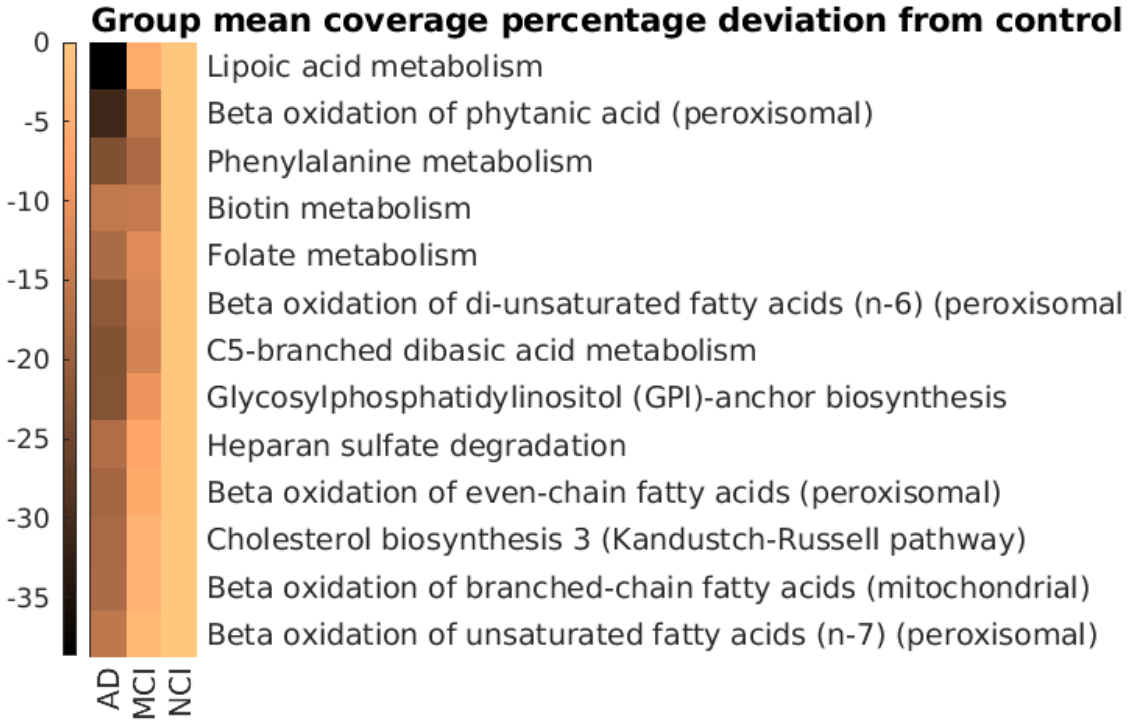


Can AD be identified through metabolic models?

Using ROSMAP bulk brain RNA-seq data to create single-sample genome-scale metabolic models of persons with and without Alzheimer's Disease

Master's thesis in Complex Adaptive Systems

GASTON SANDSTIG

DEPARTMENT OF LIFE SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Master's thesis 2024

Can AD be identified through metabolic models?

Using ROSMAP bulk brain RNA-seq data to create single-sample genome-scale metabolic models of persons with and without Alzheimer's Disease

Gaston Sandstig

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Division of Systems and Synthetic Biology
Polster Group
Chalmers University of Technology
Gothenburg, Sweden 2024

Can AD be identified through metabolic models?
Using ROSMAP bulk brain RNA-seq data to create single-sample genome-scale metabolic models of persons with and without Alzheimer's Disease
GASTON SANDSTIG

Supervisor: Annikka Polster, Department of Life Sciences
Examiner: Eduard Kerkhoven, Department of Life Sciences

Master's Thesis 2024
Department of Life Sciences
Division of Systems and Synthetic Biology
Polster Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Heatmap showing possible impairment of metabolic subsystems in those affected with Alzheimer's Disease. Abbreviations AD, MCI, and NCI stand for Alzheimer's Disease, minor, and no cognitive impairment respectively. Further discussed in 3.3.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2024

Can AD be identified through metabolic models?
Analysing single-sample genome-scale metabolic models of brain tissue with and without Alzheimer's Disease
Gaston Sandstig
Department of Life Sciences
Chalmers University of Technology

Abstract

The most common form of dementia is Alzheimer's Disease (AD), a progressive neurodegenerative disorder, but so far no methods of curing or preventing it are known. Amyloid-β proteins have been the focus of therapeutic development but clinical trials focused on reducing amyloid-β production or preventing its aggregation have failed, leading to increased interests in other areas of AD pathology. There is evidence of impaired metabolism in those with AD, and in this project gene expression data from the ROSMAP study was used to construct single-sample genome-scale metabolic models in an effort to see what differences can be discerned between models from those with AD and not. Although there was no observed separation between the groups when using PCA and t-SNE, the aggregate metabolic coverage between groups differed in several metabolic subsystems. The differences found had support in literature, but subsystems C5-branched dibasic acid metabolism, GPI-anchor biosynthesis, heparan sulphate degradation, and lipoic acid metabolism had not been identified as differing in comparable *in silico* metabolic AD model studies.

Keywords: Alzheimer's Disease, Genome-Scale Metabolic Model, Metabolic Dysfunction, RNA-seq

Acknowledgements

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG36836 (RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic).

In addition I would like to thank the many people who have helped me in this project from SysBio. Of course special thanks to my supervisor Annikka, my examiner Ed, as well as Danish and all the other colleagues who have been part of Polster lab this previous year.

Gaston Sandstig, Gothenburg, January 2024

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AD	Alzheimer's Disease
GEM	Genome-scale metabolic model
MCI	Minor Cognitive Impairment
NCI	No Cognitive Impaiment
TPM	Transcripts per million

Contents

List of Acronyms **vii**

1 Introduction **1**
1.1 Background . . . 1
1.1.1 Alzheimer's Disease . . . 1
1.1.2 Personalised treatment of AD . . . 2
1.1.3 AD metabolism . . . 2
1.1.4 Modelling metabolism with GEMs . . . 2
1.2 Aim . . . 3

2 Theory **4**
2.1 Normalising count data . . . 4
2.1.1 TPM . . . 4
2.1.2 Quantile normalisation . . . 5
2.2 Modelling metabolism with GEMs . . . 5
2.2.1 How does a GEM work? . . . 5
2.2.2 GEM construction algorithm . . . 7
2.3 Clustering and comparison of GEMs . . . 8
2.3.1 Clustering by distance . . . 8
2.3.2 Subsystem coverage comparison . . . 8

3 Results **9**
3.1 Data analysis . . . 9
3.1.1 Excluding samples . . . 9
3.1.2 Association of metadata . . . 9
3.1.3 Correlation of count data to metadata . . . 12
3.2 Normalisation . . . 13
3.2.1 TPM and gene distributions . . . 13
3.2.2 Relative log expression plots and quantile normalisation . . . 15
3.3 Structural GEM differences . . . 16
3.3.1 PCA and tSNE clustering . . . 16
3.3.2 Differences in subsystem coverage . . . 17
3.3.3 Comparison to other *in silico* AD metabolism studies . . . 19

4 Conclusion **21**
4.1 Subsystem coverage . . . 21
4.2 MCI modeling . . . 21

4.3 Clustering and normalisation 22
4.4 Gene filtration 22
4.5 Future use 22

Bibliography **23**

A Appendix **I**

1
Introduction

1.1 Background

Improved health standards and medical advances have caused life expectancy worldwide to nearly double in the last 100 years [1]. In the intervening time we have also gone from understanding dementia as an inevitable part of ageing – i.e. senility – to a disease that could potentially be avoided, prevented, or cured [2]. The increased life expectancy in combination with low fertility rates in most developed countries has also resulted in older people accounting for a growing share of the total population [3]. Because of this, dementia is increasing in prevalence and finding ways to prevent it is important both from a health and societal perspective [4].

1.1.1 Alzheimer's Disease

The most common form of dementia is Alzheimer's Disease (AD), a progressive neurodegenerative disorder. More than 10 million people are diagnosed with AD annually, but so far no methods of curing or preventing it are known [5]. Although sporadic ("typical", late-onset) AD is estimated to be 70% heritable, its genetic basis is not well understood in contrast to familial (autosomal dominant, early-onset) cases which are caused by mutations in the amyloid-β precursor protein (APP) and its processing proteins [6]. Since aggregations of amyloid-β and tau proteins in the brain are characteristic of AD [7], and amyloid-β or APP seemed to be the primary cause of familial AD, it prompted the hypothesis that amyloid-β was the cause of sporadic AD as well and that any tau pathology was a downstream effect [8]. APP and amyloid-β proteins have been the focus of therapeutic development but clinical trials focused on reducing amyloid-β production or preventing its aggregation have failed which points to a more complicated relationship between AD and amyloid-β [9]. Only one gene (APOE) has been repeatedly shown as a risk or protective factor for sporadic AD cases in genetic association studies while other genes have not been replicated by subsequent studies [10]. While the lack of replicable gene associations may be explained by gene-gene, gene-environment, or epigenetic interactions, it is also important to consider the significant heterogeneity among AD phenotypes for which different genes may be more or less important [11, 12]. Furthermore, due to the late onset there is also an increased chance of comorbidities that interact with AD to create an even more unique clinical phenotype in each patient [13]. If each patient is afflicted with their own version of AD, what might treatment look like?

1.1.2 Personalised treatment of AD

Tailoring medical interventions to each individual patient is evidently good. But the degree to which it can be tailored is limited by our understanding of the patient and the disease, as well as what data that can be measured. An example would be the blood thinning medication Wafarin, which is very effective but also hard to dose appropriately. Two genes – CYP2C9 and VKORC1 – were found to account for 35-50% of dosing variability [14]. In order to utilise the finding we both need to measure what gene variants the patient has with gene sequencing technology, and we need to know how to implement that information into a model, here a dosing algorithm. What models are used for AD, and what data is needed to use them? As was mentioned before, previous hypotheses and models of AD, mainly focused around amyloid-β, are currently being reevaluated due to failure in clinical trials [15, 16]. In this project I have therefore chosen to conveniently sidestep this issue by evaluating if metabolism modelling can distinguish those with AD from non-cognitively impaired (NCI) individuals. Metabolism modelling of AD could subsequently aid in personalising treatment to the patient by evaluating the effect of medicine or intervention to them specifically, taking into account AD phenotype and comorbidities.

1.1.3 AD metabolism

There are also further reasons to investigate metabolism in relation to AD. At an early stage in the disease, studies have found impairments to glucose metabolism [17, 18] and proteostasis abnormalities [19, 20, 21] likely caused by oxidative stress [22]. Additionally, impaired lipid metabolism has been identified as a contributing factor to AD [23]. Some of these symptoms can even be detected in brain peripheral tissue, which may be a potential method of diagnosis [24]. If the models in this project are accurate, we should see these changes reflected in our results. But how will the models be constructed?

1.1.4 Modelling metabolism with GEMs

What is a GEM?

On a micro-level, metabolism is the way a cell picks up nutrients from the blood and uses them to produce proteins and to sustain itself, while sending any waste back into the bloodstream. But cells have different needs depending on their specialisation and purpose, location in the body, and signals from other cells, all of which changes across time. A versatile model of cell metabolism should have the potential to represent any one of them, even if it is only a snapshot.

In the simplest terms, the process described above is just a bunch of reactions. Through their shared metabolites the reactions create the metabolic network, where the products of one reaction are used as the substrate of another. Luckily for us, in the body almost every reaction is catalysed by an enzyme, which is translated from mRNA which in turn is transcribed from a gene. Hence a model with all

the metabolic genes should contain all of human metabolic potential; it would be a genome-scale metabolic model (GEM). Metabolic genes can be found in the genome by consulting enzyme databases, but reactions without enzymes – such as spontaneous reactions – also need to be implemented into the model. The same goes for information on enzyme kinetics, which compartment in the cell the reaction occurs in, etc. [25].

How are context-specific GEMs constructed?

In most practical cases, all of the metabolic network is not in use at the same time. In order to analyse metabolic conditions in a sample, we need to measure what reactions are present in the metabolic network to create a GEM specific to that context – whether that context is the type of tissue, the individual, or a disease. By measuring what enzymes are present, or inferring that information from gene expression, it is possible to estimate what parts of the metabolic network are available and functional, as well as the throughput of those parts [25, 26, 27]. From modelling the entire metabolic system, we are also able to see effects that might be far up- or downstream from any enzyme dysfunction, or interactions between several changes to gene expression. Such effects would be hard do discover from looking at the individual gene or enzyme.

While constructing GEMs based on the aggregate results of several samples is common, the same methods are applicable to individual samples. In such a way the GEM can be tailored to represent tissue in an individual patient, and the effects of treatments can be tested *in silico*, making them suitable for personalised and precision medicine [28, 29, 30].

1.2 Aim

In the thesis I aim to take post-mortem bulk brain (dorsolateral prefrontal cortex) gene expression (RNA-seq) data from the ROSMAP study [31] to build GEMs. A model will be constructed from a single tissue sample from each participant, creating a single-sample GEM (ssGEM). The models will subsequently be used to investigate associations between it and the clinical phenotype of the individual. I will be looking at differences between AD, minor and no cognitive impairment (MCI, NCI) models to see if they can be distinguished from each other. In contrast to a GEM constructed from the average of several samples, ssGEMs could potentially show heterogeneity inside each group of AD, MCI, and NCI, which could be analysed to see if there are patterns or clusters within the groups. This study will evaluate the viability of such analysis as there have not been previous ssGEM AD studies.

2
Theory

2.1 Normalising count data

Over the last 20 years DNA and RNA sequencing technology has developed rapidly and it is now possible to sequence at low cost and large scale. Commonly the new methods that have enabled this development are grouped under the term next-generation sequencing. RNA-seq is a workflow that uses next-generation sequencing to obtain estimates of the relative abundance of RNA transcripts in a sample [32]. The level of an RNA transcript is both a direct indication of how much the corresponding gene is expressed, and in the case of mRNA it can be a proxy for protein abundance [33]. The level of expression is commonly referred to as the 'count' or number of 'reads' for that gene or transcript. But before the data can be used in downstream applications, it needs to be normalised.

2.1.1 TPM

In RNA-seq each read represents a 25 bp fragment being matched to a gene. As such, longer genes will have more reads [34]. Additionally, due to differences in experimental design (mainly sequencing depth) as well as both natural and technical variance, total transcript abundance – the sum of all counts or reads – varies between samples [35]. To be able to compare measurements within and between samples, they need to be normalised to some common standard. Originally converting to reads per kilobase million (RPKM) was a common way to normalise for length and abundance, but it is gradually being superseded by transcripts per million (TPM) normalisation [34, 36]. The following formula summarises TPM normalisation for a gene i in a sample:

$$\mathrm{TPM}_i = \frac{\mathrm{RPK}_i}{\sum \mathrm{RPK}_i} 10^6, \ \mathrm{RPK}_i = \frac{\text{reads}}{\text{gene length}}$$

Step-by-step, TPM normalisation entails:

1. Dividing the reads of each gene with its length to obtain the reads per kilobase (RPK)
2. Calculating the fraction of each gene in the sample
3. Multiplying with 1 million to obtain per million counts (for readability).

A caveat to the method above is that RNA is only composed of intronic regions (non-coding regions or non-transcribed regions) of a gene, making it shorter than

gene length. Hence transcript length – the combined length of all intronic regions – is used in practice.

2.1.2 Quantile normalisation

As will be described in the results, the distribution of gene counts in samples diverged significantly after TPM normalisation – even within a single batch of samples. mRNA values will typically follow a log-normal distribution, and we would expect samples to be drawn from the same or similar distributions as a minority of genes are differentially expressed [37]. To correct for the divergence, which is assumed to be from technical variation, quantile normalisation (QN) was applied. The goal of QN is quite simple: making the gene counts of each sample follow the same statistical distribution. In QN, counts are normalised based on the their rank within their sample, such that the count of the n:th most expressed gene in each sample is set to the average across the samples. Afterwards the rank of genes in the sample will be the same as before normalisation, but quantiles (e.g. max and min, 25th and 75th, median, etc.) are now identical across samples [38, 39].

2.2 Modelling metabolism with GEMs

2.2.1 How does a GEM work?

A cell must convert nutrients from the blood into energy and materials to survive. By looking at the reactions that occur along the way – metabolism – we can simplify a complex real world process into a network of reactions connected by their substrates and products (metabolites). Since almost every reaction in the body needs an enzyme, and every enzyme is a protein, every reaction will correspond to one or more genes. Figure 2.1 shows the citrate cycle as a metabolic network where the nodes are metabolites and the edges are enzymes. Thanks to the combined work of several studies and projects, these networks are available for human metabolism as well as other model organisms [40, 41, 42]. Genome-scale metabolic models (GEMs) are similar to figure 2.1 but more extensive. Genome-scale means that they are no longer a 'local' map, but a 'global' map that encompasses the entire metabolic network by including every enzyme. GEMs will also model the cell compartments and transport reactions between them, so not every metabolite is available to the entire cell at once. In downstream analysis attempting to estimate the "use" of or flux through each reaction GEMs are considered steady-state models: they assume that no metabolite accumulates in the cell. Every metabolite either enters or exits the cell as a boundary condition or is consumed by a psuedo-reaction such as biomass synthesis – i.e a sink in the model which simulates accumulation by removing metabolites from the system [43, 44].

But of course not every tissue or cell can do all metabolic tasks, and to model the metabolism of a tissue in an individual we need a way to estimate the network for them and that tissue specifically: a context-specific GEM. As was mentioned, there is a relation between gene, protein (enzyme), and reaction, which is formalised

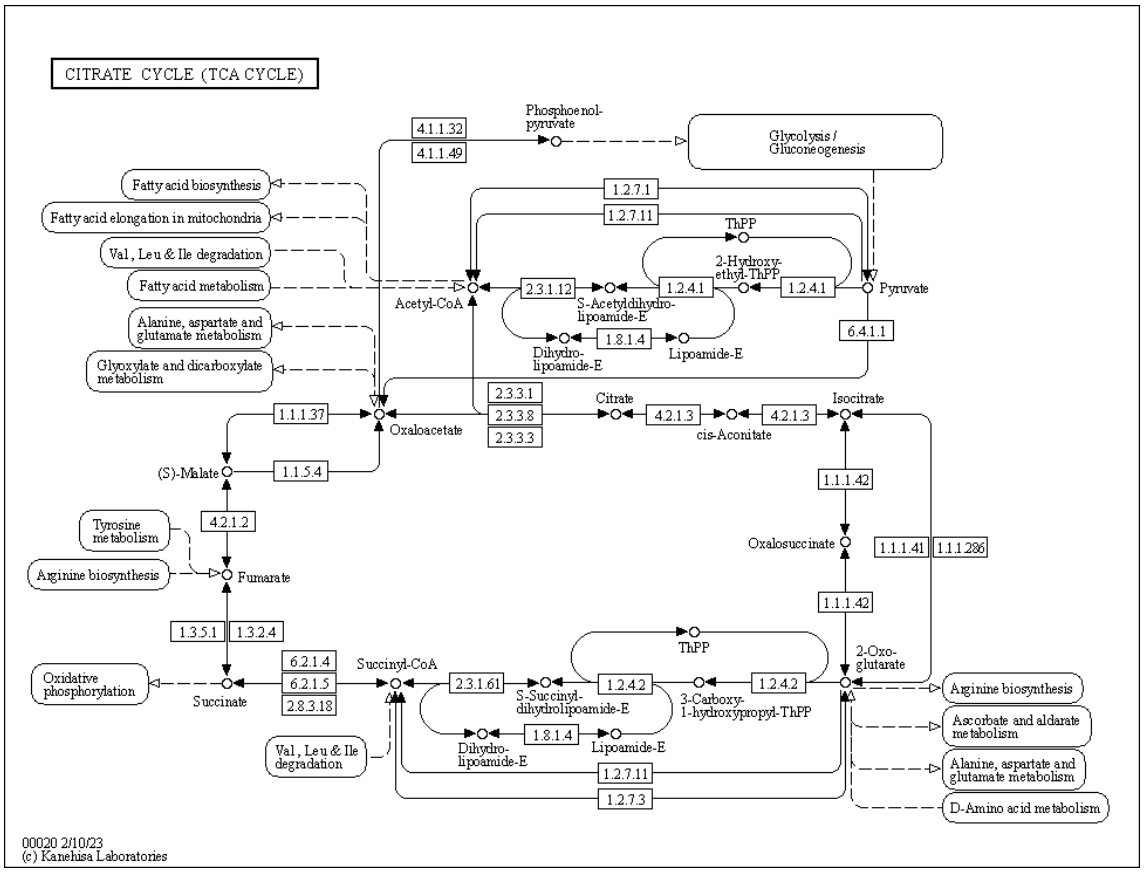


Figure 2.1: Reference pathway map of the citrate cycle (map00020) taken from KEGG [40, 41, 42]. The map is not specific to any one organism. Labels on the edges signify an enzyme that can catalyse a reaction that converts one metabolite into the other.

as the gene-protein-reaction (GPR) association. Due to cases where reactions could be catalysed by one enzyme or another, and cases where the enzyme is a complex of several proteins, GPR associations are not 1:1:1 but more accurately described by logical statements. Take an imaginary reaction R that could be catalysed by two similar enzyme complexes Rα and Rβ. The complexes share 2 proteins but require a different third protein. The logical statement to determine if reaction R is present in the network could be written as a gene rule:

(gene A and gene B and gene C) or (gene A and gene B and gene D)

If these gene rules are known we can determine what enzymes are likely present by measuring the gene expression in a tissue sample and evaluating the gene rule [45]. Since gene rules are included in most GEMs, it is not necessary to construct a context-specific GEM from scratch. Instead a reference GEM is tailored by using the expression data to determine what reactions to keep.

2.2.2 GEM construction algorithm

This project will use the model 'Human1' as a reference [46]. From the model single-sample GEMs (ssGEM) are constructed by filtering for reactions that are supported in the ROSMAP expression data. More specifically the ssGEMs will be constructed according to the 'ftINIT'-algorithm from the RAVEN MATLAB package [47, 48], which can be roughly described by the following steps:

1. Choose a reference model
2. Assign a score to each reaction by comparing expression data of the associated genes against a threshold. If the genes exceed the threshold it is given a positive score, and a negative score is given to genes below the threshold.
3. Exclude reactions from the reference model such that the summed score of remaining reactions is maximal, but still making sure that the final model has no isolated reactions or dead ends.
4. Test if the model can perform essential tasks for cell growth and survival. If not, add the needed reactions while still fulfilling step 3.

Because the GEMs are assumed to be steady-state, the dead ends mentioned in step 3 are useless since no metabolite is allowed to accumulate at or can be used from that dead end. If a part of the model cannot be used, it should not be included. The same goes for isolated reactions.

When scoring the reactions in step 2 the threshold is set to the average TPM value of that gene across all samples, hence why it is so important to normalise the values before construction. Gene expression data is given a score based on the logarithmic fold change against the threshold, so a gene twice as expressed as the threshold would get a score proportional to $\ln(2) \approx +0.69$, while one half as expressed would get a score proportional to $\ln(\frac{1}{2}) \approx -0.69$. The scores are also bounded such that:

$$-1 \leq \ln(\text{fold change}) \leq 2$$

After performing all the scoring, summing, and task-testing, the result of the algorithm is a 'structural GEM': a road map of the system's metabolism. But much like a road map, it only contains qualitative information – what pathways can or cannot be used – and it does not contain quantitative information, e.g. what pathways are more or less used. It is possible to estimate the usage with methods such as flux balance analysis, but we can also discern a lot about the similarity between models from only the structural information.

2.3 Clustering and comparison of GEMs

2.3.1 Clustering by distance

To measure the similarity or dissimilarity of two data points, we need to assert a metric – a 'distance' between data points. With a distance measurement we can subsequently group data points which all are close to each other, and look how the data distributes itself between the groups (ex. hierarchical clustering or stochastic neighbour embedding). The structural GEMs can be compared by their Hamming distance as they can mapped to a binary array: 1 if a reaction is present, 0 if it is not. Two models that are identical except for one reaction would be 1 step apart, identical except for two reactions would 2 steps apart, and so on [49, 50]. Through the binary arrays we can also use variance analysis such as PCA to cluster the models.

2.3.2 Subsystem coverage comparison

In addition, we can compare subsystem coverage in the models. Similar to the concept of a metabolic pathway, a metabolic subsystem is a collection of reactions aimed at a specific task or several related tasks [51]. Subsystem coverage is based on the number of reactions present from that subsystem, and if coverage is worse than the control group – those with no cognitive impairment (NCI) or those with no plaques – the functioning of that subsystem is likely impaired [46, 50, 51]. By averaging coverage across the groups and comparing with NCI, we can infer what subsystems are likely impaired in those with minor cognitive impairment (MCI) and AD.

3
Results

3.1 Data analysis

3.1.1 Excluding samples

The data acquired from ROSMAP was analysed in order to detect missing critical metadata or other problems in samples. Table 3.1 shows a breakdown of excluded data and reason for exclusion. The metadata considered critical was the AD diagnosis of the participant, their age at death, and the quality (RNA integrity number) of the sample. Overall 63 of the original 639 samples were discarded. Additionally, of the 60,725 genes in the data set 118 genes were excluded from the subsequent analysis as we lacked their transcript length – which is needed for normalisation – and 38,025 genes were pre-filtered due to low expression (less than 10 reads in a majority of samples before normalisation). For the most part samples were annotated with all critical metadata and as can be seen from table 3.1 the largest exclusion of samples was due to participants having other cognitive impairments (CI). Not only do we not know what other kinds of CI affected these participants, but these cohorts are also small leading to less conclusive results. To streamline the analysis these were excluded.

3.1.2 Association of metadata

Each sample was annotated with a lot of associated metadata. All metadata that was the same across all samples after exclusion was discarded and the annotations that varied were kept. The annotations that varied and what they measure are explained in table 3.2. Association between metadata in the samples was calculated based on methods by Yoon et al. [55] and is shown in a heat-map in figure 3.1. Associated annotations provide (statistical) information about each other. For example, if a participant last visited at age 85 we immediately know that they could not have

Reason for exclusion	Amount
Sample duplication	1
Lacking critical metadata	7
Diagnosed with non-AD cognitive impairment	55

Table 3.1: Number of samples excluded from the data set and reason for exclusion

Annotation	Explanation
cts_mmse30_lv	MMSE cognitive test result at last visit
cogdx	Clinical consensus diagnosis of cognitive status at time of death
dcfdx_lv	Clinical diagnosis of cognitive status at last visit
braaksc	Semiquantitative measure of severity of tau protein pathology
ceradsc	Semiquantitative measure of amyloid-β plaques
ceradsc_binary	Simplification of ceradsc into a binary (AD or not) classification
msex	Male/Female
age_at_visit_max	Age at last visit
age_death	Age at death
race	Self-reported race
spanish	Self-reported Spanish/Hispanic/Latino origin
Study	Whether patient is from ROS or MAP study
educ	Years of education
apoe_genotype	APOE genotype
pmi	Postmortem interval
PCT_PF_READS_ALIGNED	Percentage of reads in sample that aligned to human DNA; Picard metric
N_unmapped	Number of unmapped reads in sample; STAR metric
RIN	RNA integrity number
Batch	Sample sequencing batch
PCT_RIBOSOMAL_BASES	Fraction of bases in sample aligned to ribosomal RNA; Picard metric
N_multimapping	Number of multimapped reads in sample; STAR metric
N_ambiguous	Number of ambiguously reads in sample; STAR metric
N_noFeature	Number of reads in sample aligning but not mapping to a gene; STAR metric
PCT_CODING_BASES	Fraction of bases in sample aligned to coding regions; Picard metric
PCT_INTERGENIC_BASES	Fraction of bases in sample aligned to intergenic regions; Picard metric
PCT_INTRONIC_BASES	Fraction of bases in sample aligned to intronic regions; Picard metric

Table 3.2: Sample metadata annotations and what they measure[52, 53, 54].

died before the age of 85, and there should be a heavy association between the annotations [56]. An association of 1 gives complete information while a 0 gives no information. Based on the hierarchical clustering dendrogram in figure 3.1, the annotations can be divided into three groups.

Clinical diagnoses, age, and sex

At the top right of the heat-map the various ways of diagnosing AD are clustered. cogdx is the diagnosis given when considering all information up to the time of death. Because the clinician bases it on all data gathered prior to death it is not surprising that it is heavily associated with cts_mmse30_lv and dcfdx_lv which are pre-mortem assessments of AD. Similarly, ceradsc_binary is just a binning of the 4 ceradsc categories of amyloid-β aggregation into 2 categories which makes them heavily associated. The last way of diagnosing AD is braaksc which doesn't actually diagnose AD but measures tau protein pathology which is thought to be the cause of neuronal degradation in AD [57]. Because the diagnoses methods are less then fully associated, which one is used to analyse the data will affect the results. In recognition of the debated role of amyloid-β and tau proteins in AD, we reasoned that cogdx is a more important diagnosis since it: (1) avoids cases where the person had NCI in life but is post-mortem classified as AD due to finding amyloid-β and tau proteins, and (2) captures the metabolism of MCI participants which is important for early diagnosis and – hopefully in the future – treatment.

Compared to the association between the different diagnoses methods, the association between them and the age of death or age at last visit is not that strong, which is a bit surprising as age is a major risk factor for developing AD [58]. The sex of the participant being clustered with age is expected since men have a lower life expectancy.

Technical metadata

In the lower right of the heat-map are several annotations measuring the collection of RNA, sequencing of the complementary DNA, and mapping of the sequenced DNA to human DNA. Some must be associated, such as number of bases aligned to introns (PCT_INTRONIC_BASES) with the number of bases aligned to genes (PCT_INTERGENIC_BASES) since introns are part of genes. The association between batch and RIN is explained later and shown in figure 3.4, but because RIN is a measure of RNA degradation it should be and is associated with the number of bases aligned with ribosomal RNA (PCT_RIBOSOMAL_BASES) due to it being more stable than mRNA [59]. However, it is surprising that PCT_RIBOSOMAL_BASES is more associated with batch than RIN. A strong but expected association can be seen between number of reads aligned to human DNA (PCT_PF_READS_ALIGNED) with the number of reads that did not map onto a gene (N_unmapped) since N_unmapped is not given as a fraction of the total and therefore should rise with the total number of reads aligned.

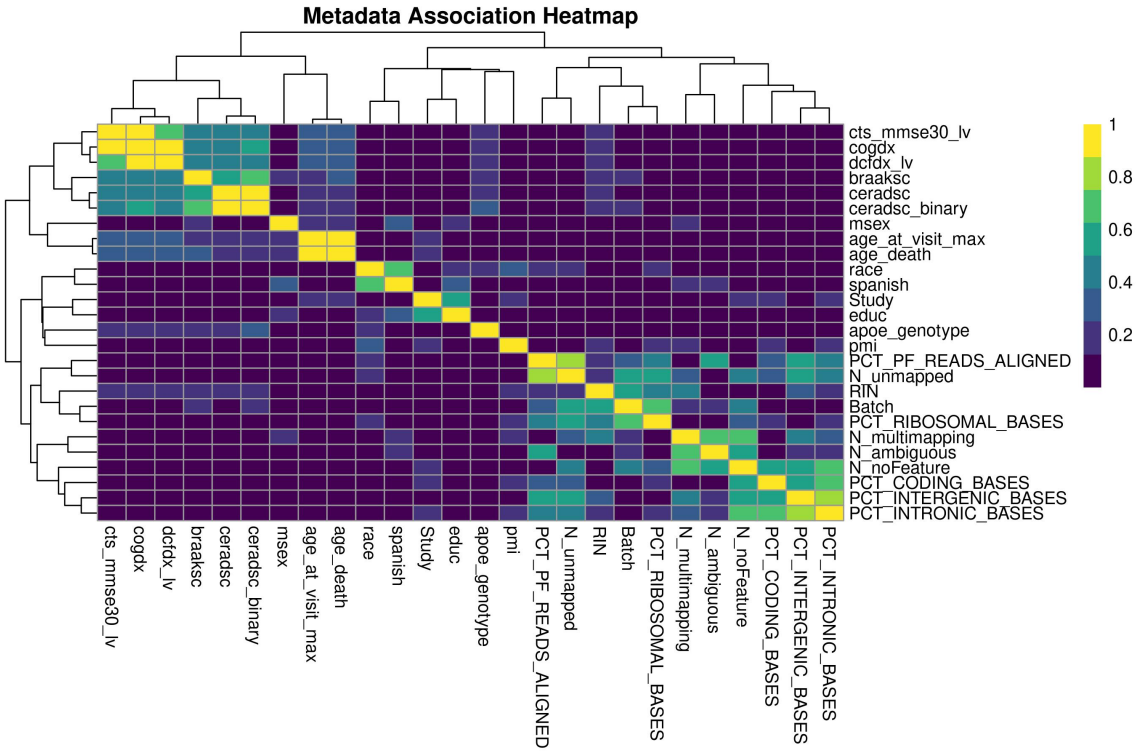


Figure 3.1: Heat-map of association between sample metadata annotations along with a hierarchical clustering dendrogram

Other metadata

In the center of the heat-map are various metadata annotations which are weakly associated to other annotations or not associated at all, with the exception of two pairs of annotations. Years of education – educ – is strongly associated with which study the person belonged to since ROS and MAP have tried to aim at different social groups: while both are from the US, ROS is composed of Catholic nuns, priests, and brothers but MAP tried to get a diversity of participants and aimed for a third of participants to have 12 or fewer years of education [60]. The other pair of strongly associated annotations are self-reported race and Spanish, Hispanic, or Latino origin which is understandable from demographic differences between the US and Spain or South America. Surprisingly there is no strong association between APOE genotype and the various AD diagnoses methods, even though it can significantly alter the risk of getting AD [10].

3.1.3 Correlation of count data to metadata

Just as associations between metadata annotations can tell us about trends in the metadata, we can analyse the relationship between metadata and the RNA-seq count data to see trends. With the help of DeSeq2 [61] principal component analysis (PCA) was performed on the count data and can be seen in figure 3.2. The PCA plots have been annotated with different metadata to see if they correlate with principal component (PC) 1 or 2. Out of the different metadata annotations,

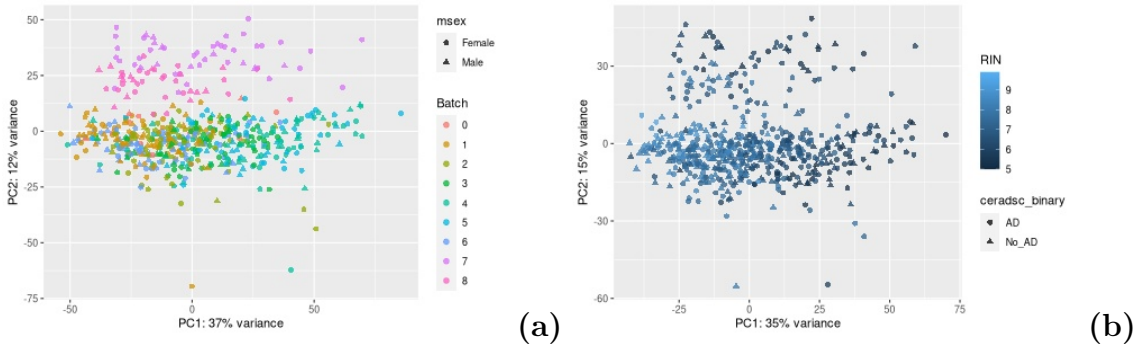


Figure 3.2: Principal Component plots of variance stabilised sample gene count data using only genes that are differentially expressed: (a) showing M/F and batch annotations (b) showing RIN and a post-mortem AD assessment (ceradsc).

only batch and RIN visibly correlated with the PCs as seen in figures 3.2a and 3.2b.

For the batches there is some correlation with PC1, but the batches still overlap a lot on this axis. There is however a clear separation along PC2 of batches 7 and 8 from the rest. RIN does not have a clear separation along PC2, but has a clearly visible correlation with PC1. These results indicate that there is technical (unwanted) variation in the data which might affect downstream analysis. Since batches were separated based on RIN by the experimenters [31] as visualised in figure 3.4 it is hard to know whether the variation is due to RIN or a batch effect. Looking at figure 3.2 it seems likely that variation along PC2 is more of a batch effect since batches 7 and 8 are totally separate but RIN seems to be uncorrelated with PC2.

To avoid any batch effect subsequent analysis and processing of the data was done batch-wise. Batches had roughly 70-80 samples each (with the exception of the deep coverage reference batch 0) and would hopefully still yield significant results even with the smaller size.

Most importantly annotations of AD diagnoses, shown in figures 3.3a and 3.2b, did not correlate to PC1 or 2, which shows us that there is no simple and easy way to differentiate NCI, MCI, and AD on the basis of the data. This is reflected in the literature as genetic association studies are unable to find genes consistently associated with AD (except APOE) [10].

3.2 Normalisation

3.2.1 TPM and gene distributions

Count data was converted into TPM to account for differences in gene length and total reads per sample. As described in section 2.1 RNA-seq results would ideally represent samples drawn from the same gene expression distribution as only a minority of genes are differentially expressed [37]. A couple of the batches are displayed in figures 3.5 (not all batches, see Appendix A), and show right-skewed log-normal

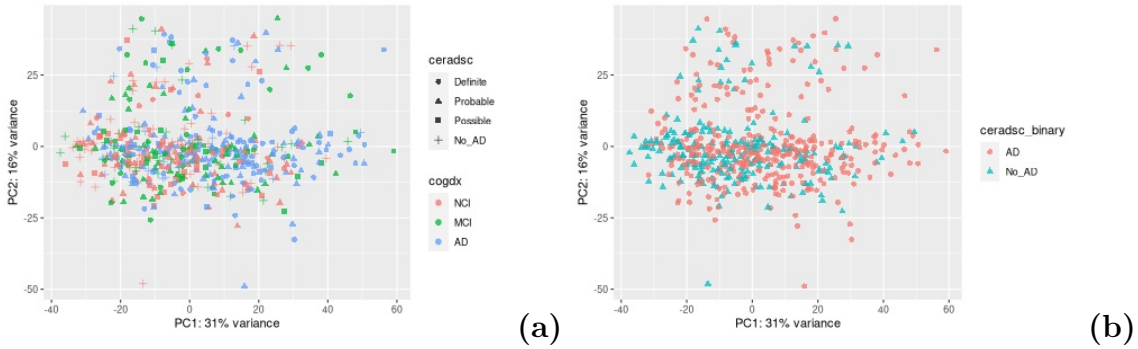


Figure 3.3: Principal Component plots of variance stabilised sample gene count data using only genes that are differentially expressed: (a) showing post-mortem (ceradsc) and peri-mortem (cogdx) assesment (b) showing simplified post-mortem assessment (binary ceradsc).

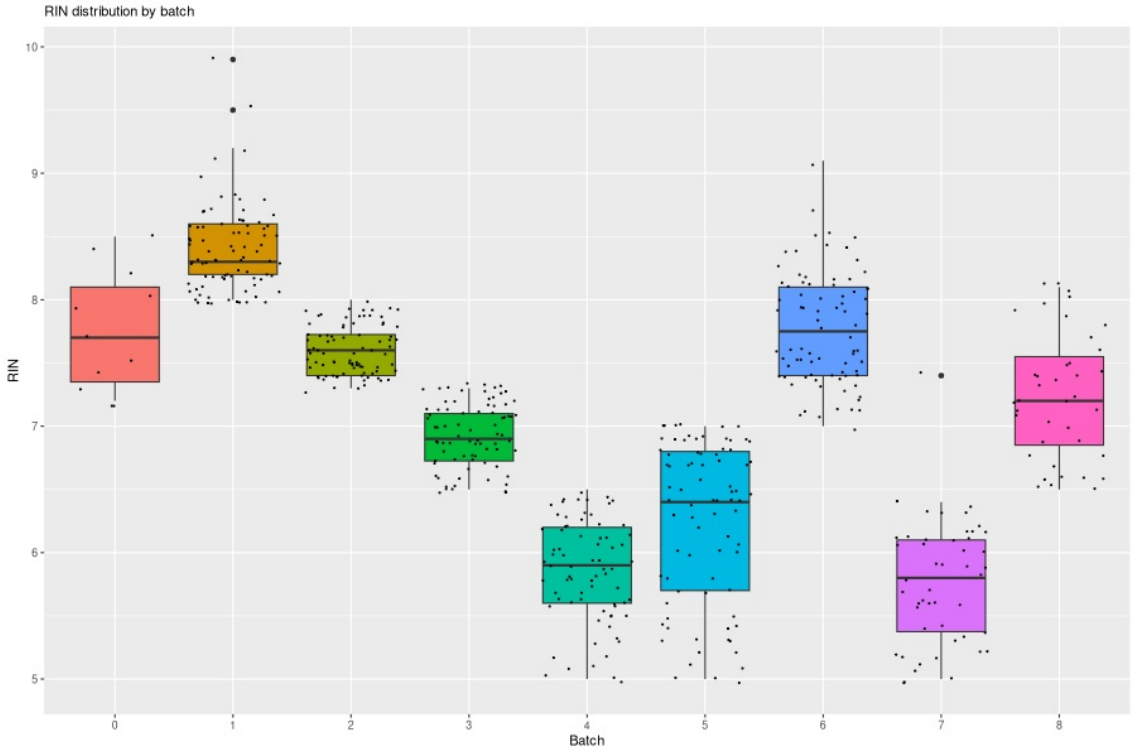


Figure 3.4: Boxplot showing the pooling of RIN in different batches reflecting what is described by experimenters Mostafavi et al. [31].

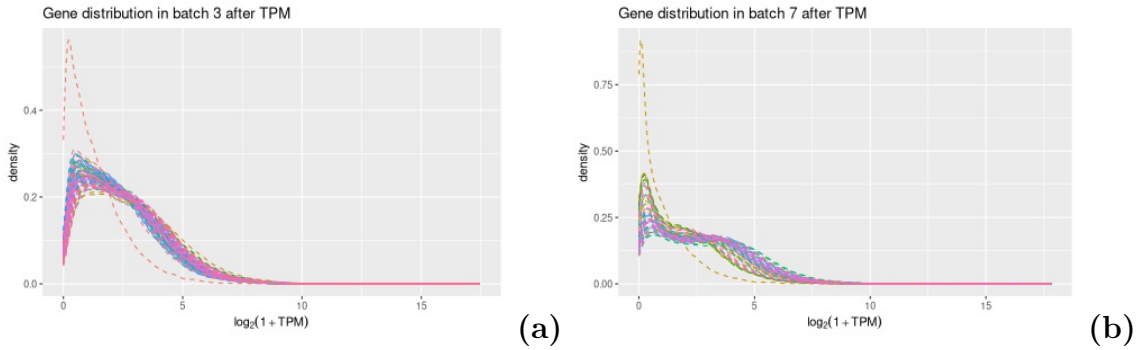


Figure 3.5: Sample gene count distributions after TPM and log-transformed for visualisation, with batches (a) 3 and (b) 7 as examples

distributions: normal distributions when TPM-values are log-transformed, but with long-tails towards higher values. We can also see that the skewness varies between the samples; some samples have a distribution with a taller peak while other samples have a more equal spread over the range of TPM values. So although the samples seem to be drawn from the same type of distribution, the parameters of that distribution diverge between them. Since the threshold in the ftINIT-algorithm (see section 2.2) was set as the average across the samples, it is even more important that samples are directly comparable.

3.2.2 Relative log expression plots and quantile normalisation

Another way of visualising the sample gene distributions is with relative log expression (RLE) plots, which show how much a sample deviates from others [62]. In contrast to gene distributions which don't care which of the genes are highly or lowly expressed, RLE plots compare the expression of a gene against the median expression of that gene and creates box-plots of those differences per sample, reminiscent of error box-plots in regression analysis. As an example, a sample with overall higher expression – mostly positive differences against the median – would have a box with a mean higher than 0, but if some amount of the genes are less expressed than the median – a negative difference – the whiskers of the box will still stretch beneath 0. Figures 3.6a and 3.6b show RLE plots from a couple of batches, and we can see that even in NCI (control) samples deviate from each other. To correct for the divergent distributions quantile normalisation (QN) was performed on each batch. Although QN modifies the the quantities of the data a lot, it has been shown to perform well in comparisons with other common normalisation methods [38].

After QN the gene distributions in a batch are identical, and we can see the effects on the RLE in figure 3.7. Each batch now has the same mean deviation for all its samples, and we can see that some samples deviate from the overall mean more than others, most prominently batch 4 and 7 (excluding batch 0 because of the smaller size). Since batch 4 and 7 had the lowest mean RIN, as shown in figure 3.4, that might be the cause of the deviation. However, batch 5 seems not to deviate

that much even though it also has a slightly lower RIN than other batches.

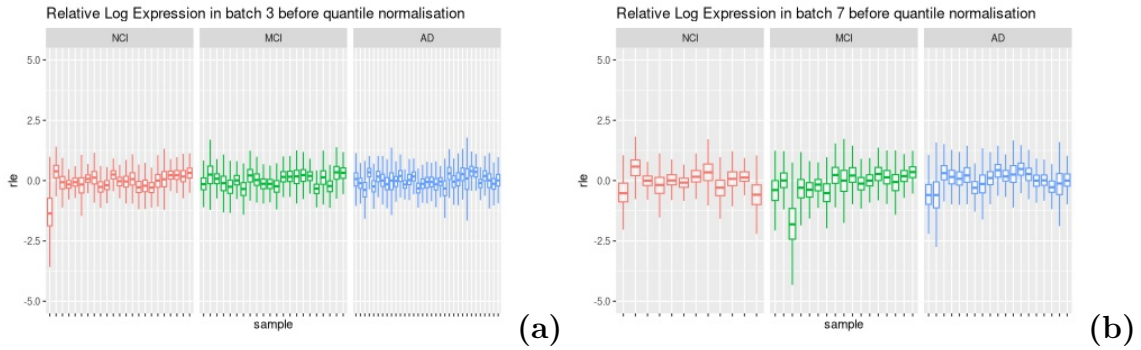


Figure 3.6: RLE plots by premortem AD diagnosis, with batches (a) 3 and (b) 7 as examples

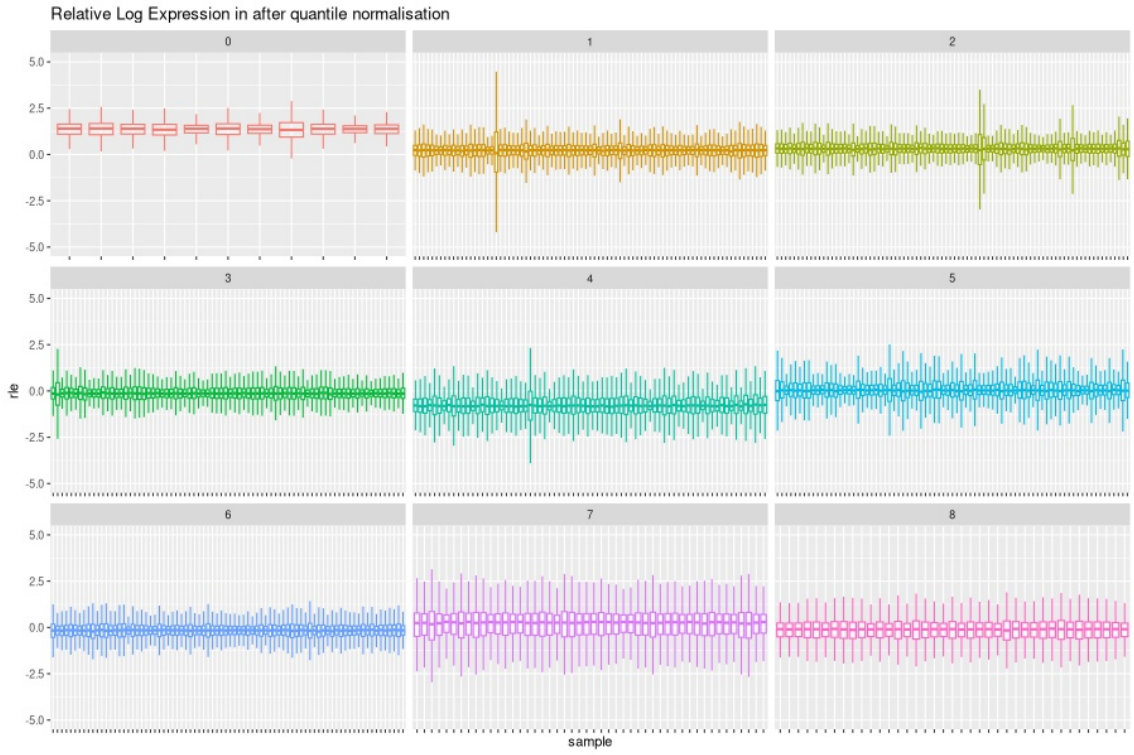


Figure 3.7: RLE plot comparing all batches

3.3 Structural GEM differences

3.3.1 PCA and tSNE clustering

From the normalised data single-sample genome-scale metabolic models (ssGEMs) were constructed according to the ftINIT-algorithm. After construction, the models were analysed based on which reactions had been included and excluded. As described in section 2.3 the reactions of a GEM can be mapped onto a binary array

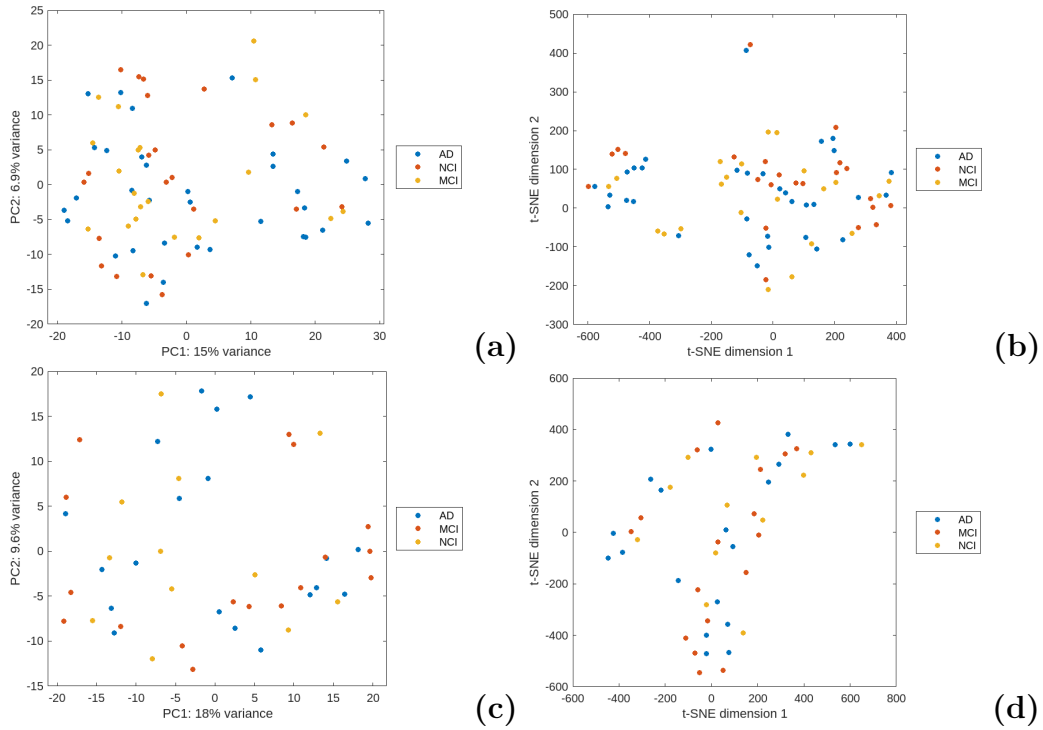


Figure 3.8: PCA and t-SNE plots annotated with premortem AD diagnosis, with batches (a,b) 3 and (c,d) 7 as examples

with a 1 if included and 0 if excluded. Based on the variance in reaction presence and the Hamming distance between arrays, PCA and t-SNE plots of each batch were constructed and a couple are shown in figure 3.8 annotated with the AD diagnosis at death. When looking at these plots, no clustering of AD, MCI, or NCI models can be seen. Because later analysis shows group differences between the models, it is surprising that these cannot be seen in the PCA or t-SNE. Plots were annotated with other metadata as well but no pattern emerged from these either (not shown, see appendix A).

3.3.2 Differences in subsystem coverage

An aggregate measure of the model differences can be obtained from comparing subsystem coverage in the ssGEMs. A subsystem is a part of the metabolic network with a certain function or group of functions, and we assume that when a large number of reactions are absent from that system – i.e. the coverage is low – the functioning of it will be impaired [51]. A heat-map of the subsystem coverage in each group is shown in figure 3.9. The heat-map compares the mean coverage in AD and MCI models against the mean coverage in the control group of NCI models. The subsystems in figure 3.9 are those where coverage in a group deviated from the overall mean of all models by more than 10%.

Many of the subsystems shown recapitulate what was mentioned in the introduction. Lipoic acid is an antioxidant and important cofactor to proteins in mitochondrial energy generation [63, 64]. β-oxidation of fatty acids are an intermediary step for lipids

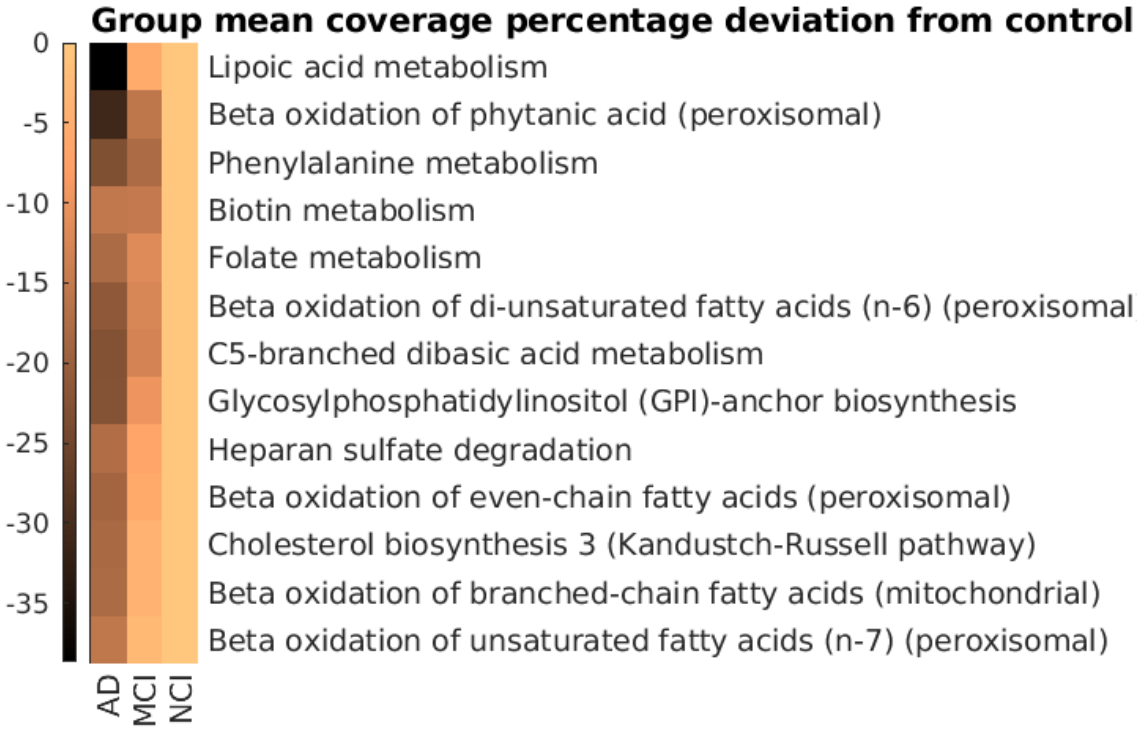


Figure 3.9: Heatmap of the mean subsystem coverage score in each peri-mortem diagnosis compared to the NCI mean.

to be processed into energy [65], and similarly C5-branched dibasic acid metabolism is related to glycolysis and the TCA cycle as well as the metabolism of a number of amino acids, which are of course needed in protein synthesis [66]. Phenylalanine is an essential amino acid [67], and folate metabolism is important for the synthesis of some amino acids as well as maintaining oxidative balance [68]. As for poor coverage of GPI-anchor [69] and cholesterol biosynthesis [70, 71], heparan sulfate degradation [72, 73], and biotin metabolism [74], they have all been directly or tangentially associated with AD in previous studies.

While it is good that the significant findings agree with literature, some consideration should be given to what was not found to be significant. One of the most supported metabolic impacts of AD is glucose hypo-metabolism [75, 76], but few of the subsystems found significant are related to glucose metabolism. Of course the complex interactions of the different subsystems could cause this effect indirectly, but nevertheless. Impairments may also be "falsely" missed since subsystem classification is, to some degree, arbitrary. For example, there is some evidence of impaired glucose transport in AD [77] but these reactions are all part of the 'Transport reactions'-subsystem where they only constitute a small part: 6 of 4187 possible reactions in Human-GEM according to Metabolic Atlas [46, 78]. Even with all transport reactions absent such a difference would not be seen as significant with this method.

3.3.3 Comparison to other *in silico* AD metabolism studies

To determine if these results had been previously found in *in silico* metabolic AD models a literature search of articles with "Alzheimer" and "metabolic model" in their abstract was performed. Not many studies have been performed on *in silico* AD metabolism, and it was necessary that pathways or subsystems were evaluated in the study to compare with the above results. The search yielded 22 results of which 5 were accessible and comparable to the results in this project. Table 3.3 shows the articles and whether their findings match this project. Notably none of the articles find an alteration of C5-branched dibasic acid metabolism, GPI-anchor biosynthesis, and heparan sulphate degradation. Bayraktar et al. show that there is less lipoic acid metabolism coverage in ROSMAP AD and NCI samples – which are taken from the dorsolateral prefrontal cortex (DLPFC) – in comparison to samples from the cerebellum (CBE) and temporal cortex (TCX) , but do not show any difference between AD and NCI [79]. Another point of difference is that MCI models are only considered in Stempler et al. [80] and this project, while there is no discussion of them in the other articles.

	Stempler et al. [80]	Lam et al. [81]	Li et al. [82]	Kim et al. [83]	Bayraktar et al. [79]
MCI modelling	✓				
Lipoic acid metabolism					✓
Fatty acid oxidation (peroxisomal)	✓		✓	✓	
Phenylalanine metabolism	✓		✓		
Biotin metabolism	✓				
Folate metabolism	✓		✓		
C5-branched dibasic acid metabolism					
GPI-anchor biosynthesis					
Heparan sulfate degradation					
Cholesterol biosynthesis 3		✓	✓		
Fatty acid oxidation (mitochondrial)		✓	✓	✓	

Table 3.3: Comparing the subsystem coverage findings to findings in other *in silico* metabolic AD modelling articles

4

Conclusion

This project looked at single-sample genome-scale metabolic models (ssGEMs) for modelling of Alzheimer's Disease (AD) metabolism. As could be seen in the results there have not been many *in silico* metabolic AD models in general, and to my knowledge none have been constructed from individual samples. The results from the subsystem coverage analysis in this study and various analyses by the other comparable studies show that metabolic differences can be found between AD, MCI, and NCI *in silico* models.

4.1 Subsystem coverage

Subsystem coverage measured in the AD and MCI ssGEMs was significantly lower than NCI models in several subsystems, all of which had support in the literature. MCI models also had a coverage in-between AD and NCI which lends additional support to the findings. The decreased coverage of lipoic acid metabolism, GPI-anchor biosynthesis, and heparan sulfate degradation had not been found in comparable studies, but have been the main subject of other AD articles. C5-branched dibasic acid metabolism had not been found in the comparable studies, and has not been the highlighted much in other AD articles, though I suspect the reason to be that its function is too vague to have been of interest. Perhaps that is even more of a sign that it should be further investigated in relation to AD.

4.2 MCI modeling

Regarding the MCI models, it was surprising to find that it was so rare to model this in-between step of AD. Many of the studies try to find metabolomic markers for AD or therapeutic targets, which would be much more important at an early stage of AD, hopefully even before MCI. Granted, many studies may not want to use pre-mortem assessments to determine sample diagnosis since post-mortem assessments are more objective, but perhaps a similar early-stage AD category can be used for the post-mortem assessment as well. To fully understand AD, modelling the disease at different stages of progression seems essential and I hope that future studies are encouraged to explore this option.

4.3 Clustering and normalisation

The project was not able to identify AD or MCI in individual models based on clustering of variance in the structural information (presence or lack of metabolic reactions) or Hamming distances between models. As group differences were shown in subsystem coverage, the lack of clustering indicates that other variance within the groups is overshadowing this difference. Here it worth examining whether the choice of normalisation and analysis could have been improved. As can be seen from the post-QN RLE plot in figure 3.7, some samples are likely outliers and will affect the threshold described in section 2.2. Another potential issue is the batch sizes. In a batch each group of NCI, MCI, and AD contains around 10-20 samples, which is possibly too small to highlight the similarities between them. Combining all batches but removing those batches that deviated significantly might be a better choice. Alternatively samples could have been pooled into bigger groups based on RIN. Regardless, this exploration of single-sample GEMs does not substantiate their usefulness in AD metabolism research.

4.4 Gene filtration

Another aspect that could be improved is the gene filtration. The main reason for filtering out low expression genes is to speed up downstream analysis, but the exclusion criteria still leaves us with roughly 20,000 genes. The total amount of metabolic genes (genes in Human1) is 2920 so simply choosing these and discarding the rest would achieve this goal while also giving a PCA that is more representative of the data that are later used in model construction, and the same goes for TPM distribution and RLE plots. It also eliminates the chance of a gene being highly expressed in a small cohort or single sample but being discarded due to very low expression in the rest of the samples. That will not matter for most models, but seems more suited to the single-sample nature of the project.

4.5 Future use

If the models were to be further analysed in the future, the next step would be to perform flux balance analysis (FBA) on the models to evaluate if the impaired subsystems induce energy hypo-metabolism i.e. lowered ATP generation. Through such analysis we would be able to see the aggregate effects of all the impaired subsystems, at least when it comes to that one aspect of AD metabolism. Since FBA maximises some flux, it would likely not be suitable for targets such as proteostasis or oxidative stress where a balance between several fluxes is important. Here random sampling of flux values seems more appropriate to show what effects the impaired subsystems have on them.

Bibliography

[1] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, and M. Roser, "Life Expectancy," *Our World in Data*, Dec. 2023.

[2] W. Pedersen, "Senile Dementia," in *xPharm: The Comprehensive Pharmacology Reference* (S. J. Enna and D. B. Bylund, eds.), pp. 1–18, New York: Elsevier, Jan. 2007.

[3] H. Ritchie and M. Roser, "Age Structure," *Our World in Data*, Dec. 2023.

[4] "Global burden of disease study (2019) – processed by our world in data," tech. rep., IHME, 2019.

[5] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. R. Menezes, E. Rimmer, M. Scazufca, and Alzheimer's Disease International, "Global prevalence of dementia: a Delphi consensus study," *Lancet (London, England)*, vol. 366, pp. 2112–2117, Dec. 2005.

[6] D. Avramopoulos, "Genetics of Alzheimer's disease: recent advances," *Genome Medicine*, vol. 1, p. 34, Mar. 2009.

[7] "Consensus Recommendations for the Postmortem Diagnosis of Alzheimer's Disease," *Neurobiology of Aging*, vol. 18, pp. S1–S2, July 1997.

[8] R. van der Kant, L. S. B. Goldstein, and R. Ossenkoppele, "Amyloid--independent regulators of tau pathology in Alzheimer disease," *Nature Reviews Neuroscience*, vol. 21, pp. 21–35, Jan. 2020. Number: 1 Publisher: Nature Publishing Group.

[9] E. Karran, M. Mercken, and B. D. Strooper, "The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics," *Nature Reviews Drug Discovery*, vol. 10, pp. 698–712, Sept. 2011. Number: 9 Publisher: Nature Publishing Group.

[10] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database," *Nature Genetics*, vol. 39, pp. 17–23, Jan. 2007. Number: 1 Publisher: Nature Publishing Group.

[11] B. C. Dickerson, S. M. McGinnis, C. Xia, B. H. Price, A. Atri, M. E. Murray, M. F. Mendez, and D. A. Wolk, "Approach to atypical Alzheimer's disease and case studies of the major subtypes," *CNS Spectrums*, vol. 22, pp. 439–449, Dec. 2017. Publisher: Cambridge University Press.

[12] J. Graff-Radford, K. X. X. Yong, L. G. Apostolova, F. H. Bouwman, M. Carrillo, B. C. Dickerson, G. D. Rabinovici, J. M. Schott, D. T. Jones, and M. E. Murray, "New insights into atypical Alzheimer's disease in the era of biomarkers," *The Lancet Neurology*, vol. 20, pp. 222–234, Mar. 2021.

[13] J. X. Hu, C. E. Thomas, and S. Brunak, "Network biology concepts in complex disease comorbidities," *Nature Reviews Genetics*, vol. 17, pp. 615–629, Oct. 2016. Number: 10 Publisher: Nature Publishing Group.

[14] M. T. M. Lee and T. E. Klein, "Pharmacogenetics of warfarin: challenges and opportunities," *Journal of Human Genetics*, vol. 58, pp. 334–338, June 2013. Number: 6 Publisher: Nature Publishing Group.

[15] K. Herrup, "The case for rejecting the amyloid cascade hypothesis," *Nature Neuroscience*, vol. 18, pp. 794–799, June 2015. Number: 6 Publisher: Nature Publishing Group.

[16] M. Tolar, S. Abushakra, and M. Sabbagh, "The path forward in Alzheimer's disease therapeutics: Reevaluating the amyloid cascade hypothesis," *Alzheimer's & Dementia*, vol. 16, no. 11, pp. 1553–1560, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2019.09.075.

[17] E. Croteau, C. A. Castellano, M. Fortier, C. Bocti, T. Fulop, N. Paquet, and S. C. Cunnane, "A cross-sectional comparison of brain glucose and ketone metabolism in cognitively healthy older adults, mild cognitive impairment and early Alzheimer's disease," *Experimental Gerontology*, vol. 107, pp. 18–26, July 2018.

[18] C. M. Weise, K. Chen, Y. Chen, X. Kuang, C. R. Savage, E. M. Reiman, and Alzheimer's Disease Neuroimaging Initiative, "Left lateralized cerebral glucose metabolism declines in amyloid- positive persons with mild cognitive impairment," *NeuroImage. Clinical*, vol. 20, pp. 286–296, 2018.

[19] V. Bonet-Costa, L. C.-D. Pomatto, and K. J. A. Davies, "The Proteasome and Oxidative Stress in Alzheimer's Disease," *Antioxidants & Redox Signaling*, vol. 25, pp. 886–901, Dec. 2016.

[20] J. N. Keller, K. B. Hanni, and W. R. Markesbery, "Impaired Proteasome Function in Alzheimer's Disease," *Journal of Neurochemistry*, vol. 75, no. 1, pp. 436–439, 2000. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1471-4159.2000.0750436.x.

[21] B. P. Tseng, K. N. Green, J. L. Chan, M. Blurton-Jones, and F. M. LaFerla, "Abeta inhibits the proteasome and enhances amyloid and tau accumulation," *Neurobiology of Aging*, vol. 29, pp. 1607–1618, Nov. 2008.

[22] D. A. Butterfield and B. Halliwell, "Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease," *Nature Reviews Neuroscience*, vol. 20, pp. 148–160, Mar. 2019. Number: 3 Publisher: Nature Publishing Group.

[23] Y. Zeng, S. Cao, N. Li, J. Tang, and G. Lin, "Identification of key lipid metabolism-related genes in Alzheimer's disease," *Lipids in Health and Disease*, vol. 22, p. 155, Sept. 2023.

[24] K. Leuner, K. Schulz, T. Schütt, J. Pantel, D. Prvulovic, V. Rhein, E. Savaskan, C. Czech, A. Eckert, and W. E. Müller, "Peripheral Mitochondrial Dysfunction in Alzheimer's Disease: Focus on Lymphocytes," *Molecular Neurobiology*, vol. 46, pp. 194–204, Aug. 2012.

[25] L. Liu, R. Agren, S. Bordel, and J. Nielsen, "Use of genome-scale metabolic models for understanding microbial physiology," *FEBS Letters*, vol. 584, no. 12, pp. 2556–2564, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2010.04.052.

[26] J. D. Orth, I. Thiele, and B. Palsson, "What is flux balance analysis?," *Nature Biotechnology*, vol. 28, pp. 245–248, Mar. 2010. Number: 3 Publisher: Nature Publishing Group.

[27] N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson, "Genome-scale microbial in silico models: the constraints-based approach," *Trends in Biotechnology*, vol. 21, pp. 162–169, Apr. 2003.

[28] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen, "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling," *Molecular Systems Biology*, vol. 10, p. 721, Mar. 2014. Publisher: John Wiley & Sons, Ltd.

[29] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen, "Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT," *PLOS Computational Biology*, vol. 8, p. e1002518, May 2012. Publisher: Public Library of Science.

[30] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Current status and applications of genome-scale metabolic models," *Genome Biology*, vol. 20, p. 121, June 2019.

[31] S. Mostafavi, C. Gaiteri, S. E. Sullivan, C. C. White, S. Tasaki, J. Xu, M. Taga, H.-U. Klein, E. Patrick, V. Komashko, C. McCabe, R. Smith, E. M. Bradshaw, D. E. Root, A. Regev, L. Yu, L. B. Chibnik, J. A. Schneider, T. L. Young-Pearse, D. A. Bennett, and P. L. De Jager, "A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease," *Nature Neuroscience*, vol. 21, pp. 811–819, June 2018. Number: 6 Publisher: Nature Publishing Group.

[32] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, Jan. 2009. Number: 1 Publisher: Nature Publishing Group.

[33] Y. Liu, A. Beyer, and R. Aebersold, "On the Dependency of Cellular Protein Levels on mRNA Abundance," *Cell*, vol. 165, pp. 535–550, Apr. 2016.

[34] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, pp. 621–628, July 2008. Number: 7 Publisher: Nature Publishing Group.

[35] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and on behalf of The French StatOmique Consortium, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, pp. 671–683, Nov. 2013.

[36] G. P. Wagner, K. Kin, and V. J. Lynch, "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples," *Theory in Biosciences*, vol. 131, pp. 281–285, Dec. 2012.

[37] C. Lu and R. D. King, "An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems," *Bioinformatics*, vol. 25, pp. 2020–2027, 06 2009.

[38] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics (Oxford, England)*, vol. 19, pp. 185–193, Jan. 2003.

[39] "Quantile normalization." `https://en.wikipedia.org/w/index.php?title=Quantile_normalization&oldid=1138433182`, Feb. 2023. Page Version ID: 1138433182.

[40] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, Jan. 2000.

[41] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Science: A Publication of the Protein Society*, vol. 28, pp. 1947–1951, Nov. 2019.

[42] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, "KEGG for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Research*, vol. 51, pp. D587–D592, Jan. 2023.

[43] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 1777–1782, Feb. 2007. Publisher: Proceedings of the National Academy of Sciences.

[44] F. Santos, J. Boele, and B. Teusink, "Chapter twenty-four - A Practical Guide to Genome-Scale Metabolic Models and Their Analysis," in *Methods in Enzymology* (D. Jameson, M. Verma, and H. V. Westerhoff, eds.), vol. 500 of *Methods in Systems Biology*, pp. 509–532, Academic Press, Jan. 2011.

[45] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nature Biotechnology*, vol. 26, pp. 1003–1010, Sept. 2008. Number: 9 Publisher: Nature Publishing Group.

[46] J. L. Robinson, P. Kocabaş, H. Wang, P.-E. Cholley, D. Cook, A. Nilsson, M. Anton, R. Ferreira, I. Domenzain, V. Billa, A. Limeta, A. Hedin, J. Gustafsson, E. J. Kerkhoven, L. T. Svensson, B. O. Palsson, A. Mardinoglu, L. Hansson, M. Uhlén, and J. Nielsen, "An atlas of human metabolism," *Science Signaling*, vol. 13, p. eaaz1482, Mar. 2020. Publisher: American Association for the Advancement of Science.

[47] J. Gustafsson, M. Anton, F. Roshanzamir, R. Jörnsten, E. J. Kerkhoven, J. L. Robinson, and J. Nielsen, "Generation and analysis of context-specific genome-scale metabolic models derived from single-cell RNA-Seq data," *Proceedings of the National Academy of Sciences*, vol. 120, p. e2217868120, Feb. 2023. Publisher: Proceedings of the National Academy of Sciences.

[48] H. Wang, S. Marcišauskas, B. J. Sánchez, I. Domenzain, D. Hermansson, R. Agren, J. Nielsen, and E. J. Kerkhoven, "RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor," *PLOS Computational Biology*, vol. 14, p. e1006541, Oct. 2018. Publisher: Public Library of Science.

[49] A. Cabbia, P. A. J. Hilbers, and N. A. W. van Riel, "A Distance-Based Framework for the Characterization of Metabolic Heterogeneity in Large Sets of

Genome-Scale Metabolic Models," *Patterns*, vol. 1, p. 100080, Sept. 2020.

[50] J. L. Robinson, P. Kocabaş, H. Wang, P.-E. Cholley, D. Cook, A. Nilsson, M. Anton, R. Ferreira, I. Domenzain, V. Billa, A. Limeta, A. Hedin, J. Gustafsson, E. J. Kerkhoven, L. T. Svensson, B. O. Palsson, A. Mardinoglu, L. Hansson, M. Uhlén, and J. Nielsen, "Supplementary data and scripts for "An Atlas of Human Metabolism"," Dec. 2019.

[51] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, "The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes," *Nucleic Acids Research*, vol. 33, no. 17, pp. 5691–5702, 2005.

[52] "Variables | RADC." `https://www.radc.rush.edu/docs/var/variables.htm`, Jan. 2024.

[53] "Picard Metrics Definitions." `https://broadinstitute.github.io/picard/picard-metric-definitions.html`, Jan. 2024.

[54] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan. 2013.

[55] G. Yoon, C. L. Müller, and I. Gaynanova, "Fast Computation of Latent Correlations," *Journal of Computational and Graphical Statistics*, vol. 30, pp. 1249–1256, Oct. 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2021.1882468.

[56] N. Altman and M. Krzywinski, "Association, correlation and causation," *Nature Methods*, vol. 12, pp. 899–900, Oct. 2015. Number: 10 Publisher: Nature Publishing Group.

[57] M. Williams, "Progress in Alzheimer's disease drug discovery: an update," *Current Opinion in Investigational Drugs (London, England: 2000)*, vol. 10, pp. 23–34, Jan. 2009.

[58] S. Gao, H. C. Hendrie, K. S. Hall, and S. Hui, "The Relationships Between Age, Sex, and the Incidence of Dementia and Alzheimer Disease: A Meta-analysis," *Archives of General Psychiatry*, vol. 55, pp. 809–815, Sept. 1998.

[59] Z. Li, S. Reimers, S. Pandit, and M. P. Deutscher, "RNA quality control: degradation of defective transfer RNA," *The EMBO Journal*, vol. 21, pp. 1132–1138, Mar. 2002.

[60] D. A. Bennett, J. A. Schneider, A. S. Buchman, L. L. Barnes, P. A. Boyle, and R. S. Wilson, "Overview and Findings from the Rush Memory and Aging Project," *Current Alzheimer research*, vol. 9, pp. 646–663, July 2012.

[61] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, Dec. 2014.

[62] L. C. Gandolfo and T. P. Speed, "RLE plots: Visualizing unwanted variation in high dimensional data," *PLoS One*, vol. 13, p. e0191629, Feb. 2018.

[63] L. Packer, E. H. Witt, and H. J. Tritschler, "Alpha-lipoic acid as a biological antioxidant," *Free Radical Biology and Medicine*, vol. 19, pp. 227–250, Aug. 1995.

[64] S. S. Hardas, R. Sultana, A. M. Clark, T. L. Beckett, L. I. Szweda, M. P. Murphy, and D. A. Butterfield, "Oxidative modification of lipoic acid by HNE in Alzheimer disease brain," *Redox Biology*, vol. 1, pp. 80–85, Jan. 2013.

[65] M. Edwards and S. S. Mohiuddin, "Biochemistry, Lipolysis," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023.

[66] "KEGG PATHWAY: map00660."

[67] M. J. Lopez and S. S. Mohiuddin, "Biochemistry, Essential Amino Acids," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023.

[68] M. M. Zarou, A. Vazquez, and G. Vignir Helgason, "Folate metabolism: a re-emerging therapeutic target in haematological cancers," *Leukemia*, vol. 35, pp. 1539–1551, June 2021. Number: 6 Publisher: Nature Publishing Group.

[69] K. Sambamurti, D. Sevlever, T. Koothan, L. M. Refolo, I. Pinnix, S. Gandhi, L. Onstead, L. Younkin, C. M. Prada, D. Yager, Y. Ohyagi, C. B. Eckman, T. L. Rosenberry, and S. G. Younkin, "Glycosylphosphatidylinositol-anchored Proteins Play an Important Role in the Biogenesis of the Alzheimer's Amyloid -Protein *," *Journal of Biological Chemistry*, vol. 274, pp. 26810–26814, Sept. 1999. Publisher: Elsevier.

[70] V. R. Varma, H. Büşra Lüleci, A. M. Oommen, S. Varma, C. T. Blackshear, M. E. Griswold, Y. An, J. A. Roberts, R. O'Brien, O. Pletnikova, J. C. Troncoso, D. A. Bennett, T. Çakır, C. Legido-Quigley, and M. Thambisetty, "Abnormal brain cholesterol homeostasis in Alzheimer's disease—a targeted metabolomic and transcriptomic study," *NPJ Aging and Mechanisms of Disease*, vol. 7, p. 11, June 2021.

[71] E. Staurenghi, S. Giannelli, G. Testa, B. Sottero, G. Leonarduzzi, and P. Gamba, "Cholesterol Dysmetabolism in Alzheimer's Disease: A Starring Role for Astrocytes?," *Antioxidants*, vol. 10, p. 1890, Nov. 2021.

[72] G.-l. Zhang, X. Zhang, X.-m. Wang, and J.-P. Li, "Towards Understanding the Roles of Heparan Sulfate Proteoglycans in Alzheimer's Disease," *BioMed Research International*, vol. 2014, p. e516028, July 2014. Publisher: Hindawi.

[73] I. Ozsan McMillan, J.-P. Li, and L. Wang, "Heparan sulfate proteoglycan in Alzheimer's disease: aberrant expression and functions in molecular pathways related to amyloid- metabolism," *American Journal of Physiology-Cell Physiology*, vol. 324, pp. C893–C909, Apr. 2023. Publisher: American Physiological Society.

[74] K. M. Lohr, B. Frost, C. Scherzer, and M. B. Feany, "Biotin rescues mitochondrial dysfunction and neurotoxicity in a tauopathy model," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 33608–33618, Dec. 2020. Publisher: Proceedings of the National Academy of Sciences.

[75] D. Schubert, "Glucose metabolism and Alzheimer's disease," *Ageing Research Reviews*, vol. 4, pp. 240–257, May 2005.

[76] S. Cunnane, S. Nugent, M. Roy, A. Courchesne-Loyer, E. Croteau, S. Tremblay, A. Castellano, F. Pifferi, C. Bocti, N. Paquet, H. Begdouri, M. Bentourkia, E. Turcotte, M. Allard, P. Barberger-Gateau, T. Fulop, and S. I. Rapoport,

"Brain fuel metabolism, aging, and Alzheimer's disease," *Nutrition*, vol. 27, pp. 3–20, Jan. 2011.

[77] L. Xu, R. Liu, Y. Qin, and T. Wang, "Brain metabolism in Alzheimer's disease: biological mechanisms of exercise," *Translational Neurodegeneration*, vol. 12, p. 33, June 2023.

[78] F. Li, Y. Chen, M. Anton, and J. Nielsen, "GotEnzymes: an extensive database of enzyme parameter predictions," *Nucleic Acids Research*, vol. 51, pp. D583–D586, Jan. 2023.

[79] A. Bayraktar, S. Lam, O. Altay, X. Li, M. Yuan, C. Zhang, M. Arif, H. Turkez, M. Uhlén, S. Shoaie, and A. Mardinoglu, "Revealing the Molecular Mechanisms of Alzheimer's Disease Based on Network Analysis," *International Journal of Molecular Sciences*, vol. 22, p. 11556, Oct. 2021.

[80] S. Stempler, K. Yizhak, and E. Ruppin, "Integrating Transcriptomics with Metabolic Modeling Predicts Biomarkers and Drug Targets for Alzheimer's Disease," *PLOS ONE*, vol. 9, p. e105383, Aug. 2014. Publisher: Public Library of Science.

[81] S. Lam, N. Hartmann, R. Benfeitas, C. Zhang, M. Arif, H. Turkez, M. Uhlén, C. Englert, R. Knight, and A. Mardinoglu, "Systems Analysis Reveals Ageing-Related Perturbations in Retinoids and Sex Hormones in Alzheimer's and Parkinson's Diseases," *Biomedicines*, vol. 9, p. 1310, Oct. 2021. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

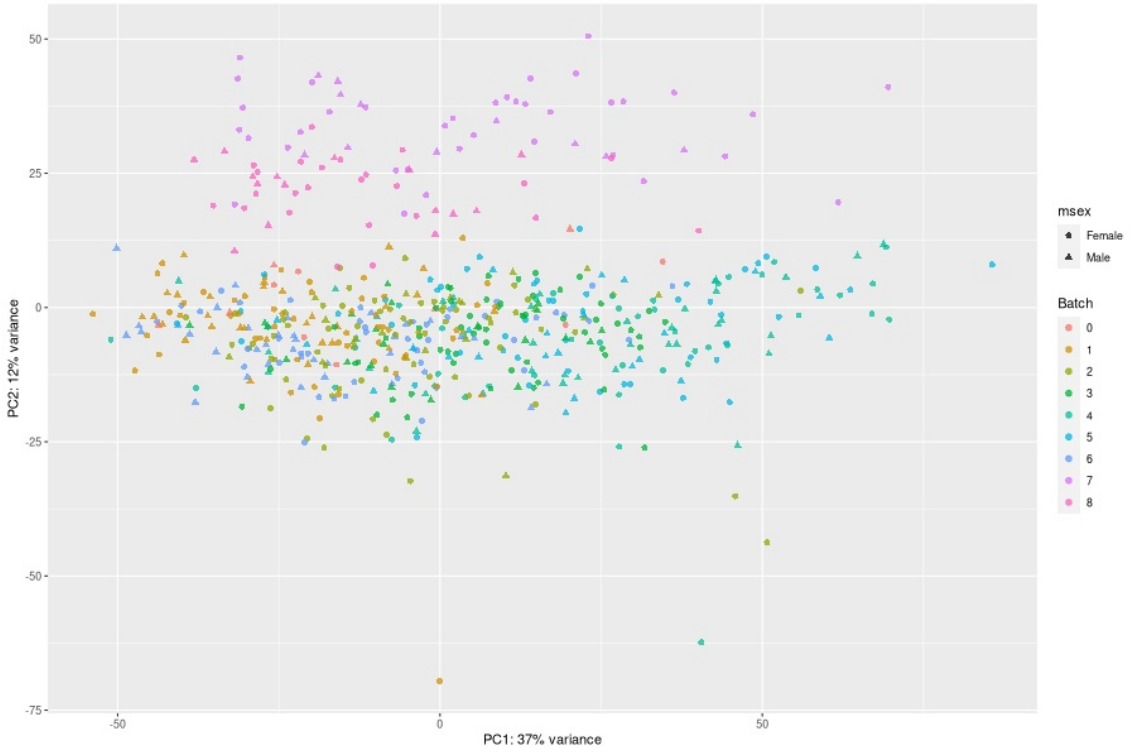
[82] W.-X. Li, G.-H. Li, X. Tong, P.-P. Yang, J.-F. Huang, L. Xu, and S.-X. Dai, "Systematic metabolic analysis of potential target, therapeutic drug, diagnostic method and animal model applicability in three neurodegenerative diseases," *Aging*, vol. 12, pp. 9882–9914, May 2020.

[83] S. Kim, Y. Nam, M.-j. Kim, S.-h. Kwon, J. Jeon, S. J. Shin, S. Park, S. Chang, H. U. Kim, Y. Y. Lee, H. S. Kim, and M. Moon, "Proteomic analysis for the effects of non-saponin fraction with rich polysaccharide from Korean Red Ginseng on Alzheimer's disease in a mouse model," *Journal of Ginseng Research*, vol. 47, pp. 302–310, Mar. 2023.

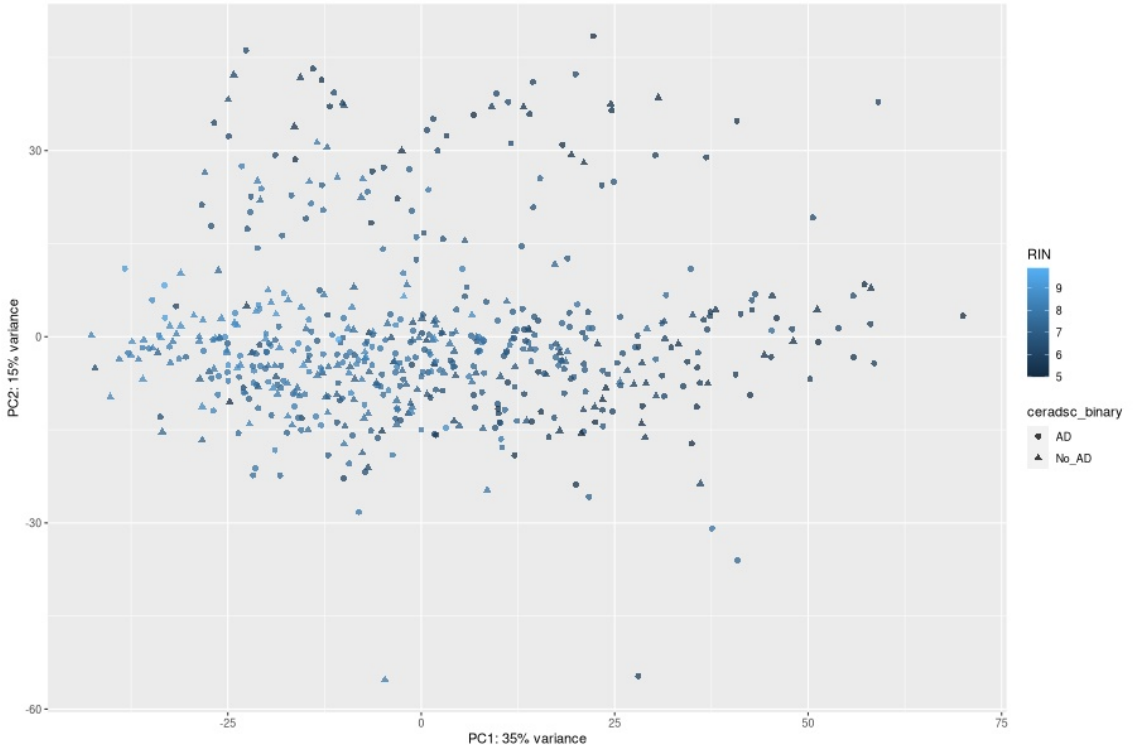
A

Appendix

In this appendix are large equivalents of all small plots found in the results. If you are interested in equivalent plots from other batches, or other annotations of the plots shown, they are available on the Polster-lab GitHub: https://github.com/Polster-lab/ssADGEM.

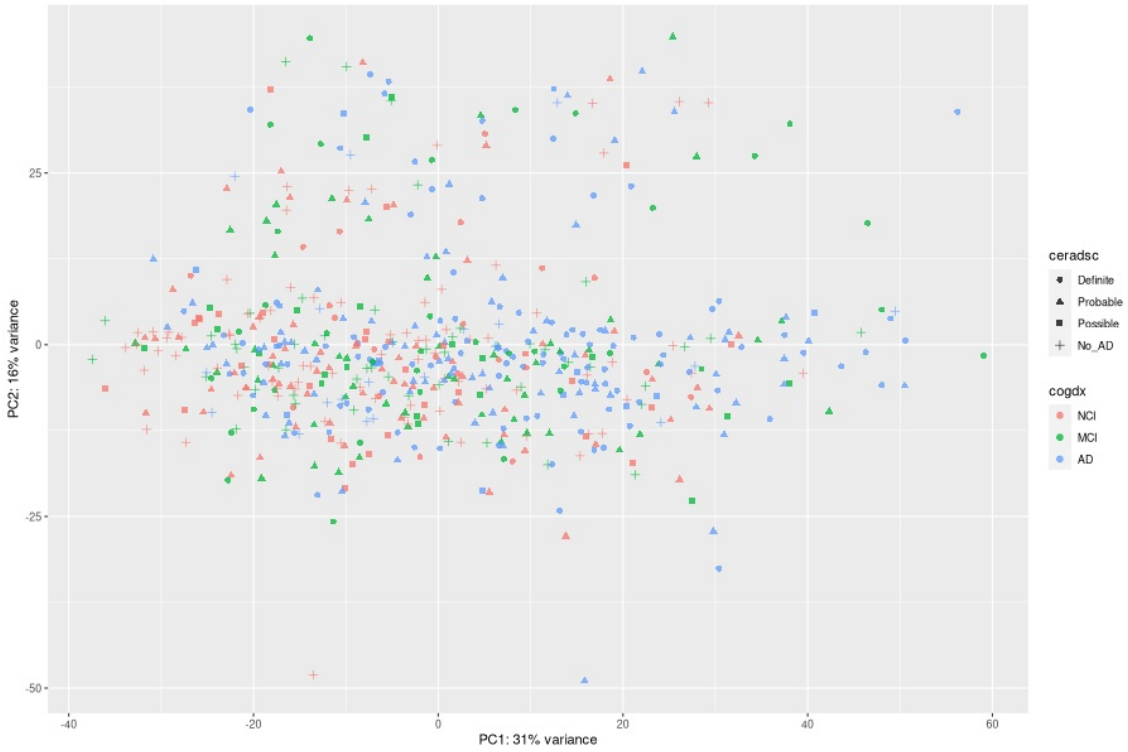


(a)



(b)

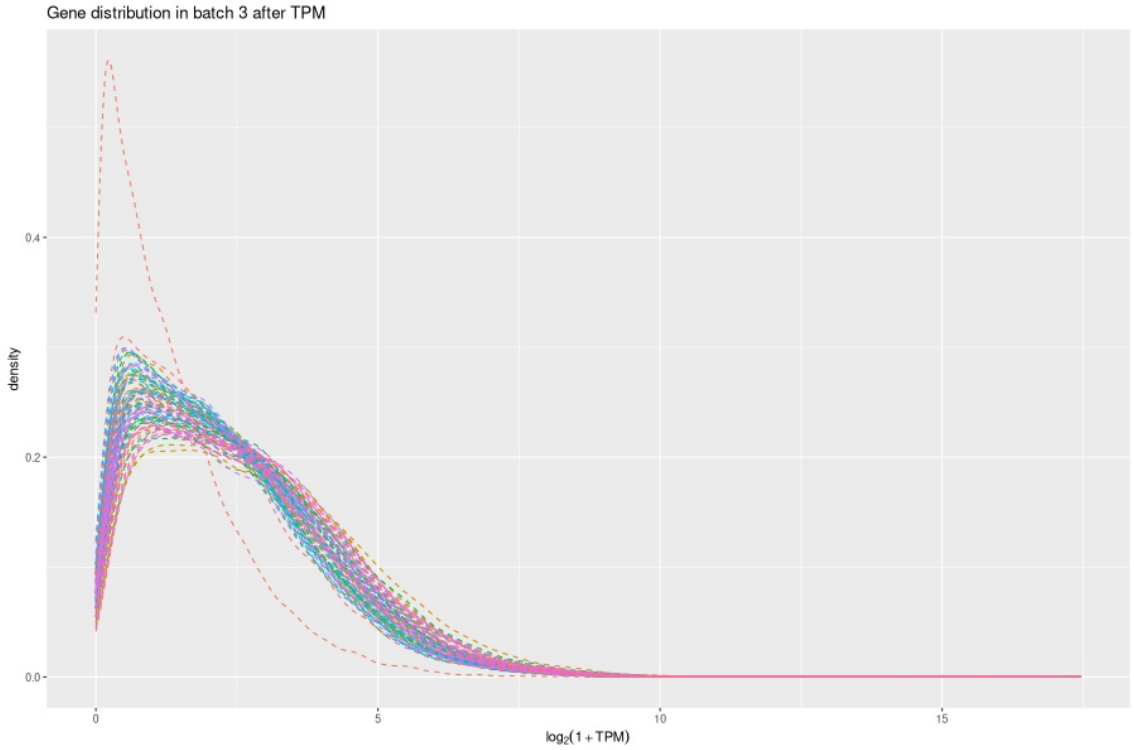
Figure A.1: Principal Component plots of variance stabilised sample gene count data using only genes that are differentially expressed: (a) showing M/F and batch annotations (b) showing RIN and a post-mortem AD assessment (ceradsc).



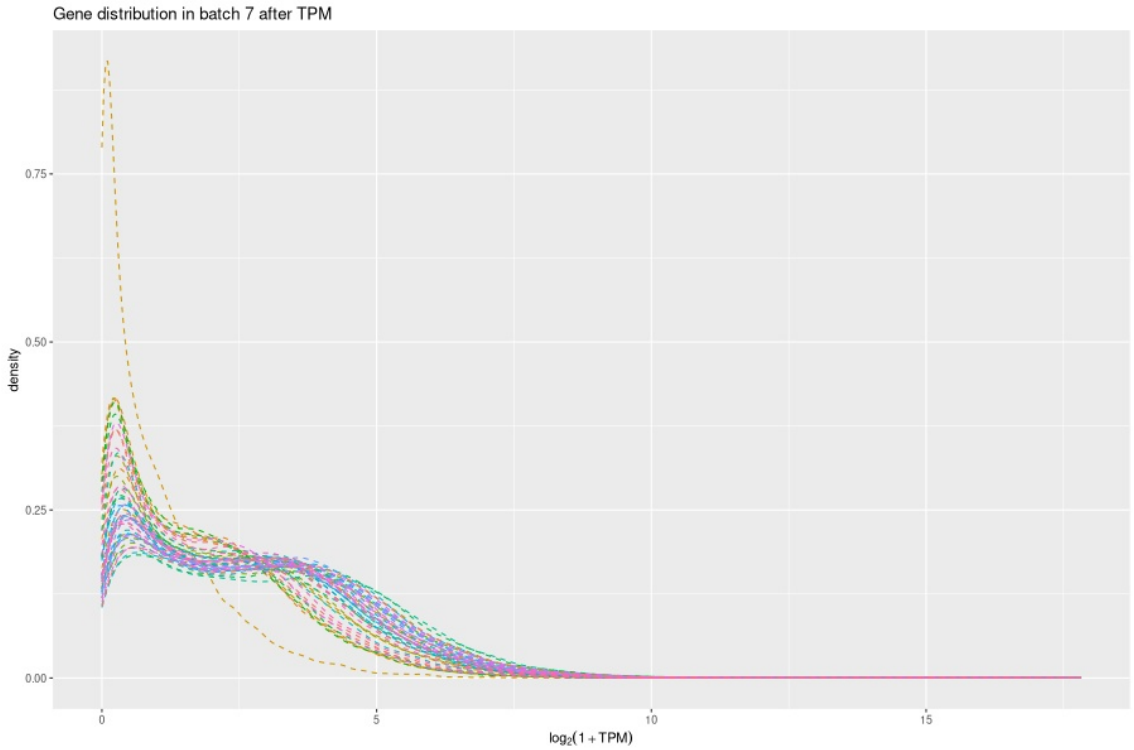
(a)

(b)

Figure A.2: Principal Component plots of variance stabilised sample gene count data using only genes that are differentially expressed: (a) showing post-mortem (ceradsc) and peri-mortem (cogdx) assesment (b) showing simplified post-mortem assessment (binary ceradsc).

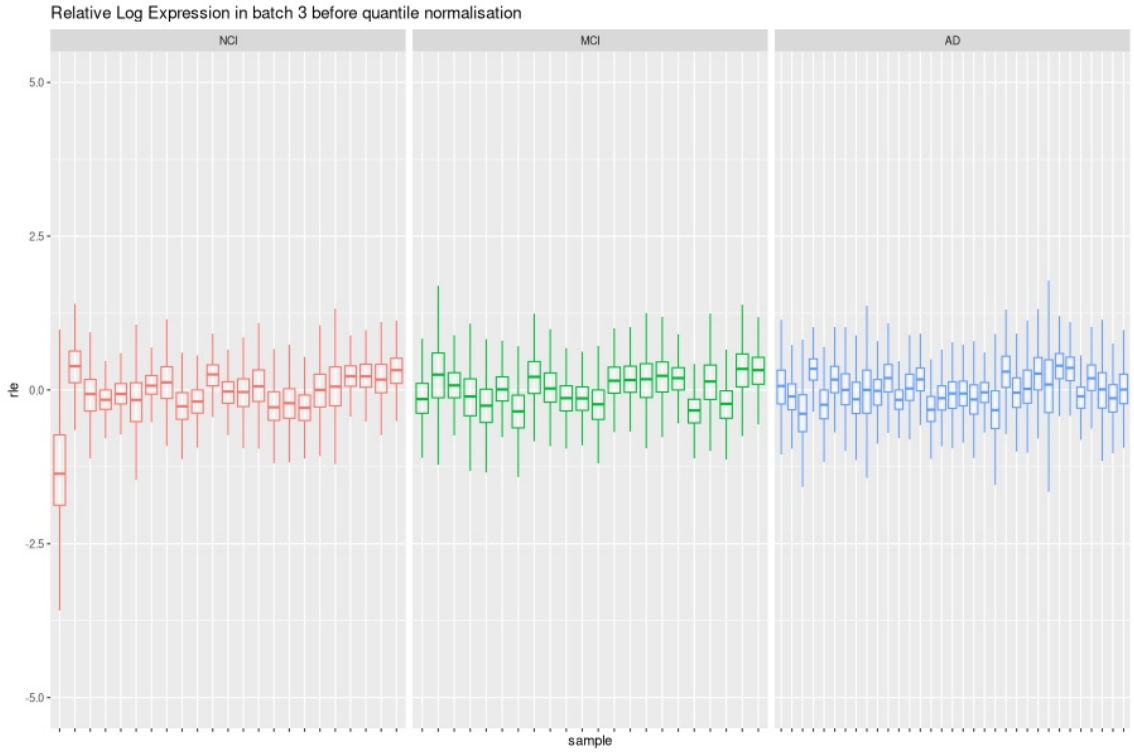


(a)

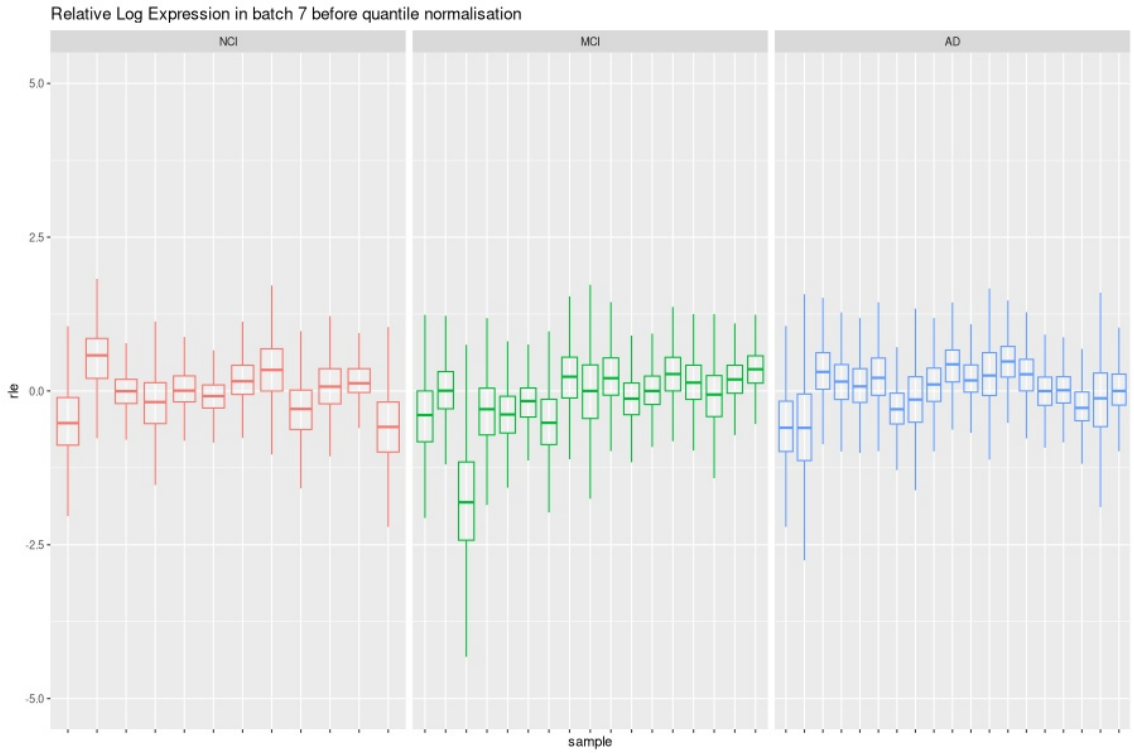


(b)

Figure A.3: Sample gene count distributions after TPM and log-transformed for visualisation, with batches (a) 3 and (b) 7 as examples



(a)



(b)

Figure A.4: RLE plots by premortem AD diagnosis, with batches (a) 3 and (b) 7 as examples

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY