BINF-F401 - Computational Methods for Functional Genomics

Course Project - Analysis of the associations between the transcriptome and morphological features of the thyroid gland

Lionel Lenoir

Edwige Loems

David Perez

group 4

Course teacher : Vincent Detours

June 2022

Contents

1 Introduction

1.1 The Canguilhem study

There has been in the past decades a problem of overdiagnosis and overtreating of thyroid cancer [**1**]. In response to this problem, it could be interesting to learn from the french philosopher – and physician – Georges Canguilhem, who in his essay of medical epistemology "*Le normal et le pathologique*", criticises the concept of pathology as being a deviation from a normal state.

It is within this frame of mind that the project of the Canguilhem study was launched. The project consists in a survey of the diversity of normal thyroid morphology, using a novel unbiased and quantitative approach to measure morphology. In order to achieve that, thyroid histological slices from the GTEx database are analysed. GTEx stands for Genotype-Tissue Expression, this project aims to collect samples for different non-diseased tissues across thousand of individuals and to study the gene expression and regulation in those different tissues. Among the goals of the study, the most relevant ones in the context of this project are the survey of the correlation of morphological variations with clinical data and gene expression.

1.2 Quantitative morphology

Quantitative morphology is based on the quantitative information that can be extracted from an image. In contrast, qualitative morphology relies on the interpretation of an image and its characteristics which implies cognitive and scientific biases. In this context quantitative morphology allows to get rid of those bias. This ties in with Canghuilhem's criticism of the pathological state being a deviation from normal state define by scientific and cognitive bias.

In order to achieve the goals of the study, a quantitative and unbiased approach was developed to characterise the morphology of a thyroid slice. The general strategy consists in dividing the image of the slices in tiles and to use unsupervised AI to gather these tiles into different clusters. These clusters are meant to correspond to morphological categories, which should hold a meaningful biological significance hence different genes expression for each morphological categories.

The way this was achieved is by using the tiles to train deep neural networks using a contrastive learning framework. This is used to produce a compressed version of the images which exist in a space with much less dimensions than the number of pixels of the original images and which is called the *latent representation* of the image. The images in their latent representation are then collapsed to 2 dimensions – using the UMAP method. In the latent space, the tiles are grouped in 64 clusters, using an unsupervized clustering method.

Once each tile is appointed to a cluster, the number of tiles from each cluster can be computed for each thyroid slice and the gene expression for each cluster can be computed too. Thus, a quantitative morphological description of the thyroid slice is achieved, and it was obtained without human supervision, or annotation.

A thyroid as any other biological tissue or organ is composed of different cells with different morphological features, moreover the images used to create the tiles come from microscopic instrument. The preparation of the tissues in order to observe them under the microscope implies technical bias, e.g.: coloration default, tear in the tissues or photographic flaws.

1.3 Overview of the project

The data used in this project describes 136 samples – thyroid slices – from the GTEx database. 3 tables contain the data describing these samples : a table with the clinical data, containing information about the individuals from whom the thyroids were extracted, a table with the counts of the 64 morphological clusters characterising the slices, and finally a table with gene expression data – RNA read counts – for each sample. More information on these tables will be given in the appropriate sections of this report.

The first part of the project consists in an overview of the clinical data. In particular, the distribution of the clinical variables are plotted and commented on, correlations and dependences between the different variables are computed, and confounding technical variables are looked for.

In the second section, associations between the clinical variables and the cluster counts are determined and represented.

Finally, in the third part of the project, associations between morphological clusters, gene expression and REACTOME gene sets are analysed. The REACTOME gene set gives us information about biological pathways and processes and the genes associated to each one.

2 Exploration of the clinical data

2.1 Distribution of the clinical variables

The clinical data used in this project is contained in 8 variables, 5 demographic variables and 3 technical ones. The 5 demographic variables are sex, age, height, weight and BMI (Body Mass Index). The 3 technical variables are the cohort (whether the individual was a postmortem donor or organ donor), the ischemic time (time between death and the start of the GTEx procedure) and a variable called the Hardy scale, categorizing individuals by type of death. More extensive information about these variables can be found at the url : https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000424/phs000424.v8.p2/pheno_variable_summaries/phs000424.v8.pht002742.v8.GTEx_Subject_Phenotypes.data_dict.xml.

The first step of the exploration of these variables is the visualisation of their distribution.

2.1.1 Sex

This distribution is represented on the figure 2.1

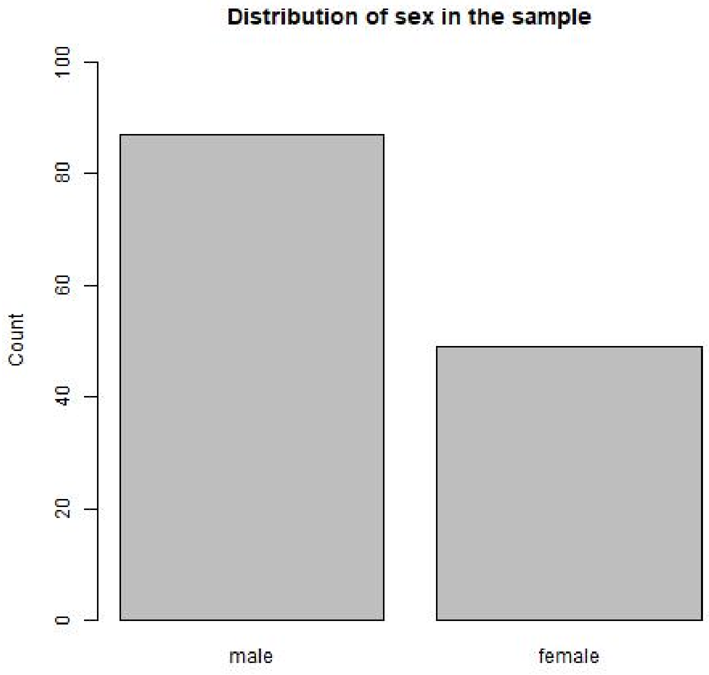


Figure 2.1: Repartition of both sexes in the sample

The distribution between male and female is unbalanced in our sample : there are 87 males for 49 females (which represents 64% and 36% respectively).

2.1.2 Age

An histogram of the distribution of ages in the sample is represented in figure 2.2

The statistical summary of this distribution is the following :

Minimum : 21 years, maximum : 49 years, median : 38 years and mean : 37.53 years

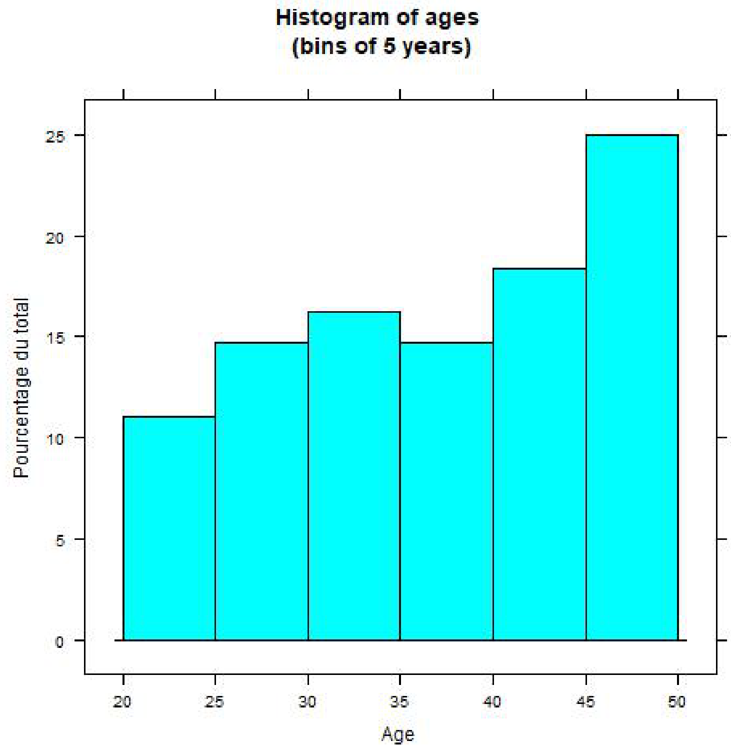


Figure 2.2: Distribution of ages

It should be noted that the sample does not contain anyone older than 50 years old, which is a relatively young age for a cutoff in the sample. This should be kept in mind later, when the age variable and its correlation on morphological cluster counts will be analysed.

2.1.3 Heights

An histogram of the distribution of heights in the sample is represented in figure 2.3

The statistical summary of this distribution is the following : Minimum : 58 inches, maximum : 76 inches, median : 68 inches and mean : 68.18 inches

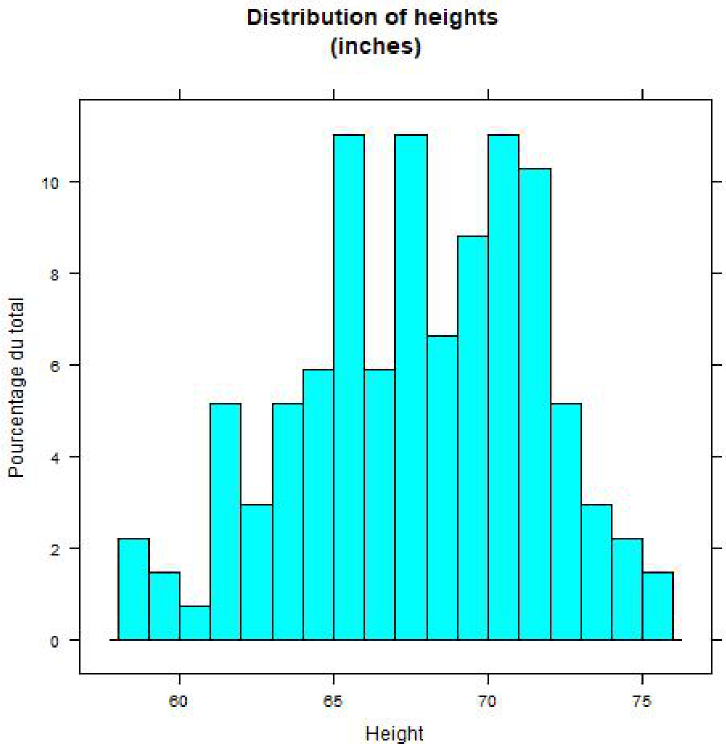


Figure 2.3: Distribution of heights

This distribution seems to roughly follows a bell curve.

2.1.4 Weights and BMI

An histogram of the distribution of weights in the sample is represented in figure 2.4, and the distribution of BMIs is represented on figure 2.5

The statistical summary of the distribution of the weights is the following : Minimum : 92 pounds, maximum : 264 pounds, median : 184.8 pounds and mean : 182.9 pounds.

Regarding the BMI it is as follow : Minimum : 18.58, maximum : 34.86, median : 27.75 and mean : 27.33.

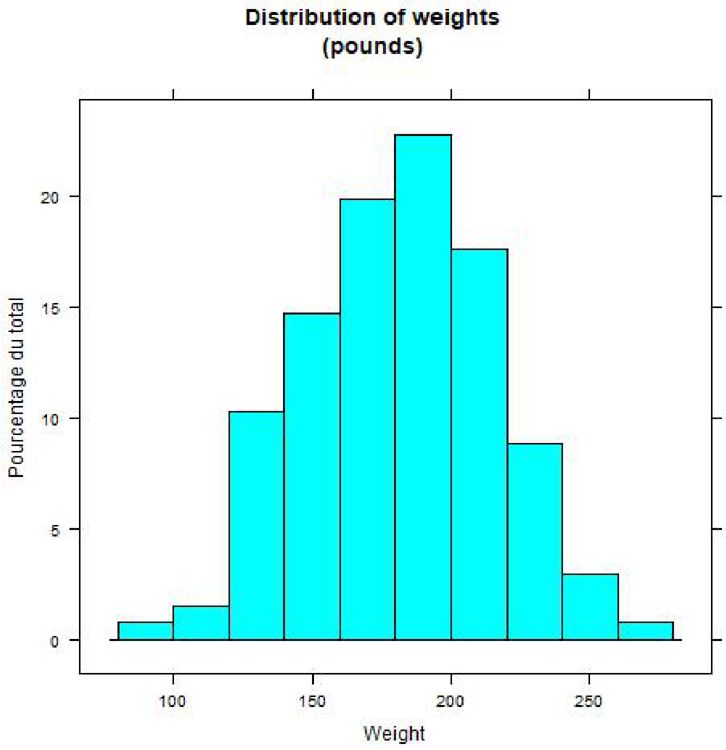


Figure 2.4: Distribution of weights

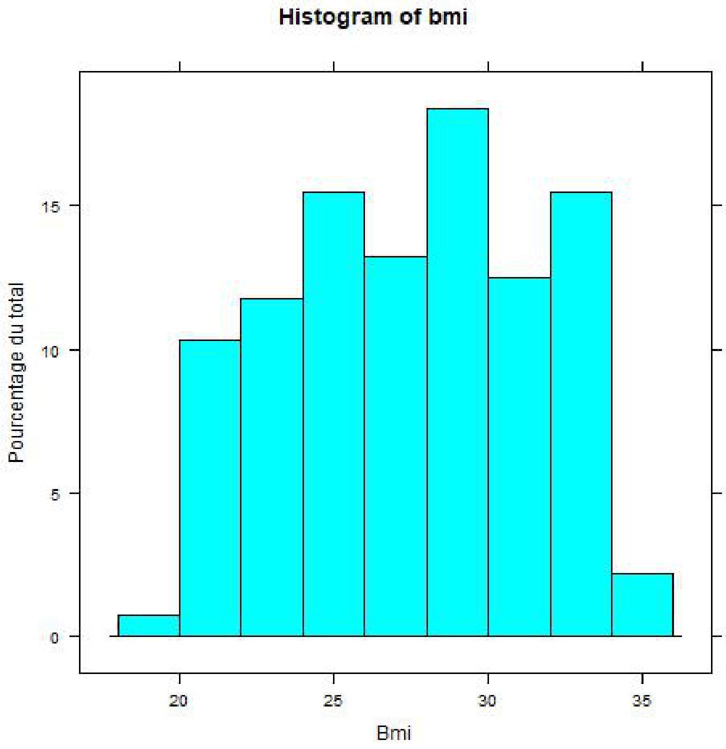


Figure 2.5: Distribution of BMI

While the weights – as for the heights – are distributed around more frequent central values, with decreasing frequency as the value of the weight gets further from the mean, the BMIs are more uniformly distributed, with a few extreme values.

2.1.5 Cohort

First of all, some clarifications concerning the technical variable "cohort". In the framework of this project, the cohort includes two categories: organ donor and post-mortem donor. As mentioned before the GTEx project aims to collect non-diseased tissues in order to do so eligibility criteria have been set up to ensure that organ donor could be used to describe non-diseased tissues[**2**]. Thus organ donor refers to subject who complies to the standard of GTEx, per se non-diseased thyroids, and postmortem refers to subject who do not complies to the standards of GTEx, thus one can expect to observe morphological features or gene expression uncharacteristic of non-diseased thyroid in the forthcoming analysis.

This distribution is represented on the figure 2.6

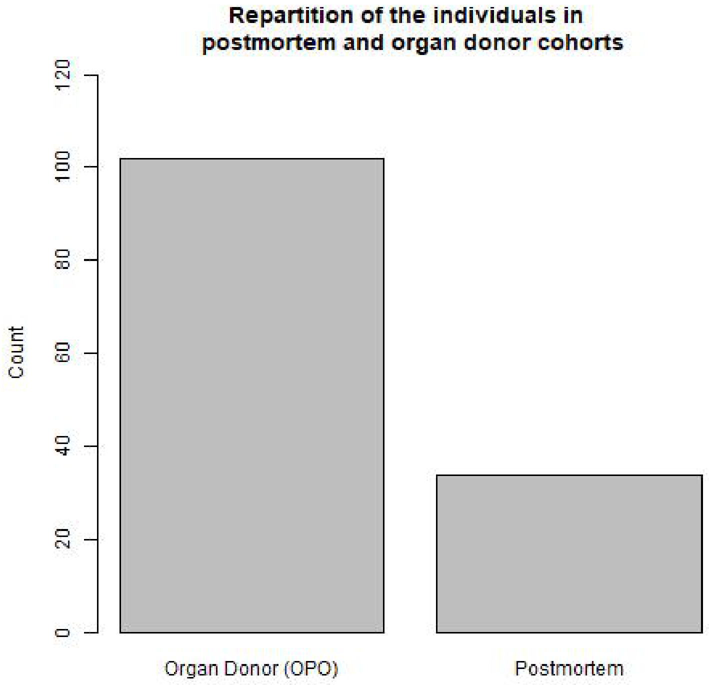


Figure 2.6: Number of individuals in each cohort

As shown in figure 2.6, many more people in the sample were part of the cohort of organ donors than those who were postmortem donors. The percentages are respectively 75% and 25%.It is important to note that a majority of the samples that are going to be analysed could be used to characterise non-diseased thyroid which is in line with the perspective of the Canguilhem study.

2.1.6 Ischemic time

As can be seen on the figure 2.7, the shortest ischemic times are the most frequent, with a frequency decreasing as the ischemic time increases. However there are cases where there were several hours – up to more than a day in some cases – between the time of death and the procedure to extract the thyroid slice.

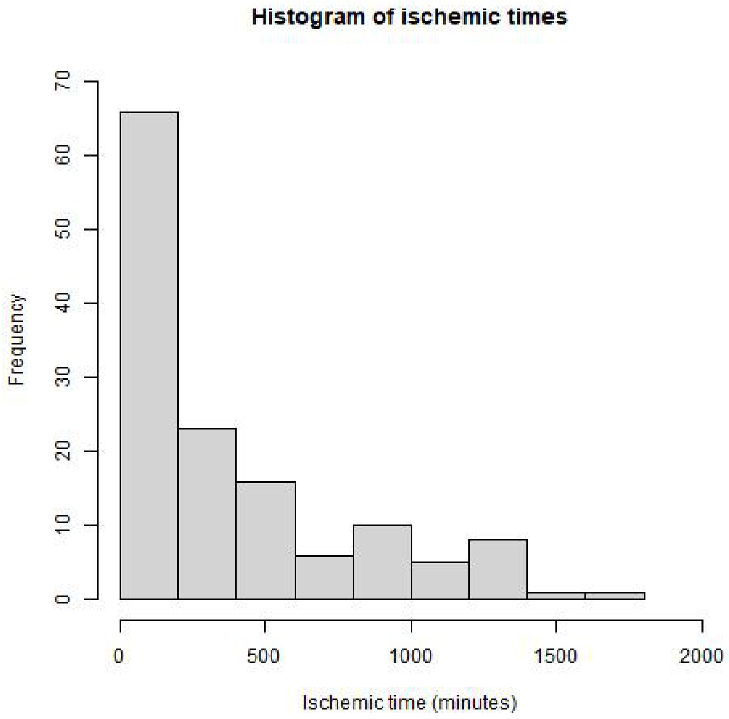


Figure 2.7: Distribution of ischemic times

In the frame of this project, it is useful to keep in mind that longest ischemic times implies potential tissue deterioration before their collection thus potential unwanted technical features.

2.1.7 Hardy scale

The Hardy scale is a system of categorisation of the type of death. The scale goes from 0 to 4 with 0 representing cases on a ventilator immediately before death and :

1. Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min.

2. Fast death of natural causes Sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr.

3. Intermediate death Death after a terminal phase of 1 to 24 hrs

4. Slow death Death after a long illness, with a terminal phase longer than 1 day

As can be observed on figure 2.8, most of the cases here are ventilator cases (more than 75%) and the categories 1, 3 and 4 having less than 10 cases.

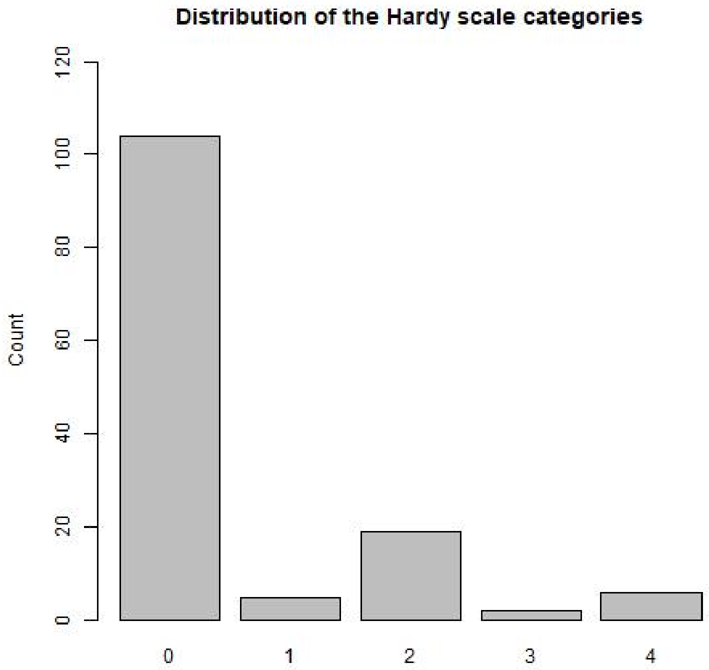


Figure 2.8: Repartition of cases in the Hardy scale

This uneven distribution with a majority of the subject being ventilator cases is not surprising since the goal of this study is to gather morphological and genes expression information about normal thyroid and not about thyroid that have been exposed to disease or illness, as for the 3rd and 4th scales of the Hardy scale.

2.2 Correlation between variables

In this section, correlation between different variables are analysed, first between demographic variables, then between technical and demographic variables and more specifically regarding technical variables that may be confounding demographic variables.

2.2.1 Correlation between demographic variables

Since the sex is the only demographic variable that is not numerical, it has to be analysed in a different section than the other numerical variable – age, weight, height and BMI.

Sex

The graphical representation of the distribution of clinical variable between male and female is shown in figure 2.9.

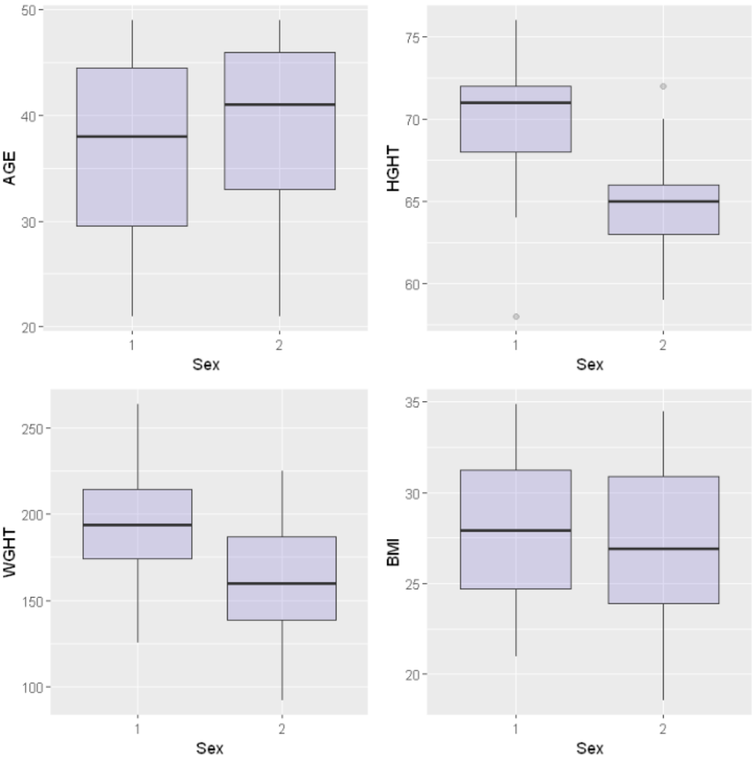


Figure 2.9: Boxplot representing the distribution of clinical variables by sex

The graphical representation show a possible correlation between the weight, the height and the sex of the subject. In order to formalise our analysis, a Mann-Whitney's test is performed for each numeric demographic variables – age, weight, height and BMI – considering the two sex as two independent categories. The Mann-Whitney's test allows to evaluate if difference in the distribution of the variable taken into account in each test is due to the sex or independent of it. The result of the Mann-Whitney's test are p-values, for the records p-values measures the proportion of false positives among the positive results.

1. P-value for age : 0.131321452047474

2. P-value for height : 3.51914176038072e-16

3. P-value for weight : 1.48960430486491e-07

4. P-value for BMI : 0.377965307143104

P-values greater than 0.05 for the age and BMI does not allow the hypothesis of independence of the variables to be rejected. On the contrary P-values lower than 0.05 for the height and weight allow to reject the hypothesis of independence thus height and weight are in fact correlated to the sex of the subject

Age, weight, height and BMI

Since those values are continuous it is possible to directly calculate the correlation between and plot them against each other to obtain a more graphical representation. Those results are shown in figure 2.10 and 2.11

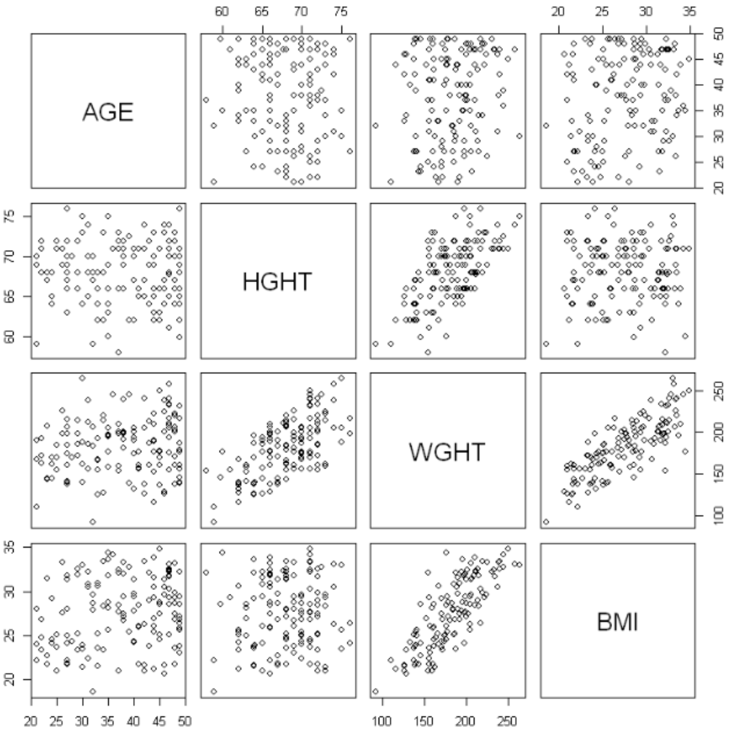
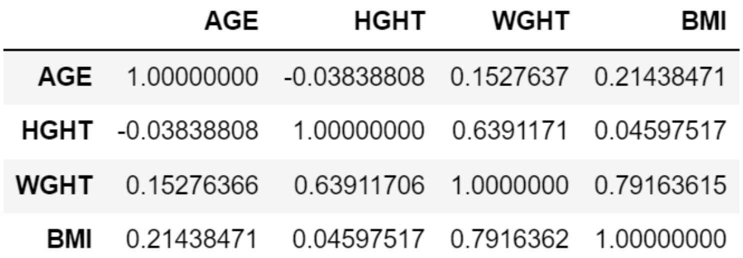


Figure 2.10: Distribution of the different clinical variables with respect to each other



	AGE	HGHT	WGHT	BMI
AGE	1.00000000	-0.03838808	0.1527637	0.21438471
HGHT	-0.03838808	1.00000000	0.6391171	0.04597517
WGHT	0.15276366	0.63911706	1.0000000	0.79163615
BMI	0.21438471	0.04597517	0.7916362	1.00000000

Figure 2.11: Matrix of the correlation between each numerical clinical variable

The significance is computed using the Pearson's test to obtain P-values and correct them with the Bonferroni correction.

1. P-value for age against height : 0.657252579129967

2. P-value for age against weight : 0.0758141030285219

3. P-value for age against BMI : 0.0122012961169879

4. P-value for height against weight : 5.62213707188937e-17

5. P-value for height against BMI : 0.595079094701712

6. P-value for weight against BMI : 1.84504113881356e-30

P-values lower than 0.05 highlight the dependence between the two variables from which it derive. In this case :

1. Age and BMI

2. Height and weight

3. Weight and BMI

2.2.2 Correlation between demographic variables and technical variables

The Fisher test applied to the two categorical variable of the dataset – sex and cohort – provided a p-value of 0.8373, greater than 0.05 which does not allow the hypothesis of independence of the variables to be rejected. The Fisher test have been chosen because is applicable on small count data and thus fits this case perfectly.

The graphical representation of figure 2.12 does not show a clear correlation between demographic variables and the cohort. With the aim to showcase potentials correlation a Mann-Whitney's test is performed for demographic variables – age, weight, height and BMI – considering the two cohort. The outcomes are the following P-values :

1. P-value for age : 0.000715351356559735

2. P-value for height : 0.0793189642162195

3. P-value for weight : 0.0113078041343223

4. P-value for BMI : 0.105604145672584

P-values greater than 0.05 for the height and BMI does not allow the hypothesis of independence of the variables to be rejected. On the contrary P-values lower than 0.05 for the age and weight allow to reject the hypothesis of independence thus age and weight are in fact correlated to the cohort of the subject

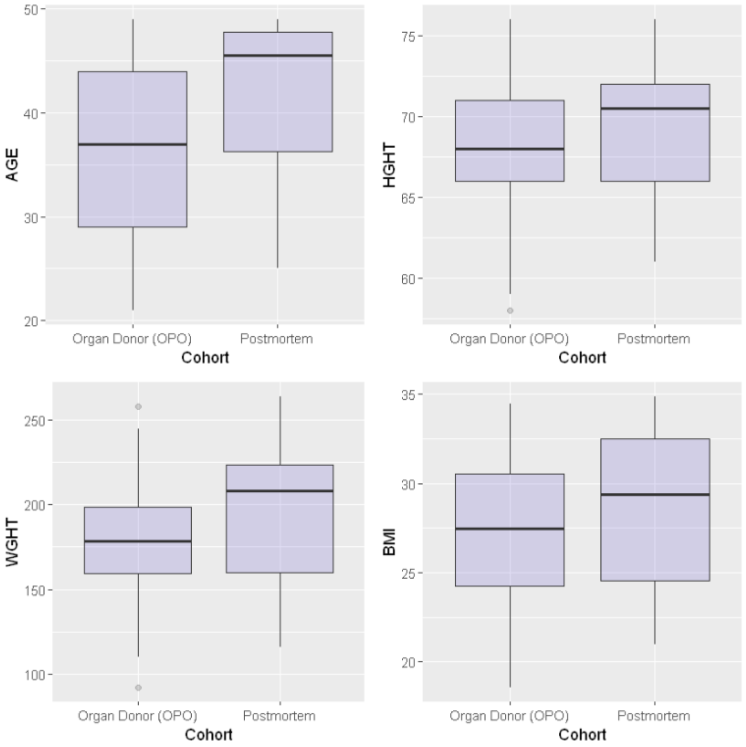


Figure 2.12: Distribution of the clinical variables between the two cohorts

In order to facilitate the representation and analysis, the "fast deaths" case from the Hardy scale – 1 and 2 – are regrouped in a single ’1’ class and the "slow deaths" case – 0, 2 and 3 – in a ’0’ class. The graphical representation of figure 2.13 does not show a clear correlation between demographic variables and the different classes of the Hardy scale.

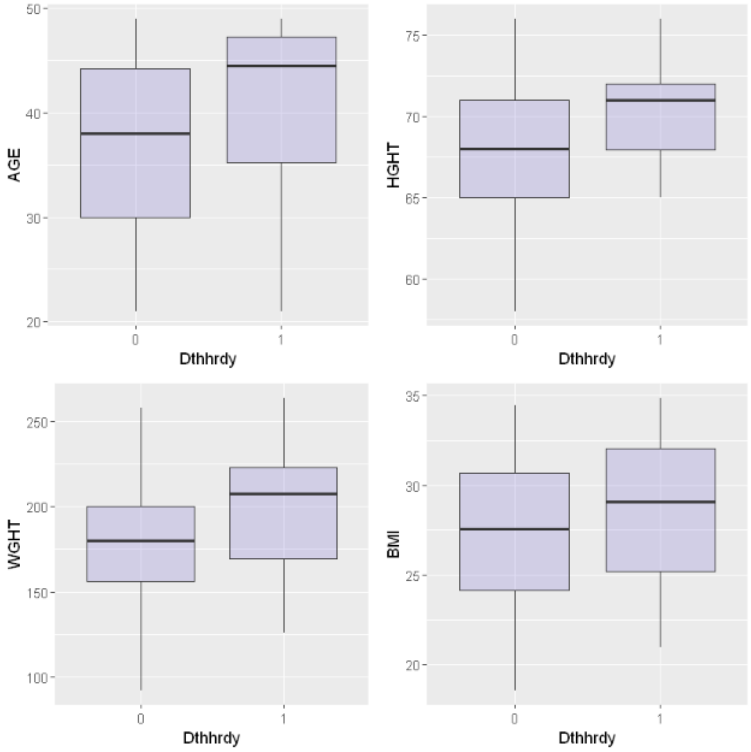


Figure 2.13: Distribution of the clinical variables on the Hardy scale

In order to showcase potentials correlation a Mann-Whitney's test is performed for demographic variables considering the two newly formed group from the Hardy scale. The outcomes are the following P-values :

1. P-value for age : 0.0281955890446734
2. P-value for height : 0.00422709674948223
3. P-value for weight : 0.00692956857751305
4. P-value for BMI : 0.195035914673178

P-values lower than 0.05 for the height, weight and BMI allow to reject the hypothesis of independence thus height, weight and BMI are correlated with the Hardy scale. Since BMI is a product of the height and weight of the subject it is obvious that its correlation with the Hardy scale flows from the correlations of the height and weight with the Hardy scale.

Confounding technical variable

The variable 'AGE' in the dataset is correlated to the variables 'COHORT' and 'HARDY', the variable 'HEIGHT' is correlated with 'HARDY' and the variable 'WEIGHT' is correlated with the 'COHORT' as shown in the analysis above. These demographic data – age, height, weight – will need to be adjusted to these confounding technical variable – cohort, Hardy – in order to avoid spurious correlation in further statistical analysis.

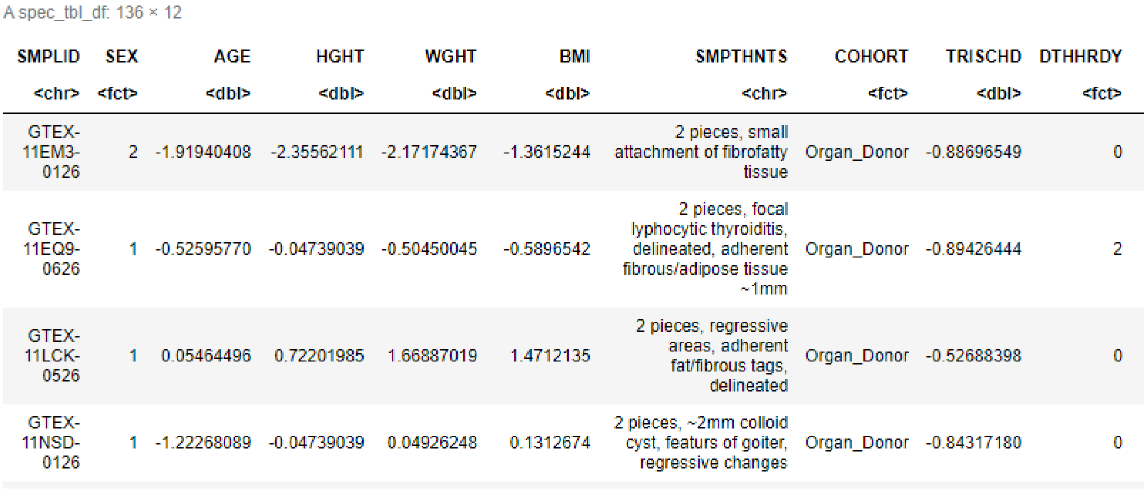
3

Associations between clinical data and morphology

3.1 Differential expression computations

The purpose of this section is to compute associations between the clinical variables and the morphological features of the thyroid sections – the counts associated to the 64 morphological clusters. As this essentially consists in differential expression analysis where cluster counts are used instead of gene expression, tools for differential gene expression analysis can be used. It is why the library DESeq2 was used for the computations (with R version 4.1.3).

Among the 8 clinical variables, 5 of them are continuous numerical variables (age, height, weight, bmi, ischemic time) and 3 are categorical (sex, cohort and Hardy scale). The first step was to normalize the numerical variables. This was done with the *scale()* function of R which centers and scales the columns of a numeric matrix – substract the mean and divide by the standard deviation. Once this computation is made, the clinical data finds itself in a dataframe of which a sample is shown in figure 3.1. It is noteworthy to point out that the type of the catogorical variables is changed to *factor*, in order to make sure that their value are not misinterpreted as quantitative information.

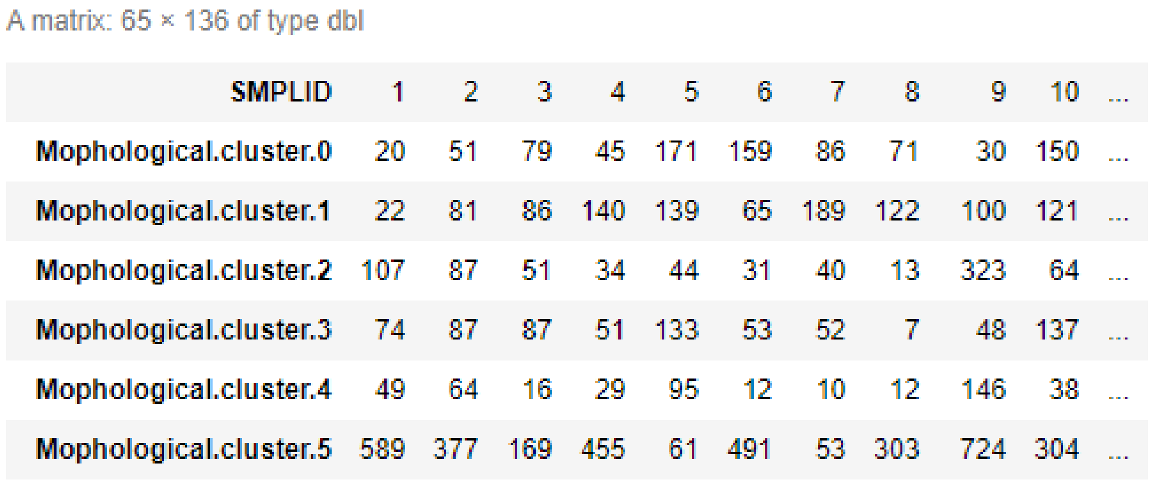


A spec_tbl_df: 136 × 12

SMPLID	SEX	AGE	HGHT	WGHT	BMI	SMPTHNTS	COHORT	TRISCHD	DTHHRDY
<chr>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<fct>	<dbl>	<fct>
GTEX-11EM3-0126	2	-1.91940408	-2.35562111	-2.17174367	-1.3615244	2 pieces, small attachment of fibrofatty tissue	Organ_Donor	-0.88696549	0
GTEX-11EQ9-0626	1	-0.52595770	-0.04739039	-0.50450045	-0.5896542	2 pieces, focal lyphocytic thyroiditis, delineated, adherent fibrous/adipose tissue ~1mm	Organ_Donor	-0.89426444	2
GTEX-11LCK-0526	1	0.05464496	0.72201985	1.66887019	1.4712135	2 pieces, regressive areas, adherent fat/fibrous tags, delineated	Organ_Donor	-0.52688398	0
GTEX-11NSD-0126	1	-1.22268089	-0.04739039	0.04926248	0.1312674	2 pieces, ~2mm colloid cyst, featurs of goiter, regressive changes	Organ_Donor	-0.84317180	0

Figure 3.1: Part of the dataframe with normalized clinical data

The data of the morphological cluster counts used is a tsv with a matrix of the count of tiles in each morphological cluster. In order to be able to use it as an input for DESeq2, it is transposed, to obtain a matrix partially shown on figure 3.2.



A matrix: 65 × 136 of type dbl

SMPLID	1	2	3	4	5	6	7	8	9	10	...
Mophological.cluster.0	20	51	79	45	171	159	86	71	30	150	...
Mophological.cluster.1	22	81	86	140	139	65	189	122	100	121	...
Mophological.cluster.2	107	87	51	34	44	31	40	13	323	64	...
Mophological.cluster.3	74	87	87	51	133	53	52	7	48	137	...
Mophological.cluster.4	49	64	16	29	95	12	10	12	146	38	...
Mophological.cluster.5	589	377	169	455	61	491	53	303	724	304	...

Figure 3.2: Part of the matrix with morphological clusters counts

The computation of the fold changes of the different morphological clusters with respect to the different variables is done with the following few lines of code (in this case for the variable 'SEX') :

```
#creating the dataset object
dds = DESeqDataSetFromMatrix(t(data.matrix(mc)),colData=cd, design = ~ SEX)
# computation of differential expressions
dds <- DESeq(dds, quiet=TRUE)
res <- results(dds)
#selection of clusters with adjusted p-values < 0.05
resSig <- subset(res, padj < 0.05)
```

Where *t(data.matrix(mc))* corresponds to the matrix represented in figure 3.2. This code enables to compute the dataframe ***res***, which contains the different computed results and in particular the ***log2FloldChange***, the logarithm in base 2 of the fold change. As explained in the DESeq2 package documentation : "For a particular gene, a log2 fold change of -1 for condition treated vs untreated means that the treatment induces a multiplicative change in observed gene expression level of $2^{-1} = 0.5$ compared to the untreated condition." For a continuous variable, the fold change is computed per unit of change of that variable [**3**] – which is also why it is important to normalize the variables, to not have a fold change merely dependent of the scale of the variables, or the units used.

Another important output stored in ***res*** is the adjusted p-value – adjusted for multiple testing – obtained with the Wald test. This value enables to select clusters – or genes – for which the fold change is statistically significant.

To store our results, a dataframe (***assoc_table***) is defined, for which each column corresponds to a morphological cluster and each line to a clinical variable. All the elements of the dataframe corresponding to a cluster/variable couple with an adjusted p-value lower than 0.05 are filled with the value of the log2 fold change. The other elements of the matrix are left NULL as the p-value above 0.05 indicates that the hypothesis of independence between the variable and the cluster can not be rejected.

First, it is interesting to look if any cluster does not cross the threshold of an adjusted p-value under 0.05 with any clinical variable. It is indeed the case for the clusters 8, 11, 13, 22, 24, 30, 31, 37, 53 and 55 (so 10 different clusters). In the cases where there is a statistically significant association between the morphological cluster and the clinical variable, we can count the number of "down/up regulated" clusters for each variable, to see which clinical variables impact the morphology the most. This counting is represented on figure 3.3.

The 3 technical variables are associated to the highest number of morphological clusters – 29 for the ischemic time, 34 for the Hardy scale and 37 for the cohort. This suggests that these variables are actually the ones having the largest impact on the morphology of the slices. Among the "demographic" variables, the age is the one having the impact on the highest number of clusters (19). The BMI is the variable with the least impact on the morphological clusters – only 1 : the cluster 14.

It could also be interesting to visualise for each cluster, how many clinical variables have a significant impact on it. This is represented on figure 3.4. The cluster 4 is the cluster impacted by the highest (6) number of variables.

In order to summarize the information computed in this section, the matrix ***assoc_table*** is represented on figure 3.5 as a heatmap, where the coloring depends on the log2 fold change (in the cases where the p-value is under 0.05) with negative values beeing darker and higher positive values beeing brighter, see figure 3.6 for more information. The figure is obtained with the command *heatmap.2*, using the library gplots (with the parameter scale='none'). The clusters without any association to a clinical variable are not represented. Looking at the heatmap, several observations can be made :

- The technical variables are not only the ones affecting the highest number of clusters, but they are also the ones with the strongest impact in terms of fold change.
- Clusters of clusters can be made, of clusters associated with the same variables.
- The cluster 12 has the highest fold change for the 3 technical variables

3.2 Controlling for confounding technical variables

In the first section of the project, it is determined that the variable ’AGE’ in the dataset is correlated to the variables ’COHORT’ and ’HARDY’, as is the variable ’Weight’. The variable ’HEIGHT’ is correlated with ’HARDY’. In order to cancel the effects of these correlations, the computations shown above are made again, but this time while controlling ’AGE’, ’HEIGHT’ and ’WEIGHT’ for the variation of their confounding technical variables. This is done in the definition of the DESeq dataset object, for example with age :

```
dds = DESeqDataSetFromMatrix(t(data.matrix(mc)),colData=cd,
    design = ~ AGE + COHORT + DTHHRDY)
```

We can observe on the figures 3.7 and 3.8 that the number of clusters associated with age and weight has been drastically reduced (to 2 and 3, respectively), perhaps due to the dataset becoming to small to obtain high enough p-values. The height variable, however, is now correlated with 32 morphological clusters.

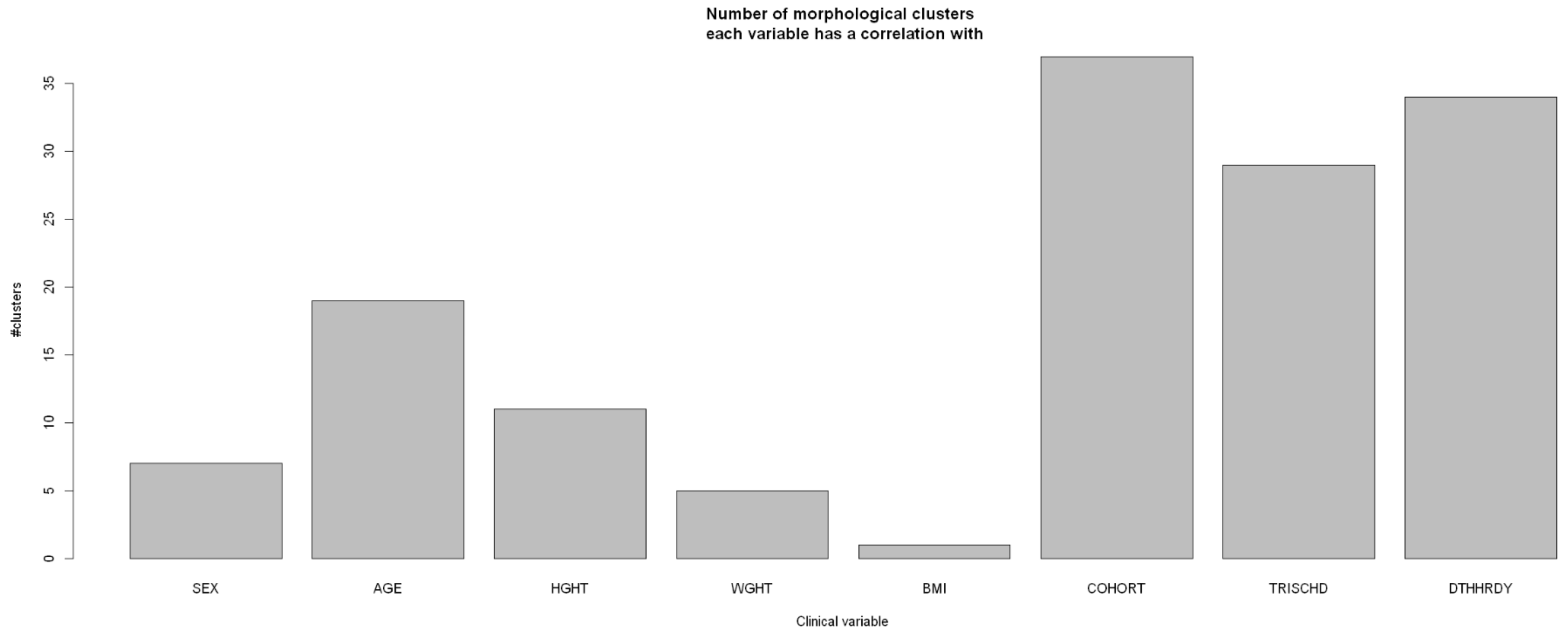


Figure 3.3: Number of morphological clusters each variable has a correlation with

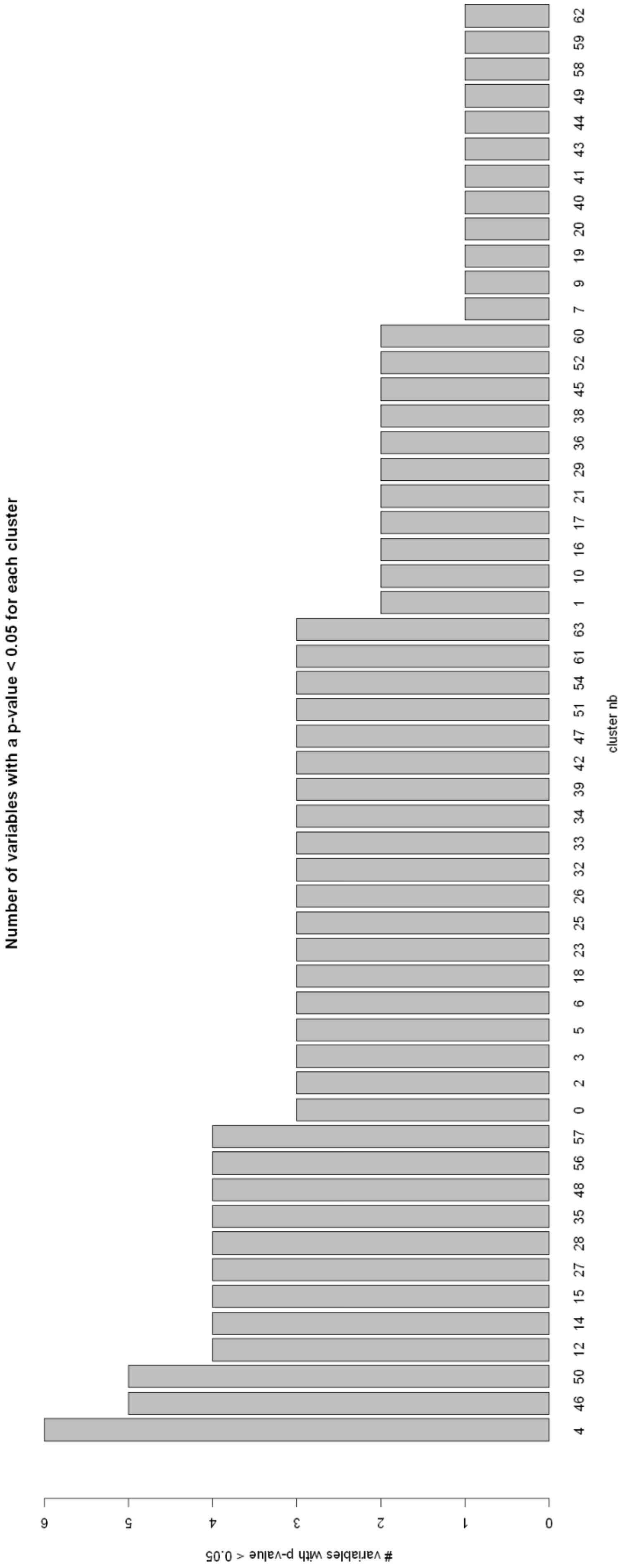


Figure 3.4: Number of variables with a p-value < 0.05 for each cluster

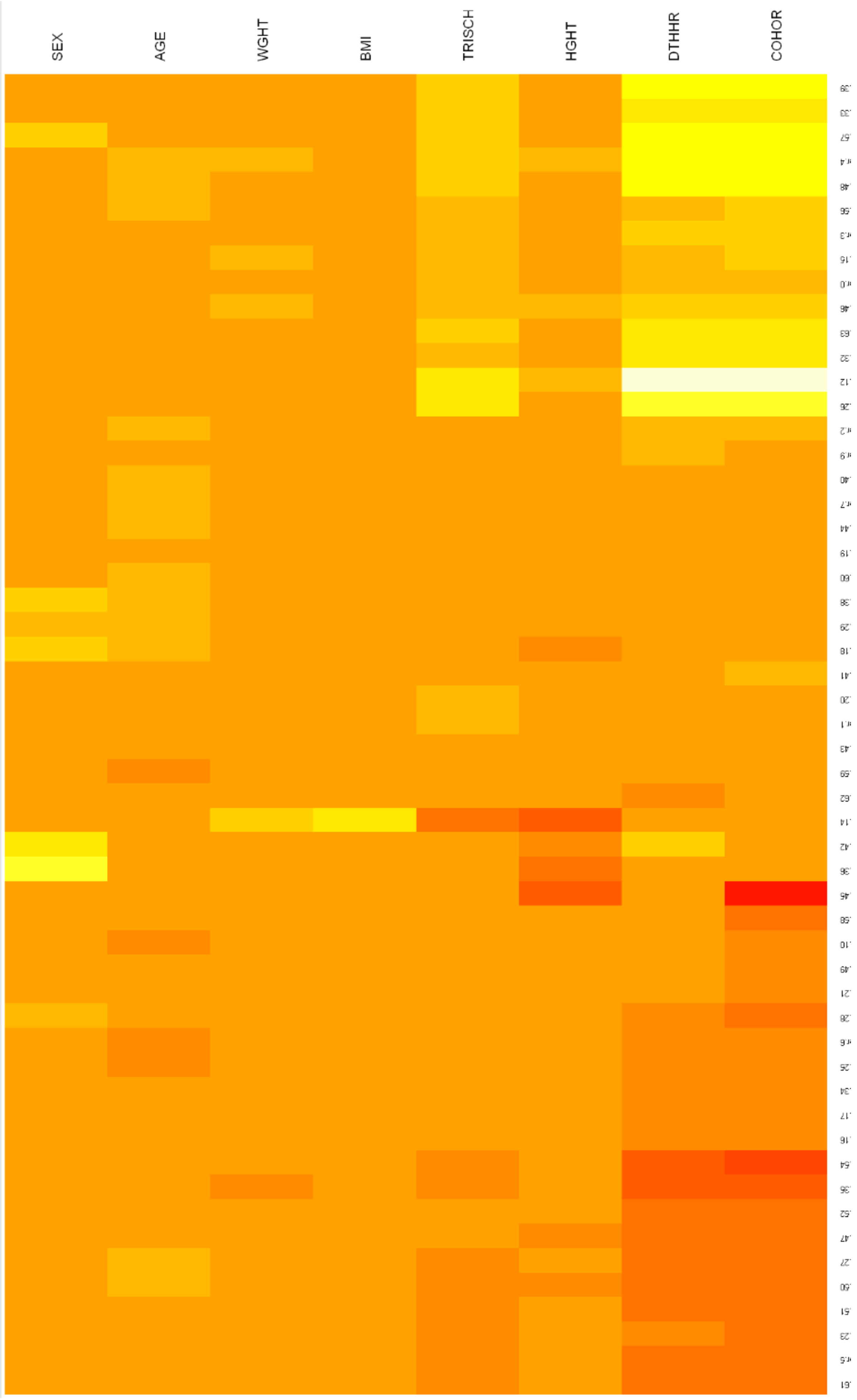


Figure 3.5: Heatmap representing the associations between morphological clusters and clinical variables

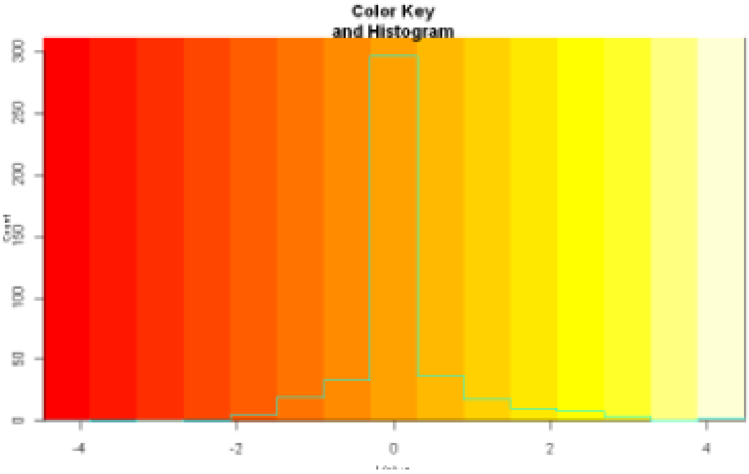


Figure 3.6: Legend for the figure 3.5

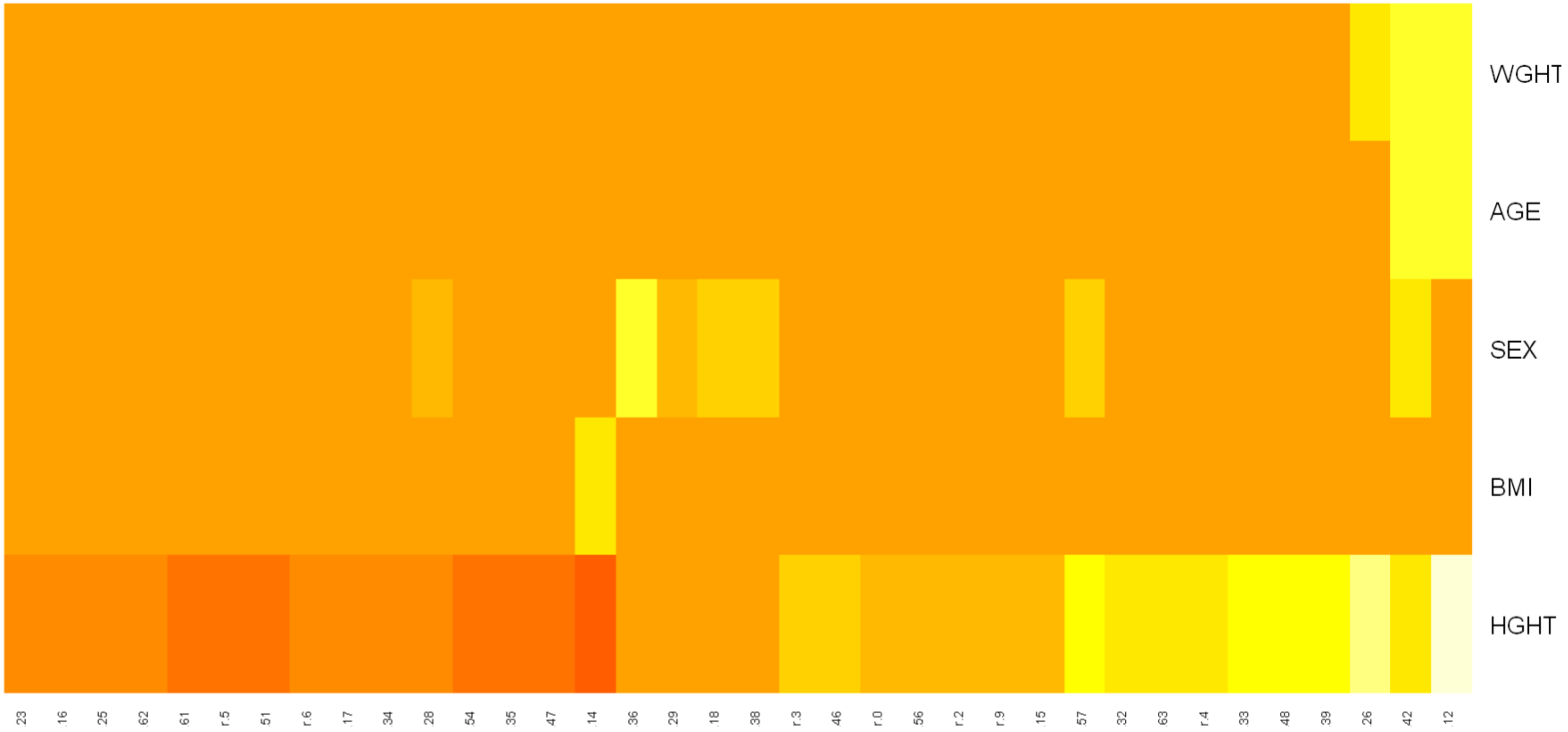


Figure 3.7: Heatmap representing the associations between morphological clusters and clinical variables (controlled for their confounding variables

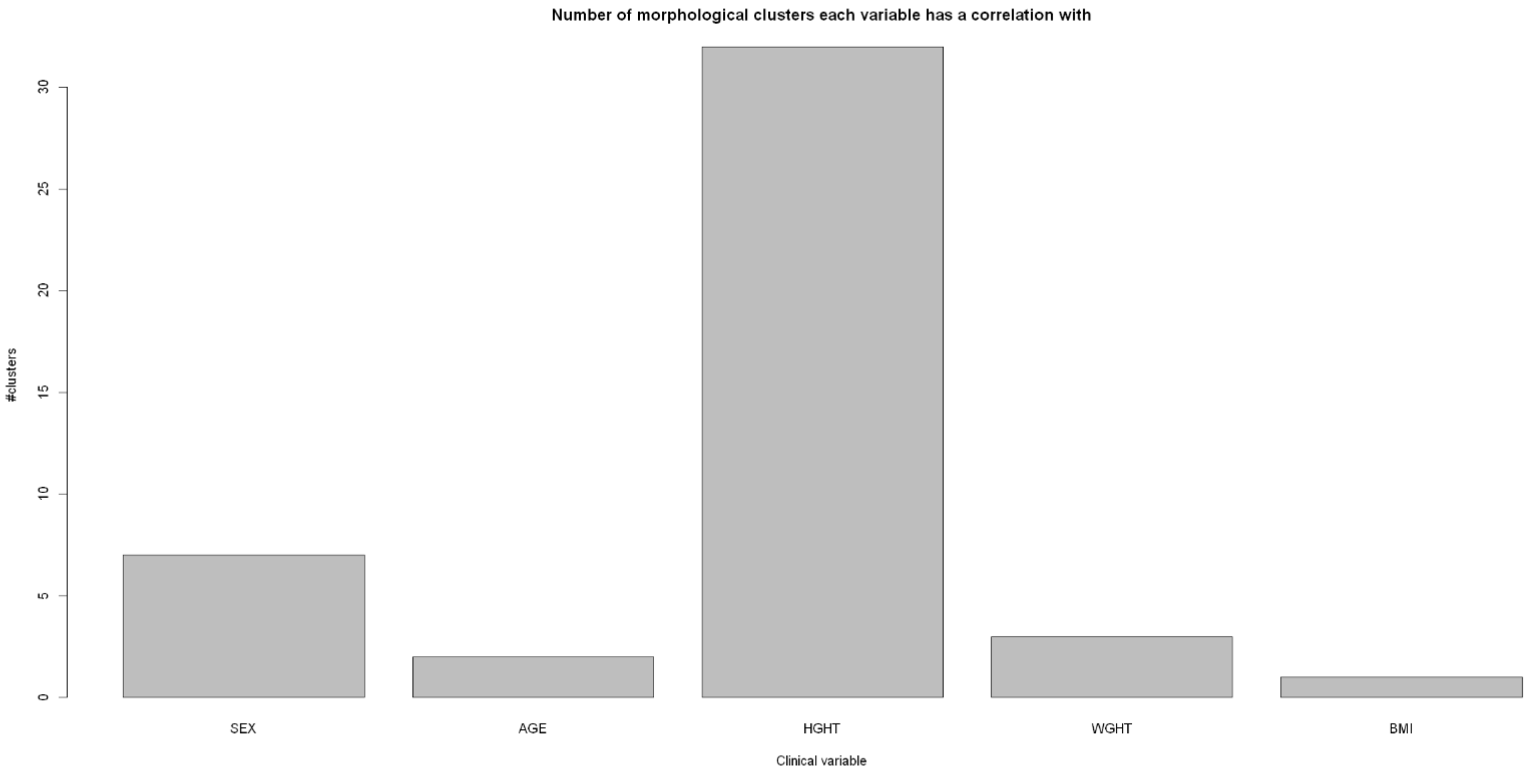


Figure 3.8: Number of morphological clusters each variable (controlled for its confounding variables has a correlation with

4

Associations between morphology and gene expression

4.1 Introduction

The project's overall goal is to compare the morphological description of tissues, and thyroid glands, to their transcriptome. The expression of morphological cluster are provided by the dataset ***mc :morphological-counts.tsv.*** The expression of genes across the samples are provided by the dataset ***rna :RNA-read-counts.tsv***.

All the dataframes mentioned in chapter 4 of the report are available on the jupyter notebook under the title question3.

In the third part of our project, we use DESeq2 to perform a differential expression analysis to detect deferentially expressed genes among our samples associated with each morphological cluster. We are provided with gene sets from subsection C2 of the MSigDB database. DESseq2 performs statistical tests based on negative binomial distributions that have notably reasonable control of false-positive errors with comparable specificity and sensitivity resulting from the tests. [**4**].

We use the package fgsea to perform fast preranked gene set enrichment analysis with the gene ranking obtained by the differential analyses as input. The enrichment scores will reflect the degree to which our gene sets are over represented at the top or bottom of our ranked list of genes [**5**].

4.2 Significant down-regulated and up-regulated genes associated with each morphological cluster

To obtain the significant down-regulated and up-regulated genes associated with each morphological cluster, we perform a differential expression analysis using DESeq2. The *rna* dataset (expression of genes across the samples) is used as the count matrix. The sample information table is provided with the dataset *mc*. DESeq2 design represents how to model the sample and will be provided by the morphological cluster column of *mc* for which we perform the differential expression analysis [**6**].

The *mc* dataset is the count matrix of the morphologies identified by the AI, and it gives the expression of the morphological clusters. The columns are the morphological clusters indexed from 0 to 63, and the rows are the samples indexed by SMPLID. For each sample, the cluster count (number of tiles in each morphological cluster) is directly related to image size, which depends on how the pathologist cut the organ as much as its actual volume. The raw counts for each sample were replaced by the morphological cluster proportions to review the number of genes associated with each morphological cluster.

The *rna* dataset gives the expression of genes across the samples. The transcript *ENSEMBL IDs* are the rownames, and the columns are composed of the *GTEX samples id.* It is un-normalized as DESeq2 expects count data obtained (from RNA-seq) [**6**]. Filtering was performed to reduce the number of transcripts. As suggested, the transcripts not or little expressed in the thyroid are not needed, and the focus was made on the transcripts showing high variability. The filtering was made by selecting only the genes in the top quartile of expression. As a result, the number of transcripts dropped from 56200 to 14050.

To perform the differential expression analysis for each cluster, we created an empty dataframe *cluster_table* for our results. The rows were set as the *transcript id* and the columns as the *log2 fold changes* values for each morphological clusters. A conditional function was used to fill the dataframe with the results obtained for each cluster. There was no missing value in the dataframe.

A similar method was used to perform each cluster's fast pre-ranked gene set enrichment analysis. The results were extracted from a result table generated using the function results, which extracts a results table with log2 fold changes, p values, and adjusted p values [**6**].

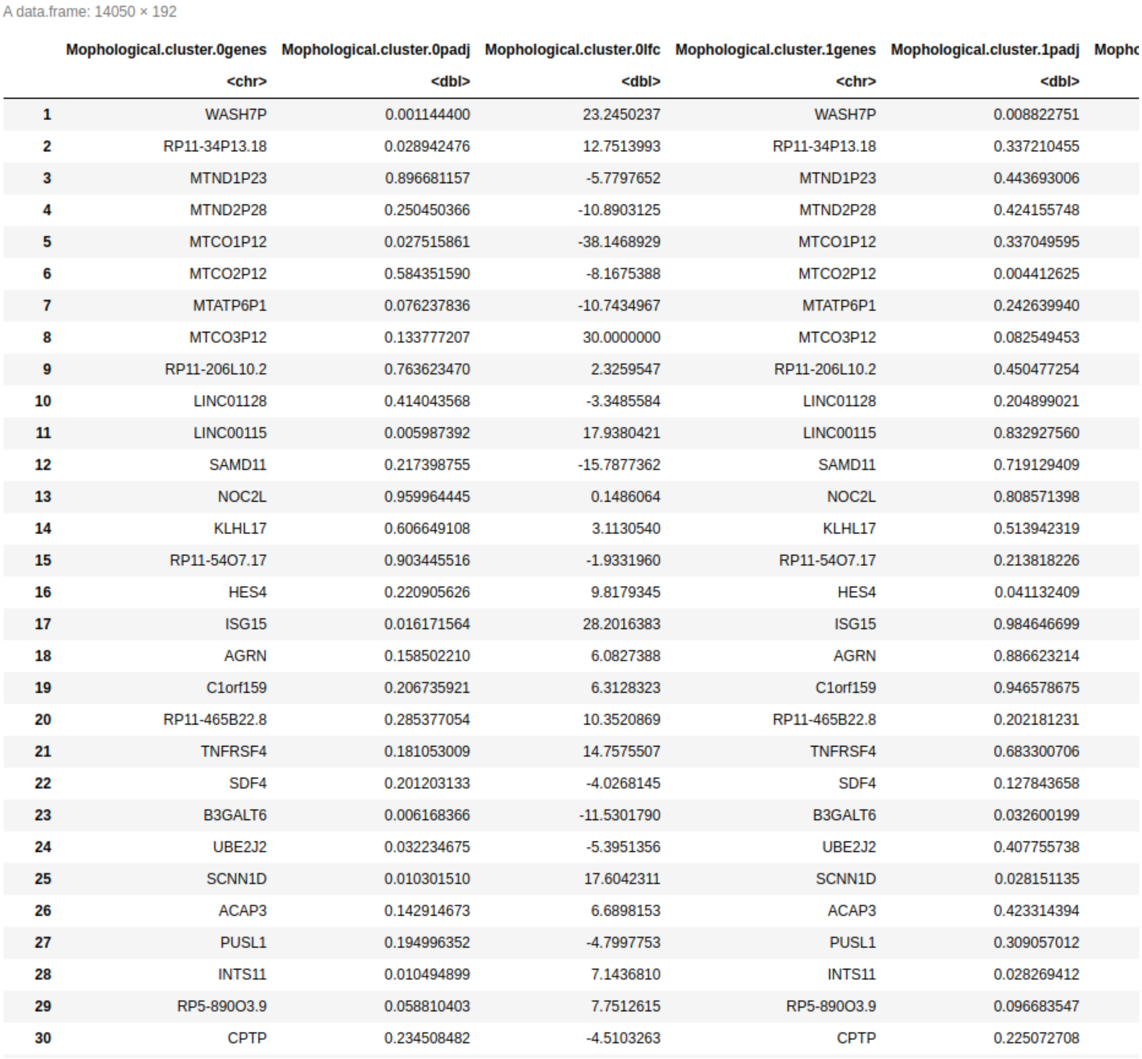
To obtain the ranks to perform the fgsea analysis of the gene sets, we initially focused on obtaining only the log2 fold changes values for each transcript. We chose to use the log 2 fold change value (LFC) to rank the genes associated with the transcripts. The log 2 fold changes are the log ratio of the observed gene expression level compared to different conditions. The condition (design) is set to the morphological cluster proportions (adjusted raw cluster counts) for which were are conducting the differential expression analysis. The p-value was disregarded as a ranking criteria because small changes, even if statistically highly significant, might not be the most interesting candidates for further investigation. The variance of LFC estimates for genes with low read count is usually strong. [**7**] DESeq2 allows for the shrinkage of the LFC estimates toward zero when the information for a gene is low. [**6**] With the use of the LFC shrinkage, the number of genes called significantly differentially expressed depends as much on the sample size and other aspects of experimental design as it does on the biology of the experiment. [**7**] The differential expression analysis for each cluster was not performed with the LFC shrinkage. For most bulk RNA-seq experiments, the LFC shrinkage did not affect statistical testing. Specific sequencing datasets show better performance with the testing separated from using the LFC prior. [**6**]. A comparison of the gene's ranking obtained with the differential expression analysis performed by DESeq2 with and without shrinkage could be interesting in the future.

To obtain the number of significant up-regulated and down-regulated genes associated with each cluster, a new empty dataframe *cluster_table_padj* was created to contain both the adjusted p-value and the log 2 fold changes values of each transcript for each cluster. The rownames are still defined by the *transcript ENSEMBL IDs.* The genes with an adjusted p-value below 0.05 and log2 fold change greater than one were selected as significantly up-regulated genes. The genes with an adjusted p-value below 0.05 and a log2 fold change lower than one were selected as significant down-regulated genes. [**8**]

The number of significant up-regulated and down-regulated genes for each cluster was reported on the dataframe *significant_1.* The top ten up-regulated genes are reported on the dataframe *significant_2.* When performing the differential expression analysis, we defined the rownames of the dataset *rna* by the *transcript id.* We reported the corresponding gene symbols to generate the list of the up-regulated and down-regulated genes.

The morphological cluster 14 corresponds to muscular tissues. When sampling out thyroid tissue, surgeons sometimes take muscle adjacent to the thyroid. Cluster 14 was used as a continuous covariate. When using a continuous covariate with DESeq2, the average effect of each group is controlled for, regardless of the trend over the continuous covariate. [**6**]

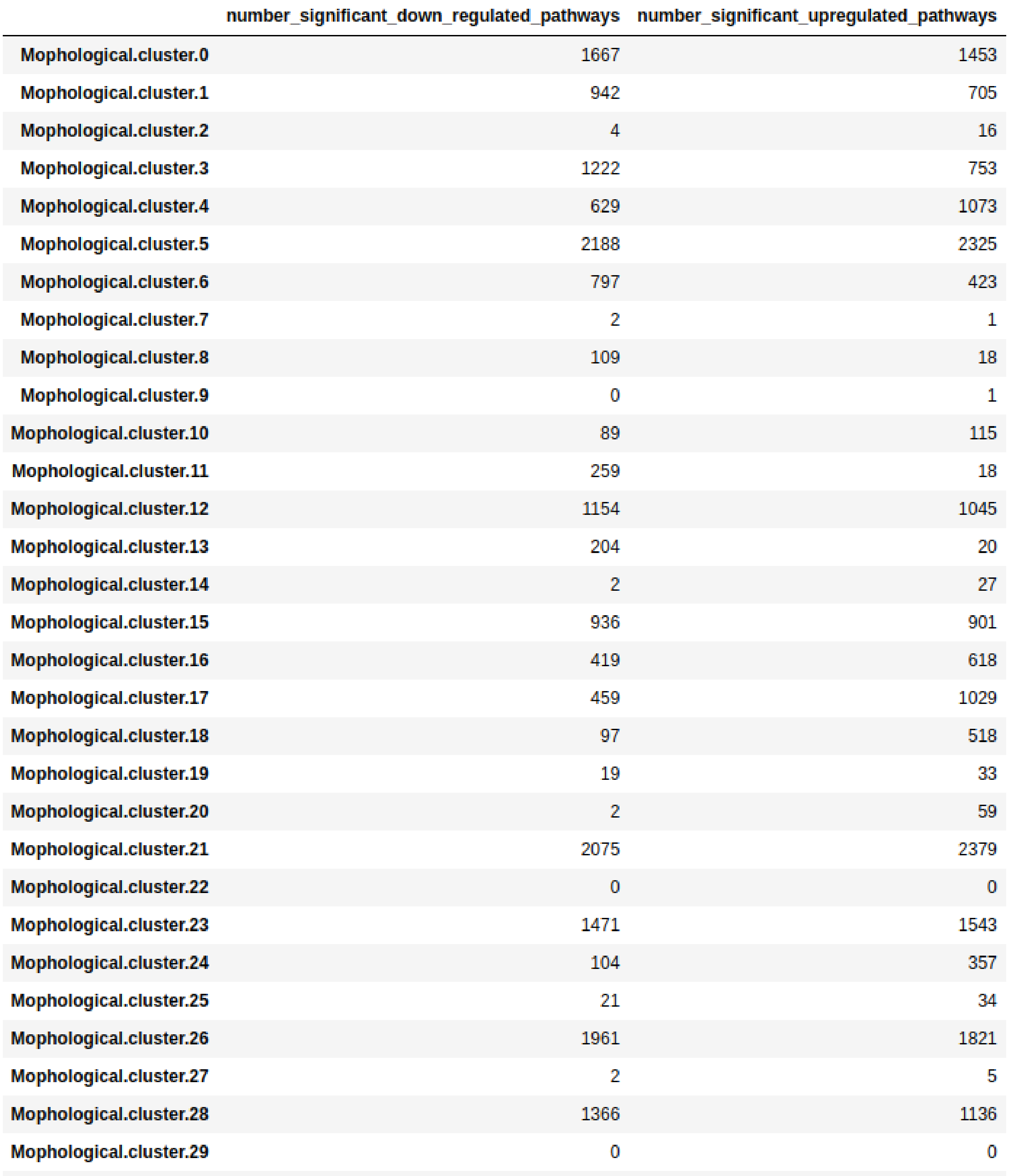
The differential expression analysis was performed again with the addition of cluster 14 as a continuous covariate, and it was added with each cluster as the design for DESeq2 analysis. As expected, we observed that the number of significant up-regulated and down-regulated genes for each cluster was slightly smaller when adding the covariate. The log 2 fold changes and adjusted p values are reported on the dataframe *cluster_table_cnd.* An overview of the table is shown in figure 4.1. The number of significant up-regulated and down-regulated genes associated with each cluster is listed in the dataframe *significantcnd_1.* Figure 4.2 displays the number of significant up-regulated and down-regulated genes associated with the cluster 0 to 29. The top ten significant up-regulated genes are listed in the dataframe *significantcnd_2* and displayed in table 4.3 for the cluster 0 to 20. The top ten most-upregulated genes seemed unchanged compared to our prior results.



A data.frame: 14050 × 192

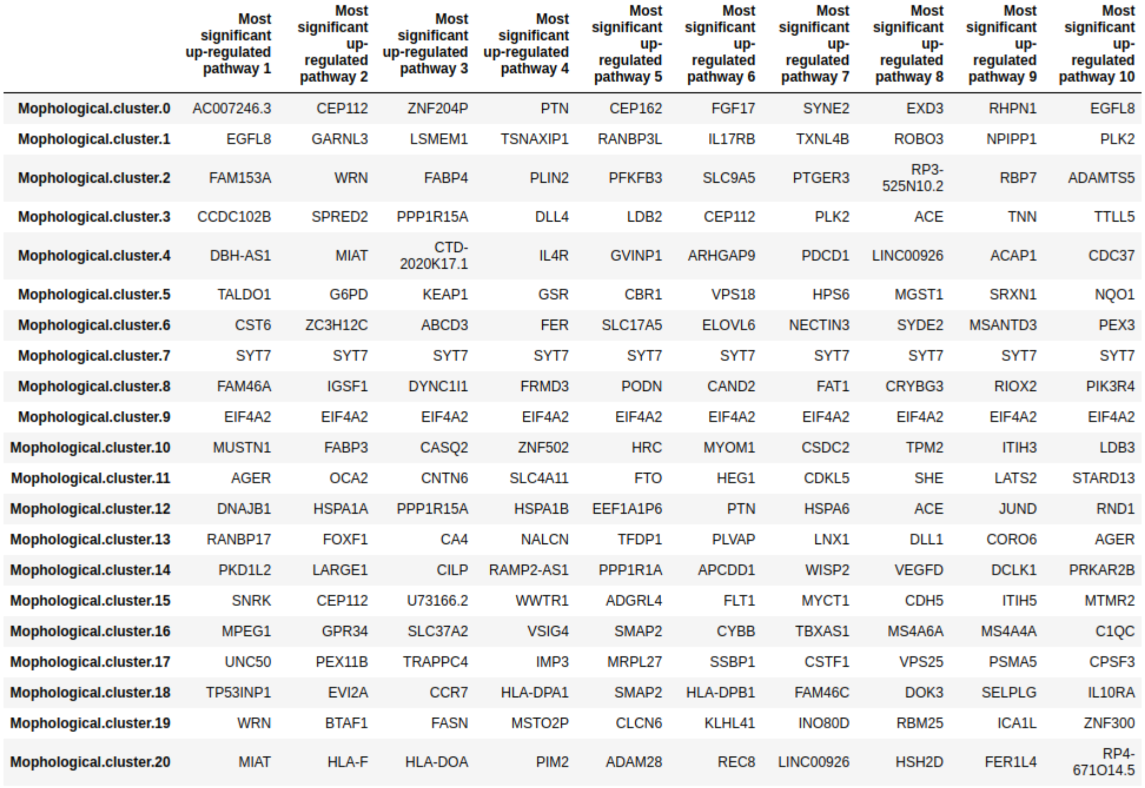
	Mophological.cluster.0genes	Mophological.cluster.0padj	Mophological.cluster.0lfc	Mophological.cluster.1genes	Mophological.cluster.1padj	Mophc
	<chr>	<dbl>	<dbl>	<chr>	<dbl>	
1	WASH7P	0.001144400	23.2450237	WASH7P	0.008822751	
2	RP11-34P13.18	0.028942476	12.7513993	RP11-34P13.18	0.337210455	
3	MTND1P23	0.896681157	-5.7797652	MTND1P23	0.443693006	
4	MTND2P28	0.250450366	-10.8903125	MTND2P28	0.424155748	
5	MTCO1P12	0.027515861	-38.1468929	MTCO1P12	0.337049595	
6	MTCO2P12	0.584351590	-8.1675388	MTCO2P12	0.004412625	
7	MTATP6P1	0.076237836	-10.7434967	MTATP6P1	0.242639940	
8	MTCO3P12	0.133777207	30.0000000	MTCO3P12	0.082549453	
9	RP11-206L10.2	0.763623470	2.3259547	RP11-206L10.2	0.450477254	
10	LINC01128	0.414043568	-3.3485584	LINC01128	0.204899021	
11	LINC00115	0.005987392	17.9380421	LINC00115	0.832927560	
12	SAMD11	0.217398755	-15.7877362	SAMD11	0.719129409	
13	NOC2L	0.959964445	0.1486064	NOC2L	0.808571398	
14	KLHL17	0.606649108	3.1130540	KLHL17	0.513942319	
15	RP11-54O7.17	0.903445516	-1.9331960	RP11-54O7.17	0.213818226	
16	HES4	0.220905626	9.8179345	HES4	0.041132409	
17	ISG15	0.016171564	28.2016383	ISG15	0.984646699	
18	AGRN	0.158502210	6.0827388	AGRN	0.886623214	
19	C1orf159	0.206735921	6.3128323	C1orf159	0.946578675	
20	RP11-465B22.8	0.285377054	10.3520869	RP11-465B22.8	0.202181231	
21	TNFRSF4	0.181053009	14.7575507	TNFRSF4	0.683300706	
22	SDF4	0.201203133	-4.0268145	SDF4	0.127843658	
23	B3GALT6	0.006168366	-11.5301790	B3GALT6	0.032600199	
24	UBE2J2	0.032234675	-5.3951356	UBE2J2	0.407755738	
25	SCNN1D	0.010301510	17.6042311	SCNN1D	0.028151135	
26	ACAP3	0.142914673	6.6898153	ACAP3	0.423314394	
27	PUSL1	0.194996352	-4.7997753	PUSL1	0.309057012	
28	INTS11	0.010494899	7.1436810	INTS11	0.028269412	
29	RP5-890O3.9	0.058810403	7.7512615	RP5-890O3.9	0.096683547	
30	CPTP	0.234508482	-4.5103263	CPTP	0.225072708	

Figure 4.1: head of cluster_table_cnd



	number_significant_down_regulated_pathways	number_significant_upregulated_pathways
Mophological.cluster.0	1667	1453
Mophological.cluster.1	942	705
Mophological.cluster.2	4	16
Mophological.cluster.3	1222	753
Mophological.cluster.4	629	1073
Mophological.cluster.5	2188	2325
Mophological.cluster.6	797	423
Mophological.cluster.7	2	1
Mophological.cluster.8	109	18
Mophological.cluster.9	0	1
Mophological.cluster.10	89	115
Mophological.cluster.11	259	18
Mophological.cluster.12	1154	1045
Mophological.cluster.13	204	20
Mophological.cluster.14	2	27
Mophological.cluster.15	936	901
Mophological.cluster.16	419	618
Mophological.cluster.17	459	1029
Mophological.cluster.18	97	518
Mophological.cluster.19	19	33
Mophological.cluster.20	2	59
Mophological.cluster.21	2075	2379
Mophological.cluster.22	0	0
Mophological.cluster.23	1471	1543
Mophological.cluster.24	104	357
Mophological.cluster.25	21	34
Mophological.cluster.26	1961	1821
Mophological.cluster.27	2	5
Mophological.cluster.28	1366	1136
Mophological.cluster.29	0	0

Figure 4.2: Number of significant up-regulated and down-regulated genes associated with each cluster (cluster 0 to 29)



	Most significant up-regulated pathway 1	Most significant up-regulated pathway 2	Most significant up-regulated pathway 3	Most significant up-regulated pathway 4	Most significant up-regulated pathway 5	Most significant up-regulated pathway 6	Most significant up-regulated pathway 7	Most significant up-regulated pathway 8	Most significant up-regulated pathway 9	Most significant up-regulated pathway 10
Mophological.cluster.0	AC007246.3	CEP112	ZNF204P	PTN	CEP162	FGF17	SYNE2	EXD3	RHPN1	EGFL8
Mophological.cluster.1	EGFL8	GARNL3	LSMEM1	TSNAXIP1	RANBP3L	IL17RB	TXNL4B	ROBO3	NPIPP1	PLK2
Mophological.cluster.2	FAM153A	WRN	FABP4	PLIN2	PFKFB3	SLC9A5	PTGER3	RP3-525N10.2	RBP7	ADAMTS5
Mophological.cluster.3	CCDC102B	SPRED2	PPP1R15A	DLL4	LDB2	CEP112	PLK2	ACE	TNN	TTLL5
Mophological.cluster.4	DBH-AS1	MIAT	CTD-2020K17.1	IL4R	GVINP1	ARHGAP9	PDCD1	LINC00926	ACAP1	CDC37
Mophological.cluster.5	TALDO1	G6PD	KEAP1	GSR	CBR1	VPS18	HPS6	MGST1	SRXN1	NQO1
Mophological.cluster.6	CST6	ZC3H12C	ABCD3	FER	SLC17A5	ELOVL6	NECTIN3	SYDE2	MSANTD3	PEX3
Mophological.cluster.7	SYT7	SYT7	SYT7	SYT7	SYT7	SYT7	SYT7	SYT7	SYT7	SYT7
Mophological.cluster.8	FAM46A	IGSF1	DYNC1I1	FRMD3	PODN	CAND2	FAT1	CRYBG3	RIOX2	PIK3R4
Mophological.cluster.9	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2	EIF4A2
Mophological.cluster.10	MUSTN1	FABP3	CASQ2	ZNF502	HRC	MYOM1	CSDC2	TPM2	ITIH3	LDB3
Mophological.cluster.11	AGER	OCA2	CNTN6	SLC4A11	FTO	HEG1	CDKL5	SHE	LATS2	STARD13
Mophological.cluster.12	DNAJB1	HSPA1A	PPP1R15A	HSPA1B	EEF1A1P6	PTN	HSPA6	ACE	JUND	RND1
Mophological.cluster.13	RANBP17	FOXF1	CA4	NALCN	TFDP1	PLVAP	LNX1	DLL1	CORO6	AGER
Mophological.cluster.14	PKD1L2	LARGE1	CILP	RAMP2-AS1	PPP1R1A	APCDD1	WISP2	VEGFD	DCLK1	PRKAR2B
Mophological.cluster.15	SNRK	CEP112	U73166.2	WWTR1	ADGRL4	FLT1	MYCT1	CDH5	ITIH5	MTMR2
Mophological.cluster.16	MPEG1	GPR34	SLC37A2	VSIG4	SMAP2	CYBB	TBXAS1	MS4A6A	MS4A4A	C1QC
Mophological.cluster.17	UNC50	PEX11B	TRAPPC4	IMP3	MRPL27	SSBP1	CSTF1	VPS25	PSMA5	CPSF3
Mophological.cluster.18	TP53INP1	EVI2A	CCR7	HLA-DPA1	SMAP2	HLA-DPB1	FAM46C	DOK3	SELPLG	IL10RA
Mophological.cluster.19	WRN	BTAF1	FASN	MSTO2P	CLCN6	KLHL41	INO80D	RBM25	ICA1L	ZNF300
Mophological.cluster.20	MIAT	HLA-F	HLA-DOA	PIM2	ADAM28	REC8	LINC00926	HSH2D	FER1L4	RP4-671O14.5

Figure 4.3: Top 10 significant up-regulated genes for the cluster 0 to 20

4.3 Reactome gene sets

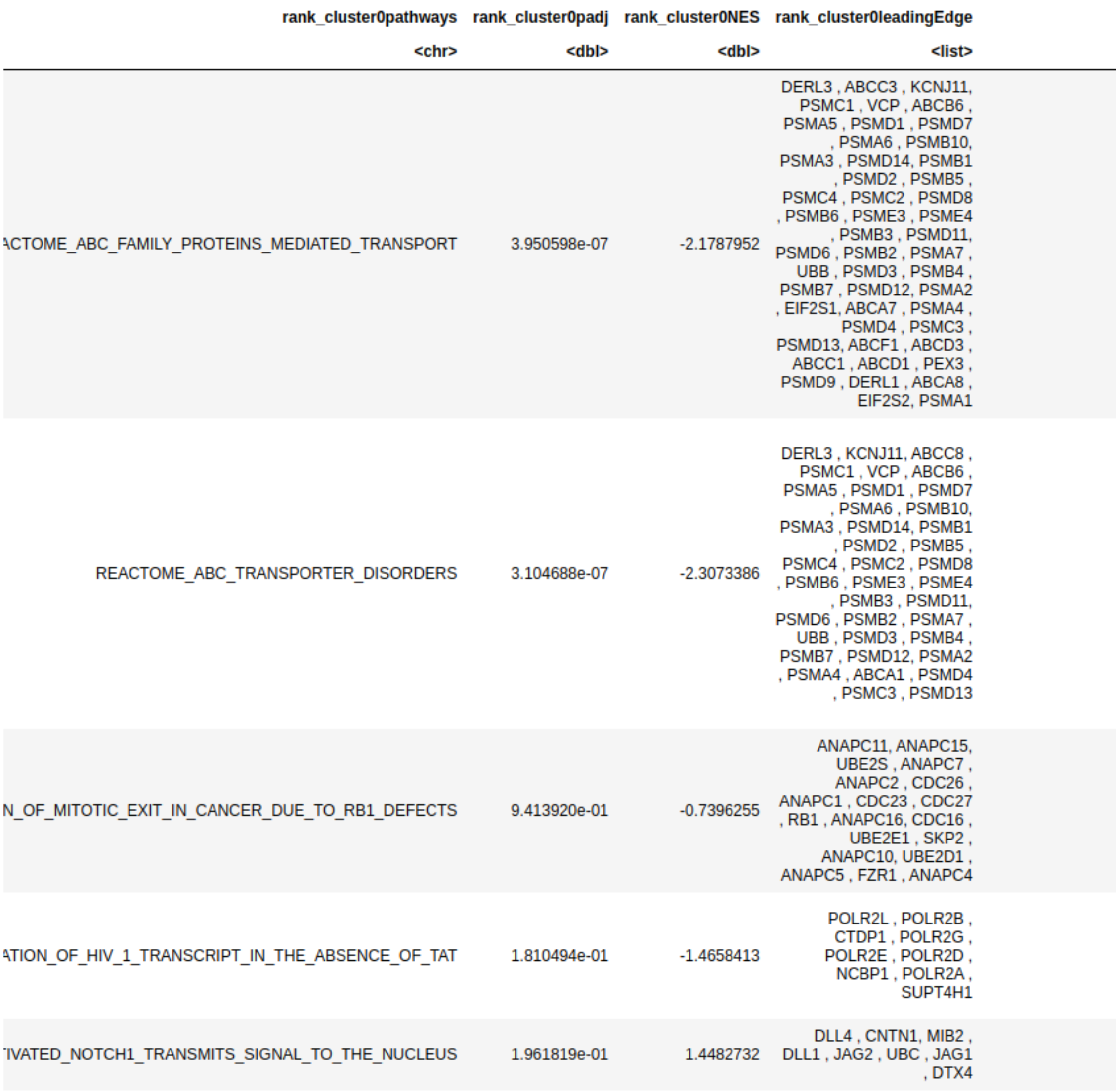
As previously mentioned, we are provided with gene sets from subsection C2 of the MSigDB database and use the package fgsea to perform a fast pre-ranked gene set enrichment analysis with the gene ranking obtained by the differential analyses as input. The enrichment scores will reflect the degree to which our gene sets are overrepresented at the top or bottom of our ranked list of genes [**5**].

The fgsea function gmtPathways was used to read and load the gene sets. The provided Reactome was copied directly to the folder extdata of the fgsea package. The fgsea package provides example pathways and example ranks to allow the first use of the function with example data. We used both files to determine and confirm the type and class of our input data. We used the information gathered in the dataframe *cluster_table_cnd* to generate the ranking of the genes for each cluster. The genes were sorted by decreasing value of the log 2 fold changes, and it was not necessary to provide ranked inputs.

As before, we create an empty dataframe *fgsea_table* to centralise our results for each cluster. There are four columns for each cluster, containing the pathway's name, adjusted p-value (padj), the normalized enrichment score (NES), and the leading edge (list of genes that contributed to the enrichment score). As the number of obtained results differed between the cluster, we set an initial dataframe having an excessive number of rows and later removed the rows containing only NA values. The reduced dataframe is called *fgsea_table_reduced.* The head of the dataframe is displayed in figure 4.4.

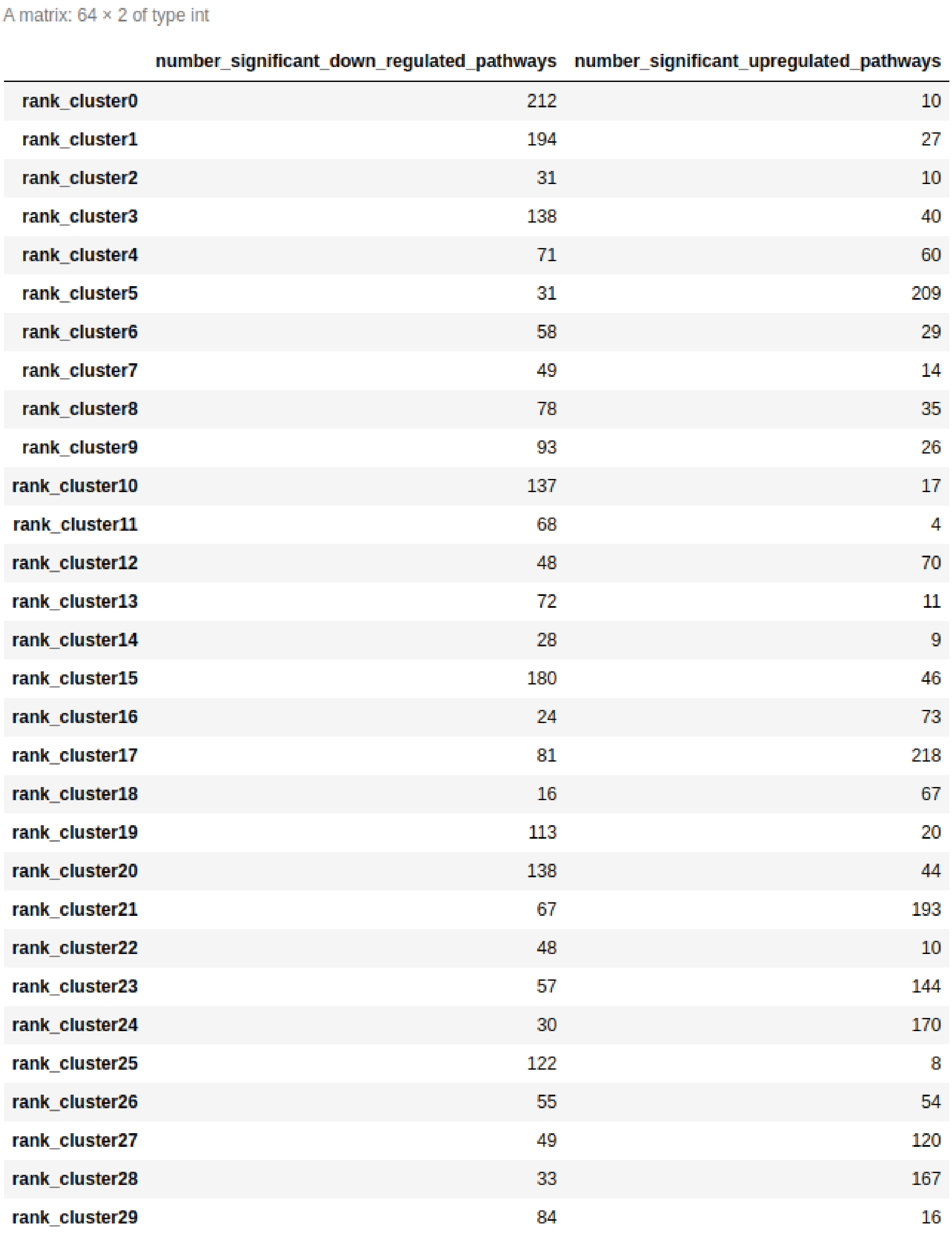
The enrichment score is computed by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero encountered in walking the list. A positive ES indicates gene set enrichment at the top of the ranked list; a negative ES indicates gene set enrichment at the bottom of the ranked list.[**8**].

The significant pathways enriched at the top and bottom of the ranked list are determined by their adjusted p-value and enrichment score. After setting a threshold at an adjusted p-value inferior to 0.05, the significant pathways enriched at the top of the list are selected as the pathways with an enrichment score greater than 0. In comparison, the pathways enriched at the bottom of the list are selected as the pathways with an enrichment score lower than 0. Figure 4.5 regroups the number of significant pathways enriched at the top and bottom of the list associated with each cluster (clusters 0 to 29). Figure 4.6 list the top pathway enriched at the top of the list for cluster 0 to 29. The complete dataframes are available on the Jupiter notebook.



rank_cluster0pathways	rank_cluster0padj	rank_cluster0NES	rank_cluster0leadingEdge
<chr>	<dbl>	<dbl>	<list>
ACTOME_ABC_FAMILY_PROTEINS_MEDIATED_TRANSPORT	3.950598e-07	-2.1787952	DERL3 , ABCC3 , KCNJ11, PSMC1 , VCP , ABCB6 , PSMA5 , PSMD1 , PSMD7 , PSMA6 , PSMB10, PSMA3 , PSMD14, PSMB1 , PSMD2 , PSMB5 , PSMC4 , PSMC2 , PSMD8 , PSMB6 , PSME3 , PSME4 , PSMB3 , PSMD11, PSMD6 , PSMB2 , PSMA7 , UBB , PSMD3 , PSMB4 , PSMB7 , PSMD12, PSMA2 , EIF2S1, ABCA7 , PSMA4 , PSMD4 , PSMC3 , PSMD13, ABCF1 , ABCD3 , ABCC1 , ABCD1 , PEX3 , PSMD9 , DERL1 , ABCA8 , EIF2S2, PSMA1
REACTOME_ABC_TRANSPORTER_DISORDERS	3.104688e-07	-2.3073386	DERL3 , KCNJ11, ABCC8 , PSMC1 , VCP , ABCB6 , PSMA5 , PSMD1 , PSMD7 , PSMA6 , PSMB10, PSMA3 , PSMD14, PSMB1 , PSMD2 , PSMB5 , PSMC4 , PSMC2 , PSMD8 , PSMB6 , PSME3 , PSME4 , PSMB3 , PSMD11, PSMD6 , PSMB2 , PSMA7 , UBB , PSMD3 , PSMB4 , PSMB7 , PSMD12, PSMA2 , PSMA4 , ABCA1 , PSMD4 , PSMC3 , PSMD13
N_OF_MITOTIC_EXIT_IN_CANCER_DUE_TO_RB1_DEFECTS	9.413920e-01	-0.7396255	ANAPC11, ANAPC15, UBE2S , ANAPC7 , ANAPC2 , CDC26 , ANAPC1 , CDC23 , CDC27 , RB1 , ANAPC16, CDC16 , UBE2E1 , SKP2 , ANAPC10, UBE2D1 , ANAPC5 , FZR1 , ANAPC4
ATION_OF_HIV_1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	1.810494e-01	-1.4658413	POLR2L , POLR2B , CTDP1 , POLR2G , POLR2E , POLR2D , NCBP1 , POLR2A , SUPT4H1
IVATED_NOTCH1_TRANSMITS_SIGNAL_TO_THE_NUCLEUS	1.961819e-01	1.4482732	DLL4 , CNTN1, MIB2 , DLL1 , JAG2 , UBC , JAG1 , DTX4

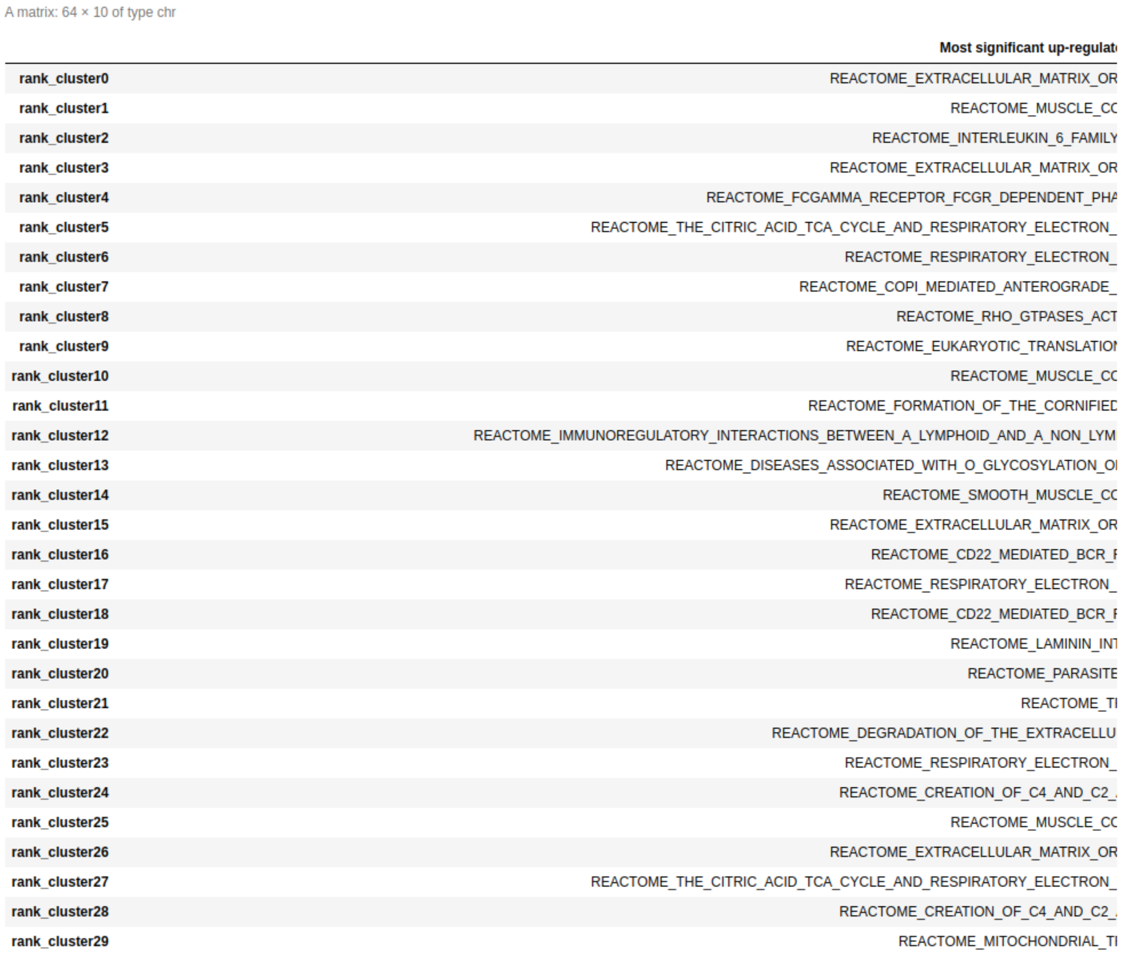
Figure 4.4: head of fgsea_table_reduced



A matrix: 64 × 2 of type int

	number_significant_down_regulated_pathways	number_significant_upregulated_pathways
rank_cluster0	212	10
rank_cluster1	194	27
rank_cluster2	31	10
rank_cluster3	138	40
rank_cluster4	71	60
rank_cluster5	31	209
rank_cluster6	58	29
rank_cluster7	49	14
rank_cluster8	78	35
rank_cluster9	93	26
rank_cluster10	137	17
rank_cluster11	68	4
rank_cluster12	48	70
rank_cluster13	72	11
rank_cluster14	28	9
rank_cluster15	180	46
rank_cluster16	24	73
rank_cluster17	81	218
rank_cluster18	16	67
rank_cluster19	113	20
rank_cluster20	138	44
rank_cluster21	67	193
rank_cluster22	48	10
rank_cluster23	57	144
rank_cluster24	30	170
rank_cluster25	122	8
rank_cluster26	55	54
rank_cluster27	49	120
rank_cluster28	33	167
rank_cluster29	84	16

Figure 4.5: Number of significant pathways enriched at the top and bottom of the list associated with each cluster (cluster 0 to 29)



A matrix: 64 × 10 of type chr

	Most significant up-regulat
rank_cluster0	REACTOME_EXTRACELLULAR_MATRIX_OR
rank_cluster1	REACTOME_MUSCLE_CC
rank_cluster2	REACTOME_INTERLEUKIN_6_FAMILY
rank_cluster3	REACTOME_EXTRACELLULAR_MATRIX_OR
rank_cluster4	REACTOME_FCGAMMA_RECEPTOR_FCGR_DEPENDENT_PHA
rank_cluster5	REACTOME_THE_CITRIC_ACID_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_
rank_cluster6	REACTOME_RESPIRATORY_ELECTRON_
rank_cluster7	REACTOME_COPI_MEDIATED_ANTEROGRADE_
rank_cluster8	REACTOME_RHO_GTPASES_ACT
rank_cluster9	REACTOME_EUKARYOTIC_TRANSLATION
rank_cluster10	REACTOME_MUSCLE_CC
rank_cluster11	REACTOME_FORMATION_OF_THE_CORNIFIED
rank_cluster12	REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYM
rank_cluster13	REACTOME_DISEASES_ASSOCIATED_WITH_O_GLYCOSYLATION_OI
rank_cluster14	REACTOME_SMOOTH_MUSCLE_CC
rank_cluster15	REACTOME_EXTRACELLULAR_MATRIX_OR
rank_cluster16	REACTOME_CD22_MEDIATED_BCR_I
rank_cluster17	REACTOME_RESPIRATORY_ELECTRON_
rank_cluster18	REACTOME_CD22_MEDIATED_BCR_I
rank_cluster19	REACTOME_LAMININ_INT
rank_cluster20	REACTOME_PARASITE
rank_cluster21	REACTOME_TI
rank_cluster22	REACTOME_DEGRADATION_OF_THE_EXTRACELLU
rank_cluster23	REACTOME_RESPIRATORY_ELECTRON_
rank_cluster24	REACTOME_CREATION_OF_C4_AND_C2_
rank_cluster25	REACTOME_MUSCLE_CC
rank_cluster26	REACTOME_EXTRACELLULAR_MATRIX_OR
rank_cluster27	REACTOME_THE_CITRIC_ACID_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_
rank_cluster28	REACTOME_CREATION_OF_C4_AND_C2_
rank_cluster29	REACTOME_MITOCHONDRIAL_TI

Figure 4.6: Top pathway enriched at the top of the list for the cluster 0 to 29

4.4 Discussion of the results, technically and biologically

The morphological cluster 58 contains tissue edges and various artifacts, and the morphological cluster 45 contains dense lymphocyte aggregates (aggregation of lymphocytic cells in the stroma). Both clusters can be used as a positive and negative control for our analysis. Cluster 58 contains 0 significant up-regulated genes and one significant down-regulated gene, and cluster 45 contains 1825 significant up-regulated genes and 1878 significant down-regulated genes. These results were expected considering the nature of the morphological clusters. By examining the sample's morphological cluster proportion associated with cluster 58, we observe that most of the proportions are below 0.1%. That means most samples have a low number of tiles in cluster 58. Since cluster 58 contains tissue edges and artifacts, we expected the cluster to have a low correlation with gene expression. In comparison, the opposite reasoning can be done for cluster 45.

The morphological variation was computed across 893 thyroid slices from GTEx and rests on unsupervised deep learning models. To examine the correlation of these morphological variations with gene expression, differential expression analysis was computed with DESseq2 to rank the genes associated with each morphological cluster and extract statistically and differentially expressed genes among our samples associated with each morphological cluster.

Fast Gene Set Enrichment Analysis methods were performed with fgsea to attribute an enrichment score to each pathway. The score depends on the ranked gene's correlation with the phenotype. The results can be verified by using the list of the top 10 pathways enriched at the top of the list associated with cluster 58. The Aggregation of lymphocytes is often an outcome of the activation of various signaling pathways and is associated with fundamental cellular responses such as the establishment of immune synapses [**9**]. The top three pathways associated with cluster 58 enriched at the top of the list are pathways involved in immune response (REACTOME_FCERI_MEDIATED_NF_KB_ACTIVATION, REACTOME_PARASITE_INFECTION, REACTOME_FCGAMMA_RECEPTOR_FCGR_DEPENDENT_PHAGOCYTOSIS).

One of the project's objectives is to survey the diversity of normal thyroid morphologies. All our samples are from asymptomatic thyroids. To estimate the prevalence of inflammation among our samples, we can review the pathways associated with inflammation present in our gene sets and their enrichment score. We only found one significantly enriched pathway associated with inflammation in cluster 1 (REACTOME_ADORA2B_MEDIATED_ANTI_INFLAMMATORY_CYTOKINES_PRODUCTION). There were four other pathways involved in inflammation present in the gene set but not significantly enriched (REACTOME_CLEC7A_INFLAMMASOME_PATHWAY, REACTOME_DEX_H_BOX_HELICASES_ACTIVATE_TYPE_I _IFN_AND_INFLAMMATORY_CYTOKINES_PRODUCTION, REACTOME_INFLAMMASOMES, REACTOME _THE_NLRP3_INFLAMMASOME) [**10**]. With this information, we can assume that the prevalence of inflammation in the collection of asymptomatic thyroids is low.

Bibliography

[1] Jegerlehner S, Bulliard J-L, Aujesky D, Rodondi N, Germann S, Konzelmann I, et al. (2017) Overdiagnosis and overtreatment of thyroid cancer: A population-based temporal trend study. PLoS ONE 12(6): e0179387. https://doi.org/10.1371/journal.pone.017938

[2] Carithers LJ, Ardlie K, Barcus M, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank. 2015;13(5):311-319. doi:10.1089/bio.2015.0032

[3] Michael I. Love, Simon Anders, and Wolfgang Huber. (2022) Analyzing RNA-seq data with DESeq2. Last consulted : June 9, 2022. Url : http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

[4] Anjum, Arfa, Seema Jaggi, Eldho Varghese, Shwetank Lall, Arpan Bhowmik, and Anil Rai. “Identification of Differentially Expressed Genes in RNA-Seq Data of Arabidopsis Thaliana: A Compound Distribution Approach.” Journal of Computational Biology 23, no. 4 (April 1, 2016): 239–47. https://doi.org/10.1089/cmb.2015.0205.

[5] “GSEA User Guide.” Accessed June 9, 2022.
https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html.

[6] Love, Michael, Simon Anders, and Wolfgang Huber. “Analyzing RNA-Seq Data with DESeq2.” Bioconductor, n.d.

[7] Love, Michael I., Wolfgang Huber, and Simon Anders. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology 15, no. 12 (December 5, 2014): 550. https://doi.org/10.1186/s13059-014-0550-8.

[8] ang, Lifang, Hanye Wang, Panpan Wang, Mingju Gao, Luqi Huang, Xiuming Cui, and Yuan Liu. “De Novo and Comparative Transcriptomic Analysis Explain Morphological Differences in Panax Notoginseng Taproots.” BMC Genomics 23 (January 31, 2022): 86. https://doi.org/10.1186/s12864-021-08283-w.

[8] Dezorella, Nili, Sigi Kay, Shoshana Baron, Mika Shapiro, Ziv Porat, Varda Deutsch, Yair Herishanu, and Ben-Zion Katz. "Measurement of Lymphocyte Aggregation by Flow Cytometry-Physiological Implications in Chronic Lymphocytic Leukemia: Cell Aggregation Measured by Flow Cytometry." Cytometry Part B: Clinical Cytometry 90, no. 3 (May 2016): 257–66. https://doi.org/10.1002/cyto.b.21263.

[9] Cassel, Suzanne L., Sophie Joly, and Fayyaz S. Sutterwala. "The NLRP3 Inflammasome: A Sensor of Immune Danger Signals." Seminars in Immunology 21, no. 4 (August 2009): 194–98. https://doi.org/10.1016/j.smim.2009.05.002.