Quantification of human sexual dimorphism with the dimorphism index in subcutaneous adipose tissue

Youmna Ayadi Kevin Straatman

2025-06-22

Table of contents

1 Introduction

Many animal species are subject to sexual dimorphism. In humans, sexual dimorphism can be expressed as observable differences in physical traits between male and female individuals such as individual height,

breast tissue development, muscular mass or hip-to-waist ratio. Furthermore, these dimorphic traits differences were extended to organs and health observations such as disease incidence or treatment response in recent years [1], leading to new considerations aligned with a more personalized medicine. Thus, health and clinical implications from sexual dimorphism knowledge cannot be underestimated.

Among all organs, the adipose tissue presents interesting multi-faceted metabolic, endocrinologic and immune roles with important medical consequences that were shown to differ between men and women. For example, differences in fat tissue distribution (under hormone control) between men and women have been shown with men storing more visceral fat than women, which is a known risk factor for cardiovascular diseases, type 2 diabetes, cancers and other chronic diseases [2]. This justifies the attention of this project for the adipose tissue. But what methods to use?

With the advent of new computing technologies and the recent rise of more data-driven approaches to study biological systems, researchers are now able to use more quantitative and less biased approaches to study gene expression in organisms with the help of new developped analysis tools and big database projects such as GTEx (Genotype-Tissue Expression) [3].

These new tools and approaches have led to the observation of a continuous sexual dimorphism spectrum within each sex, further complexifying the dimorphism problem, with now each human (and each of its organs) being more or less sexually dimorphic on the male or female end of this spectrum, as opposed to the previous strictly binary categorization.

Following this trend and knowing the importance of adipose tissue as an organ, this project aims to test new methods to characterize gene expression differences related to sexual dimorphism in the adipose tissue using patient gene expression data from the GTEx Database, quantitative analysis techniques such as Differential Gene Expression Analysis or Gene Set Enrichment Analysis (GSEA), and a new metric called the Dimorphism Index. The Dimorphism Index we use later in this project is a metric that takes into account the sexual dimorphism as a continuous parameter. It is defined as such : given the matrix of similarities between the same organ from all individuals in the population, the dimorphism index of person A's organ is defined as the median across all pairs [person A, person X] of the following ratio : proportion of males that have this organ more similar to the person A's organ than person X's organisms divided by the proportion of females that have this organ more similar to the person A's organ than person X's organisms It is to be noted that similarity in this context refers to the Spearman's correlation across all genes of any pair of genome-wide expression profiles.

More intuitively, it represents the probability that a subject's organ is more similar to the same organ of a same-sex subject in the population than the same organ of an individual of the opposite sex in the population. In other words, for a male, the higher this ratio is, the more "masculine" the organ of the subject is. For a female, the higher the ratio is, the more "feminine" the organ of the subject is. This way of defining the dimorphism index presents 4 main advantages. It makes no prior assumptions about dimorphism traits and is thus unbiased in that regard. Moreover, it gives a quantitative continuous value corresponding to the new realities of sexual dimorphism. Also, this approach has the benefit of being universal, as it is applied/applicable to the entire phenotypic space and at all scales (populations, organs, etc.). Finally, it is meaningful to individuals and not limited to population-level investigations only.

2 Methods

2.1 Available Data and Preprocessing

The data at the basis of the analysis in this report is an RNA read count matrix containing RNA counts of 18749 genes measured in postmortem subcutaneous adipose tissue of 648 donors. The raw data comes from the GTEx project and the effect of technical covariates was removed from the expression data. Those adjusted counts can be found in the `RNA_counts_adjusted.csv` file.

Additionally, clinical and technical data related to the samples are also available with pre-computed dimorphism indices in the `covariates.csv` file.

The sample donor's ages were discretized by decade. However, for the differential gene expression analysis (see Section 2.3 and Section 3.2) we considered this covariate to be a continuous variable and thus added an `AGE_CONTINUOUS` covariate which is 35 for the 30-39 decade, 45 for the 40-49 decade, etc.

Finally, 12 samples had missing values for the Hardy scale. For simplicity, we decided to remove those samples because there cannot be missing values in a variable used in the design of the `DESeq` function, used for differential gene expression analysis. Even though this decrease in sample size could lead to lower statistical power, this method was chosen since the number of samples with missing values is relatively small compared to the total number of samples (660). Other, more sophisticated methods could have been used, such as replacing the missing values by the most common one (`HARDY_SCALE` $= 0$ with this data set), or creating a classifier model to predict the missing values using the other covariates. We could also have replaced the "NA" value by a new category called "unknown". However, since there are only 12 samples in that category, and there are quite a lot of variables in the design (see Section 3.2), we would risk having a low statistical power (very few degrees of freedom).

All the analyses were performed with R (version 4.4.1 (2024-06-14)). For more details about the R version information, the OS and attached or loaded packages, see Section 5.3.

CODE and DATA AVAILABILITY?

2.2 Principal Component Analysis

The PCAs in Section 3.1 were performed on the counts after a variance stabilizing transformation (using the `vst` function from `DESeq2` with its default arguments). Since the log geometric means could not be computed because every gene contained at least one zero, a pseudo-count of 1 was added to every entry of the count matrix.

The PCAs were performed by using the `DESeq2` `plotPCA` function and setting the `ntop` parameter to 500.

2.3 Differential Gene Expression Analysis

The differential gene expression (DGE) analysis was performed using the `DESeq2` package. First, half of the genes with the lowest variability were filtered out. The variability was determined by computing the median absolute deviation (MAD), defined as the median of the absolute deviations from the data's median $\tilde{X} = \text{median}(X)$:

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|)$$

with $X = X_1, X_2, ..., X_n$, and n being the number of samples.

The Hardy scale and age were included as covariates in the generalized linear model (GLM) for the count data by using the following design: `~ HARDY_SCALE + AGE_CONTINUOUS + SEX + SEX:RNA_DI`. For a justification of this design, see Section 3.1.

Afterwards, the continuous variable `AGE_CONTINUOUS` was centered and scaled to improve convergence of the GLM, as recommended by `DESeq2`. The `RNA_DI` variable was not centered nor scaled as the value `RNA_DI` $= 0$ has a biological meaning, and centering that variable would erase this meaning, especially since there are many more males than females in our data set.

To identify genes associated with the dimorphism indices, with sex or with age, the Wald significance test was used. To identify genes associated with the Hardy scale covariate, the likelihood ratio test (LRT) was used instead. The LRT can be used when evaluating expression change across more than two levels (the Hardy scale has 5 levels). For this test, a full model, containing the covariate of interest (i.e., `HARDY_SCALE`), is compared to a reduced model, with the covariate of interest removed. If a significant amount of variation is explained by the covariate of interest, this suggests that the gene is differentially expressed across the different levels. The reduced model used in our analysis is: `~ AGE_CONTINUOUS + SEX + SEX:RNA_DI` [7].

Finally, a Benjamini-Hochberg correction was applied to all the p-values to control the false discovery rate at a level of 0.05.

2.4 Identification of Hormone Receptors Associated with the Male and Female Dimorphism Indices and with Sex

In order to identify hormone receptors that are associated with the female and male dimorphism indices and sex, a list of 78 hormonal receptors was created. This list contains the protein names, the corresponding gene names, the HGNC gene IDs, the family names and alias.

To get a comprehensive list of hormone receptors expressed in the human genome, we used the HGNC (HUGO Gene Nomenclature Committee) [1] as a resource to find approved gene nomenclature. We searched on the database with keywords “hormone receptor” and filtered by protein families. We then downloaded the results as a `.txt` file, converted it to a `.csv` file to easily extract the family IDs. We then used the HGNC Biomart tool [6] (for families) with the appropriate filters to return all proteins parts of requested family IDs and downloaded the results as a csv file. This csv file contained a column with the gene approved nomenclature and another column with the associated approved gene complete name (which corresponds most often to the protein receptor name).

At first, this method yielded 200 protein groups and a total of more than 19000 protein entries in the final csv, which we considered too high of a number and with a lot of false positives (ie. non-hormonal receptors). We thus decided to conduct a more strict selection and only include the results including explicitly the word “hormone”, namely the first 9 families with ID : 71,199,227,235,266,270,1175,1896,2070. This yielded only 87 protein entries, which was more manageable and interpretable for this project.

Finally, it was used as a reference to map every gene of interest in our dataset to a complete approved gene name.

Based on the significant (FDR < 0.05) genes found by the Wald tests for the dimorphism indices and sex in the DGE analysis, we determined which ones were in the list of hormonal receptors.

2.5 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis was performed to identify pathways associated with the dimorphism indices and with sex. To this end, the REACTOME gene sets were used (in the `c2.cp.reactome.v2024.1.Hs.symbols.gmt` file) with the `fgsea` R package. The genes tested for differential expression in Section 3.2 were ranked by decreasing order of the log2 Fold Change (for male and female dimorphism indices, and sex). Gene sets with less than 15 genes were not tested (`minSize` argument of the `fgsea` function). The number of permutations to do was set to 10000 (`nperm` argument of the `fgsea` function).

3 Results

3.1 Descriptive Analysis

The distribution of all the covariates is shown in Figure 1. The number of male sample donor's was twice that of female sample donor's. The two classes were therefore unbalanced. The death classification, based on the Hardy scale, was "0" for the majority of donors, which corresponds to cases on a ventilator immediately before death. The second most common death classification was a Hardy scale of "2": the sudden unexpected deaths of people who had been reasonably healthy. Additionally, the vast majority of samples had unspecified pathology categories (missing values). The second most common pathology category was fibrosis.

Moreover, the majority of donors, regardless of sex, were in the 50-69 age range. The average RNA integrity number was 6.86, with a minimum at 5.5, and seemed to be similarly distributed regardless of the sex.

The ischemic time (i.e., the time interval between death of withdrawal from life-suport and the preservation of the sample by fixation or freezing) varied between 83 minutes and 1683 minutes (1.38 hours), with a mean of 607.74. This variable also seemed to be similarly distributed across the two sexes.

The male and female dimorphism indices in log scale were similarly distributed between male and females. The female dimorphism index had a mean of 0.17 and a variance of 0.08, while the male dimorphism index had a mean of 0.18 and a variance of 0.08. Interestingly, the dimorphism indices calculated from histology profiles showed a very different distribution. The male histology profile-based dimorphism index seemed to have a bimodal distribution, with a mean of 0.12 and a variance of 0.29. The female histology profile-based dimorphism index had a mean of 0.39 and a variance of 0.23.

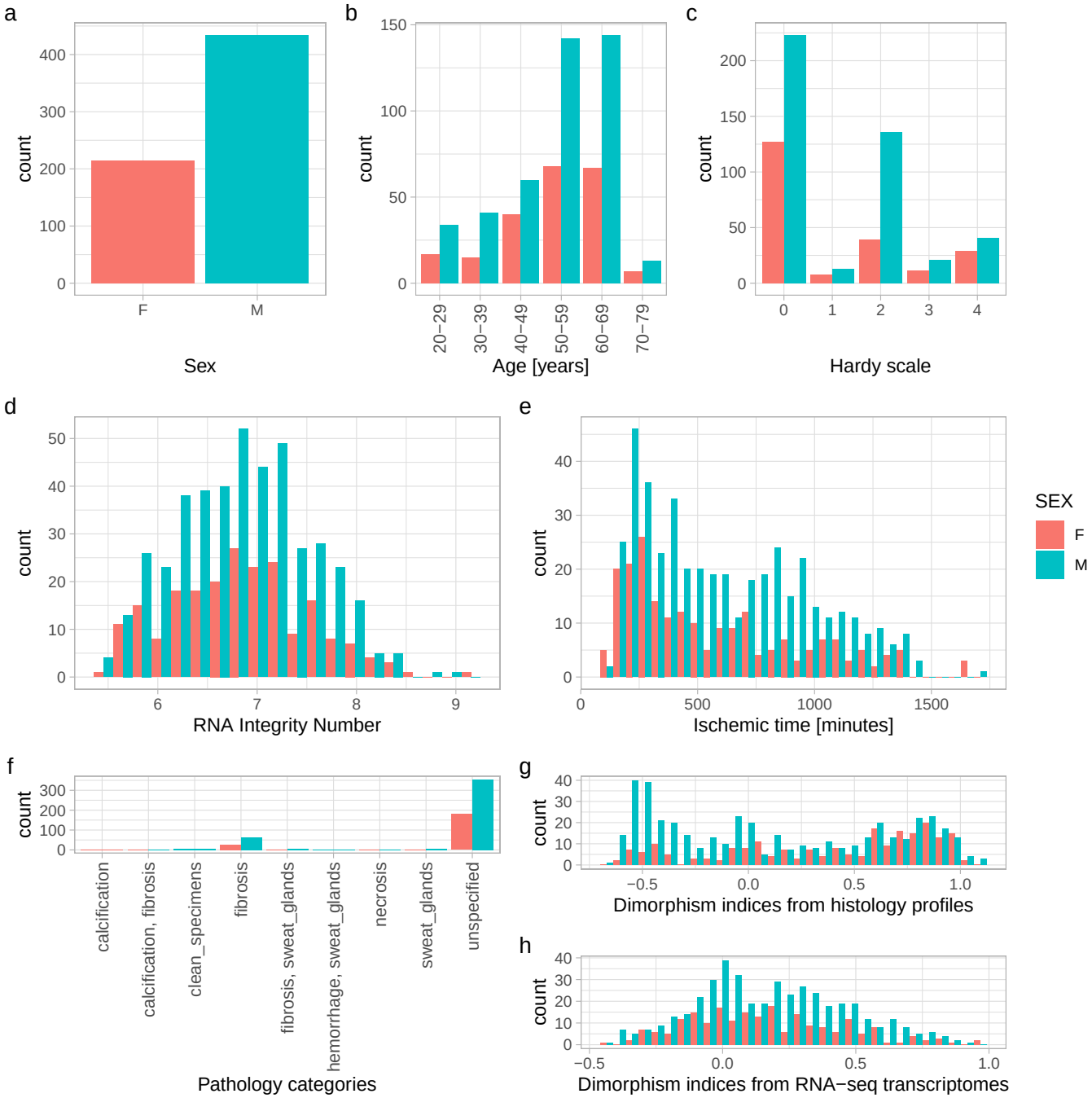

Figure 1: Distribution of the clinical and technical variables, and the dimorphism indices of the samples separated by sex (F = Female, M = Male). **(a)** Number of Female and Male subcutaneous adipose tissue sample donors. **(b)** Distribution of ages (in decades) of the donors. **(c)** Death classification of the sample donors with the Hardy scale. **(d)** Distribution of the sample RNA integrity numbers. **(e)** Distribution of the sample ischemic time. **(f)** Medical conditions diagnosed post-mortem from histology slides. **(g)** Distribution of the transcriptional dimorphism indices computed from whole RNA-seq transcriptomes. **(h)** Distribution of the dimorphism indices computed from AI-based histology profiles.

Figure 2 shows the Pearson correlation coefficients between each continuous clinical and technical covariate. It also shows the pairwise scatter plots of those variables. We observed that the RNA- and histology-

based dimorphism indices were significantly correlated with one another, even though the correlation was quite low. Since the two indices measure the same phenomenon, it was expected to observe a positive correlation between those variables. However, the low value of this correlation could indicate that there was a high error in one or both measurements of the dimorphism indices or that the two variables did not capture exactly the same information. From the pairwise scatter plot, however, it does not appear that the low correlation was due to a nonlinear relationship between the two indices.

We also noticed that the age was negatively correlated with the RNA integrity number (RIN) and positively correlated with the ischemic time. The latter variable was negatively correlated with the RNA integrity number. This last result is easily interpreted: the RIN is a measure of RNA quality in the sample. If the ischemic time is large, the sample has more time to degrade, thus decreasing the RNA integrity.

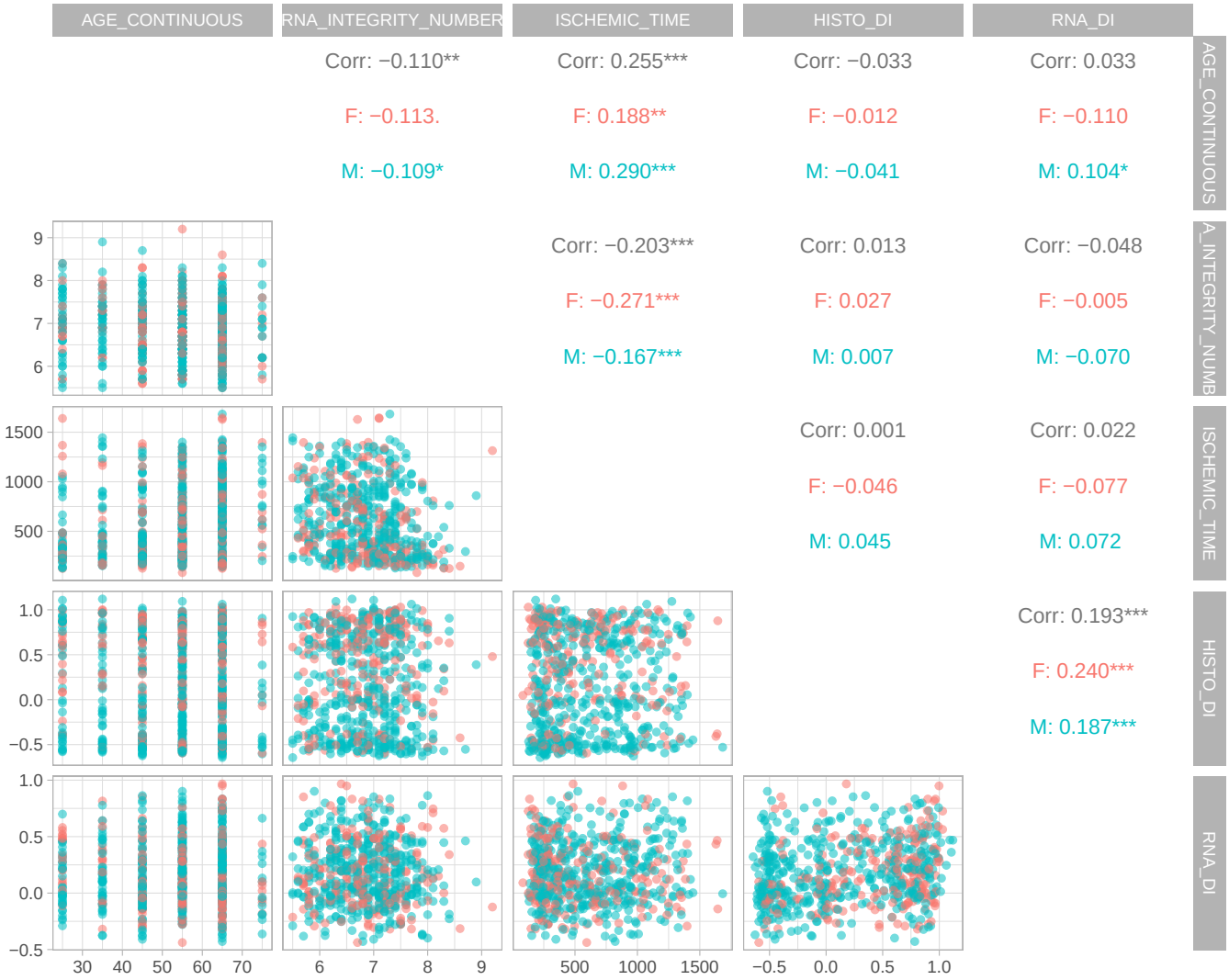

Figure 2: Pairs plot of the continuous clinical and technical covariates for the samples separated by sex. The lower diagonal shows the scatter plots of each pair of variables. The upper diagonal indicates the Pearson correlation coefficients of the variabes considering all the samples (grey), or for the samples separated by sex (F = Female, M = Male). Statistical significance of the correlation is indicated by an asterisk (*** for p-value $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and . for $p < 0.10$).

To check whether some technical variables were possibly confounding the dimorphism indices and clinical

variables, a PCA was performed on the 500 genes with the highest variance. Moreover, a variance stabilizing transformation was applied to the counts after adding a pseudo-count of 1 to every entry of the count matrix (to avoid calculating log(0)). Afterwards, the first two principal components were plotted, and the points, representing the samples, were colored according to the variable of interest. If a clustering (for categorical variables) or a clear gradient (for continuous variables) appears along the first and/or second principal component(s), this indicates that the variable of interest is probably a significant source of variation that affects the counts.

First, in Figure 3a, we see a separation of the samples according to their sex along the first principal component. In Figure 3b, the samples were colored based on their dimorphism index (calculated from whole RNA-seq transcriptomes). For each sex, a clear gradient can be seen along the first principal component. We also observe that, for the males, the dimorphism decreases along PC1, while the opposite occurs for females. This aligns with the definition of the dimorphism indices: males that are more similar to other males have a larger dimorphism index. Females that are more similar to other females also have a larger dimorphism index. But, females/males that are more similar to the opposite sex have low dimorphism indices, so they are found in the center of the PCA plot, in the region where male and female samples overlap the most.

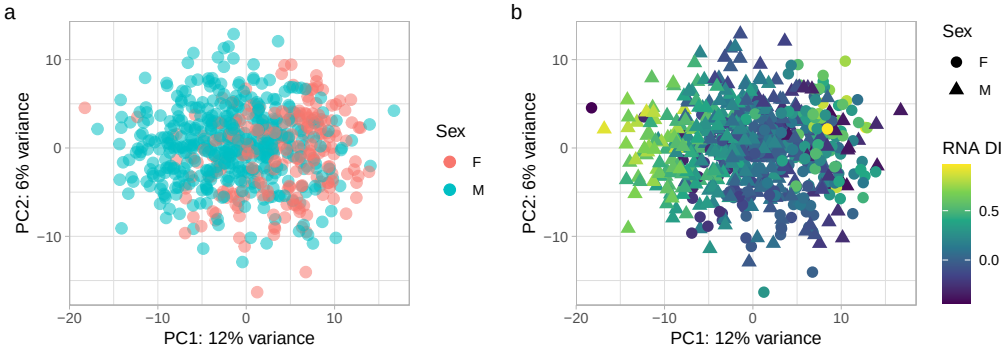

Figure 3: First two principal components of the PCA performed on the 500 most variant genes. The samples are colored either by sex **(a)** or by the RNA-derived dimorphism index **(b)**. For subplot **(b)**, the samples are colored by sex.

In Figure 4a and Figure 4b, the samples were first separated by their sex before performing the PCA. The samples were then colored by the corresponding donor's age. For the males subset, no gradient was visible along PC1 and PC2 according to age. On the other hand, women of age below 40 seemed to cluster more to the right of the plot on Figure 4b. Therefore, age could be a confounding factor for the female dimorphism index.

On Figure 4c, the dots were colored by the Hardy scale, which is related to the cause of death. A separation between samples with a Hardy scale of 0 and 2 can be seen. Since this separation occured along the first principal component like the male and female dimorphism indices and the sex, the Hardy scale could be a confounding factor as differences in the data due to the Hardy scale value could be attributed to the sex, and vice versa.

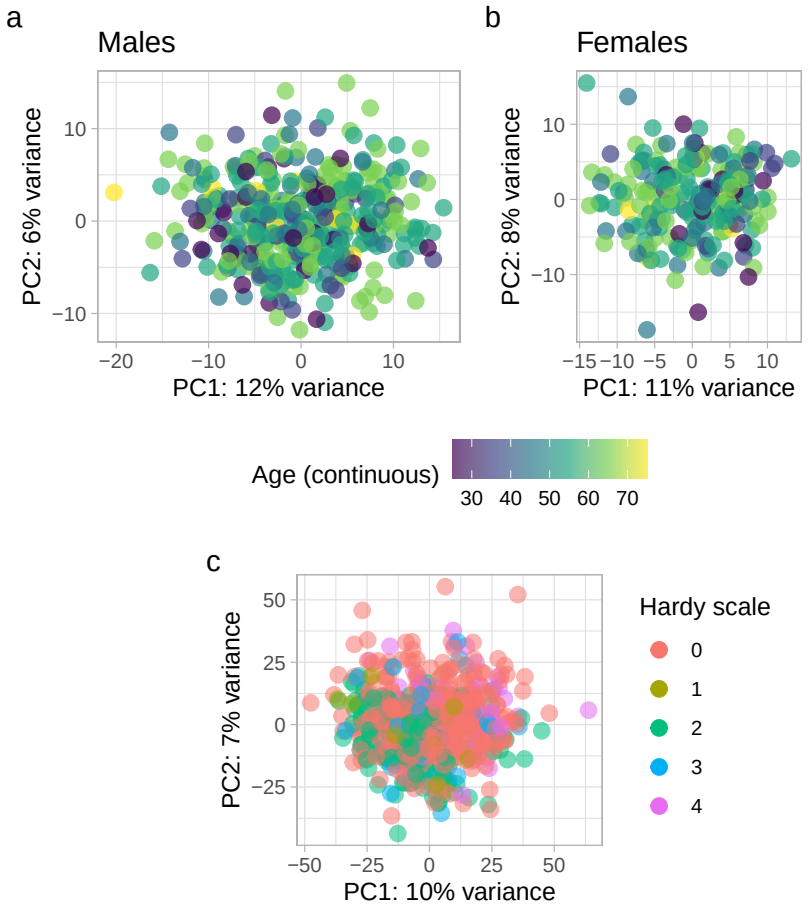

Figure 4: First two principal components of the PCA performed on the 500 most variant genes in the male and female subsets (**a** and **b**) or in the entire data set **(c)**. The samples are colored either by age (**a** and **b**) or by Hardy scale (**c**).

In conclusion, the following technical variables were taken into account for the differential gene expression analysis in the following section: `SEX`, `RNA_DI`, `AGE_CONTINUOUS` and `HARDY_SCALE`. The first two variables are of particular interest to us for this study, and the last two variables are potential confounders that are important to take into account and adjust for.

3.2 Differential Gene Expression Analysis

In this section, we performed a differential gene expression analysis to identify genes that exhibited statistically significant differences in expression levels between two or more groups, or genes whose expression levels changed in relation to a continuous variable.

In this study, we were most interested in identifying the genes associated with the male and female dimorphism indices. We then aimed to identify genes that were differentially expressed in men compared to women. Finally, we determined which genes were associated with the confounding factors identified in the previous section, namely the age and the Hardy scale.

To obtain the following results, the design: `~ HARDY_SCALE + AGE_CONTINUOUS + SEX + SEX:RNA_DI` was used in `DESeq2`.

3.2.1 Genes associated with the female and male dimorphism indices

All the genes tested for differential expression are shown as dots on the volcano plots in Figure 5. For both male and female dimorphism indices (DIs), the majority of genes tested were statistically significantly associated with the DIs (77.73% of genes for the male DI, and 70.23% for the female DI). We also noticed that, for the female DI, the majority of differently expressed genes had expression levels that were negatively associated with the female DI, while the majority of differently expressed genes were positively associated with the male DI. Among the down-regulated genes with the lowest adjusted p-values for the female DI, some were also found to be among the up-regulated genes with the lowest adjusted p-values for the male DI. This included genes such as GARNL3, ARHGEF25 or DBNDD2.

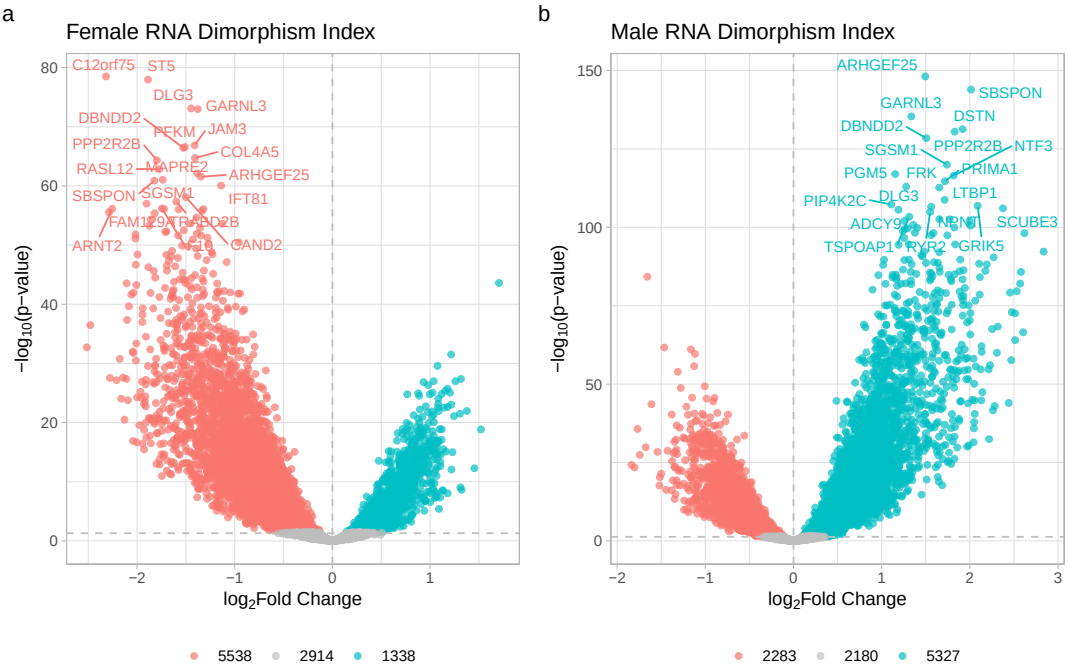

Figure 5: Volcano plots showing differential gene expression analysis results for the female **(a)** and male **(b)** dimorphism indices. Differential expression was assessed using Wald tests with multiple testing correction (Benjamini-Hochberg method). Genes meeting the significance threshold (FDR < 0.05) are color-coded according to the direction of change: up-regulated genes with a positive log2 fold change (FC) are colored in blue, while down-regulated genes with a negative log2 FC are colored in red. For both plots, the 20 genes with the lowest adjusted p-values are labelled with their IDs.

To find whether this overlap between genes that are negatively associated with the female DI and positively associated with the male DI was also true for the other significant genes, we plotted a Venn diagram (see Figure 6) showing the overlap between the up- and down-regulated genes for both male and female DIs. And indeed, there is a substantial overlap of 4676 genes that were positively associated with the male DI but negatively associated with the female DI. Conversely, 1112 genes were negatively associated with the male DI but positively with the female DI. The number of genes that were regulated the same way according to the two DIs was much lower, with just 12 genes positively associated with both male and female DIs, and 99 genes down-regulated according to both DIs.

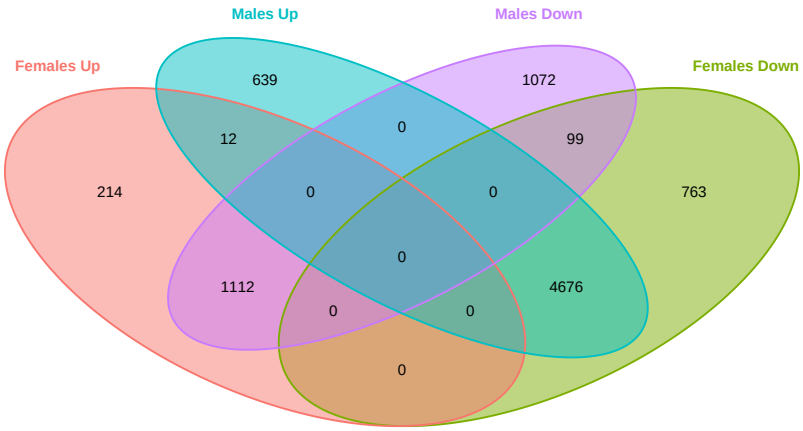

Figure 6: Venn diagram showing the overlap in genes that were up- and down-regulated in men and women according to the male and female dimoprhism indices.

3.2.2 Genes associated with sex, and other technical and clinical covariates

Figure 7a shows the baseline difference between males and females, i.e., the genes that were differently expressed when the dimorphism index is null. It can be interpreted as the differentially expressed genes for individuals that are equally similar to other same-sex and opposite-sex individuals in terms of their overall transcriptomes. In order to determine the main effect of sex, i.e., determine which genes were differentially expressed in males and females regardless of the value of the dimorphism indices, another design, `~ HARDY_SCALE + AGE_CONTINUOUS + SEX` would have to be used. However, the results for the Hardy scale and the age would also change, making the results more difficult to interpret. The results using this alternative design are shown in the annex in Section 5.1.

When the dimorphism indices were equal to 0, many genes were still considered to be differentially expressed: 914 genes were up-regulated in males compared to females, while 3786 genes were more strongly expressed in females compared to males. However, the amplitude of the log2 fold change is below 1 (in absolute value) for most of the significant genes, meaning that, while the difference in gene expression may be statistically significant, it may not be biologically relevant.

Conversely, a relatively small number of genes (9.3% of tested genes) were associated with age (Figure 7b).

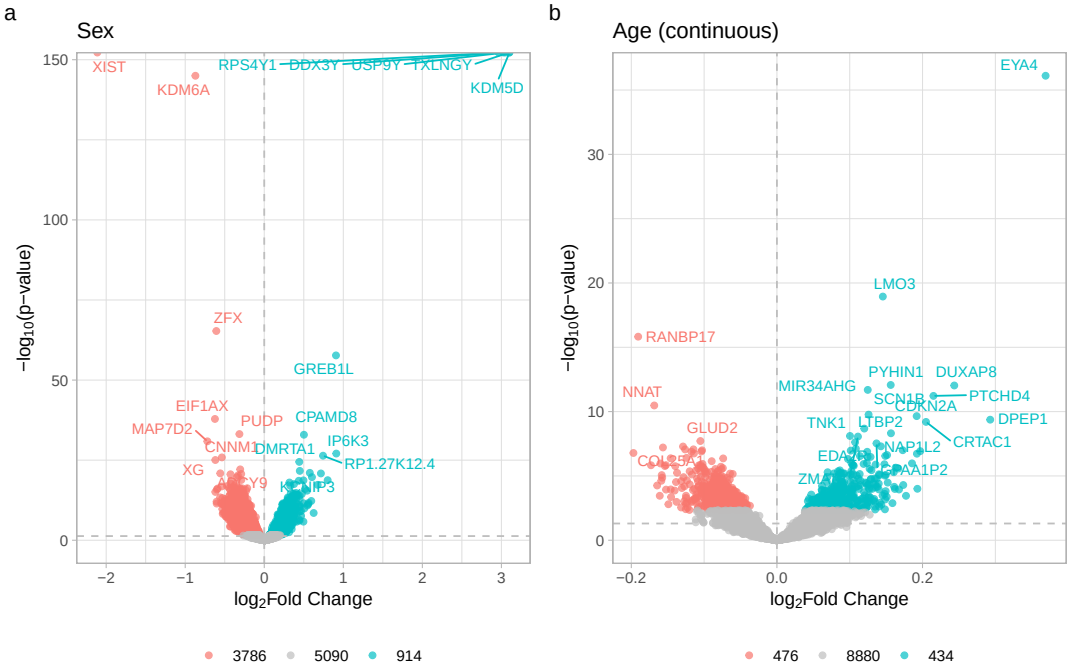

Figure 7: Volcano plots showing differential gene expression analysis results for the sex **(a)** and age (continuous) **(b)** covariates. Differential expression was assessed using Wald tests with multiple testing correction (Benjamini-Hochberg method). Genes meeting the significance threshold (FDR < 0.05) are color-coded according to the direction of change: up-regulated genes with a positive log2 fold change (FC) are colored in blue, while down-regulated genes with a negative log2 FC are colored in red. For both plots, the 20 genes with the lowest adjusted p-values are labelled with their IDs. A positive log2 Fold Change for the sex covariate indicates that genes are more expressed in males than females, and a negative log2 Fold Change indicates the opposite.

Finally, with the likelihood ratio test (LRT), 5890 genes were considered significantly associated with the Hardy scale, and 3900 genes were not. Figure 8 shows boxplots of the counts of the 9 genes with the lowest adjusted p-values for the LRT. We see that for all those genes, except for GPX3, the median counts decrease as the Hardy scale increases, except for the level "0" which tends to have median counts more similar to the level "3".

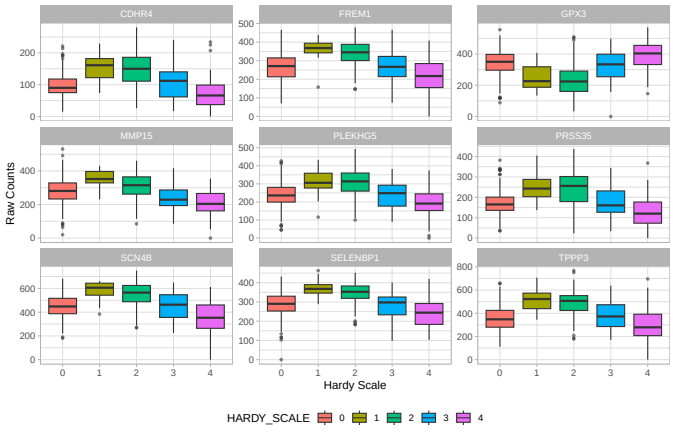

Figure 8: Boxplots of the counts of the 9 genes with the lowest adjusted p-value (by Benjamini-Hochberg method) for the likelihood ratio test on the Hardy scale covariate.

3.3 Interpretation of the DI-associated transcriptome

3.3.1 Hormone Receptors Associated with the Dimorphism Indices and with Sex

Among the 78 hormone receptors in our curated list, 29 were significantly associated with the female DI, 28 with the male DI, and 20 with the sex (Figure 9). The majority of those hormone receptors were associated with all three variables as 33 hormone receptors in total were associated with either of the three variables, and 26 of them were shared between the male and female dimorphism indices.

We noticed that most of those receptors had positive associations with the male DI but negative associations with the female DI. The log2 fold changes of the hormone receptors associated with the sex (when RNA_DI = 0) were much closer to zero compared to the dimorphism indices.

Among this list of significant receptors, we found some related to sex hormones, such as the androgen receptor (androgens refer to testosterone and dihydrotestosterone (DHT), the male sex hormones), the estrogen receptors 1 and 2, the progesterone receptor, and the luteinizing hormone/choriogonadotropin receptor, which is a protein associated with disorders of male secondary sexual character development when mutations are present in the gene [8]. Interestingly, the androgen receptor was positively associated with the female DI and sex, but not with the male DI.

Some of those significant receptors do not have a direct sexual function, such as the nuclear receptor subfamily 1 group D members 1 and 2, which play roles in circadian rhythms and carbohydrate and lipid metabolism [9]. The RAR related orphan receptors A and B also regulate the expression of genes involved in the circadian rhythm [10].

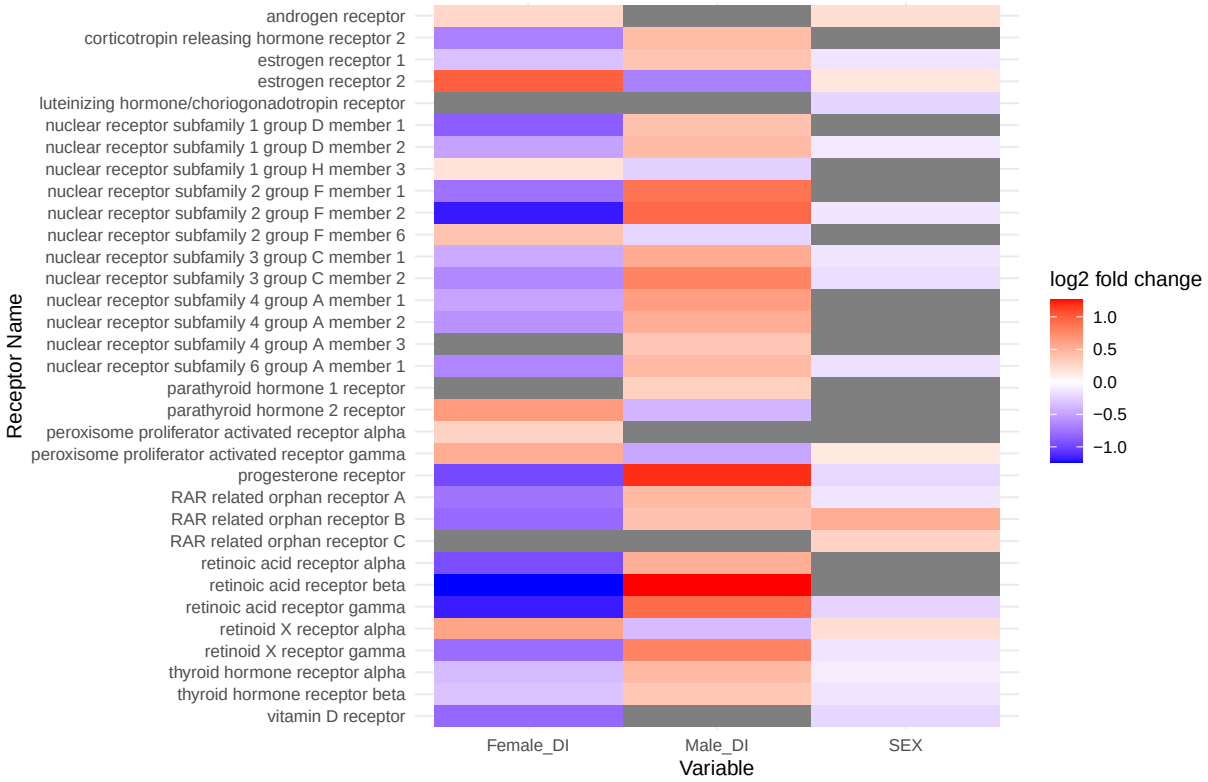

Figure 9: Hormonal receptors associated with the female and male dimorphism indices and the sex (when the RNA_DI variable is null). The colors correspond to the log2 fold changes, determined in the DGE analysis. A grey color is given to the receptors that were not significantly (FDR $\geqslant 0.05$) associated with the variable.

3.3.2 Gene Set Enrichment Analysis

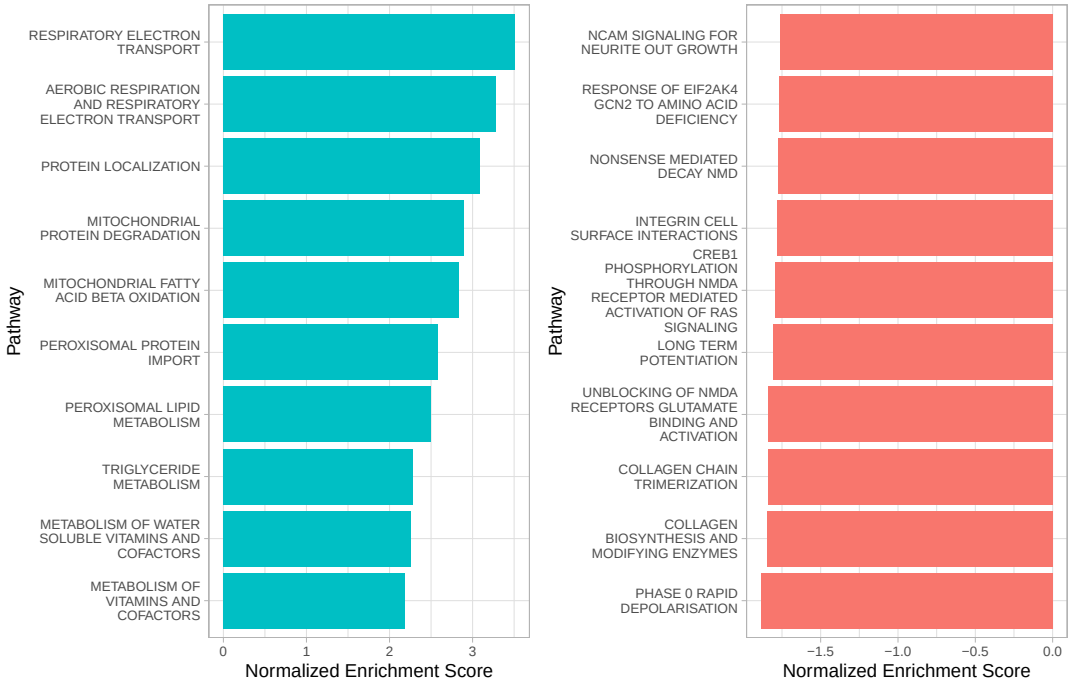

Figure 10: 10 significant Reactome pathways with the highest (in blue) and lowest (in red) enrichment scores from the GSEA for the female dimorphism index. A positive Normalized Enrichment Score (NES) indicates that the gene set is enriched in genes that are positively associated (positive log2 Fold Change) with the female dimorphism index. A negative NES indicates a gene set enriched in genes that are negatively associated with the female dimorphism index.

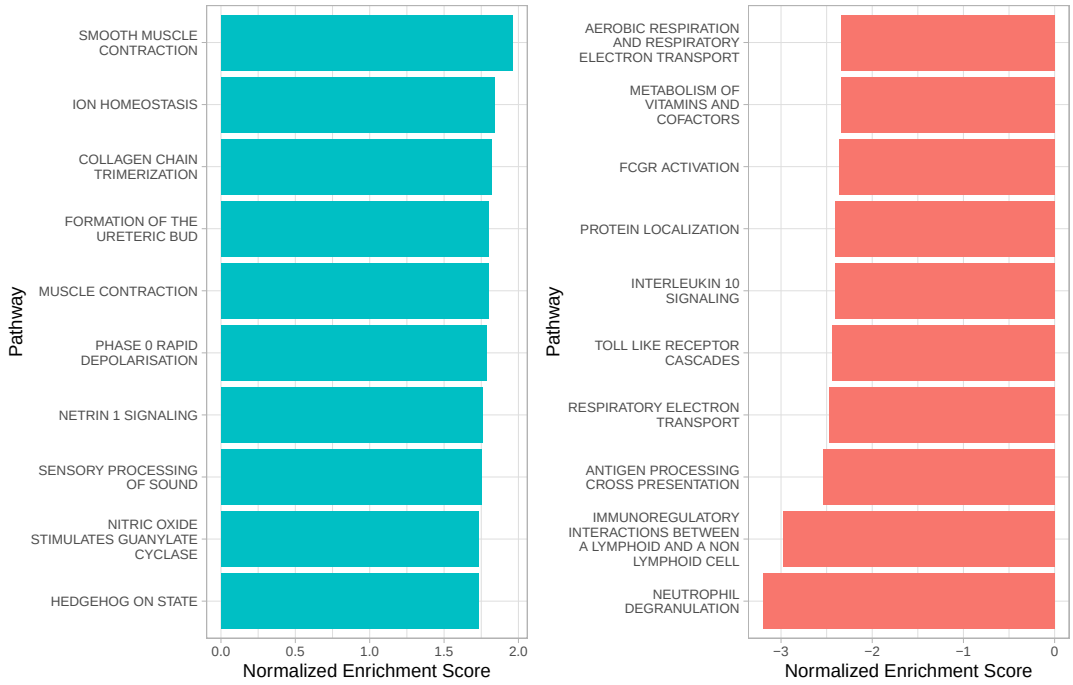

Figure 11: 10 significant Reactome pathways with the highest (in blue) and lowest (in red) enrichment scores from the GSEA for the male dimorphism index. A positive Normalized Enrichment Score (NES) indicates that the gene set is enriched in genes that are positively associated (positive log2 Fold Change) with the male dimorphism index. A negative NES indicates a gene set enriched in genes that are negatively associated with the male dimorphism index.

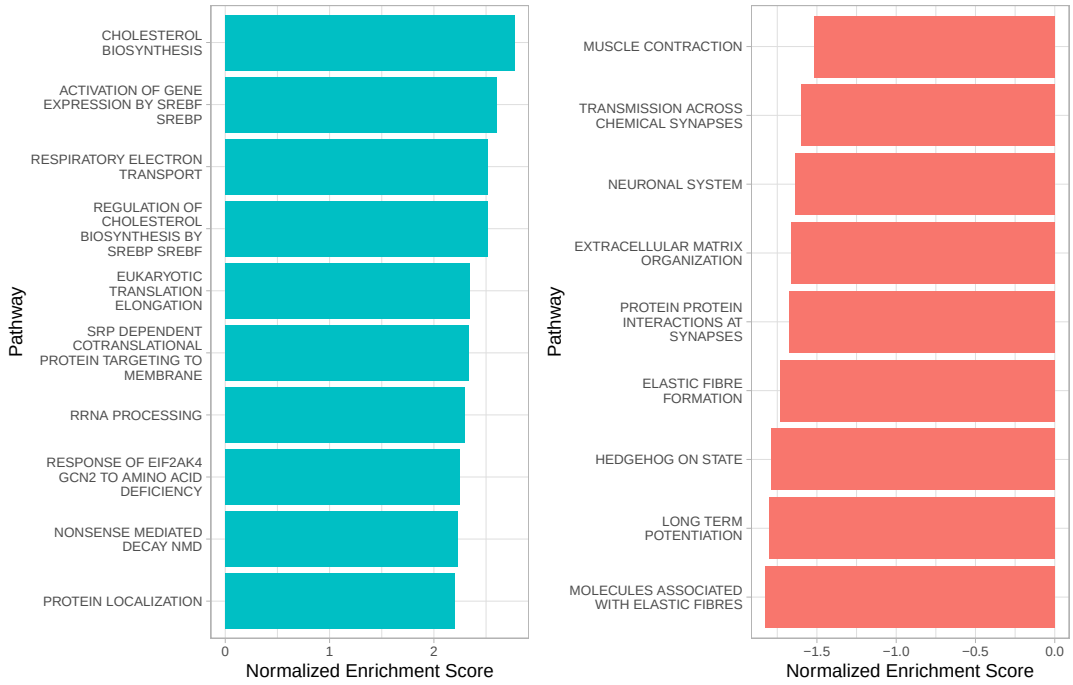

Figure 12: 10 Reactome pathways with the highest (in blue) and lowest (in red) enrichment scores from the GSEA for the sex. A positive Normalized Enrichment Score (NES) indicates that the gene set is enriched in genes present at the top of the ranked list. The genes were ranked by log2 Fold Change, so a positive NES indicates that many genes that were up-regulated in males are present in the gene set. A negative NES indicates that many genes with very low (negative) log2 Fold Changes, at the bottom of the ranked list, are present in the gene set.

4 Discussion and Conclusion

- The fact that the sexual dimorphism index has a different meaning for men and women makes its use in statistical "tests" more difficult, such as the choice of the design for DESeq2 DGE analysis.
- GSEA: the ranking of the genes passed to the fgsea algorithm can be based on different statistics (log2FC, test statistic, adjusted p-value, log2FC * (log10(padj))), and the choice of this ranking changes the results of GSEA, making interpretation of results difficult.

5 Annexes

5.1 Alternative design for DGE analysis

The plots in Figure 13 show the results of the Wald tests when using an alternative design to determine the overall effect of sex, without controlling for the dimorphism indices: `~ HARDY_SCALE + AGE_CONTINUOUS + SEX`.

For the sex variable (Figure 13a), while the number of significant genes, both up-regulated and down-regulated has changed compared to the results in Figure 7b, the identity of the most significant genes (the genes with the lowest adjusted p-values) remain very similar. XIST and KDM6A are still the most

strongly down-regulated genes, and RPS4Y1, DDX3Y USP9Y, TXLNGY and KDM5D are still the most up-regulated genes (in males compared to females).

For the age (Figure 13b), the conclusion is reached: the number of significant genes has changed drastically, but the identity of the most significant genes remain the same (EYA4, LMO3, RANBP17, NNAT, etc.). With this alternative design, the number of genes positively associated with age as been increased by a factor of 3.78, and by a factor of 1.7 for the genes that were negatively associated with age.

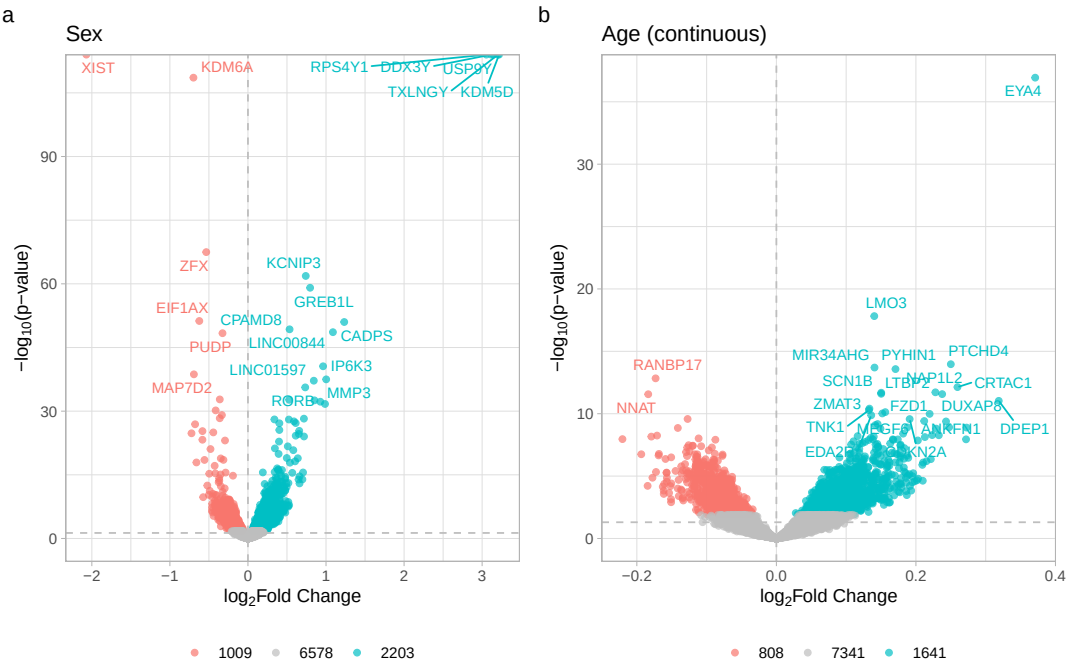

Figure 13: Volcano plots showing differential gene expression analysis results for the age (continuous) **(a)** and sex **(b)** covariates. Differential expression was assessed using Wald tests with multiple testing correction (Benjamini-Hochberg method). Genes meeting the significance threshold (FDR < 0.05) are color-coded according to the direction of change: up-regulated genes with a positive log2 fold change (FC) are colored in blue, while down-regulated genes with a negative log2 FC are colored in red. For both plots, the 20 genes with the lowest adjusted p-values are labelled.

For the Hardy scale, the number of significant genes has also changed with the alternative design compared to the original design, but the change, in relative terms, is less significant. With this alternative design 6007 genes were associated with the Hardy scale covariate. The identity of the 9 most significant genes has changed for just one gene: in the alternative design, the TRIL gene replaces the MMP15 gene.

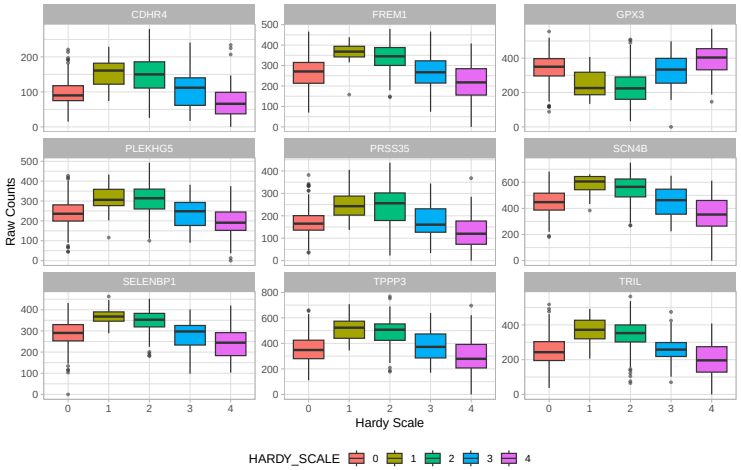

Figure 14: Top 9 genes with lowest adjusted p-value for the likelihood ratio test on Hardy scale covariate

In conclusion, when changing the original design by the alternative design to determine the genes that are differently expressed in males and females, the number of differently expressed genes for all the covariates in the design change. However, the identity of the most significant (lowest adjusted p-values) genes remains very similar between the two designs. Therefore, the biological conclusions that we would reach would overall remain the same when limiting ourselves to those most significant genes.

5.2 References

[0] Parts of the introduction were heavily inspired by the BINF-F401-2025 project presentation slides.

[1] Mauvais-Jarvis F, Bairey Merz N, Barnes PJ, Brinton RD, Carrero JJ, DeMeo DL, De Vries GJ, Epperson CN, Govindan R, Klein SL, Lonardo A, Maki PM, McCullough LD, Regitz-Zagrosek V, Regensteiner JG, Rubin JB, Sandberg K, Suzuki A. Sex and gender: modifiers of health, disease, and medicine. *Lancet.* 2020 Aug 22;396(10250):565-582. doi: 10.1016/S0140-6736(20)31561-0. Erratum in: *Lancet.* 2020 Sep 5;396(10252):668. doi: 10.1016/S0140-6736(20)31827-4. PMID: 32828189; PMCID: PMC7440877.

[2] Karastergiou, K., Smith, S.R., Greenberg, A.S. et al. Sex differences in human adipose tissues – the biology of pear shape. *Biol Sex Differ* 3, 13 (2012). https://doi.org/10.1186/2042-6410-3-13

[3] https://www.gtexportal.org/home/

[4] Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, Bruford EA. Genenames.org: the HGNC resources in 2023. Nucleic Acids Res. PMID: 36243972 DOI: 10.1093/nar/gkac888

[5] https://www.genenames.org/tools/search/#!/?query=hormone%20receptor&filter=document_type:%22group%22&start=0&rows=20

[6] https://biomart.genenames.org/martform/#!/default/HGNC?datasets=hgnc_gene_mart_2025_06_20

[7] https://hbctraining.github.io/DGE_workshop_salmon/lessons/08_DGE_LRT.html

[8] https://www.ncbi.nlm.nih.gov/gene/3973

[9] https://www.ncbi.nlm.nih.gov/gene/9975

[10] https://www.ncbi.nlm.nih.gov/gene/6096

5.3 Session Information

```
R version 4.4.1 (2024-06-14)
Platform: aarch64-apple-darwin20
Running under: macOS Sonoma 14.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPAC

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Europe/Brussels
tzcode source: internal

attached base packages:
[1] grid      stats4    stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] fgsea_1.30.0                corrplot_0.95
 [3] VennDiagram_1.7.3           futile.logger_1.4.3
 [5] viridis_0.6.5               viridisLite_0.4.2
 [7] GGally_2.2.1                scales_1.4.0
 [9] pheatmap_1.0.12             DESeq2_1.44.0
[11] SummarizedExperiment_1.34.0 Biobase_2.64.0
[13] MatrixGenerics_1.16.0       matrixStats_1.5.0
[15] GenomicRanges_1.56.2        GenomeInfoDb_1.40.1
[17] IRanges_2.38.1              S4Vectors_0.42.1
[19] BiocGenerics_0.50.0         patchwork_1.3.0
[21] ggrepel_0.9.6               gt_1.0.0
[23] magrittr_2.0.3              lubridate_1.9.4
[25] forcats_1.0.0               stringr_1.5.1
[27] dplyr_1.1.4                 purrr_1.0.4
[29] readr_2.1.5                 tidyr_1.3.1
[31] tibble_3.2.1                ggplot2_3.5.2
[33] tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] tidyselect_1.2.1        farver_2.1.2            fastmap_1.2.0
 [4] digest_0.6.37           timechange_0.3.0        lifecycle_1.0.4
 [7] compiler_4.4.1          rlang_1.1.6             tools_4.4.1
[10] yaml_2.3.10             data.table_1.17.2       knitr_1.50
[13] lambda.r_1.2.4          labeling_0.4.3          S4Arrays_1.4.1
```

```
[16] DelayedArray_0.30.1     plyr_1.8.9              xml2_1.3.8
[19] RColorBrewer_1.1-3      abind_1.4-8             BiocParallel_1.38.0
[22] withr_3.0.2             colorspace_2.1-1        tinytex_0.57
[25] cli_3.6.5               rmarkdown_2.29          crayon_1.5.3
[28] generics_0.1.4          rstudioapi_0.17.1       httr_1.4.7
[31] tzdb_0.5.0              zlibbioc_1.50.0         parallel_4.4.1
[34] formatR_1.14            XVector_0.44.0          vctrs_0.6.5
[37] Matrix_1.7-3            jsonlite_2.0.0          hms_1.1.3
[40] locfit_1.5-9.12         glue_1.8.0              ggstats_0.9.0
[43] codetools_0.2-20        cowplot_1.1.3           stringi_1.8.7
[46] gtable_0.3.6            UCSC.utils_1.0.0        pillar_1.10.2
[49] htmltools_0.5.8.1       GenomeInfoDbData_1.2.12 R6_2.6.1
[52] evaluate_1.0.3          lattice_0.22-7          futile.options_1.0.1
[55] fastmatch_1.1-6         Rcpp_1.0.14             gridExtra_2.3
[58] SparseArray_1.4.8       xfun_0.52               pkgconfig_2.0.3
```