RNA-seq Co-Expression Analysis Across Tissues and Ageing

Transcriptional Module Co-Expression Preservation Within Tissues and Age-Related Decline in Gene-Gene and GO term Relationships

Francisco José Calheiros Craveiro Lopes

Thesis to obtain the Master of Science Degree in

Integrated Master in Biological Engineering

Supervisors: Prof. Dr. Andreas Beyer
Prof. Dra. Susana de Almeida Mendes Vinga Martins

Examination Committee

Chairperson: Prof. Dr. Gabriel António Amaro Monteiro
Supervisor: Prof. Dra. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Dr. Miguel Nobre Parreira Cacho Teixeira

October 2021

Preface

The work presented in this thesis was performed at the CellNet Group from CECAD (Cluster of Excellence for Aging Research) of Cologne Graduate School of Ageing Research (Cologne, Germany), during the period February-October 2021, under the supervision of Prof. Dr. Andreas Beyer, and within the frame of the Erasmus Placement programme. The thesis was co-supervised at Instituto Superior Técnico by Prof. Dra. Susana Vinga.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Acknowledgments

I offer my first word of thanks to my mentor Ana Carolina Leote, for her guidance, availability and words of encouragement throughout this project.

I would like to acknowledge my dissertation supervisors Prof. Dr. Andreas Beyer and Prof. Dra Susana Vinga, for their insight, support, and knowledge sharing that has made this Thesis possible.

A special thanks to Telma, my girlfriend, for the unconditional support provided over the last months.

I would also like to thank my parents for their friendship, encouragement and caring over all these years, for always being there for me through thick and thin and without whom this project would not be possible.

To each and every one of you – Thank you.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 13/08/2020.

Abstract

There is still no comprehensive understanding concerning co-expression preservation across tissues and concerning co-expression decline across ageing regarding age-related transcriptional dysregulation.

One objective was to assess the co-expression preservation of cross-tissue highly correlated gene modules within specific tissues to infer the underlying gene regulatory network stability between tissues. Modules were learnt by hierarchical clustering with Pearson correlation in human RNA-seq data. GO enrichment analyses were applied to interpret the obtained modules. Some modules stably conserved moderate to high co-expression within several specific tissues in line with the expectation that gene co-expression networks are not entirely rearranged between tissues. Providing additional support that many tissue-specific data and studies can be much more unified.

Additionally, genes and modules co-expression decline across ageing was evaluated, further deriving a kind of ”hub genes of ageing”. Gene-gene relevance for ageing was inferred by PCA variable loadings, specifically describing the co-expression variance in the direction of ageing. The sum of loadings per gene provided a kind of ”hubness of ageing” measure. GO GSEA was applied to interpret the sum of loadings. A heavy and consensual GO term representation of the immune system and proteostasis was obtained, as well as cell cycle regulation, respiratory chain, keratin-associated proteins, and cellular proliferation, locomotion, and structure. It was proposed that the corresponding gene-gene relationships might be interesting to delve into to assess the underlying mechanism of the respective systems decline during ageing. This may be useful for developing intervention strategies to delay or prevent ageing phenotypes such as immune senescence.

Keywords

Ageing; Tissue-specific regulation; Gene co-expression; RNA-seq data analysis.

Resumo

Ainda não existe um apoio abrangente relativamente à preservação da co-expressão ao longo de tecidos nem relativamente ao declínio da co-expressão ao longo do envelhecimento quanto à desregulação transcripcional relacionada com a idade.

Avaliou-se a preservação da co-expressão de módulos genéticos altamente correlacionados em tecidos específicos, inferindo a estabilidade da rede regulatória genética subjacente entre tecidos. Aplicou-se agrupamento hierárquico com correlação de Pearson em dados *RNA-seq* humanos. Os módulos interpretaram-se com análises de enriquecimento *GO*. Alguns módulos conservaram co-expressão moderada a alta, estavelmente, em vários tecidos específicos, apoiando a expectativa de que redes de co-expressão genética não estão completamente reorganizadas entre tecidos, e portanto, muitos dados e estudos *tissue-specific* podem ser mais unificados.

Adicionalmente, o declínio da co-expressão de genes e módulos foi avaliado ao longo do envelhecimento, derivando ainda uma espécie de ”*hub genes of ageing*”. A relevância das relações gene-gene para o envelhecimento foi inferida por *variable loadings* de *PCA*, descrevendo a variância de co-expressão na direcção do envelhecimento. A soma das *loadings* por gene forneceu uma espécie de medida de ”*hubness of ageing*”. A soma das *loadings* interpretaram-se com *GSEA*. Obteve-se uma pesada e consensual representação de termos GO do sistema imunitário, proteostase, ciclo celular, cadeia respiratória, queratina, e proliferação, locomoção e estrutura celular. Propõe-se que as correspondentes relações gene-gene devam ser relevantes para aprofundar a avaliação dos mecanismos subjacentes ao declínio dos respectivos sistemas durante o envelhecimento, na esperança de desenvolver estratégias de intervenção para atrasar ou prevenir fenótipos do envelhecimento, tais como a senescência imunitária.

Palavras Chave

Envelhecimento; Regulação ao longo de tecidos; Co-expressão genética; Análise de dados de *RNA-seq*

Contents

List of Figures

Acronyms

AMD	Age-related Macular Degeneration
BH	Benjamini-Hochberg
B2M	beta-2-microglobulin
BC	Bladder Cancer
CTLs	Cytotoxic T Cells
ER	Endoplasmatic Reticulum
EGF	Epidermal Growth Factor
FDR	False Discovery Rate
PC1	first Principal Component
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GTEx	Genotype-Tissue Expression
GO-BP	GO Biological Processes
GO-CC	GO Cellular Component
GO-MF	GO Molecular Function
HSC	Hematopoietic Stem Cell
IMF	Intermyofibrillar
KAPs	Keratin Associated Proteins
KIFs	Keratin Intermediate Filaments
LR	Linear Regression
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA

MHC	Major Histocompatibility Complex
miRNA	micro RNA
NK	Natural Killer
PCA	Principal Component Analysis
ROS	Reactive Oxygen Species
RIN	RNA Integrity Number
snRNA	small nuclear RNA
SSCs	Spermatogonial Stem Cells
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

1

Introduction

Contents

1.1 Gene Modules Co-expression Across and Within Tissues

It is already well known that a great portion of genes has tissue-specific expression levels and that genes with higher expression levels in a subset of tissues relative to the baseline expression across all tissues often play critical roles in the biological functions unique to those tissues [1]. Identification of tissue-specific genes has provided a deeper molecular insight of tissue functions [2–4], has led to uncovering key genetic regulatory elements [5, 6], and helped to define the molecular basis of several human diseases [7]. Nevertheless, even though tissue specificities are often described based on gene expression levels, it is recognized that, by itself, it does not adequately capture the variety of processes that distinguish different tissues [1] neither common processes across tissues.

To this end, it is increasingly implemented methods of Interaction networks based on genes co-expression, protein-protein interaction, localization and sequence to capture functional information, significantly improving prediction of gene function and characterization of interactions among gene products [8].

Because similar expression patterns between genes might reflect a similar or shared function, gene-gene relationships are frequently assessed via the degree of coordination of their gene expression level variation across samples, also known as gene co-expression [9]. More specifically, a shared expression profile between genes might mean that the same factors drive their activity or that they are functionally related [10].

Networks built from bulk gene expression data have been extensively observed to recapitulate known gene functions, resulting in numerous genomic applications of co-expression analyses [11].

For example, co-expression analyses have been employed to infer about the binding between transcription factors and causal regulation of their downstream targets [12], characterize disease and deduce interventions [13], and to understand inter-tissue molecular interactions [14].

Biological systems can be functionally organized in partly separated functional modules of genes defined by some particular association between them, such as metabolic or signaling pathways and protein interaction and regulatory networks [15]. Clearly separated networks do not exist in biological systems [15] since different functional modules are, to some extent, interconnected, influencing each other, or progressing together.

More precisely, a functional module can be defined as a group of genes or their protein products which are in some way related. For example genes that share similar regulatory pathways (co-regulated), genes that share similar expression patterns (co-expressed), genes whose proteins compose the same protein complex or genes that participate in the same metabolic or signalling pathway [16, 17].

Detection of co-expression modules is frequently used to infer about gene-gene interactions and functional genome annotation through the guilt-by-association principle and allow for a better understanding of disease origin and progression [18] by comparing changes in gene-gene interactions learnt in both healthy and diseased samples individually.

Several approaches and algorithms have been used for module detection in gene expression data. The one utilised in the present work and most popular approach is classical clustering, which has been used since the beginning of gene expression quantification and is still the most widely used [18].

Gene modules also shed light on tissue-specificity, in which cells perform different functions despite possessing practically identical DNA. Tissue-specificity is believed to be partially achieved through tissue-dependent mechanisms of gene regulation, including epigenetic modification and transcriptional and post-transcriptional regulation. Co-expression modules or networks can, to some extent, capture those tissue-dependent mechanisms [19].

However, many tissue-specific or even cell-type-specific regulatory studies have too few samples to accurately infer about the millions of parameters that would define a co-expression or regulatory network [19]. Thus, one solution that the present work attempts to assess its plausibility would be to to learn a single consensus network for all or most tissues by integrating available samples from different tissues. Actually, it has been observed [20] that networks learnt from different tissues ”share far more links than would be expected by chance, and learning links across multiple tissues” appears to be ”less noisy than learning links using a single tissue”.

Not only the present work attempts to assess the plausibility of merging cross-tissue data for GRN inferring when there is lack of samples, but it also elucidates on the feasibility for such a GRN to predict gene expression across different tissues based on a single model. This was achieved by assessing the degree to which highly correlated cross-tissue gene modules preserve their co-expression within specific tissues, e.g. elucidating to what extent the regulatory relationships between genes are maintained across several tissues.

Here it is proposed the idea that independent of cell type, there is an underlying omnipresent regulatory network that mainly returns different gene expression levels because of different levels of gene activation and silencing so that depending on the cell type, different parts of the network are used. Nevertheless, there should be some gene modules that make use of this network in a rather stable way across tissues and cell types in a manner that they should be enough to build a more tissue encompassing regulatory network and to predict a great portion of genes expression levels across tissues and cell types. In fact, to our knowledge, there is no evidence that gene co-expression networks are completely rearranged between cell types and consequently by tissue type.

Actually, there have been studies [19] showing that consistent modules across tissues are especially prone to be enriched for Gene Ontology functions, and that these functions tend to be those which are essential to all tissues (e.g. mitosis). The referenced study applied an GNAT (Gene Network Analysis Tool) algorithm, to construct co-expression networks for each 35 distinct human tissues, using a tissue similarity hierarchy to encourage nearby tissues (in the hierarchy) to have similar networks. However, the mentioned study uses a very sparse amount of samples ranging from tissues with only 12 samples to tissues with 157 samples being the median 25 samples.

A more recent study [21] also shows that those kinds of consistent network modules across tissues are significantly correlated between them, indicating a general similar network pattern across tissues. The study also shows that physically closer tissues seem to be more similar in their co-expression networks. Their network modules were enriched in tissue-common functions like organelle membrane or immune-related functions and tissue-specific functions like renal functions in the kidney. The referred study used the weighted gene co-expression network analysis (WGCNA) approach and performed maximal clique analyses to retrieve modules conserved across tissues and also tissue-specific modules. The study used 52 human tissues, with the number of samples for each tissue varying from 71 to around 500.

Another recent study [22], identified regulon modules that globally regulate multiple cell groups and tissues across mouse cell atlases, and observed that cell type–specific regulons are characterised by distinct composition and activity, critical for their definition. The referred study collects regulons by applying ”GRNboost” algorithm using a list of TFs (Transcription Factors) and gathering their direct target genes harbouring significant TF motif enrichment.

A 2021 study [23] shows that ”14,636 out of 33,488 genetically regulated genes are co-regulated together with at least one other gene, resulting in 14,727 unique expression clusters across 49 analyzed human tissues”. This study defined regulatory clusters as a group of genes, located within the same genomic region, that are regulated by the same eQTL signal. These regulatory clusters were calculated for each tissue separately based on linear regression models using ”FastQTL” algorithm.

In the present work, a gene clustering is done only one time in a cross-tissue approach and used an equilibrated and substantial amount of samples (455) per tissue: Muscle - Skeletal, Whole Blood, Skin - Sun Exposed (Lower leg), Artery - Tibial, Adipose - Subcutaneous, Thyroid, Skin - Not Sun Exposed (Suprapubic), Nerve - Tibial, Lung, Esophagus - Mucosa, Adipose - Visceral (Omentum) and Esophagus - Muscularis. In light of this, the present work has a more robust and equilibrated amount of human tissue samples for the co-expression measures than any other study (to our knowledge). Additionally, the applied approach of assessing cross-tissue-learnt-clusters in specific tissues is novel, at least in this specific topic.

1.2 Co-expression Changes Across Ageing And Within Age Groups

A second part of the present work revolves around the idea that the mentioned gene modules and their co-expression analyses might also provide insight into the underlying mechanisms of ageing. Thus, instead of delving into tissue-specificity, we can similarly delve into age-group-specificity and explore changes in gene modules or co-expression networks across ageing.

Ageing occurs in all living organisms and is a natural process that can be defined as a deterioration of the cell functioning [24] thought to be through a series of mechanisms namely the loss of genomic stability, epigenetic alterations, loss of proteostasis, deregulated nutrient signalling, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, deviant intercellular communication and telomere shortening [25]. These 9 age-related phenotypes that appear to be conserved among species are called the 9 hallmarks of ageing and have been consensual to this day [26]. As a consequence of this decline of cell functioning, ageing erodes every physiological function of our organism [27] leading to a progressive fitness decline that brings life to a close [28].

Regardless of this generally accepted idea that ageing is a multifactorial process, many theories have emerged that try to explain ageing with a single predominant age-related change.

On the one hand, a popular theory of ageing is the “Stochastic Theory”. This theory suggests that ageing results from the accumulation of random damage outside or within cells accompanied by continuous decline of damage repairing systems [24].

On the other hand, theories support the concept that ageing is a genetically regulated process. One example of this regulation derives from the telomere-shortening hallmark of ageing. Telomere length decreases upon mitosis, thus across ageing. They are thought to play a DNA-protective role at the end of each chromosome, avoiding decreasing the remaining DNA length upon mitosis. When telomeres get dangerously short, apoptosis or cellular senescence is triggered [29].

Even so, there is one increasingly controversial theory which is the Free Radicals (or Oxidative Stress) Theory of Aging. This theory suggests an accumulation of Reactive Oxygen Species (ROS) across ageing by antioxidant defences decline and ROS increased production due to mitochondrial dysfunction. This accumulation is popularly thought to lead to increased oxidative damage of biomolecules with age, causing a decline in cellular function [24].

Substantial evidence supported this theory for a long time because it points to an age-related ROS increase and oxidative damage, accompanied by gradual loss of mitochondrial function, which in turn enhances ROS production [30]. Additionally, ”several age-dependent diseases are associated with severe increases in oxidative stress” [30].

Nevertheless, there has been an increasing amount of results that have forced an intense re-evaluation of the oxidative stress theory of ageing [25]. Of particular impact, there has been unexpected evidence that high ROS production not only does not cause accelerated ageing but is actually correlated to increased longevity in multiple organisms, as well as the evidence that decreasing the ROS production has failed to increase lifespan [30]. Actually, there has been accumulated solid evidence for the role of ROS in triggering proliferative and survival signals in response to physiological signals and stress conditions, suggesting the role of ROS as a stress-elicited survival signal aimed at compensating for the progressive deterioration associated with ageing [25].

A new conceptual framework tries to accommodate, the seemingly, conflicting evidence of ROS in ageing, hypothesising that the ROS increase across ageing might be an attempt to maintain survival until it betrays the original purpose and eventually aggravates, rather than alleviate, the age-associated damage [25].

As mentioned, gene modules and their co-expression analyses might also provide insight into the underlying mechanisms of ageing. The concept of transcriptional dysregulation has been proposed as a possible central mechanism of functional decline during ageing; until now, its generality has not been comprehensively empirically supported.

There is already some evidence in terms of increased transcriptional variability in scRNA-seq across ageing in mice [31,32], and human pancreas [33]. This transcriptional noise levels that increase with age are suggested as a possible consequence of the accumulation of mutations or and epimutations [34].

Then there is one recent study [35] that applies a novel clever new measure called global coordination level (GCL) that measures the average multivariate dependency between expression levels of random gene subsets of single cells. This study performed GCL analyses of 19 cohorts of scRNA-seq data from mice and fruit flies, finding a significant age-related decrease in the GCL across cell types and organisms. Furthermore, this work demonstrates that loss of gene-to-gene coordination is associated with ageing-related DNA damage.

Another study [36] delves into age-related deregulation of gene expression and protein synthesis, characterising those changes in both the transcriptome and translatome of mouse tissues and identified several involved processes related to inflammation, extracellular matrix, lipid metabolism, regulation of blood pressure, proteasomal protein degradation, mitochondrial activity, and oxidative stress.

Additionally, it has been previously describe the decrease in gene co-expression within genetic modules in bulk microarray data across 16 different mice tissues [37].

The goal of the present work was to determine, by analysing RNA-seq profiles across 26 different tissues within several human age groups, whether transcriptional dysregulation, as manifested in the gene-gene co-expression, is a characteristic phenomenon in ageing.

Altogether, investigating co-expression across ageing allows deriving general knowledge about the underlying topological and functional properties and eventual key driver genes of ageing that might be associated with age-related diseases and lifespan, which could be used as diagnostic biomarkers and drug targets.

1.3 Work Outlook

This work is organized as follows: In the ‘Methodologies’ section it is presented the exploited data sets and the inference process of the gene modules and gene-gene relationships across ageing. Then, it is described the module comparison approach, whether between tissues or age groups. Next, the section ‘Results and Discussion’ summarizes, describes and discusses the obtained results. Finally, the section ‘Conclusions’ ends the work with some final remarks.

2

Methodologies

Contents

2.1 Data Loading

This project used publicly available [38] RNA-seq gene expression data collected from the Genotype-Tissue Expression (GTEx) consortium official website. GTEx samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays, including Whole-Genome Sequencing (WGS), Whole-Exome Sequencing (WES), and RNA-Seq. GTEx's gene read counts dataset (v8) contains data from 838 postmortem donors comprising 17382 RNA-seq samples of 56200 genes across 54 tissue sites and two cell lines.

2.2 Data Filtering

2.2.1 Gene Filtering

Because this project works mainly with gene-gene correlations, reducing the number of genes has a significant impact on the computation power needed for computing the correlation matrices. Additionally, lowly expressed genes should also be removed since their quantification might be noisy. Thus, genes whose mean expression was below 1 were filtered out. This approximately corresponds to the 37% quantile of the genes mean expression. This threshold of mean expression equal to 1 can be observed in figure 2.1 where the gene mean expression density plot is represented.

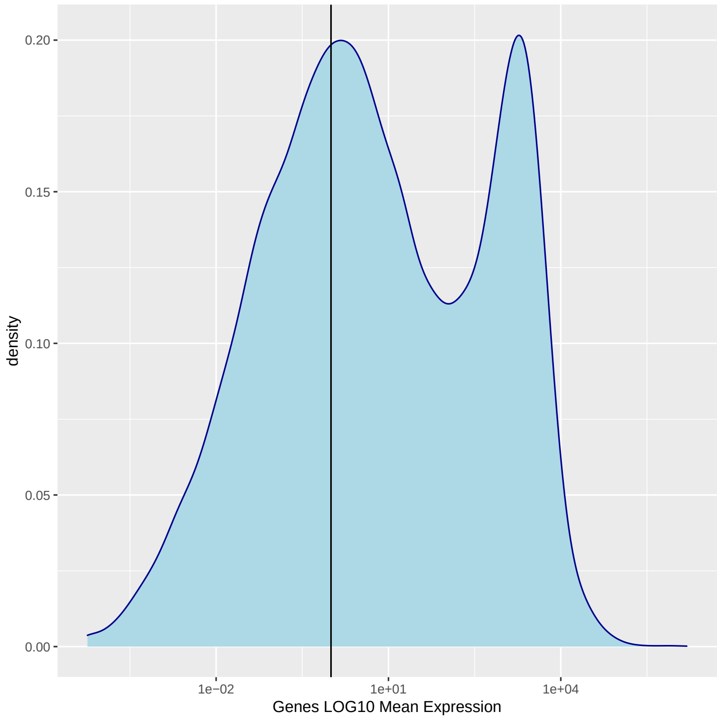
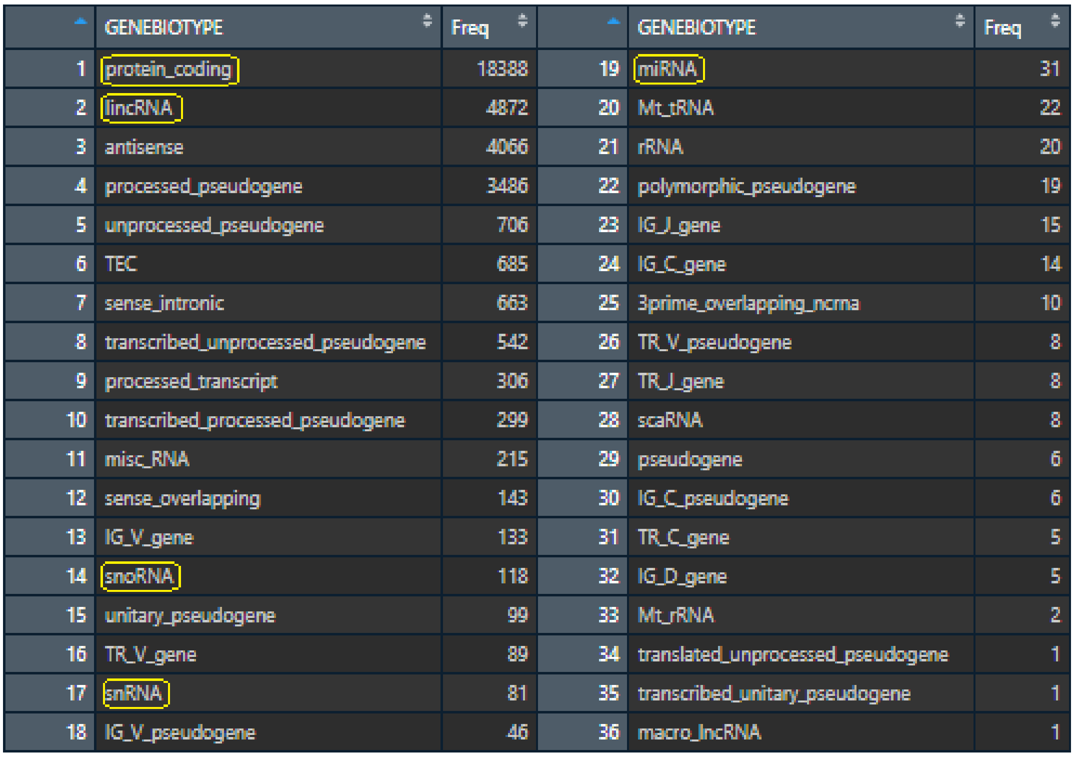


Figure 2.1: Gene mean expression density plot obtained from raw GTEx data (v8) with a black vertical line representing the threshold applied of gene mean expression < 1 for gene filtering. The gene mean expression X axis is transformed into a log10 scale.

Additionally, the kind of gene biotypes present in the data and their frequencies were evaluated (figure 2.2). The gene biotypes that were considered relevant and kept after filtering were protein-coding, long intergenic non-coding RNA (lincRNA), small nuclear RNA (snRNA), micro RNA (miRNA), and small nucleolar RNA (snoRNA).



	GENEBIOTYPE	Freq
1	protein_coding	18388
2	lincRNA	4872
3	antisense	4066
4	processed_pseudogene	3486
5	unprocessed_pseudogene	706
6	TEC	685
7	sense_intronic	663
8	transcribed_unprocessed_pseudogene	542
9	processed_transcript	306
10	transcribed_processed_pseudogene	299
11	misc_RNA	215
12	sense_overlapping	143
13	IG_V_gene	133
14	snoRNA	118
15	unitary_pseudogene	99
16	TR_V_gene	89
17	snRNA	81
18	IG_V_pseudogene	46
19	miRNA	31
20	Mt_tRNA	22
21	rRNA	20
22	polymorphic_pseudogene	19
23	IG_J_gene	15
24	IG_C_gene	14
25	3prime_overlapping_ncrna	10
26	TR_V_pseudogene	8
27	TR_J_gene	8
28	scaRNA	8
29	pseudogene	6
30	IG_C_pseudogene	6
31	TR_C_gene	5
32	IG_D_gene	5
33	Mt_rRNA	2
34	translated_unprocessed_pseudogene	1
35	transcribed_unitary_pseudogene	1
36	macro_lncRNA	1

Figure 2.2: Gene biotype frequencies (Freq) in raw GTEx (v8) data with the selected biotypes from filtering circled in yellow.

2.2.2 Sample Filtering

One of the annotations of the samples from GTEx is the RNA Integrity Number (RIN) which gives a measurement of how much the RNA from each sample has been digested by the presence of the nearly ubiquitous RNase enzymes [39]. RNA rapid digestion results in shorter fragments that commonly occur in the samples and can potentially compromise results of downstream applications [39].

The GTEx Portal [40] clarifies that RNA Integrity Number was measured by Agilent Bioanalyzer and states that all samples with a RIN of 6.0 or higher qualify for RNA Sequence analysis. Therefore, following GTEx Portal recommendations, samples with RIN smaller than 6.0 were filtered out.

Additionally, GTEx data contains two cell line samples, namely the ”Cells - Cultured fibroblasts” (504 samples) and ”Cells - EBV-transformed lymphocytes” (174 samples). These samples were also removed as they are not the focus of the current project.

After sample filtering, there remain 15030 samples.

2.3 DESeq2 Data Normalisation

The raw counts of mapped reads for each gene should be proportional to the expression of RNA, but this kind of data carries some factors that must be accounted for and corrected. Normalisation is the process of scaling these raw count values to account for those factors. This way, the expression levels are more comparable between and within samples.

The main factors often considered during normalisation are sequencing depth, gene length and RNA composition. However, for this kind of analysis between samples and not within-sample comparisons, what is required is to consider sequencing depth and RNA composition. Thus an adequate method would be DESeq2 [41].

Sequencing depth is the read counts sum of all genes within a sample. For example, supposing the sequencing method overall read twice as much counts in one sample than another. In that case, some genes might misleadingly reveal themselves as doubly overexpressed if sequencing depth is not taken into account.

In this case, RNA composition refers to taking into account mainly highly differentially expressed genes between samples. This is because genes that are highly deferentially expressed between samples might skew normalised gene expression when the sequencing depth is considered [41].

To normalise for sequencing depth and RNA composition, DESeq2 uses the median of ratios method. Simplistically, it divides counts by sample-specific size factors determined by the median ratio of gene counts relative to geometric mean per gene.

DESeq2 normalisation is implemented as a package for the R statistical environment (used R version 3.4.3) and is available [42] as part of the Bioconductor project [43].

2.4 Batch Effect Correction

In bioinformatics, it is essential to do a step of confounder removal because, in gene expression datasets, there are a plethora of heterogeneity sources, and some might be uninteresting. Some of its heterogeneity is good and desired because it enables gene expression variance across samples for a correlation between genes to be capturable. Those desirable factors can be related to differences from donors to donors, such as lifestyles or ethnicity. However, there are technical confounders such as the ischemic time that should be removed so as not to skew the data and not capture relationships derived from artefacts [44].

It was used a Linear Regression (LR) adjustment for known confounders given that LR has been described [45] as the most adequate, outperforming other adjustment methods when explicitly applied to the GTEx dataset when assessing if the removal of unwanted technical variation harmed the biological signal that is of interest to the researcher [45].

This way, LR was used to regress out the known covariates ischemic time (SMTSISCH representing the interval in minutes between the time of donor death and sample collection), experimental batch (SMGEBTCH) and death type (DTHHRDY), fitting the model for each gene separately. Known covariates were regressed out using the R statistical environment's built-in ”lm” function. These batch effects are among the ones usually regressed out in datasets as GTEx [44] [45].

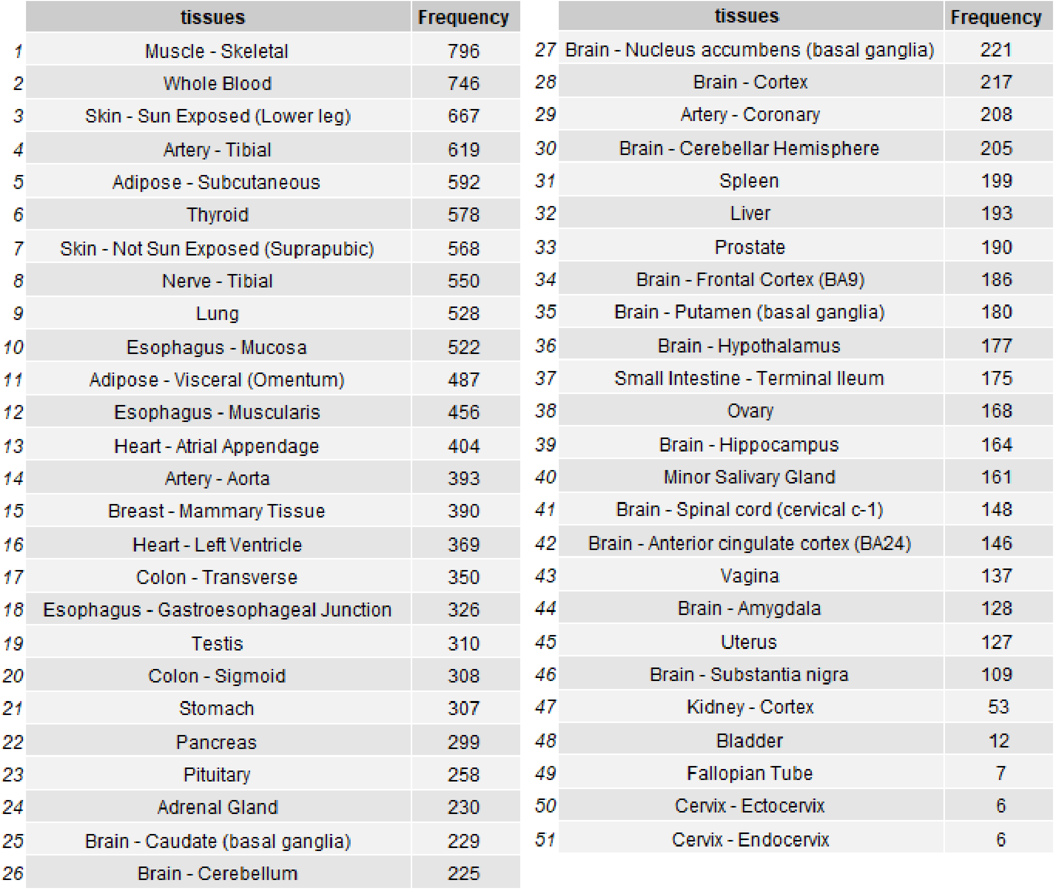
It should be noted that LR is a method that transforms absolute expression values into residuals [46]. When correcting for a batch effect of a covariate, for each gene, LR separates the samples into the corresponding batches and swaps the genes' absolute expression by the residual value of the genes' absolute expression relative to the corresponding batch's average expression for each of the samples.

Additionally, in gene expression projects it is usually done a logarithmic transformation of the data. One of the main reasons derives from the library preparation cDNA amplification of RNA-Seq PCR step. This cDNA amplification results in exponential scales. Thus the natural fold change does not follow a normal distribution, whereas the log2 or log10 transformed one is closer. Another reason is that we are modelling proportional changes rather than additive changes when using log-transformed expression values. This is typically biologically more relevant.

The log2 transformation was applied to the data, with the addition of a pseudo count of one. Log2 was used because it is the one most commonly applied. Given that batch effect removal transforms the data, this logarithmic transformation was done before those steps.

2.5 Sample Subsetting

After the preprocessing of gene filtering, batch effect correction and DESeq2 normalisation, the data represented in figure 2.3 is the base data that is then differently subsetted for each of the analysis in the following sections.



	tissues	Frequency
1	Muscle - Skeletal	796
2	Whole Blood	746
3	Skin - Sun Exposed (Lower leg)	667
4	Artery - Tibial	619
5	Adipose - Subcutaneous	592
6	Thyroid	578
7	Skin - Not Sun Exposed (Suprapubic)	568
8	Nerve - Tibial	550
9	Lung	528
10	Esophagus - Mucosa	522
11	Adipose - Visceral (Omentum)	487
12	Esophagus - Muscularis	456
13	Heart - Atrial Appendage	404
14	Artery - Aorta	393
15	Breast - Mammary Tissue	390
16	Heart - Left Ventricle	369
17	Colon - Transverse	350
18	Esophagus - Gastroesophageal Junction	326
19	Testis	310
20	Colon - Sigmoid	308
21	Stomach	307
22	Pancreas	299
23	Pituitary	258
24	Adrenal Gland	230
25	Brain - Caudate (basal ganglia)	229
26	Brain - Cerebellum	225
27	Brain - Nucleus accumbens (basal ganglia)	221
28	Brain - Cortex	217
29	Artery - Coronary	208
30	Brain - Cerebellar Hemisphere	205
31	Spleen	199
32	Liver	193
33	Prostate	190
34	Brain - Frontal Cortex (BA9)	186
35	Brain - Putamen (basal ganglia)	180
36	Brain - Hypothalamus	177
37	Small Intestine - Terminal Ileum	175
38	Ovary	168
39	Brain - Hippocampus	164
40	Minor Salivary Gland	161
41	Brain - Spinal cord (cervical c-1)	148
42	Brain - Anterior cingulate cortex (BA24)	146
43	Vagina	137
44	Brain - Amygdala	128
45	Uterus	127
46	Brain - Substantia nigra	109
47	Kidney - Cortex	53
48	Bladder	12
49	Fallopian Tube	7
50	Cervix - Ectocervix	6
51	Cervix - Endocervix	6

Figure 2.3: Sample frequency per tissue obtained after raw GTEx sample filtering by removing samples with RIN<6 and Cultured Cells.

This is the base data that is differently subseted for each project after further preprocessing of gene filtering, DESeq2 normalisation and batch effect correction.

2.5.1 Tissue-specific and cross-tissue analysis

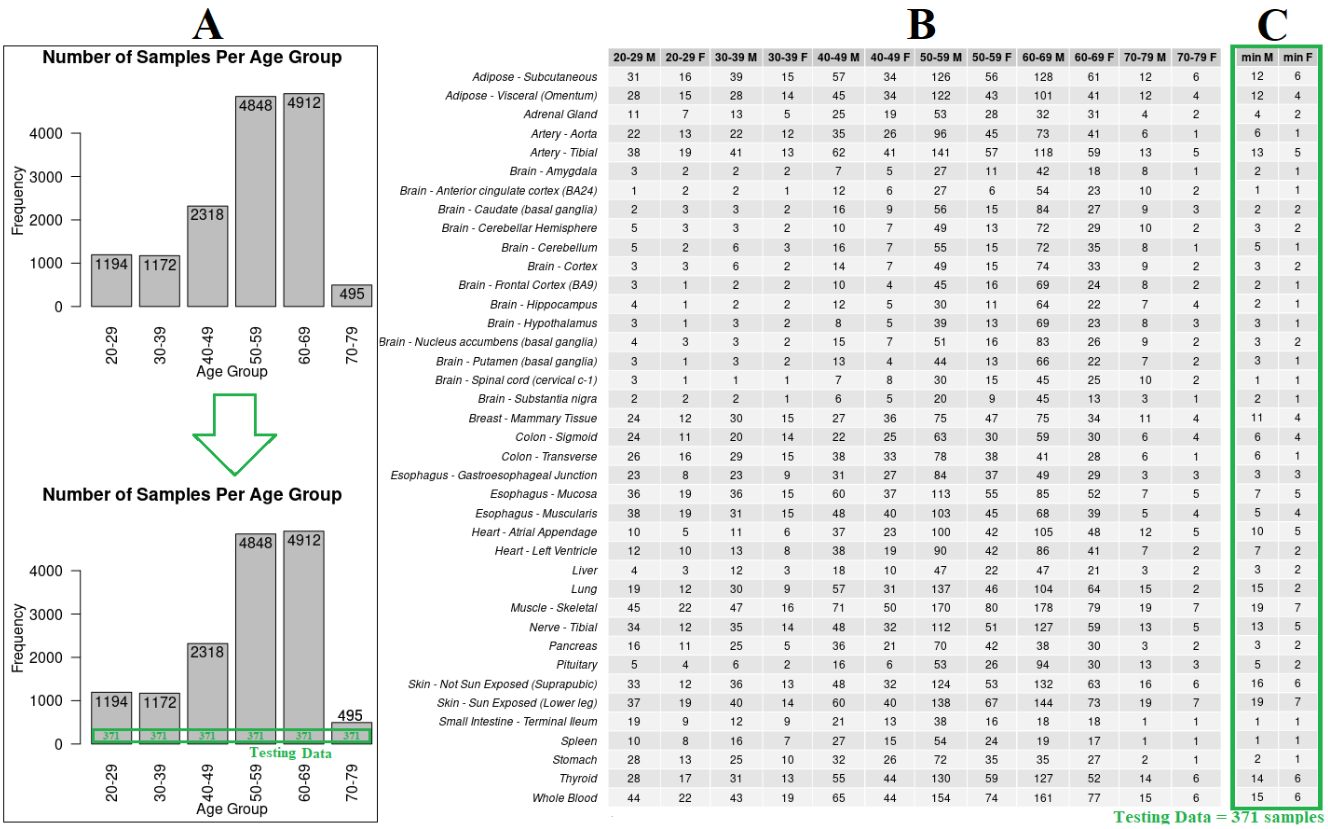
In figure 2.3 we have the number of samples per tissue, but to ensure the tissue-specific and cross-tissue analyses were comparable, the same total number of samples per tissue was used in each tissue subset (455 samples). Additionally, the same number of samples from each tissue was used in the cross-tissue analysis (37 samples).

This balance in samples required the discarding of all tissues with fewer than 455 samples. The resulting tissues are the Muscle - Skeletal, Whole Blood, Skin - Sun Exposed (Lower leg), Artery - Tibial, Adipose - Subcutaneous, Thyroid, Skin - Not Sun Exposed (Suprapubic), Nerve - Tibial, Lung, Esophagus - Mucosa, Adipose - Visceral (Omentum) and Esophagus - Muscularis.

2.5.2 Age group analysis

Similar to the tissue analysis, in the age group analysis, it is desired to obtain (”learning”) gene modules by clustering algorithm within a cross-age sample subset and then to reevaluate (”testing”) those modules within different age groups sample subsets.

The amount of samples per age group (20-29, 30-39, 40-49, 50-59, 60-69 and 70-79) was determined in the base data (figure 2.4.A superior panel). In this analysis, it is fundamental for the ”testing” age groups sample subsets to have as many samples as possible and in figure 2.4.A it can be seen that the smallest age group (70-79) has 495 samples. It is not desired to decrease this amount by sparing samples to the cross-age subset, so the ”70-79” age group was set aside from being used in the ”learning” cross-age subset.

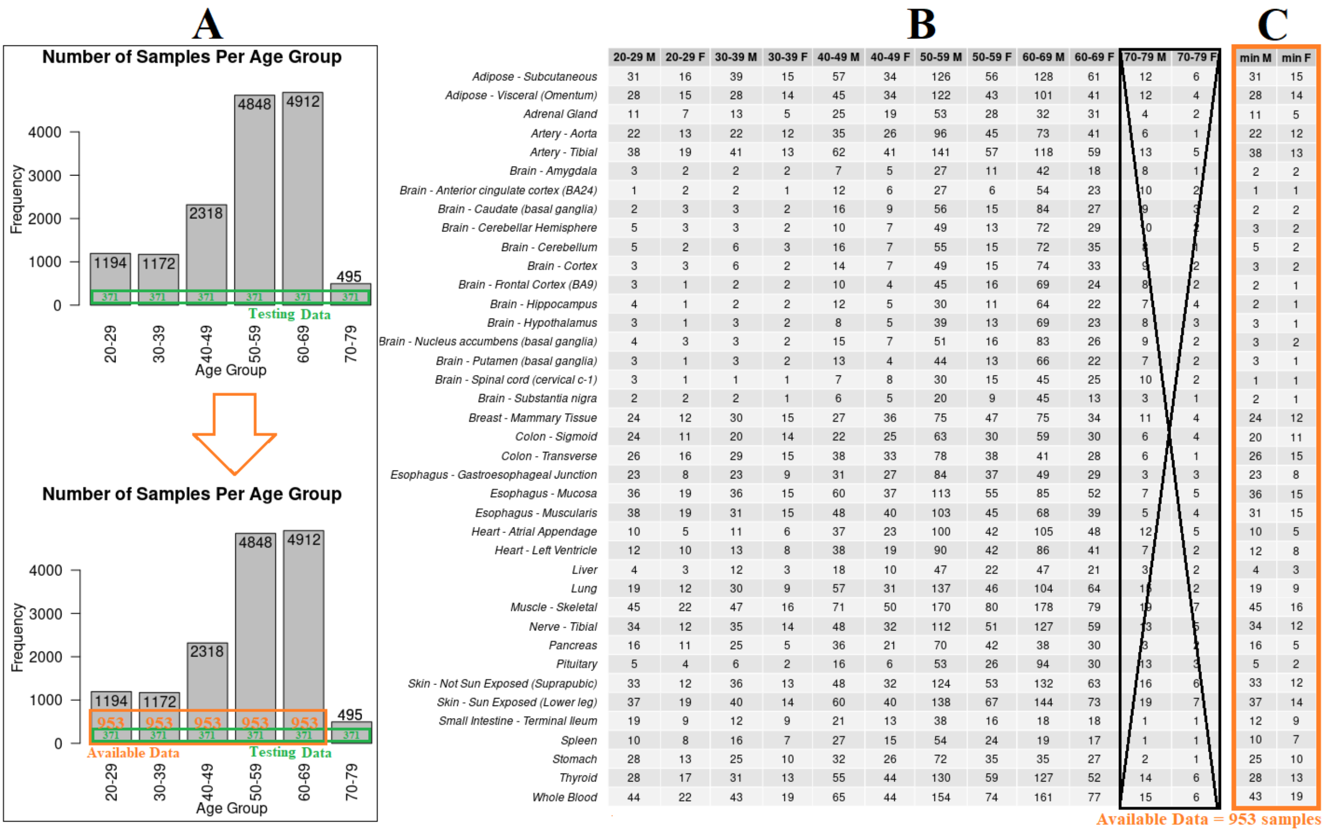


	20-29 M	20-29 F	30-39 M	30-39 F	40-49 M	40-49 F	50-59 M	50-59 F	60-69 M	60-69 F	70-79 M	70-79 F	min M	min F
Adipose - Subcutaneous	31	16	39	15	57	34	126	56	128	61	12	6	12	6
Adipose - Visceral (Omentum)	28	15	28	14	45	34	122	43	101	41	12	4	12	4
Adrenal Gland	11	7	13	5	25	19	53	28	32	31	4	2	4	2
Artery - Aorta	22	13	22	12	35	26	96	45	73	41	6	1	6	1
Artery - Tibial	38	19	41	13	62	41	141	57	118	59	13	5	13	5
Brain - Amygdala	3	2	2	2	7	5	27	11	42	18	8	1	2	1
Brain - Anterior cingulate cortex (BA24)	1	2	2	1	12	6	27	6	54	23	10	2	1	1
Brain - Caudate (basal ganglia)	2	3	3	2	16	9	56	15	84	27	9	3	2	2
Brain - Cerebellar Hemisphere	5	3	3	2	10	7	49	13	72	29	10	2	3	2
Brain - Cerebellum	5	2	6	3	16	7	55	15	72	35	8	1	5	1
Brain - Cortex	3	3	6	2	14	7	49	15	74	33	9	2	3	2
Brain - Frontal Cortex (BA9)	3	1	2	2	10	4	45	16	69	24	8	2	2	1
Brain - Hippocampus	4	1	2	2	12	5	30	11	64	22	7	4	2	1
Brain - Hypothalamus	3	1	3	2	8	5	39	13	69	23	8	3	3	1
Brain - Nucleus accumbens (basal ganglia)	4	3	3	2	15	7	51	16	83	26	9	2	3	2
Brain - Putamen (basal ganglia)	3	1	3	2	13	4	44	13	66	22	7	2	3	1
Brain - Spinal cord (cervical c-1)	3	1	1	1	7	8	30	15	45	25	10	2	1	1
Brain - Substantia nigra	2	2	2	1	6	5	20	9	45	13	3	1	2	1
Breast - Mammary Tissue	24	12	30	15	27	36	75	47	75	34	11	4	11	4
Colon - Sigmoid	24	11	20	14	22	25	63	30	59	30	6	4	6	4
Colon - Transverse	26	16	29	15	38	33	78	38	41	28	6	1	6	1
Esophagus - Gastroesophageal Junction	23	8	23	9	31	27	84	37	49	29	3	3	3	3
Esophagus - Mucosa	36	19	36	15	60	37	113	55	85	52	7	5	7	5
Esophagus - Muscularis	38	19	31	15	48	40	103	45	68	39	5	4	5	4
Heart - Atrial Appendage	10	5	11	6	37	23	100	42	105	48	12	5	10	5
Heart - Left Ventricle	12	10	13	8	38	19	90	42	86	41	7	2	7	2
Liver	4	3	12	3	18	10	47	22	47	21	3	2	3	2
Lung	19	12	30	9	57	31	137	46	104	64	15	2	15	2
Muscle - Skeletal	45	22	47	16	71	50	170	80	178	79	19	7	19	7
Nerve - Tibial	34	12	35	14	48	32	112	51	127	59	13	5	13	5
Pancreas	16	11	25	5	36	21	70	42	38	30	3	2	3	2
Pituitary	5	4	6	2	16	6	53	26	94	30	13	3	5	2
Skin - Not Sun Exposed (Suprapubic)	33	12	36	13	48	32	124	53	132	63	16	6	16	6
Skin - Sun Exposed (Lower leg)	37	19	40	14	60	40	138	67	144	73	19	7	19	7
Small Intestine - Terminal Ileum	19	9	12	9	21	13	38	16	18	18	1	1	1	1
Spleen	10	8	16	7	27	15	54	24	19	17	1	1	1	1
Stomach	28	13	25	10	32	26	72	35	35	27	2	1	2	1
Thyroid	28	17	31	13	55	44	130	59	127	52	14	6	14	6
Whole Blood	44	22	43	19	65	44	154	74	161	77	15	6	15	6

Figure 2.4: A - Sample frequency histogram for each available age group with a schematic representation of the amount of samples per age group that should be used as testing data in the lower part of the figure. B - Sample frequency per age group in each of the tissues for both female (F) and male (M). C - The minimum amount of samples found across the several age groups in each of the tissues for both female (F) and male (M). Obs: tissues which had zero samples at least in one of the age groups or gender were omitted from this figure.

So one option would be to use 495 randomly chosen samples from each age group for the ”testing” step because that is the minimum number of samples found across age groups. Nevertheless, the data has a heterogeneous distribution of samples across the age groups regarding the number of samples per tissue and gender. This can be observed in figure 2.4.B. So, as to proportionate an equilibrium between age groups regarding the amount of samples per tissue and gender, the minimum number of samples per tissue in each of the age groups for both genders was determined (figure 2.4.C). This reasoning applied to all tissues for both genders results in 371 samples which can be used as testing data. Figure 2.4.A lower panel shows a schematic representation of this amount of 371 samples per age group that should be used as testing data.

Regarding the data available for learning, the same process was applied but, as mentioned, disregarding the samples from ”70-79” age group. Then, applying the same algorithm to the remaining age groups, the available data for the learning step was obtained (figure 2.5.C).

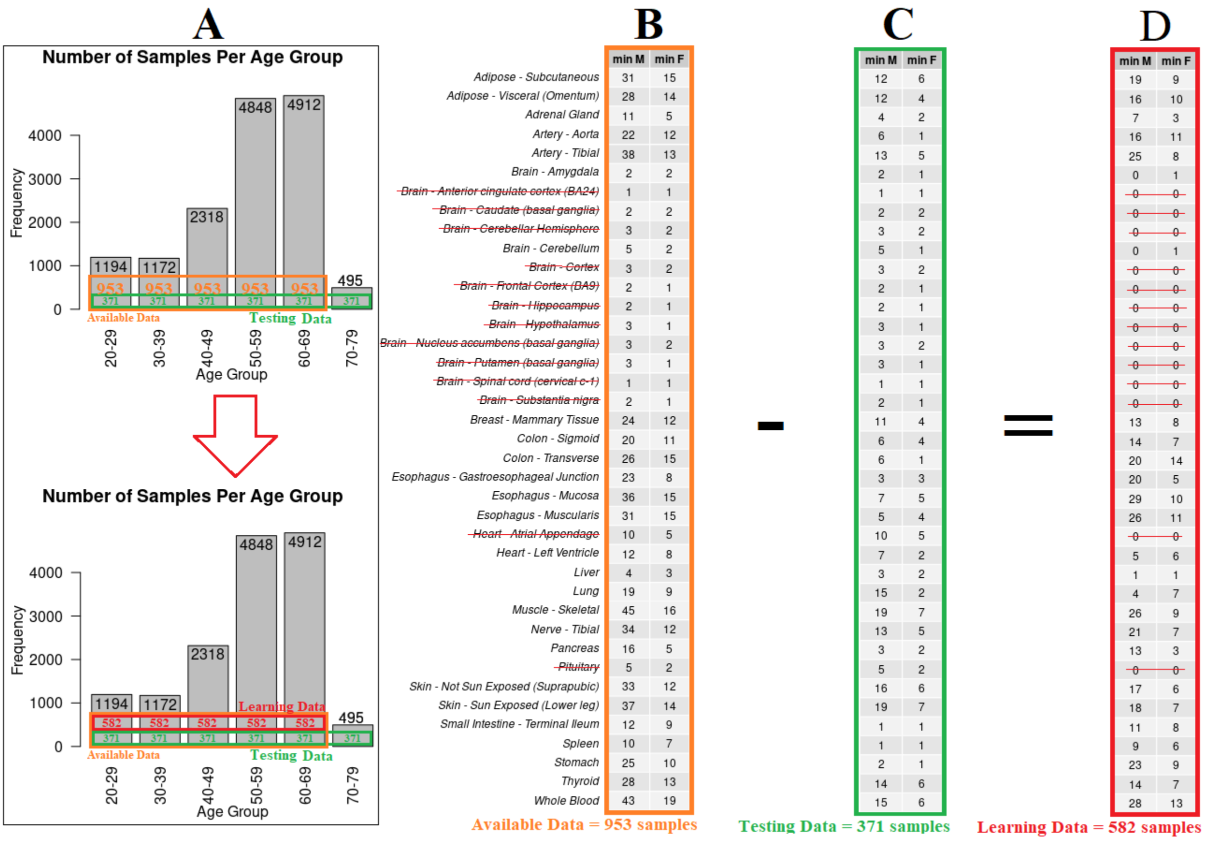


	20-29 M	20-29 F	30-39 M	30-39 F	40-49 M	40-49 F	50-59 M	50-59 F	60-69 M	60-69 F	70-79 M	70-79 F	min M	min F
Adipose - Subcutaneous	31	16	39	15	57	34	126	56	128	61	12	6	31	15
Adipose - Visceral (Omentum)	28	15	28	14	45	34	122	43	101	41	12	4	28	14
Adrenal Gland	11	7	13	5	25	19	53	28	32	31	4	2	11	5
Artery - Aorta	22	13	22	12	35	26	96	45	73	41	6	1	22	12
Artery - Tibial	38	19	41	13	62	41	141	57	118	59	13	5	38	13
Brain - Amygdala	3	2	2	2	7	5	27	11	42	18	8	1	2	2
Brain - Anterior cingulate cortex (BA24)	1	2	2	1	12	6	27	6	54	23	10	2	1	1
Brain - Caudate (basal ganglia)	2	3	3	2	16	9	56	15	84	27	9	3	2	2
Brain - Cerebellar Hemisphere	5	3	3	2	10	7	49	13	72	29	10	2	3	2
Brain - Cerebellum	5	2	6	3	16	7	55	15	72	35	8	1	5	2
Brain - Cortex	3	3	6	2	14	7	49	15	74	33	9	2	3	2
Brain - Frontal Cortex (BA9)	3	1	2	2	10	4	45	16	69	24	8	2	2	1
Brain - Hippocampus	4	1	2	2	12	5	30	11	64	22	7	4	2	1
Brain - Hypothalamus	3	1	3	2	8	5	39	13	69	23	8	3	3	1
Brain - Nucleus accumbens (basal ganglia)	4	3	3	2	15	7	51	16	83	26	9	2	3	2
Brain - Putamen (basal ganglia)	3	1	3	2	13	4	44	13	66	22	7	2	3	1
Brain - Spinal cord (cervical c-1)	3	1	1	1	7	8	30	15	45	25	10	2	1	1
Brain - Substantia nigra	2	2	2	1	6	5	20	9	45	13	3	1	2	1
Breast - Mammary Tissue	24	12	30	15	27	36	75	47	75	34	11	4	24	12
Colon - Sigmoid	24	11	20	14	22	25	63	30	59	30	6	4	20	11
Colon - Transverse	26	16	29	15	38	33	78	38	41	28	6	1	26	15
Esophagus - Gastroesophageal Junction	23	8	23	9	31	27	84	37	49	29	3	3	23	8
Esophagus - Mucosa	36	19	36	15	60	37	113	55	85	52	7	5	36	15
Esophagus - Muscularis	38	19	31	15	48	40	103	45	68	39	5	4	31	15
Heart - Atrial Appendage	10	5	11	6	37	23	100	42	105	48	12	5	10	5
Heart - Left Ventricle	12	10	13	8	38	19	90	42	86	41	7	2	12	8
Liver	4	3	12	3	18	10	47	22	47	21	3	2	4	3
Lung	19	12	30	9	57	31	137	46	104	64	15	2	19	9
Muscle - Skeletal	45	22	47	16	71	50	170	80	178	79	19	7	45	16
Nerve - Tibial	34	12	35	14	48	32	112	51	127	59	13	5	34	12
Pancreas	16	11	25	5	36	21	70	42	38	30	3	2	16	5
Pituitary	5	4	6	2	16	6	53	26	94	30	13	3	5	2
Skin - Not Sun Exposed (Suprapubic)	33	12	36	13	48	32	124	53	132	63	16	6	33	12
Skin - Sun Exposed (Lower leg)	37	19	40	14	60	40	138	67	144	73	19	7	37	14
Small Intestine - Terminal Ileum	19	9	12	9	21	13	38	16	18	18	1	1	12	9
Spleen	10	8	16	7	27	15	54	24	19	17	1	1	10	7
Stomach	28	13	25	10	32	26	72	35	35	27	2	1	25	10
Thyroid	28	17	31	13	55	44	130	59	127	52	14	6	28	13
Whole Blood	44	22	43	19	65	44	154	74	161	77	15	6	43	19

Available Data = 953 samples

Figure 2.5: A - Sample frequency histogram for each available age group and in the lower part of the figure a schematic representation of the amount of samples per age group that are available looking at all the age groups except for the 70-79 age group. B - Sample frequency per age group in each of the tissues for both female (F) and male (M). C - The minimum amount of samples found across all the age groups, except for the 70-79 age group, in each of the tissues for both female (F) and male (M), that can be used as testing data. Obs: tissues which had zero samples at least in one of the age groups or gender were omitted from this figure as well as tissues who became available by disregarding the minimum amount of samples present at the 70-79 age group, i.e. tissues that were omitted from testing data in figure 2.4.

It should be noted that this available data schematised in figure 2.5 is not the actual data used for learning since this data still contains the testing data from figure 2.4. The actual data used for learning is the subtraction of the samples used for testing (figure 2.6.C) from the available data (figure 2.6.B) which leaves the samples for the learning step present in figure 2.6.D. Figure 2.6.A (lower panel) schematises a histogram of the available data containing both the testing and learning data in the respective age groups.



	B		C		D	
	min M	min F	min M	min F	min M	min F
Adipose - Subcutaneous	31	15	12	6	19	9
Adipose - Visceral (Omentum)	28	14	12	4	16	10
Adrenal Gland	11	5	4	2	7	3
Artery - Aorta	22	12	6	1	16	11
Artery - Tibial	38	13	13	5	25	8
Brain - Amygdala	2	2	2	1	0	1
~~Brain - Anterior cingulate cortex (BA24)~~	1	1	1	1	~~0~~	~~0~~
~~Brain - Caudate (basal ganglia)~~	2	2	2	2	~~0~~	~~0~~
~~Brain - Cerebellar Hemisphere~~	3	2	3	2	~~0~~	~~0~~
Brain - Cerebellum	5	2	5	1	0	1
~~Brain - Cortex~~	3	2	3	2	~~0~~	~~0~~
~~Brain - Frontal Cortex (BA9)~~	2	1	2	1	~~0~~	~~0~~
~~Brain - Hippocampus~~	2	1	2	1	~~0~~	~~0~~
~~Brain - Hypothalamus~~	3	1	3	1	~~0~~	~~0~~
~~Brain - Nucleus accumbens (basal ganglia)~~	3	2	3	2	~~0~~	~~0~~
~~Brain - Putamen (basal ganglia)~~	3	1	3	1	~~0~~	~~0~~
~~Brain - Spinal cord (cervical c-1)~~	1	1	1	1	~~0~~	~~0~~
~~Brain - Substantia nigra~~	2	1	2	1	~~0~~	~~0~~
Breast - Mammary Tissue	24	12	11	4	13	8
Colon - Sigmoid	20	11	6	4	14	7
Colon - Transverse	26	15	6	1	20	14
Esophagus - Gastroesophageal Junction	23	8	3	3	20	5
Esophagus - Mucosa	36	15	7	5	29	10
Esophagus - Muscularis	31	15	5	4	26	11
~~Heart - Atrial Appendage~~	10	5	10	5	~~0~~	~~0~~
Heart - Left Ventricle	12	8	7	2	5	6
Liver	4	3	3	2	1	1
Lung	19	9	15	2	4	7
Muscle - Skeletal	45	16	19	7	26	9
Nerve - Tibial	34	12	13	5	21	7
Pancreas	16	5	3	2	13	3
~~Pituitary~~	5	2	5	2	~~0~~	~~0~~
Skin - Not Sun Exposed (Suprapubic)	33	12	16	6	17	6
Skin - Sun Exposed (Lower leg)	37	14	19	7	18	7
Small Intestine - Terminal Ileum	12	9	1	1	11	8
Spleen	10	7	1	1	9	6
Stomach	25	10	2	1	23	9
Thyroid	28	13	14	6	14	7
Whole Blood	43	19	15	6	28	13

Available Data = 953 samples

Testing Data = 371 samples

Learning Data = 582 samples

Figure 2.6: A - Sample frequency histogram for each available age group and in the lower part of the figure a schematic representation of the amount of samples per age group that are available (orange) for the analysis, containing both the testing (green) and learning (red) data in the respective age groups . B - The minimum amount of samples found across all the age groups, except for the 70-79 age group, in each of the tissues for both female (F) and male (M), that is available for the analysis. C - The minimum amount of samples found across all the age groups in each of the tissues for both female (F) and male (M), that can be used as testing data. D - The minimum amount of samples found across all the age groups, except for the 70-79 age group, in each of the tissues for both female (F) and male (M), that can be used as learning data.

2.6 Correlation Matrices

In each analysis, whether in tissue subsets or age group subsets, the aim was to learn gene-gene relationships. Thus, in this biological context, it was desired to use similarity measures such as Pearson Correlation that captures similarities between patterns (across samples), disregarding value intensities. This way, the correlation matrices were computed for all the tissue subsets and age group subsets using the R statistical environment's built-in ”cor” function. In this context, whether two genes are directly (positively) correlated or inversely (negatively) correlated, they are of interest in both cases. The squared correlations were used to simplify the analysis, which is also a common practice in this field. Given that we are working with high dimension data regarding the amount of gene-gene pair correlations, it was necessary to use strategies that would minimise the size of the resulting correlation matrices. The most adequate approach was found to be multiplying the correlation values by 100, rounding the values to the ones. This way, if a squared correlation value is, for example, 0.72, it becomes 72.

2.7 Hierarchical Clustering of Genes

In each analysis, whether in tissue or age group correlation matrices from learning subsets, a gene clustering step was done by the hierarchical clustering complete linkage method.

Correlation matrices were transformed into distances matrices by subtracting their values from 100, and clustering trees were computed using the R statistical environment built-in ”hclust” function with the complete linkage method. Then, the hierarchical trees were cut at a 0.40 distance threshold (0.60 squared correlation) with the R statistical environment built-in ”cutree” function to define the clusters. Finally, a minimum cluster size filter of 10 genes was used to control the number of clusters obtained. This filtering was done to avoid obtaining many small clusters that are not big enough to be meaningfully considered a biological module or at least are not interesting when we are working with big data.

After clustering, the average squared correlation of all the pairwise combinations within each cluster was computed and named as within-cluster correlation or co-expression from this point on. The within-cluster correlation was also computed in the testing subsets, which is the reason behind those subsets' correlation matrices. All these computed values allowed the visualisation of changes in within-cluster correlation across subsets in a heatmap.

2.8 Heatmapping

The heatmaps allowed to visualise changes in the within-cluster correlation across subsets in a convenient way. Heatmapping was achieved with the ”pheatmap” function, which is implemented as a package for the R statistical environment (R version 4.0.2) and is available [47] as part of the CRAN R repository project.

The default parameters were used except that columns were clustered in the tissue analysis by complete linkage hierarchical clustering with Pearson Correlation between columns as vectors. Moreover, in the age analysis, columns were grouped manually.

2.9 Gene Ontology Enrichment

After heatmapping, some clusters might reveal interesting to delve into. To that end, a Gene Ontology (GO) enrichment was computed for all clusters in both analysis. Gene annotation from the GTEx dataset (v8) was provided as Ensembl ID, which had to be converted to gene symbol, a unique short abbreviation for the gene name. This conversion was done using the ”mapIds” function, which is implemented by the ”AnnotationDbi” package for the R statistical environment and is available [43] as part of the Bioconductor project. For that end, it was used mainly the annotation package ”org.Hs.eg.db” [48] to get ”SYMBOL”, ”GENEBIOTYPE” and ”FULLNAME” annotations, and symbol ID’s were also complemented by the annotation package ”EnsDb.Hsapiens.v79” [49] whenever that correspondence was not found with ”org.Hs.eg.db”. Both annotation packages are available as part of the Bioconductor project [43].

GO enrichment step was done using the ”topGO” package [50] for the R statistical environment available as part of the Bioconductor project [43].

Here follows a description of the used parameters:

- The ”fisher” statistic test to compute the number of significantly annotated genes for each GO term.
- The ”weight01” algorithm to deal with the GO graph structure.
- The gene-to-GO mappings annotation was ”annFUN.org”.
- A node size of 20 to prune the GO hierarchy from the terms with less than 20 annotated genes.
- A p-value cutoff of 0.01.
- An enrichment cutoff of 0.5. Enrichment is computed by the log2 of the quotient of the number of significant genes of a given GO term in a cluster by the expected value given a random chance based on all the genes available.

Three types of GO enrichments were computed: The GO Biological Processes (GO-BP) enrichment (e.g., signal transduction), the GO Molecular Function (GO-MF) enrichment (e.g., ATPase activity) and the GO Cellular Component (GO-CC) enrichment (e.g., ribosome).

The respective Gene Ontology Enrichment’s were Plotted in bar plots with ”ggplot2” package [51] for the R statistical environment available as part of the CRAN R repository project. The bars colour transparency was set proportional to the -log10 of the p-values of each GO term in a given cluster. Thus, the smaller the p-value, the more significantly enriched a GO term is, and the less transparent the respective bar of a GO term is.

2.10 Cluster Correlation Slope with Age Analysis

In the age analysis, it was obtained the within-clusters correlation across several age groups. Then, those correlation values were used to estimate their slope against age, where for each age group, it was assigned a median value. The built-in ”lm” function of the R statistical environment was used to estimate the slope and p-values assuming a $y=m\cdot x+b$ regression type, where ”y” is the vector of a clusters’ within-cluster correlation across ageing and ”x” is the vector of median age values representing each correspondent age group, x=(25,35,45,55,65,75).

2.11 Age Principal Component Analysis

Each correlation matrix corresponding to an age group (20-29, 30-39, 40-49, 50-59, 60-69 and 70-79) was transformed into a single vector of correlations. Then, each age group vector of correlations was inserted as a row of a matrix. Thus, each row of the resulting matrix is a whole age group correlation matrix, and each column is the corresponding gene-gene pair.

A Principal Component Analysis (PCA) was applied to this combined matrix where the data was interpreted as six samples (age groups) with hundreds of millions of features (variables) that are the gene-gene correlations in each of the samples.

PCA was done using the R statistical environment built-in ”prcomp” function with variables being shifted to be zero centred (center=True) and with the variables being scaled to have unit variance (scale=True). Then PC’s were plotted using scatter plot from ”ggplot2” package [51] for the R statistical environment available as part of the CRAN R repository project.

2.12 Age GO Gene Set Enrichment Analysis

After the PCA analysis, the variable loadings of each gene pair in the first Principal Component (PC1) are obtained. These variable loadings are the linear combination coefficients, of the gene-gene pairs, that best describes the data variance. Then these gene-gene pairs are ”split”, and for each gene, the several variable loadings absolute values are summed. So, for each gene, a value is obtained representing all the variable loadings coefficients added up together, representing that genes’ cumulative description of the PC1 from the several gene pairs it composes.

A GO Gene Set Enrichment Analysis (GSEA) is applied to this vector of the sum of variable loadings for each gene. A gene with a higher value means it is one whose interactions with other genes greatly contribute to the data variance in the PC1 direction. Applying GO GSEA is a well-established approach to understand which GO terms are over-represented in the genes with high values of some interesting variable (in this case, the sum of variable loadings). This method is a more consistent approach than arbitrarily filtering the genes by choosing a sum of variable loadings threshold and then applying a GO enrichment analysis to the resulting subset of genes. In fact, sometimes it is challenging to choose a threshold.

GO GSEA was applied using the ”gseGO” function implemented in ”clusterProfiler” package for the R statistical environment (R version 4.0.2) and is available [52] as part of the Bioconductor project [43], using the annotation package ”org.Hs.eg.db” [48]. The GO-BP was assessed using a minimum gene set size of 20 (minGSSize=20). This minGSSize means that the algorithm will not evaluate GO terms which its representing gene set in the data have fewer than 20 genes. It is also used a maximum gene set size of 50 (maxGSSize=50). This maxGSSize means that the algorithm will not evaluate GO terms which its representing gene set in the data have more than 50 genes.

Choosing the minGSSize or maxGSSize is a way to control the amount of GO terms that are returned as enriched and control the GO level that is being assessed. For example, if a particular GO term is represented in the data with 80 genes, using a maxGSSize of 50 means that we want a less broad Go term than that for this analysis and maybe its lower branches might be more specific enough not to be filtered out. The p-value adjustment Benjamini-Hochberg (BH) method was used to limit the False Discovery Rate (FDR) given that the data is big enough for the FDR to be concerning. A p-value cutoff of 0.05 was used.

2.13 Age Graph Network

By applying the before-mentioned approach, the GO GSEA analysis returns the set of GO terms that are enriched when looking at their genes' contributions for the first principal component. For each significantly enriched GO term, GO GSEA also returns the core genes responsible for the enrichment of each GO term. Based on the analysis objectives, the obtained core enriching genes are selected and plotted in graph networks where the interactions (edges) between each gene pair (node) is the respective variable loading value returned from the PC1.

Groups of GO terms were also represented in a graph network with their respective core enriching genes. For this, GO terms from the GSEA were grouped based on similarity by the authors discernment. Regarding this grouping of GO terms, it should be noted that many represented GO terms share genes, and for a gene per GO term group representation to be possible, a gene assignment to each GO term group needed to be done. From the list of all genes, genes were assigned to each GO term group one by one and removed from the list of genes in a specific order. The order was the following: ”Respiratory_ chain”, ”Mitochondrial_ fusion”, ”Protein_ regulation_ Folding”, ”Vitamin_ biosynthetic_ process”, ”Liver_ regeneration”, ”Blood-brain_ barrier”, ”Keratan_ sulfate_ process”, ”Autophagosome”, ”Fibroblast_ proliferation”, ”Membrane_ biogenesis”, ”Cholesterol_ sterol_ process”, ”Carbohydrate_ metabolism”, ”Erythrocyte_ differentiation”, ”Cell_ polarity”, ”Amyloid_Brain_Neurons”, ”B_ T_ cell_ apoptose_ cellcycle_ immune”, ”Epidermal_ growth_ factor”, ”Histone_ modification”, ”Virus”, ”DNA_ metabolism”, ”RNA”, ”Cytoplasmic_ translational_ initiation”, ”Gene_ silencing”, ”Cellular_ responses”, ”Actin”, ”Protein_ de_ auto_ phosphorilation”, ”NAD_ metabolic_ process”, ”GTPase_ regulation”, ”Glycoprotein_ metabolic_ process”, ”Respiratory_ burst”, ”Exocytosis_ vesicle_ docking”, ”Homeostasis”, ”Protein_ import”, ”Cell_ locomotion”, ”Lamellipodium” and ”Behavior”.

The graph networks were plotted using ”qgraph” implemented as a package [53] for the R statistical environment (R version 4.0.2) and is available as part of the CRAN R repository project.

3

Results and Discussion

Contents

3.1 Gene Modules Across and Within Tissues

As introduced, the present work attempts to assess the plausibility of unifying, much more than it already is, tissue-specific data and studies. This can be relevant in cases where there is lack of samples, and also elucidates on the feasibility of predicting gene expression across different tissues or cell types based on a single model.

The data subsets described in chapter 2.5.1 were used in this section, and gene modules were captured by gene-gene squared Pearson correlation and hierarchical clustering as explained in section 2.7.

3.1.1 Three Main Types of Module Behaviour Across Tissues

As described in section 2.6, the correlation between all possible gene pairs was computed using samples from different tissues (”CrossTissue”). Then the analysis was focused on the highly correlated clusters learnt by hierarchical clustering utilising the mentioned correlations as a similarity measure.

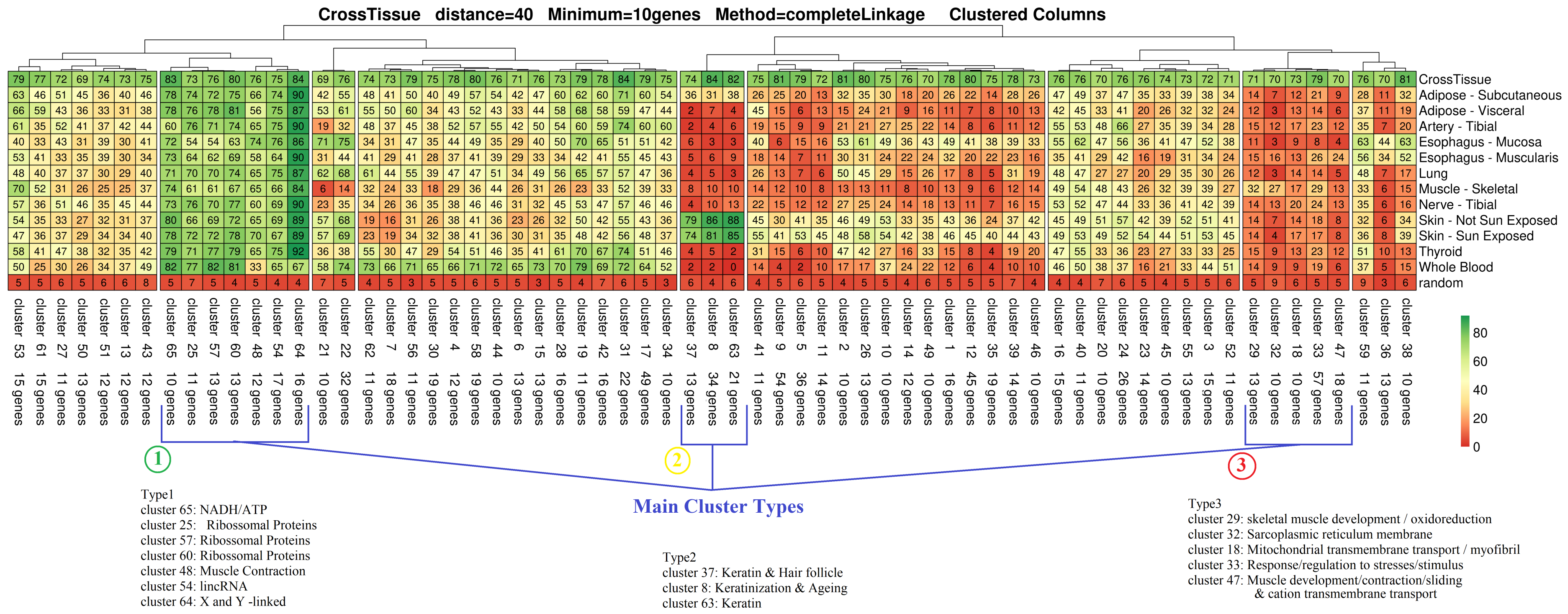
Sixty-five highly correlated clusters across different tissues were obtained. Then, it was analysed how conserved the correlation between cluster members was within specific tissues. This analysis resulted in the heatmap represented in figure 3.1 by following the method described in section 2.8.

This approach was expected to detect mainly three types of clusters regarding their within-cluster correlation conservation within the different tissues. Those are the ones further analysed and highlighted in figure 3.1 by the clustering of the columns (clusters) from the heatmap.

The 3 main types of expected clusters are:

- Clusters with high and stable correlation across most of the tissues (referred to as ”Type1” in this document).
- Clusters with high correlation in some tissues and very low correlation in others (referred to as ”Type2” in this document).
- Clusters with very low correlation in all of the individual tissues (referred to as ”Type3” in this document).

The heatmap obtained in figure 3.1 has its columns clustered in 9 groups by complete linkage and Pearson correlation between columns as vectors. One type of expected clusters (”Type1” from figure 3.1) capture tightly regulated modules of genes that keep their good coordination across most or all the tissues. Finding this kind of modules matches the expectation that gene co-expression networks are not entirely rearranged between tissues and probably cell types. The clusters highlighted as ”Type1” are related to ribosomal proteins, NADH and ATP metabolism, muscle contraction, development and differentiation, lincRNAs, and X and Y linked genes. This relation is based on the GO analyses (section



	cluster 53 15 genes	cluster 61 15 genes	cluster 27 11 genes	cluster 50 13 genes	cluster 51 12 genes	cluster 13 12 genes	cluster 43 12 genes	cluster 65 10 genes	cluster 25 11 genes	cluster 57 13 genes	cluster 60 13 genes	cluster 48 12 genes	cluster 54 17 genes
CrossTissue	79	77	72	69	74	73	75	83	73	76	80	76	75
Adipose - Subcutaneous	63	46	51	45	36	40	46	78	74	72	75	66	74
Adipose - Visceral	66	59	43	36	33	31	38	78	76	78	81	56	75
Artery - Tibial	61	35	52	41	37	42	44	60	76	71	74	65	75
Esophagus - Mucosa	40	33	43	31	39	30	41	72	54	54	63	74	76
Esophagus - Muscularis	53	41	33	35	39	30	34	73	64	62	69	58	64
Lung	48	40	37	37	30	29	40	71	70	70	74	65	75
Muscle - Skeletal	70	52	31	26	25	25	37	74	61	61	67	65	66
Nerve - Tibial	57	36	51	46	35	45	44	73	76	70	77	60	69
Skin - Not Sun Exposed	54	35	33	27	32	31	37	80	66	69	72	65	69
Skin - Sun Exposed	47	36	37	29	34	34	40	78	72	72	78	70	67
Thyroid	58	41	47	38	32	35	42	79	71	77	79	65	75
Whole Blood	50	25	30	26	34	37	49	82	77	82	81	33	65
random	5	5	6	5	6	6	8	5	7	5	4	5	4

	cluster 64 16 genes	cluster 21 10 genes	cluster 22 32 genes	cluster 62 11 genes	cluster 7 18 genes	cluster 56 11 genes	cluster 30 19 genes	cluster 4 12 genes	cluster 58 19 genes	cluster 44 10 genes	cluster 6 13 genes	cluster 15 13 genes	cluster 28 16 genes
CrossTissue	84	69	76	74	73	79	75	78	80	76	71	76	73
Adipose - Subcutaneous	90	42	55	48	41	50	40	49	57	54	42	47	60
Adipose - Visceral	87	53	61	55	50	60	34	43	52	43	33	44	58
Artery - Tibial	90	19	32	48	37	45	38	52	57	55	42	50	54
Esophagus - Mucosa	86	71	75	34	31	47	35	44	49	35	29	40	50
Esophagus - Muscularis	90	31	44	41	29	41	28	37	41	39	29	33	34
Lung	87	62	68	61	44	55	39	47	47	51	34	49	56
Muscle - Skeletal	84	6	14	32	24	33	18	29	36	44	26	31	26
Nerve - Tibial	90	23	35	34	26	36	35	38	46	46	31	32	45
Skin - Not Sun Exposed	89	57	68	19	16	31	26	38	41	36	23	26	32
Skin - Sun Exposed	89	57	69	23	19	34	32	38	41	36	30	31	35
Thyroid	92	36	38	55	30	47	29	54	51	46	35	41	61
Whole Blood	67	58	74	73	66	71	65	66	71	72	65	73	70
random	4	7	5	4	5	3	5	5	6	5	5	3	5

	cluster 19 11 genes	cluster 42 16 genes	cluster 31 22 genes	cluster 17 49 genes	cluster 34 10 genes	cluster 37 13 genes	cluster 8 34 genes	cluster 63 21 genes	cluster 41 11 genes	cluster 9 54 genes	cluster 5 36 genes	cluster 11 14 genes	cluster 2 10 genes
CrossTissue	79	78	84	79	75	74	84	82	75	81	79	72	81
Adipose - Subcutaneous	62	60	71	60	54	36	31	38	26	25	20	13	32
Adipose - Visceral	68	58	59	47	44	2	7	4	45	15	6	13	15
Artery - Tibial	60	59	74	60	60	2	4	6	19	15	9	9	21
Esophagus - Mucosa	70	65	51	51	42	6	3	3	40	6	15	16	53
Esophagus - Muscularis	42	41	55	45	43	5	6	9	18	14	7	11	30
Lung	65	57	57	47	36	4	5	3	26	13	7	6	50
Muscle - Skeletal	23	33	52	39	33	8	10	10	14	12	10	8	13
Nerve - Tibial	53	47	57	46	46	4	10	13	22	15	12	12	27
Skin - Not Sun Exposed	50	42	55	43	36	79	86	88	45	30	41	35	46
Skin - Sun Exposed	48	42	56	48	37	74	81	85	55	41	53	45	48
Thyroid	70	67	74	51	46	4	5	2	31	15	6	10	47
Whole Blood	79	69	72	64	52	2	2	0	14	4	2	10	17
random	4	7	6	5	3	6	4	6	4	5	6	5	4

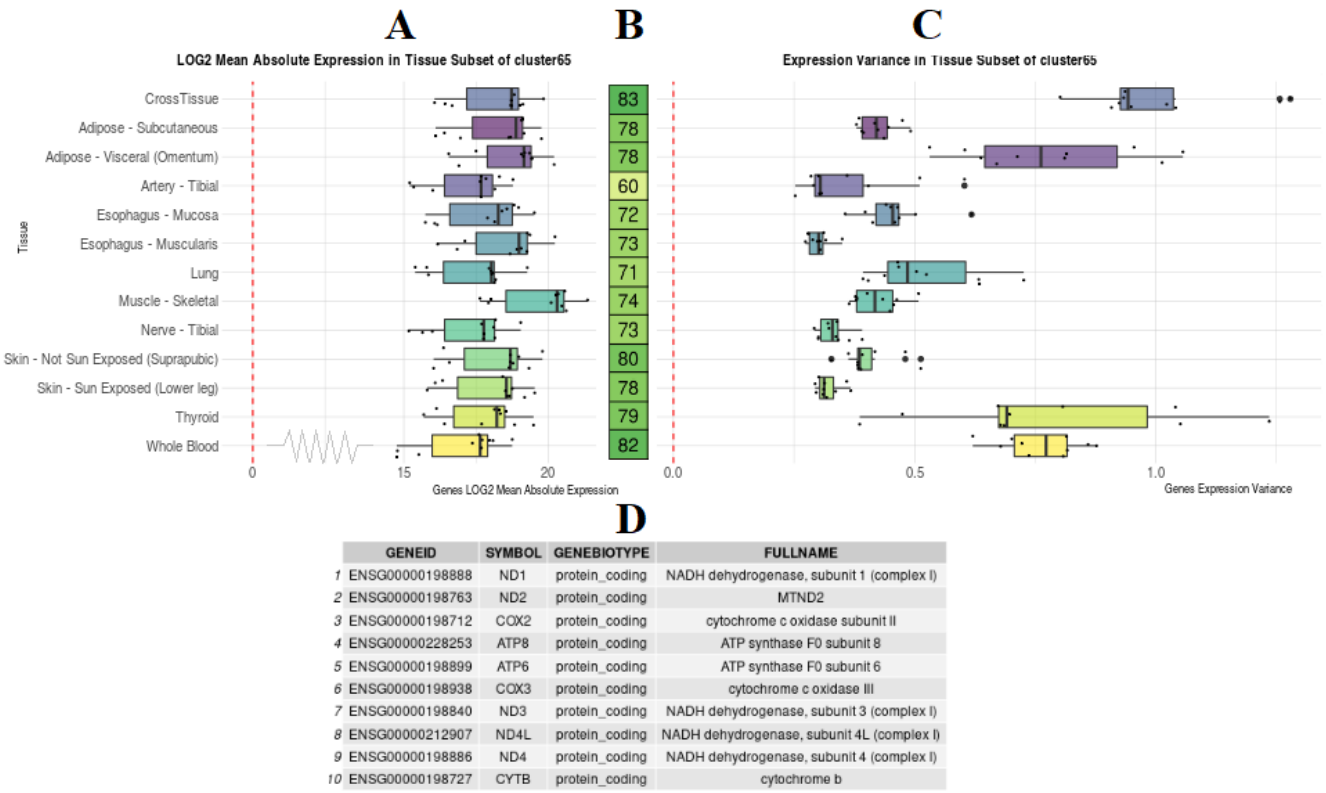
	cluster 26 13 genes	cluster 10 10 genes	cluster 14 12 genes	cluster 49 10 genes	cluster 1 16 genes	cluster 12 45 genes	cluster 35 19 genes	cluster 39 14 genes	cluster 46 10 genes	cluster 16 15 genes	cluster 40 11 genes	cluster 20 10 genes	cluster 24 26 genes
CrossTissue	80	75	76	70	78	80	75	78	73	76	76	70	76
Adipose - Subcutaneous	35	30	18	20	26	22	14	28	26	47	49	37	47
Adipose - Visceral	24	21	9	16	11	7	8	10	13	42	45	33	41
Artery - Tibial	21	27	25	22	14	8	6	11	12	55	53	48	66
Esophagus - Mucosa	61	49	36	43	49	41	38	39	33	55	62	47	56
Esophagus - Muscularis	31	24	22	24	32	20	22	23	16	35	41	29	42
Lung	45	29	15	26	17	8	5	31	19	48	47	27	27
Muscle - Skeletal	13	11	8	10	13	9	6	12	14	49	54	48	43
Nerve - Tibial	25	24	14	18	13	11	7	16	15	53	52	47	44
Skin - Not Sun Exposed	49	53	33	35	43	36	24	37	42	45	49	51	57
Skin - Sun Exposed	58	54	42	38	46	45	40	44	43	49	53	49	52
Thyroid	42	27	16	33	15	8	4	19	20	47	55	33	36
Whole Blood	17	37	24	22	12	6	4	10	10	46	50	38	37
random	5	5	4	6	5	5	5	7	4	4	4	7	6

	cluster 23 14 genes	cluster 45 10 genes	cluster 55 13 genes	cluster 3 15 genes	cluster 52 11 genes	cluster 29 13 genes	cluster 32 10 genes	cluster 18 10 genes	cluster 33 57 genes	cluster 47 18 genes	cluster 59 11 genes	cluster 36 13 genes	cluster 38 10 genes
CrossTissue	76	74	73	72	71	71	70	73	79	70	76	70	81
Adipose - Subcutaneous	35	33	39	38	34	14	7	12	21	9	28	11	32
Adipose - Visceral	20	26	32	32	24	12	3	13	14	6	37	11	19
Artery - Tibial	27	35	39	34	28	15	12	17	23	12	35	7	20
Esophagus - Mucosa	36	41	47	52	38	11	3	9	8	4	63	44	63
Esophagus - Muscularis	16	19	31	34	24	15	16	13	26	24	56	34	52
Lung	20	29	35	30	26	12	3	14	14	5	48	7	17
Muscle - Skeletal	26	32	39	39	27	32	27	17	29	13	33	6	15
Nerve - Tibial	33	36	41	42	39	14	13	20	24	13	35	6	16
Skin - Not Sun Exposed	42	39	52	51	41	14	7	14	18	8	32	6	34
Skin - Sun Exposed	54	44	51	53	45	14	4	17	17	8	36	8	39
Thyroid	23	27	33	33	25	15	9	13	23	12	51	10	13
Whole Blood	16	21	33	44	51	14	9	9	19	6	37	5	15
random	5	4	5	5	6	5	9	6	5	5	9	3	6

Figure 3.1: Heatmap of within-cluster squared Pearson correlation (x100) within different tissues. Colour scale reflects the correlation values. The 65 gene clusters were learnt by complete linkage hierarchical clustering in the ”CrossTissue” sample subset (first line of the heatmap) with a minimum cluster size filter of 10 genes and a squared Pearson correlation clustering threshold of 0.60. Within-cluster correlation was computed in the remaining tissue sample subsets (lines 2-13 of the heatmap) and the last line of the heatmap carries the within-cluster correlation of equally sized random clusters from ”CrossTissue” subset. Heatmap columns are clustered into 9 groups by complete linkage and Pearson correlation between columns. Beneath the heatmap 3 main types of clusters are identified according to their correlation patterns across tissues. A very broad identification of clusters is assigned to the main cluster types based on their respective enriched GO terms and gene annotation available in supplemental material sections A.2 and A.1, respectively.

A.2) and gene annotation (section A.1), of the corresponding clusters, present in the supplemental material. These clusters were captured as highly correlated across tissues and within tissues. For this to be possible, they should be active genes in those tissues (well detectable expression) and/or have enough expression variance in those tissues for correlation to be adequately captured if there is actual co-expression. NADH/ATP and ribosomal protein clusters were expected to have been captured in this highly coordinated fashion because they represent housekeeping genes.

The mean expression and variance of genes in cluster 65 in each tissue subset (and cross-tissue) is represented in the boxplots in figure 3.2, also exemplifying the similarly behaved remaining clusters from ”Type1” group of clusters, except for cluster 54. Cluster 54 is mainly composed of lincRNAs with extremely low expression values. It is though that cluster 54 expression values do not distinguish themselves from RNA-seq noise. So the question is if the captured high correlation values are noise-driven or biological-signal-driven. If it is biological-signal-driven, this might be an interesting functional module to delve into. Otherwise, if it is noise-driven, the only proposed explanation is that some technical factors might influence these low expressed genes in a consistent way across samples.



	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000198888	ND1	protein_coding	NADH dehydrogenase, subunit 1 (complex I)
2	ENSG00000198763	ND2	protein_coding	MTND2
3	ENSG00000198712	COX2	protein_coding	cytochrome c oxidase subunit II
4	ENSG00000228253	ATP8	protein_coding	ATP synthase F0 subunit 8
5	ENSG00000198899	ATP6	protein_coding	ATP synthase F0 subunit 6
6	ENSG00000198938	COX3	protein_coding	cytochrome c oxidase III
7	ENSG00000198840	ND3	protein_coding	NADH dehydrogenase, subunit 3 (complex I)
8	ENSG00000212907	ND4L	protein_coding	NADH dehydrogenase, subunit 4L (complex I)
9	ENSG00000198886	ND4	protein_coding	NADH dehydrogenase, subunit 4 (complex I)
10	ENSG00000198727	CYTB	protein_coding	cytochrome b

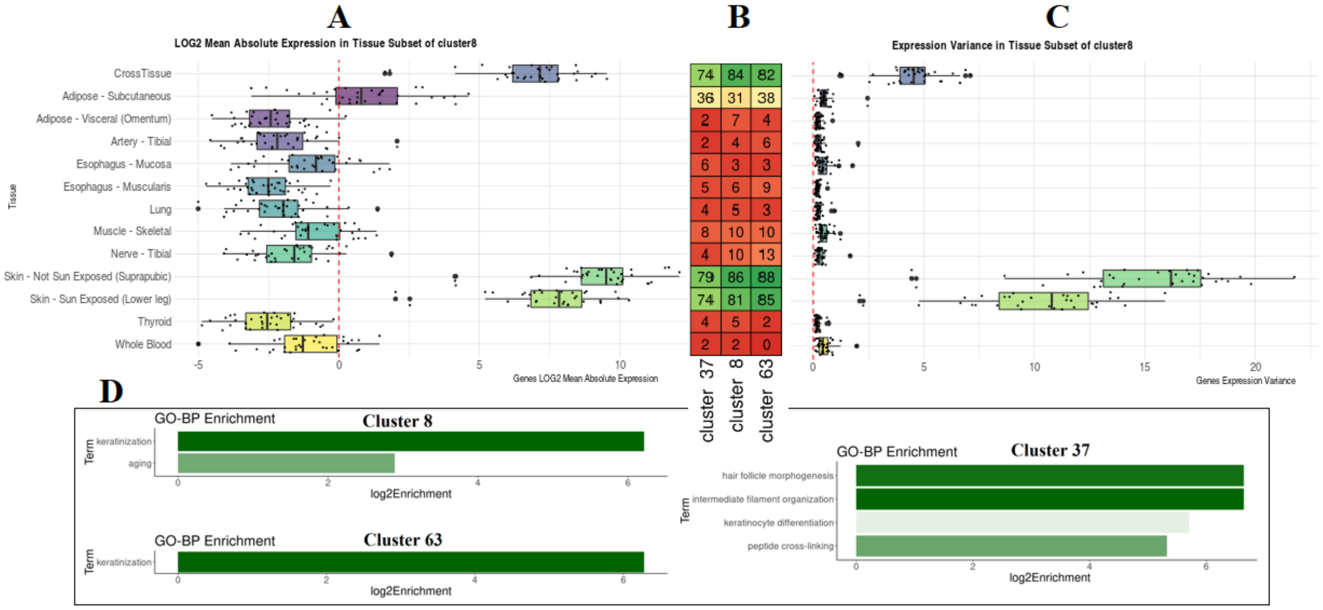
Figure 3.2: Characterization of gene cluster 65. A - Boxplots with the respective data points, where each data point is a gene from cluster 65. Gene's values are the absolute expression averaged across the corresponding tissue sample subsets in a log2 scale. B - Within-cluster 65 correlation across and within tissues taken from figure 3.1. This heatmap column has its rows aligned with the respective sample subsets (tissues) from panel A. C - Boxplots with the respective data points, where each data point is a gene from cluster 65. Gene's values are the expression variance across the corresponding tissue sample subsets. This plot (C) has its rows aligned with the respective sample subsets (tissues) from panel A. D - Cluster 65 genes annotation.

As shown in figures 3.2.A and 3.2.C, NADH/ATP related genes of cluster 65 have high expression levels in all tissues but a fairly low variance, meaning that even with low variance, their expression levels are so tightly coordinated that a high correlation can still be found. Figure 3.2.D shows the annotation of genes in cluster 65. Cluster 65 genes include cytochrome, NADH dehydrogenases and ATP synthase genes implicated in respiratory electron transport. Here, it can be observed that genes participating in functional modules such as this one can be co-expressed even at the tissue scale and within several tissues. This is consistent with the expectation that many cellular processes which require a specific stoichiometry of their molecular components to be operational, independently of tissue type, must be universally co-regulated.

Observing figure 3.1, it can be seen that there is a great portion of captured clusters with stable co-expression across tissues, even if with moderate correlation values. In this analysis, 7 out of 65 clusters were captured as stable in several tissues with extremely high correlation values within the tissues.

However, a gene cluster does not need a correlation as high as 0.60 for its genes to be considered co-expressed. Therefore, gene clusters with stable correlation values between 0.30 and 0.40 within several tissues are still potentially co-regulated. And, as expected, a considerable amount of that kind of clusters was found. This might mean that many tissue-specific data and studies can be unified to some extent much more than is currently done.

It should be noted that the squared Pearson correlation of random equally sized clusters roams the 0.05 values. Additionally, some thresholds were applied in this analysis; consequently, there are expected to be several more clusters that would behave similarly in terms of a stable co-expression preservation within specific tissues. Meaning that, probably, with less stringent thresholds, it could have been detected possibly co-regulated clusters with stable correlation values within several tissues but without correlation values as high as the ones that were captured in the ”CrossTissue” subset. The same applies to clusters with less than 10 genes that could have been captured.



cluster 37	cluster 8	cluster 63
74	84	82
36	31	38
2	7	4
2	4	6
6	3	3
5	6	9
4	5	3
8	10	10
4	10	13
79	86	88
74	81	85
4	5	2
2	2	0

Figure 3.3: Characterization of gene clusters 37, 8 and 63. A - Boxplots with the respective data points, where each data point is a gene from cluster 8. Gene’s values are the absolute expression averaged across the corresponding tissue sample subsets in a log2 scale. B - Within-cluster 37, 8 and 63 correlation across and within tissues taken from figure 3.1. These heatmap columns have their rows aligned with the respective sample subsets (tissues) from panel A. C - Boxplots with the respective data points, where each data point is a gene from cluster 8. Gene’s values are the expression variance across the corresponding tissue sample subsets. This plot (C) has its rows aligned with the respective sample subsets (tissues) from panel A. D - Cluster 37, 8 and 63 respective GO-BP enrichment analysis.

Regarding a second type of expected clusters (”Type2” from figure 3.1), they would have a high within-cluster correlation in some tissues and a very low correlation in others. This type is the case of clusters 37, 8 and 63 present in figure 3.3. These clusters are mainly composed of keratins and keratin-associated protein genes. This can be observed in supplemental material section A.1.

Keratins are the major structural proteins of the vertebrate epidermis, constituting up to 85% of a fully differentiated keratinocyte, forming Keratin Intermediate Filaments (KIFs) which are a critical component of the *stratum corneum*, the outermost layer of the epidermis [54]. Keratin Associated Proteins (KAPs) are responsible for forming the protein matrix between the KIFs [55].

Figure 3.3.A shows that these clusters of genes (using cluster 8 as an example) are highly expressed in skin tissues (sun and non-sun exposed) and as presumed highly correlated. They also have a high variance in skin tissues (figure 3.3.C). The remaining mean expression and variance plots are available in supplemental material section A.3. Except for skin tissues and ”Adipose - Subcutaneous” tissue the correlation of these clusters is as low as random equally sized clusters which appears to be a consequence of their very low expression levels that might mean that the respective genes are inactive or just don't have variance enough for correlation to be captured.

Interestingly, clusters 37, 8 and 63 genes are moderately correlated in ”Adipose - Subcutaneous” tissue accompanied by moderate expression levels even though there is minimal variance. This pattern might be explained by the physical proximity of the ”Adipose - Subcutaneous” tissue with the skin tissue. They might share a similar microenvironment, and there might be some signalling molecules that can make Subcutaneous adipose tissue cells have expression coordination patterns within these clusters genes. Furthermore, keratins and KAPs are also the most abundant structural proteins in hair, and their intermediate filaments are primarily responsible for hair's mechanical properties [55]. Actually, during hair growth, hair follicles delve deep into the rich dermal macroenvironment where adipocyte progenitor cells are activated to proliferate and form new mature adipocytes that surround the hair follicle [56]. This event of hair follicle-adipocyte communication happens in intradermal adipose tissue, which is in very close proximity with subcutaneous adipose tissue. Otherwise, it can be sample contamination from a neighbouring tissue (such as skin tissue) upon extraction of the sample.

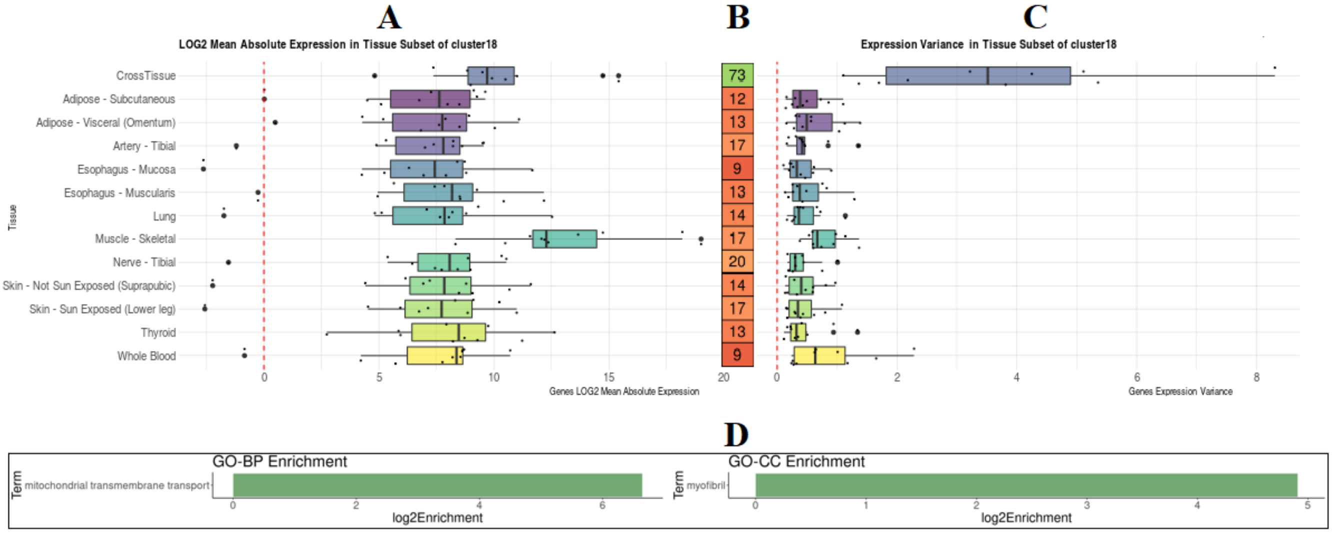


Figure 3.4: Characterization of gene cluster 18. A - Boxplots with the respective data points, where each data point is a gene from cluster 18. Gene's values are the absolute expression averaged across the corresponding tissue sample subsets in a log2 scale. B - Within-cluster 18 correlation across and within tissues taken from figure 3.1. This heatmap column has its rows aligned with the respective sample subsets (tissues) from panel A. C - Boxplots with the respective data points, where each data point is a gene from cluster 18. Gene's values are the expression variance across the corresponding tissue sample subsets. This plot (C) has its rows aligned with the respective sample subsets (tissues) from panel A. D - Cluster 8 GO-BP and GO-CC enrichment analysis.

Then there is a third type of expected clusters ("Type3" from figure 3.1) where the correlation would be low in all of the tissues. This can be because they either have very low variance within tissues, or because the genes just aren't that much co-expressed within tissues. Hereupon, these clusters would only have been captured because they change expression levels in a coordinated enough fashion between tissues for that pattern to be captured as a good correlation across tissues ("CrossTissue" sample subset) within the chosen threshold.

Actually, the "Type3" grouped clusters in figure 3.1 appear to be related to muscle development and function (figure 3.4.D and supplemental material sections A.1 and A.2). In fact, if we look at cluster 18 mean expression in figure 3.4.A, it is clearly overexpressed in skeletal muscle tissue compared with the other tissues. This coordinated overexpression in skeletal muscle tissue creates a 'step' in expression from other tissues to muscle tissue when looking at the 'CrossTissue' subset. That coordinated overexpression must be a main driver of the high squared correlation (0.73) captured for this cluster 18 across tissues. The remaining mean expression and variance plots are available in supplemental material section A.3.

If we look at cluster 18 GO enrichment analysis (figure 3.4.D), it is observed enrichment in the GO-BP term of "mitochondrial transmembrane transport" as well as the GO-CC term of "myofibril". Both terms (and respective genes) might not be directly related, but they can be part of muscle-specific functions;

thus, both having an increase in expression in the muscle tissue creates expression coordination across tissues.

This coordination is reasonable because myofibril is a rod-like organelle of a muscle cell responsible for muscle contraction, comprising approximately 80% of the volume of a whole muscle [57] and regarding ”mitochondrial transmembrane transport”, skeletal muscle has different types of mitochondria than most tissues which possess subtle differences in biochemical and functional properties and distinct subcellular regions [58]. The most abundant mitochondria type in skeletal muscle is Intermyofibrillar (IMF) mitochondria, located in close contact with the myofibril and found to have higher rates of protein synthesises, enzyme activities, and respiration [58].

This observed pattern between the biological process GO term of ”mitochondrial transmembrane transport” and the GO-CC term of ”myofibril” is rather interesting and might be an exemplar case of genes that, by participating in associated functional modules, need to be upregulated in a concerted way at very different cellular scales, including at the level of entire organelles (e.g. mitochondria and myofibrils).

3.2 Gene Modules Across Ageing And Within Age Groups

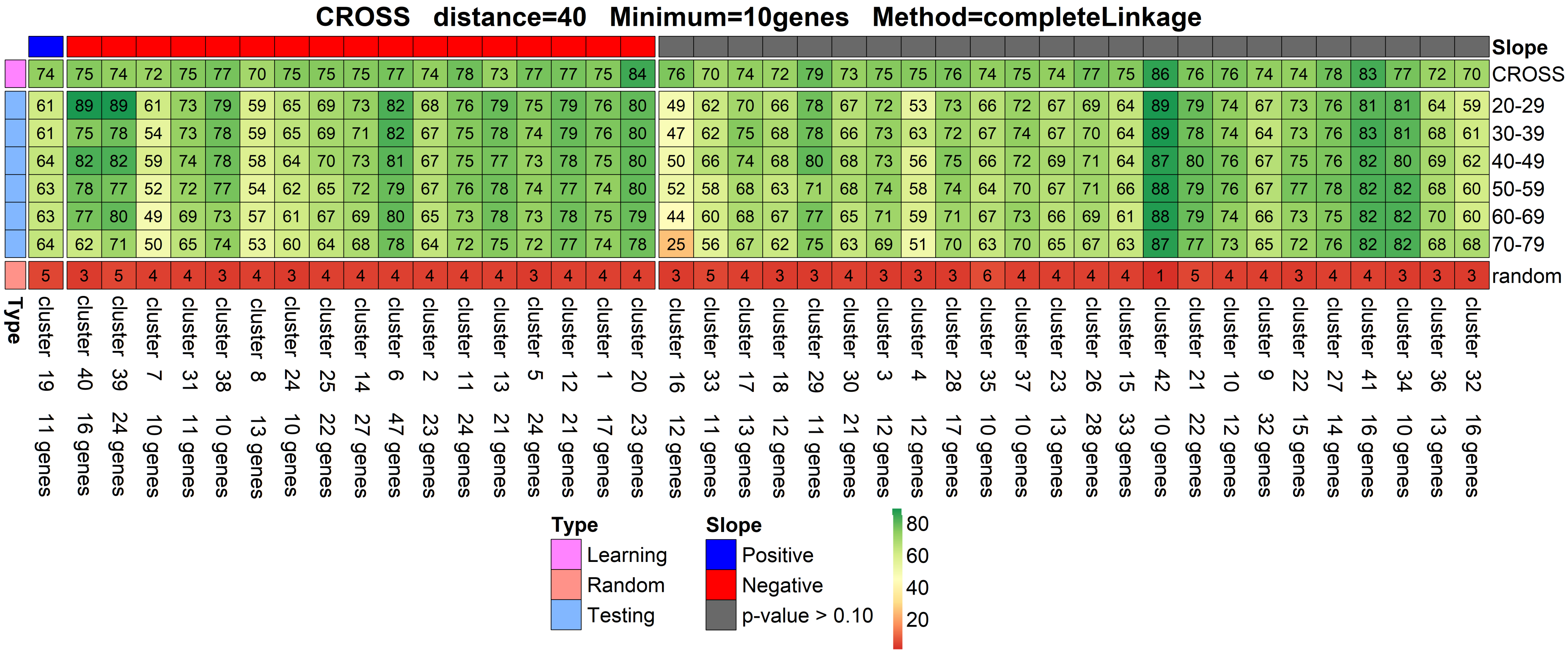
As explained in the methodologies section, gene modules were learnt in a cross-tissue and cross-age approach and then evaluated in the several age group subsets present in figure 3.5.

In figure 3.5, the clusters that significantly ($p\text{-value}<0.10$) decreased their within-cluster correlation with age are represented in the red tab columns, and the further to the left, the higher the decrease in within-cluster correlation across ageing (linear regressed slope against age group vector $\{25,35,45,55,65,75\}$). A p-value of 0.10 was considered as significant because the decrease in within-cluster correlation across ageing is not expected to be strictly linear. Or at least it was desired to capture decreases in correlation that could slightly deviate from a linear pattern.

As expected, it was obtained several clusters with a significant decrease in correlation across ageing. However, even though some clusters have only a slight decrease in correlation, it was expected to be slight because what is reasonable is small inefficiencies accumulated in the eventual cellular pathways and not actual changes in the regulatory network that would cause abrupt changes in correlation. The subsequent analysis will characterise the six most prominent clusters that decrease correlation with ageing, trying to understand its meaning and establishing some hypotheses.

3.2.1 Keratin Clusters

The first 2 clusters with the highest decrease in correlation across ageing are cluster 39 (16 genes) and cluster 40 (24 genes) which are all keratin-associated proteins in both clusters. Their respective



	cluster 19 11 genes	cluster 40 16 genes	cluster 39 24 genes	cluster 7 10 genes	cluster 31 11 genes	cluster 38 10 genes	cluster 8 13 genes	cluster 24 10 genes	cluster 25 22 genes	cluster 14 27 genes	cluster 6 47 genes	cluster 2 23 genes	cluster 11 24 genes	cluster 13 21 genes	cluster 5 24 genes	cluster 12 21 genes	cluster 1 17 genes	cluster 20 23 genes	cluster 16 12 genes	cluster 33 11 genes	cluster 17 13 genes	cluster 18 12 genes	cluster 29 11 genes	cluster 30 21 genes	cluster 3 12 genes	cluster 4 12 genes	cluster 28 17 genes	cluster 35 10 genes	cluster 37 10 genes	cluster 23 13 genes	cluster 26 28 genes	cluster 15 33 genes	cluster 42 10 genes	cluster 21 22 genes	cluster 10 12 genes	cluster 9 32 genes	cluster 22 15 genes	cluster 27 14 genes	cluster 41 16 genes	cluster 34 10 genes	cluster 36 13 genes	cluster 32 16 genes
Slope	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛
CROSS	74	75	74	72	75	77	70	75	75	75	77	74	78	73	77	77	75	84	76	70	74	72	79	73	75	75	76	74	75	74	77	75	86	76	76	74	74	78	83	77	72	70
20-29	61	89	89	61	73	79	59	65	69	73	82	68	76	79	75	79	76	80	49	62	70	66	78	67	72	53	73	66	72	67	69	64	89	79	74	67	73	76	81	81	64	59
30-39	61	75	78	54	73	78	59	65	69	71	82	67	75	78	74	79	76	80	47	62	75	68	78	66	73	63	72	67	74	67	70	64	89	78	74	64	73	76	83	81	68	61
40-49	64	82	82	59	74	78	58	64	70	73	81	67	75	77	73	78	75	80	50	66	74	68	80	68	73	56	75	66	72	69	71	64	87	80	76	67	75	76	82	80	69	62
50-59	63	78	77	52	72	77	54	62	65	72	79	67	76	78	74	77	74	80	52	58	68	63	71	68	74	58	74	64	70	67	71	66	88	79	76	67	77	78	82	82	68	60
60-69	63	77	80	49	69	73	57	61	67	69	80	65	73	78	73	78	75	79	44	60	68	67	77	65	71	59	71	67	73	66	69	61	88	79	74	66	73	75	82	82	70	60
70-79	64	62	71	50	65	74	53	60	64	68	78	64	72	75	72	77	74	78	25	56	67	62	75	63	69	51	70	63	70	65	67	63	87	77	73	65	72	76	82	82	68	68
random	5	3	5	4	4	3	4	3	4	4	4	4	4	4	3	4	4	4	3	5	4	3	3	4	3	3	3	6	4	4	4	4	1	5	4	4	3	4	4	3	3	3

Figure 3.5: Heatmap of within-cluster squared Pearson correlation (x100) within different age group subsets. The colour scale reflects the correlation values. The 42 gene clusters were learnt by complete linkage hierarchical clustering in the ”Cross” sample subset (first line of the heatmap) with a minimum cluster size filter of 10 genes and a squared Pearson correlation clustering threshold of 0.60. Within-clusters correlation was computed in each of the age group sample subsets (lines 2-7 of the heatmap) and the last line of the heatmap carries the within-cluster correlation of equally sized random clusters from the ”Cross” subset. Heatmap columns are grouped into 3 groups according to the columns vector linear regressed slope against age group vector ({25,35,45,55,65,75}). Blue means positive slope with the linear regression p-value<0.10; red means negative slope with the linear regression p-value<0.10; grey is any slope that has the linear regression p-value>0.10. Significant negative sloped (red) columns are ordered with increasing absolute slope values to the left.

GO enrichment analysis is available in figure 3.6 and the genes annotation is available in supplemental material in section A.4.

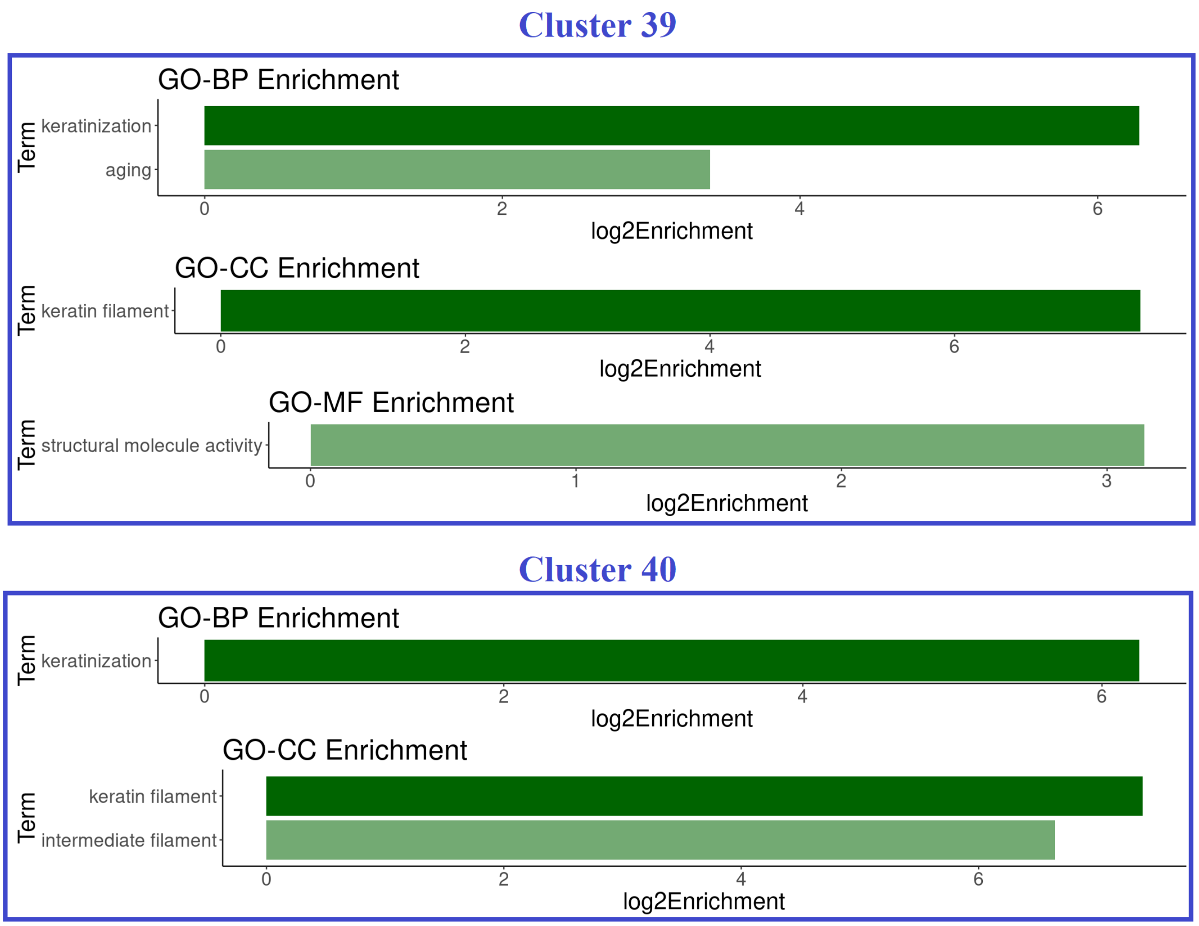


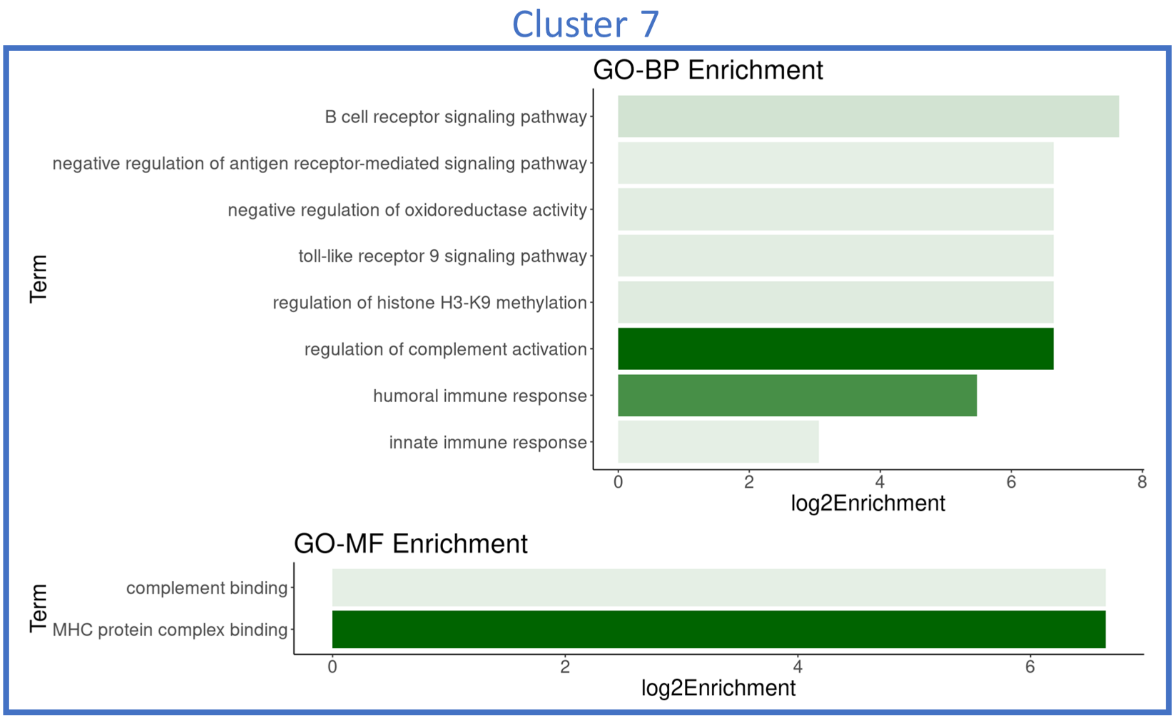
Figure 3.6: Keratin related clusters 39 and 40 that have been derived from the clustering illustrated in figure 3.5. Cluster 39 genes GO-BP, GO-CC and GO-MF enrichment analysis results and cluster 40 genes GO-BP and GO-CC enrichment analysis results.

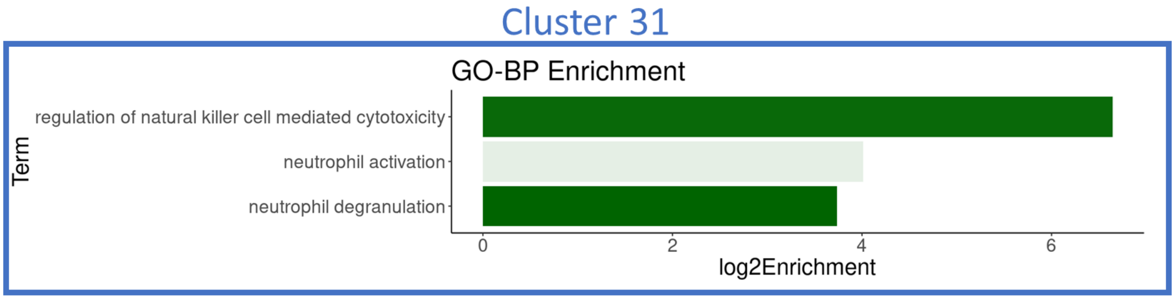
Remembering, keratins are the major structural proteins of the vertebrate epidermis, constituting up to 85% of a fully differentiated keratinocyte, forming KIFs which are a critical component of the *stratum corneum*, the outermost layer of the epidermis [54]. *Stratum corneum* KIFs are of major importance for the barrier properties of skin, the water-holding capacity of the skin, the mechanical strength and elastic resilience of skin, and skin pathologies [59]. A decline of those skin properties, as well as wrinkle formation, is a common sign of ageing [54]. In addition, studies [60] found relationships between fine wrinkle formation, loss of elastic properties of the epidermis and KIFs disruption that might be caused by alteration of keratin expressions which are strictly regulated in a keratinocyte proliferation/differentiation-specific manner. In the present study, it is observed that this strict regulation appears to loosen across ageing, at least in some of the keratin-associated genes. Therefore, it might be insightful to assess which gene pairs within clusters 39 and 40 drive the most decrease of the within-cluster correlation of the said clusters.

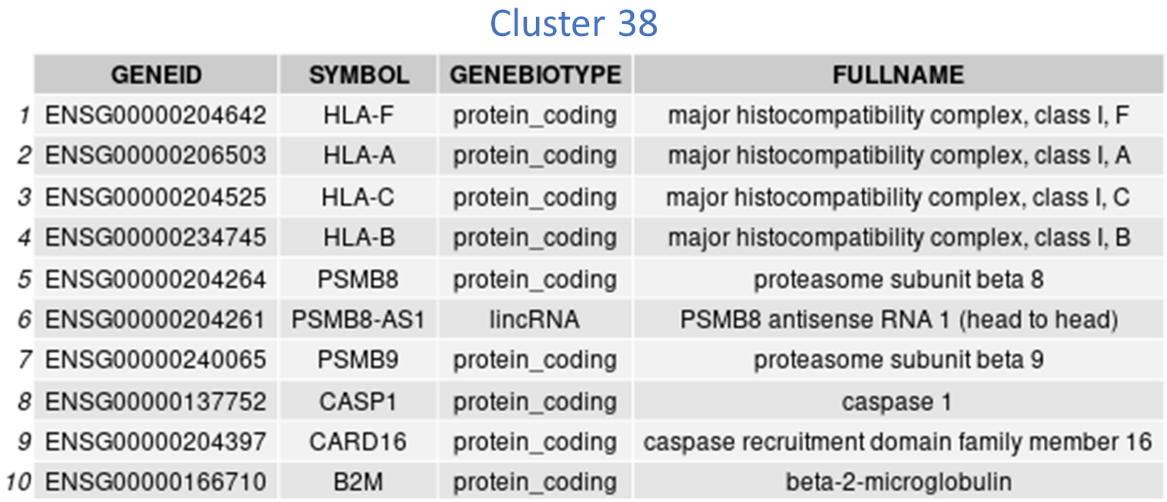
Additionally, keratins are the most abundant structural proteins in hair, and their intermediate filaments are primarily responsible for hair's mechanical properties, being that their appropriate synthesis should be a requirement to maintain hair's juvenescent properties [55]. The same goes for KAPs such as the ones present in clusters 39 and 40, responsible for forming the protein matrix between the keratin intermediate filaments, equally playing a crucial role in forming a strong hair shaft. Actually, KAP4 gene family, which is abundant in cluster 39, not only represents the largest KAP family but it is also suggested that their substantial decline in gene expression reduces hair shaft stability and flexibility [55]. Keratins and keratin-associated proteins are also profoundly related to cell polarity, shape, mitotic activity, cell signalling, and intracellular vesicle transport [61].

3.2.2 Immune System Clusters

After the keratin clusters, a set of immune-related clusters were obtained. This observation is consistent with the consensual decline of immune system functionality across ageing [62].







	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000204642	HLA-F	protein_coding	major histocompatibility complex, class I, F
2	ENSG00000206503	HLA-A	protein_coding	major histocompatibility complex, class I, A
3	ENSG00000204525	HLA-C	protein_coding	major histocompatibility complex, class I, C
4	ENSG00000234745	HLA-B	protein_coding	major histocompatibility complex, class I, B
5	ENSG00000204264	PSMB8	protein_coding	proteasome subunit beta 8
6	ENSG00000204261	PSMB8-AS1	lincRNA	PSMB8 antisense RNA 1 (head to head)
7	ENSG00000240065	PSMB9	protein_coding	proteasome subunit beta 9
8	ENSG00000137752	CASP1	protein_coding	caspase 1
9	ENSG00000204397	CARD16	protein_coding	caspase recruitment domain family member 16
10	ENSG00000166710	B2M	protein_coding	beta-2-microglobulin

Figure 3.7: Cluster 7 genes GO-BP and GO-MF enrichment analysis results; cluster 31 genes GO-BP enrichment analysis results; cluster 38 genes annotation. Those are immune system-related clusters derived from the clustering illustrated in figure 3.5.

Observing cluster 7 GO enrichment analysis in figure 3.7, it appears that it mainly involves immune responses by the complement system. The complement system is a major component of the innate immune system and also plays an important role in adaptive immunity, such as the humoral one present in the GO-BP terms [63]. Its main biological function is to recognise damaged or altered "self" components, such as apoptotic and necrotic cells, abnormal protein assemblies (e.g. amyloids, clots or antibody aggregates), or "foreign" materials such as particles, macromolecules or microorganisms, promoting their elimination either by opsonisation (enhancing their uptake by phagocytic cells) or, if they have a lipid bilayer membrane (e.g. bacteria) by lysing them [63]. However, an overactive system can cause autoimmune and inflammatory diseases such as Age-related Macular Degeneration (AMD), whereas an inactive complement system results in an increased risk for infection [63].

For instance, despite great progress in uncovering its genetic links, AMD remains an incurable disease. Maybe because AMD is not entirely a genetic disease, but also has equally important risk factors like physiological changes that occur with age and lifestyle, such as smoking and nutrition [64].

The present study might be able to give insights into the primary cause of the ageing physiological changes that contribute to AMD establishing a link with genetic reasoning. Differential expression studies might not be enough to perceive the genetic links between or within pathways that are increasingly impaired with age, decreasing it's co-expression.

As for cluster 31, it refers to Natural Killer (NK) cells and Neutrophils. Neutrophils, the most abundant cell type in human blood, are phagocytic leukocytes that comprise the first line of host immune response against invading pathogens, being important effector cells in the innate arm of the immune system [65]. Their three main antimicrobial mechanisms are phagocytosis, degranulation, and the release of nuclear material in the form of neutrophil extracellular traps [66]. More recently, it was discovered that neutrophils possess a much broader set of roles that go beyond antimicrobial responses. Actually, neutrophils respond to multiple signals by producing several cytokines and other inflammatory factors that influence and regulate inflammation and also the immune system homeostasis and even actively participate in several diseases including cancer [66].

Regarding ageing, neutrophils mediate the immediate host response to bacterial and fungal infections, which are largely responsible for the higher rates of mortality and morbidity in the elderly population [67]. Neutrophil function has been described [67] to decline with age and to be a significant factor in immune senescence, but little is known about the molecular basis of this loss of function.

NK cells are one of the major mediators of cellular cytotoxicity. This is the ability to kill other cells, which is an important effector mechanism of the immune system to combat viral infections and cancer [68].

With age, significant impairments have been reported in the main mechanisms by which NK cells confer host protection [62]. Actually, the age-associated decline in NK cell function has been associated with slower resolution of inflammatory responses, increased susceptibility of bacterial, viral and fungal infections, being that NK cells are also involved in the recognition, and elimination of senescent cells [62].

It is proposed that looking into the genes composing cluster 31 and their decrease in correlation with age might give insight into the molecular basis of NK cells and neutrophils age-related loss of function.

Regarding cluster 38, it is evident that it represents the system of Major Histocompatibility Complex (MHC) class I. Class I MHC molecules bind peptides generated mainly from the degradation of cytosolic proteins by the proteasome and display those peptides to the cell's exterior by being inserted in the external plasma membrane. This external display of peptides has the intent of exhibiting them to Cytotoxic T Cells (CTLs).

The repertoire of peptides presented by MHC class I molecules in a given set of cells is termed the immunopeptidome. This action of displaying the immunopeptidome has mainly three objectives. One is to display peptides from normal cellular protein turnover for the cells to be recognized as not foreign (compatible) by CTLs [69]. A second one is for the CTLs to recognize tumour cells by displaying malignant characteristic immunopeptidomes [70]. And the third one is for the CTLs to recognize virus-infected cells that display foreign peptides in their immunopeptidome [69].

Both in the case of tumour cells or virally infected cells, CTLs release cytotoxins into the target cells triggering the caspase cascade, eventually leading to apoptosis (programmed cell death) [71]. And that is why it is reasonable for the CASP1 (caspase 1) and CARD16 (caspase recruitment domain family member 16) to be present in cluster 38.

In cluster 38 (figure 3.7), beta-2-microglobulin (B2M) gene also makes sense because it is a component of MHC class I [72]. The PSMB8 (proteasome subunit beta 8), PSMB9 (proteasome subunit beta 9) and PSMB8-AS1 (PSMB8 antisense RNA) regard to components of the proteasome that, when recruited by interferon-γ, make the proteasome become an immunoproteasome, which is the one that generates the peptides that constitute the immunopeptidome [73]. Interestingly, B2M was reported [31] with significant ($p<0.0001$) age-related increase in cell-to-cell gene expression variation in young versus old mouse cardiomyocytes.

As discussed before, ageing is associated with an increasingly insufficient immune response, and MHC I decrease in coordination across ageing may play an important part in this process.

The age-related insufficient immune response may lead to the initiation and progression of various malignancies [74]. For example, in Bladder Cancer (BC), which is prevalent in elderly patients, there is much interest in the activation of patients' CTLs and efficient presentation of BC antigens by MHC class I molecules.

Additionally, MHC class I proteins were very recently found to be critical for maintaining neuronal structural complexity in the ageing brain [75]. During ageing, there are substantial changes in neuronal complexity, structural reorganization of dendritic spines and disturbances in synaptic signalling. These changes are thought to underlie age-related impairments in learning and memory that may occur during healthy ageing. However, the molecular mechanisms that account for these age-related structural and synaptic alterations have not been fully illuminated, and little is known about MHC class I function in the ageing brain [75]. Looking at the gene-gene pairs of MHC class I that decrease the most in correlation across ageing might provide insight into this matter.

Curiously, it was also very recently found [76] that accumulation of CTLs in the ageing mouse central nervous system leads to axon degeneration and contributes to cognitive and motor decline. So here it is proposed that nervous system functional ageing decline might, in part, be caused by irregular action of CTLs in the brain as a consequence of MHC class I deregulation across ageing.

3.3 Genome-Wide Gene-Gene Relationships Across Ageing

3.3.1 Principal Component Analysis Across Ageing

This section makes use of a matrix very similar to the table present in figure 3.5 where each row represents an age group, but each column now represents a single gene-gene pair. The values are the respective gene-gene pair squared Pearson correlation in each of the age groups. Thus, rows of this matrix contain all the values from the respective tissue correlation matrix.

As explained in the methods section 2.11 a PCA was applied to the rows (age groups) of this matrix where the variables were the hundreds of millions of gene-gene pairs. The resulting PC1 from this analysis follows in figure 3.8.B as well as the respective scree plot of the PCs in figure 3.8.A.

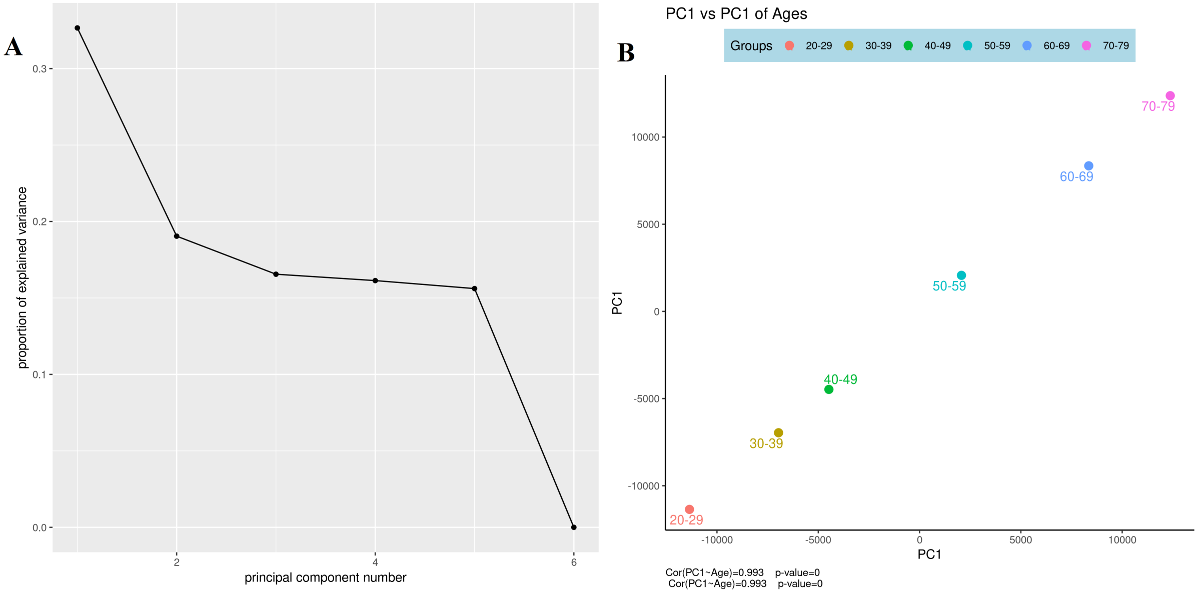


Figure 3.8: A- Scree Plot of the PCA of age groups correlations, where the variables are the gene-gene pairs squared Pearson correlation within those age groups. B- PC1 plot (PC1 in both of the 2D axis) of the PCA of age groups, where the variables are the gene-gene pairs squared Pearson correlation. The data points are the age groups whose coordinates in PC1 are calculated by a linear combination of their gene-gene pairs squared Pearson correlation. This linear combination is characterised by coefficients that give more or less weight to the respective gene-gene pairs in explaining the data variation in PC1 direction by means of its squared Pearson correlation. Bellow the plot it is provided the Pearson correlation between PC1 age groups coordinates and mean age groups vector ({25,35,45,55,65,75}) suggesting that PC1 direction aligns significantly with ageing.

Gratifyingly, the principal component that most describes the data variance (>30%) is the one and only that accurately describes the greatest variance of the data in the direction of ageing. Actually, age groups PC1 coordinates have a Pearson correlation with ageing (mean age groups vector {25,35,45,55,65,75}) of 0.993 with a p-value of 0.00008. The remaining PCs are represented in supplemental material figure A.18. This is accurate enough to interpret PC1 variable (gene-gene pairs correlation) loadings as a way to measure the contribution that a specific gene-gene pair provides in explaining ageing data variation by means of its genes squared correlation across ageing. From this point on, a gene-gene pair with high positive PC1 loading is regarded as one that accurately describes data variation in the direction of ageing through its squared Pearson correlation across age groups. A gene-gene pair with high negative PC1 loading is regarded as one that accurately describes data variation in the opposite direction of ageing by means of its squared Pearson correlation across age groups. Finally, a gene-gene pair with low absolute PC1 loading is regarded as one that is not relevant to explain ageing data variation, at least, by means of its squared Pearson correlation across age groups.

Having the variable loadings, the next step is to explore the respective values in an attempt to highlight the gene-gene pairs that are the most relevant to ageing according to this approach. To that end, in figure 3.9, there is the plot of all the loading values.

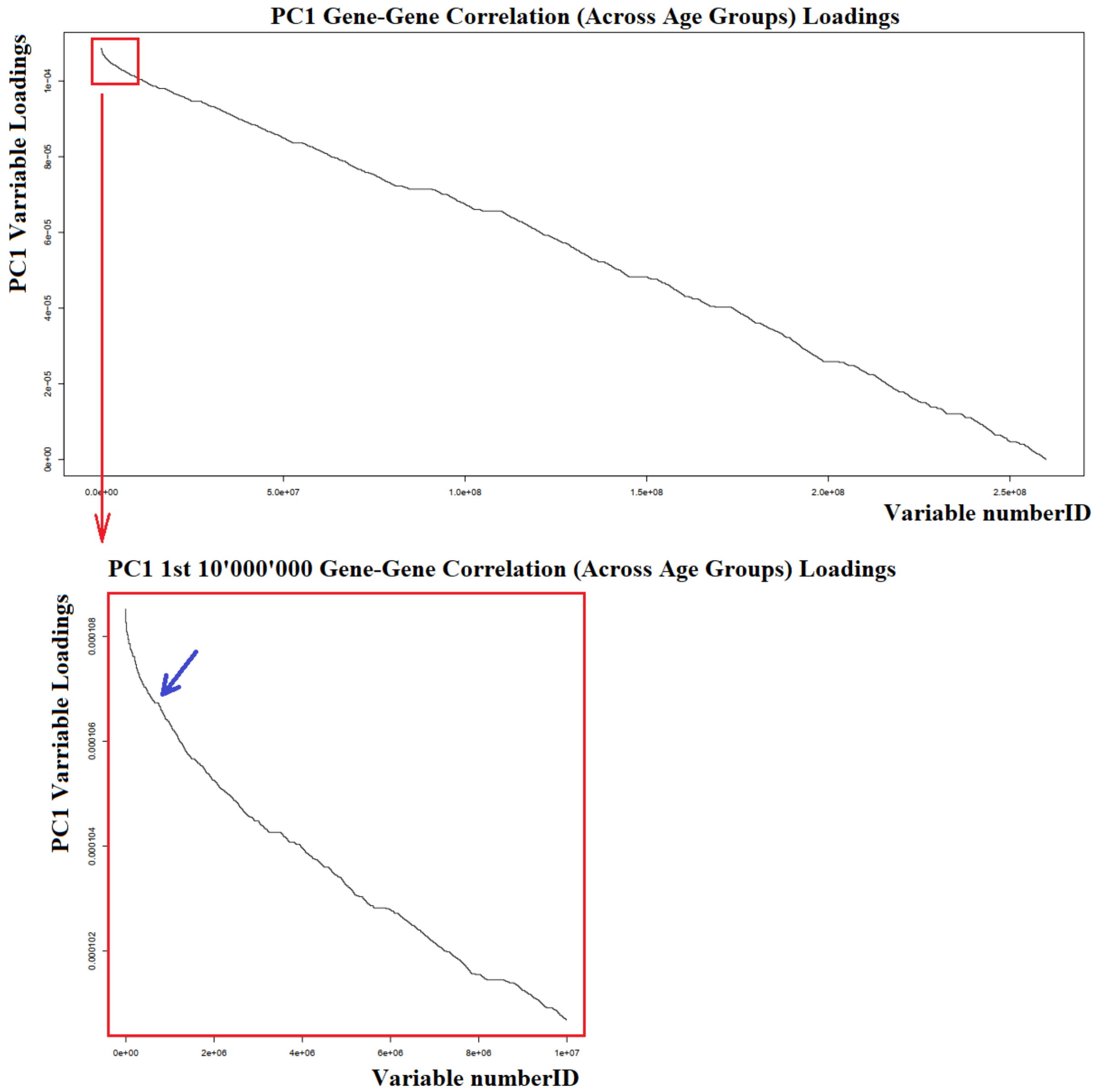


Figure 3.9: PC1 variable loadings of each gene-gene pair ordered by absolute value. PC1 was derived from the PCA (figure 3.8) of the age groups subsets gene-gene pairs squared Pearson correlation. In the right side there is a zoom of the plot of the first 10,000,000 variable loadings.

Unfortunately, the trend of loading values is fairly linear, making it challenging to choose a meaningful threshold. It may seem that in the zone of higher loading values, there might be a decisive point to choose a threshold, but after zooming in the red square of figure 3.9, it is still difficult to make a meaningful decision, and even if it was chosen a threshold such as the one indicated by the blue arrow it would still mean to highlight millions of gene-gene pair values.

3.3.2 Hub Genes of Ageing

To just analyse the top 100 or top 10 gene-gene pair loadings (from figure 3.8) would be an acceptable approach, but that was not the one followed. The developed strategy was to sum all of the loading values in which a particular gene participates, for all of the genes, as is illustrated in figure 3.10.

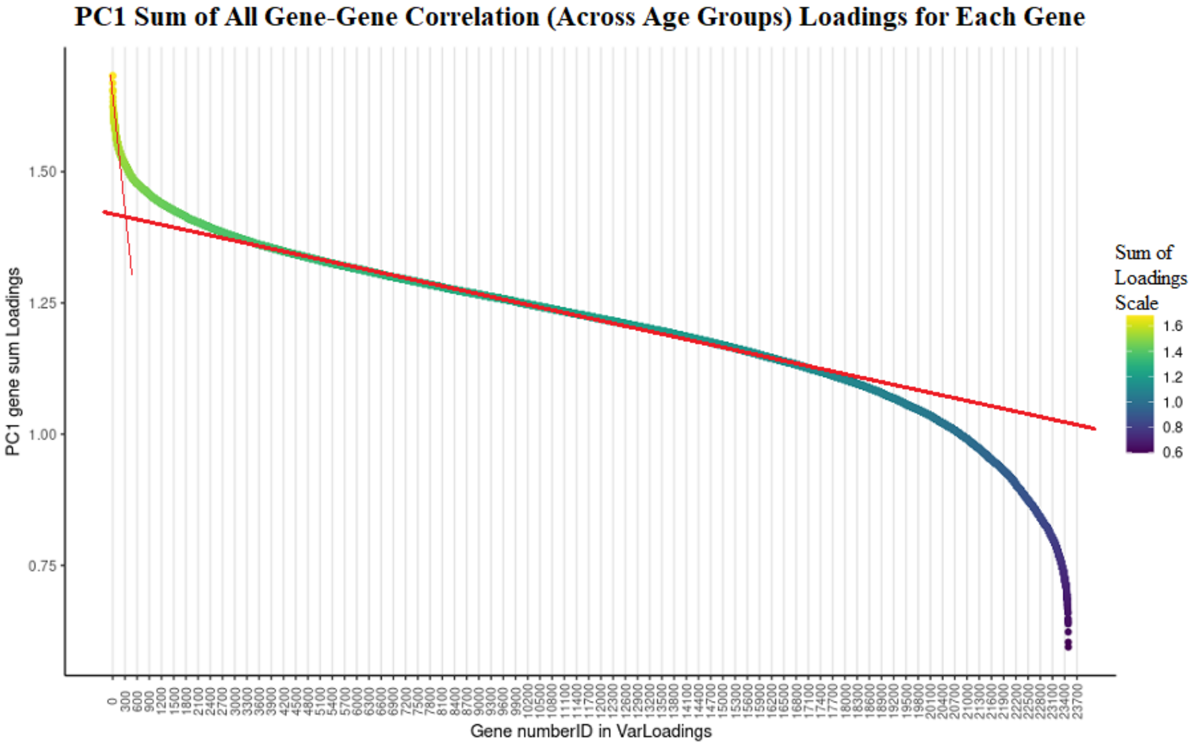


Figure 3.10: Sum of PC1 variable loading values in which each gene participates. This PC1 was derived from the PCA (figure 3.8) of the age groups subsets gene-gene pairs squared Pearson correlation. The 2 red lines represent an attempt to find a threshold sum of loadings value.

This way, instead of having hundreds of millions of variables, there is only about 20'000 genes. Conveniently, this sum of loading values acquired an interesting pattern represented in figure 3.10. This approach can be interpreted as evaluating the hubness of genes regarding their interaction's relevance in describing ageing data variance. Observing figure 3.10, it is much feasible than before to choose a threshold. By means of intersecting the two red lines in the figure, it can be chosen as a threshold the first 300 genes. According to their relationship's relevance in describing ageing data variation, these 300 genes can be interpreted as ”hub genes of ageing”. These 300 genes with the highest sum of loadings might be interesting to explore, and a GO enrichment analysis was applied. The results are available in figure 3.11, and the genes annotation is attached in the appendix A.

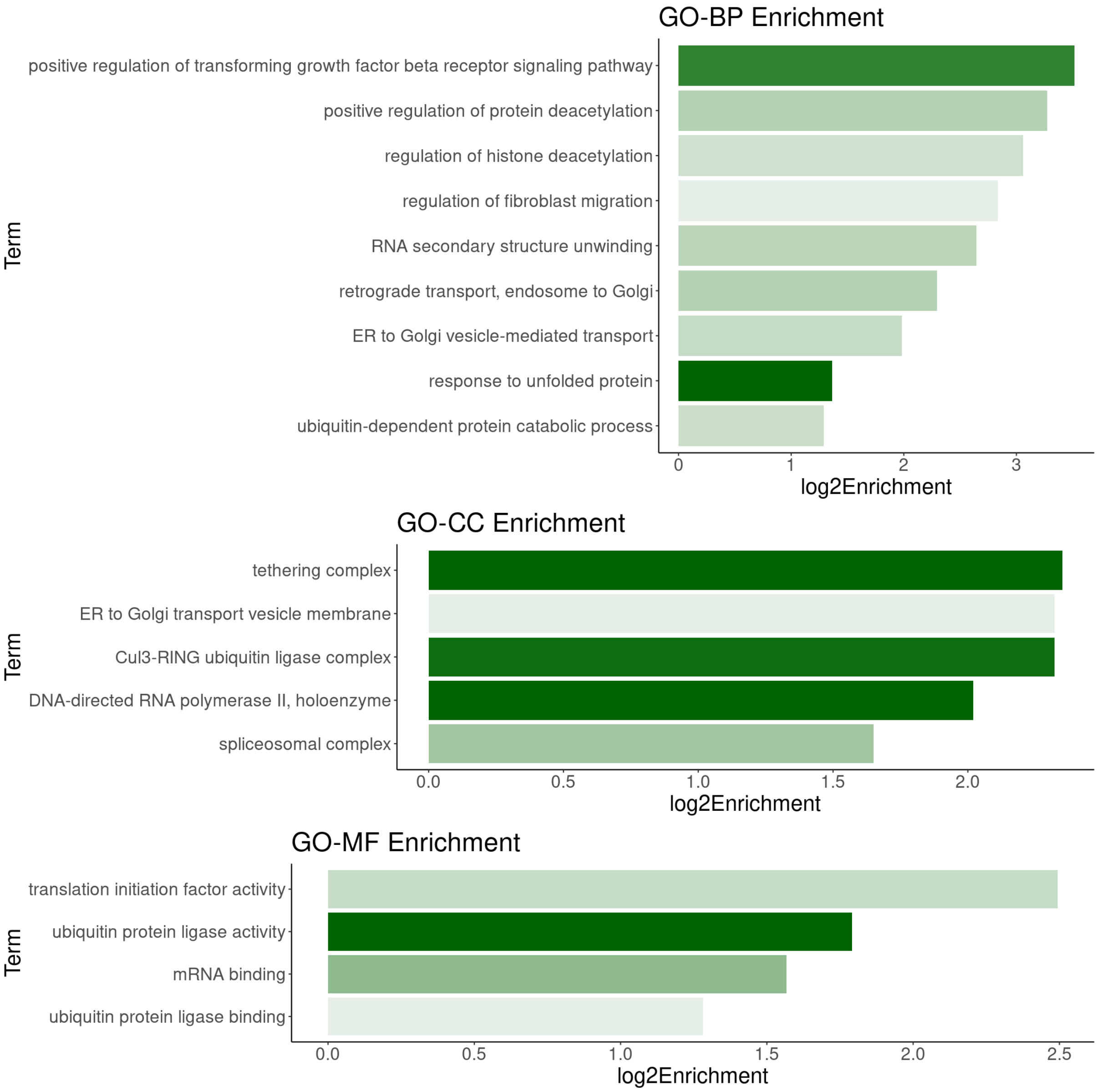


Figure 3.11: GO-BP, GO-CC and GO-MF enrichment analysis results of the top 300 genes with the highest sum of PC1 loadings values derived from the PCA represented in figure 3.8.

These top 300 ”Hub Genes of Ageing” were expected to encompass a very diffuse variety of genes, given that not only ageing might have several causes but also probably affect most molecular and organellar systems of a cell. This could make finding enriched GO terms in these 300 top genes a challenging task. These genes are listed in figure **??** in supplemental material section A.6. Nevertheless, there were captured some plausible GO terms which means that the captured ones should be heavily related to ageing.

Looking at figure 3.11, the most commonly associated terms with ageing are responses to unfolded protein, ubiquitin-dependent protein catabolic processes and Endoplasmatic Reticulum (ER) and Golgi Apparatus related transports. These results are clear indications of the loss of proteostasis (protein homeostasis) hallmark of ageing. This hallmark of ageing [25] means that ageing and some ageing-related diseases are linked to impaired proteostasis. Proteostasis involves mechanisms for the stabilisation of correctly folded proteins and mechanisms for the degradation of proteins. The Autophagy-lysosomal system and the ubiquitin-proteasome system are the two central proteolytic systems implicated in protein quality control, and both decline with ageing [25]. The results strongly suggest that some genes responsible for this process of protein quality control through protein degradation by the ubiquitin-proteasome system have considerable changes in their coordination across ageing. This decrease in coordination might be natural and healthy, just meaning a healthy or intended change in gene-gene relationships and not a decline in regulation across ageing due to some kind of damage accumulation, thus pertinent to ascertain. Actually, it could represent intended healthy adaptation changes in response to ageing. Otherwise, we could age much more aggressively.

The second most consensual hallmark of ageing here observable is the Epigenetic Alterations by Histone Modification [25] here represented by the GO term ”regulation of histone deacetylation”. There are several histone acetyltransferases and deacetylases highly associated with the process of ageing [77], therefore it is highly consistent for a histone deacetylation GO term to reveal himself in this analysis. Histone acetylation results in the neutralisation of the positive charges within histones, weakening the interaction with DNA. The resultant decondensed chromatin structure is one of the required steps for transcription activation. Histone deacetylation has the opposite effect leading to transcription inactivation. Therefore, if histone deacetylation becomes less efficient, there will be less transcriptional regulation in terms of gene silencing.

It is important to note that we should not be overly confident when interpreting the GO term results across the whole present work. It might be possible to more or less link almost every gene or GO term to ageing or to consider them interesting in this regard.

With this in mind, additionally, there are some less apparently related with ageing GO terms, that after looking into, revealed themselves interesting. One of them is ”positive regulation of transforming growth factor-beta receptor signalling pathway”. Transforming growth factor β (TGF-β) is a highly pleiotropic cytokine that plays an essential role in wound healing, angiogenesis, immunoregulation and cancer. While TGF-β might be underproduced in some autoimmune diseases, it is overproduced in many pathological conditions [78]. This means it is essential for TGF-β to be minutiously regulated according to its healthy demand, suggesting that it might be relevant to analyse the genes that contributed to the enrichment of this GO term and to ascertain their interactions in terms of correlation over ageing. Additionally, for future work, it may be pertinent to search for suggestive correlation pairs with the rest of the genome in order to try to identify which pathways are affected when TGF-β pathways are deregulated.

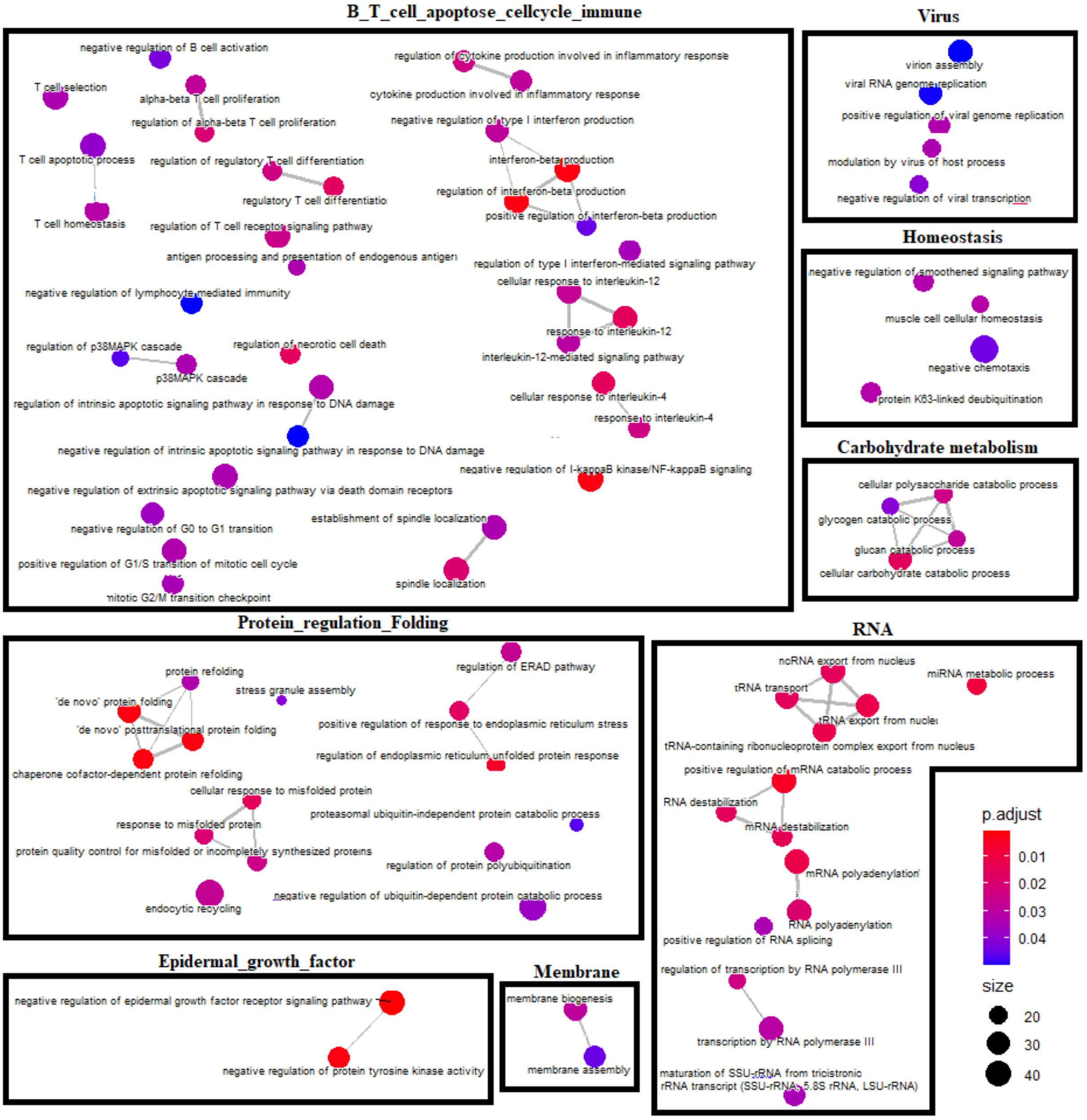
Another interesting GO term is ”regulation of fibroblast migration” because it is known [79] that across ageing, the loss of proliferative and migratory activity of fibroblasts is coupled with the loss of wound closure ability and skin repair, which are consensual signs of ageing. Regarding this migratory activity, there are studies [80] suggesting that ”increased vimentin assembly may underlay the aberrant biophysical properties progressively observed at the cellular level in the course of human ageing and propose vimentin as a potential therapeutic target for ageing-related diseases”. Accordingly, by analysing these results of fibroblast migration relationships, it could unveil relevant information that might as well indicate potential therapeutic targets for the mentioned ageing-related issues. This could also be a goal for further studies.

Lastly, ”RNA secondary structure unwinding”. Many mRNAs have an extensive secondary structure within their coding sequences, and even random sequence RNA has been found to be 50% base-paired [81] posing a potential control in protein synthesis. Hence, an unwinding capacity of those RNA secondary structures play an important part in translational regulation [81]. Thus, a decline in this kind o translational regulation can be another inherent aspect of ageing that should be researched.

The remaining genes that did not enrich any particular GO term should not be excluded as they are no less important and are equally interesting targets for analysis in future work.

3.3.3 GO GSEA Weighted by Genome-Wide Cumulative Gene Co-Expression Related Changes Across Ageing

After applying GO GSEA to the variable PC1 loadings depicted in figure 3.10 it was obtained 168 enriched GO terms in the zones of higher values of loadings. These 168 GO terms were then manually organised into 22 groups of similar or closely related GO terms and 14 isolated GO terms that were specific enough not to be grouped. This grouping is schematised in figure 3.12 where for each group it was given a general name just for reference.


B_T_cell_apoptose_cellcycle_immune
negative regulation of B cell activation
T cell selection
alpha-beta T cell proliferation
regulation of alpha-beta T cell proliferation
T cell apoptotic process
regulation of regulatory T cell differentiation
regulatory T cell differentiatio
T cell homeostasis
regulation of T cell receptor signaling pathway
antigen processing and presentation of endogenous antigen
negative regulation of lymphocyte mediated immunity
regulation of p38MAPK cascade
regulation of necrotic cell death
p38MAPK cascade
regulation of intrinsic apoptotic signaling pathway in response to DNA damage
negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage
negative regulation of extrinsic apoptotic signaling pathway via death domain receptors
negative regulation of G0 to G1 transition
establishment of spindle localization
positive regulation of G1/S transition of mitotic cell cycle
spindle localization
mitotic G2/M transition checkpoint
regulation of cytokine production involved in inflammatory response
cytokine production involved in inflammatory response
negative regulation of type I interferon production
interferon-beta production
regulation of interferon-beta production
positive regulation of interferon-beta production
regulation of type I interferon-mediated signaling pathway
cellular response to interleukin-12
response to interleukin-12
interleukin-12-mediated signaling pathway
cellular response to interleukin-4
response to interleukin-4
negative regulation of I-kappaB kinase/NF-kappaB signaling
Virus
virion assembly
viral RNA genome replication
positive regulation of viral genome replication
modulation by virus of host process
negative regulation of viral transcription
Homeostasis
negative regulation of smoothened signaling pathway
muscle cell cellular homeostasis
negative chemotaxis
protein K63-linked deubiquitination
Carbohydrate metabolism
cellular polysaccharide catabolic process
glycogen catabolic process
glucan catabolic process
cellular carbohydrate catabolic process
Protein_regulation_Folding
protein refolding
stress granule assembly
'de novo' protein folding
'de novo' posttranslational protein folding
chaperone cofactor-dependent protein refolding
regulation of ERAD pathway
positive regulation of response to endoplasmic reticulum stress
regulation of endoplasmic reticulum unfolded protein response
cellular response to misfolded protein
response to misfolded protein
proteasomal ubiquitin-independent protein catabolic process
protein quality control for misfolded or incompletely synthesized proteins
regulation of protein polyubiquitination
endocytic recycling
negative regulation of ubiquitin-dependent protein catabolic process
RNA
ncRNA export from nucleus
tRNA transport
miRNA metabolic process
tRNA export from nuclei
tRNA-containing ribonucleoprotein complex export from nucleus
positive regulation of mRNA catabolic process
RNA destabilization
mRNA destabilization
mRNA polyadenylation
RNA polyadenylation
positive regulation of RNA splicing
regulation of transcription by RNA polymerase III
transcription by RNA polymerase III
maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
Epidermal_growth_factor
negative regulation of epidermal growth factor receptor signaling pathway
negative regulation of protein tyrosine kinase activity
Membrane
membrane biogenesis
membrane assembly
p.adjust
0.01
0.02
0.03
0.04
size
20
30
40

Continues in the next page.

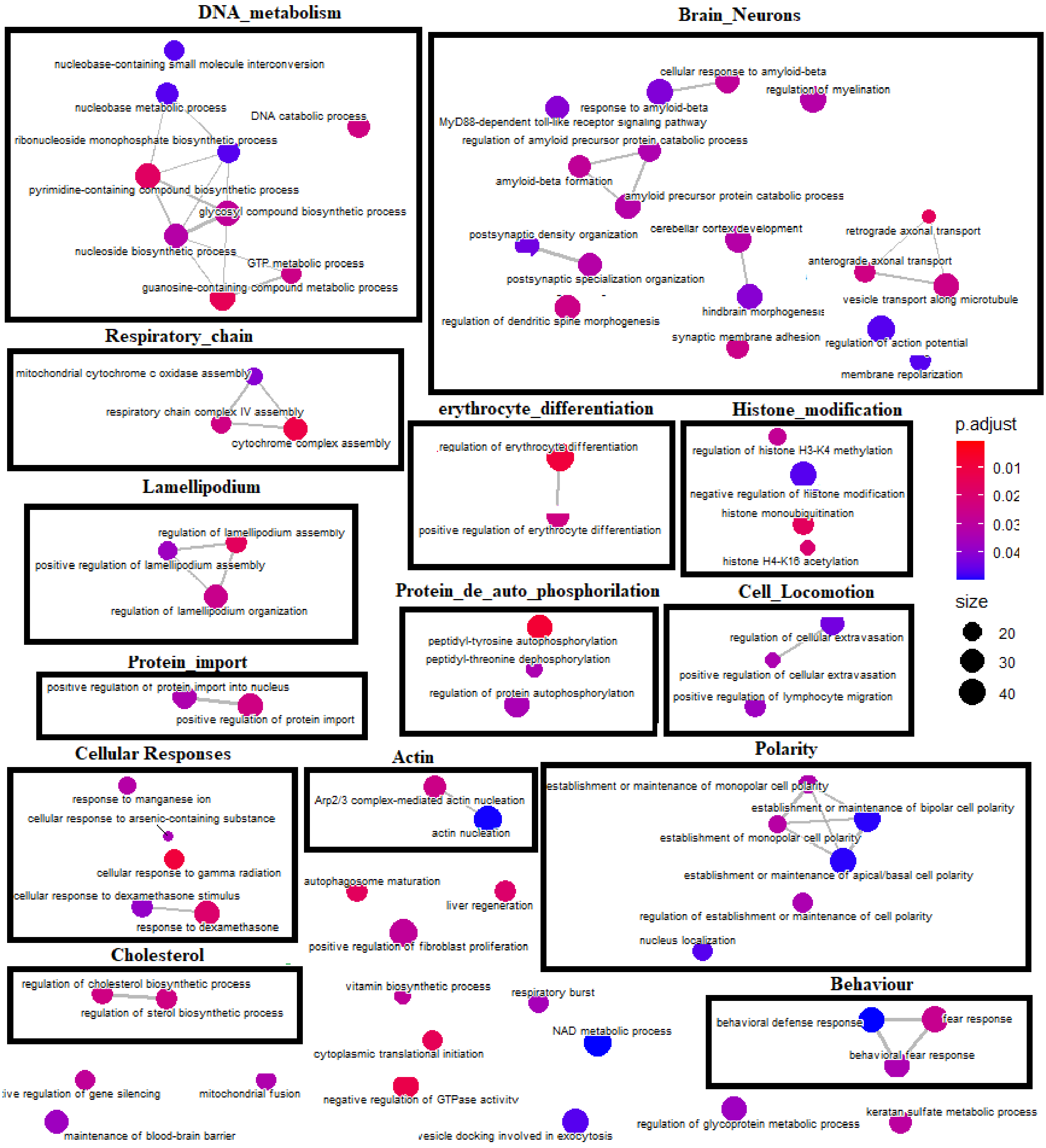

DNA_metabolism
nucleobase-containing small molecule interconversion
nucleobase metabolic process
DNA catabolic process
ribonucleoside monophosphate biosynthetic process
pyrimidine-containing compound biosynthetic process
glycosyl compound biosynthetic process
nucleoside biosynthetic process
GTP metabolic process
guanosine-containing compound metabolic process
Brain_Neurons
cellular response to amyloid-beta
regulation of myelination
response to amyloid-beta
MyD88-dependent toll-like receptor signaling pathway
regulation of amyloid precursor protein catabolic process
amyloid-beta formation
amyloid precursor protein catabolic process
postsynaptic density organization
cerebellar cortex development
retrograde axonal transport
anterograde axonal transport
postsynaptic specialization organization
vesicle transport along microtubule
hindbrain morphogenesis
regulation of dendritic spine morphogenesis
synaptic membrane adhesion
regulation of action potential
membrane repolarization
Respiratory_chain
mitochondrial cytochrome c oxidase assembly
respiratory chain complex IV assembly
cytochrome complex assembly
erythrocyte_differentiation
regulation of erythrocyte differentiation
positive regulation of erythrocyte differentiation
Histone_modification
regulation of histone H3-K4 methylation
negative regulation of histone modification
histone monoubiquitination
histone H4-K16 acetylation
p.adjust
0.01
0.02
0.03
0.04
size
20
30
40
Lamellipodium
regulation of lamellipodium assembly
positive regulation of lamellipodium assembly
regulation of lamellipodium organization
Protein_de_auto_phosphorilation
peptidyl-tyrosine autophosphorylation
peptidyl-threonine dephosphorylation
regulation of protein autophosphorylation
Cell_Locomotion
regulation of cellular extravasation
positive regulation of cellular extravasation
positive regulation of lymphocyte migration
Protein_import
positive regulation of protein import into nucleus
positive regulation of protein import
Cellular Responses
response to manganese ion
cellular response to arsenic-containing substance
cellular response to gamma radiation
cellular response to dexamethasone stimulus
response to dexamethasone
Actin
Arp2/3 complex-mediated actin nucleation
actin nucleation
Polarity
establishment or maintenance of monopolar cell polarity
establishment or maintenance of bipolar cell polarity
establishment of monopolar cell polarity
establishment or maintenance of apical/basal cell polarity
regulation of establishment or maintenance of cell polarity
nucleus localization
autophagosome maturation
liver regeneration
positive regulation of fibroblast proliferation
Cholesterol
regulation of cholesterol biosynthetic process
regulation of sterol biosynthetic process
vitamin biosynthetic process
respiratory burst
NAD metabolic process
cytoplasmic translational initiation
Behaviour
behavioral defense response
fear response
behavioral fear response
ive regulation of gene silencing
mitochondrial fusion
negative regulation of GTPase activity
maintenance of blood-brain barrier
vesicle docking involved in exocytosis
regulation of glycoprotein metabolic process
keratan sulfate metabolic process

Figure 3.12: Enrichment map which organizes enriched terms into a network with edges connecting overlapping gene sets. In this way, mutually overlapping gene sets tend to cluster together, making it easy to identify functional modules. Here we have 168 GO terms returned as top enriched by the sum of loadings values from figure 3.10 as a result of GO GSEA. GO GSEA enriched terms are manually organised into 22 groups of similar or closely related GO terms and 14 isolated GO terms that were specific enough not to be grouped.

The GO GSEA is applied to this vector of the sum of variable loadings for each gene. And it should be reminded that a gene with a higher value means it is one whose interactions with other genes greatly contribute to the data variance in the PC1 direction. Additionally, applying GO GSEA is a well-established approach to understand which gene sets (GO terms) are over-represented in the genes with higher values of some interesting variable (in this case, the sum of variable loadings). For each significantly enriched GO term it is returned the set of core enriching genes that significantly ”top” enriched the respective GO term taking into account the sum of variable loadings values. If a GO term has 200 genes present in the sum of loadings list of more than 23,000 genes and if it is top enriched in this list by 20 core enriching genes, it means that 20 genes from this GO term have loading values high enough to enrich this GO term in the top of the list.

In this GO GSEA step, it is obtained mostly terms that represent or are related to already known areas affected by the consensual hallmarks of ageing [25].

Regarding the ”B_T_cell_apoptose_cellcycle_immune” GO terms group, it was already discussed its relevance for ageing in section 3.2.2 due to the immune system's complement system, neutrophils, natural killer cells and the major histocompatibility complex. Nevertheless, here it can be seen an additional strong representation of induced cell death accompanied by cell cycle regulation as well as a representation of T cell regulation at different levels. One of the major and well-known hallmarks of ageing is mitochondrial dysfunction [25], which in T cells leads to the acquisition of a proinflammatory phenotype through a combination of several molecular mechanisms, including the accumulation of inflammatory metabolites, epigenetic alterations, post-transcriptional protein modifications, and the release of mtDNA to the cytoplasm that activates specific pathways, culminating in the activation of the inflammasome and the transactivation of genes coding for proinflammatory cytokines [27]. Interestingly, here was obtained a well-represented group of GO terms related to the respiratory chain and a ”NAD metabolic process” GO term, which could have been grouped with the respiratory chain group. Indeed, meddling with respiratory chain complexes accelerates immunosenescence in human T cells [27] and destabilisation of the said complexes is one of the central mechanisms causing defective mitochondrial bioenergetics across ageing [25]. Moreover, other mechanisms causing defective mitochondrial bioenergetics across ageing are alterations in mitochondrial dynamics resulting from an imbalance of fission and fusion events and defective quality control by mitophagy, an organelle-specific form of macroautophagy that targets defective mitochondria for proteolytic degradation and also changes in the lipid composition of mitochondrial membranes [25].

In fact, all these mentioned mechanisms might be here (figure 3.12) represented. There is represented a ”mitochondrial fusion” GO term. Fusion is a crucial element in maintaining mitochondrial physiology, enabling content mixing within a mitochondrial population, preventing permanent loss of essential components. Cells with a decline in mitochondrial fusion, as a consequence, acquire a subpopulation of mitochondria that lack mtDNA nucleoids leading to respiration-deficient mitochondria [82]. Mitochondrial fusion appears to play a protective role in neurodegeneration, and studies [83] discuss its current evidence of their role in the ageing of multicellular organisms and how these connect to cell cycle regulation, quality control, and transmission of energy status. There is also represented an ”autophagosome maturation” GO term which might be correlated with the mentioned mitochondrial dysfunction due to defective quality control by mitophagy in which autophagosomes enclose whole mitochondria [84]. Actually, mitophagy shares the core molecular machinery with general macroautophagy [84] consolidating the link between the extensively described [24] decline in the autophagic activity across ageing and the ageing hallmark of mitochondrial dysfunction. As mentioned, a decline in the autophagic activity across ageing is described [24] and its contribution to the accumulation of damaged macromolecules and organelles during ageing worsens ageing-associated diseases, such as neurodegeneration or cancer, among others. Also, about mitochondrial dysfunction across ageing, it was mentioned changes in the mitochondrial membrane's composition, and within the GO GSEA obtained terms, there is also ”membrane biogenesis” and ”membrane assembly” terms. Actually, it is also a topic within ageing and cellular senescence the dynamics of mitochondria-associated membranes [85]. Furthermore, it is explored the possibility that modifications in the physicochemical properties of the plasma membrane resulting from changes in its lipid composition and the distribution and function of lipid raft might be a unifying cause for the decreased efficiency of immune responses in older people, consequence of alterations in T lymphocyte functions, caused by modifications in the early events of signal transduction [86]. In truth, this membrane group of GO terms has lipid raft related genes as core enrichment genes such as the RFTN1 gene and, again, recent studies have shown a close relationship between lipid rafts and the age-associated decline and dysregulation of cellular signalling pathways, such as T-cell receptor signalling and cellular senescence-related signalling [87] as well as the role of cholesterol in lipid raft functions across ageing in T lymphocytes [86] being that cholesterol is one of the main constituents of lipid raft.

As a matter of fact, in the obtained results, there is also a group of GO terms related to cholesterol and its biosynthetic process regulation. It is reported that ageing dysregulates cholesterol metabolism via a number of mechanisms and that the ratio between the so-called ”good cholesterol” (HDL-C) and ”bad cholesterol” (LDL-C) decreases across ageing, significantly impacting the risk of cardiovascular disease in older people [88]. At the same time, there is considerable literature [89] describing the active role of cholesterol during liver regeneration either by its function as a structural lipid and regulation of changes in membrane fluidity as well as its signalling role, functioning as secondary messenger along with other lipids and as a precursor for new messengers that diffuse from the plasma membrane into the nucleus to affect the transcription of genes that induce changes in the homeostasis.

Coming back to the ”B_T_cell_apoptose_cellcycle_immune” group of GO terms, it is of particular importance for there to be the ”negative regulation of I-κB kinase/NF-κB signaling” GO term because a study [37] points the affected activity of transcription factor NF-κB as a possible causal mechanism for loss of gene co-expression across ageing. They state that old age may affect the activity of transcription factor NF-κB, in a way that its direct targets may decrease their correlation with age. This is expected because NF-κB is involved with inflammation, which, as already mentioned, increases with age in all tissues [90], deeply implicating the NF-κB transcription factor with ageing.

In figure 3.12 it is also seen an ”Epidermal_growth_factor” roup of GO terms. Epidermal Growth Factor (EGF) plays important roles in normal wound healing involved in inflammation, wound cell migration and mitosis, neovascularisation, and regeneration of the extracellular matrix [91]. By stimulating cell growth and differentiation, these GO terms end up being related with the also represented ”positive regulation of fibroblast proliferation” because indeed EGF promotes fibroblast proliferation [92]. The implied GO terms end up having similar roles in ageing as the previously described effects regarding ”regulation of fibroblast migration” in chapter 3.3.2. Looking back on to ”regulation of fibroblast migration”, here it is also represented two groups of GO terms clearly related to cellular migration (”Cell_Locomotion” and ”Lamellipodium”). The lamellipodium is essential for cell motility, the organisation of membrane domains, phagocytosis and the development of substrate adhesions [93]. In fact, the lamellipodium is born of actin nucleation in the plasma membrane and is the primary area of actin incorporation, or microfilament formation of a cell [93], which is consistent with the also represented group of GO terms ”Actin”. Actin also makes part of actomyosin, which its contractility in fibroblasts decreases substantially [94] during ageing. This loss of fibroblast contractility leads to reduced connective tissue stiffness. As it was previously discussed, it is described [94] that throughout the ageing process, fibroblasts lose contractility, leading to reduced connective-tissue stiffness. The stiffness of the extracellular environment has a role in orienting cell division, maintaining tissue boundaries, directing cell migration, driving differentiation and also maintaining normal tissue homeostasis [95]. Concerning actin’s role, it is also deeply related to the represented group of GO terms ”Polarity” [96]. Curiously, loss of epigenetic polarity is a hallmark of hematopoietic stem cell ageing; thus, their functional deterioration and consequently loss of tissue homeostasis [97]. *Polarity* can be defined as the uneven distribution of molecules and organelles within a cell and is necessary for a multitude of processes like cell motility; asymmetric inheritance; cell-type-specific functions as oriented vesicle secretion in axons and dendrids of neurons; tissue orientation as the apicobasal polarity of the epithelium and tissue specialisation [97]. Many of the mentioned processes are compromised during ageing and cellular senescence. For example, permeability epithelium barriers are leakier during ageing; elderly people have impaired vascular function and increased frequency of cancer, and asymmetrical inheritance is compromised in senescent cells, including stem cells [98].

Polarity is especially relevant for Hematopoietic Stem Cell (HSC) ageing because it is required for asymmetric cell division [97]. *Asymmetric cell division* is a mechanism that balances HSC self-renewal and differentiation, where the unequal inheritance of cell fate determinants into daughter cells is determined by mitosis linked mechanism [99]. In fact, there is also a representation of Erythrocyte differentiation GO terms, which is one of the HSC differentiation products. It is known that the process of ageing in HSCs leads to a reduction in blood cell production [100]. A reduction in blood cell production implies reduced oxygen supply to tissues, partly responsible for age-related physical and cognitive decline. With the decrease in the number of red blood cells or haemoglobin in the blood, anaemia is actually common in the elderly, and its prevalence increases with age [101].

The represented ”respiratory burst” makes sense in the context of the neutrophil impact on ageing already discussed in chapter 3.2.2, because the respiratory burst is named as the neutrophils prodigious respiratory burst-induced production of superoxide, important in pathogen degradation in lysophagossomes or tumour cells apoptosis [65].

The represented ”Histone_modification” group of GO terms relevance for ageing was already discussed in chapter 3.3.2 regarding the top 300 possible hub genes of ageing GO enrichment analysis. Actually, there is a study [37] which points to the deterioration of chromatin structure as a possible causal mechanism for loss of gene co-expression in old age. As discussed before, chromatin histone modifications manage gene expression permitting and prevention. The referred study states that chromatin domains would become less well-defined if these histone modifications were to deteriorate across ageing. Genes entirely repressed and strongly activated at a young age would show either high basal expression or low activated expression in old age. This phenomenon would result in lower co-expression levels with other genes in the network. The ”negative regulation of gene silencing” GO term also represented is expected to have a similar influence at the level of gene activation and inactivation.

Similarly to the ”Histone_modification” group of GO terms, the ”RNA” group of GO terms might be representing factors that would affect the decrease in co-expression with ageing in an equivalent way as chromatin deterioration might by influencing gene expression.

Within ”RNA” group of GO terms it can be seen GO terms as ”positive regulation of mRNA catabolic process”, ”mRNA distabilization”, ”tRNA transport”, ”mRNA polyadenylation”, ”positive regulation of RNA splicing” or ”regulation of transcription by RNA polymerase III”. Deterioration of any of these processes can gene expression and consequently change co-expression across ageing. Actually the also represented ”cytoplasmatic translational initiation” GO term could also have been grouped here, by being expected to have a similar influence.

In truth, it was already shown [36] that mRNAs encoding protein synthesis machinery components have its translation decreased with age in both liver and kidney, in mice. And, of course, this correlates well with the previously observed decline in overall protein synthesis with age [102]

The represented ”Brain_Neurons” group of GO terms Will be explored further in the discussion.

One of the most crucial aspects of ageing is ”Loss of Proteostasis”. Proteostasis includes the correct folding of proteins and their stabilisation mainly via the heat-shock family of proteins, as well as protein degradation, when they become faulty or in excess, by the proteasome or lysosomes. These mechanisms are heavily represented in the GO terms obtained in the group of ”Protein Regulation” present in figure 3.12 as, for example, ”’*de Novo* protein folding”, ”chaperone cofactor-dependent protein refolding”, ”response to misfolded protein”, ”proteasomal ubiquitin-independent protein catabolic process” and the more encompassing but still important ”autophagosome maturation” GO term.

Many studies have demonstrated that ageing is associated with perturbed proteostasis, that genetic manipulations which improve proteostasis delay ageing in mammals, and that experimental perturbation of proteostasis or chronic expression of unfolded, misfolded or aggregated proteins contributes to the development of age-associated pathologies such as Alzheimer’s disease, Parkinson’s disease and cataracts [25]. Therefore, this GO GSEA step of the analysis reinforces and is in concordance with the described by displaying the mentioned GO terms as Hub terms in the explanation of the data variance in the increasing age group direction, which in fact is the direction of greatest variance in the data as discussed before.

All these mentioned features of ageing are highly interconnected, meaning that they progress together, influence each other, and give rise to common features of the aged phenotype. Furthermore, in figure 3.12 there are still some GO terms or groups of, that were left from being discussed as ”Homeostasis”, ”Cellular_Responses”, ”DNA_metabolism”, ”Carbohydrate metabolism”, ”Protein_de_auto_phosphorilation”, ”Protein_import”, between others that should also be interesting to delve into but should end up being interconnected with some of the already discussed terms.

3.3.4 Gene Graph Network Weighted by Gene Co-Expression-Related-Changes Across Ageing

Not only the analysis and discussion done until now are in concordance with the literature, but it further allows to check, within or between GO terms, which are the gene-gene associations that best describe ageing, whether it is because their coordination decreases or increases across ageing. For example, figure 3.13 illustrates those relationships across ageing between the core enriching genes of the GO GSEA enriched terms related to protein folding that have the word ”folding” in their names.

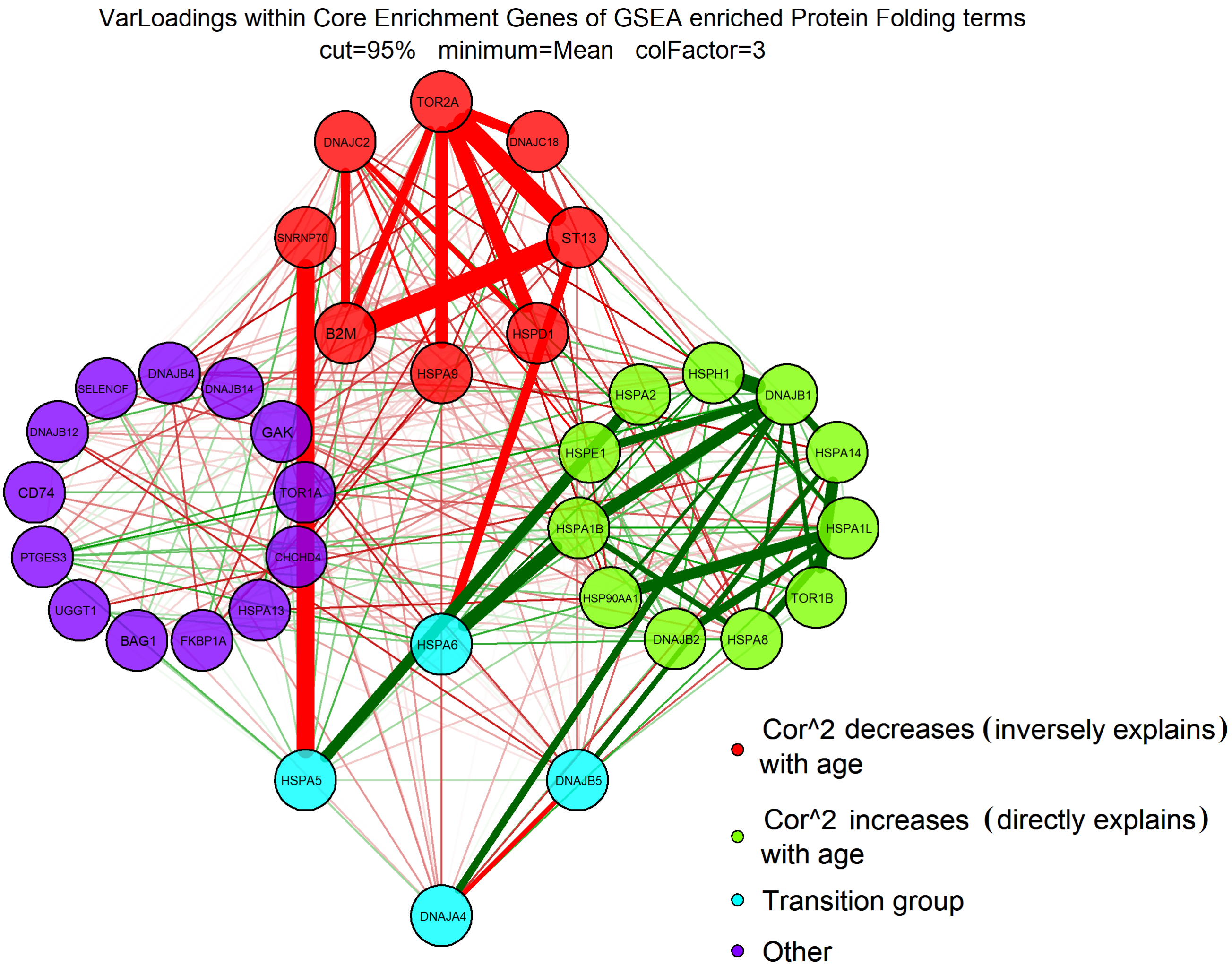


Figure 3.13: Node network graph between the core enriching genes (nodes) of the obtained (figure 3.12) GO GSEA enriched terms related to protein folding. The underlying GO terms are ”protein refolding”, ”’*de novo*’ protein folding”, ”’*de novo*’ posttranslational protein folding” and ”chaperone cofactor-dependent protein refolding”. The thickness and colour intensity of the edges (lines) reflect the respective gene-gene loadings of the first principal component (figure 3.8.B) that best describes the data variance which actually is the greatest in the direction of ageing. Across ageing means across 20-29, 30-39, 40-49, 50-59, 60-69 and 70-79 age group sample subsets. Green edges indicate gene-gene relationships which their coordination across age groups increases, thus explaining the data variance in the direction of ageing, and red edges indicate gene-gene relationships which their coordination across age groups decreases, thus inversely explaining the data variance in the direction of ageing. Regarding the plot parameters, the ”minimum” is the mean of the loadings and is the value at which the smaller edges are not shown. The ”cut” is the 95% quantile of the loadings and is the value at which higher value edges become wider and more colour intense and lower value edges become thinner and less colour intense. ”colfactor” of 3 is the exponential factor at which the colour intensity changes according to the loading’s values. The group of red nodes tries to bring together gene-gene pairs whose negative loadings are the highest, the group of green nodes tries to bring together gene-gene pairs whose positive loadings are the highest, the group of blue nodes tries to bring together genes that have both high positive and high negative loadings interactions, and the group of purple nodes groups the remaining genes that didn’t fit in any of the other groups.

In the node network graph present in figure 3.13 genes are grouped by the ones who mainly have relationships of decreasing correlation with age (in red) and increasing correlation with age (in green). In the Blue group, there are the transition genes that both have substantial decreasing and increasing correlations across age. These representations allow for interesting analysis; for example, it is possible to observe that gene HSPA6, across ageing, decreases its coordination with gene ST13 to increase its coordination with gene HSPA1B. The protein encoded by the ST13 gene is an adaptor protein that mediates the association of heat shock proteins with several targets [103]. The expression of this gene is reported to be downregulated in colorectal carcinoma tissue and is suggested as a candidate tumour suppressor gene [104]. Additionally, the major factor that increases a person's risk for colorectal cancer is increasing age [105]. This simple analysis enables hypotheses such as hypothesising that some types of cancer incidence in old age might be due to ST13 decreasing co-expression with HSPA6. But then, we can also observe that ST13 decreases its coordination with the B2M gene, which was already discussed in section 3.2.2 in cluster 38 (figure 3.7) that was a module captured in a cross-tissue and cross-age approach that significantly decreased its within-cluster squared correlation across age. MHC class-I molecules comprise B2M and act as tumour suppressors [106], and interestingly B2M loss is involved in the loss of the MHC class-I antigens for the CTLs recognition of colorectal cancer cells [107]. All of this information begs the hypothesis that cancers as colorectal cancer might be increasingly present in older ages because ST13 is decreasing its correlation with B2M and HSPA6, not directing HSPA6 efficiently enough to the B2M for it to operate in a correctly folded state. Actually, it is described that for B2M to assemble with the MHC class-I molecules, it may need the help of chaperone proteins as HSPA5 [108] and here it is being proposed that it actually may also need HSPA6 by the guidance of ST13.

Additionally, HSPA5 appears to lose coordination with SNRNP70 to gain coordination with HSPA2 across ageing. HSPA5 is involved in the correct folding of proteins and degradation of misfolded proteins [109]. SNRNP70 is a small nuclear ribonucleoprotein, and a core component of the spliceosome [110]. SNRNP70 has been reported to form detergent-insoluble aggregates in both sporadic and familial human cases of Alzheimer's disease co-localising with Tau in neurofibrillary tangles in Alzheimer's disease [111]. A study [111] states that the mechanisms underlying SNRNP70 aggregation are unknown and suggests that it might be the aggregated SNRNP70 itself or other biopolymers (e.g. proteins or nucleic acids) that interact with and sequester natively folded soluble SNRNP70 into insoluble aggregates. According to the results in figure 3.13 it is hypothesizable that SNRNP70 aggregates might be due to its decreasing co-expression with HSPA5. Maybe resulting from a decrease of SNRNP70 correct folding maintenance by chaperone HSPA5 or, more likely, a decrease in the correct processing of misfolded SNRNP70 by HSPA5 chaperone. This is relevant because, as explained before, reduced protein homeostasis leads to increased protein instability which is a common molecular feature of ageing, even though it still remains unclear whether this is a cause or consequence of the ageing process [112]. Furthermore, protein aggregation is a specific form of protein instability described during normal ageing and demonstrated to accelerate the functional decline of different tissues during normal ageing [112]. Interestingly, HSPA5 levels are reduced in the brains of Alzheimer's disease patients [113], supporting the proposed hypothesis that HSPA5 discoordination with SNRNP70 might be a reason for those aggregates in Alzheimer's disease and maybe a reason for an eventual spliceosomal malfunctioning in general, leading to transcriptional alteration which is one of the hallmarks of ageing.

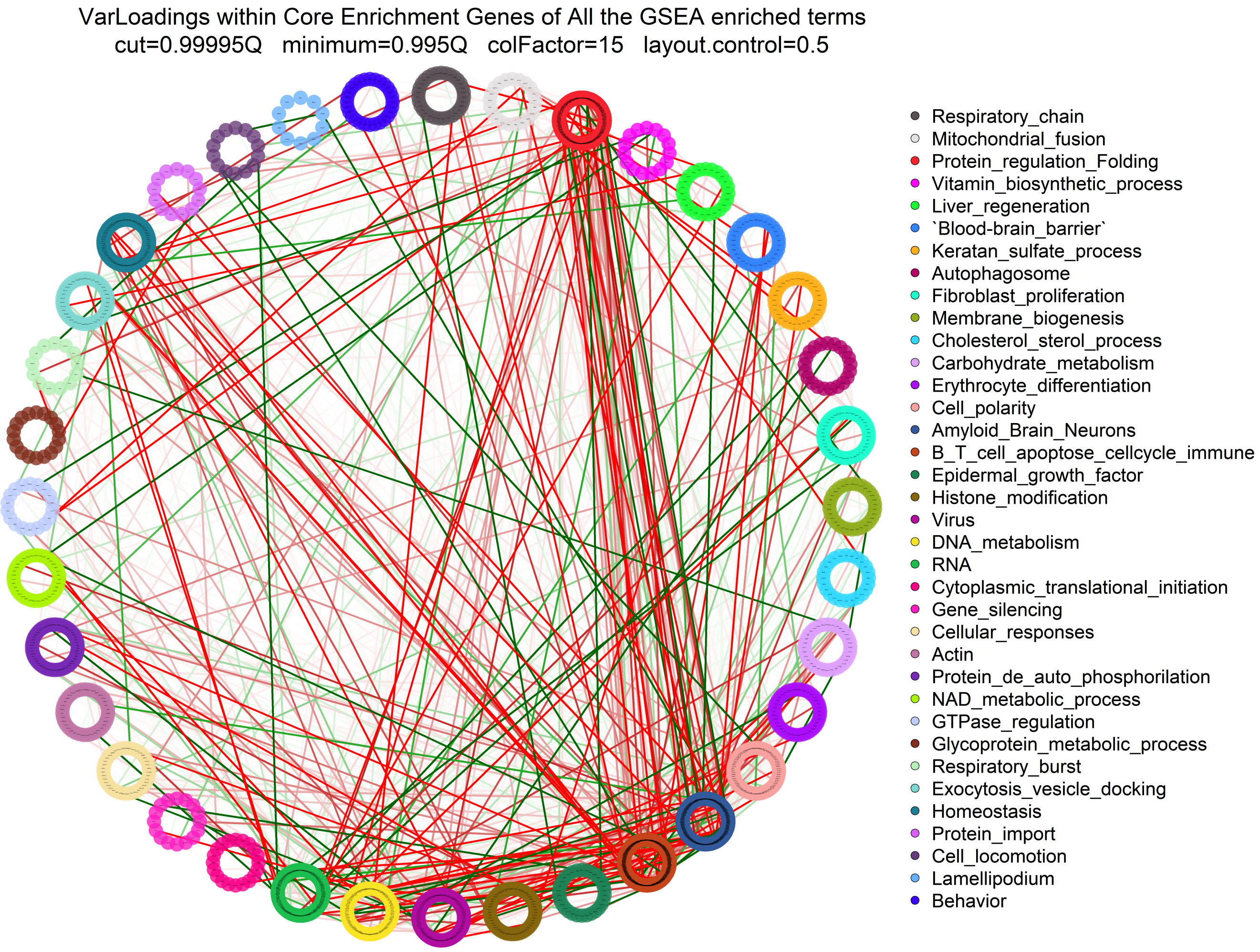

VarLoadings within Core Enrichment Genes of All the GSEA enriched terms
cut=0.99995Q minimum=0.995Q colFactor=15 layout.control=0.5
● Respiratory_chain
● Mitochondrial_fusion
● Protein_regulation_Folding
● Vitamin_biosynthetic_process
● Liver_regeneration
● `Blood-brain_barrier`
● Keratan_sulfate_process
● Autophagosome
● Fibroblast_proliferation
● Membrane_biogenesis
● Cholesterol_sterol_process
● Carbohydrate_metabolism
● Erythrocyte_differentiation
● Cell_polarity
● Amyloid_Brain_Neurons
● B_T_cell_apoptose_cellcycle_immune
● Epidermal_growth_factor
● Histone_modification
● Virus
● DNA_metabolism
● RNA
● Cytoplasmic_translational_initiation
● Gene_silencing
● Cellular_responses
● Actin
● Protein_de_auto_phosphorilation
● NAD_metabolic_process
● GTPase_regulation
● Glycoprotein_metabolic_process
● Respiratory_burst
● Exocytosis_vesicle_docking
● Homeostasis
● Protein_import
● Cell_locomotion
● Lamellipodium
● Behavior

Figure 3.14: Node network graph between the core enriching genes (nodes) of the obtained (figure 3.12) GO GSEA enriched terms groups. The colour intensity of the edges (lines) reflect the respective gene-gene loadings of the first principal component (figure 3.8.B) that best describes the data variance which actually is the greatest in the direction of ageing (age groups: 20-29, 30-39, 40-49, 50-59, 60-69 and 70-79 years old). Green edges indicate gene-gene relationships which their coordination across age groups increases, thus explaining the data variance in the direction of ageing, and red edges indicate gene-gene relationships which their coordination across age groups decreases, thus inversely explaining the data variance in the direction of ageing. Regarding the plot parameters, the "minimum" is the mean of the loadings and is the value at which the smaller edges are not shown. The "cut" is the 99995% quantile of the loadings and is the value at which higher value edges become more colour intense and lower value edges become less colour intense. "colfactor" of 15 is the exponential factor at which the colour intensity changes according to the loading's values.

With this GO GSEA approach, it is also possible to assess inter GO term relationships as in figure 3.14. Given the filters that were necessary to apply to conveniently illustrate only the most intense relationships, it is clear that there is a vast amount of interplay between different GO terms regarding their linear changes in gene coordination across ageing. It is observable that there are some prominent interactions between GO Terms. Mainly, the group of GO terms here named as ”Protein_regulation_Folding” seems to have a relatively high amount of gene-gene relationships of decreasing coordination across ageing with genes from the groups of GO terms called ”B_T_cell_apoptose_cellcycle_immune” and ”Amyloid_Brain_Neurons”. This is rather interesting because it allows hypothesising that a lot of age-related problems regarding the immune system and cell-cycle regulation might be due to a decrease in co-expression with protein folding and regulation modules. And the same for the Brain problems as amyloid-related ones. Coincidentally the pairs protein regulation/immune system and protein regulation/brain diseases were the ones explored upon analysis of within protein folding relationships regarding figure 3.13.

It should be noted that many represented GO terms share genes, and for this figure 3.14 representation to be possible, a gene assignment to each GO term group needed to be done. From the list of all genes, genes were assigned to each GO term group one by one and removed from the list of genes in a specific order described in the methodologies section 2.13.

The results of this study might also be consistent with the long-standing hypothesis of ageing as dysdifferentiation, where cells start losing their proper state of differentiation. This hypothesis actually goes back to the 1970s when Richard Cutler proposed the idea based on observations of active genes in aged tissues that should typically be silent in that tissue [114]

Although certainly vast and seemingly complex, genome-wide correlations and a focused GO term networks such as this one provide an organisational framework that helps the process of hypothesis generation and testing that look beyond where our current knowledge ends to develop a more encompassing view of the problems posed by ageing and ageing disorders and their potential solutions.

4

Conclusions

Contents

4.1 Work Limitations

Conclusions about the present work should always have in mind the several underlying limitations. First of all, the work uses bulk (tissue) data which could be creating co-expression connections that potentially obscure co-regulatory modules because of the heterogeneity of cell types in RNA-seq data [10]. Additionally, in a similar way, co-expression itself has its own limitations when inferring about co-regulation because the same gene might be involved in different cellular processes and because some functions can be 'served' by alternative genes in different conditions, cell types or tissues. In a simplistic way, one of the disadvantages of using correlation between gene pairs is that it doesn't take into account the relations that those genes might have with other genes.

Regarding the age analysis, age might affect different tissues differently, and the current analyses presumably won't capture these tissue-specific interactions relevant to ageing.

Regarding GO term enrichment analyses, we should be careful with the over-interpretation because these annotations have pitfalls. Sometimes one might be finding terms as significantly enriched because some of its member genes might also be involved in another process that would actually make more biological sense [115]. Additionally, ageing is a process that affects multiple systems, and because within an organism, many systems are interconnected, it is possible to be over-interpreting processes as possible sources of ageing when they are not.

In the present work, a peculiar way to apply PCA was used. And one of the requirements was to use few data points, which were the six that represented the six different age groups. This should be kept in mind, but also that a robust amount (thousands of millions) of variables represented each data point.

Upon interpretation of the ageing analyses, it should be taken into account that the PCA was scaled. Relationships whose co-expression is extremely low but still decreases or increases linearly with age are though to be here highlighted with a big variable loading value as big as the linearity of the co-expression change. In future work, it might be advisable to remove from analysis gene pairs whose mean co-expression across ageing is near the random values observed for the equally sized random modules from the module analysis.

Finally, some of the decreases in gene-gene coordination with age might be natural and healthy, just meaning a healthy or intended change in gene-gene relationships and not an actual decline in regulation across ageing due to some kind of damage accumulation. Indeed, it could represent intended healthy adaptation changes in response to ageing. Otherwise, we could age much more aggressively.

4.2 Gene Modules Across and Within Tissues

It was obtained 65 highly correlated gene modules across tissues. It was observed that some preserve its high correlation within several specific tissues in a stable way. Others displayed lower but still considerable correlation values within several specific tissues in a stable manner. Some clusters would only preserve considerable or high correlation values within a small set of specific tissues. Moreover, some clusters exhibited very low values within any of the utilised tissues. This analysis appears to be in agreement with the expectation that gene co-expression networks are not entirely rearranged between tissues.

Gene co-expression analyses are widely used to infer gene modules associated with diseases. Nevertheless, a systematic view and comparison of gene co-expression networks and modules across tissues are more or less ignored. The present work provides additional support that many tissue-specific data and studies can be unified to some extent, much more than is currently done. Furthermore, this kind of functional module analysis is essential to evaluate tissue heterogeneity, and commonalities, which is critical for tissue-specific disease studies and drug design [21].

4.3 Gene-Gene Relationships Across Ageing

Among highly correlated clusters captured in a cross-age sample subset, some revealed a significant decrease in correlation across the several age group subsets. The ones with the most decrease were keratin-related clusters hypothesized to play a part in the decline of the healthy proprieties of skin and hair shaft during ageing. One other group of clusters with a significant decrease in correlation across ageing were immune system-related clusters. These clusters mainly comprised GO terms associated with complement binding, neutrophils, natural killer cells and major histocompatibility complex class I molecules. All of these systems were discussed to have declining functionality across ageing with impactful consequences. It was proposed that these cluster's gene-gene relationships might be interesting to delve into as means to assess the underlying mechanism of the respective systems decline during ageing.

Additionally, a more genome-wide approach allowed to evaluate which gene-gene relationships explain the most the data variance in the direction of ageing, as well as the 'hubness' of genes in the same perspective as 'hub genes of ageing'. Several deeply interconnected GO enriched terms were obtained with both analysis, which revealed to be consistent with the consensual hallmarks of ageing. As a matter of fact, the most prominent obtained GO terms were related to proteostasis, immune system, cell cycle regulation, respiratory chain, cellular proliferation, locomotion, and structure.

This analysis further allows the evaluation of the gene-gene pairs most relevant for ageing within GO terms such as Protein Folding related ones or even between distinct GO terms as between protein folding and immune system-related ones in graph network.

In this work it was shown that there are large-scale changes in gene co-expression associated with the ageing process. The implemented analysis is an approach that might prove helpful for the increasing research effort focused on identifying age-related changes in areas such as the immune function in the hope of developing intervention strategies to delay or prevent ageing phenotypes such as immune senescence. Additionally, these observations provide additional evidence of transcriptional dysregulation, and if transcriptional dysregulation is a mechanism that links damage accumulation with the decline of tissue function, ageing therapies should not only focus on fixing the known specific mechanisms but also face the more substantial challenge of preventing or slowing down the damage accumulation.

4.4 Future Work

The remaining genes from the top 300 ”hub genes of ageing” that did not enrich any particular GO term should not be overlooked as they are no less important and are equally interesting targets for analysis in future work.

In fact, lincRNAs were not explored because they predominantly lacked GO term annotation. But lincRNAs are part of the long non-coding RNA (lncRNA)s which modulate gene expression patterns at the transcriptional, post-transcriptional, and post-translational levels affecting key cellular processes such as differentiation, quiescence, proliferation, senescence, the cellular response to stress and immune agents, and many others cellular functions as relevant to the biology of ageing [116]. The objective will also be to assess the co-expression decline of lincRNAs gene-gene pairs whether it be within lincRNAs, or between lincRNAs and GO or GO GSEA enriched terms, or between lincRNAs and any other gene.

The next, ongoing, objective is to repeat the assessment of the ”hub genes of ageing” analysis not with PCA loadings, but with a more straightforward metric of the gene-gene co-expression decline across ageing. The metric already being implemented is the linear regression slope (of the correlations) against an age vector ($\{25,35,45,55,65,75\}$) and the respective p-value. This way, it is expected to be more practical to select the relevant interactions for ageing, *i.e.*, the ones with the highest slope and an acceptable p-value. The slope values, contrary to the correlation or scaled PCA, gives insight into how accentuate is the co-expression change across ageing and the p-value also serves to statistically test whether the impact of a specific relationship on ageing is significant or not.

The succeeding objective is to assess which of those relevant interactions attenuate the most the decline of their co-expression across ageing in gonad tissues. More specifically, testis tissue because it is the one with the most available samples and because it comprises many Spermatogonial Stem Cells (SSCs).

Since the genetic information contained within germ cells is passed from generation to generation, the germline (like SSCs) is often referred to as immortal. Therefore, germ cells may possess unique strategies to protect and transmit the genetic information contained within them indefinitely [117]. This might prove to be rather interesting given that there is evidence [118] that immortality in stem cells (such as SSCs) should be regulated by increased proteostasis. And proteostasis was recurrently observed as relevant for ageing in the present work.

Here, testis tissue is thought to be the best option for two main reasons:

1. Testis must be one of the tissues with a greater concentration of stem cells.
2. Testis Stem cells, even if they age in some way, still need to give rise to healthy sperm line cells without any of the normal age-related patterns (e.g. accumulated damage), otherwise upon fertilization, humans would give rise to a zygote cell with age patterns as accumulated damage.

Summarising, the objective is to assess the gene-gene interactions in which their decline in co-expression (slope) across age attenuates the most in the testis tissues when compared to the cross-tissue (somatic) approach. Thus, it might be a way to identify gene-gene relationships fundamental to maintaining a young cellular phenotype. So, gene-gene relationships that would significantly decrease their coordination across age in somatic tissues but less in testis.

Since the accumulation of damaged proteins is linked to many neurodegenerative disorders and other age-related disorders, a better understanding of the processes of stem cell function and proteostasis could lead to better treatment of those illnesses.

Bibliography

[1] R. Y. Yang, J. Quan, R. Sodaei, F. Aguet, A. V. Segrè, J. A. Allen, T. A. Lanz, V. Reinhart, M. Crawford, S. Hasson, G. Consortium, K. G. Ardlie, R. Guigó, and H. S. Xi, "A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation," *bioRxiv*, p. 311563, 4 2018. [Online]. Available: https://www.biorxiv.org/content/10.1101/311563v1https://www.biorxiv.org/content/10.1101/311563v1.abstract

[2] D. M. Hwang, A. A. Dempsey, R.-X. Wang, M. Rezvani, J. D. Barrans, MHSc, K.-S. Dai, H.-Y. Wang, H. Ma, E. Cukerman, Y.-Q. Liu, J.-R. Gu, J.-H. Zhang, S. K. W. Tsui, M. M. Y. Waye, K.-P. Fung, C.-Y. Lee, and C.-C. Liew, "A Genome-Based Resource for Molecular Cardiovascular Medicine," *Circulation*, vol. 96, no. 12, pp. 4146–4203, 12 1997. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.CIR.96.12.4146

[3] E. L. Gautier, T. Shay, J. Miller, M. Greter, C. Jakubzick, S. Ivanov, J. Helft, A. Chow, K. G. Elpek, S. Gordonov, A. R. Mazloom, A. Ma'ayan, W.-J. Chua, T. H. Hansen, S. J. Turley, M. Merad, and G. J. Randolph, "Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages," *Nature Immunology 2012 13:11*, vol. 13, no. 11, pp. 1118–1128, 9 2012. [Online]. Available: https://www.nature.com/articles/ni.2419

[4] M. W. Painter, S. Davis, R. R. Hardy, D. Mathis, C. Benoist, and T. I. G. P. Consortium, "Transcriptomes of the B and T Lineages Compared by Multiplatform Microarray Profiling," *The Journal of Immunology*, vol. 186, no. 5, pp. 3047–3057, 3 2011. [Online]. Available: https://www.jimmunol.org/content/186/5/3047https://www.jimmunol.org/content/186/5/3047.abstract

[5] T. F. C. (DGT), the RIKEN PMI, and CLST, "A promoter-level mammalian expression atlas," *Nature*, vol. 507, no. 7493, p. 462, 1 2014. [Online]. Available: /pmc/articles/PMC4529748//pmc/articles/PMC4529748/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4529748/

[6] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel,

B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin, "An atlas of active enhancers across human cell types and tissues," *Nature 2014 507:7493*, vol. 507, no. 7493, pp. 455–461, 3 2014. [Online]. Available: https://www.nature.com/articles/nature12787

[7] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak, "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes," *Proceedings of the National Academy of Sciences*, vol. 105, no. 52, pp. 20 870–20 875, 12 2008. [Online]. Available: https://www.pnas.org/content/105/52/20870https://www.pnas.org/content/105/52/20870.abstract

[8] R. J. Schaefer, R. Briskine, N. M. Springer, and C. L. Myers, "Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser," *PLoS ONE*, vol. 9, no. 6, 6 2014.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, 1998.

[10] B. D. Harris, M. Crow, S. Fischer, J. Gillis Correspondence, and J. Gillis, "Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain ll Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain," *Cell Systems*, vol. 12, 2021. [Online]. Available: https://doi.org/10.1016/j.cels.2021.04.010

[11] M. Crow, A. Paul, S. Ballouz, Z. J. Huang, and J. Gillis, "Exploiting single-cell expression to characterize co-expression replicability," *Genome Biology 2016 17:1*, vol. 17, no. 1, pp. 1–19, 5 2016. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0964-6

[12] S. L, H. SC, W. A, C. R, N. JR, C. H, W. M, T. J, B.-J. Z, and E. JR, "A transcription factor hierarchy defines an environmental stress response network," *Science (New York, N.Y.)*, vol. 354, no. 6312, 11 2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27811239/

[13] A. Day, J. Dong, V. A. Funari, B. Harry, S. P. Strom, D. H. Cohn, and S. F. Nelson, "Disease Gene Characterization through Large-Scale Co-Expression Analysis," *PLOS ONE*, vol. 4, no. 12, p. e8491, 2009. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008491

[14] Q. Long, C. Argmann, S. M. Houten, T. Huang, S. Peng, Y. Zhao, Z. Tu, and J. Zhu, "Inter-tissue coexpression network analysis reveals DPP4 as an important gene in heart to blood communication," *Genome Medicine 2016 8:1*, vol. 8, no. 1, pp. 1–15, 2 2016. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0268-1

[15] S. Tornow and H. W. Mewes, "Functional modules by relating protein interaction networks and gene expression," *Nucleic Acids Research*, vol. 31, no. 21, p. 6283, 11 2003. [Online]. Available: /pmc/articles/PMC275479//pmc/articles/PMC275479/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC275479/

[16] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–1133, 2 2003. [Online]. Available: https://www.pnas.org/content/100/3/1128https://www.pnas.org/content/100/3/1128.abstract

[17] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature 1999 402:6761*, vol. 402, no. 6761, pp. C47–C52, 12 1999. [Online]. Available: https://www.nature.com/articles/35011540

[18] W. Saelens, R. Cannoodt, and Y. Saeys, "A comprehensive evaluation of module detection methods for gene expression data," *Nature Communications 2018 9:1*, vol. 9, no. 1, pp. 1–12, 3 2018. [Online]. Available: https://www.nature.com/articles/s41467-018-03424-4

[19] E. Pierson, t. G. Consortium, D. Koller, A. Battle, and S. Mostafavi, "Sharing and Specificity of Co-expression Networks across 35 Human Tissues," *PLOS Computational Biology*, vol. 11, no. 5, p. e1004220, 5 2015. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004220

[20] R. M. Piro, U. Ala, I. Molineris, E. Grassi, C. Bracco, G. P. Perego, P. Provero, and F. Di Cunto, "An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction," *European Journal of Human Genetics 2011 19:11*, vol. 19, no. 11, pp. 1173–1180, 6 2011. [Online]. Available: https://www.nature.com/articles/ejhg201196

[21] B. He, J. Xu, Y. Tian, B. Liao, J. Lang, H. Lin, X. Mo, Q. Lu, G. Tian, and P. Bing, "Gene Coexpression Network and Module Analysis across 52 Human Tissues," *BioMed Research International*, vol. 2020, 2020.

[22] A. Fønss Møller and K. Nath Natarajan, "Predicting gene regulatory networks from cell atlases," *Life Science Alliance*, 2020. [Online]. Available: http://doi.org/10.26508/lsa.202000658

[23] T. Strunz, M. Kellner, C. Kiel, and B. H. F. Weber, "Assigning Co-Regulated Human Genes and Regulatory Gene Clusters," *Cells 2021, Vol. 10, Page 2395*, vol. 10, no. 9,

p. 2395, 9 2021. [Online]. Available: https://www.mdpi.com/2073-4409/10/9/2395/htmhttps://www.mdpi.com/2073-4409/10/9/2395

[24] M. C. Barbosa, R. A. Grosso, and C. M. Fader, “Hallmarks of Aging: An Autophagic Perspective,” *Frontiers in Endocrinology*, vol. 0, no. JAN, p. 790, 2019.

[25] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The Hallmarks of Aging,” *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013.

[26] F. Guerville, P. De Souto Barreto, I. Ader, S. Andrieu, L. Casteilla, C. Dray, N. Fazilleau, S. Guyonnet, D. Langin, R. Liblau, A. Parini, P. Valet, N. Vergnolle, Y. Rolland, and B. Vellas, “Revisiting the Hallmarks of Aging to Identify Markers of Biological Age,” *The Journal of Prevention of Alzheimer's Disease 2019 7:1*, vol. 7, no. 1, pp. 56–64, 12 2019. [Online]. Available: https://link.springer.com/article/10.14283/jpad.2019.50

[27] M. Mittelbrunn and G. Kroemer, “Hallmarks of T cell aging,” *Nature Immunology 2021 22:6*, vol. 22, no. 6, pp. 687–698, 5 2021. [Online]. Available: https://www.nature.com/articles/s41590-021-00927-z

[28] J. Vijg, “Loss of gene coordination as a stochastic cause of ageing,” *Nature Metabolism*, 2020. [Online]. Available: https://doi.org/10.1038/s42255-020-00295-2

[29] A. G and L. PM, “Telomeres and aging,” *Physiological reviews*, vol. 88, no. 2, pp. 557–579, 4 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18391173/

[30] S. Hekimi, J. Lapointe, and Y. Wen, “Taking a “good” look at free radicals in the aging process,” *Trends in cell biology*, vol. 21, no. 10, p. 569, 10 2011. [Online]. Available: /pmc/articles/PMC4074523//pmc/articles/PMC4074523/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4074523/

[31] B. R, H. CH, R. KA, D. AD, B. RA, D. ME, C. RB, C. GB, P. BH, K. CA, and V. J, “Increased cell-to-cell variation in gene expression in ageing mouse heart,” *Nature*, vol. 441, no. 7096, pp. 1011–1014, 6 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16791200/

[32] I. Angelidis, L. M. Simon, I. E. Fernandez, M. Strunz, C. H. Mayr, F. R. Greiffo, G. Tsitsiridis, M. Ansari, E. Graf, T.-M. Strom, M. Nagendran, T. Desai, O. Eickelberg, M. Mann, F. J. Theis, and H. B. Schiller, “An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics,” *Nature Communications 2019 10:1*, vol. 10, no. 1, pp. 1–17, 2 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-08831-9

[33] E. M, A. HE, M. M, B. J, B. R, K. SK, and Q. SR, “Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns,” *Cell*, vol. 171, no. 2, pp. 321–330, 10 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28965763/

[34] V. J and D. X, “Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging,” *Cell*, vol. 182, no. 1, pp. 12–23, 7 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32649873/

[35] O. Levy, G. Amit, D. Vaknin, T. Snir, S. Efroni, P. Castaldi, Y.-Y. Liu, H. Y. Cohen, and A. Bashan, “Age-related loss of gene-to-gene transcriptional coordination among single cells,” *Nature Metabolism 2020 2:11*, vol. 2, no. 11, pp. 1305–1315, 11 2020. [Online]. Available: https://www.nature.com/articles/s42255-020-00304-4

[36] A. S. Anisimova, M. B. Meerson, M. V. Gerashchenko, I. V. Kulakovskiy, S. E. Dmitriev, and V. N. Gladyshev, “Multifaceted deregulation of gene expression and protein synthesis with age,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 581–15 590, 7 2020. [Online]. Available: https://www.pnas.org/content/117/27/15581https://www.pnas.org/content/117/27/15581.abstract

[37] L. K. Southworth, A. B. Owen, and S. K. Kim, “Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules,” *PLoS Genet*, vol. 5, no. 12, p. 1000776, 2009. [Online]. Available: www.plosgenetics.org

[38] “GTEx Portal.” [Online]. Available: https://www.gtexportal.org/home/

[39] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, “The RIN: an RNA integrity number for assigning integrity values to RNA measurements,” *BMC Molecular Biology 2006 7:1*, vol. 7, no. 1, pp. 1–14, 1 2006. [Online]. Available: https://bmcmolbiol.biomedcentral.com/articles/10.1186/1471-2199-7-3

[40] “GTEx Portal.” [Online]. Available: https://www.gtexportal.org/home/documentationPage#staticTextSampleCollection

[41] R. Khetani, “Introduction to DGE: count normalization with DESeq2,” 2017. [Online]. Available: https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

[42] “Bioconductor - DESeq2.” [Online]. Available: http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html

[43] G. RC, C. VJ, B. DM, B. B, D. M, D. S, E. B, G. L, G. Y, G. J, H. K, H. T, H. W, I. S, I. R, L. F, L. C, M. M, R. AJ, S. G, S. C, S. G, T. L, Y. JY, and Z. J, “Bioconductor: open software development

for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, 2004. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15461798/

[44] P. G. Ferreira, M. Muñoz-Aguirre, F. Reverter, C. P. Sá Godinho, A. Sousa, A. Amadoz, R. Sodaei, M. R. Hidalgo, D. Pervouchine, J. Carbonell-Caballero, R. Nurtdinov, A. Breschi, R. Amador, P. Oliveira, C. Çubuk, J. Curado, F. Aguet, C. Oliveira, J. Dopazo, M. Sammeth, K. G. Ardlie, and R. Guigó, "The effects of death and post-mortem cold ischemia on human tissue transcriptomes," *Nature Communications*, vol. 9, no. 1, pp. 1–15, 12 2018. [Online]. Available: www.nature.com/naturecommunications

[45] J. Somekh, S. S. Shen-Orr, and I. S. Kohane, "Batch correction evaluation framework using a-priori gene-gene associations: Applied to the GTEx dataset," *BMC Bioinformatics*, vol. 20, no. 1, p. 268, 5 2019. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2855-9

[46] S. N. Gorb and S. Reports, "Interactions and contrasts," pp. 1–17, 2013. [Online]. Available: https://genomicsclass.github.io/book/pages/interactions_and_contrasts.html

[47] "CRAN - Package pheatmap." [Online]. Available: https://cran.r-project.org/web/packages/pheatmap/index.html

[48] "Bioconductor - org.Hs.eg.db." [Online]. Available: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

[49] "Bioconductor - EnsDb.Hsapiens.v79." [Online]. Available: https://bioconductor.org/packages/release/data/annotation/html/EnsDb.Hsapiens.v79.html

[50] "Bioconductor - topGO." [Online]. Available: https://bioconductor.org/packages/release/bioc/html/topGO.html

[51] "Create Elegant Data Visualisations Using the Grammar of Graphics [R package ggplot2 version 3.3.5]," 6 2021. [Online]. Available: https://cran.r-project.org/package=ggplot2

[52] "Bioconductor - clusterProfiler." [Online]. Available: https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

[53] "CRAN - Package qgraph." [Online]. Available: https://cran.r-project.org/web/packages/qgraph/index.html

[54] T. Sano, T. Kume, T. Fujimura, H. Kawada, S. Moriwaki, and Y. Takema, "The formation of wrinkles caused by transition of keratin intermediate filaments after repetitive UVB exposure,"

Archives of Dermatological Research 2004 296:8, vol. 296, no. 8, pp. 359–365, 12 2004. [Online]. Available: https://link.springer.com/article/10.1007/s00403-004-0533-9

[55] G. M, G. S, H. O, F. G, K. A, and P. D, "Ageing processes influence keratin and KAP expression in human hair follicles," *Experimental dermatology*, vol. 20, no. 9, pp. 759–761, 9 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21569108/

[56] B. Schmidt and V. Horsley, "Unraveling hair follicle-adipocyte communication," *Experimental dermatology*, vol. 21, no. 11, p. 827, 2012. [Online]. Available: /pmc/articles/PMC3507425//pmc/articles/PMC3507425/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3507425/

[57] C. N. Joyce Chen, L. D. Thompson, and L. A. Snow, "Muscle Structure and Function," *Orthopaedic Physical Therapy Secrets: Third Edition*, pp. 1–9, 1 2017.

[58] A. Damirchi, P. Babaei, M. Gholamali, and K. Ranjbar, "Mitochondrial Biogenesis in Skeletal Muscle: Exercise and Aging," *Skeletal Muscle - From Myogenesis to Clinical Relations*, 8 2012. [Online]. Available: https://www.intechopen.com/chapters/38419

[59] L. Norlén and A. Al-Amoudi, "Stratum Corneum Keratin Structure, Function, and Formation: The Cubic Rod-Packing and Membrane Templating Model," *Journal of Investigative Dermatology*, vol. 123, no. 4, pp. 715–732, 10 2004.

[60] T. Sano, T. Kume, T. Fujimura, H. Kawada, and K. Higuchi, "Keratin alterations could be an early event of wrinkle formation," *Journal of Dermatological Science*, vol. 53, no. 1, pp. 77–79, 1 2009. [Online]. Available: http://www.jdsjournal.com/article/S0923181108002363/fulltexthttp://www.jdsjournal.com/article/S0923181108002363/abstracthttps://www.jdsjournal.com/article/S0923-1811(08)00236-3/abstract

[61] H. H. Bragulla and D. G. Homberger, "Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia," *Journal of Anatomy*, vol. 214, no. 4, p. 516, 2009. [Online]. Available: /pmc/articles/PMC2736122//pmc/articles/PMC2736122/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2736122/

[62] J. Hazeldine and J. M. Lord, "The impact of ageing on natural killer cell function and potential consequences for health in older adults," *Ageing Research Reviews*, vol. 12, no. 4, p. 1069, 9 2013. [Online]. Available: /pmc/articles/PMC4147963//pmc/articles/PMC4147963/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147963/

[63] M. V. Carroll and R. B. Sim, "Complement in health and disease," *Advanced Drug Delivery Reviews*, vol. 63, no. 12, pp. 965–975, 9 2011.

[64] A. Armento, M. Ueffing, and S. J. Clark, “The complement system in age-related macular degeneration,” *Cellular and Molecular Life Sciences 2021 78:10*, vol. 78, no. 10, pp. 4487–4505, 3 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00018-021-03796-9

[65] E. Mortaz, S. D. Alipoor, I. M. Adcock, S. Mumby, and L. Koenderman, “Update on Neutrophil Function in Severe Inflammation,” *Frontiers in Immunology*, vol. 0, no. OCT, p. 2171, 10 2018.

[66] C. Rosales, “Neutrophil: A Cell with Many Roles in Inflammation or Several Cell Types?” *Frontiers in Physiology*, vol. 0, no. FEB, p. 113, 2 2018.

[67] S. Butcher, H. Chahel, and J. M. Lord, “Ageing and the neutrophil: no appetite for killing?” *Immunology*, vol. 100, no. 4, p. 411, 2000. [Online]. Available: /pmc/articles/PMC2327031//pmc/articles/PMC2327031/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2327031/

[68] I. Prager and C. Watzl, “Mechanisms of natural killer cell-mediated cellular cytotoxicity,” *Journal of Leukocyte Biology*, vol. 105, no. 6, pp. 1319–1329, 6 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/JLB.MR0718-269Rhttps://onlinelibrary.wiley.com/doi/abs/10.1002/JLB.MR0718-269Rhttps://jlb.onlinelibrary.wiley.com/doi/10.1002/JLB.MR0718-269R

[69] E. W. Hewitt, “The MHC class I antigen presentation pathway: strategies for viral immune evasion,” *Immunology*, vol. 110, no. 2, pp. 163–169, 10 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-2567.2003.01738.xhttps://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2567.2003.01738.xhttps://onlinelibrary.wiley.com/doi/10.1046/j.1365-2567.2003.01738.x

[70] D. Dersh, J. Hollý, and J. W. Yewdell, “A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion,” *Nature Reviews Immunology 2020 21:2*, vol. 21, no. 2, pp. 116–128, 8 2020. [Online]. Available: https://www.nature.com/articles/s41577-020-0390-6

[71] J. A. Rudd-Schmidt, A. W. Hodel, T. Noori, J. A. Lopez, H.-J. Cho, S. Verschoor, A. Ciccone, J. A. Trapani, B. W. Hoogenboom, and I. Voskoboinik, “Lipid order and charge protect killer T cells from accidental death,” *Nature Communications*, vol. 10, no. 1, 12 2019. [Online]. Available: /pmc/articles/PMC6881447//pmc/articles/PMC6881447/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6881447/

[72] “B2M beta-2-microglobulin [Homo sapiens (human)] - Gene - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=567

[73] M. Basler, C. J. Kirk, and M. Groettrup, “The immunoproteasome in antigen processing and other immunological functions,” *Current Opinion in Immunology*, vol. 25, no. 1, pp. 74–80, 2 2013.

[74] E. Wieczorek and M. A. Garstka, “Recurrent bladder cancer in aging societies: Importance of major histocompatibility complex class I antigen presentation,” *International Journal of Cancer*, vol. 148, no. 8, pp. 1808–1820, 4 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.33359https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.33359https://onlinelibrary.wiley.com/doi/10.1002/ijc.33359

[75] M. J. Lazarczyk, J. E. Kemmler, B. A. Eyford, J. A. Short, M. Varghese, A. Sowa, D. R. Dickstein, F. J. Yuk, R. Puri, K. E. Biron, M. Leist, W. A. Jefferies, and D. L. Dickstein, “Major Histocompatibility Complex class I proteins are critical for maintaining neuronal structural complexity in the aging brain,” *Scientific Reports 2016 6:1*, vol. 6, no. 1, pp. 1–13, 5 2016. [Online]. Available: https://www.nature.com/articles/srep26199

[76] J. Groh, K. Knöpper, P. Arampatzi, X. Yuan, L. Lößlein, A.-E. Saliba, W. Kastenmüller, and R. Martini, “Accumulation of cytotoxic T cells in the aged CNS leads to axon degeneration and contributes to cognitive and motor decline,” *Nature Aging 2021 1:4*, vol. 1, no. 4, pp. 357–367, 4 2021. [Online]. Available: https://www.nature.com/articles/s43587-021-00049-z

[77] S.-J. Yi and K. Kim, “New Insights into the Role of Histone Changes in Aging,” *International Journal of Molecular Sciences*, vol. 21, no. 21, pp. 1–20, 11 2020. [Online]. Available: /pmc/articles/PMC7662996//pmc/articles/PMC7662996/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7662996/

[78] G. J. Prud’homme, “Pathobiology of transforming growth factor β in cancer, fibrosis and immunologic disease, and therapeutic considerations,” *Laboratory Investigation 2007 87:11*, vol. 87, no. 11, pp. 1077–1091, 8 2007. [Online]. Available: https://www.nature.com/articles/3700669

[79] D. Kim, S. Y. Kim, S. K. Mun, S. Rhee, and B. J. Kim, “Epidermal growth factor improves the migration and contractility of aged fibroblasts cultured on 3D collagen matrices,” *International Journal of Molecular Medicine*, vol. 35, no. 4, pp. 1017–1025, 4 2015. [Online]. Available: http://www.spandidos-publications.com/10.3892/ijmm.2015.2088/abstracthttps://www.spandidos-publications.com/10.3892/ijmm.2015.2088

[80] S. K and G. N, “Vimentin Plays a Crucial Role in Fibroblast Ageing by Regulating Biophysical Properties and Cell Migration,” *Cells*, vol. 8, no. 10, 9 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31569795/

[81] S. Takyar, R. P. Hickerson, and H. F. Noller, “mRNA Helicase Activity of the Ribosome,” *Cell*, vol. 120, no. 1, pp. 49–58, 1 2005. [Online].

Available: http://www.cell.com/article/S0092867404011468/fulltexthttp://www.cell.com/article/S0092867404011468/abstracthttps://www.cell.com/cell/abstract/S0092-8674(04)01146-8

[82] H. Chen and D. C. Chan, "Physiological functions of mitochondrial fusion," *Annals of the New York Academy of Sciences*, vol. 1201, no. 1, pp. 21–25, 7 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1749-6632.2010.05615.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2010.05615.xhttps://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2010.05615.x

[83] Y. J. Liu, R. L. McIntyre, G. E. Janssens, and R. H. Houtkooper, "Mitochondrial fission and fusion: A dynamic role in aging and potential target for age-related disease," *Mechanisms of Ageing and Development*, vol. 186, p. 111212, 3 2020.

[84] G. Chen, G. Kroemer, and O. Kepp, "Mitophagy: An Emerging Role in Aging and Age-Associated Diseases," *Frontiers in Cell and Developmental Biology*, vol. 0, p. 200, 3 2020.

[85] J. Janikiewicz, J. Szymański, D. Malinska, P. Patalas-Krawczyk, B. Michalska, J. Duszyński, C. Giorgi, M. Bonora, A. Dobrzyn, and M. R. Wieckowski, "Mitochondria-associated membranes in aging and senescence: structure, function, and dynamics," *Cell Death & Disease 2018 9:3*, vol. 9, no. 3, pp. 1–12, 2 2018. [Online]. Available: https://www.nature.com/articles/s41419-017-0105-5

[86] T. Fulop, A. L. Page, H. Garneau, N. Azimi, S. Baehl, G. Dupuis, G. Pawelec, and A. Larbi, "Aging, immunosenescence and membrane rafts: the lipid connection," *Longevity & Healthspan*, vol. 1, no. 1, p. 6, 12 2012. [Online]. Available: /pmc/articles/PMC3886260//pmc/articles/PMC3886260/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3886260/

[87] Y. Ohno-Iwashita, Y. Shimada, M. Hayashi, and M. Inomata, "Plasma membrane microdomains in aging and disease," *Geriatrics & Gerontology International*, vol. 10, no. SUPPL. 1, pp. S41–S52, 7 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1447-0594.2010.00600.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1447-0594.2010.00600.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1447-0594.2010.00600.x

[88] A. E. Morgan, K. M. Mooney, S. J. Wilkinson, N. A. Pickles, and M. T. Mc Auley, "Cholesterol Metabolism: A Review of How Ageing Disrupts the Biological Mechanisms Responsible for its Regulation," *Ageing Research Reviews*, 2016. [Online]. Available: http://dx.

[89] D.-C. B, B.-O. MA, M.-S. M, and M.-O. J, "Cholesterol: recapitulation of its active role during liver regeneration," *Liver international : official journal of the International Association for the Study of the Liver*, vol. 31, no. 9, pp. 1271–1284, 10 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21745289/

[90] M. PL and M. EG, "Inflammation and the degenerative diseases of aging," *Annals of the New York Academy of Sciences*, vol. 1035, pp. 104–116, 2004. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15681803/

[91] S. G, R. DS, and C. W, "EGF and TGF-alpha in wound healing and repair," *Journal of cellular biochemistry*, vol. 45, no. 4, pp. 346–352, 1991. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2045428/

[92] Y. A, M. Y, T. A, U. E, and K. Y, "Effect of EGF and bFGF on fibroblast proliferation and angiogenic cytokine production from cultured dermal substitutes," *Journal of biomaterials science. Polymer edition*, vol. 23, no. 10, pp. 1315–1324, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21722419/

[93] J. Small, T. Stradal, E. Vignal, and K. Rottner, "The lamellipodium: where motility begins," *Trends in Cell Biology*, vol. 12, no. 3, pp. 112–120, 3 2002. [Online]. Available: http://www.cell.com/article/S0962892401022371/fulltexthttp://www.cell.com/article/S0962892401022371/abstracthttps://www.cell.com/trends/cell-biology/abstract/S0962-8924(01)02237-1

[94] B. Roy, L. Yuan, Y. Lee, A. Bharti, A. Mitra, and G. V. Shivashankar, "Fibroblast rejuvenation by mechanical reprogramming and redifferentiation," *Proceedings of the National Academy of Sciences*, vol. 117, no. 19, pp. 10 131–10 141, 5 2020. [Online]. Available: https://www.pnas.org/content/117/19/10131https://www.pnas.org/content/117/19/10131.abstract

[95] A. M. Handorf, Y. Zhou, M. A. Halanski, and W.-J. Li, "Tissue Stiffness Dictates Development, Homeostasis, and Disease Progression," *Organogenesis*, vol. 11, no. 1, p. 1, 1 2015. [Online]. Available: /pmc/articles/PMC4594591//pmc/articles/PMC4594591/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4594591/

[96] R. Li and G. G. Gundersen, "Beyond polymer polarity: how the cytoskeleton builds a polarized cell," *Nature Reviews Molecular Cell Biology 2008 9:11*, vol. 9, no. 11, pp. 860–873, 11 2008. [Online]. Available: https://www.nature.com/articles/nrm2522

[97] E. Mejia-Ramirez, H. Geiger, and M. C. Florian, "Loss of epigenetic polarity is a hallmark of hematopoietic stem cell aging," *Human Molecular Genetics*, vol. 29, no. R2, pp. R248–R254, 10 2020. [Online]. Available: https://academic.oup.com/hmg/article/29/R2/R248/5894945

[98] H. Soares, H. S. Marinho, C. Real, and F. Antunes, "Cellular polarity in aging: role of redox regulation and nutrition," *Genes & Nutrition*, vol. 9, no. 1, 1 2014. [Online]. Available: /pmc/articles/PMC3896621//pmc/articles/PMC3896621/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3896621/

[99] D. Loeffler and T. Schroeder, "ASYMMETRIC CELL DIVISION CONTROLS HEMATOPOIETIC STEM CELL METABOLIC ACTIVATION AND DIFFERENTIATION," *Experimental Hematology*, vol. 76, p. S76, 8 2019.

[100] L. Zhang, R. Mack, P. Breslin, and J. Zhang, "Molecular and cellular mechanisms of aging in hematopoietic stem cells and their niches," *Journal of Hematology & Oncology 2020 13:1*, vol. 13, no. 1, pp. 1–22, 11 2020. [Online]. Available: https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-00994-z

[101] D. L. Smith, "Anemia in the Elderly," *American Family Physician*, vol. 62, no. 7, pp. 1565–1572, 10 2000.

[102] A. AS, A. AI, M. NE, G. VN, and D. SE, "Protein synthesis and quality control in aging," *Aging*, vol. 10, no. 12, pp. 4269–4288, 12 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30562164/

[103] W. Li and L. Zhang, *Regulation of ATG and Autophagy Initiation*, 2019, vol. 1206.

[104] B. R, S. Z, Z. JW, L. D, Z. YL, and Z. S, "ST13, a proliferation regulator, inhibits growth and migration of colorectal cancer cell lines," *Journal of Zhejiang University. Science. B*, vol. 13, no. 11, pp. 884–893, 11 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23125081/

[105] "Colorectal Cancer Prevention (PDQ®)–Health Professional Version - National Cancer Institute." [Online]. Available: https://www.cancer.gov/types/colorectal/hp/colorectal-prevention-pdq#section/all

[106] C. Garrido, L. Paco, I. Romero, E. Berruguilla, J. Stefansky, A. Collado, I. Algarra, F. Garrido, and A. M. Garcia-Lora, "MHC class I molecules act as tumor suppressor genes regulating the cell cycle gene expression, invasion and intrinsic tumorigenicity of melanoma cells," *Carcinogenesis*, vol. 33, no. 3, pp. 687–693, 3 2012. [Online]. Available: https://academic.oup.com/carcin/article/33/3/687/2464120

[107] K. L, G. KC, H. AB, M. N, H. AL, K. P, M. A, B. JG, and B. WF, "Loss of HLA class-I alleles, heavy chains and beta 2-microglobulin in colorectal cancer," *International journal of cancer*, vol. 51, no. 3, pp. 379–385, 1992. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1592528/

[108] Y. IA and R. KL, "Antigen processing and presentation by the class I major histocompatibility complex," *Annual review of immunology*, vol. 14, pp. 369–396, 1996. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8717519/https://pubmed.ncbi.nlm.nih.gov/8717519/?dopt=Abstract

[109] N. DT, W. SS, and L. RA, “Analysis in vivo of GRP78-BiP/substrate interactions and their role in induction of the GRP78-BiP gene,” *Molecular biology of the cell*, vol. 3, no. 2, pp. 143–155, 1992. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1550958/

[110] “SNRNP70 Gene - GeneCards — RU17 Protein — RU17 Antibody.” [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SNRNP70

[111] I. Diner, C. M. Hales, I. Bishof, L. Rabenold, D. M. Duong, H. Yi, O. Laur, M. Gearing, J. Troncoso, M. Thambisetty, J. J. Lah, A. I. Levey, and N. T. Seyfried, “Aggregation Properties of the Small Nuclear Ribonucleoprotein U1-70K in Alzheimer Disease,” *The Journal of Biological Chemistry*, vol. 289, no. 51, p. 35296, 12 2014. [Online]. Available: /pmc/articles/PMC4271217//pmc/articles/PMC4271217/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4271217/

[112] C. Huang, S. Wagner-Valladolid, A. D. Stephens, R. Jung, C. Poudel, T. Sinnige, M. C. Lechler, N. Schlörit, M. Lu, R. F. Laine, C. H. Michel, M. Vendruscolo, C. F. Kaminski, G. S. K. Schierle, and D. C. David, “Intrinsically aggregation-prone proteins form amyloid-like aggregates and contribute to tissue aging in Caenorhabditis elegans,” *eLife*, vol. 8, 5 2019. [Online]. Available: /pmc/articles/PMC6524967//pmc/articles/PMC6524967/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6524967/

[113] K. T, I. K, S. N, M. K, K. T, H. J, M. T, Y. T, G. F, M. Y, N. Y, T. J, T. T, I. Y, M. O, T. A, S. G.-H. P, T. M, and T. M, “Presenilin-1 mutations downregulate the signalling pathway of the unfolded-protein response,” *Nature cell biology*, vol. 1, no. 8, pp. 479–485, 1999. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10587643/

[114] T. Ono and R. G. Cutler, “Age-dependent relaxation of gene repression: Increase of endogenous murine leukemia virus-related and globin-related RNA in brain and liver of mice,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 75, no. 9, p. 4431, 1978. [Online]. Available: /pmc/articles/PMC336129/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC336129/

[115] P. Gaudet and C. Dessimoz, “Gene Ontology: Pitfalls, Biases, and Remedies,” *Methods in Molecular Biology*, vol. 1446, pp. 189–205, 2017. [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-4939-3743-1_14

[116] I. Grammatikakis, A. C. Panda, K. Abdelmohsen, and M. Gorospe, “Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging,” *Aging (Albany NY)*, vol. 6, no. 12, p. 992, 2014. [Online]. Available: /pmc/articles/PMC4298369//pmc/articles/PMC4298369/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4298369/

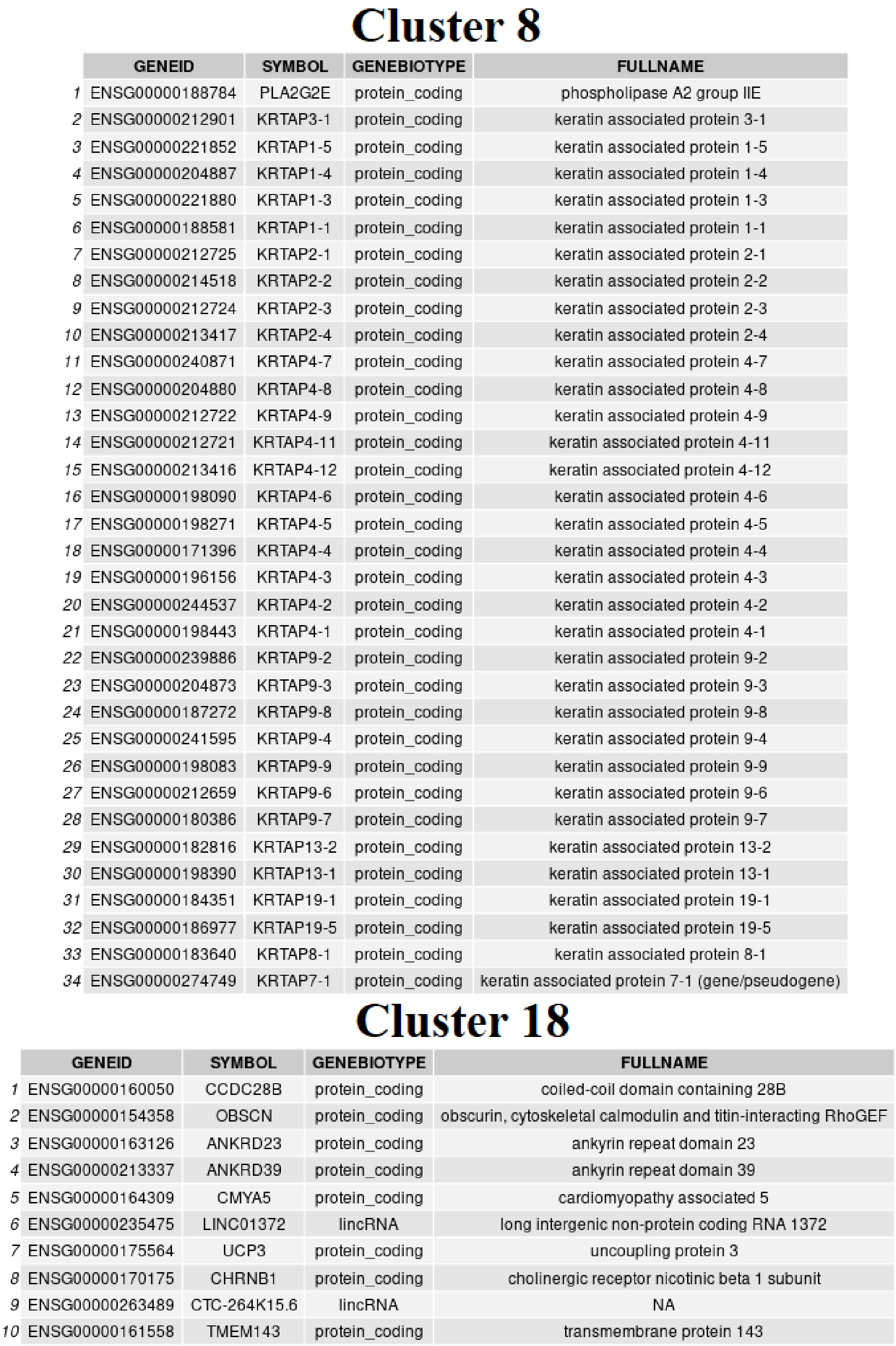
[117] J. DL, “Aging and the germ line: where mortality and immortality meet,” *Stem cell reviews*, vol. 3, no. 3, pp. 192–200, 9 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17917132/

[118] I. Saez, S. Koyuncu, R. Gutierrez-Garcia, C. Dieterich, and D. Vilchez, “Insights into the ubiquitin-proteasome system of human embryonic stem cells,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–21, 3 2018. [Online]. Available: https://www.nature.com/articles/s41598-018-22384-9

A

Suplemental Materials

A.1 Tissue Analysis: Clusters Annotation



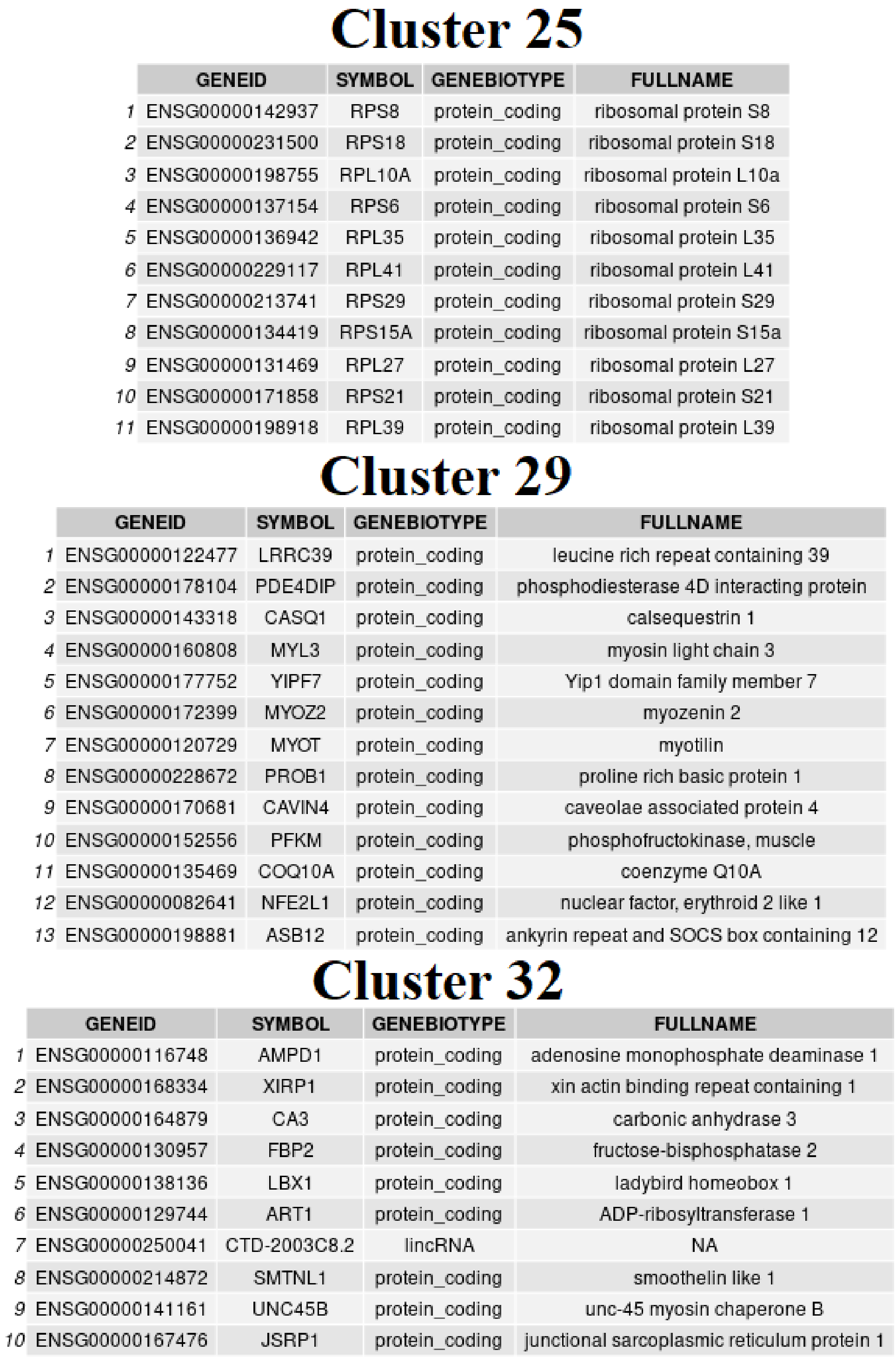
Cluster 8

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000188784	PLA2G2E	protein_coding	phospholipase A2 group IIE
2	ENSG00000212901	KRTAP3-1	protein_coding	keratin associated protein 3-1
3	ENSG00000221852	KRTAP1-5	protein_coding	keratin associated protein 1-5
4	ENSG00000204887	KRTAP1-4	protein_coding	keratin associated protein 1-4
5	ENSG00000221880	KRTAP1-3	protein_coding	keratin associated protein 1-3
6	ENSG00000188581	KRTAP1-1	protein_coding	keratin associated protein 1-1
7	ENSG00000212725	KRTAP2-1	protein_coding	keratin associated protein 2-1
8	ENSG00000214518	KRTAP2-2	protein_coding	keratin associated protein 2-2
9	ENSG00000212724	KRTAP2-3	protein_coding	keratin associated protein 2-3
10	ENSG00000213417	KRTAP2-4	protein_coding	keratin associated protein 2-4
11	ENSG00000240871	KRTAP4-7	protein_coding	keratin associated protein 4-7
12	ENSG00000204880	KRTAP4-8	protein_coding	keratin associated protein 4-8
13	ENSG00000212722	KRTAP4-9	protein_coding	keratin associated protein 4-9
14	ENSG00000212721	KRTAP4-11	protein_coding	keratin associated protein 4-11
15	ENSG00000213416	KRTAP4-12	protein_coding	keratin associated protein 4-12
16	ENSG00000198090	KRTAP4-6	protein_coding	keratin associated protein 4-6
17	ENSG00000198271	KRTAP4-5	protein_coding	keratin associated protein 4-5
18	ENSG00000171396	KRTAP4-4	protein_coding	keratin associated protein 4-4
19	ENSG00000196156	KRTAP4-3	protein_coding	keratin associated protein 4-3
20	ENSG00000244537	KRTAP4-2	protein_coding	keratin associated protein 4-2
21	ENSG00000198443	KRTAP4-1	protein_coding	keratin associated protein 4-1
22	ENSG00000239886	KRTAP9-2	protein_coding	keratin associated protein 9-2
23	ENSG00000204873	KRTAP9-3	protein_coding	keratin associated protein 9-3
24	ENSG00000187272	KRTAP9-8	protein_coding	keratin associated protein 9-8
25	ENSG00000241595	KRTAP9-4	protein_coding	keratin associated protein 9-4
26	ENSG00000198083	KRTAP9-9	protein_coding	keratin associated protein 9-9
27	ENSG00000212659	KRTAP9-6	protein_coding	keratin associated protein 9-6
28	ENSG00000180386	KRTAP9-7	protein_coding	keratin associated protein 9-7
29	ENSG00000182816	KRTAP13-2	protein_coding	keratin associated protein 13-2
30	ENSG00000198390	KRTAP13-1	protein_coding	keratin associated protein 13-1
31	ENSG00000184351	KRTAP19-1	protein_coding	keratin associated protein 19-1
32	ENSG00000186977	KRTAP19-5	protein_coding	keratin associated protein 19-5
33	ENSG00000183640	KRTAP8-1	protein_coding	keratin associated protein 8-1
34	ENSG00000274749	KRTAP7-1	protein_coding	keratin associated protein 7-1 (gene/pseudogene)

Cluster 18

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000160050	CCDC28B	protein_coding	coiled-coil domain containing 28B
2	ENSG00000154358	OBSCN	protein_coding	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF
3	ENSG00000163126	ANKRD23	protein_coding	ankyrin repeat domain 23
4	ENSG00000213337	ANKRD39	protein_coding	ankyrin repeat domain 39
5	ENSG00000164309	CMYA5	protein_coding	cardiomyopathy associated 5
6	ENSG00000235475	LINC01372	lincRNA	long intergenic non-protein coding RNA 1372
7	ENSG00000175564	UCP3	protein_coding	uncoupling protein 3
8	ENSG00000170175	CHRNB1	protein_coding	cholinergic receptor nicotinic beta 1 subunit
9	ENSG00000263489	CTC-264K15.6	lincRNA	NA
10	ENSG00000161558	TMEM143	protein_coding	transmembrane protein 143

Figure A.1: Gene annotation of clusters 8 and 18 from tissue analysis.



Cluster 25

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000142937	RPS8	protein_coding	ribosomal protein S8
2	ENSG00000231500	RPS18	protein_coding	ribosomal protein S18
3	ENSG00000198755	RPL10A	protein_coding	ribosomal protein L10a
4	ENSG00000137154	RPS6	protein_coding	ribosomal protein S6
5	ENSG00000136942	RPL35	protein_coding	ribosomal protein L35
6	ENSG00000229117	RPL41	protein_coding	ribosomal protein L41
7	ENSG00000213741	RPS29	protein_coding	ribosomal protein S29
8	ENSG00000134419	RPS15A	protein_coding	ribosomal protein S15a
9	ENSG00000131469	RPL27	protein_coding	ribosomal protein L27
10	ENSG00000171858	RPS21	protein_coding	ribosomal protein S21
11	ENSG00000198918	RPL39	protein_coding	ribosomal protein L39

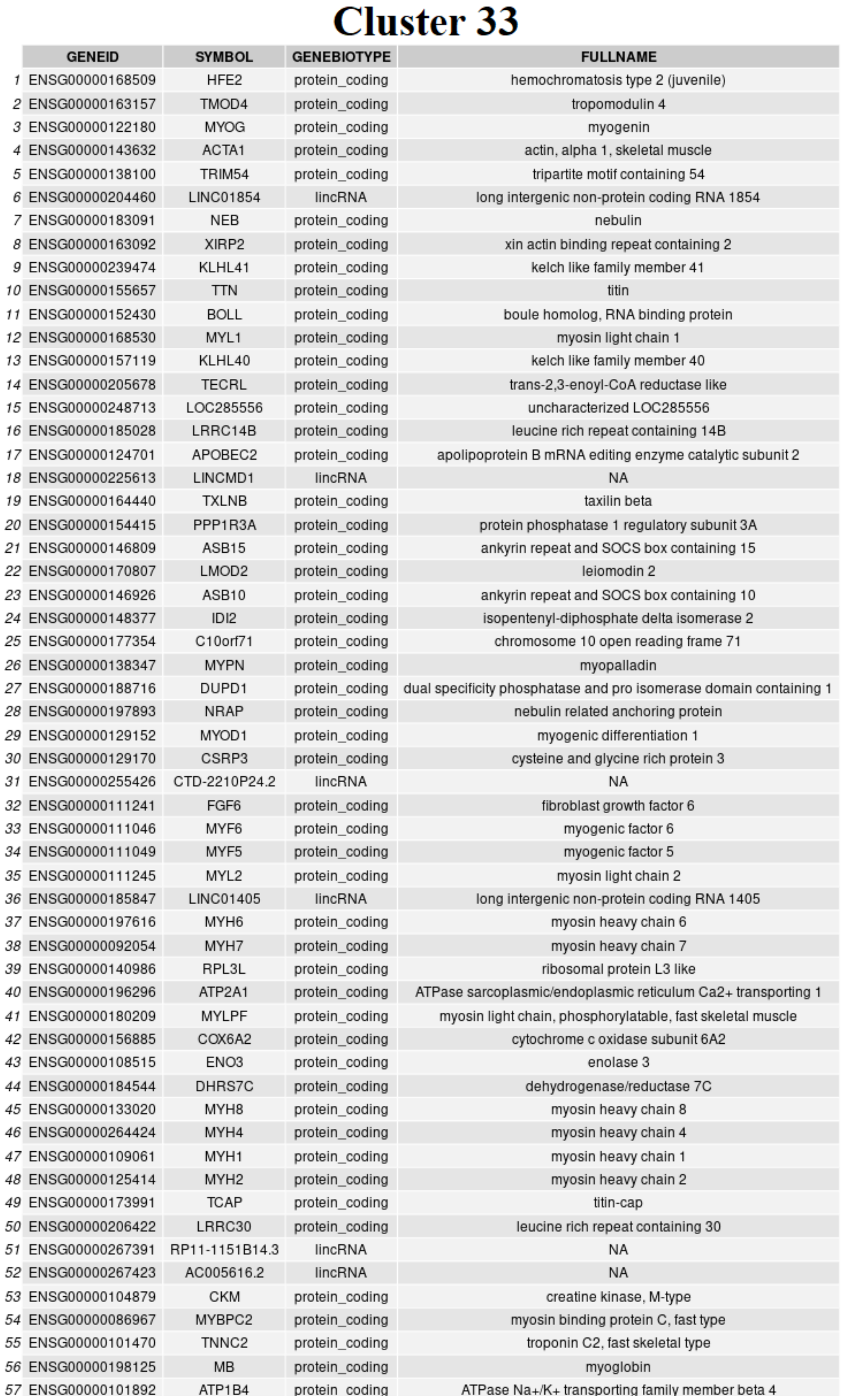
Cluster 29

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000122477	LRRC39	protein_coding	leucine rich repeat containing 39
2	ENSG00000178104	PDE4DIP	protein_coding	phosphodiesterase 4D interacting protein
3	ENSG00000143318	CASQ1	protein_coding	calsequestrin 1
4	ENSG00000160808	MYL3	protein_coding	myosin light chain 3
5	ENSG00000177752	YIPF7	protein_coding	Yip1 domain family member 7
6	ENSG00000172399	MYOZ2	protein_coding	myozenin 2
7	ENSG00000120729	MYOT	protein_coding	myotilin
8	ENSG00000228672	PROB1	protein_coding	proline rich basic protein 1
9	ENSG00000170681	CAVIN4	protein_coding	caveolae associated protein 4
10	ENSG00000152556	PFKM	protein_coding	phosphofructokinase, muscle
11	ENSG00000135469	COQ10A	protein_coding	coenzyme Q10A
12	ENSG00000082641	NFE2L1	protein_coding	nuclear factor, erythroid 2 like 1
13	ENSG00000198881	ASB12	protein_coding	ankyrin repeat and SOCS box containing 12

Cluster 32

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000116748	AMPD1	protein_coding	adenosine monophosphate deaminase 1
2	ENSG00000168334	XIRP1	protein_coding	xin actin binding repeat containing 1
3	ENSG00000164879	CA3	protein_coding	carbonic anhydrase 3
4	ENSG00000130957	FBP2	protein_coding	fructose-bisphosphatase 2
5	ENSG00000138136	LBX1	protein_coding	ladybird homeobox 1
6	ENSG00000129744	ART1	protein_coding	ADP-ribosyltransferase 1
7	ENSG00000250041	CTD-2003C8.2	lincRNA	NA
8	ENSG00000214872	SMTNL1	protein_coding	smoothelin like 1
9	ENSG00000141161	UNC45B	protein_coding	unc-45 myosin chaperone B
10	ENSG00000167476	JSRP1	protein_coding	junctional sarcoplasmic reticulum protein 1

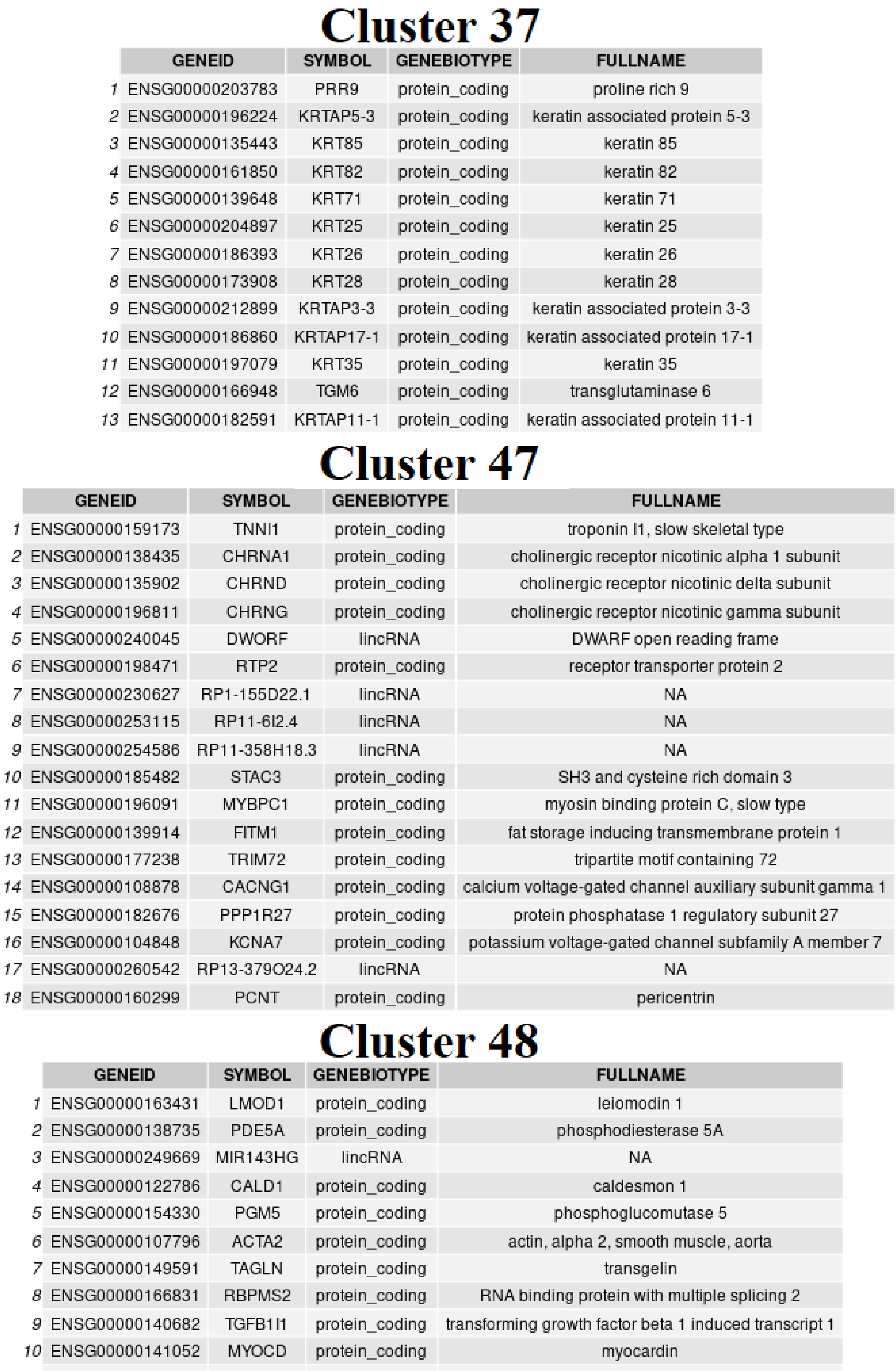
Figure A.2: Gene annotation of clusters 25, 29 and 32 from tissue analysis.



Cluster 33

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000168509	HFE2	protein_coding	hemochromatosis type 2 (juvenile)
2	ENSG00000163157	TMOD4	protein_coding	tropomodulin 4
3	ENSG00000122180	MYOG	protein_coding	myogenin
4	ENSG00000143632	ACTA1	protein_coding	actin, alpha 1, skeletal muscle
5	ENSG00000138100	TRIM54	protein_coding	tripartite motif containing 54
6	ENSG00000204460	LINC01854	lincRNA	long intergenic non-protein coding RNA 1854
7	ENSG00000183091	NEB	protein_coding	nebulin
8	ENSG00000163092	XIRP2	protein_coding	xin actin binding repeat containing 2
9	ENSG00000239474	KLHL41	protein_coding	kelch like family member 41
10	ENSG00000155657	TTN	protein_coding	titin
11	ENSG00000152430	BOLL	protein_coding	boule homolog, RNA binding protein
12	ENSG00000168530	MYL1	protein_coding	myosin light chain 1
13	ENSG00000157119	KLHL40	protein_coding	kelch like family member 40
14	ENSG00000205678	TECRL	protein_coding	trans-2,3-enoyl-CoA reductase like
15	ENSG00000248713	LOC285556	protein_coding	uncharacterized LOC285556
16	ENSG00000185028	LRRC14B	protein_coding	leucine rich repeat containing 14B
17	ENSG00000124701	APOBEC2	protein_coding	apolipoprotein B mRNA editing enzyme catalytic subunit 2
18	ENSG00000225613	LINCMD1	lincRNA	NA
19	ENSG00000164440	TXLNB	protein_coding	taxilin beta
20	ENSG00000154415	PPP1R3A	protein_coding	protein phosphatase 1 regulatory subunit 3A
21	ENSG00000146809	ASB15	protein_coding	ankyrin repeat and SOCS box containing 15
22	ENSG00000170807	LMOD2	protein_coding	leiomodin 2
23	ENSG00000146926	ASB10	protein_coding	ankyrin repeat and SOCS box containing 10
24	ENSG00000148377	IDI2	protein_coding	isopentenyl-diphosphate delta isomerase 2
25	ENSG00000177354	C10orf71	protein_coding	chromosome 10 open reading frame 71
26	ENSG00000138347	MYPN	protein_coding	myopalladin
27	ENSG00000188716	DUPD1	protein_coding	dual specificity phosphatase and pro isomerase domain containing 1
28	ENSG00000197893	NRAP	protein_coding	nebulin related anchoring protein
29	ENSG00000129152	MYOD1	protein_coding	myogenic differentiation 1
30	ENSG00000129170	CSRP3	protein_coding	cysteine and glycine rich protein 3
31	ENSG00000255426	CTD-2210P24.2	lincRNA	NA
32	ENSG00000111241	FGF6	protein_coding	fibroblast growth factor 6
33	ENSG00000111046	MYF6	protein_coding	myogenic factor 6
34	ENSG00000111049	MYF5	protein_coding	myogenic factor 5
35	ENSG00000111245	MYL2	protein_coding	myosin light chain 2
36	ENSG00000185847	LINC01405	lincRNA	long intergenic non-protein coding RNA 1405
37	ENSG00000197616	MYH6	protein_coding	myosin heavy chain 6
38	ENSG00000092054	MYH7	protein_coding	myosin heavy chain 7
39	ENSG00000140986	RPL3L	protein_coding	ribosomal protein L3 like
40	ENSG00000196296	ATP2A1	protein_coding	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1
41	ENSG00000180209	MYLPF	protein_coding	myosin light chain, phosphorylatable, fast skeletal muscle
42	ENSG00000156885	COX6A2	protein_coding	cytochrome c oxidase subunit 6A2
43	ENSG00000108515	ENO3	protein_coding	enolase 3
44	ENSG00000184544	DHRS7C	protein_coding	dehydrogenase/reductase 7C
45	ENSG00000133020	MYH8	protein_coding	myosin heavy chain 8
46	ENSG00000264424	MYH4	protein_coding	myosin heavy chain 4
47	ENSG00000109061	MYH1	protein_coding	myosin heavy chain 1
48	ENSG00000125414	MYH2	protein_coding	myosin heavy chain 2
49	ENSG00000173991	TCAP	protein_coding	titin-cap
50	ENSG00000206422	LRRC30	protein_coding	leucine rich repeat containing 30
51	ENSG00000267391	RP11-1151B14.3	lincRNA	NA
52	ENSG00000267423	AC005616.2	lincRNA	NA
53	ENSG00000104879	CKM	protein_coding	creatine kinase, M-type
54	ENSG00000086967	MYBPC2	protein_coding	myosin binding protein C, fast type
55	ENSG00000101470	TNNC2	protein_coding	troponin C2, fast skeletal type
56	ENSG00000198125	MB	protein_coding	myoglobin
57	ENSG00000101892	ATP1B4	protein_coding	ATPase Na+/K+ transporting family member beta 4

Figure A.3: Gene annotation of cluster 33 from tissue analysis.



Cluster 37

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000203783	PRR9	protein_coding	proline rich 9
2	ENSG00000196224	KRTAP5-3	protein_coding	keratin associated protein 5-3
3	ENSG00000135443	KRT85	protein_coding	keratin 85
4	ENSG00000161850	KRT82	protein_coding	keratin 82
5	ENSG00000139648	KRT71	protein_coding	keratin 71
6	ENSG00000204897	KRT25	protein_coding	keratin 25
7	ENSG00000186393	KRT26	protein_coding	keratin 26
8	ENSG00000173908	KRT28	protein_coding	keratin 28
9	ENSG00000212899	KRTAP3-3	protein_coding	keratin associated protein 3-3
10	ENSG00000186860	KRTAP17-1	protein_coding	keratin associated protein 17-1
11	ENSG00000197079	KRT35	protein_coding	keratin 35
12	ENSG00000166948	TGM6	protein_coding	transglutaminase 6
13	ENSG00000182591	KRTAP11-1	protein_coding	keratin associated protein 11-1

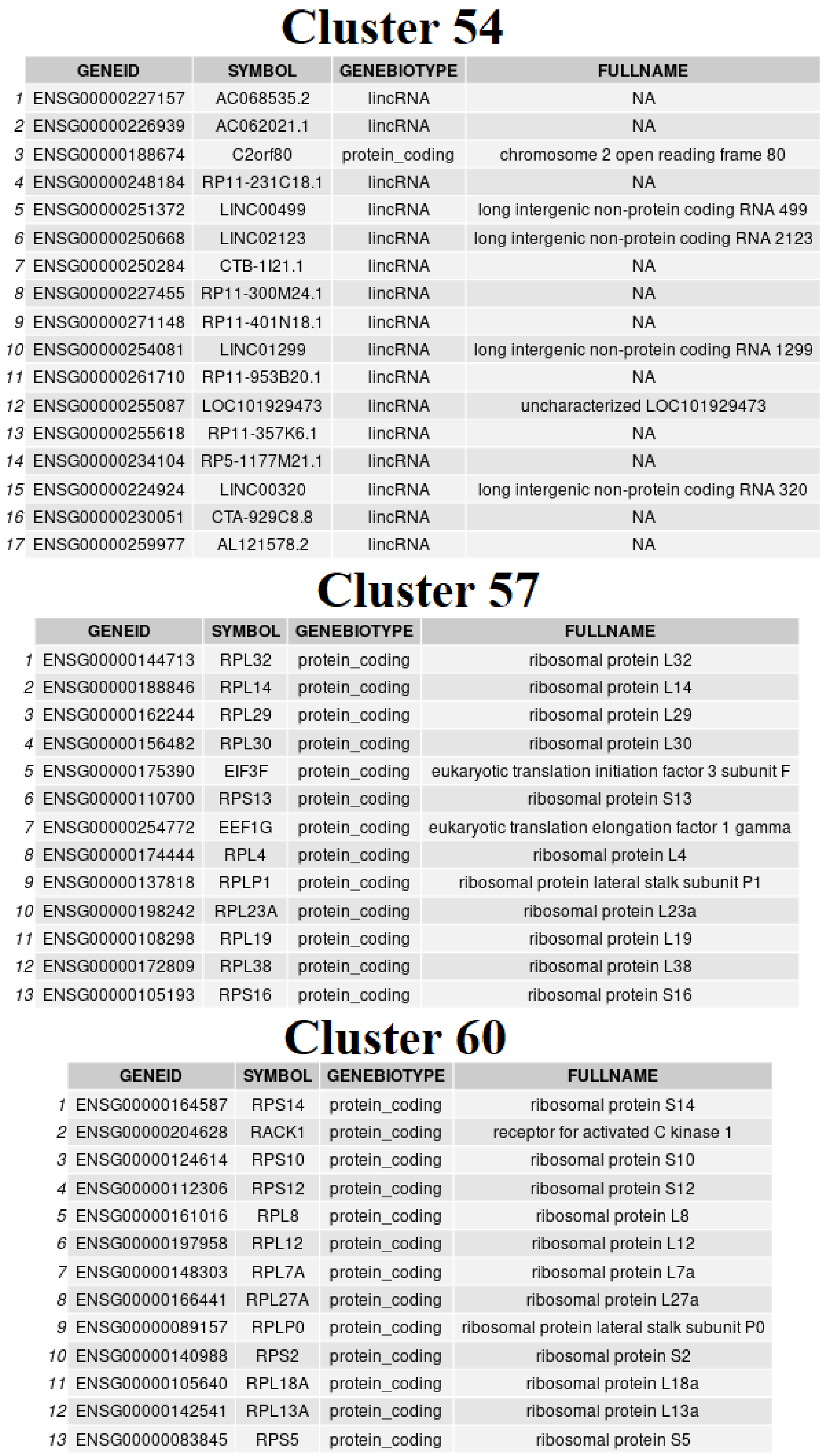
Cluster 47

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000159173	TNNI1	protein_coding	troponin I1, slow skeletal type
2	ENSG00000138435	CHRNA1	protein_coding	cholinergic receptor nicotinic alpha 1 subunit
3	ENSG00000135902	CHRND	protein_coding	cholinergic receptor nicotinic delta subunit
4	ENSG00000196811	CHRNG	protein_coding	cholinergic receptor nicotinic gamma subunit
5	ENSG00000240045	DWORF	lincRNA	DWARF open reading frame
6	ENSG00000198471	RTP2	protein_coding	receptor transporter protein 2
7	ENSG00000230627	RP1-155D22.1	lincRNA	NA
8	ENSG00000253115	RP11-6I2.4	lincRNA	NA
9	ENSG00000254586	RP11-358H18.3	lincRNA	NA
10	ENSG00000185482	STAC3	protein_coding	SH3 and cysteine rich domain 3
11	ENSG00000196091	MYBPC1	protein_coding	myosin binding protein C, slow type
12	ENSG00000139914	FITM1	protein_coding	fat storage inducing transmembrane protein 1
13	ENSG00000177238	TRIM72	protein_coding	tripartite motif containing 72
14	ENSG00000108878	CACNG1	protein_coding	calcium voltage-gated channel auxiliary subunit gamma 1
15	ENSG00000182676	PPP1R27	protein_coding	protein phosphatase 1 regulatory subunit 27
16	ENSG00000104848	KCNA7	protein_coding	potassium voltage-gated channel subfamily A member 7
17	ENSG00000260542	RP13-379O24.2	lincRNA	NA
18	ENSG00000160299	PCNT	protein_coding	pericentrin

Cluster 48

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000163431	LMOD1	protein_coding	leiomodin 1
2	ENSG00000138735	PDE5A	protein_coding	phosphodiesterase 5A
3	ENSG00000249669	MIR143HG	lincRNA	NA
4	ENSG00000122786	CALD1	protein_coding	caldesmon 1
5	ENSG00000154330	PGM5	protein_coding	phosphoglucomutase 5
6	ENSG00000107796	ACTA2	protein_coding	actin, alpha 2, smooth muscle, aorta
7	ENSG00000149591	TAGLN	protein_coding	transgelin
8	ENSG00000166831	RBPMS2	protein_coding	RNA binding protein with multiple splicing 2
9	ENSG00000140682	TGFB1I1	protein_coding	transforming growth factor beta 1 induced transcript 1
10	ENSG00000141052	MYOCD	protein_coding	myocardin

Figure A.4: Gene annotation of clusters 37, 47 and 48 from tissue analysis.



Cluster 54

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000227157	AC068535.2	lincRNA	NA
2	ENSG00000226939	AC062021.1	lincRNA	NA
3	ENSG00000188674	C2orf80	protein_coding	chromosome 2 open reading frame 80
4	ENSG00000248184	RP11-231C18.1	lincRNA	NA
5	ENSG00000251372	LINC00499	lincRNA	long intergenic non-protein coding RNA 499
6	ENSG00000250668	LINC02123	lincRNA	long intergenic non-protein coding RNA 2123
7	ENSG00000250284	CTB-1I21.1	lincRNA	NA
8	ENSG00000227455	RP11-300M24.1	lincRNA	NA
9	ENSG00000271148	RP11-401N18.1	lincRNA	NA
10	ENSG00000254081	LINC01299	lincRNA	long intergenic non-protein coding RNA 1299
11	ENSG00000261710	RP11-953B20.1	lincRNA	NA
12	ENSG00000255087	LOC101929473	lincRNA	uncharacterized LOC101929473
13	ENSG00000255618	RP11-357K6.1	lincRNA	NA
14	ENSG00000234104	RP5-1177M21.1	lincRNA	NA
15	ENSG00000224924	LINC00320	lincRNA	long intergenic non-protein coding RNA 320
16	ENSG00000230051	CTA-929C8.8	lincRNA	NA
17	ENSG00000259977	AL121578.2	lincRNA	NA

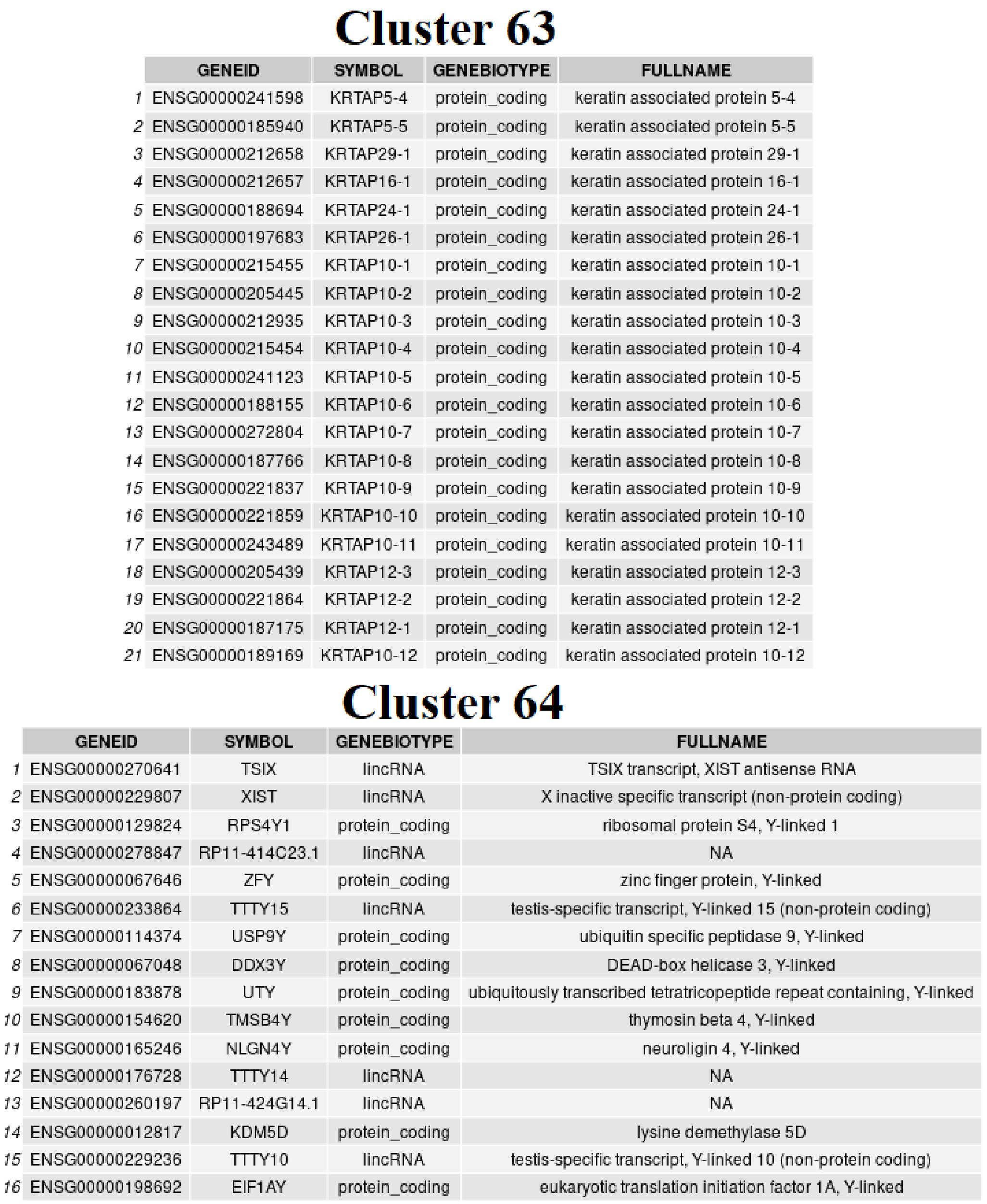
Cluster 57

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000144713	RPL32	protein_coding	ribosomal protein L32
2	ENSG00000188846	RPL14	protein_coding	ribosomal protein L14
3	ENSG00000162244	RPL29	protein_coding	ribosomal protein L29
4	ENSG00000156482	RPL30	protein_coding	ribosomal protein L30
5	ENSG00000175390	EIF3F	protein_coding	eukaryotic translation initiation factor 3 subunit F
6	ENSG00000110700	RPS13	protein_coding	ribosomal protein S13
7	ENSG00000254772	EEF1G	protein_coding	eukaryotic translation elongation factor 1 gamma
8	ENSG00000174444	RPL4	protein_coding	ribosomal protein L4
9	ENSG00000137818	RPLP1	protein_coding	ribosomal protein lateral stalk subunit P1
10	ENSG00000198242	RPL23A	protein_coding	ribosomal protein L23a
11	ENSG00000108298	RPL19	protein_coding	ribosomal protein L19
12	ENSG00000172809	RPL38	protein_coding	ribosomal protein L38
13	ENSG00000105193	RPS16	protein_coding	ribosomal protein S16

Cluster 60

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000164587	RPS14	protein_coding	ribosomal protein S14
2	ENSG00000204628	RACK1	protein_coding	receptor for activated C kinase 1
3	ENSG00000124614	RPS10	protein_coding	ribosomal protein S10
4	ENSG00000112306	RPS12	protein_coding	ribosomal protein S12
5	ENSG00000161016	RPL8	protein_coding	ribosomal protein L8
6	ENSG00000197958	RPL12	protein_coding	ribosomal protein L12
7	ENSG00000148303	RPL7A	protein_coding	ribosomal protein L7a
8	ENSG00000166441	RPL27A	protein_coding	ribosomal protein L27a
9	ENSG00000089157	RPLP0	protein_coding	ribosomal protein lateral stalk subunit P0
10	ENSG00000140988	RPS2	protein_coding	ribosomal protein S2
11	ENSG00000105640	RPL18A	protein_coding	ribosomal protein L18a
12	ENSG00000142541	RPL13A	protein_coding	ribosomal protein L13a
13	ENSG00000083845	RPS5	protein_coding	ribosomal protein S5

Figure A.5: Gene annotation of clusters 54, 57 and 60 from tissue analysis.



Cluster 63

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000241598	KRTAP5-4	protein_coding	keratin associated protein 5-4
2	ENSG00000185940	KRTAP5-5	protein_coding	keratin associated protein 5-5
3	ENSG00000212658	KRTAP29-1	protein_coding	keratin associated protein 29-1
4	ENSG00000212657	KRTAP16-1	protein_coding	keratin associated protein 16-1
5	ENSG00000188694	KRTAP24-1	protein_coding	keratin associated protein 24-1
6	ENSG00000197683	KRTAP26-1	protein_coding	keratin associated protein 26-1
7	ENSG00000215455	KRTAP10-1	protein_coding	keratin associated protein 10-1
8	ENSG00000205445	KRTAP10-2	protein_coding	keratin associated protein 10-2
9	ENSG00000212935	KRTAP10-3	protein_coding	keratin associated protein 10-3
10	ENSG00000215454	KRTAP10-4	protein_coding	keratin associated protein 10-4
11	ENSG00000241123	KRTAP10-5	protein_coding	keratin associated protein 10-5
12	ENSG00000188155	KRTAP10-6	protein_coding	keratin associated protein 10-6
13	ENSG00000272804	KRTAP10-7	protein_coding	keratin associated protein 10-7
14	ENSG00000187766	KRTAP10-8	protein_coding	keratin associated protein 10-8
15	ENSG00000221837	KRTAP10-9	protein_coding	keratin associated protein 10-9
16	ENSG00000221859	KRTAP10-10	protein_coding	keratin associated protein 10-10
17	ENSG00000243489	KRTAP10-11	protein_coding	keratin associated protein 10-11
18	ENSG00000205439	KRTAP12-3	protein_coding	keratin associated protein 12-3
19	ENSG00000221864	KRTAP12-2	protein_coding	keratin associated protein 12-2
20	ENSG00000187175	KRTAP12-1	protein_coding	keratin associated protein 12-1
21	ENSG00000189169	KRTAP10-12	protein_coding	keratin associated protein 10-12

Cluster 64

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000270641	TSIX	lincRNA	TSIX transcript, XIST antisense RNA
2	ENSG00000229807	XIST	lincRNA	X inactive specific transcript (non-protein coding)
3	ENSG00000129824	RPS4Y1	protein_coding	ribosomal protein S4, Y-linked 1
4	ENSG00000278847	RP11-414C23.1	lincRNA	NA
5	ENSG00000067646	ZFY	protein_coding	zinc finger protein, Y-linked
6	ENSG00000233864	TTTY15	lincRNA	testis-specific transcript, Y-linked 15 (non-protein coding)
7	ENSG00000114374	USP9Y	protein_coding	ubiquitin specific peptidase 9, Y-linked
8	ENSG00000067048	DDX3Y	protein_coding	DEAD-box helicase 3, Y-linked
9	ENSG00000183878	UTY	protein_coding	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked
10	ENSG00000154620	TMSB4Y	protein_coding	thymosin beta 4, Y-linked
11	ENSG00000165246	NLGN4Y	protein_coding	neuroligin 4, Y-linked
12	ENSG00000176728	TTTY14	lincRNA	NA
13	ENSG00000260197	RP11-424G14.1	lincRNA	NA
14	ENSG00000012817	KDM5D	protein_coding	lysine demethylase 5D
15	ENSG00000229236	TTTY10	lincRNA	testis-specific transcript, Y-linked 10 (non-protein coding)
16	ENSG00000198692	EIF1AY	protein_coding	eukaryotic translation initiation factor 1A, Y-linked

Figure A.6: Gene annotation of clusters 63 and 64 from tissue analysis.

A.2 Tissue Analysis: Clusters GO Enrichment Analysis

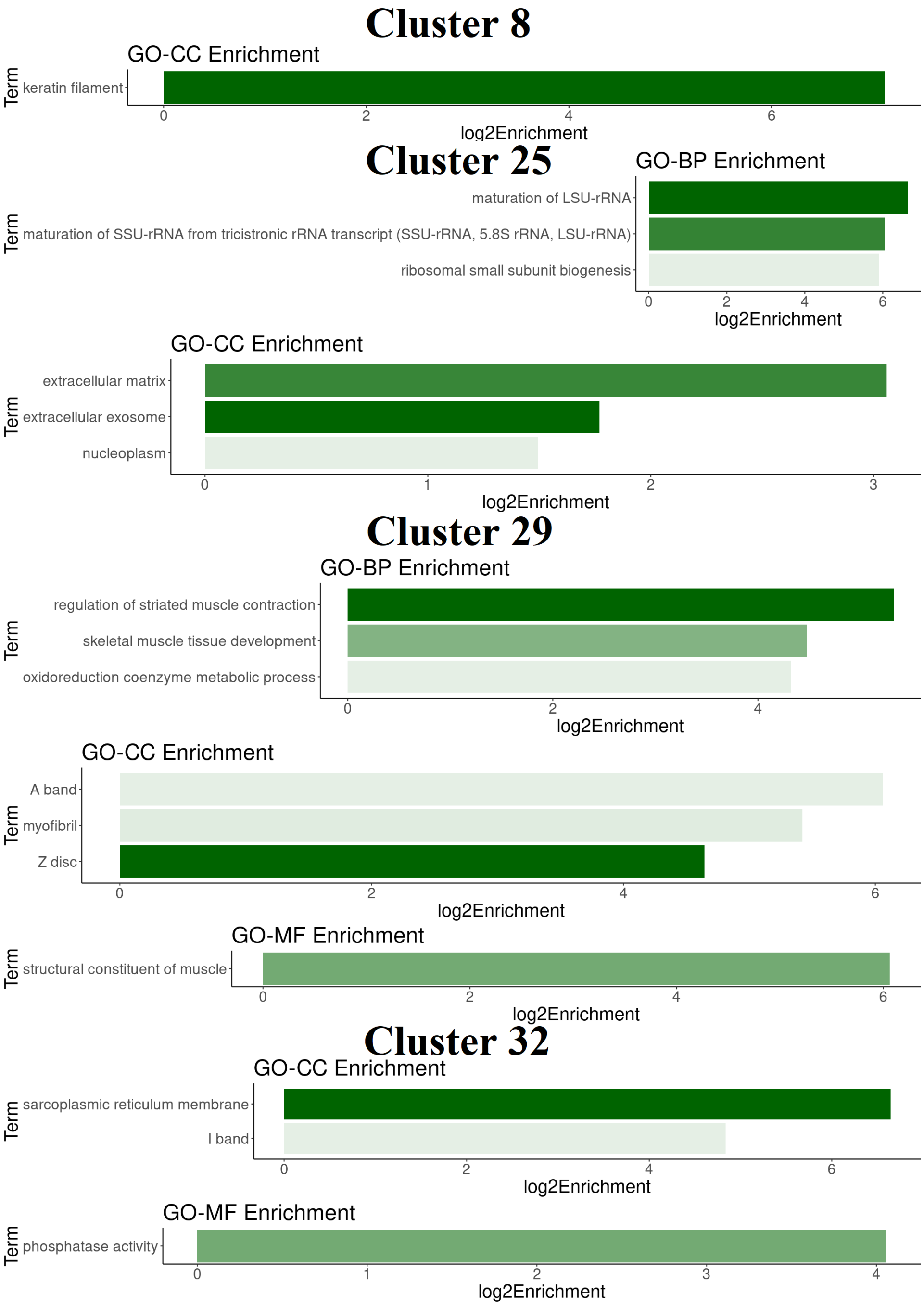


Figure A.7: GO enrichment analysis of clusters 8, 25, 29 and 32 from tissue analysis. GO-BP means Biological Processes GO; GO-CC means Cellular Component GO; GO-MF means Molecular Function GO.

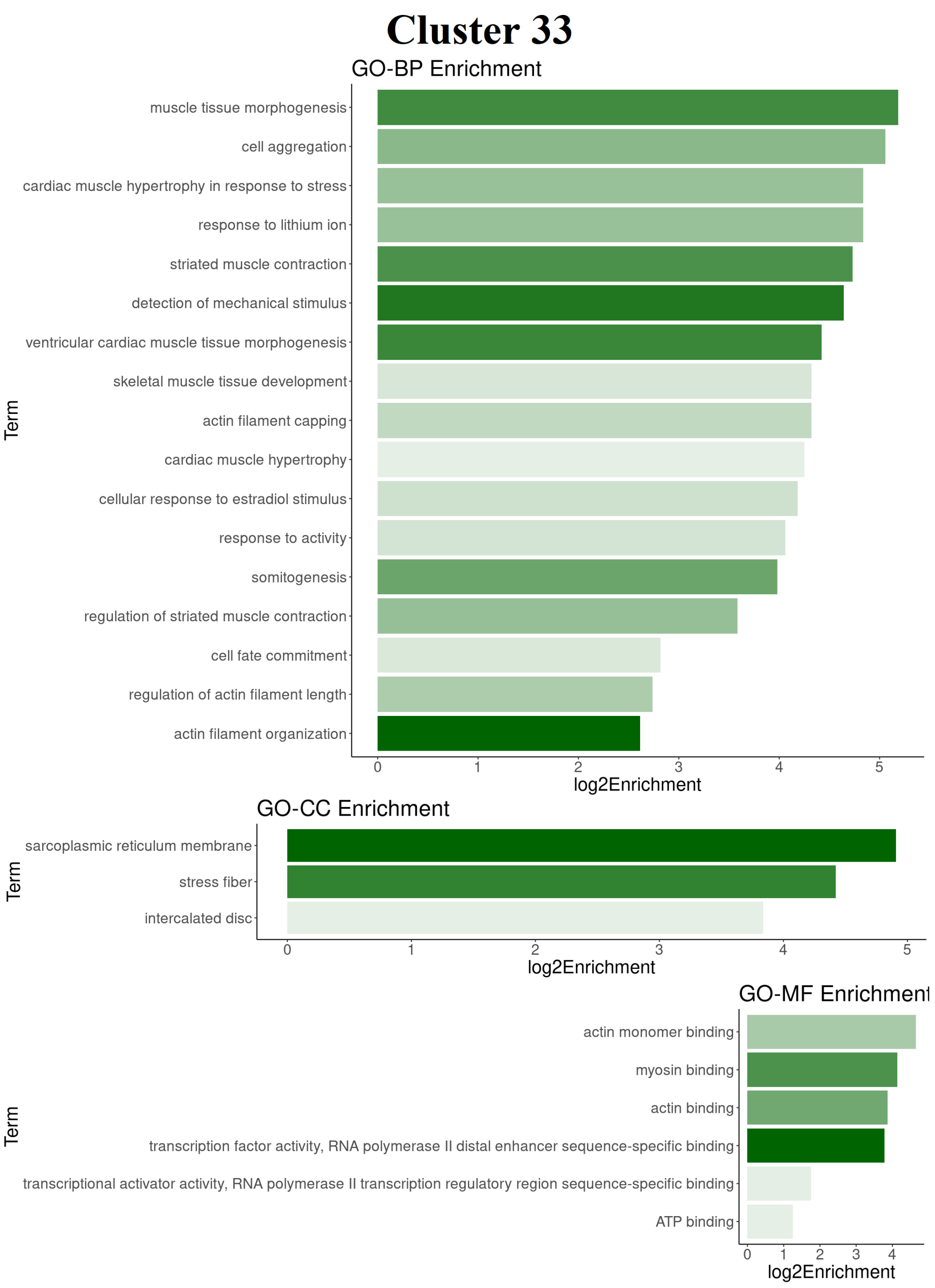


Figure A.8: GO enrichment analysis of clusters 33 from tissue analysis. GO-BP means Biological Processes GO; GO-CC means Cellular Component GO; GO-MF means Molecular Function GO.

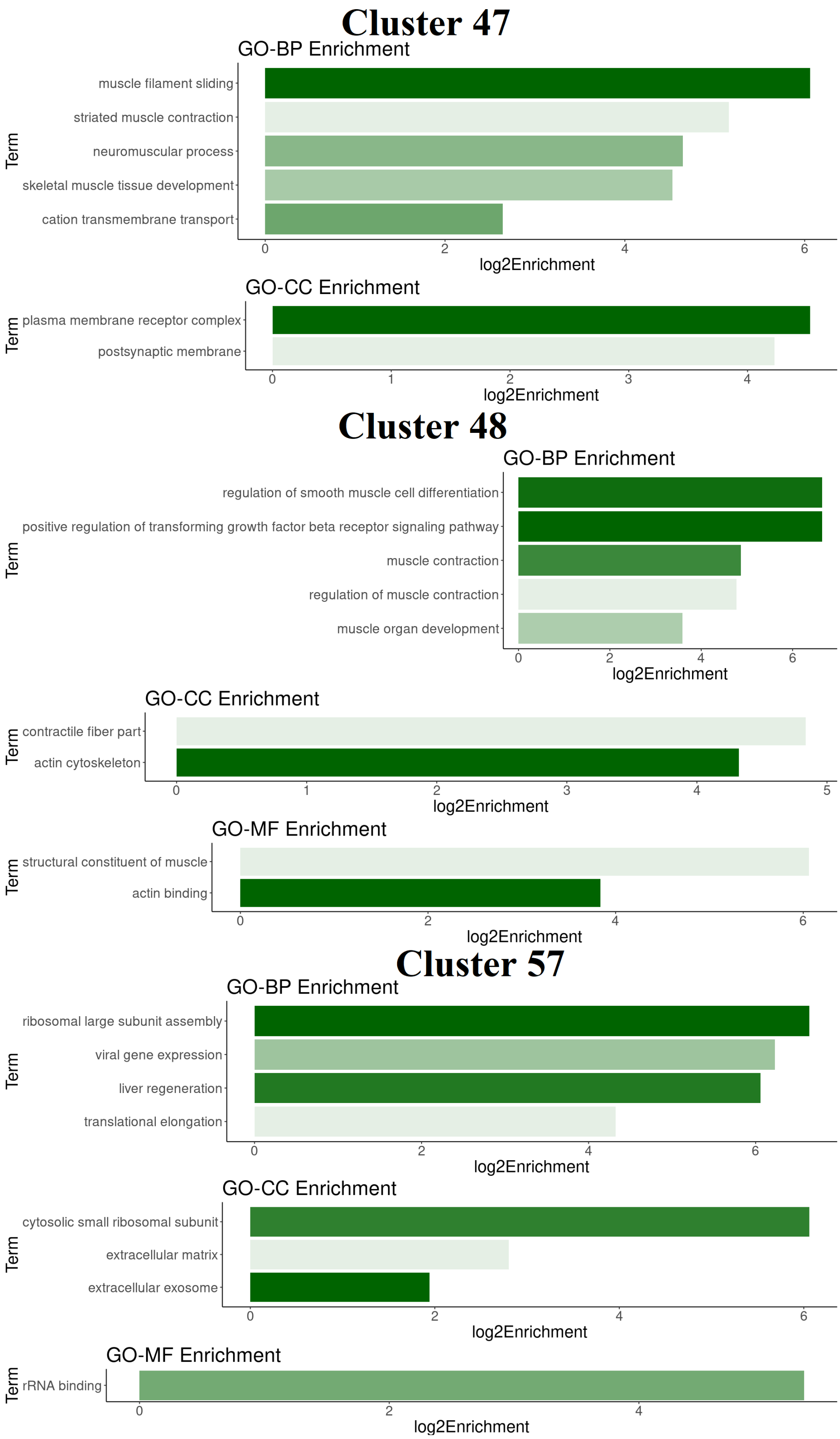


Figure A.9: GO enrichment analysis of clusters 47, 48 and 57 from tissue analysis. GO-BP means Biological Processes GO; GO-CC means Cellular Component GO; GO-MF means Molecular Function GO.

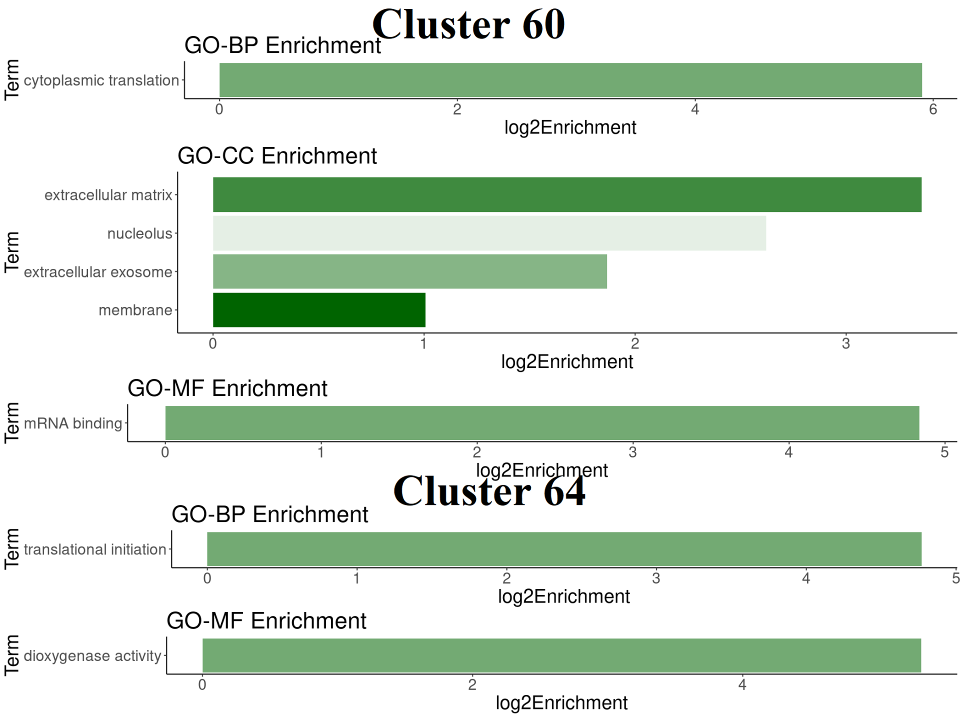


Figure A.10: GO enrichment analysis of clusters 60 and 64 from tissue analysis. GO-BP means Biological Processes GO; GO-CC means Cellular Component GO; GO-MF means Molecular Function GO.

A.3 Tissue Analysis: Clusters Gene Mean Expression and Variance Boxplots

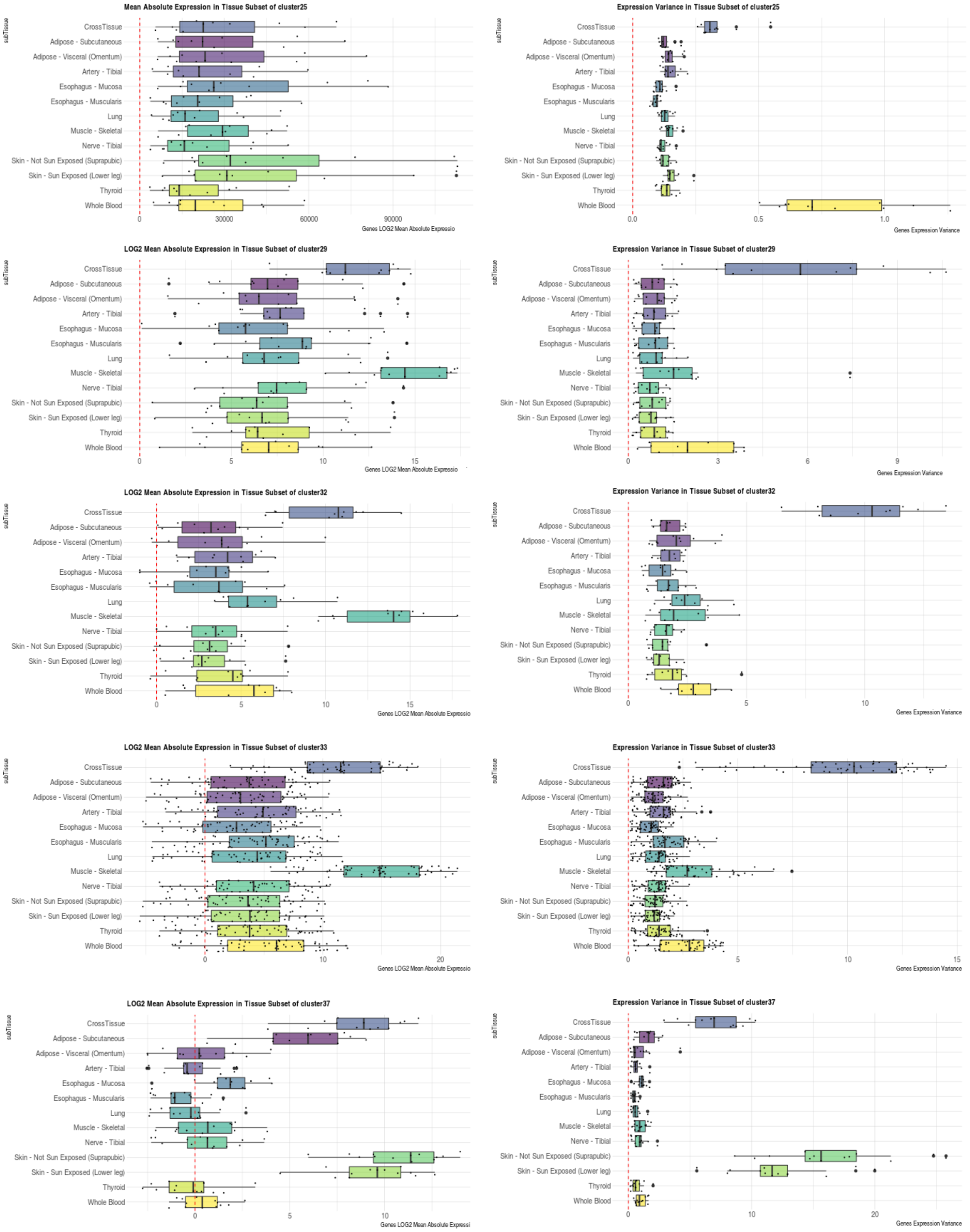


Figure A.11: Gene mean expression and variance boxplots of clusters 25, 29, 32, 33 and 37 from tissue analysis.

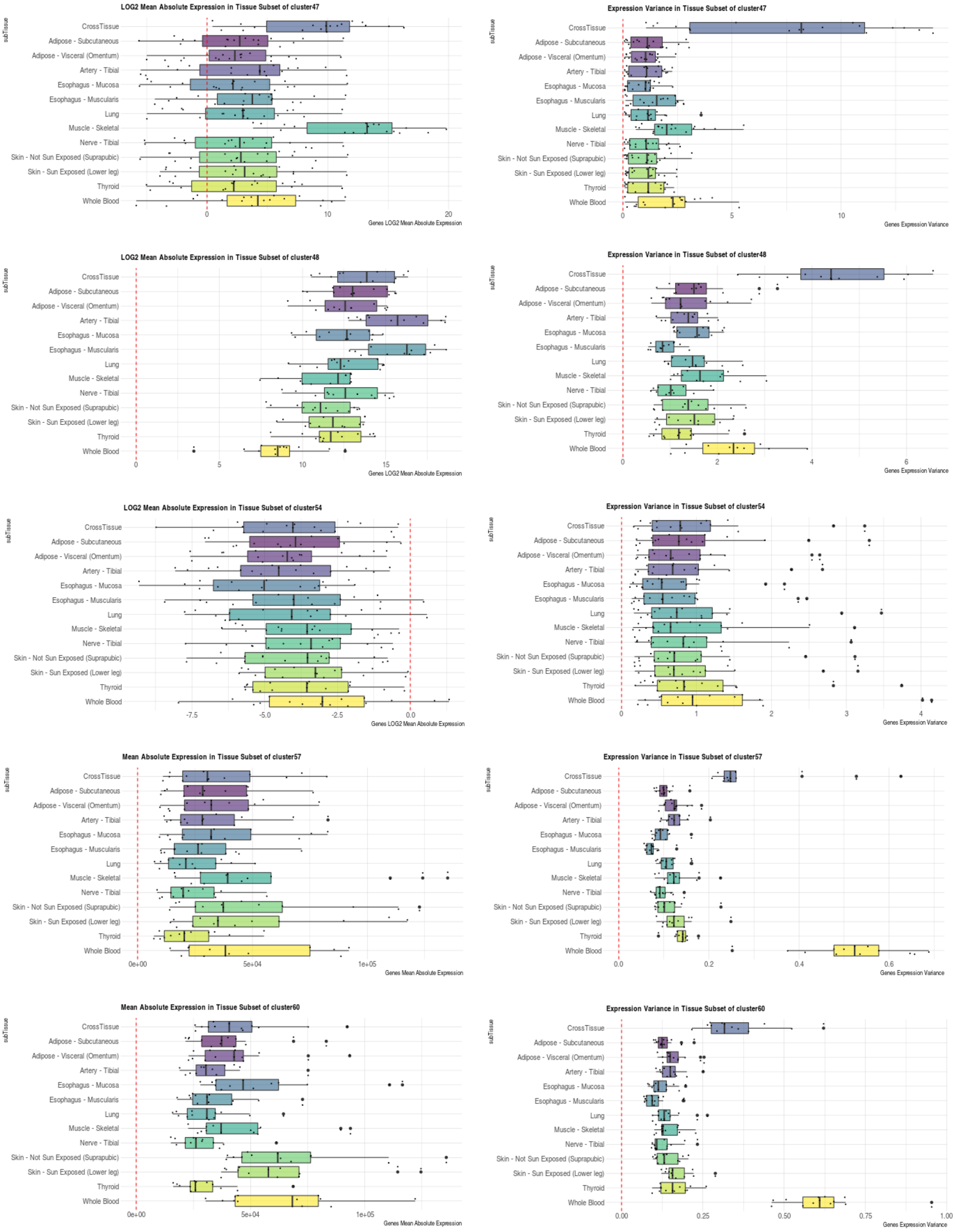


Figure A.12: Gene mean expression and variance boxplots of clusters 47, 48, 54, 57 and 60 from tissue analysis.

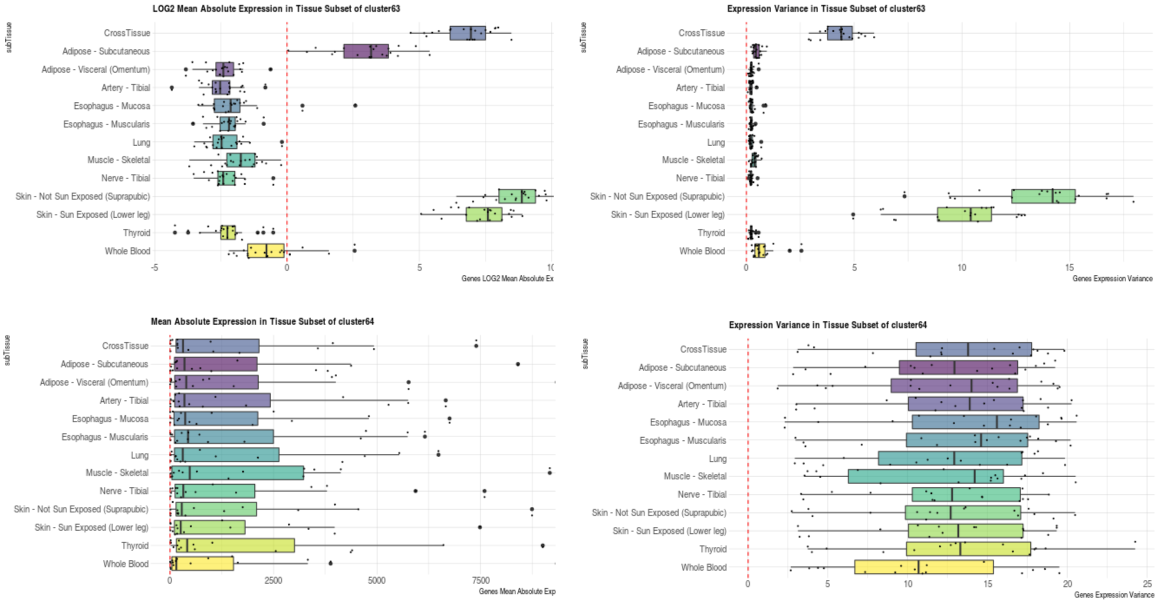
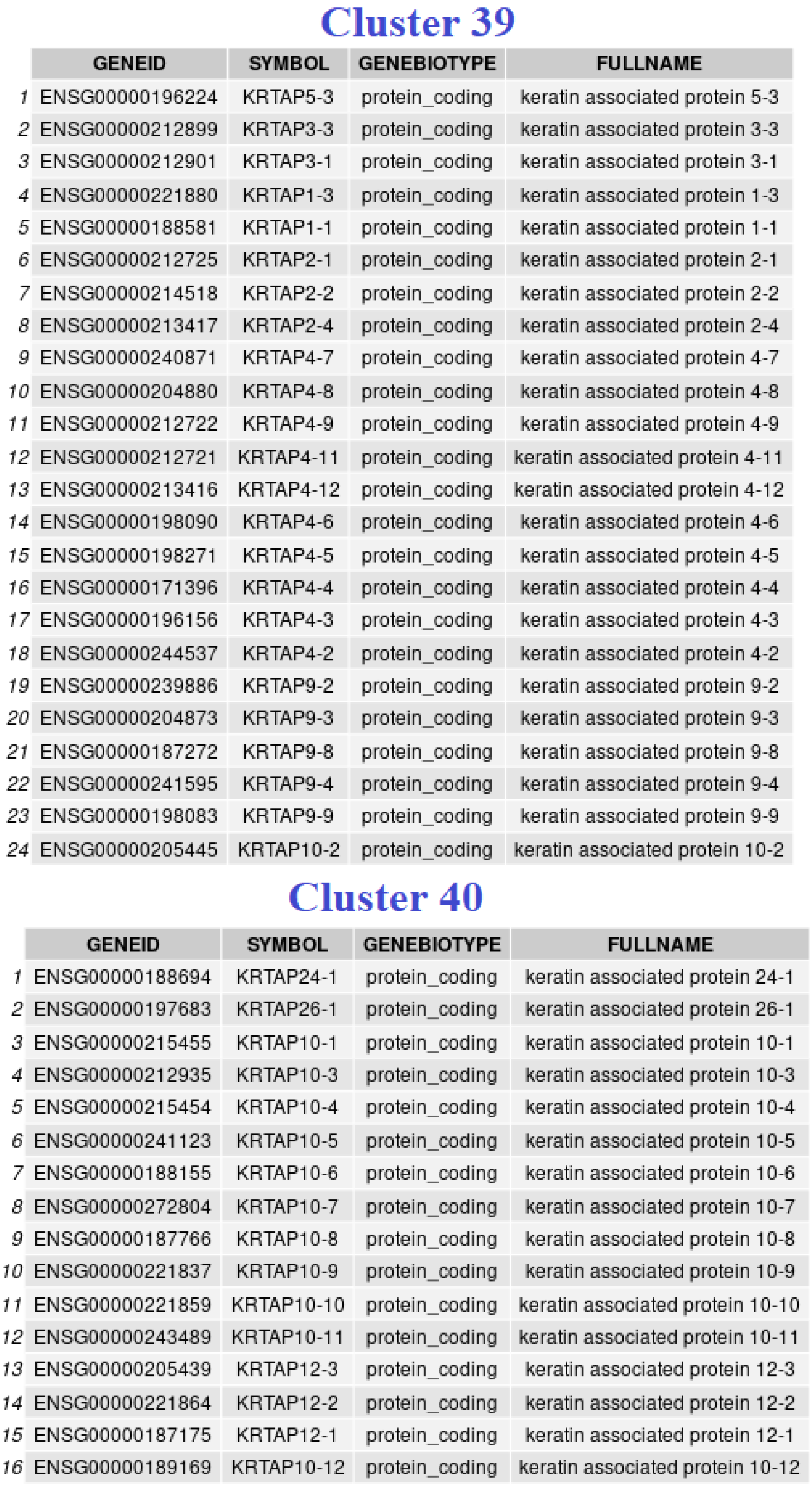


Figure A.13: Gene mean expression and variance boxplots of clusters 63 and 64 from tissue analysis.

A.4 Age Analysis: Clusters Annotation



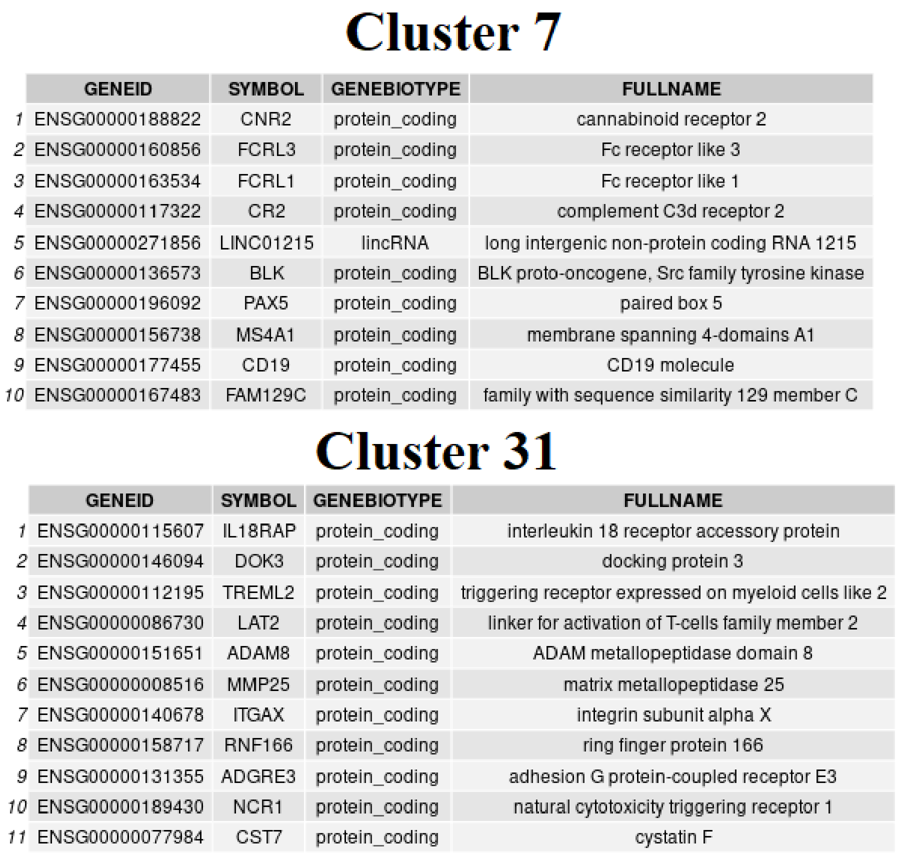
Cluster 39

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000196224	KRTAP5-3	protein_coding	keratin associated protein 5-3
2	ENSG00000212899	KRTAP3-3	protein_coding	keratin associated protein 3-3
3	ENSG00000212901	KRTAP3-1	protein_coding	keratin associated protein 3-1
4	ENSG00000221880	KRTAP1-3	protein_coding	keratin associated protein 1-3
5	ENSG00000188581	KRTAP1-1	protein_coding	keratin associated protein 1-1
6	ENSG00000212725	KRTAP2-1	protein_coding	keratin associated protein 2-1
7	ENSG00000214518	KRTAP2-2	protein_coding	keratin associated protein 2-2
8	ENSG00000213417	KRTAP2-4	protein_coding	keratin associated protein 2-4
9	ENSG00000240871	KRTAP4-7	protein_coding	keratin associated protein 4-7
10	ENSG00000204880	KRTAP4-8	protein_coding	keratin associated protein 4-8
11	ENSG00000212722	KRTAP4-9	protein_coding	keratin associated protein 4-9
12	ENSG00000212721	KRTAP4-11	protein_coding	keratin associated protein 4-11
13	ENSG00000213416	KRTAP4-12	protein_coding	keratin associated protein 4-12
14	ENSG00000198090	KRTAP4-6	protein_coding	keratin associated protein 4-6
15	ENSG00000198271	KRTAP4-5	protein_coding	keratin associated protein 4-5
16	ENSG00000171396	KRTAP4-4	protein_coding	keratin associated protein 4-4
17	ENSG00000196156	KRTAP4-3	protein_coding	keratin associated protein 4-3
18	ENSG00000244537	KRTAP4-2	protein_coding	keratin associated protein 4-2
19	ENSG00000239886	KRTAP9-2	protein_coding	keratin associated protein 9-2
20	ENSG00000204873	KRTAP9-3	protein_coding	keratin associated protein 9-3
21	ENSG00000187272	KRTAP9-8	protein_coding	keratin associated protein 9-8
22	ENSG00000241595	KRTAP9-4	protein_coding	keratin associated protein 9-4
23	ENSG00000198083	KRTAP9-9	protein_coding	keratin associated protein 9-9
24	ENSG00000205445	KRTAP10-2	protein_coding	keratin associated protein 10-2

Cluster 40

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000188694	KRTAP24-1	protein_coding	keratin associated protein 24-1
2	ENSG00000197683	KRTAP26-1	protein_coding	keratin associated protein 26-1
3	ENSG00000215455	KRTAP10-1	protein_coding	keratin associated protein 10-1
4	ENSG00000212935	KRTAP10-3	protein_coding	keratin associated protein 10-3
5	ENSG00000215454	KRTAP10-4	protein_coding	keratin associated protein 10-4
6	ENSG00000241123	KRTAP10-5	protein_coding	keratin associated protein 10-5
7	ENSG00000188155	KRTAP10-6	protein_coding	keratin associated protein 10-6
8	ENSG00000272804	KRTAP10-7	protein_coding	keratin associated protein 10-7
9	ENSG00000187766	KRTAP10-8	protein_coding	keratin associated protein 10-8
10	ENSG00000221837	KRTAP10-9	protein_coding	keratin associated protein 10-9
11	ENSG00000221859	KRTAP10-10	protein_coding	keratin associated protein 10-10
12	ENSG00000243489	KRTAP10-11	protein_coding	keratin associated protein 10-11
13	ENSG00000205439	KRTAP12-3	protein_coding	keratin associated protein 12-3
14	ENSG00000221864	KRTAP12-2	protein_coding	keratin associated protein 12-2
15	ENSG00000187175	KRTAP12-1	protein_coding	keratin associated protein 12-1
16	ENSG00000189169	KRTAP10-12	protein_coding	keratin associated protein 10-12

Figure A.14: Gene annotation of clusters 39 and 40 from age analysis.



Cluster 7

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000188822	CNR2	protein_coding	cannabinoid receptor 2
2	ENSG00000160856	FCRL3	protein_coding	Fc receptor like 3
3	ENSG00000163534	FCRL1	protein_coding	Fc receptor like 1
4	ENSG00000117322	CR2	protein_coding	complement C3d receptor 2
5	ENSG00000271856	LINC01215	lincRNA	long intergenic non-protein coding RNA 1215
6	ENSG00000136573	BLK	protein_coding	BLK proto-oncogene, Src family tyrosine kinase
7	ENSG00000196092	PAX5	protein_coding	paired box 5
8	ENSG00000156738	MS4A1	protein_coding	membrane spanning 4-domains A1
9	ENSG00000177455	CD19	protein_coding	CD19 molecule
10	ENSG00000167483	FAM129C	protein_coding	family with sequence similarity 129 member C

Cluster 31

	GENEID	SYMBOL	GENEBIOTYPE	FULLNAME
1	ENSG00000115607	IL18RAP	protein_coding	interleukin 18 receptor accessory protein
2	ENSG00000146094	DOK3	protein_coding	docking protein 3
3	ENSG00000112195	TREML2	protein_coding	triggering receptor expressed on myeloid cells like 2
4	ENSG00000086730	LAT2	protein_coding	linker for activation of T-cells family member 2
5	ENSG00000151651	ADAM8	protein_coding	ADAM metallopeptidase domain 8
6	ENSG00000008516	MMP25	protein_coding	matrix metallopeptidase 25
7	ENSG00000140678	ITGAX	protein_coding	integrin subunit alpha X
8	ENSG00000158717	RNF166	protein_coding	ring finger protein 166
9	ENSG00000131355	ADGRE3	protein_coding	adhesion G protein-coupled receptor E3
10	ENSG00000189430	NCR1	protein_coding	natural cytotoxicity triggering receptor 1
11	ENSG00000077984	CST7	protein_coding	cystatin F

Figure A.15: Gene annotation of clusters 7 and 31 from age analysis.

A.5 Age Analysis: Cluster GO Enrichment Analysis

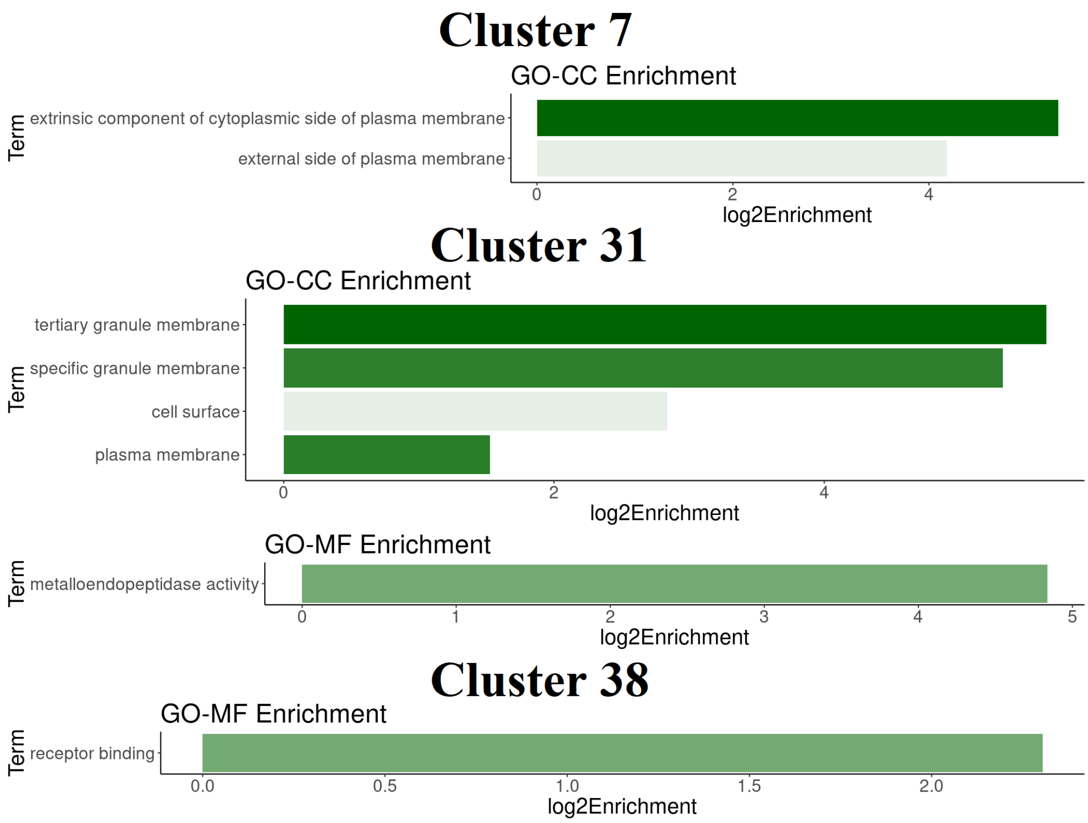
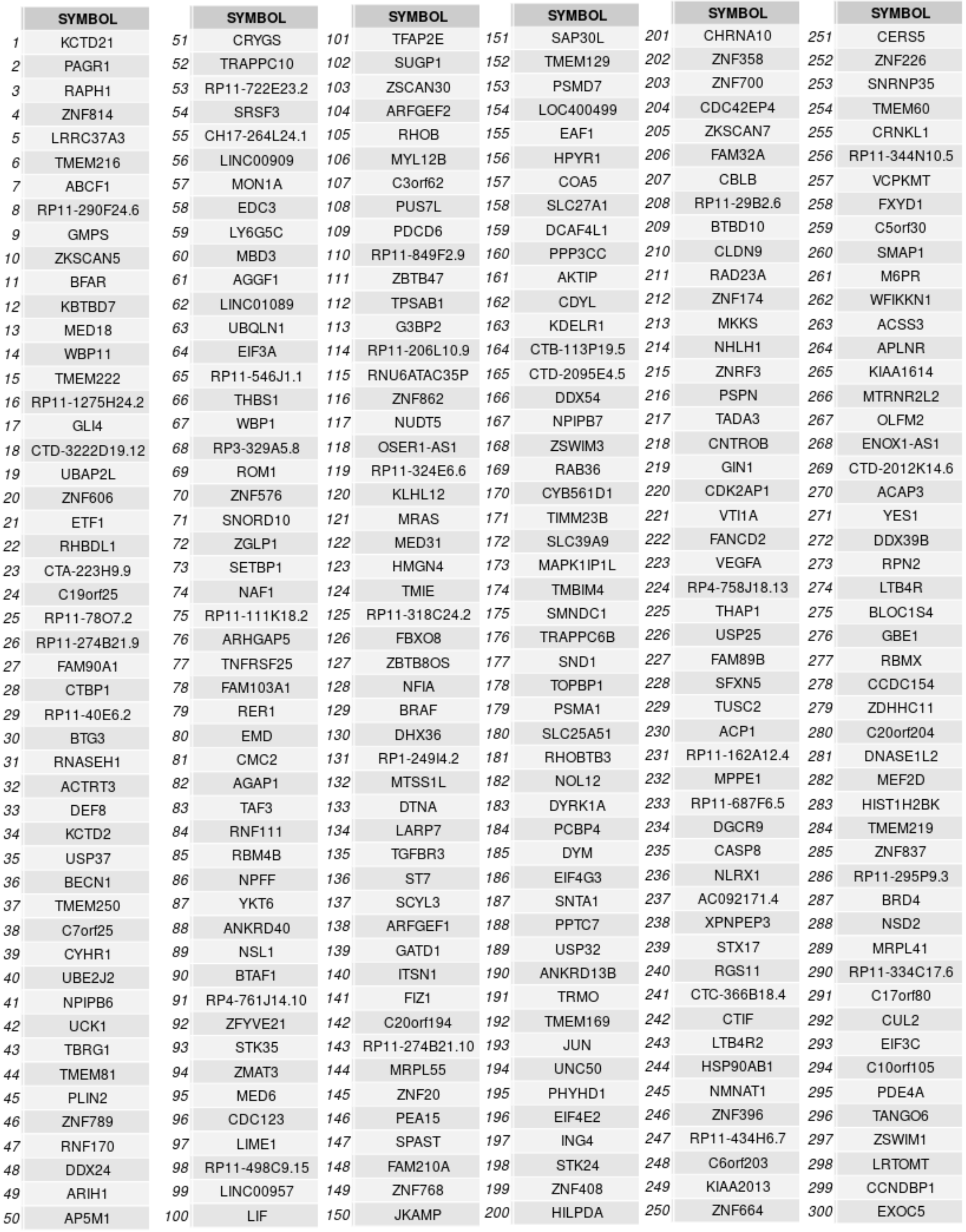


Figure A.16: GO enrichment analysis of clusters 7, 31 and 38 from tissue analysis. GO-BP means Biological Processes GO; GO-CC means Cellular Component GO; GO-MF means Molecular Function GO.

A.6 Age Analysis: 300 ”hub genes of ageing” symbol names



	SYMBOL		SYMBOL		SYMBOL		SYMBOL		SYMBOL		SYMBOL
1	KCTD21	51	CRYGS	101	TFAP2E	151	SAP30L	201	CHRNA10	251	CERS5
2	PAGR1	52	TRAPPC10	102	SUGP1	152	TMEM129	202	ZNF358	252	ZNF226
3	RAPH1	53	RP11-722E23.2	103	ZSCAN30	153	PSMD7	203	ZNF700	253	SNRNP35
4	ZNF814	54	SRSF3	104	ARFGEF2	154	LOC400499	204	CDC42EP4	254	TMEM60
5	LRRC37A3	55	CH17-264L24.1	105	RHOB	155	EAF1	205	ZKSCAN7	255	CRNKL1
6	TMEM216	56	LINC00909	106	MYL12B	156	HPYR1	206	FAM32A	256	RP11-344N10.5
7	ABCF1	57	MON1A	107	C3orf62	157	COA5	207	CBLB	257	VCPKMT
8	RP11-290F24.6	58	EDC3	108	PUS7L	158	SLC27A1	208	RP11-29B2.6	258	FXYD1
9	GMPS	59	LY6G5C	109	PDCD6	159	DCAF4L1	209	BTBD10	259	C5orf30
10	ZKSCAN5	60	MBD3	110	RP11-849F2.9	160	PPP3CC	210	CLDN9	260	SMAP1
11	BFAR	61	AGGF1	111	ZBTB47	161	AKTIP	211	RAD23A	261	M6PR
12	KBTBD7	62	LINC01089	112	TPSAB1	162	CDYL	212	ZNF174	262	WFIKKN1
13	MED18	63	UBQLN1	113	G3BP2	163	KDELR1	213	MKKS	263	ACSS3
14	WBP11	64	EIF3A	114	RP11-206L10.9	164	CTB-113P19.5	214	NHLH1	264	APLNR
15	TMEM222	65	RP11-546J1.1	115	RNU6ATAC35P	165	CTD-2095E4.5	215	ZNRF3	265	KIAA1614
16	RP11-1275H24.2	66	THBS1	116	ZNF862	166	DDX54	216	PSPN	266	MTRNR2L2
17	GLI4	67	WBP1	117	NUDT5	167	NPIPB7	217	TADA3	267	OLFM2
18	CTD-3222D19.12	68	RP3-329A5.8	118	OSER1-AS1	168	ZSWIM3	218	CNTROB	268	ENOX1-AS1
19	UBAP2L	69	ROM1	119	RP11-324E6.6	169	RAB36	219	GIN1	269	CTD-2012K14.6
20	ZNF606	70	ZNF576	120	KLHL12	170	CYB561D1	220	CDK2AP1	270	ACAP3
21	ETF1	71	SNORD10	121	MRAS	171	TIMM23B	221	VTI1A	271	YES1
22	RHBDL1	72	ZGLP1	122	MED31	172	SLC39A9	222	FANCD2	272	DDX39B
23	CTA-223H9.9	73	SETBP1	123	HMGN4	173	MAPK1IP1L	223	VEGFA	273	RPN2
24	C19orf25	74	NAF1	124	TMIE	174	TMBIM4	224	RP4-758J18.13	274	LTB4R
25	RP11-78O7.2	75	RP11-111K18.2	125	RP11-318C24.2	175	SMNDC1	225	THAP1	275	BLOC1S4
26	RP11-274B21.9	76	ARHGAP5	126	FBXO8	176	TRAPPC6B	226	USP25	276	GBE1
27	FAM90A1	77	TNFRSF25	127	ZBTB8OS	177	SND1	227	FAM89B	277	RBMX
28	CTBP1	78	FAM103A1	128	NFIA	178	TOPBP1	228	SFXN5	278	CCDC154
29	RP11-40E6.2	79	RER1	129	BRAF	179	PSMA1	229	TUSC2	279	ZDHHC11
30	BTG3	80	EMD	130	DHX36	180	SLC25A51	230	ACP1	280	C20orf204
31	RNASEH1	81	CMC2	131	RP1-249I4.2	181	RHOBTB3	231	RP11-162A12.4	281	DNASE1L2
32	ACTRT3	82	AGAP1	132	MTSS1L	182	NOL12	232	MPPE1	282	MEF2D
33	DEF8	83	TAF3	133	DTNA	183	DYRK1A	233	RP11-687F6.5	283	HIST1H2BK
34	KCTD2	84	RNF111	134	LARP7	184	PCBP4	234	DGCR9	284	TMEM219
35	USP37	85	RBM4B	135	TGFBR3	185	DYM	235	CASP8	285	ZNF837
36	BECN1	86	NPFF	136	ST7	186	EIF4G3	236	NLRX1	286	RP11-295P9.3
37	TMEM250	87	YKT6	137	SCYL3	187	SNTA1	237	AC092171.4	287	BRD4
38	C7orf25	88	ANKRD40	138	ARFGEF1	188	PPTC7	238	XPNPEP3	288	NSD2
39	CYHR1	89	NSL1	139	GATD1	189	USP32	239	STX17	289	MRPL41
40	UBE2J2	90	BTAF1	140	ITSN1	190	ANKRD13B	240	RGS11	290	RP11-334C17.6
41	NPIPB6	91	RP4-761J14.10	141	FIZ1	191	TRMO	241	CTC-366B18.4	291	C17orf80
42	UCK1	92	ZFYVE21	142	C20orf194	192	TMEM169	242	CTIF	292	CUL2
43	TBRG1	93	STK35	143	RP11-274B21.10	193	JUN	243	LTB4R2	293	EIF3C
44	TMEM81	94	ZMAT3	144	MRPL55	194	UNC50	244	HSP90AB1	294	C10orf105
45	PLIN2	95	MED6	145	ZNF20	195	PHYHD1	245	NMNAT1	295	PDE4A
46	ZNF789	96	CDC123	146	PEA15	196	EIF4E2	246	ZNF396	296	TANGO6
47	RNF170	97	LIME1	147	SPAST	197	ING4	247	RP11-434H6.7	297	ZSWIM1
48	DDX24	98	RP11-498C9.15	148	FAM210A	198	STK24	248	C6orf203	298	LRTOMT
49	ARIH1	99	LINC00957	149	ZNF768	199	ZNF408	249	KIAA2013	299	CCNDBP1
50	AP5M1	100	LIF	150	JKAMP	200	HILPDA	250	ZNF664	300	EXOC5

Figure A.17: Top 300 genes with higher sum of loading values from PC1 of PCA. It is represented the genes symbol ID.

A.7 Age Analysis: PCA

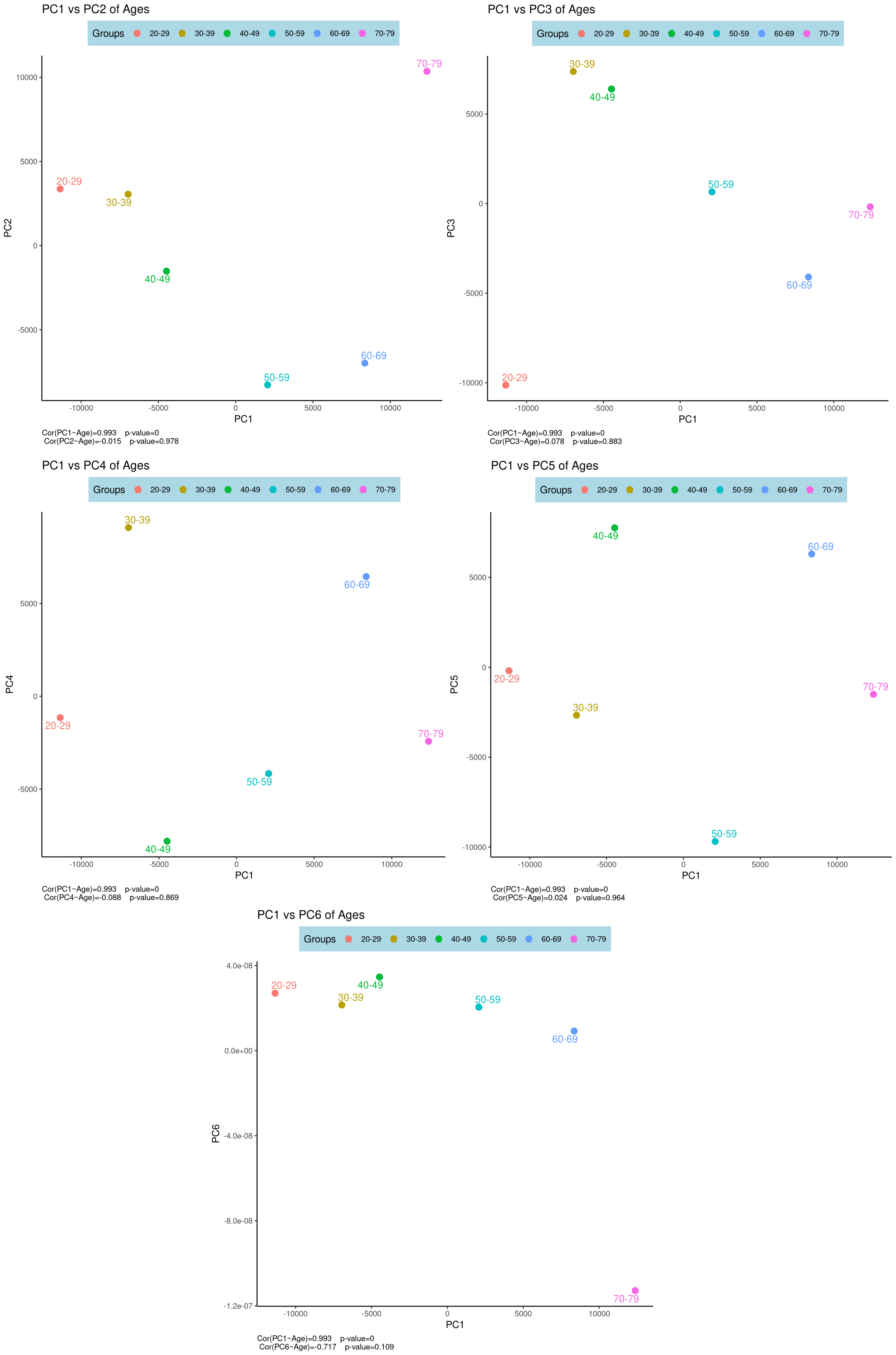


Figure A.18: Principal components plot of age group subsets. Each age group subset carries the gene correlation matrix values.