Functional Enrichment Analysis of the Effects of Long-term Endurance Training on the Sex-Specific Skeletal Muscle Transcriptome

Erik Dassoff

2023-12-08

GitHub Link

Introduction

There is increasing interest in delivering personalized guidance to improve sports performance and overall health. One potential direction is to identify sex-specific differences in exercise adaptations to optimize training regimens for improved health and performance goals. Of particular interest is identifying strategies to optimize cardiorespiratory fitness in different populations, given that cardiorespiratory fitness is increasingly recognized as a better predictor of all-cause mortality than levels of physical activity (Davidson et al. 2018). Additionally, female sex has been indicated to be protective against pathological exercise-induced cardiac adaptations seen at the highest doses of physical activity (Petek et al. 2023). It is useful to consider these differences, to better understand potential mechanisms behind the greater risk profile seen in males. It is further important to identify any potential sex differences, considering that women have historically been under-represented in exercise science research.

Considering that exercise results in widespread physiological adaptations, bioinformatic analysis can be used as a starting point to identify potential sex-specific adaptations, by simultaneously assessing many pathways that are acutely or chronically up or down-regulated by exercise. Landen et al. (2019) previously reviewed many sex-specific alterations in muscle gene expression, DNA methylation, histone modification, and non-coding RNA expression, pointing to differences in muscle proteolysis, oxidative phosphorylation, and tissue remodeling (Landen et al. 2019). Additionally, females are reported to have higher baseline levels of oxidative-type muscle fibers and increased intramuscular triglyceride stores, which also influences oxidative metabolism and energy utilization during exercise (Landen et al. 2019). These have been previously reported to increase oxidative metabolic adaptation in females following endurance exercise training, whereas males were more likely to have increased cardiac adaptations (Landen et al. 2019). These differences could also relate to differential cardiac risk profiles seen between males and females following long-term endurance exercise. However, long-term chronic adaptations have not been thoroughly investigated or are inconsistent.

The objectives of this project are to explore gene expression data to identify sex-specific differentially expressed genes (DEGs) and functionally enriched pathways, by comparing muscle tissue from long-term well-trained participants with untrained controls. Additionally, this project seeks to identify gene expression networks altered by endurance training that differ between these groups. Throughout this process, the goal of this project is also to learn novel tools used in functional enrichment analysis.

Description of Data Set

Chapman et al. (2020) investigated sex-specific differences in the skeletal muscle transcriptome, comparing groups with over 15 years of endurance or strength-training history to untrained controls. This gene expression data contains a wealth of information on chronic adaptations from endurance training, including potential sex-specific differences. The dataset used in this project was obtained on 25 Nov 2023, as an excel file from supplementary differential gene expression (DGE) data (Chapman et al. 2020). The data contains information on gene names, p-values from differential expression analysis, fold change values, and direction of regulation. Prior steps conducted by Chapman et al. (2020) involved DGE analysis and quality control via the DESeq2 workflow. All groups are initially compared via histograms and volcano plots, to explore the entire available data set; however, enrichment analysis focuses specifically on addressing the question of sex-specific endurance training adaptations.

Section 1: Data Acquisition, Exploration, Filtering, and Quality Control

```
# suppressing warnings to suppress warnings from package loading.

# help with comment wrapping:
# https://stackoverflow.com/questions/60163629/knitr-to-pdf-not-wrapping-comments

# attributions: This script was adapted from vignettes for enhanced volcano
# plots (Blighe et al. 2018) and functional enrichment analysis via the
# pathfindR package (Ulgen et al. 2019). Additionally, data was obtained from
# Chapman et al. (2020) for differential gene expression analysis. Additional R
# tutorials were consulted throughout, as indicated via hyperlinks throughout
# the script.----

# purpose of script: This script is intended to investigate differential gene
# expression and to investigate functional comparisons between two groups of
# differential expression data.

# loading libraries----
library(readxl)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(data.table)
# BiocManager::install('EnhancedVolcano')
library(EnhancedVolcano)
library(gridExtra)
library(pathfindR)
library(microbenchmark)
library(utils)
library(cowplot)


# setting adjustable parameters---- p-value for statistical significance,
# considering p-value already adjusted for false discovery rate. 0.05 is
# selected to consider a level of statistical significance that limits the
# false discovery rate of all significant results to less than 5%.
```

```r
pCutoff <- 0.05

# cutoff for fold change for volcano plots. Two is selected to highlight
# reasonably large differentially expressed genes that are likely to have high
# biological relevance.
FCcutoff <- 2

# p-value cutoff for significant enrichment. 0.05 selected as a standard level
# and to not be overly exclusive.
enrichThresh <- 0.05

# adjustment method for enrichment analysis (i.e., bonferroni or fdr). Using
# fdr to be consistent with differential expression analysis by originial
# authors.
adjMethod <- "fdr"

# number of iterations for protein sub-network search. Start with ten and
# increase for greater coverage of functional pathways.
iterations <- 10

# gene set for enrichment analysis (choose from: 'Kegg', 'Reactome',
# 'BioCarta', 'GO-all', 'Go-BP', 'GO-CC', 'GO-MF', 'cell_markers', 'mmu_KEGG').
# Starting with KEGG, because it contains biological pathways of interest with
# good interpretability.
geneSet <- "KEGG"

# setting the protein interaction network database to search to create gene
# interaction networks (choose from: 'Biogrid', 'STRING', 'KEGG', 'GeneMania',
# 'IntAct', 'mmu_STRING'). Starting with KEGG to be consistent with enrichment
# analysis choice.
pinName <- "KEGG"

# setting greedy search algorithm, since it is useful to identify multiple
# subnetworks and doesn't focus on trying to find one optimized subnetwork as
# others do. This can help with identifying un-linked pathways. It is also
# recommended for computational efficiency, so it was also chosen for this
# purpose.
searchMethod <- "GR"

# input the terms that are desired to be compared.
comparison1 <- "FEvsFC"
comparison2 <- "MEvsMC"


# loading dataset---- Muscle tissue differential gene expression data from the
# following article as a downloadable excel file in the xlsx format:
# https://www.sciencedirect.com/science/article/pii/S2211124720307890#abs0015

# insert path to file
path <- "C:/Users/Erik/OneDrive - University of Guelph/Documents/BINF/enduranceTrainingDEG.xlsx"

# the data is being read into R with the readxl package. lapply is used to load
# every sheet in the excel file.
```

```
DEGlist <- lapply(excel_sheets(path), read_excel, path = path)

# check
class(DEGlist)
```

```
[1] "list"
```

```
# finding sheet names. Help:
# https://readxl.tidyverse.org/reference/excel_sheets.html
sheetNames <- excel_sheets(path)

# renaming sheet names in list
names(DEGlist) <- sheetNames

# check
names(DEGlist)
```

```
[1] "FEvsFC" "MEvsMC" "MCvsFC" "MSvsME" "MEvsFE" "MSvsMC"
```

```
# combining all comparisons into one data frame to make it easier to create
# multiple figures in one grid using facet wrap, and for other downstream
# processing.
allDEGcomparisons <- rbindlist(DEGlist, idcol = "comparison")

# checking that all group comparisons are in the new data frame
unique(allDEGcomparisons$comparison)
```

```
[1] "FEvsFC" "MEvsMC" "MCvsFC" "MSvsME" "MEvsFE" "MSvsMC"
```

```
head(allDEGcomparisons)
```

```
   comparison  gene_name Log2FoldChange  P-Adjusted Direction
1:     FEvsFC AC131097.2       2.799144 0.001136582        UP
2:     FEvsFC       CRYM       2.323224 0.009487791        UP
3:     FEvsFC      SCN1A       2.319971 0.002326664        UP
4:     FEvsFC     DIRAS1       2.216412 0.000160436        UP
5:     FEvsFC AL139011.2       2.154747 0.001265842        UP
6:     FEvsFC     HS6ST3       2.081391 0.000401908        UP
```

```
# adding negative log10 p-values for downstream analysis
allDEGcomparisons$negLogP <- -log10(allDEGcomparisons$`P-Adjusted`)

# checking that column was added and that values are correct
names(allDEGcomparisons)
```

```
[1] "comparison"     "gene_name"      "Log2FoldChange" "P-Adjusted"
[5] "Direction"      "negLogP"
```

```r
all.equal(allDEGcomparisons$negLogP, -log10(allDEGcomparisons$`P-Adjusted`))
```

```
[1] TRUE
```

```r
# selecting variables to set up format for pathfindR downstream
selectForPathfindR <- function(x) {
    x %>%
        select(gene_name, Log2FoldChange, `P-Adjusted`)
}

DEGlistPathfindR <- lapply(DEGlist, selectForPathfindR)

# check that column names are as expected
sapply(DEGlistPathfindR, function(x) {
    names(x)
})
```

```
     FEvsFC           MEvsMC           MCvsFC           MSvsME
[1,] "gene_name"      "gene_name"      "gene_name"      "gene_name"
[2,] "Log2FoldChange" "Log2FoldChange" "Log2FoldChange" "Log2FoldChange"
[3,] "P-Adjusted"     "P-Adjusted"     "P-Adjusted"     "P-Adjusted"
     MEvsFE           MSvsMC
[1,] "gene_name"      "gene_name"
[2,] "Log2FoldChange" "Log2FoldChange"
[3,] "P-Adjusted"     "P-Adjusted"
```

```r
# exploring data----

# creating function to explore data in every element of the list
explore <- function(x) {
    class = class(x)
    names = names(x)
    dimensions = dim(x)
    return(list(class = class, names = names, dimensions = dimensions))
}

sapply(DEGlist, explore)
```

```
           FEvsFC      MEvsMC      MCvsFC      MSvsME      MEvsFE
class      character,3 character,3 character,3 character,3 character,3
names      character,4 character,4 character,4 character,4 character,4
dimensions integer,2   integer,2   integer,2   integer,2   integer,2
           MSvsMC
class      character,3
names      character,4
dimensions integer,2
```

```r
# creating function to better understand the data values
dataStats <- function(x) {
    pValStats = summary(x$"P-Adjusted")
    FCstats = summary(x$Log2FoldChange)
```

```
    numGeneNames = length(unique(x$gene_name))
    directions = table(x$Direction)
    return(list(`P Value Stats` = pValStats, `Fold Change Stats` = FCstats, `Number of Unique Genes` = n
        `Change Directions` = directions))
}

# using sapply to shorten output. lapply was used to view all features
# originally.
sapply(DEGlist, dataStats)
```

```
                       FEvsFC           MEvsMC           MCvsFC
P Value Stats          summaryDefault,6 summaryDefault,6 summaryDefault,6
Fold Change Stats      summaryDefault,6 summaryDefault,6 summaryDefault,6
Number of Unique Genes 1711             1097             452
Change Directions      table,2          table,2          table,2
                       MSvsME           MEvsFE           MSvsMC
P Value Stats          summaryDefault,6 summaryDefault,6 summaryDefault,6
Fold Change Stats      summaryDefault,6 summaryDefault,6 summaryDefault,6
Number of Unique Genes 241              135              26
Change Directions      table,2          table,2          table,2
```

```
# checking to see if there are the same number of unique gene names as gene
# names overall for each list element. If so, these can serve as the unique
# identifiers for each element in each list.
sapply(DEGlist, function(x) {
    length(x$gene_name) == length(unique(x$gene_name))
})
```

```
FEvsFC MEvsMC MCvsFC MSvsME MEvsFE MSvsMC
  TRUE   TRUE   TRUE   TRUE   TRUE   TRUE
```

```
# advice for quality checking for differential gene expression analysis:
# https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/
# lecture_notes/08_practical_DE.pdf

# help with facetwrap:http://zevross.com/blog/2019/04/02/
# easy-multi-panel-plots-in-r-using-facet_wrap-and-facet_grid-from-ggplot2/

# plotting histogram of p-values for all comparisons with color to identify
# which are up- or down-regulated. Any comparisons with only up- or
# down-regulated genes are likely to not be of high quality.
GGhist <- ggplot(allDEGcomparisons, aes(`P-Adjusted`, fill = Direction))
GGhist + geom_histogram(binwidth = 0.001, col = I("grey10")) + labs(y = "Number of Differentially Expres
    x = "Adjusted P-Value") + theme_classic() + facet_wrap(~allDEGcomparisons$comparison) +
    scale_fill_viridis_d(option = "D")
```

```
# creating volcano plots to visualize significantly differentially expressed
# genes with high significance

# creating a function for the volcano plot, to visualize significantly
# differentially expressed genes with high significance. Formatting is adjusted
```

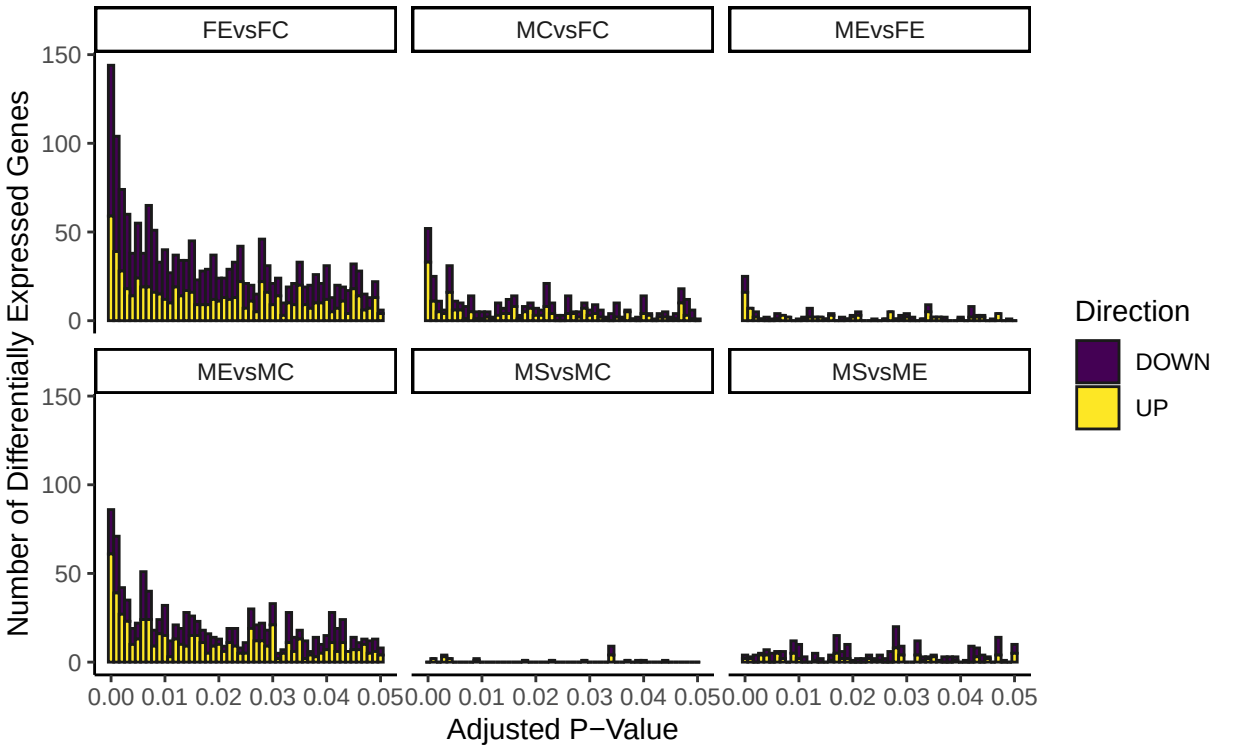

Figure 1: Amount of significantly differentially regulated genes by p-value and group comparison. Abbreviations: F - Female; M - Male; E - Endurance; C - Control.

```
# to fit in the Rmarkdown output.
volcanoPlot <- function(data, title) {
    EnhancedVolcano(data, lab = data$gene_name, x = "Log2FoldChange", y = "P-Adjusted",
        title = title, titleLabSize = 14, legendLabSize = 14, legendIconSize = 6,
        axisLabSize = 14, legendPosition = "top", pCutoff = pCutoff, FCcutoff = FCcutoff,
        pointSize = 3, labSize = 4, boxedLabels = TRUE, drawConnectors = TRUE, gridlines.minor = FALSE)
}

# plotting result for each comparison.
p1 <- volcanoPlot(DEGlist$FEvsFC, "FE vs FC")
p2 <- volcanoPlot(DEGlist$MEvsMC, "ME vs MC")
p3 <- volcanoPlot(DEGlist$MCvsFC, "MC vs FC")
p4 <- volcanoPlot(DEGlist$MSvsME, "MS vs ME")
p5 <- volcanoPlot(DEGlist$MEvsFE, "ME vs FE")
p6 <- volcanoPlot(DEGlist$MSvsMC, "MS vs MC")

# change colours to viridis scale
viridis <- function(p) {
    p + scale_color_viridis_d(option = "D")
}

p1 <- viridis(p1)
p2 <- viridis(p2)
p5 <- viridis(p5)
p3 <- viridis(p3)
```

```
# arranging selected figures into a grid
grid.arrange(p1, p2, p5, p3, ncol = 2)
```

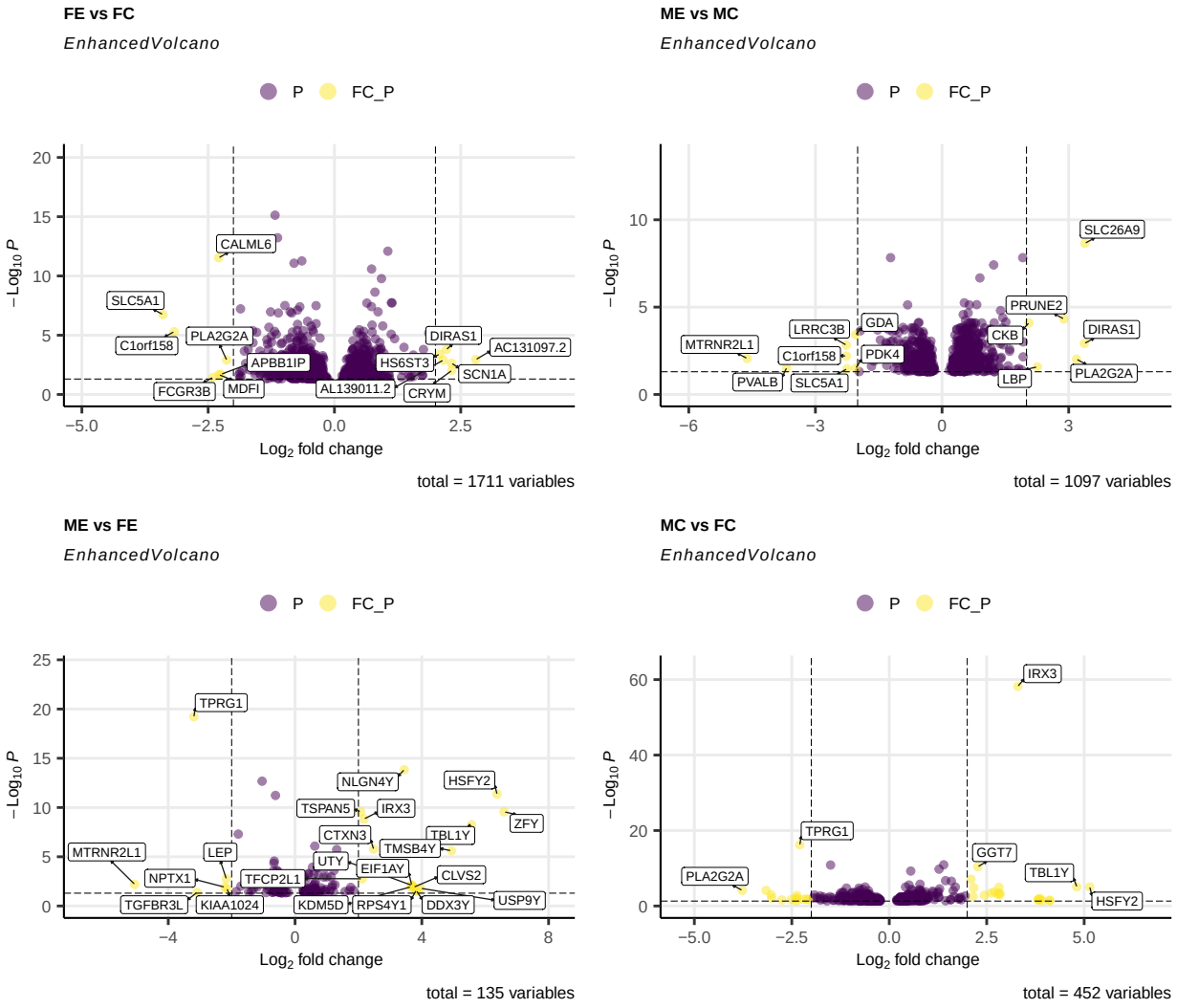

Figure 2: Volcano plots of differentially expressed genes for all combinations of male and female endurance athletes and their controls. Abbreviations: F - Female; M - Male; E - Endurance; C - Control; P - Significant (adj. $P < 0.05$) differential expression; FC_P - Significant differential expression (adj. $P < 0.05$) with fold change > 2.

Main Software Tools

The main software tool used in this analysis was pathfindR (Ulgen et al. 2019). The associated vignette provided guidance on available features, options, and code formatting. This analysis extended on the vignette by writing new composite functions and selecting tailored settings for this analysis. This software was chosen, since it has been used in well-cited papers in the nutrition and exercise science fields and provides novel improvements to functional enrichment analysis. The approach differs from the more common over-representation analysis (ORA) by collectively accounting for the significance of differentially expressed genes

by first mapping them onto optimized protein networks before enrichment analysis. This helps to better account for gene interactions and create more realistic representations of enriched pathways. I also found the tool to provide easily-interpretable visualizations and high flexibility to adjust the analysis to preferred settings. Examples of customizability include the ability to select algorithms used for identifying active gene subnetworks according to use-case and computational availability, the ability to access multiple types of protein interaction networks/gene sets for enrichment analysis, and the ability to adjust some plot features with GGplot code. Still, the additional optimization steps beyond ORA could present higher computational demands, especially for greater numbers of comparisons or longer gene lists.

Section 2: Main Analysis

```
# suppressing intermediate figure output with fig.show = hide

# full tutorial:
# https://cran.r-project.org/web/packages/pathfindR/vignettes/intro_vignette.html

# converting to data frame for all comparisons, which is an essential class
# structure to be able to run downstream pathfindR analysis
DEGlistPathfindR <- lapply(DEGlistPathfindR, function(x) {
    as.data.frame(x)
})

# check
sapply(DEGlistPathfindR, function(x) {
    class(x)
})
```

```
      FEvsFC       MEvsMC       MCvsFC       MSvsME       MEvsFE       MSvsMC
"data.frame" "data.frame" "data.frame" "data.frame" "data.frame" "data.frame"
```

```
# turning off scientific notation to match pathway analysis format
options(scipen = 999)

# check that values are not in scientific notation
summary(DEGlistPathfindR$FEvsFC$`P-Adjusted`)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.004723 0.015112 0.018202 0.029718 0.049961
```

```
# creating function for running pathfindR, which will be used more than once
# downstream. Not plotting the enrichment chart, so that it can be plotted
# separately later.
pathfindR <- function(data) {
    run_pathfindR(data, adj_method = adjMethod, gene_sets = geneSet, pin_name_path = pinName,
        search_method = "GS", enrichment_threshold = enrichThresh, plot_enrichment_chart = FALSE)
}

# creating custom function for selecting key outputs of interest: clustering
# pathway enrichment with agglomerative hierarchical clustering and plotting
# the enriched terms as well as the gene interaction networks for each
```

```r
# comparison.
enrichAnalysis <- function(data) {
    result <- run_pathfindR(data)
    clusteredEnrichment <- cluster_enriched_terms(result, plot_clusters_graph = FALSE)
    termGeneGraph <- term_gene_graph(result, use_description = TRUE)
    return(list(result = result, enrichment = clusteredEnrichment, termPlot = termGeneGraph))
}

# creating pre-specified search list
df1 <- DEGlistPathfindR[[comparison1]]
df2 <- DEGlistPathfindR[[comparison2]]
inputList <- list(df1, df2)
remove(df1, df2)

# checking that the list contains two data frames and viewing
class(inputList)
```

```
[1] "list"
```

```r
length(inputList)
```

```
[1] 2
```

```r
class(inputList[[1]])
```

```
[1] "data.frame"
```

```r
head(inputList[[1]])
```

```
   gene_name Log2FoldChange  P-Adjusted
1 AC131097.2       2.799144 0.001136582
2       CRYM       2.323224 0.009487791
3      SCN1A       2.319971 0.002326664
4     DIRAS1       2.216412 0.000160436
5 AL139011.2       2.154747 0.001265842
6     HS6ST3       2.081391 0.000401908
```

```r
# running analysis function through pre-selected comparisons
combinedResult <- lapply(inputList, enrichAnalysis)

# viewing tabular results view(combinedResult[[1]]$result)
# view(combinedResult[[1]]$enrichment) view(combinedResult[[2]]$enrichment)

# creating a subset of samples with fold enrichment >2, low p-values, and only
# the representative terms for each cluster, to simplify figure by not having
# several terms within each cluster.
selectedClusters1 <- subset(combinedResult[[1]]$enrichment, Status %in% "Representative" &
    Fold_Enrichment > 1.5 & lowest_p < 0.01)

selectedClusters2 <- subset(combinedResult[[2]]$enrichment, Status %in% "Representative" &
```

```
    Fold_Enrichment > 1.5 & lowest_p < 0.01)

# subsetting for just cluster 1 to gain detailed information, since this was a
# unique enrichment for females
selectedClusters1b <- subset(combinedResult[[1]]$enrichment, Cluster == 1)

# subsetting for top male term to match format
selectedClusters2b <- subset(combinedResult[[2]]$enrichment, Cluster == 1)

# checking that subsetting was performed
length(combinedResult[[1]]$enrichment$ID) > length(selectedClusters1$ID)
```

```
[1] TRUE
```

```
unique(selectedClusters1b$Cluster)
```

```
[1] 1
```

```
# plotting enrichment analysis
e1 <- enrichment_chart(selectedClusters1, plot_by_cluster = TRUE)
e2 <- enrichment_chart(selectedClusters2, plot_by_cluster = TRUE)
e1b <- enrichment_chart(selectedClusters1b)
e2b <- enrichment_chart(selectedClusters2b)

# help with plot_grid: https://wilkelab.org/cowplot/articles/plot_grid.html
grid1 <- plot_grid(e1, e2, labels = c(comparison1, comparison2))

plot_grid(grid1, e1b, e2b, labels = c("", paste(comparison1, "Ribosome Pathway"),
    paste(comparison2, "TCA Cycle Pathway")), nrow = 3, rel_heights = c(10, 4, 4),
    scale = c(1, 0.8, 0.8))
```

```
# performing comparison between results and looking at what fields are
# available. Plot is set to false to not automatically plot, since the
# combined_results_graph function has more options.
resultComparison <- combine_pathfindR_results(result_A = combinedResult[[1]]$result,
    result_B = combinedResult[[2]]$result, plot_common = FALSE)
names(resultComparison)

# replacing software-given terms with experimental labels to create more
# informative labels for the plot.
status <- str_replace(string = resultComparison$status, pattern = "A only", replacement = "FE vs FC only
status <- str_replace(string = status, pattern = "B only", replacement = "ME vs MC only")
resultComparison$status <- status

# check
unique(resultComparison$status)

# order by p-value so the top results with the lowest p-value are plotted below
procResultComparison <- resultComparison[order(resultComparison$combined_p), ]

# checking that p-values are in order
```

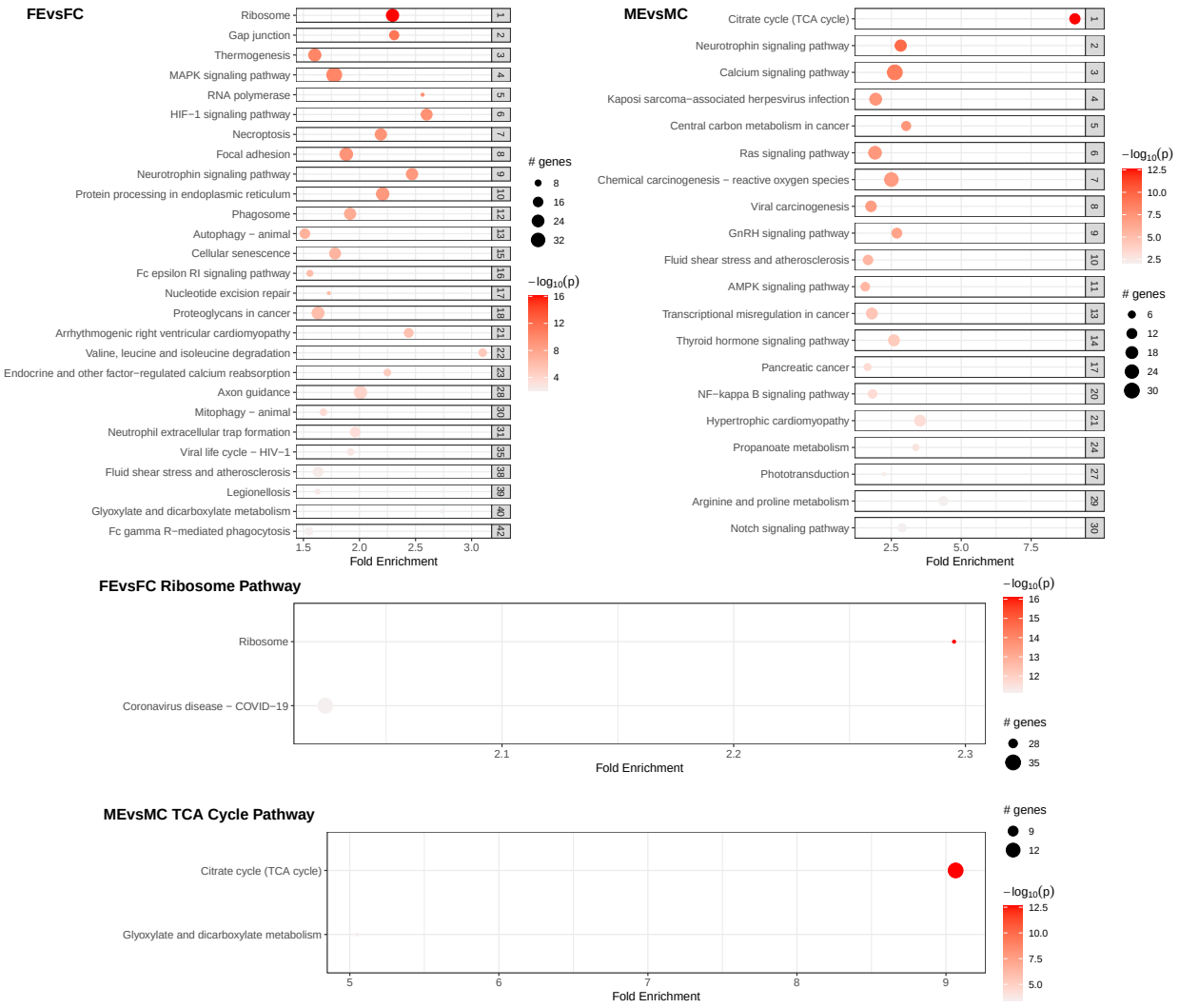

Figure 3: Functional enrichment analysis for female and male endurance athletes compared to their respective sedentary controls. All terms with enrichment p-values < 0.01 and fold change values above 1.5 are included. Abbreviations: F - Female; M - Male; E - Endurance; C - Control.

```
head(procResultComparison$combined_p)

# graphing combined gene-network results. Adding full descriptive terms to
# nodes and sizing according to p value. Plotting only the top 8 terms for
# readability.
combiNetPlot <- combined_results_graph(procResultComparison, use_description = TRUE,
    node_size = "p_val", selected_terms = procResultComparison$Term_Description[1:8])

# adjusting to make more accessible with viridis color scale
combiNetPlot + scale_colour_viridis_d(option = "D")
```

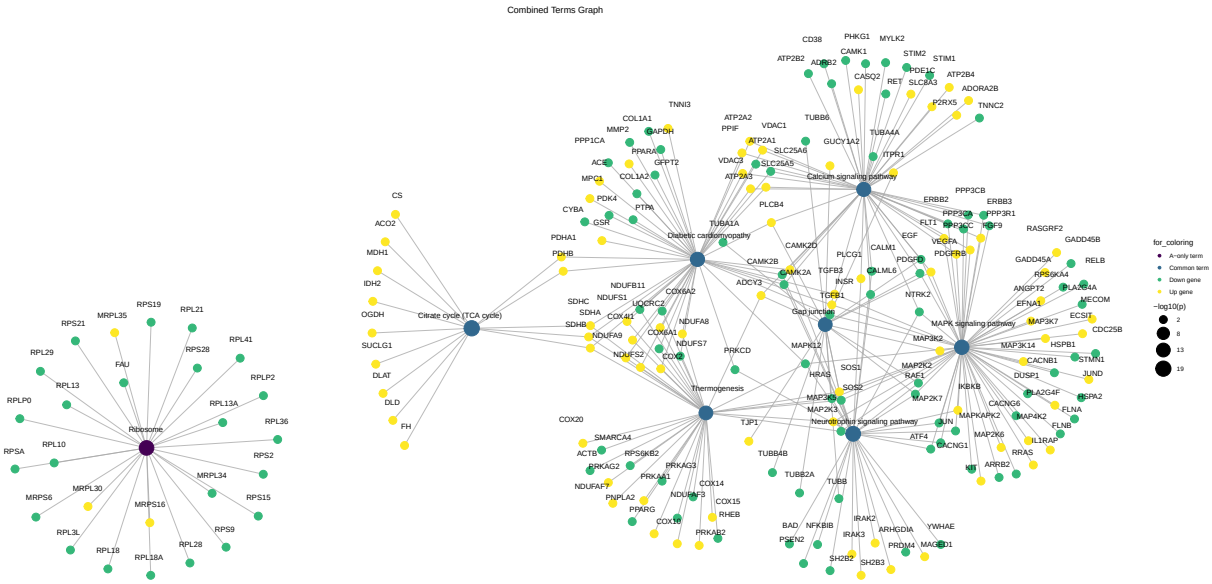

Figure 4: Comparison of functional enrichment networks between female and male endurance athletes, relative to their respective controls, and considering the top eight pathways with the lowest combined p-values. A-only represents female-only differential expression.

Results and Discussion

The objective of this study was to explore any sex-specific adaptations to long-term endurance exercise. While the results did not largely indicate differences in the most significant processes affected by endurance training in males and females, they provide preliminary evidence that there may be differences in the extent to which various pathways are activated. Histograms (Figure 1) demonstrated that there were no non-significant genes with false-discovery-rate adjusted p-values greater than 0.05, confirming that Chapman et al. (2020) had already filtered out non-significant DEGs. Additionally, results for each comparison were balanced for both up and down-regulated genes, which helps to indicate that genes were indeed differentially expressed, as opposed to merely having differences due to total gene counts per sample. Lastly, these results clearly demonstrated that the greatest numbers of significant DEGs were found for comparisons between male and female endurance athletes versus their respective controls. Volcano plots (Figure 2) further highlighted significant genes with large fold changes between groups. These were also used to examine whether some differences in gene expression between male and female endurance athletes are due to general sex-specific differences rather than sex-specific training adaptations; some, but not all, significant DEGs were similar (e.g., TPRG1, HSFY2, etc.) when comparing female versus male endurance athletes with female versus male

untrained controls. Still, the top functionally enriched pathways in resting muscle biopsies of endurance-trained participants differed between males and females. Males demonstrated the highest enrichment for the citrate cycle, neurotrophin signaling, and calcium signaling pathways, whereas females demonstrated the highest enrichment for the ribosome, gap junction, and thermogenesis pathways (Figure 3). However, when integrating gene interaction maps (Figure 4), it was seen that most pathways followed the same patterns of regulation between males and females, with the exception that ribosome pathways were uniquely affected in females. Upon further investigation, the upregulated ribosome pathway also contained gene regulation related to the Coronavirus disease (Figure 3).

This result highlights some of the limitations of this analysis. It is possible that exercise stimulates some pathways seen in disease states, although these would be expected immediately after performing exercise, versus in the resting state. It is additionally possible that some participants received a vaccination, were infected with, or were recovering from Coronavirus, creating misleading results. In general, these results are limited by comparing different groups of participants, such that differences may be due to confounding influences other than training or sex-specific differences. Additionally, differences in heart-related pathways were seen, although disease states were up-regulated in both males and females. Still, this does not necessarily indicate disease progression, as some non-pathological cardiac adaptations closely resemble pathological ones (Martinez et al. 2021) and may share some gene expression networks. In addition, skeletal muscle tissue was sampled and likely does not reflect cardiovascular adaptations. These examples highlight some challenges in interpreting true biological implications from these results. Still, there were overarching similarities between the results of this analysis and that of the original article (Chapman et al. 2020). Both analyses saw up-regulation of oxidative processes, reflected through the enrichment in the citrate cycle or other pathways (i.e., thermogenesis, production of reactive oxygen species, etc.), which would be expected of endurance training adaptations. Unexpectedly, females did not have clearly increased oxidative adaptations relative to male counterparts, as seen previously (Landen et al. 2019).

There may be a number of reasons for these differences to prior results. Differences seen between analyses reflect the use of GO terms, gene sets from MSigDB, and the PIANO R package by Chapman et al. (2020), while the newer pathfindR was used in this analysis. Indeed, future work could perform additional sensitivity analysis to investigate commonalities in results, even when considering different analytical choices (e.g., choice of gene sets and protein interaction networks or the number of sub-network search iterations). In-depth analysis of individual pathways may also be done to describe detailed mechanisms of action. Future work should additionally be replicated in multiple groups and/or using longitudinal data to help identify common biological responses, thus helping to disentangle true effects from confounding factors. A greater consensus is likely to be achieved following these steps, leading to a better clarification in sex-specific endurance training adaptations. Additional research is especially needed to clarify the role of oxidative metabolic adaptations, which could also relate to differential cardiac adaptations.

Reflection

When preparing this final project, I ran into several warning messages when trying out new packages and coding techniques. At times, it seemed like I was continuously running into these challenges. In general, this course has been a continual learning experience, where I never felt at the beginning of an assignment that I knew how to do it; there was always a lot of trial and error. I am someone who always seeks out challenges and I was glad that I tried to push myself in this course, but it can also be difficult to always feel like you're frequently running into challenges (i.e., R warning messages). However, there were also a few moments where I realized that I wasn't getting the errors I had in the past, and that meant I must be able to figure out this one too. I took a break, came back, and systematically tested out possible solutions, ultimately arriving at ideas that worked. Looking back now after finishing this assignment, I'm pretty amazed at how much I've learned during just one term, and with many other projects going on at the same time. There were a lot of things that were very new to me when I started this course, but I've managed to learn them. In addition to the specific skills I've learned, a lesson has been that it's worthwhile to challenge myself, not get too frustrated when it doesn't work, and trust that I'm always improving, even when it's hard to see that progress in the moment; with systematic patience and giving myself enough time, I can learn anything.

Acknowledgments

No outside help was consulted in this assignment, except for cheat sheets for R (Rmarkdown, GGplot, etc.), linked tutorials, and cited vignettes.

References

Blighe, K, S Rana, and M Lewis. 2018. "EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling." https://github.com/kevinblighe/EnhancedVolcano.

Chapman, Mark A., Muhammad Arif, Eric B. Emanuelsson, Stefan M. Reitzner, Maléne E. Lindholm, Adil Mardinoglu, and Carl Johan Sundberg. 2020. "Skeletal Muscle Transcriptomic Comparison between Long-Term Trained and Untrained Men and Women." *Cell Reports* 31 (12): 107808. https://doi.org/10.1016/J.CELREP.2020.107808.

Davidson, Taryn, Baruch Vainshelboim, Peter Kokkinos, Jonathan Myers, and Robert Ross. 2018. "Cardiorespiratory Fitness versus Physical Activity as Predictors of All-Cause Mortality in Men." *American Heart Journal* 196: 156–62. https://doi.org/10.1016/J.AHJ.2017.08.022.

Landen, Shanie, Sarah Voisin, Jeffrey M. Craig, Sean L. McGee, Séverine Lamon, and Nir Eynon. 2019. "Genetic and Epigenetic Sex-Specific Adaptations to Endurance Exercise." *Epigenetics* 14 (6): 523. https://doi.org/10.1080/15592294.2019.1603961.

Martinez, Matthew W., Jonathan H. Kim, Ankit B. Shah, Dermot Phelan, Michael S. Emery, Meagan M. Wasfy, Antonio B. Fernandez, et al. 2021. "Exercise-Induced Cardiovascular Adaptations and Approach to Exercise and Cardiovascular Disease: JACC State-of-the-Art Review." *Journal of the American College of Cardiology* 78 (14): 1453–70. https://doi.org/10.1016/J.JACC.2021.08.003.

Petek, Bradley J., Eugene H. Chung, Jonathan H. Kim, Rachel Lampert, Benjamin D. Levine, Dermot Phelan, Alfred Danielian, et al. 2023. "Impact of Sex on Cardiovascular Adaptations to Exercise: JACC Review Topic of the Week." *Journal of the American College of Cardiology* 82 (10): 1030–38. https://doi.org/10.1016/J.JACC.2023.05.070.

Ulgen, Ege, Ozan Ozisik, and Osman Ugur Sezerman. 2019. "PathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks." *Frontiers in Genetics* 10 (SEP). https://doi.org/10.3389/FGENE.2019.00858.

Helpful Tutorials

"Arranging Plots in a Grid • Cowplot." n.d. Accessed December 8, 2023. https://wilkelab.org/cowplot/articles/plot_grid.html.

Dündar, Friederike. 2020. "Performing Differential Gene Expression Analysis." <https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_not es/08_practical_DE.pdf>.

"Easy Multi-Panel Plots in R Using Facet_wrap() and Facet_grid() from Ggplot2 | Technical Tidbits From Spatial Analysis & Data Science." n.d. Accessed December 8, 2023. http://zevross.com/blog/2019/04/02/easy-multi-panel-plots-in-r-using-facet_wrap-and-facet_grid-from-ggplot2/.

"List All Sheets in an Excel Spreadsheet — Excel_sheets • Readxl." n.d. Accessed December 8, 2023. https://readxl.tidyverse.org/reference/excel_sheets.html.

"R - Knitr to PDF Not Wrapping Comments - Stack Overflow." n.d. Accessed December 8, 2023. https://stackoverflow.com/questions/60163629/knitr-to-pdf-not-wrapping-comments.