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
francisco.c.lopes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2021

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. 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.

1. Introduction

1.1. Gene Modules Co-expression Across and Within Tissues

Detection of co-expression gene modules is frequently used to infer about gene-gene interactions, functional annotation and allow for a better understanding of disease origin and progression [1].

It is expected that gene regulation networks are not entirely rearranged between tissues, and there is still no comprehensive support that many tissue-specific data and studies could be much more unified than it already is. Co-expression analysis can be used to explore this concept.

Actually, there have been studies [2] 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).

A more recent study [3] 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.

Another recent study [4], identified regulon modules that globally regulate multiple cell groups and tissues across mouse cell atlases.

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. 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

Ageing occurs in all living organisms and is a natural process that can be defined as a deterioration of the cell functioning [5] 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 [6]. 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.

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 [7], and human pancreas [8]. This transcriptional noise levels that increase with age are suggested as a possible consequence of the accumulation of mutations or and epimutations [9].

Then there is one recent study [10] that measures a global coordination level (GCL) metric in 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.

Additionally, it has been previously described the decrease in gene co-expression within genetic modules in bulk microarray data across 16 different mice tissues [11].

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.

2. Methodologies

2.1. Data Loading

This project used publicly available 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 WGS, 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

Genes whose mean expression was below 1 were filtered out. This approximately corresponds to the 37% quantile of the genes mean expression.

Additionally, genes were filtered by biotype. The gene biotypes that were considered relevant and kept after filtering were protein-coding, long intergenic noncoding RNA (lincRNA), small nuclear RNA (snRNA), micro RNA (miRNA), and small nucleolar RNA (snoRNA).

2.2.2 Sample Filtering

Following GTEx Portal recommendations, samples with RNA Integrity Number (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

For the present work which uses analyses between samples and not within-sample comparisons, it is required to consider sequencing depth and RNA composition uppon data normalization. Thus an adequate method is DESeq2 [12] normalization, which was the one implemented.

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

2.4. Batch Effect Correction

Before batch effect removal, a log2 transformation (the most common) was applied to the data. This is done to compensate for the exponential amplification of RNA-Seq PCR step and also because it is desired to capture proportional changes in gene expression rather than additive changes which is typically biologically more relevant.

It was used a Linear Regression (LR) adjustment for known confounders batch effect removal. The known covariates removed were ischemic time, experimental batch and death type, fitting the model for each gene separately. Known covariates were regressed out using the R statistical environment's built-in ”lm” function.

2.5. Sample Subsetting

2.5.1 Tissue-specific and cross-tissue analysis

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 filtered data. In this Age analysis, it is fundamental for the "testing" age groups sample subsets to have as many samples as possible and 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.

So one option, to equilibrate the amount of samples between comparisons, 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. 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. This reasoning applied to all tissues for both genders results in 371 samples which can be used as testing data in each of the age groups.

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 953 samples per age group.

But this available data is not the actual data used for learning since this data still contains the testing data. The actual data used for learning is the subtraction of the samples used for testing from the available data, which leaves 582 samples per age group for the learning step.

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.

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 1, and clustering trees were computed using the R statistical environment built-in "hclust" function. 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.

After clustering, the average squared correlation of all the pairwise combinations of genes 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 as part of the CRAN R repository project.

2.9. Gene Ontology Enrichment

After heatmapping, some clusters might reveal interesting to delve into. To that end, a 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 as part of the Bioconductor project. For that end, it was used mainly the ”org.Hs.eg.db” annotation to get ”SYMBOL”, ”GENEBIOTYPE” and ”FULLNAME” annotations, and symbol ID's were also complemented by the ”EnsDb.Hsapiens.v79” annotation whenever that correspondence was not found with ”org.Hs.eg.db”. Both annotation packages are available as part of the Bioconductor project.

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

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).

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·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 6 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).

3. Results and Discussion

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.

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 by means of the resulting heatmap represented in figure 1.

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 1.

One type of expected clusters (”Type1” from figure 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

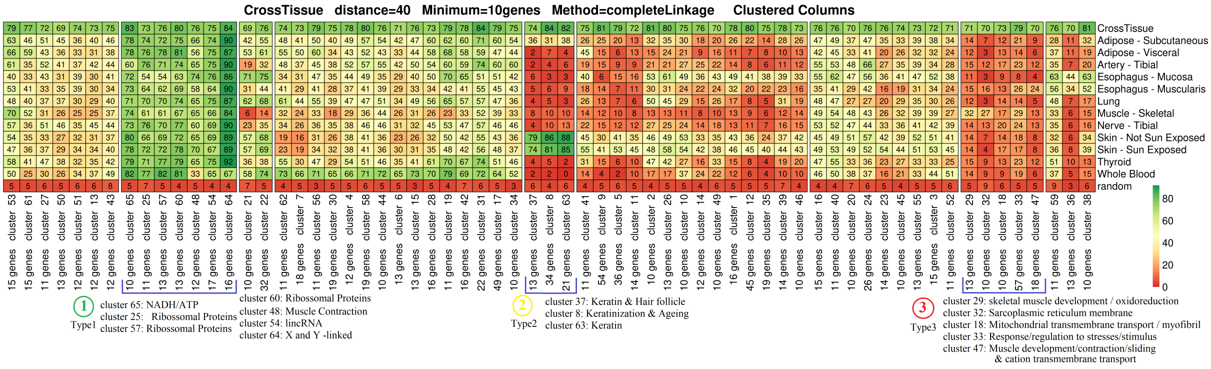


Figure 1: Heatmap of within-cluster squared Pearson correlation (x100) within different tissues. Colour scale reflects the correlation values. The 65 gene clusters were learnt by complete linkage hierarchical clustering in the ”CrossTissue” sample subset (first line of the heatmap) with a minimum cluster 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 annotations.

and ATP metabolism, muscle contraction, development and differentiation, lincRNAs, and X and Y linked genes. 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) were analysed, 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.

NADH/ATP related genes of cluster 65 were observed to 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. 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 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.

Regarding a second type of expected clusters (”Type2” from figure 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 **??** which revealed high correlation values in skin tissues, but very low values in the remaining tissues. These clusters are mainly composed of keratins and keratin-associated protein genes (KAPs). Keratins are the major structural proteins of the vertebrate epidermis, forming keratin intermediate filaments (KIFs) which are a critical component of the *stratum corneum*, the outermost layer of the epidermis [13]. KAPs are responsible for forming the protein matrix between the KIFs [14]. It was observed that these clusters genes have high expression variance and are highly expressed in skin tissues. 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.

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. Otherwise, it can be sample contamination from a neighbouring tissue (such as skin tissue) upon extraction of the sample.

Then there is a third type of expected clusters (”Type3” from figure 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).

Actually, the ”Type3” grouped clusters in figure 1 appear to be related to muscle development and function. In fact, cluster 18 genes’ mean expression show 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.

In cluster 18 GO enrichment analysis, it was 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 [15] 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 [16]. The most abundant mitochondria type in skeletal muscle is IMF mitochondria, located in close contact with the myofibril and found to have higher rates of protein synthesises, enzyme activities, and respiration [16].

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 2.

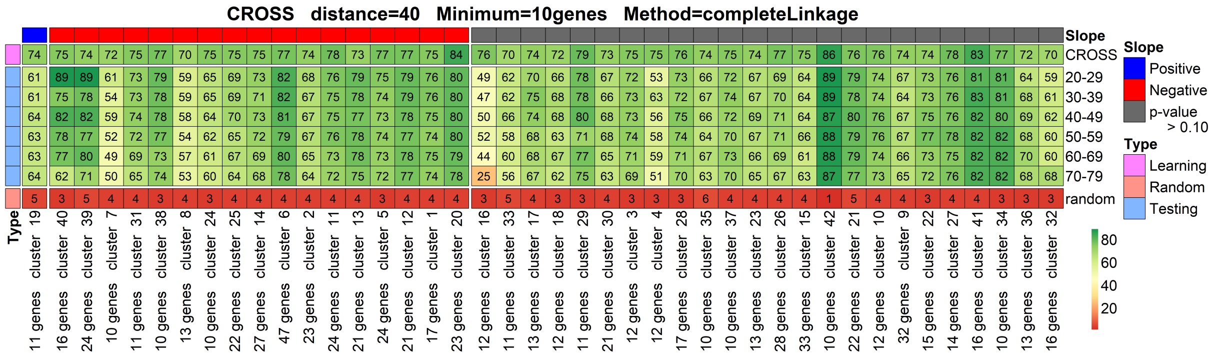
In figure 2, the clusters that significantly (p-value$<$0.10) decreased their within-cluster correlation with age are represented in the red tab columns, and the further to the left, the higher the decrease in within-cluster correlation across ageing. A p-value of 0.10 was considered as significant because 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. 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.

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 [17]. A decline of those skin properties, as well as wrinkle formation, is a common sign of ageing [13]. In addition, studies [18] 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. In the



Type	11 genes cluster 19	16 genes cluster 40	24 genes cluster 39	10 genes cluster 7	11 genes cluster 31	10 genes cluster 38	13 genes cluster 8	10 genes cluster 24	22 genes cluster 25	27 genes cluster 14	47 genes cluster 6	23 genes cluster 2	24 genes cluster 11	21 genes cluster 13	24 genes cluster 5	21 genes cluster 12	17 genes cluster 1	23 genes cluster 20	12 genes cluster 16	11 genes cluster 33	13 genes cluster 17	12 genes cluster 18	11 genes cluster 29	21 genes cluster 30	12 genes cluster 3	12 genes cluster 4	17 genes cluster 28	10 genes cluster 35	10 genes cluster 37	13 genes cluster 23	28 genes cluster 26	33 genes cluster 15	10 genes cluster 42	22 genes cluster 21	12 genes cluster 10	32 genes cluster 9	15 genes cluster 22	14 genes cluster 27	16 genes cluster 41	10 genes cluster 34	13 genes cluster 36	16 genes cluster 32	
	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	Slope
🟪	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	CROSS
🟦	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	20-29
🟦	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	30-39
🟦	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	40-49
🟦	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	50-59
🟦	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	60-69
🟦	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	70-79
🟧	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	random

Figure 2: 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.

present study, it is observed that these genes regulation appears to loosen across ageing. 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.

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 [19].

Cluster 7 GO enrichment analysis revealed immune responses by the complement system. 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 [20]. 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 [20]. For instance, despite great progress in uncovering its genetic links, AMD remains an incurable disease. The present study might be able to give insights into the primary cause of the ageing physiological changes that contribute to autoimmune diseases such as AMD establishing a link with genetic reasoning.

As for cluster 31, its GO enrichment analysis pointed to Natural Killer (NK) cells and Neutrophils. Neutrophils are phagocytic leukocytes that comprise the first line of immune response against invading pathogens [21], mediating the response to bacterial and fungal infections, which are largely responsible for the higher rates of mortality and morbidity in the elderly population [22]. Neutrophil function has been described [22] 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 [23]. With age, significant impairments have been reported in the main mechanisms by which NK cells confer host protection [19]. Actually, the age-associated decline in NK cell function has been associated with slower resolution of inflammatory responses, increased susceptibility of viral infections, being that NK cells are also involved in the recognition, and elimination of senescent cells [19].

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.

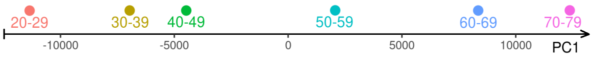
Regarding cluster 38 GO enrichment analysis, it is evident that it represents the system of Major Histocompatibility Complex (MHC) class I. Class I MHC molecules bind peptides generated mainly from the degradation of cytosolic proteins by the proteasome and display those peptides to the cell’s exterior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

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 [24]. A second one is for the CTLs to recognize tumour cells by displaying malignant characteristic immunopeptidomes [25]. And the third one is for the CTLs to recognize virus-infected cells that display foreign peptides in their immunopeptidome [24].

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.

3.3. Genome-Wide Gene-Gene Relationships Across Ageing

As explained in the methods section 2.11 a PCA was applied to the 6 age group correlation matrices where the variables were the hundreds of millions of gene-gene pairs. The resulting PC1 from this analysis follows in figure 3.



Cor(PC1~Age)=0.993 p-value=0

Figure 3: PCA of age groups co-expression plot, where the variables are the gene-gene pairs squared Pearson correlation. Corresponding variable loadings 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. 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.

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. Unfortunately, the trend of loading values is fairly linear, making it challenging to choose a meaningful threshold, and even if it was chosen a threshold.

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 4.

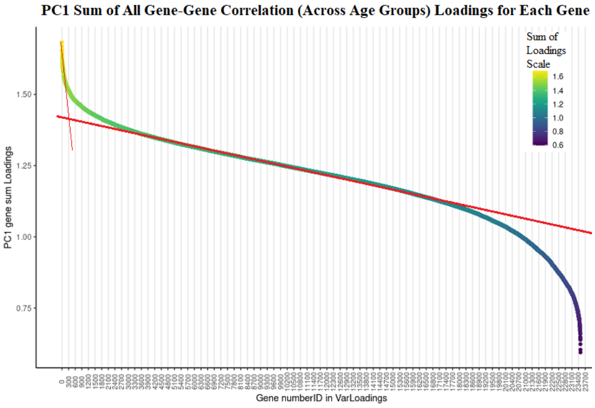


Figure 4: Sum of PC1 variable loading values in which each gene participates. This PC1 was derived from the PCA (figure 3) of the age groups subsets gene-gene pairs squared Pearson correlation. The 2 red lines represent an attempt to find a threshold sum of loadings value.

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 4. This approach can be interpreted as evaluating the hubness of genes regarding their interaction's relevance in describing ageing data variance. Observing figure 4, 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 GO enrichment analyses were applied.

From the obtained GO terms, the most commonly associated 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 [6] 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 are described to decline with ageing [6]. 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 obtained is the Epigenetic Alterations by Histone Modification [6] here represented by the GO term ”regulation of histone deacetylation”. There are several histone acetyltransferases and deacetylases highly associated with the process of ageing [26], therefore it is highly consistent for a histone deacetylation GO term to reveal himself in this analysis. If histone deacetylation becomes less efficient with ageing, there will be less transcriptional regulation in terms of gene suppression.

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 relates to transforming growth factor β (TGF-β) which 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 [27]. 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 across ageing.

Another interesting GO term is ”regulation of fibroblast migration” because it is known [28] 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. By analysing these results of fibroblast migration relationships, it could unveil relevant information that might indicate potential therapeutic targets for the mentioned ageing-related issues.

Lastly, ”RNA secondary structure unwinding” capacity play an important part in translational regulation [29]. Thus, its dysregulation can be another inherent aspect of ageing that should be researched.

4. Conclusions

4.1. 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 and supports that many tissue-specific data and studies can be unified to some extent, much more than is currently done.

4.2. 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’. Some deeply interconnected GO enriched terms were obtained, which revealed to be consistent with the consensual hallmarks of ageing. In this work it was shown that there are large-scale changes in gene co-expression associated with the ageing process.

References

[1] W. Saelens, “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

[2] E. Pierson, “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

[3] B. He, “Gene Coexpression Network and Module Analysis across 52 Human Tissues,” *BioMed Research International*, vol. 2020, 2020.

[4] A. Fønss Møller, “Predicting gene regulatory networks from cell atlases,” *Life Science Alliance*, 2020. [Online]. Available: http://doi.org/10.26508/lsa.202000658

[5] M. C. Barbosa, “Hallmarks of Aging: An Autophagic Perspective,” *Frontiers in Endocrinology*, vol. 0, no. JAN, p. 790, 2019.

[6] C. López-Otín, “The Hallmarks of Aging,” *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013.

[7] I. Angelidis, “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

[8] E. M, “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/

[9] V. J, “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/

[10] O. Levy, “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

[11] L. K. Southworth, “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

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

[13] T. Sano, “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

[14] G. M, “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/

[15] C. N. Joyce Chen, “Muscle Structure and Function,” *Orthopaedic Physical Therapy Secrets: Third Edition*, pp. 1–9, 1 2017.

[16] A. Damirchi, “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

[17] L. Norlén, “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.

[18] T. Sano, “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

[19] J. Hazeldine, “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/

[20] M. V. Carroll, “Complement in health and disease,” *Advanced Drug Delivery Reviews*, vol. 63, no. 12, pp. 965–975, 9 2011.

[21] E. Mortaz, “Update on Neutrophil Function in Severe Inflammation,” *Frontiers in Immunology*, vol. 0, no. OCT, p. 2171, 10 2018.

[22] S. Butcher, “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/

[23] I. Prager, “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

[24] 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

[25] D. Dersh, “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

[26] S.-J. Yi, “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/

[27] 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

[28] D. 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

[29] S. Takyar, “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