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

Youmna Ayadi   Kevin Straatman

2025-06-19

Table of contents

1 Abstract

2 Introduction

Definition of the dimorphism indices. It is an exploratory analysis.

3 Methods

3.1 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 / normalized out” 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 3.3 and Section 4.2) 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 to samples because there cannot be missing values in a variable used in the design of the `DESeq` function, used for differential gene expression analysis. Other, more sophisticated methods could have been used, such as replacing the missing values by the most common one (`HARDY_SCALE` = 0 with our data set), or creating a classifier model to predict the missing values using the other covariates (source?). 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 quite a lot of variables in the design (see Section 4.2), we would risk having a low statistical power (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 6.1.

CODE AVAILABILITY?

3.2 PCA

To estimate the size factors and the vst (variance stabilizing transformation) with DESeq2: every gene contains at least one zero, cannot compute log geometric means. Solution: added a pseudo-count of 1 to every entry.

The we perform a variance stabilizing transformation using the `DESeq2 vst` function.

3.3 Differential Gene Expression Analysis

The differential gene expression (DGE) analysis was performed using `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 $X = X_1, X_2, ..., X_n$.

The Hardy scale and age were included as covariates in the model by using the following design: `~ HARDY_SCALE + AGE_CONTINUOUS + SEX + SEX:RNA_DI`. For a justification of this design, see Section 4.1.

Samples with missing values for the Hardy scale covariate were removed because variables in the design formula cannot contain missing values (NAs). 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 (648) and because of its simplicity. The missing values represent 0% of the total samples. 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 (source?). 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 quite a lot of variables in the design, we would risk having a low statistical power (very few degrees of freedom).

Afterwards, the continuous variables (`AGE_CONTINUOUS`, and `RNA_DI`) were centered and scaled to improve convergence of the generalized linear model (GLM), as recommended by `DESeq2`.

To identify genes associated with either 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 was used instead (JUSTIFICATION).

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

3.4 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 4.2 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).

4 Results

4.1 Descriptive Analysis

The distribution of all the covariates is shown in Figure 1. We first notice that the number of male sample donor's is twice that of female sample donor's. The two classes are therefore unbalanced (which could cause problems if we wanted to perform classification with the data?). The death classification, based on the Hardy scale, is "0" for the majority of donors, which corresponds to cases on a ventilator immediately before death. The second most common death classification is 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, are in the 50-69 age range. The average RNA integrity number is 6.861, with a minimum at 5.5, and seems to be similarly distributed regardless of the sex. (GOOD? REASONABLE?)

The ischemic time (i.e., "time from death or withdrawal of life-support until the time the sample is placed in a fixative solution or frozen") varies between 83 minutes and 1683 minutes (1.383 hours), with a mean of 607.744. This variable also seems to be similarly distributed across the two sexes.

Concerning the male and female dimorphism indices, we observe that the ones based the RNA-seq transcriptomes are similarly distributed between male and females. The female dimorphism index has a mean of 0.17 and a variance of 0.081, while the male dimorphism index has a mean of 0.182 and a variance of 0.079. Interestingly, the dimorphism indices calculated from histology profiles have a very different distribution. The male histology profile-based dimorphism index seems to have a bimodal distribution, with a mean of 0.119 and a variance of 0.289. The female histology profile-based dimorphism index has a mean of 0.392 and a variance of 0.231.

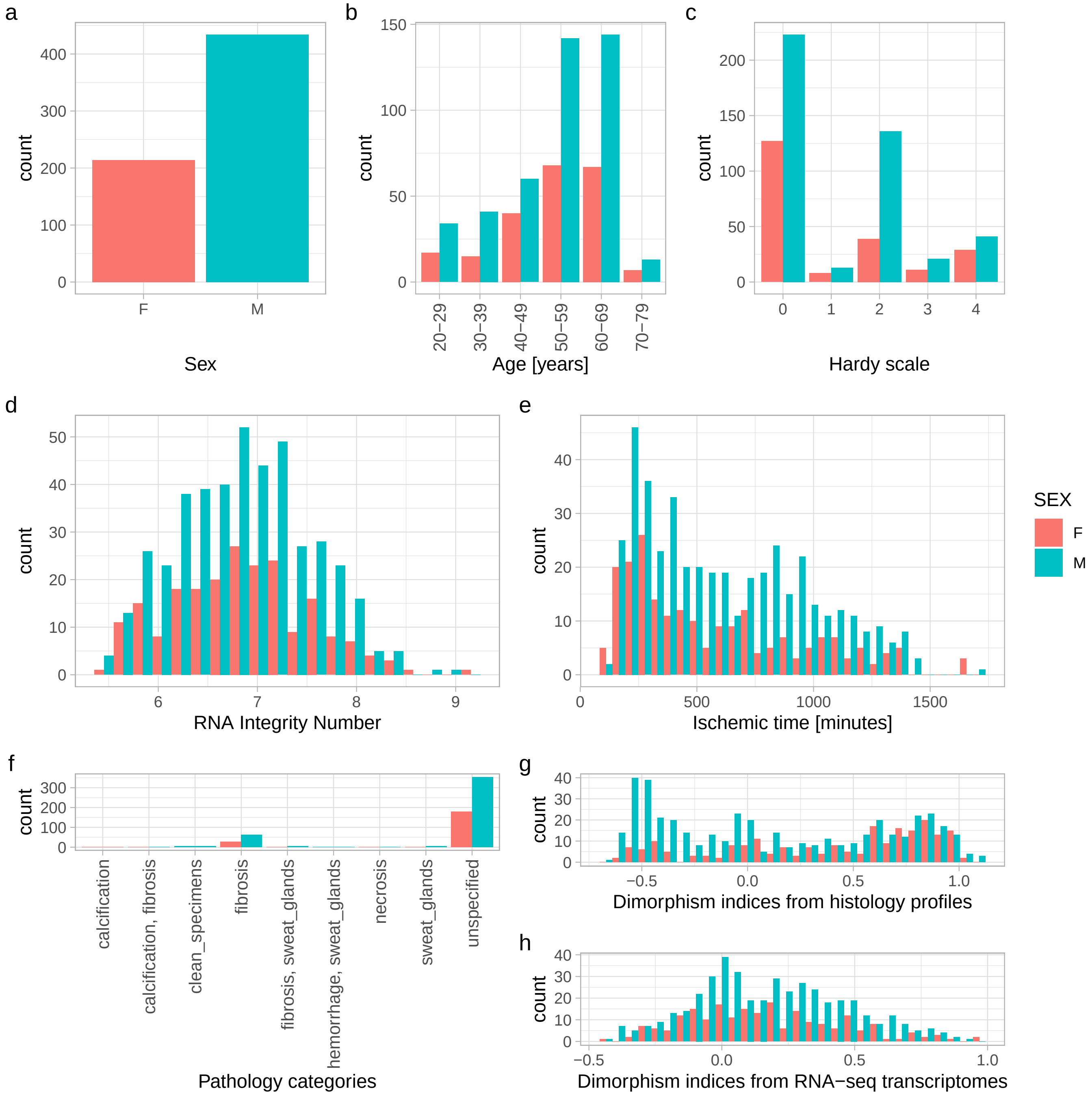

Figure 1: Distribution of the clinical and technical variables, and the dimorphism indices of the samples separated by sex (F = Female, M = Male). **(a)** Number of Female and Male subcutaneous adipose tissue sample donors. **(b)** Distribution of ages (in decades) of the donors. **(c)** Death classification of the samples' donors with the Hardy scale. **(d)** Distribution of the samples' RNA integrity numbers. **(e)** Distribution of the samples' 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 covariates. It also shows the pairwise scatter plots of those variables. We observe that the RNA- and histology-based dimorphism indices are significantly correlated with one another, even though the correlation is quite low. Since the two indices measure the same "phenomenon", it is expected to observe a positive correlation between those variables. However, the low value of this correlation could indicate that there is a high error in one or both measurements of the dimorphism indices, that the two variables do not measure exactly the same "information", or something else. From the pairwise scatter plot, however, it does not appear that the low correlation is due to a nonlinear relationship between the two indices.

We also notice that the age is negatively correlated with the RNA integrity number (RIN) and positively correlated with the ischemic time. The latter variable is negatively correlated with the RNA integrity number. This last result is easily interpretable: 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.

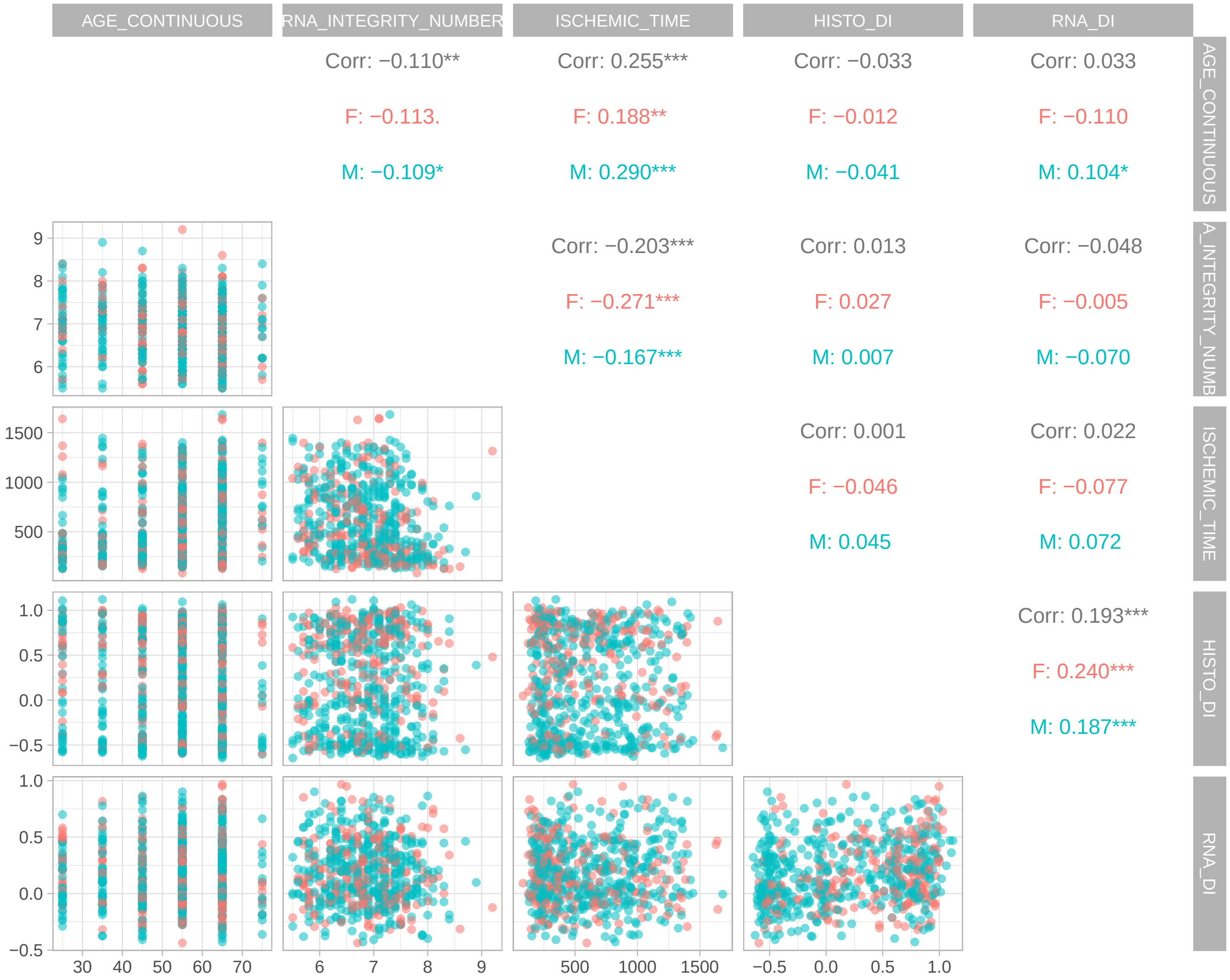

Figure 2: Pairs plot of the continuous clinical and technical covariates for the samples separated by sex. The lower diagonal shows the scatter plots of each pair of variables. The upper diagonal indicates the Pearson correlation coefficients of the variabes considering all the samples (grey), or for 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 are 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 individuals / samples / donors, 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 counts.

First, in Figure 3a, we see a separation of the samples according to their sex along the first principal component. In Figure 3b and Figure 3c, the samples were first separated by their sex before performing the PCA and coloring the samples based on their dimorphism index (calculated from whole RNA-seq transcriptomes). For both plots at the bottom, a clear gradient can be seen along the first principal component. We also notice that the male dimorphism index decreases along PC1, while the female dimorphism increases along PC1. (EXPLANATION?)

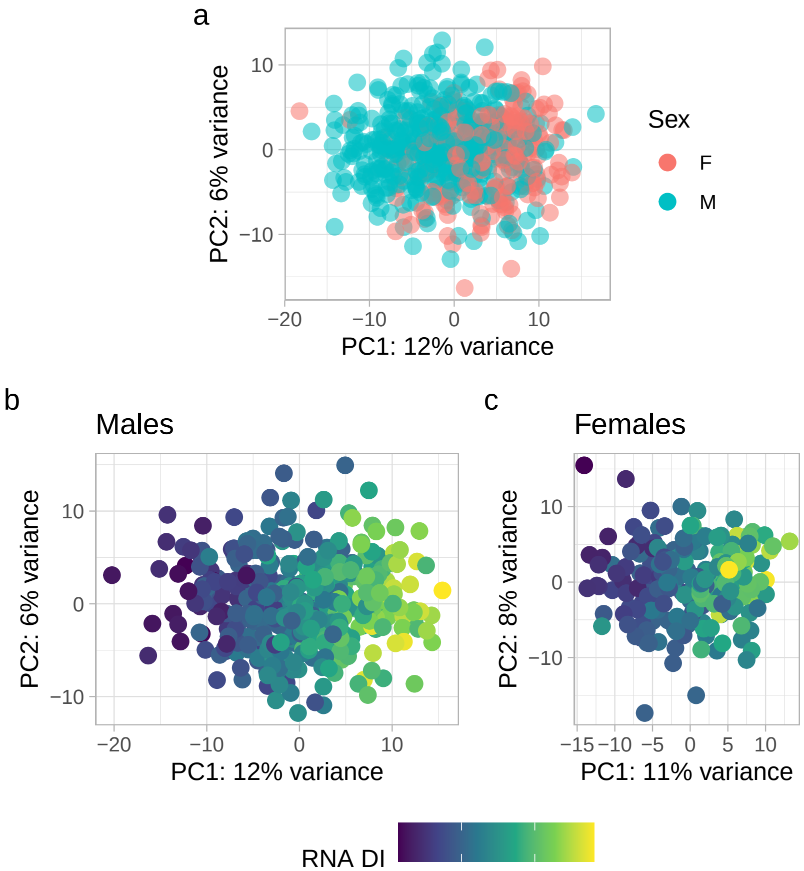

Figure 3: First two principal components of the PCA performed on the 500 most variant genes in the entire data set **(a)** or in male and feamle subsets (**b** and **c**). The samples are colored either by sex **(a)**, by the male RNA-derived dimorphism index **(b)** or female RNA-derived dimorphism index **(c)**.

In Figure 4a and Figure 4b, the same PCA was performed as in Figure 3b and Figure 3c but the samples were colored by the corresponding donor's age. For the males subset, no gradient is visible along PC1 and PC2 according to age. On the other hand, women of age below 40 seem 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 occurs along the first principal component like the male and female dimorphism indices, 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.

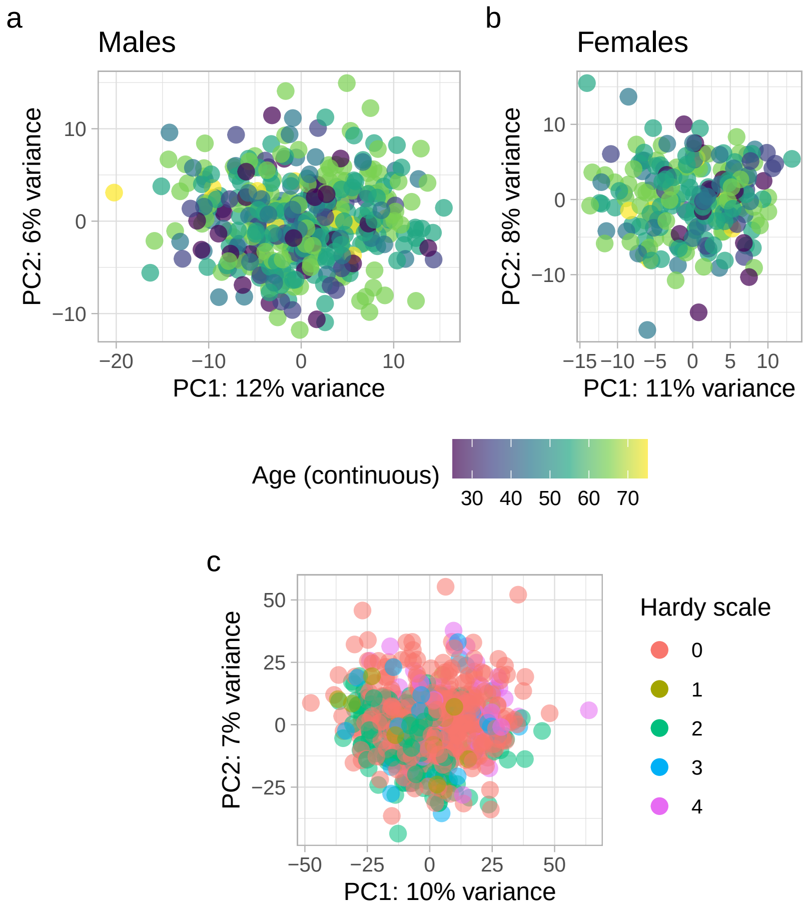

Figure 4: First two principal components of the PCA performed on the 500 most variant genes in the male and female subsets (**a** and **b**) or in the entire data set **(c)**. The samples are colored either by age (**a** and **b**) or by Hardy scale (**c**).

In conclusion, the following technical variables will be 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.

4.2 Differential Gene Expression Analysis

Contrasts that were are most interested in: “SEX_M_vs_F” (differentially expressed genes in males vs. females), “SEXF.RNA_DI” (to identify the genes associated with the female dimorphism index), and “SEXM.RNA_DI” (to find the genes associated with the male dimorphism index). In DESeq2, this corresponds to the design: ~ HARDY_SCALE + AGE_CONTINUOUS + SEX + SEX:RNA_DI (see next section).

4.3 Interpretation of the DI-associated transcriptome

4.3.1 Hormone Receptors Associated with the Dimorphism Indices and with Sex

4.3.2 Gene Set Enrichment Analysis

5 Discussion and Conclusion

6 Annexes

6.1 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
```