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

Youmna Ayadi Kevin Straatman

2025-06-17

Table of contents

1 Abstract

2 Introduction

3 Methods

3.1 Available Data

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 660 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.

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.

3.2 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 (660) and because of its simplicity. The missing values represent 1.818% 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.

3.3 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

Correlation between the continuous variables:

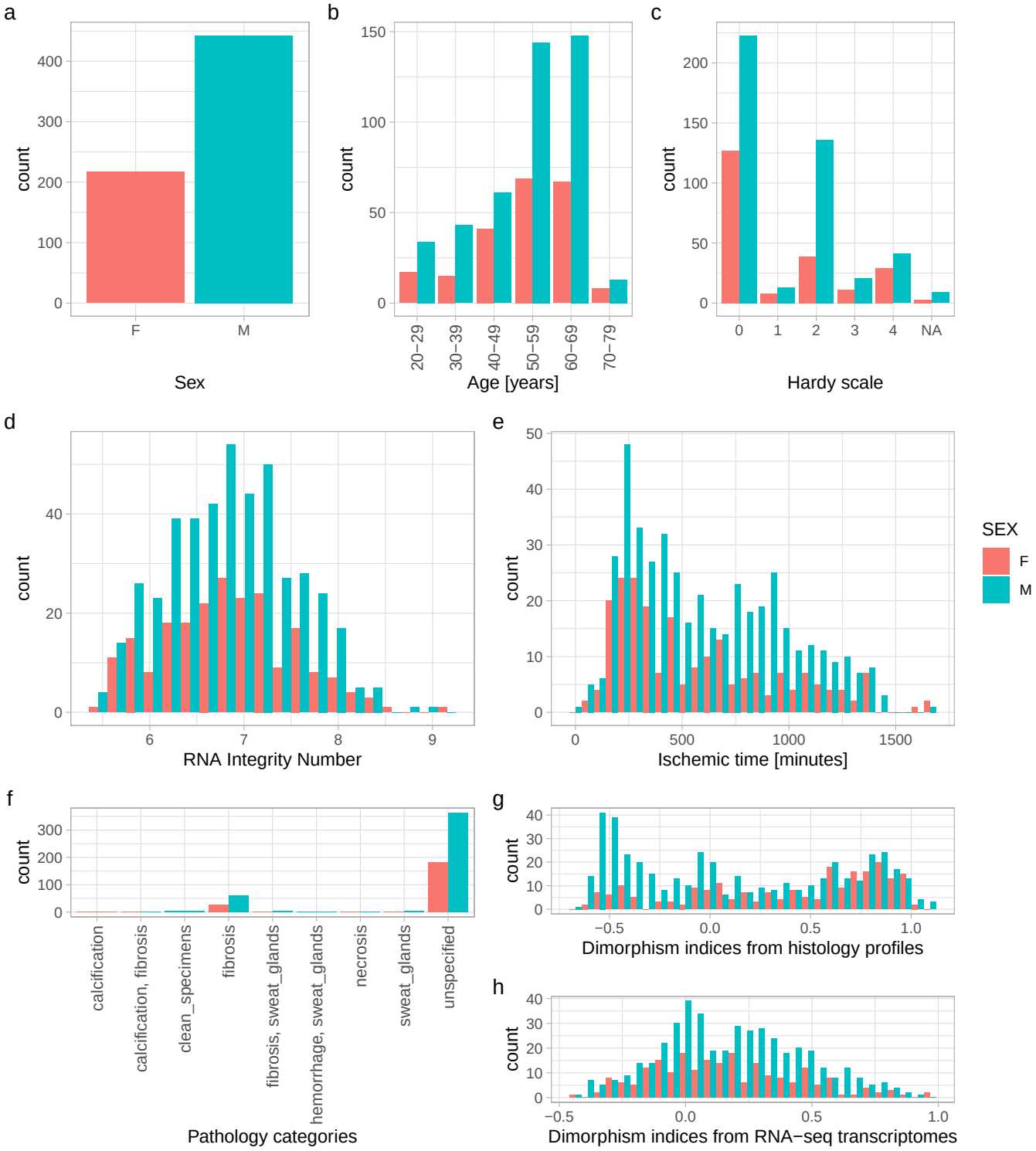

Figure 1: Distribution of the clinical and technical variables, and the dimorphism indices of the samples separated by sex (F = Female, M = Male). **(a)** Number of Female and Male subcutaneous adipose tissue sample donors. **(b)** Distribution of ages (in decades) of the donors. **(c)** Death classification 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.

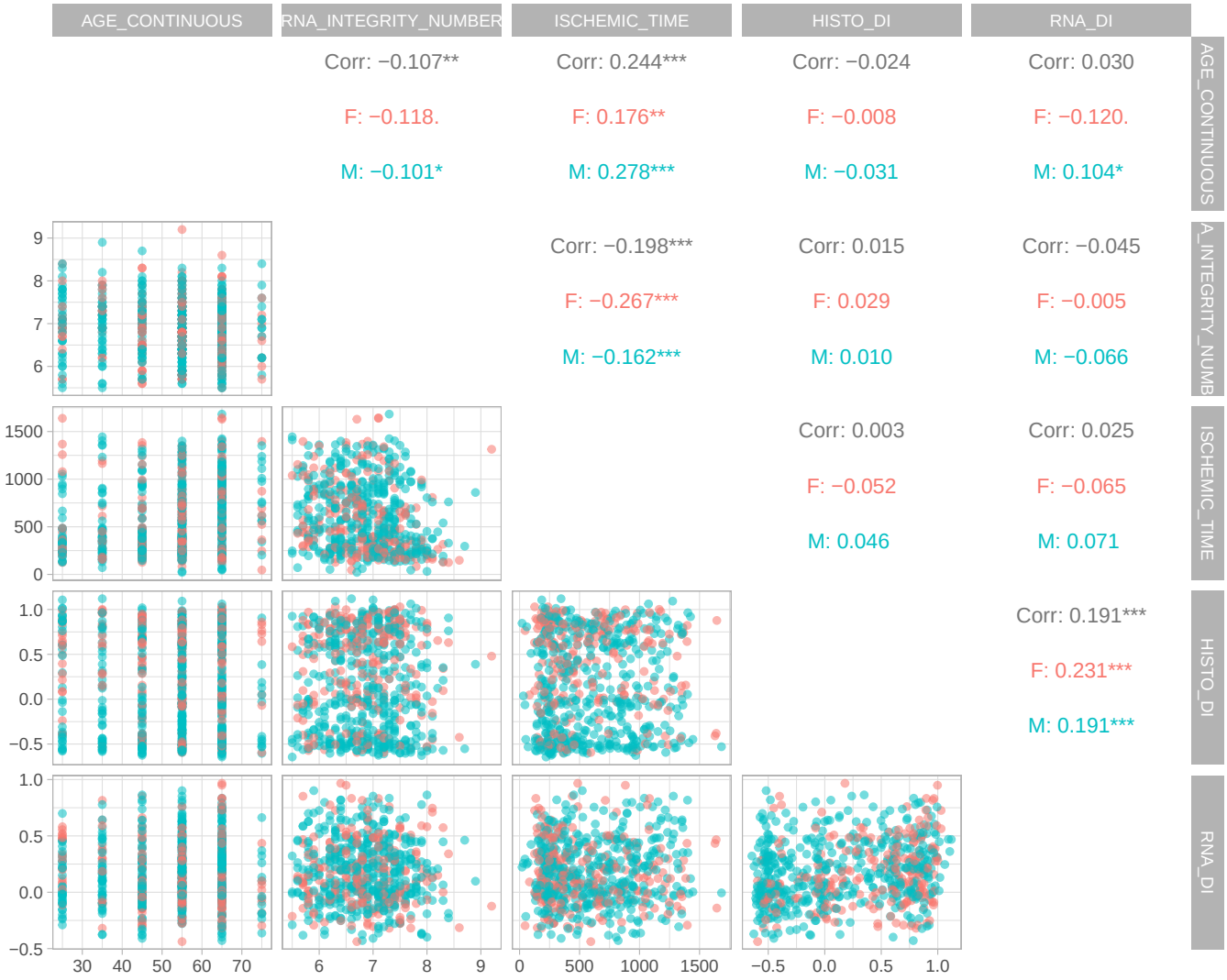

Figure 2: Pairs plot of the continuous clinical and technical covariates for the samples separated by sex. The lower diagonal shows the scatter plots of each pair of variables. The upper diagonal indicates the Pearson correlation coefficients of the variabes considering all the samples (grey), or for 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 $p < 0.10$).

4.2 Differential Gene Expression Analysis

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;

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
```