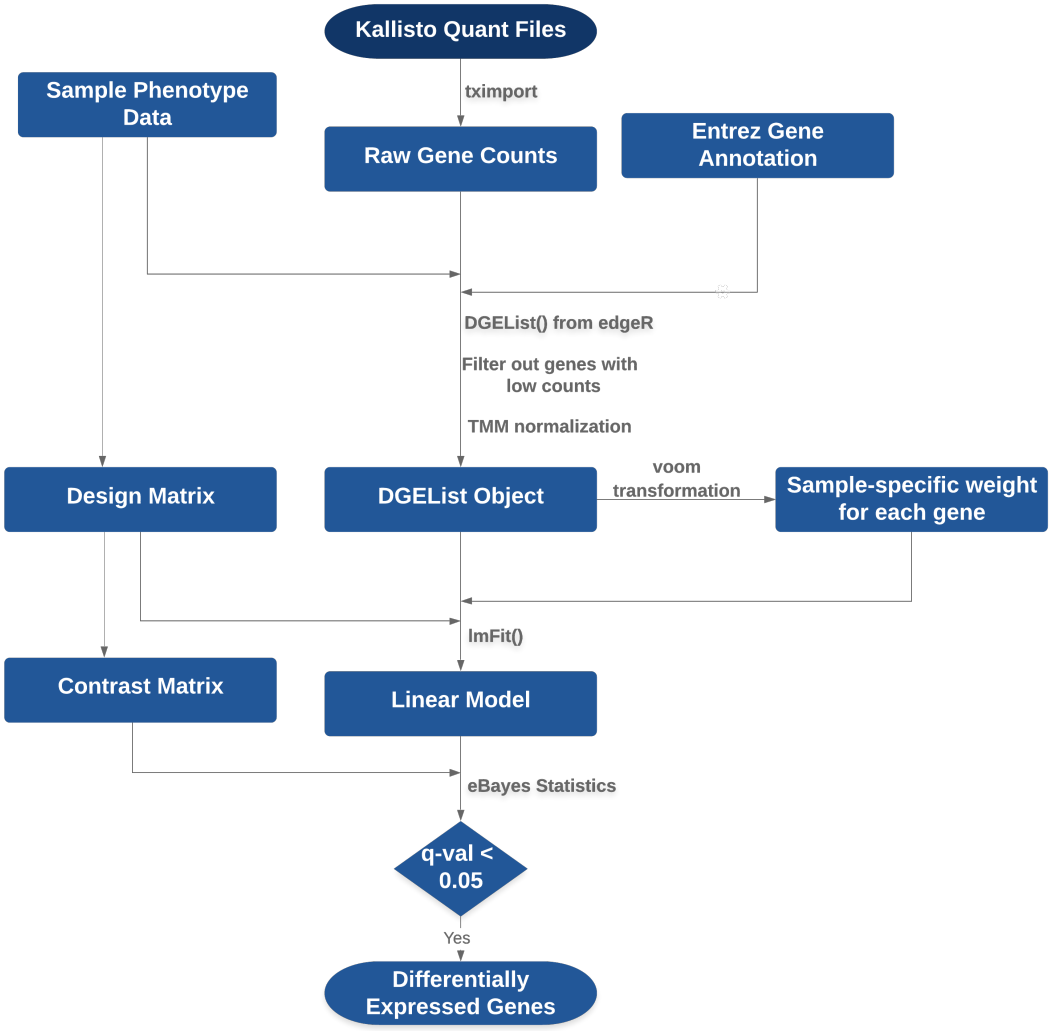

Figure 2.2: RNA-Seq analysis pipeline - R and Bioconductor packages.

2.5 Acknowledgement

Chapter 2 is coauthored with Subramaniam, Shankar; Caldwell, Andrew; Wagner, Steven; and Anantharaman, Balaji Ganesh. The thesis author was the primary author of this chapter.

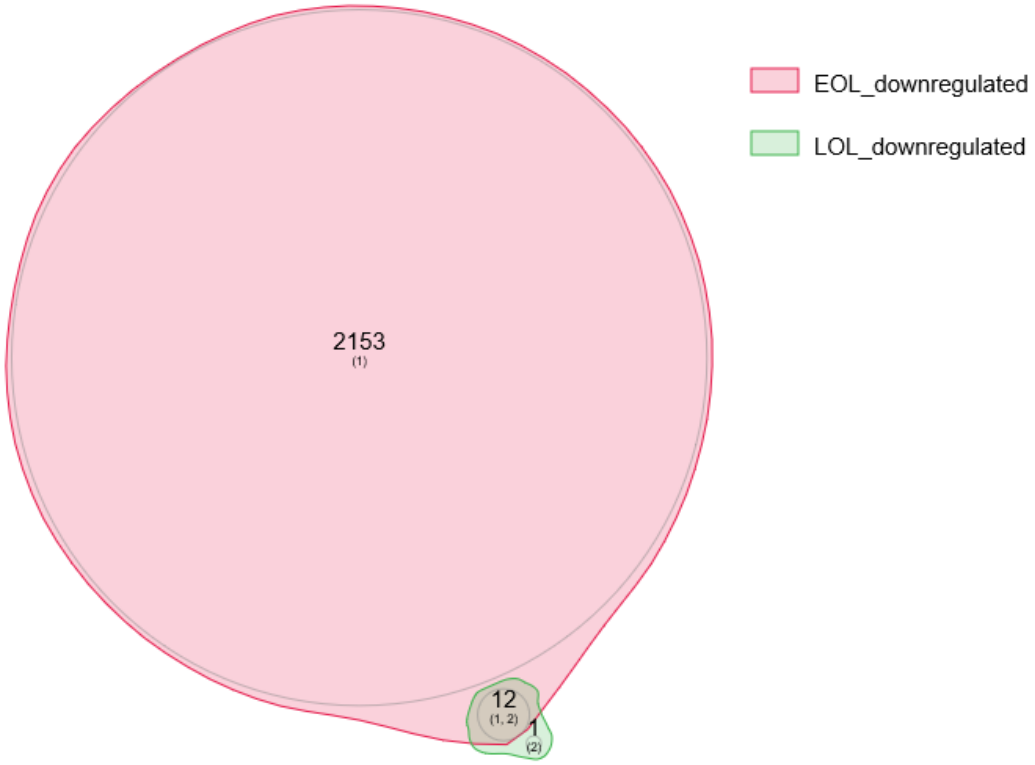

Figure 2.3: Downregulated DEGs - Early Onset and Late Onset samples; A FDR-adjusted p-value cutoff of 0.05 was used to decide differential expression, and the direction of logFC was used to determine upregulation/downregulation.

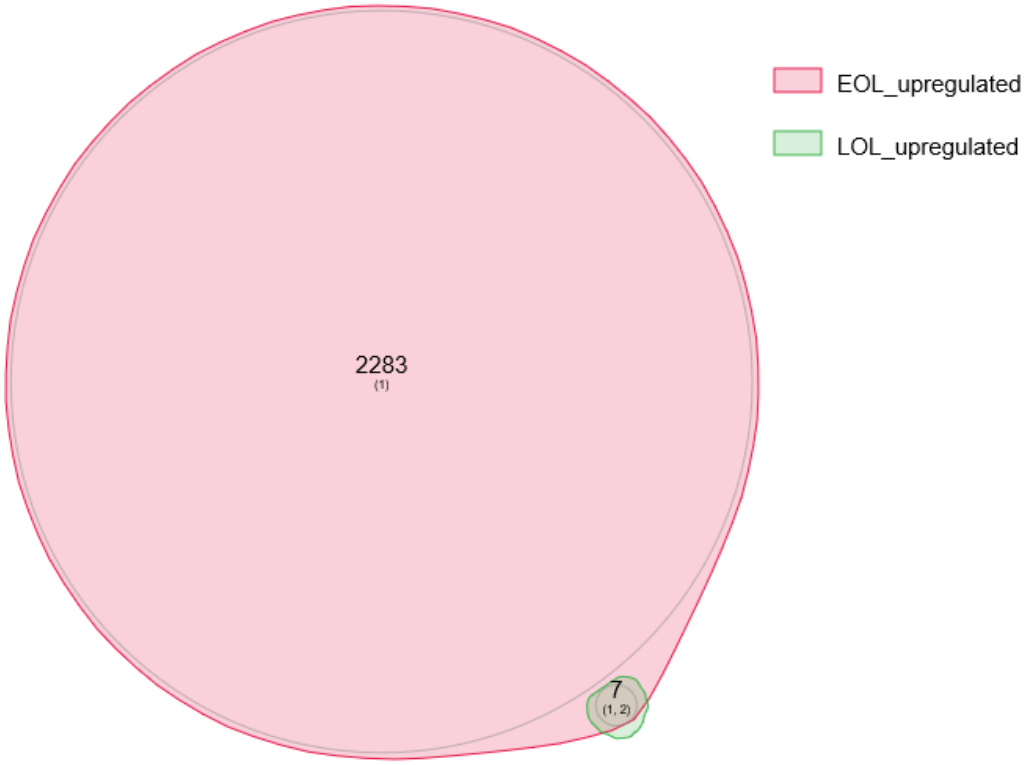

Figure 2.4: Upregulated DEGs - Early Onset and Late Onset samples; A FDR-adjusted p-value cutoff of 0.05 was used to decide differential expression, and the direction of logFC was used to determine upregulation/downregulation.

Chapter 3

Sample Clustering

3.1 Introduction

Hierarchical clustering is an unsupervised [71] learning method that partitions a given group of objects into hierarchy-forming clusters. Dendrograms are used to visualize this hierarchy. Hierarchical clustering algorithms come in two distinct flavours - 1) Agglomerative and 2) Divisive. Agglomerative clustering algorithms start off with each object residing in a cluster of its own; pairs of clusters with the highest similarity are merged sequentially until all samples agglomerate in a single cluster. Divisive clustering algorithms start off with all objects being a part of a single, all-including cluster which is then successively broken down by removing the edges between pairs of clusters that have the lowest similarity [58].

A myriad of distance-based and ratio-based methods [15] have been traditionally used to quantify the similarity between clusters and merge/split them. Single link or MIN is an agglomerative method that uses the shortest distance between two clusters as a measure of similarity; Complete link or MAX is an agglomerative method that uses the maximum distance between two clusters to compute similarity; Group average is an agglomerative method that uses the average pairwise distance between objects from different clusters to appraise similarity.

Ward's method is another agglomerative clustering method that, akin to K-means clustering, produces groups that minimize the within-group dispersion at each fusion. It does so by trying to minimize the extra sum of squares caused by the agglomeration of clusters at each step [55].

The lack of a global objective function in hierarchical clustering makes it hard to decide on an optimal number of partitions and usually, metadata associated with the objects are used to make a decision as to when to stop the clustering process. However, the similarity between traditional k-means clustering and hierarchical clustering using Ward's method allows us to use optimization methods associated with k-means clustering to find a suitable stopping point. One such method determines the best 'k' by optimizing the Silhouette coefficient, a metric that is computed by comparing the average distance between data points in the same cluster against the average of the distances between these points and data points from other clusters (Fig. 3.1).

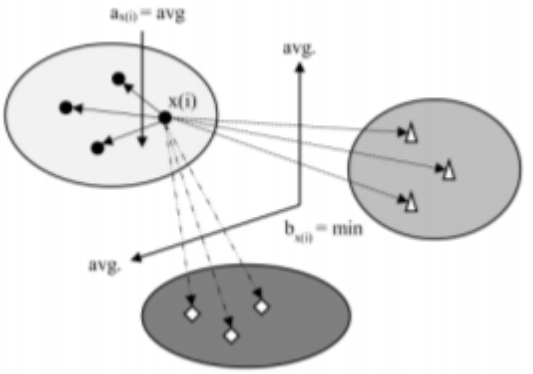

Figure 3.1: Calculating silhouette coefficient [47].

3.2 Clustering - Version 1

The categorization of our 40 AD samples into 19 early-onset sporadic AD (EOSAD) and 21 late-onset sporadic AD (LOSAD) ended up obfuscating differential expression, completely in LOSAD and partially in EOSAD, when contrasted against the Non-demented Controls (NDC). This happens despite our best efforts to weed out the noise in the expression data by filtering

out genes with low read counts. Inherent heterogeneity in the data caused by differences in the mechanisms adopted could be a possible explanation for this phenomenon. Variable gene counts caused by the degradation of the mRNA post-mortem could be another factor contributing to the heterogeneity among the samples. The latter could be accounted for by taking a look at the sequenced samples' quality metrics. Based on the percentage of reads from a sample that was mapped to a transcript, we decided to exclude one LOSAD sample, that had just 3.4% of its reads mapped, from further analysis. The dataset's small sample size motivated us to be conservative and a few AD samples that had a lowly 10-20% of their reads mapped to a transcript were retained for further analysis.

Hierarchical clustering of the 49 remaining samples, using scaled Counts Per Million (CPM) values, and Euclidean distance to compute pairwise similarity between the samples, failed to dichotomize the samples as EOSAD, LOSAD, and controls. Instead, we observe heterogeneous clusters of all three sample types (Fig. 3.2).

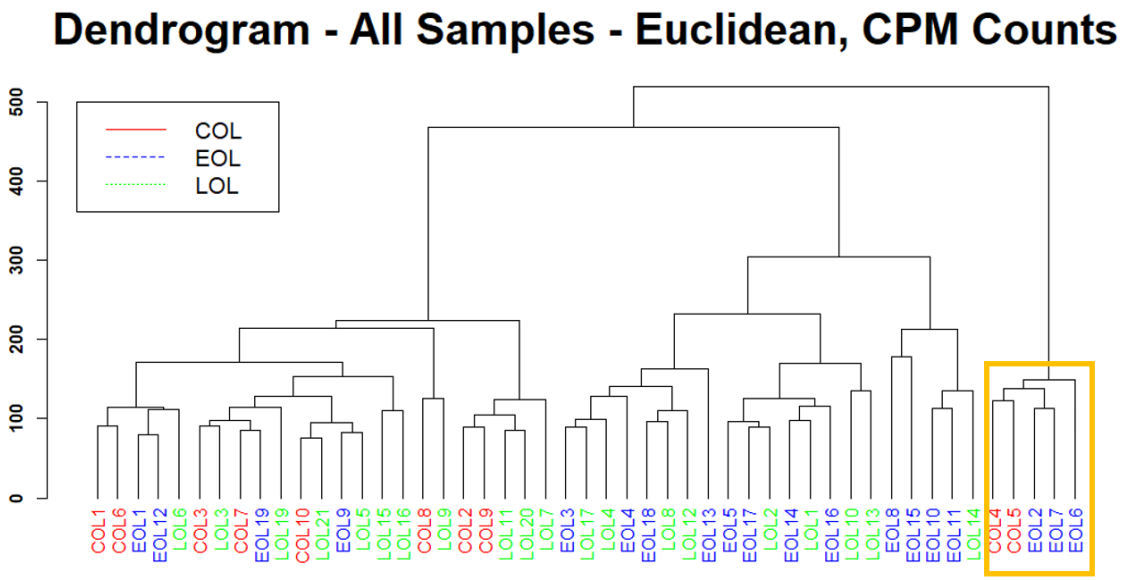

Figure 3.2: Hierarchical clustering of all 49 samples - using filtered CPM counts, Euclidean distance metric.

The clusters formed hint at the possibility that a few of the samples do not show a marked

difference in gene expression as compared to the NDC. Five samples (EOL 2,6,7; COL 4,5) that had the lowest mapped reads percentages, clustered together on the outermost arm of the dendrogram (orange box in Fig 3.2). This clustering of AD samples still holds good upon the removal of the controls from the clustering process (Fig. 3.3).

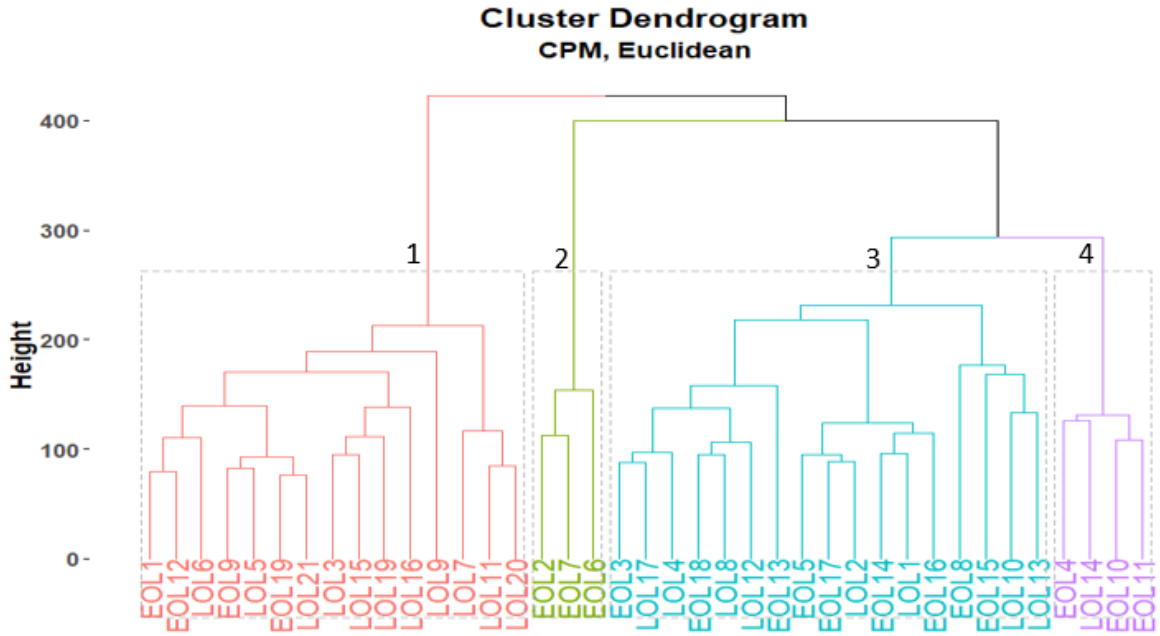

Figure 3.3: Hierarchical clustering of 39 AD samples - using filtered CPM counts, Euclidean distance metric. The four branches that form the clusters of interest are highlighted using numbers.

In order to ensure that these clusters are the result of phenotypic differences between the samples, we converted our gene expression data into geneset enrichment scores using Singscore [27], a single sample gene set scoring method, and the Gene Ontology Biological Process (GO:BP) genesets from the Molecular Signatures Database (MSigDB). Given a list of genes ranked based on their expression values, Singscore performs single-sample enrichment for a geneset and returns a normalized enrichment score which can then be used for phenotypic comparison between samples. Genesets of size 10 or less and/or those with less than 5 genes present in the prefiltered AD gene expression matrix, were excluded from the analysis. A geneset enrichment matrix, with the 39 AD samples as columns and the GO: BP genesets as rows, was created and used to

hierarchically cluster the samples. Euclidean distance was used to quantify similarity between pairs of samples. The dendrogram for this geneset-level clustering looks similar to the dendrogram we get from gene-level clustering.

Using a tanglegram (Fig. 3.4) to compare the two dendrograms, we see that up until a certain height along either dendrogram, the clusters cut from both methods are identical (indicated using the blue line). Hence, it can be reasonably concluded that the genes that drive the expression of specific phenotypes show identical patterns of expression in the AD samples and that samples with such specific gene expression paradigms naturally group together to form clusters.

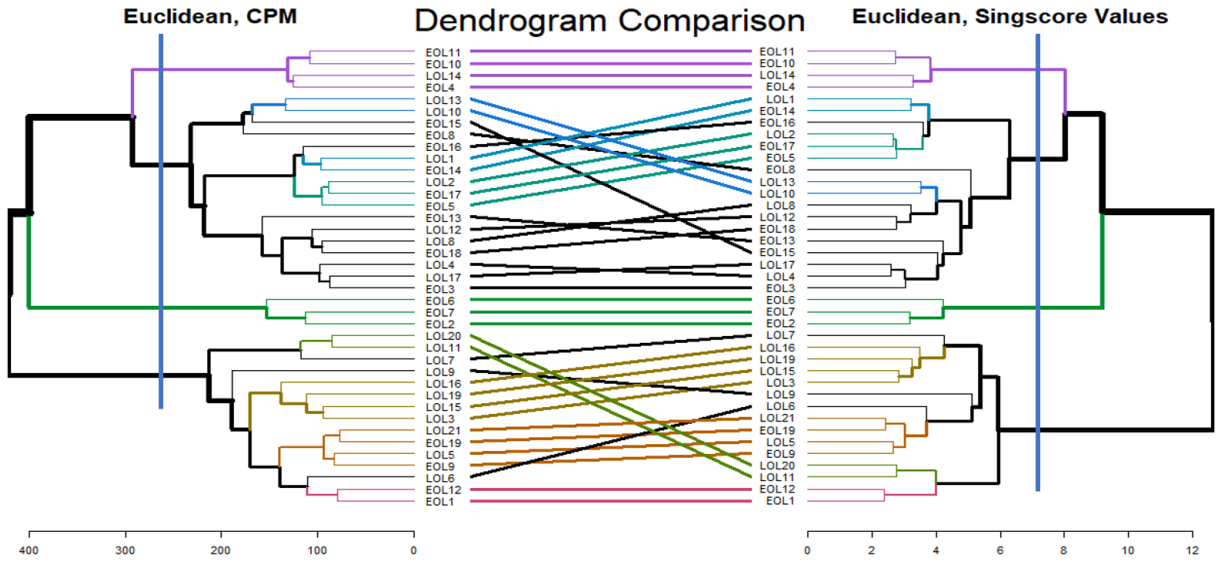

Figure 3.4: Tanglegram comparing gene-level hierarchical clustering and geneset-level hierarchical clustering; Both methods use Euclidean distance to quantify similarity. The level at which the dendrograms are cut to get identical clusters is shown using the blue line.

We decided to cut the dendrograms at a level where we get four identical clusters from either dendrograms (blue line in Fig 3.4). One of the four clusters consists of three EOSAD samples with low mapping percentages. The other three clusters are a mix of LOSAD and EOSAD samples and have 4, 15 and 17 samples. Let these clusters be called Group_1 (15 samples), Group_2 (3 low-quality samples), Group_3 (17 samples) and Group_4 (4 samples).

All four clusters give us different levels of differential expression when compared to

the 8 Non-Demented Controls (NDC). The two outlier controls with low mapping percentages were excluded from this analysis. Samples from Group_1 show no differential expression (FDR-adjusted p-value < 0.05) when compared to the non-demented controls. This is not surprising since these are the 15 samples that cluster together with the 8 NDC in Figure 2.1. Group_2, Group_3 and Group_4 have around 7500, 6500 and 8500 Differentially Expressed Genes (DEGs), respectively, when compared to the controls. A comparison of the upregulated (Fig. 3.6) and downregulated (Fig. 3.5) genes from each cluster indicates that all three groups have some unique DEGs.

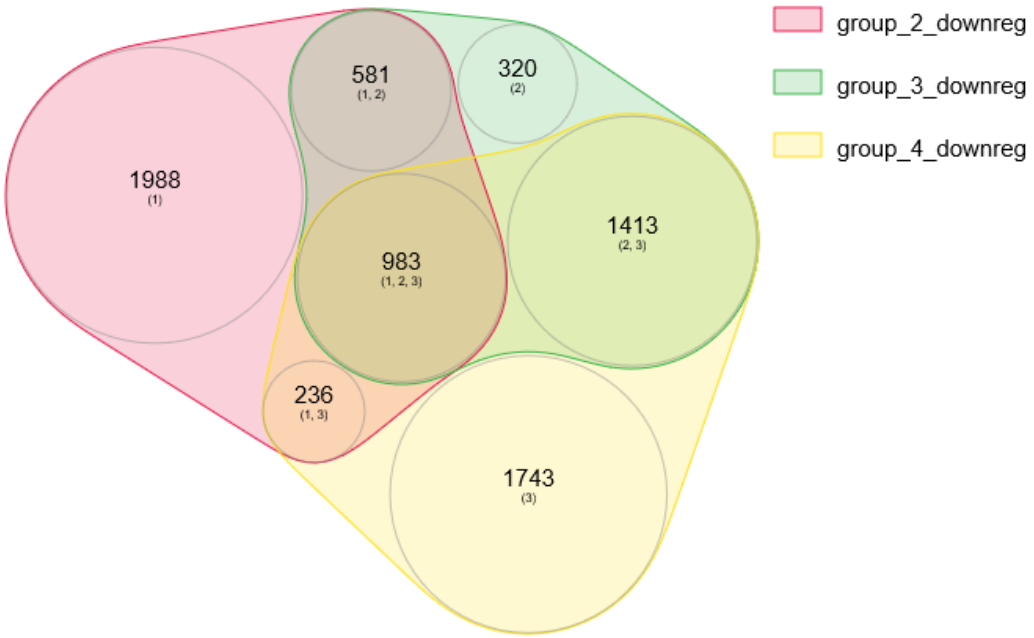

Figure 3.5: Downregulated genes from different clusters - samples from Group_1 have no genes being donwregulated compared to the NDC at a FDR-adjusted p-value cut off of 0.05.

The low RIN scores of these samples lends to the possibility of RNA-degradation - induced/-inhibited differential expression. To check for this, we used a "differential expression quality" (DEqual) plot (Fig. 3.7) [44], a diagnostic plot that shows the correlation in differential

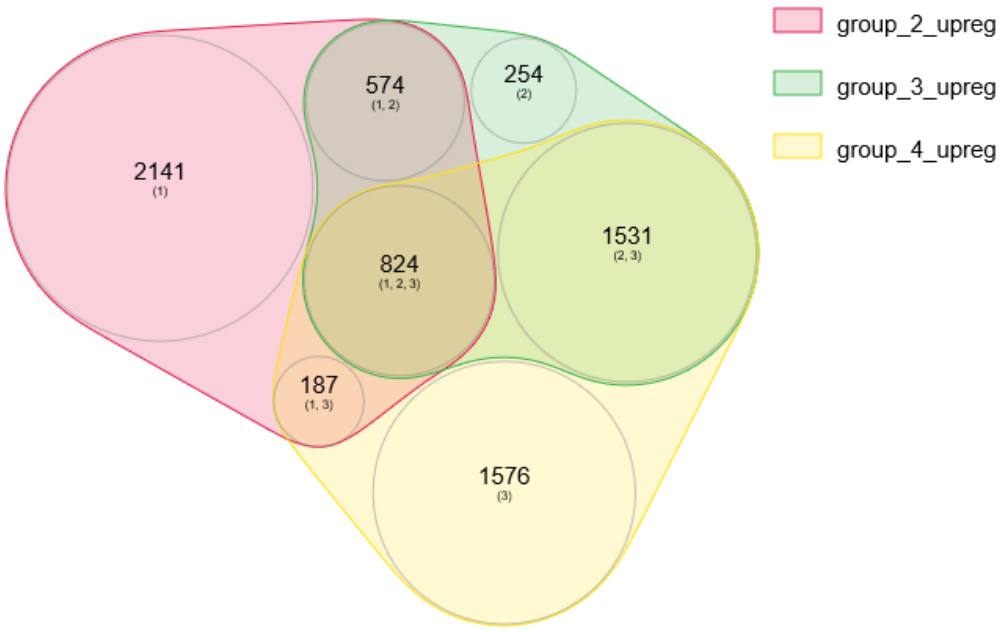

Figure 3.6: Upregulated genes from different clusters - samples from Group_1 have no genes being upregulated compared to the NDC at a FDR-adjusted p-value cut off of 0.05.

expression t-statistics between AD-induced and degradation-induced differential expression. The DEqual plots for groups 1, 3 and 4 all show weak correlations between the t-statistics for degradation-induced and AD-induced differential expression. On the other hand, the plot for Group_2 (Fig. 3.7b) which is comprised of samples with low mapping percentages, shows a strong correlation between the two differential expression statistics. Hence, it can be reasonably concluded that samples from this cluster are severely affected by degradation. These samples were excluded from further analysis.

3.3 Integrating Metadata and Reclustering

Mean values of the metadata associated with the AD samples were compared between the three groups using one-way analysis of variance (ANOVA). Subsequently, Scheffe Multiple

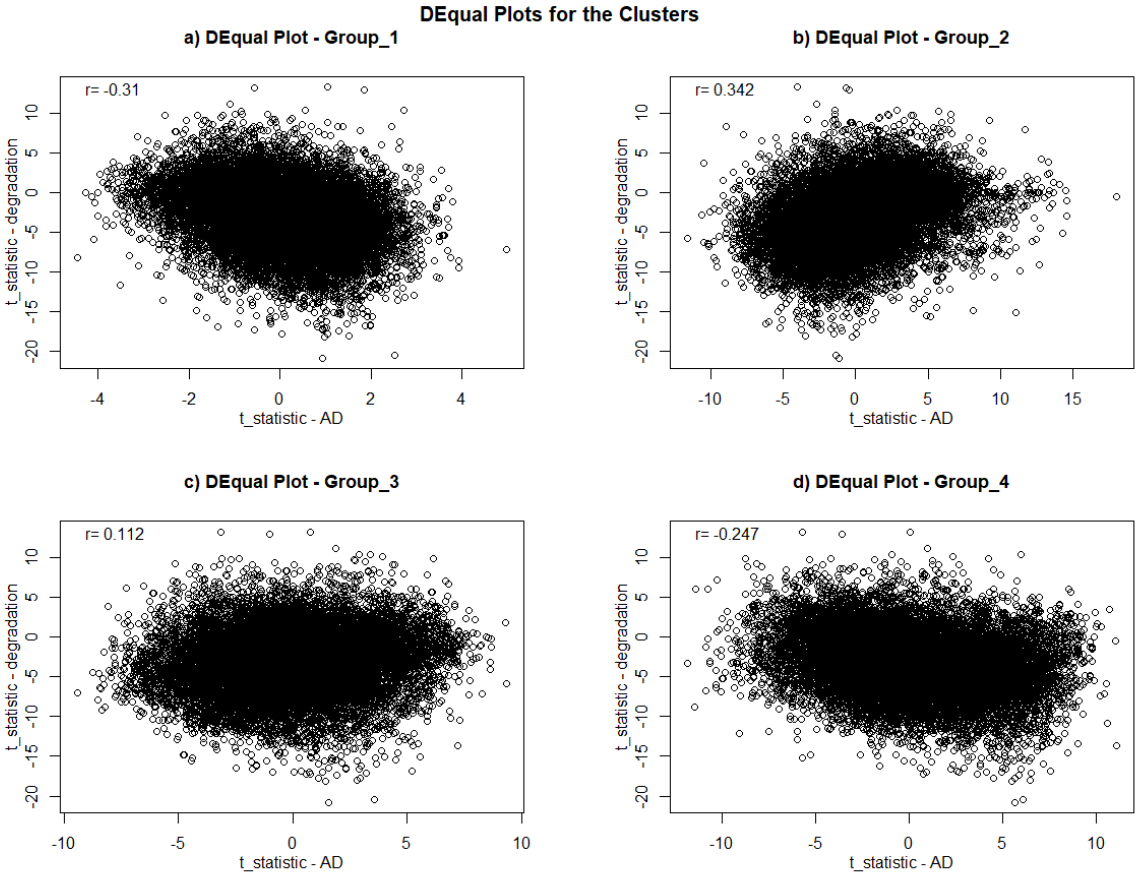

Figure 3.7: DEqual plot for differential expression in all four clusters - only samples from Group_2 (b) seem to have been affected by the degradation of mRNA, post-mortem.

Comparison Test was used to compare pairs of means. Disease-Specific Survival (DSS) is the only metadata that was found to significantly vary between the clusters ($p < 0.01$). Furthermore, pairwise comparisons show that only Group_1 and Group_4 show significant differences in mean DSS values ($p < 0.01$, CI - [1.561, 5.742]) (Fig. 3.8).

Pre-ranked Gene Set Enrichment Analysis (GSEA) takes a list of ranked genes and computes enrichment scores for different genesets using a running-sum statistic. Using permutations of this list, it then calculates a p-value that, after FDR-correction, quantifies the significance of this enrichment score. Pre-ranked GSEA was run on the DEGs from our clusters using the fgseaMultilevel function from the fgsea package v1.15.2 in R, and the MSigDB GO:BP geneset

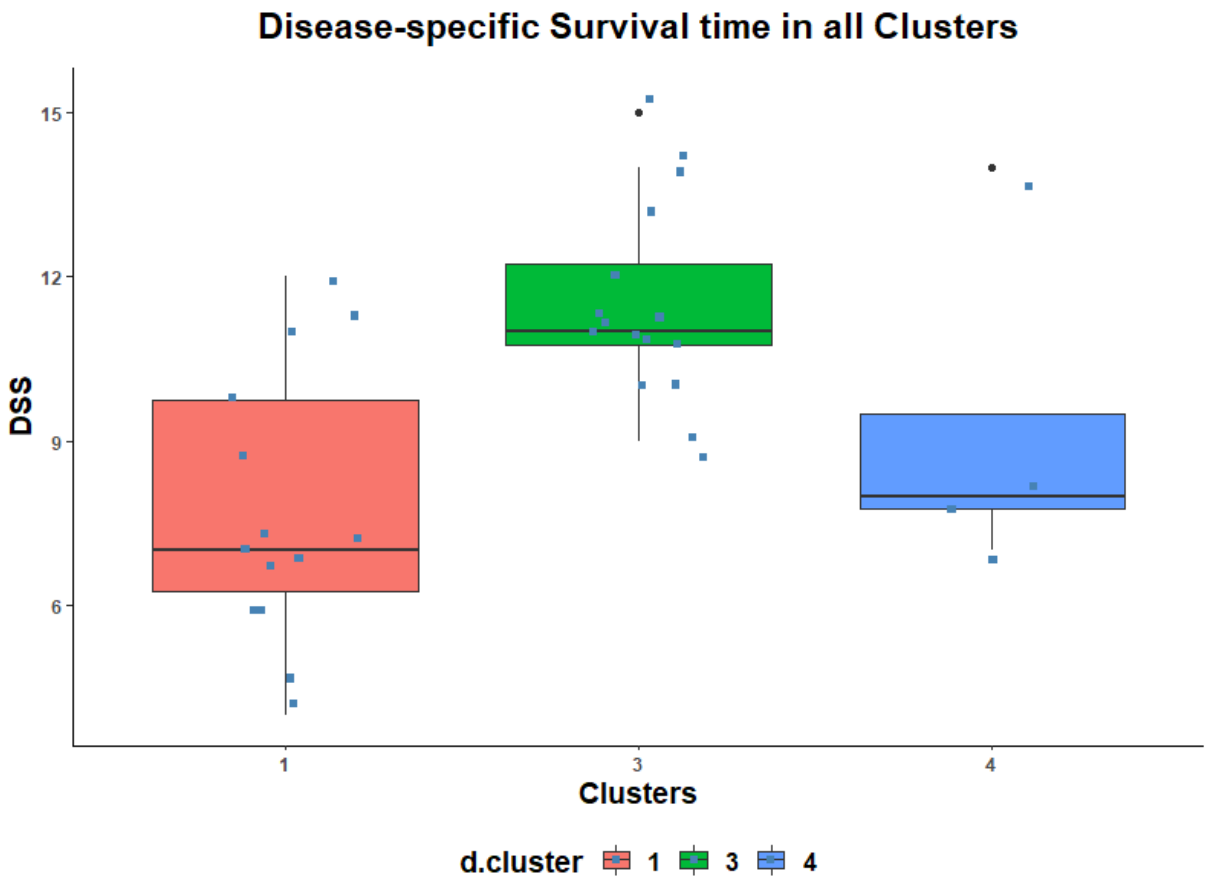

Figure 3.8: Box plot for Disease Specific Survival - One-way analysis of Variance followed by Scheffe Multiple Comparison Test shows that the mean DSS between Group_1 and Group_3 vary significantly ($p < 0.01$).

collection. GSEA results indicate a high degree of commonality among the phenotypes enriched in Group_3 and Group_4 (Fig. 3.9). The top 30 phenotypes enriched in Group_4, ranked by the adjusted p-value, are also enriched, albeit to a lesser extent, in Group_3. Group_4 seemed to consist of samples which had a more severe progression of the disease as compared to the samples from Group_3. Nevertheless, similarities in progression sets samples from these two clusters apart from the samples in Group_1 which have none of these shared phenotypes being enriched. Hence, for further analysis, we decided to combine samples in Group_3 and Group_4 into a bigger cluster, Group_3_4. Moreover, a comparison of the DSS time between Group_1 and Group_3_4 using ANOVA shows a significant difference between the mean DSS times of the two groups

($p < 0.01$) (Fig. 3.10). In addition, upon using the average silhouette, a measure of the relative closeness of a sample from one cluster to samples from other clusters, as a metric to assess the quality of clustering, we find out that splitting the data into 3 clusters gives us the highest average silhouette coefficient in both the gene-based and geneset-based clustering (Fig. 3.12).

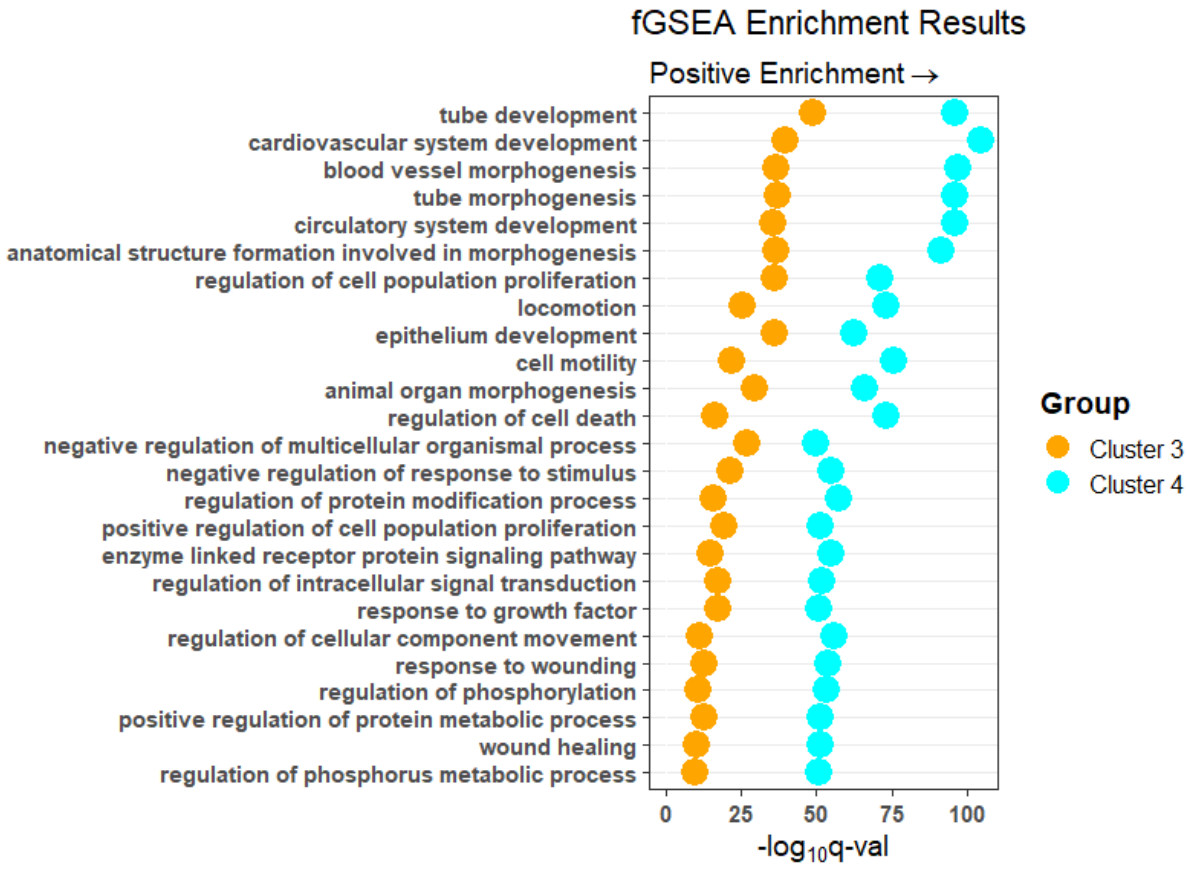

Figure 3.9: fGSEA results for samples from Group_3 and Group_4 - both groups share several enriched phenotypes, but the level of enrichment is higher in the samples from Group_4.

A DEqual plot (Fig. 3.13) of samples from the hybrid Group_3_4 shows a weak correlation between the degradation-induced and AD-induced differential expression t-statistic and hence, degradation-induced expression differences between the AD samples and NDC can be ruled out. Only samples from Group_1 and Group_3_4 were used for enrichment analysis.

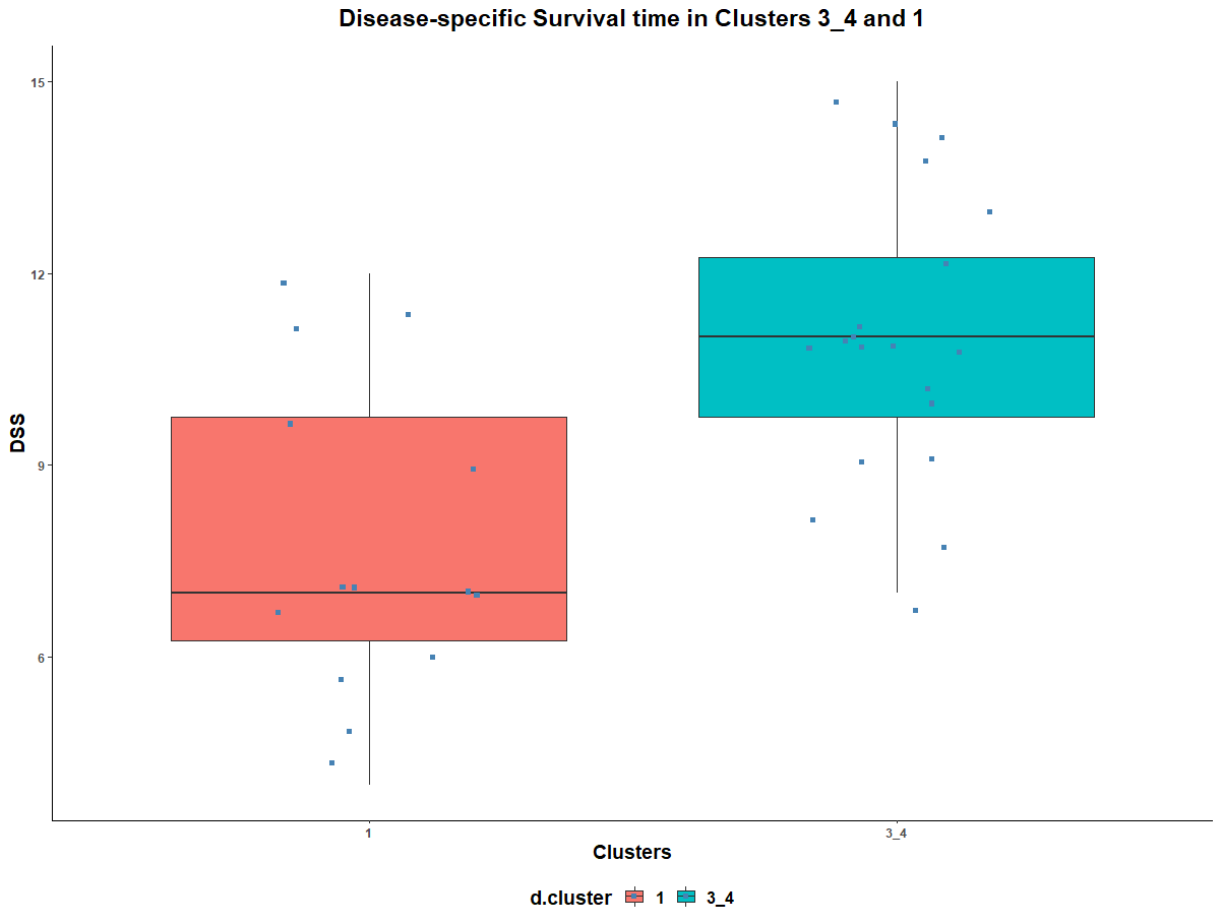

Figure 3.10: Box plot for Disease Specific Survival - One-way analysis of Variance shows that the mean DSS between Group_1 and Group_3_4 varies significantly ($p < 0.01$).

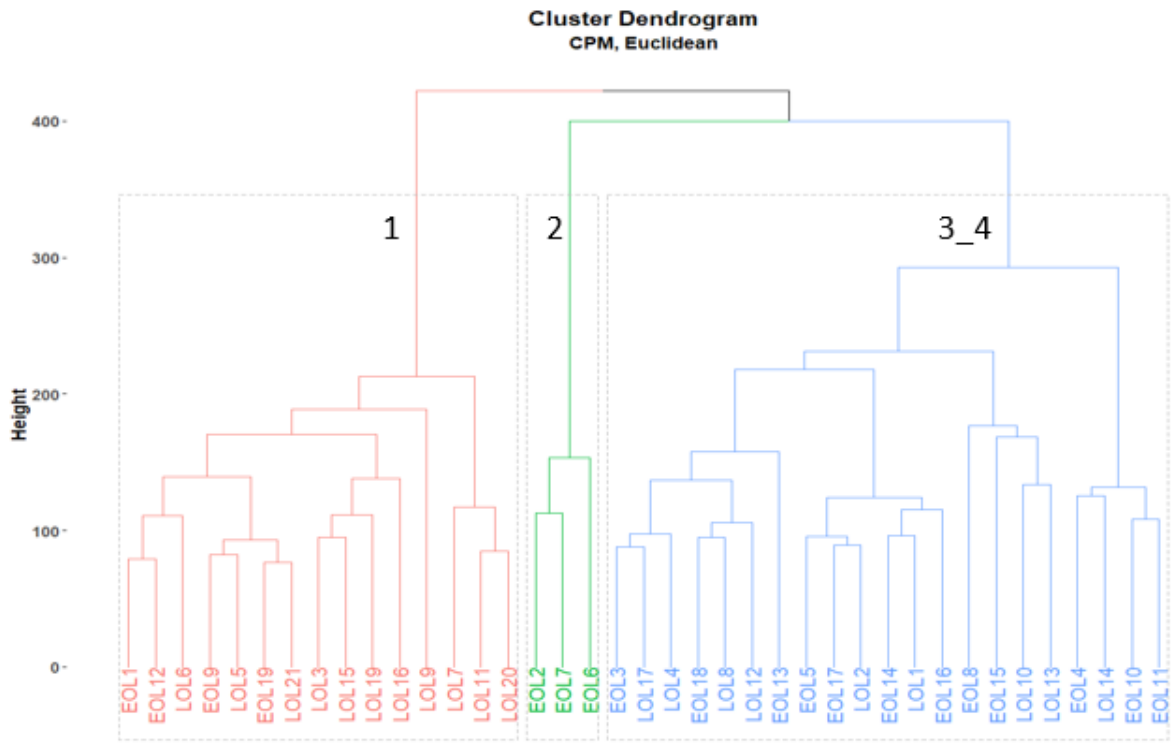

Figure 3.11: Hierarchical clustering of 39 AD samples - using filtered CPM counts, Euclidean Distance metric. The three resultant groups are highlighted using numbers.

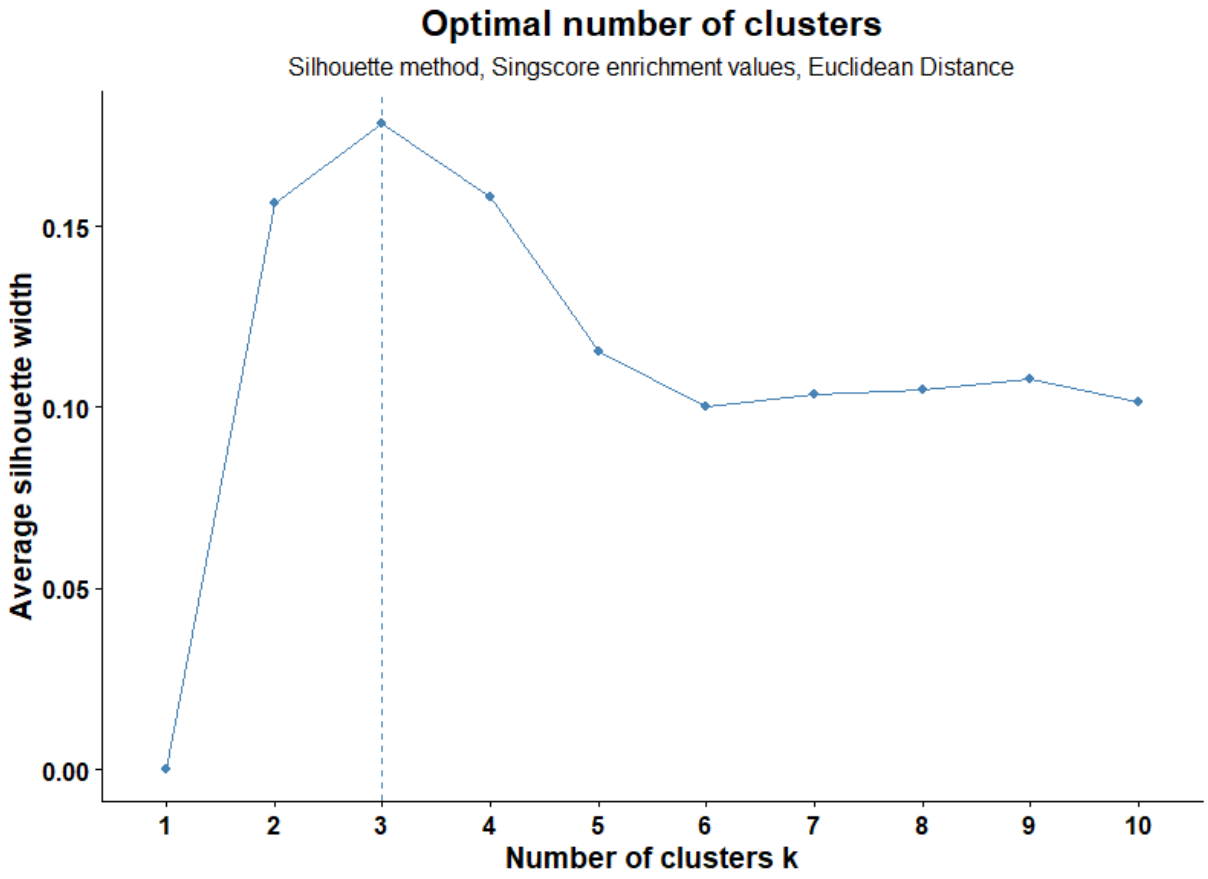

Figure 3.12: Silhouette plot for geneset-level clustering - splitting the data into three clusters gives us the highest silhouette value.

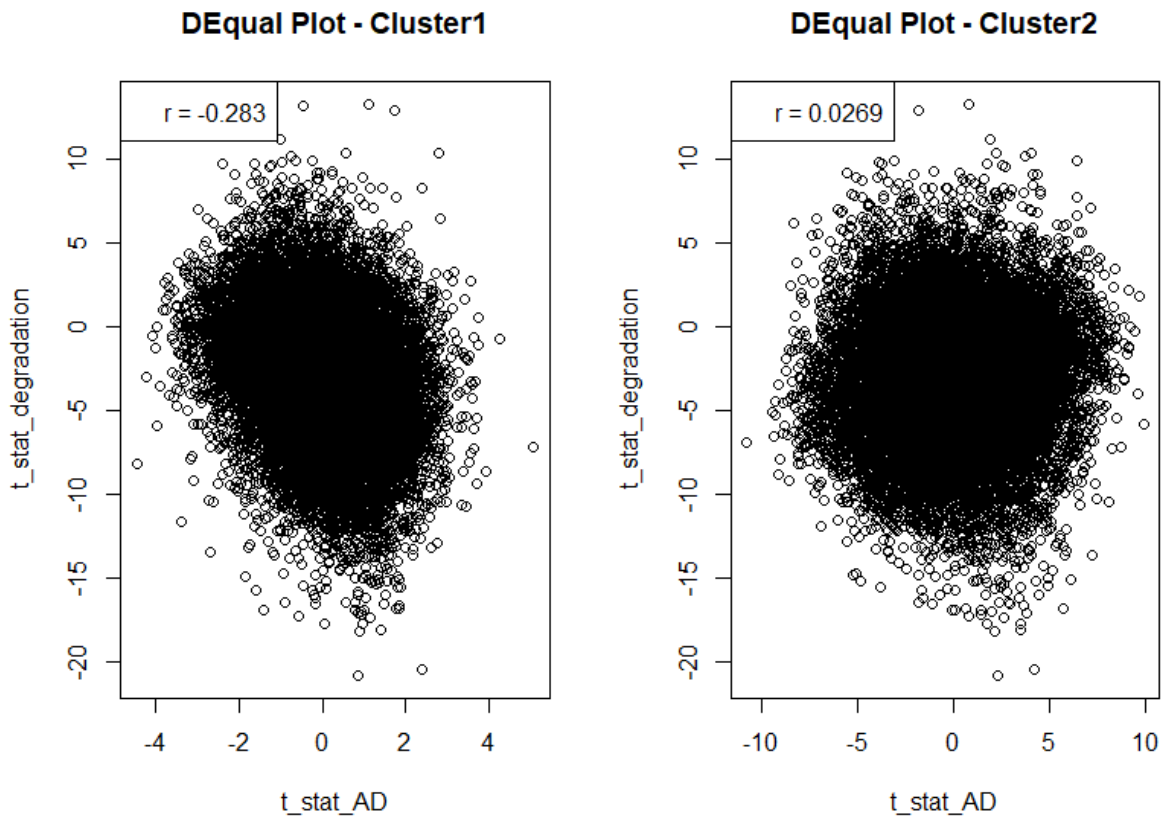

Figure 3.13: DEqual plot for differential expression in Group_1 and Group_3_4 - samples from both clusters seem to be unaffected by mRNA degradation.

Chapter 4

Enrichment Analysis

Differential expression analysis between AD samples from Group_3_4 and Group_1 against NDC identified 3611 upregulated and 3531 downregulated genes, and 0 upregulated and 0 downregulated genes, respectively, at a False Discovery Rate (FDR)-adjusted p-value cutoff of 0.05.

4.1 Methods

Integrated System for Motif Activity Response Analysis (ISMARA) [11] is a web-based tool that, given gene expression data, helps identify key Transcription Factors (TFs) and miRNAs that drive changes in expression and also predicts their mode of regulation. It does so by modelling gene expression data in terms of the computationally predicted regulatory motifs of said TFs and miRNAs. Fastq files, after quality control and adapter trimming, were uploaded to ISMARA using the command-line tool. Upon the completion of motif enrichment, sample averaging was performed and Group_3_4 samples were compared against the NDC controls. A directional Z-score is calculated by multiplying three metrics - 1) the Z-score of each motif, 2) the sign of the Pearson correlation calculated between changes in a TF/miRNA's mRNA expression and changes in the activity of the genes regulated by said TF/miRNA, and 3) the direction of change in the

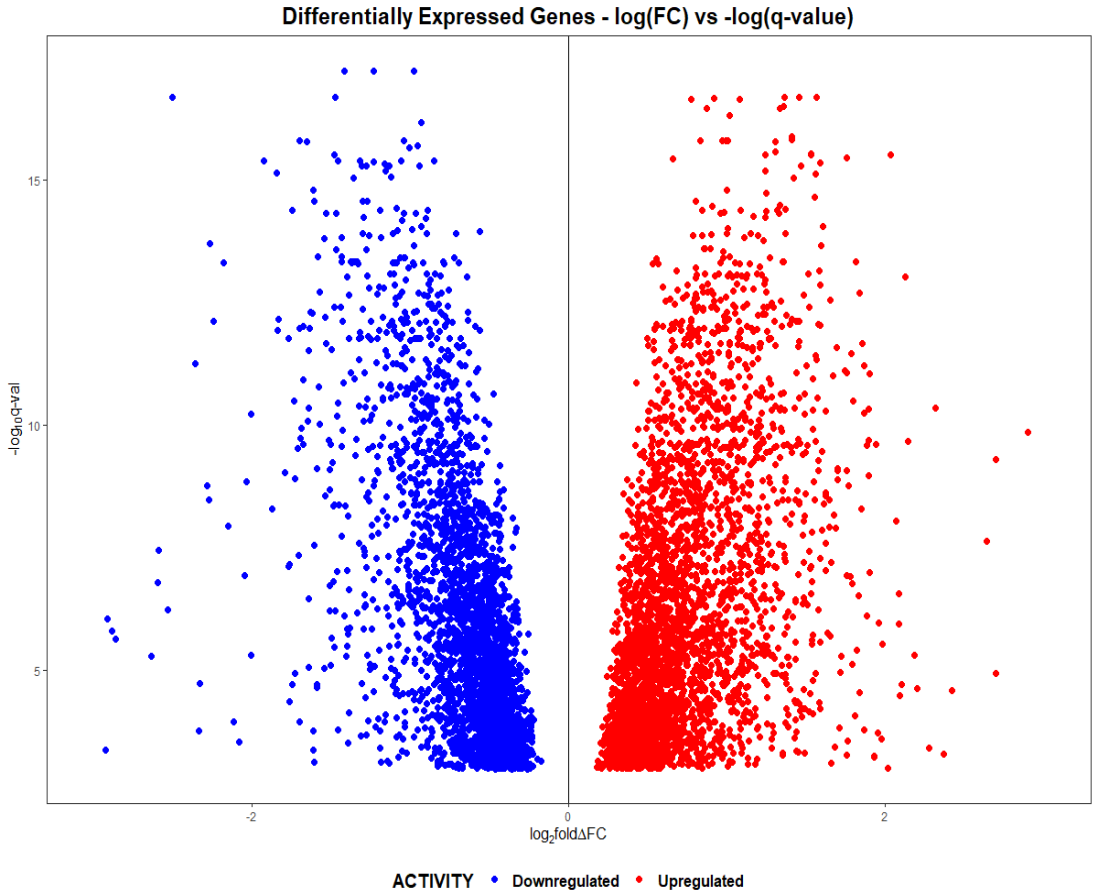

Figure 4.1: Differentially Expressed Genes(DEGs) - $\log_2$FC vs -$\log_{10}$q-val.

activity of the aforementioned genes (+1 for upregulated genes, -1 for downregulated genes). The Pearson score is also used as a proxy for determining the regulatory roles of the TFs and miRNAs - a positive correlation score would indicate that a TF/miRNA with the given motif acts as an activator and a negative Pearson score characterizes a TF/miRNA as a repressor.

ISMARA analysis of differential motif activity in samples from Group_3_4 compared to the NDC controls, helps discern TFs that cause changes in gene expression. Based on known associations with specific endotypes, the TFs were split into 5 groups - Immune/Inflammation, Cell Cycle, Dedifferentiation, Pluripotency and Neuron Lineage. Though ISMARA classifies TFs as activators/repressors based on the activity of their mRNAs, many TFs undergo substantial posttranslational modification and hence, the usage of a TF's mRNA levels to predict their mode

of action is unreliable.

The Virtual Inference of Protein-activity by Enriched Regulon (VIPER) [5] algorithm computationally infers protein activity using the expression of genes that directly interact with a given protein. These include interactions between TFs and their target genes, and VIPER infers the activity of a TF while taking into account its regulatory role(s), the confidence of the TF-target interaction, and the pleiotropic nature of TF-gene regulation. The eponymous tool that deploys this algorithm was used in conjunction with DoRothEA [30], a gene set resource consisting of signed TF-target interactions where each TF-target interaction is furnished with a confidence level based on supporting evidence, to identify key TFs that engender changes in gene expression between the AD and NDC samples.

Minimal Significant Difference (MSD) [74], calculated as the lowest possible value of the absolute log2 fold-change (logFC) within the 95% Confidence Interval (CI), is used as a metric to rank the genes for enrichment analysis. We've adopted the use of this statistic to rank our genes for downstream analysis. For VIPER, the MSD metric for each gene was multiplied by the sign of the logFC and this value was used as an estimate of the change in expression of the gene between the different sets of samples.

The regulons identified by VIPER were also stratified into 5 groups based on known associations with specific endotypes. Subsequently, pre-ranked Gene Set Enrichment analysis (GSEA) [67] was performed on the signed-MSD ranked genes using the fgseaMultilevel function from the fGSEA package v 1.15.2 in R. Using an adaptive multi-level split Monte Carlo scheme, this function allows us to swiftly and accurately calculate low GSEA p-values [48] for a given gene set collection. The MSigDB v7.1 Gene Ontology (GO): Biological Process (BP) collection was used for GSEA.

The endotype Immune/Inflammation response is upregulated in the AD samples from Group_3_4. Related GO:BP genesets such as "Cytokine Production" and "Regulation of Immune System Progress" are upregulated in group_3_4 (Fig. 4.4). Moreover, proinflammatory TFs

such as CEBPB, STAT3, and RELA [50] (Fig. 4.2a, Fig. 4.3a) show an uptick in their activity in samples from Group_3_4. However, the increase in mRNA levels of NFKB (Fig. 4.3a) need not result in an increase in the concentration of NFKB-related TFs due to the extensive posttranslational modifications the NFKB mRNA is subjected to [42].

TFs that regulate the progression of Cell Cycle, such as the E2F family of TFs [63], MYC, MYCN [14], and MAZ [6] (Fig. 4.2b, Fig. 4.3b) all show a decrease in activity in the AD samples.

Upregulation in the activity of ZEB1 promotes tumor progression since ZEB1 represses regulators of epithelial polarity [4]. ISMARA results show an increase in ZEB1's activity (Fig 4.2c) and VIPER results indicate that ZEB1's targets are downregulated (Fig. 4.3c). Similarly, ERG, a TF that's linked to the malignancy of prostate cancer [3], shows an increase in activtiy in the AD samples from Group_3_4 (Fig 4.2c). The target genes of ZBTB7A, a pleiotropic TF which when overexpressed suppresses the growth of castration-resistant Prostate Cancer [35], end up being downregulated in these AD samples (Fig 4.3 c). Other related TFs such as TEAD1 [68] and TEAD3 show similar trends as well. GSEA results highlight the upregulation of dedifferentiation related phenotypes such as "Tube Development", "Cardiovascular System Development", and "Epithelium Development" (Fig. 4.4). Pluripotency, a closely related endotype, is also upregulated with GATA2 [7] and KLF4 [33], two zinc-finger proteins that play a key role in inducing pluripotency, both showing an increase in activity (Fig. 4.2d). The target genes of NANOG and SOX2 [25], two pleiotropic TFs commonly expressed by pluripotent cells, are also donwregulated in the AD samples from Group_3_4(Fig. 4.3d).

REST, a key suppressor of neuronal genes in nonneuronal cells and a repressed gene in mature neurons, shows higher activity in AD brains compared to NDC (Fig 4.2e). RCOR1 (also known as Co-Rest), a gene that selectively represses certain genes in a mature neuron, shows lower activity in the AD brains (Fig 4.2e). Increase in REST's activity foments a large scale suppression of neuronal genes [10] in the brain and thereby, could cause lineage reversion

of mature neurons to a neural precursor state. Downregulation of GSEA-identified neuronal phenotypes such as ”Synaptic Signalling” (Fig. 4.4) and ”Synaptic Vesicle Exocytosis” (Fig. 4.4) might hint towards a net loss in functioning neurons in the AD samples.

4.2 Conclusion and Future Directions

Transcriptomic profiling of AD patients provides evidence for the presence of two distinct etiologies. One etiology is driven chiefly by dedifferentiation and REST-driven loss of neurons. The other seems to be transcriptomically no different from non-demented aging. An immediate next goal would be to build gene-protein networks using the list of enriched phenotypes and TFs, and visualize TF-gene interactions that beget the expression of different endotypes. Another goal would be to build a mathematical model that uses the metadata associated with a sample to predict the etiology that the sample has adopted.

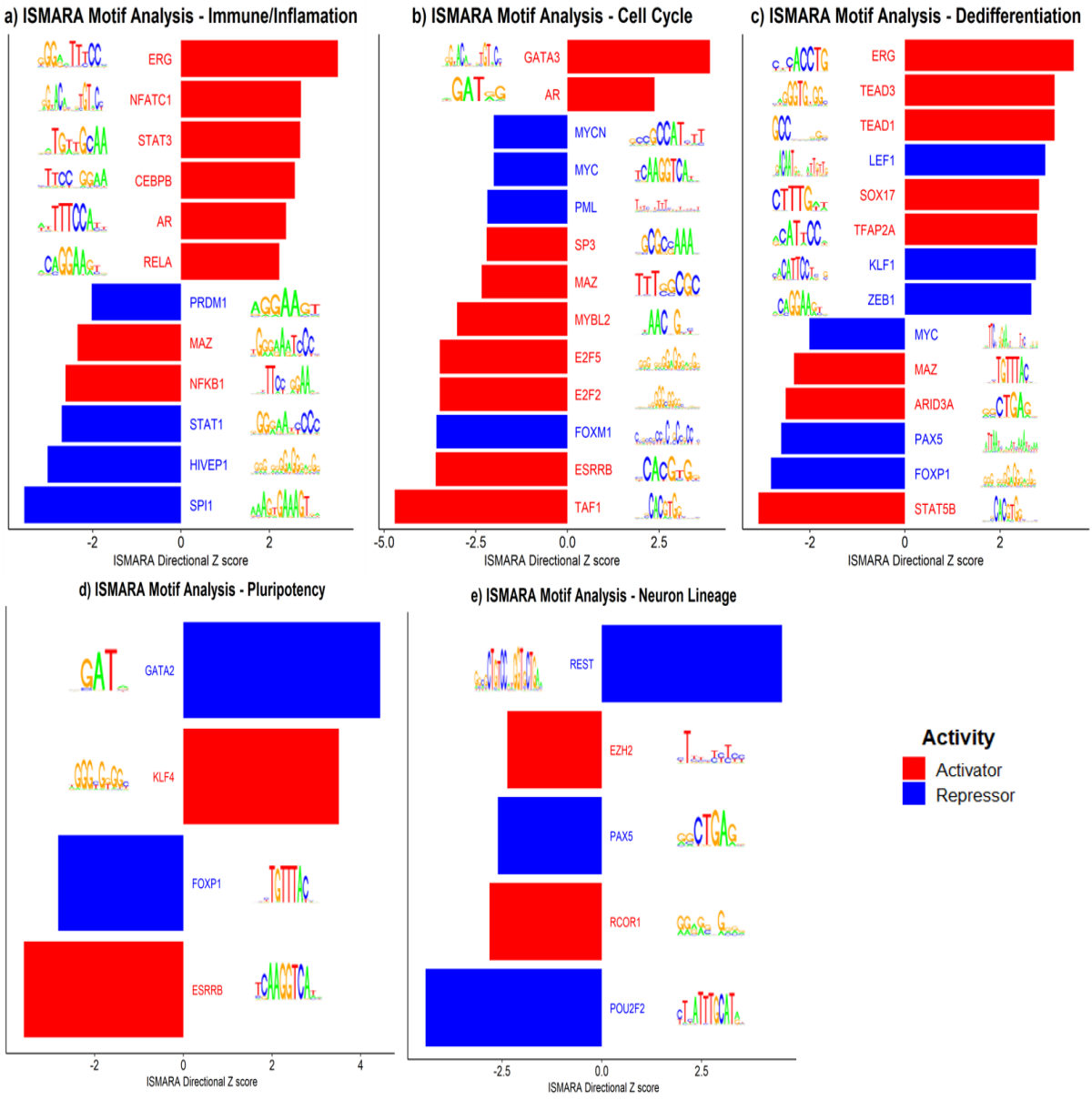

Figure 4.2: **ISMARA motif analysis - Endotypes - Group_3_4** - Differentially active Transcription Factors (TFs) [ISMARA Z-score > 2 and Pearson Correlation > 0.1] in Group_3_4 AD samples related to **a** Immune/Inflammation Response **b** Cell Cycle Regulation **c** Dedifferentiation **d** Pluripotency **e** Neuron Lineage; Directional Z-score is calculated using ISMARA Z-score, activity difference, and Pearson correlation. Also shown are the motifs to which these TFs bind.

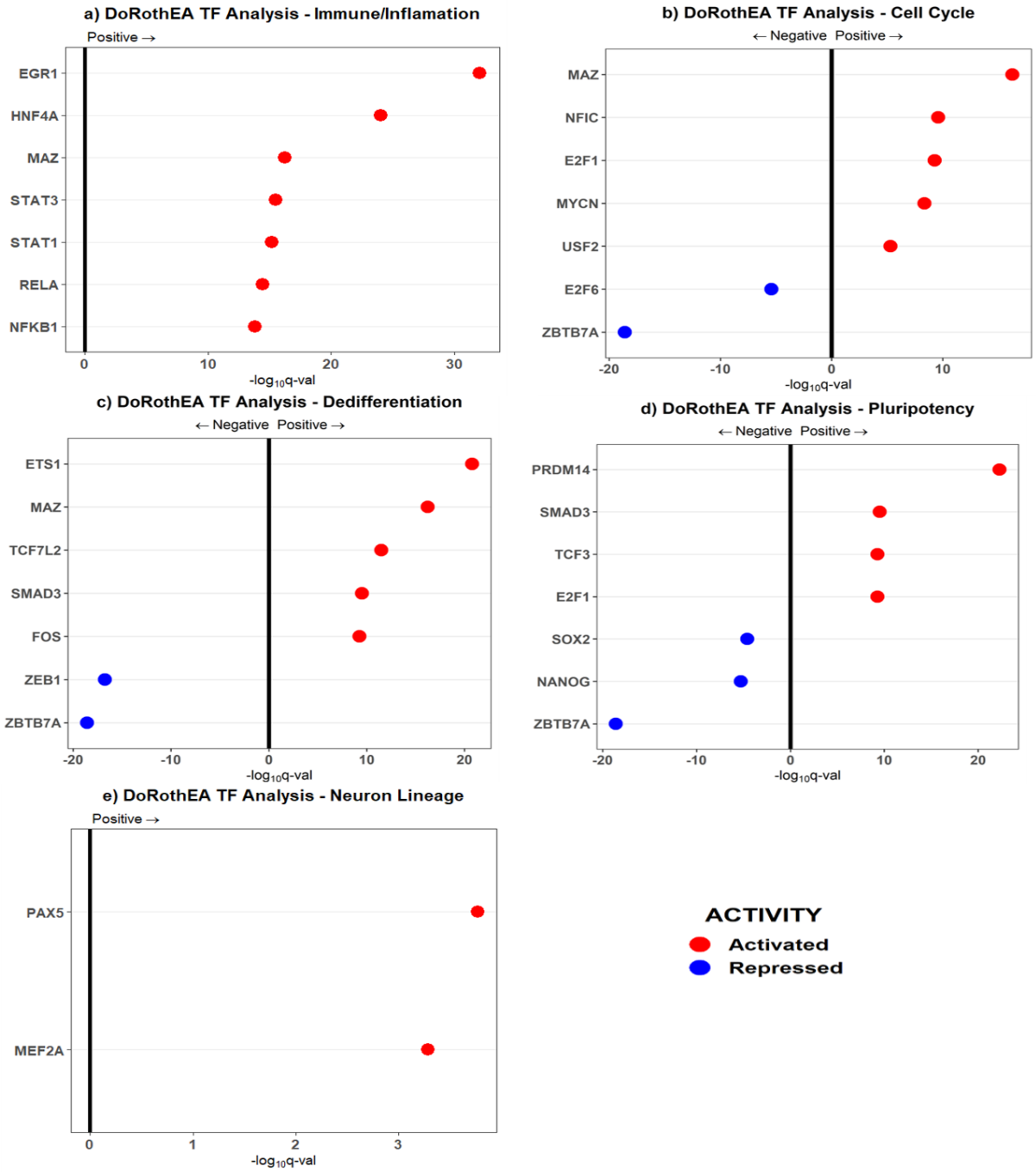

Figure 4.3: **DoRothEA TF analysis - Endotypes - Group_3_4** - VIPER enrichment [adj.p.val < 0.05] for TFs related to **a** Immune/Inflammation Response **b** Cell Cycle Regulation **c** Dedifferentiation **d** Pluripotency **e** Neuron Lineage; Score is calculated using the direction of enrichment and the adjusted p-value.

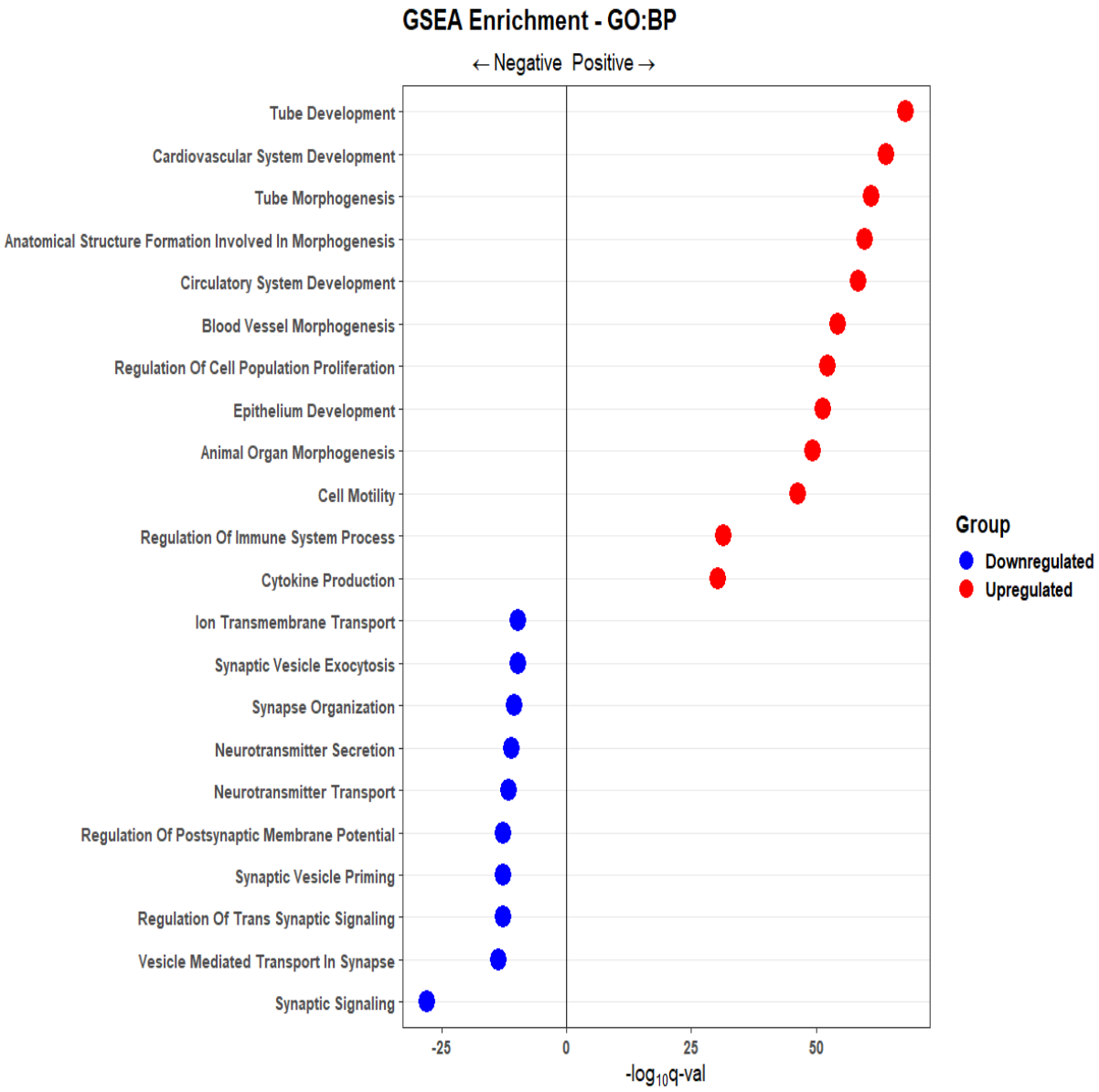

Figure 4.4: **GSEA enriched phenotypes - Group_3_4** - GSEA Enrichment using MSigDB v7.1 GO:BP Gene sets - top 10 upregulated and top 10 downregulated phenotypes; Score is calculated using the direction of enrichment and the adjusted p-value.

Bibliography

[1] Cutadapt removes adapter sequences from high-throughput sequencing reads — Martin — EMBnet.journal.

[2] 2020 Alzheimer's disease facts and figures. *Alzheimer's and Dementia*, 16(3):391–460, 3 2020.

[3] P Adamo and M R Ladomery. The oncogene ERG: a key factor in prostate cancer. *Oncogene*, 35:403–414, 2016.

[4] K. Aigner, B. Dampier, L. Descovich, M. Mikula, A. Sultan, M. Schreiber, W. Mikulits, T. Brabletz, D. Strand, P. Obrist, W. Sommergruber, N. Schweifer, A. Wernitznig, H. Beug, R. Foisner, and A. Eger. The transcription factor ZEB1 (δEF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity. *Oncogene*, 26(49):6979–6988, 10 2007.

[5] Mariano J. Alvarez, Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48(8):838–847, 8 2016.

[6] Josué Álvaro-Blanco, Katia Urso, Yuri Chiodo, Carla Martín-Cortázar, Omar Kourani, Pablo Gómez-Del Arco, María Rodríguez-Martínez, Esther Calonge, José Alcamí, Juan Miguel Redondo, Teresa Iglesias, and Miguel R. Campanero. MAZ induces MYB expression during the exit from quiescence via the E2F site in the MYB promoter. *Nucleic Acids Research*, 45(17):9960–9975, 9 2017.

[7] Zhaojun An, Peng Liu, Jiashun Zheng, Michael Q Zhang, Qi Zhou, and Sheng Ding Correspondence. Sox2 and Klf4 as the Functional Core in Pluripotency Induction without Exogenous Oct4 Graphical Abstract Highlights d Polycistronic Sox2 and Klf4 reprogrammed mouse somatic cells to iPSCs d Stoichiometry of Sox2 and Klf4 is essential for S 2A K 2A M reprogramming d 2 MEFs and NPCs adopted convergent trajectories in S 2A K 2A M reprogramming d Sox2 and Klf4 cooperatively bound to induce the pluripotency network. *Cell Reports*, 29, 2019.

[8] Mary Atz, David Walsh, Preston Cartagena, Jun Li, Simon Evans, Prabhakara Choudary, Kevin Overman, Richard Stein, Hiro Tomita, Steven Potkin, Rick Myers, Stanley J. Watson, E. G. Jones, Huda Akil, William E. Bunney, and Marquis P. Vawter. Methodological considerations for gene expression profiling of human brain. *Journal of Neuroscience Methods*, 163(2):295–309, 7 2007.

[9] Adnan A. Awada. Early and late-onset Alzheimer's disease: What are the differences?, 7 2015.

[10] Nurit Ballas, Christopher Grunseich, Diane D. Lu, Joan C. Speh, and Gail Mandel. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell*, 121(4):645–657, 5 2005.

[11] Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik Van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, 24(5):869–884, 2014.

[12] G. Blessed, B. E. Tomlinson, and M. Roth. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. *The British journal of psychiatry : the journal of mental science*, 114(512):797–811, 1968.

[13] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 5 2016.

[14] Gabriel Bretones, M. Dolores Delgado, and Javier León. Myc and cell cycle control, 5 2015.

[15] R. C. Campbell, R. R. Sokal, and F. J. Rohlf. Biometry: The Principles and Practice of Statistics in Biological Research. *Journal of the Royal Statistical Society. Series A (General)*, 133(1):102, 1970.

[16] Rudy J Castellani and Mark A Smith. Compounding artefacts with uncertainty, and an amyloid cascade hypothesis that is 'too big to fail'. *The Journal of Pathology*, 224(2):147–152, 6 2011.

[17] Sarah L. Cole and Robert Vassar. The role of amyloid precursor protein processing by BACE1, the β-secretase, in Alzheimer disease pathophysiology, 10 2008.

[18] E. H. Corder, A. M. Saunders, N. J. Risch, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, J. B. Rimmler, P. A. Locke, P. M. Conneally, K. E. Schmader, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics*, 7(2):180–184, 6 1994.

[19] Eliza Courtney, Shan Kornfeld, Karolina Janitz, and Michal Janitz. Transcriptome profiling in neurodegenerative disease, 11 2010.

[20] Alix De Calignon, Manuela Polydoro, Marc Suárez-Calvet, Christopher William, David H. Adamowicz, Kathy J. Kopeikina, Rose Pitstick, Naruhiko Sahara, Karen H. Ashe, George A. Carlson, Tara L. Spires-Jones, and Bradley T. Hyman. Propagation of Tau Pathology in a Model of Early Alzheimer's Disease. *Neuron*, 73(4):685–697, 2 2012.

[21] Michael A. Deture and Dennis W. Dickson. The neuropathological diagnosis of Alzheimer's disease, 8 2019.

[22] Fabienne Dulin, Frédéric Léveillé, Javier Becerril Ortega, Jean Paul Mornon, Alain Buisson, Isabelle Callebaut, and Nathalie Colloc'h. p3 peptide, a truncated form of Aβ devoid of synaptotoxic effect, does not assemble into soluble oligomers. *FEBS Letters*, 582(13):1865–1870, 6 2008.

[23] Stefan F. Lichtenthaler. Alpha-Secretase Cleavage of the Amyloid Precursor Protein: Proteolysis Regulated by Signaling Pathways and Protein Trafficking. *Current Alzheimer Research*, 9(2):165–177, 2 2012.

[24] Lindsay A. Farrer. Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. *JAMA*, 278(16):1349, 10 1997.

[25] Adam Filipczyk, Carsten Marr, Simon Hastreiter, Justin Feigelman, Michael Schwarzfischer, Philipp S. Hoppe, Dirk Loeffler, Konstantinos D. Kokkaliaris, Max Endele, Bernhard Schauberger, Oliver Hilsenbeck, Stavroula Skylaki, Jan Hasenauer, Konstantinos Anastassiadis, Fabian J. Theis, and Timm Schroeder. Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nature Cell Biology*, 17(10):1235–1246, 10 2015.

[26] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 11 1975.

[27] Momeneh Foroutan, Dharmesh D. Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19(1):404, 11 2018.

[28] Anthony F. Fotenos, Mark A. Mintun, Abraham Z. Snyder, John C. Morris, and Randy L. Buckner. Brain volume decline in aging: Evidence for a relation between socioeconomic status, preclinical Alzheimer disease, and reserve. *Archives of Neurology*, 65(1):113–120, 1 2008.

[29] T. Chris Gamblin, Feng Chen, Angara Zambrano, Aida Abraha, Sarita Lagalwar, Angela L. Guillozet, Meiling Lu, Yifan Fu, Francisco Garcia-Sierra, Nichole LaPointe, Richard Miller, Robert W. Berry, Lester I. Binder, and Vincent L. Cryns. Caspase cleavage of tau: Linking amyloid and neurofibrillary tangles in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):10032–10037, 8 2003.

[30] Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, 8 2019.

[31] Todd E. Golde, Christopher B. Eckman, and Steven G. Younkin. Biochemical detection of Aβ isoforms: Implications for pathogenesis, diagnosis, and treatment of Alzheimer's disease, 7 2000.

[32] Jill S. Goldman, Susan E. Hahn, Jennifer Williamson Catania, Susan Larusse-Eckert, Melissa Barber Butson, Malia Rumbaugh, Michelle N. Strecker, J. Scott Roberts, Wylie Burke, Richard Mayeux, and Thomas Bird. Genetic counseling and testing for Alzheimer disease: Joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors. *Genetics in Medicine*, 13(6):597–605, 6 2011.

[33] Andreia M. Gomes, Ilia Kurochkin, Betty Chang, Michael Daniel, Kenneth Law, Namita Satija, Alexander Lachmann, Zichen Wang, Lino Ferreira, Avi Ma'ayan, Benjamin K. Chen, Dmitri Papatsenko, Ihor R. Lemischka, Kateri A. Moore, and Carlos Filipe Pereira. Cooperative Transcription Factor Induction Mediates Hemogenic Reprogramming. *Cell Reports*, 25(10):2821–2835, 12 2018.

[34] Deborah R. Gustafson, Ingmar Skoog, Lars Rosengren, Henrik Zetterberg, and Kaj Blennow. Cerebrospinal fluid β-amyloid 1-42 concentration may predict cognitive decline in older women. *Journal of Neurology, Neurosurgery and Psychiatry*, 78(5):461–464, 2007.

[35] Dong Han, Sujun Chen, Wanting Han, Shuai Gao, Jude N. Owiredu, Muqing Li, Steven P. Balk, Housheng Hansen He, and Changmeng Cai. ZBTB7A mediates the transcriptional repression activity of the androgen receptor in prostate cancer. *Cancer Research*, 79(20):5260–5271, 10 2019.

[36] Diane P. Hanger, Brian H. Anderton, and Wendy Noble. Tau phosphorylation: the therapeutic challenge for neurodegenerative disease, 3 2009.

[37] John Hardy. Has the Amyloid Cascade Hypothesis for Alzheimers Disease been Proved? *Current Alzheimer Research*, 3(1):71–73, 2 2006.

[38] Liesi E. Hebert, Jennifer Weuve, Paul A. Scherr, and Denis A. Evans. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 5 2013.

[39] Zachary T. Herbert, Jamie P. Kershner, Vincent L. Butty, Jyothi Thimmapuram, Sulbha Choudhari, Yuriy O. Alekseyev, Jun Fan, Jessica W. Podnar, Edward Wilcox, Jenny Gipson, Allison Gillaspy, Kristen Jepsen, Sandra Splinter BonDurant, Krystalynne Morris, Maura Berkeley, Ashley LeClerc, Stephen D. Simpson, Gary Sommerville, Leslie Grimmett, Marie Adams, and Stuart S. Levine. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics*, 19(1):199, 3 2018.

[40] David M. Holtzman, Kelly R. Bales, Tanya Tenkova, Anne M. Fagan, Maia Parsadanian, Leah J. Sartorius, Brian Mackey, John Olney, Daniel McKeel, David Wozniak, and Steven M. Paul. Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 97(6):2892–2897, 3 2000.

[41] Young T. Hong, Tonny Veenith, Deborah Dewar, Joanne G. Outtrim, Vaithianadan Mani, Claire Williams, Sally Pimlott, Peter J.A. Hutchinson, Adriana Tavares, Roberto Canales, Chester A. Mathis, William E. Klunk, Franklin I. Aigbirhio, Jonathan P. Coles, Jean Claude Baron, John D. Pickard, Tim D. Fryer, William Stewart, and David K. Menon. Amyloid imaging with carbon 11 - Labeled pittsburgh compound B for traumatic brain injury. *JAMA Neurology*, 71(1):23–31, 2014.

[42] Bo Huang, Xiao Dong Yang, Acacia Lamb, and Lin Feng Chen. Posttranslational modifications of NF-κB: Another layer of regulation for NF-κB signaling pathway, 9 2010.

[43] Clifford R. Jack, Terry M. Therneau, Stephen D. Weigand, Heather J. Wiste, David S. Knopman, Prashanthi Vemuri, Val J. Lowe, Michelle M. Mielke, Rosebud O. Roberts, Mary M. Machulda, Jonathan Graff-Radford, David T. Jones, Christopher G. Schwarz, Jeffrey L. Gunter, Matthew L. Senjem, Walter A. Rocca, and Ronald C. Petersen. Prevalence of Biologically vs Clinically Defined Alzheimer Spectrum Entities Using the National Institute on Aging-Alzheimer's Association Research Framework. *JAMA Neurology*, 76(10):1174–1183, 10 2019.

[44] Andrew E. Jaffe, Ran Tao, Alexis L. Norris, Marc Kealhofer, Abhinav Nellore, Joo Heon Shin, Dewey Kim, Yankai Jia, Thomas M. Hyde, Joel E. Kleinman, Richard E. Straub, Jeffrey T. Leek, and Daniel R. Weinberger. QSVA framework for RNA quality correction in differential expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 114(27):7130–7135, 7 2017.

[45] Joanna L. Jankowsky, Daniel J. Fadale, Jeffrey Anderson, Guilian M. Xu, Victoria Gonzales, Nancy A. Jenkins, Neal G. Copeland, Michael K. Lee, Linda H. Younkin, Steven L. Wagner, Steven G. Younkin, and David R. Borchelt. Mutant presenilins specifically elevate the levels of the 42 residue β-amyloid peptide in vivo: Evidence for augmentation of a 42-specific γ secretase, 1 2004.

[46] Lesley Jones, Peter A. Holmans, Marian L. Hamshere, Denise Harold, Valentina Moskvina, Dobril Ivanov, Andrew Pocklington, Richard Abraham, Paul Hollingworth, Rebecca Sims, Amy Gerrish, Jaspreet Singh Pahwa, Nicola Jones, Alexandra Stretton, Angharad R. Morgan, Simon Lovestone, John Powell, Petroula Proitsi, Michelle K. Lupton, Carol Brayne, David C. Rubinsztein, Michael Gill, Brian Lawlor, Aoibhinn Lynch, Kevin Morgan, Kristelle S. Brown, Peter A. Passmore, David Craig, Bernadette Mcguinness, Stephen Todd, Clive Holmes, David Mann, A. David Smith, Seth Love, Patrick G. Kehoe, Simon Mead, Nick Fox, Martin Rossor, John Collinge, Wolfgang Maier, Frank Jessen, Britta Schürmann, Hendrik van den Bussche, Isabella Heuser, Oliver Peters, Johannes Kornhuber, Jens Wiltfang, Martin

Dichgans, Lutz Frölich, Hampel Harald, Michael Hüll, Dan Rujescu, Alison M. Goate, John S.K. Kauwe, Carlos Cruchaga, Petra Nowotny, John C. Morris, Kevin Mayo, Gill Livingston, Nicholas J. Bass, Hugh Gurling, Andrew Mcquillin, Rhian Gwilliam, Panos Deloukas, Ammar Al-Chalabi, Christopher E. Shaw, Andrew B. Singleton, Rita Guerreiro, Thomas W. Mühleisen, Markus M. Nöthen, Susanne Moebus, Karl Heinz Jöckel, Norman Klopp, H. Erich Wichmann, Eckhard Rüther, Minerva M. Carrasquillo, V. Shane Pankratz, Steven G. Younkin, John Hardy, Michael C. O'Donovan, Michael J. Owen, and Julie Williams. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. *PLoS ONE*, 5(11), 2010.

[47] Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop. The silhouette width criterion for clustering and association mining to select image features. *International Journal of Machine Learning and Computing*, 8(1):69–73, 2 2018.

[48] Gennady Korotkevich, Vladimir Sukhov, and Alexey Sergushichev. Fast gene set enrichment analysis. *bioRxiv*, page 060012, 10 2016.

[49] Samir Kumar-Singh, Jessie Theuns, Bianca Van Broeck, Daniel Pirici, Krist'l Vennekens, Ellen Corsmit, Marc Cruts, Bart Dermaut, Rong Wang, and Christine Van Broeckhoven. Mean age-of-onset of familial Alzheimer disease caused by presenilin mutations correlates with both increased Aβ42 and decreased Aβ40. *Human Mutation*, 27(7):686–695, 7 2006.

[50] Heehyoung Lee, Andreas Herrmann, Jie Hui Deng, Maciej Kujawski, Guilian Niu, Zhiwei Li, Steve Forman, Richard Jove, Drew M. Pardoll, and Hua Yu. Persistently Activated Stat3 Maintains Constitutive NF-κB Activity in Tumors. *Cancer Cell*, 15(4):283–293, 4 2009.

[51] Dana M. Niedowicz, Peter T. Nelson, and M. Paul Murphy. Alzheimers Disease: Pathological Mechanisms and Recent Insights. *Current Neuropharmacology*, 9(4):674–684, 12 2011.

[52] S Mattis. Dementia rating scale: professional manual. 1988.

[53] C. M. Monoranu, M. Apfelbacher, E. Grünblatt, B. Puppe, I. Alafuzoff, I. Ferrer, S. Al-Saraj, K. Keyvani, A. Schmitt, P. Falkai, J. Schittenhelm, G. Halliday, J. Kril, C. Harper, C. McLean, P. Riederer, and W. Roggendorf. PH measurement as quality control on human post mortem brain tissue: A study of the BrainNet Europe consortium. *Neuropathology and Applied Neurobiology*, 35(3):329–337, 2009.

[54] E. C. Mormino, J. T. Kluth, C. M. Madison, G. D. Rabinovici, S. L. Baker, B. L. Miller, R. A. Koeppe, C. A. Mathis, M. W. Weiner, and W. J. Jagust. Episodic memory loss is related to hippocampal-mediated β-amyloid deposition in elderly subjects. *Brain*, 132(5):1310–1323, 5 2009.

[55] Fionn Murtagh and Pierre Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295, 10 2014.

[56] Deborah A. Nickerson, Scott L. Taylor, Stephanie M. Fullerton, Kenneth M. Weiss, Andrew G. Clark, Jari H. Stengård, Veikko Salomaa, Eric Boerwinkle, and Charles F. Sing. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Research*, 10(10):1532–1545, 2000.

[57] Daniel P. Perl. Neuropathology of Alzheimer's disease, 1 2010.

[58] Harun Pirim, Burak Ekşioglu, Andy D. Perkins, and Çetin Yüceer. Clustering of high throughput gene expression data, 12 2012.

[59] Serge Przedborski, Miquel Vila, and Vernice Jackson-Lewis. Series Introduction: Neurodegeneration: What is it and where are we? *Journal of Clinical Investigation*, 111(1):3–10, 1 2003.

[60] Dorene M. Rentz, Joseph J. Locascio, John A. Becker, Erin K. Moran, Elisha Eng, Randy L. Buckner, Reisa A. Sperling, and Keith A. Johnson. Cognition, reserve, and amyloid deposition in normal aging. *Annals of Neurology*, 67(3):353–364, 3 2010.

[61] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 1 2015.

[62] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[63] Subhashini Sadasivam and James A. DeCaprio. The DREAM complex: Master coordinator of cell cycle-dependent gene expression, 8 2013.

[64] Wilfried Schroyens, Casey O'connell, and David B Sykes. Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. *New England Journal of Medicine*, 367(8):780–780, 8 2012.

[65] Jay Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7):585–587, 7 2008.

[66] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*, 4:1521, 2 2016.

[67] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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, 10 2005.

[68] Jessica Tome-Garcia, Parsa Erfani, German Nudelman, Alexander M Tsankov, Igor Katsyv, Bin Zhang, Martin Walsh, Roland H Friedel, Elena Zaslavsky, and Nadejda M Tsankova. Analysis of chromatin accessibility uncovers TEAD1 as a regulator of migration in human glioblastoma.

[69] Natalie A. Twine, Karolina Janitz, Marc R. Wilkins, and Michal Janitz. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. *PLoS ONE*, 6(1):e16266, 1 2011.

[70] Jan Verheijen and Kristel Sleegers. Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics, 6 2018.

[71] Dan Wei, Qingshan Jiang, Yanjie Wei, and Shengrui Wang. A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics*, 13(1), 7 2012.

[72] Kristin R. Wildsmith, Monica Holley, Julie C. Savage, Rebecca Skerrett, and Gary E. Landreth. Evidence for impaired amyloid β clearance in Alzheimer's disease, 7 2013.

[73] Shanrong Zhao, Wai Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1):78644, 1 2014.

[74] Joanna Zyla, Michal Marczyk, January Weiner, and Joanna Polanska. Ranking metrics in gene set enrichment analysis: Do they matter? *BMC Bioinformatics*, 18(1):1–12, 5 2017.