CS 509 - Project 2

Differential Gene Expression in the Heart of the House Mouse Based on Age and Sex

Indronil Bhattacharjee (IB) and Erica Flores (EF)

Abstract

Sexual dimorphism is a well-known phenomenon inherent to biological systems, processes beyond those involved in sexual reproduction. Due to the distribution of different chromosomes between the sexes, there is a fundamental difference in the genetic makeup that extends to differences in signal transduction and regulation. Furthermore, this leads to overall changes in biological processes, including those of disease progression and aging. Here, we explore questions about differential gene expression in the aging heart of mice, including both males and females. We utilize a public data set of RNA-seq reads from a published project, in order to confirm the results of the authors, as well as become familiar with bioinformatics work flows. We found that our overall results concur with the author, there are clear differences in aging between males and females, and there is in fact an interaction between age and sex. Many of the top aging genes that differed between the sexes are involved in immune function and homeostasis. This study highlights the importance of experimental design, particularly in disease research where the goal is to find a treatment that works for as many people as possible. Understanding the fundamental biology of the model system is critical to correctly interpreting the results and

Introduction

Bioinformatics has revolutionized the progression of biological research, allowing us to compare entire transcriptomes of specific tissues across varying conditions. In terms of experimental design, it is critical to account for all of the biological variables that can affect gene expression, including those naturally present in model organisms.

One particular variable that is prevalent in biological models is sex, as the sex chromosomes differ between males (XY) and females (XX) in mammals. When considering any research questions that involve a living model, we must consider that sex could have an effect on our results due to the differences in the underlying genome.

In terms of aging and disease, it has been shown previously that there is a clear sex bias, where women typically have higher longevity than men and tend to develop different aging-related diseases (Tower et al 2017).

According to the National Center for Health Statistics, "Heart disease is the **leading cause of death** for men, women, and people of most racial and ethnic groups in the United States." (NCHS 2018-2021). In this project, we utilize a publicly available transcriptome data set to explore the effects of aging and sex on gene expression in the heart of mice(*Mus musculus*): NIH BioProject PRJNA835826 "Effects of age and sex on gene expression in the mouse heart (house mouse)". Later, we compare our analysis to the paper published by the Han research group(Han et al 2022). 12 RNA-seq samples were used for analysis, comprised of four groups that cover the variables of age (4 months and 20 months) and sex (M and F), which are summarized in the table below (Table 1).

Table 1: **Mouse Heart Sample Key**

ID	Sex	Age (months)	Replicate
SRR19123213	M	20	R3
SRR19123214	M	20	R2
SRR19123215	M	20	R1
SRR19123216	F	20	R3
SRR19123217	F	20	R2
SRR19123218	F	20	R1
SRR19123219	M	4	R3
SRR19123220	M	4	R2
SRR19123221	M	4	R1
SRR19123222	F	4	R3
SRR19123223	F	4	R2
SRR19123224	F	4	R1

Results

Task 1. Download Data & Transcriptome Assembly

The initial steps of this project were accomplished via the command line using Linux. First, the SRA toolkit was used to download the fastq sequence data from the experiment from NCBI GEO. STAR was then used to build an index from the mouse genome and annotation files to generate the BAM and junction files. HTseq was then used to obtain the raw read counts for each annotated gene. We have included the code in the ’**Methods’** section.

Task 2. PCA Plot on Raw Read Counts

Raw read count files generated from HTseq were read into R Studio, and a count matrix for each gene in the sequencing data was generated. A principal component analysis (PCA) plot summarizing the similarity between the samples was created.

```
# Install DESeq2 (if not already installed)
suppressWarnings({ suppressMessages({
if (!requireNamespace("DESeq2", quietly = TRUE)) {
  if (!requireNamespace("BiocManager", quietly = TRUE)) {
    install.packages("BiocManager")
  }
  BiocManager::install("DESeq2")
}
})})

library(readr)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

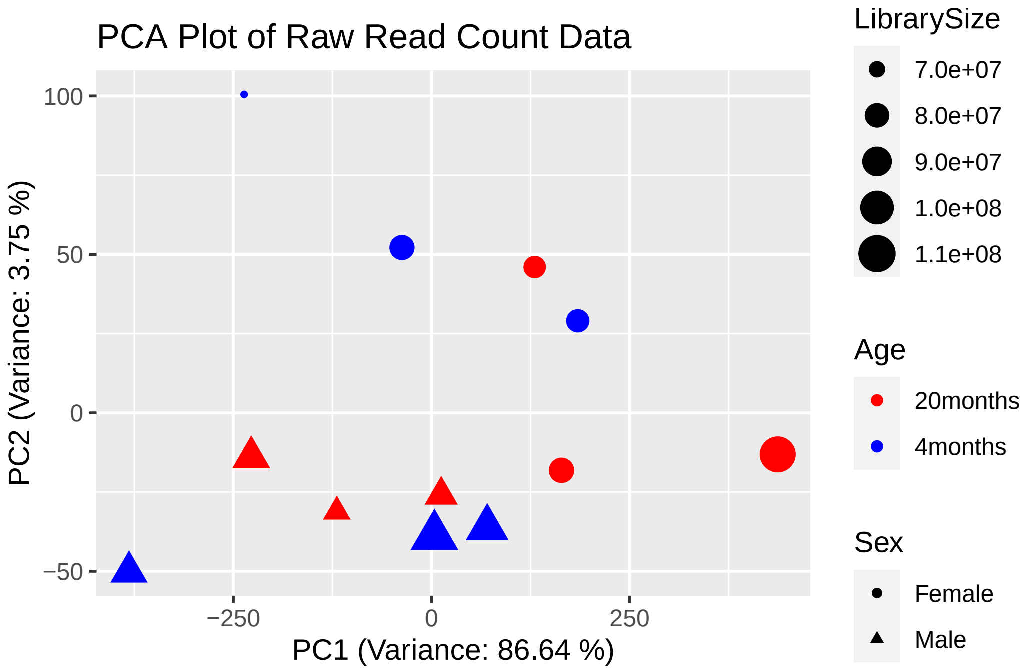
```
sample <- c()
size <- c()
# Create a list of filenames for your raw read count files
file_names <- c("raw_read_counts_SRR19123213.txt",
"raw_read_counts_SRR19123214.txt",
"raw_read_counts_SRR19123215.txt",
"raw_read_counts_SRR19123216.txt",
"raw_read_counts_SRR19123217.txt",
"raw_read_counts_SRR19123218.txt",
"raw_read_counts_SRR19123219.txt",
"raw_read_counts_SRR19123220.txt",
"raw_read_counts_SRR19123221.txt",
"raw_read_counts_SRR19123222.txt",
"raw_read_counts_SRR19123223.txt",
"raw_read_counts_SRR19123224.txt")
2
```

```
[1] 2
```

```
# Initialize an empty data frame to store read counts
raw_count_data <- data.frame()

# Loop through the files and read the data, excluding the last 5 rows
for (file in file_names) {
  data <- read.table(file, sep = "\t", col.names = c("Gene", file), stringsAsFactors = FAL
  data <- head(data, -5) # Exclude the last 5 rows

# Rename the file column to have a consistent name for joining
  col_name <- gsub(".txt","",gsub("raw_read_counts_", "", file))
  sample <- c(sample,col_name)
  colnames(data)[2] <- col_name
  size <- c(size, sum(data[, 2]))
  if (nrow(raw_count_data) == 0) {
    # If raw_count_data is empty, assign the data to it
    raw_count_data <- data
  } else {
    # Otherwise, left join the data with raw_count_data
    raw_count_data <- left_join(raw_count_data, data, by = "Gene")
  }
}
# Remove the "Gene" column as it's not needed for PCA
raw_count_data0 <- raw_count_data
raw_count_data <- dplyr::select(raw_count_data, -Gene)
# Standardize the data if necessary (centering and scaling)
raw_count_data <- scale(raw_count_data)
raw_count_data <- t(raw_count_data)
```

The following PCA plots display three covariates (sex, age, and library size) for raw and normalized read counts. The size of each point represents the library size, color indicates the age (blue = 4 months and red = 20 months), and sex is indicated by shape (circle = female and triangle = male).

```
suppressWarnings({ suppressMessages({
# Perform PCA
pca_result <- prcomp(raw_count_data, scale = TRUE)

# Create a data frame with PCA results and sample information
pca_data <- data.frame(PC1 = pca_result$x[, 1], PC2 = pca_result$x[, 2])

# Sample information (replace with your actual data)
sample_info <- data.frame(
  Sample = sample,  # Sample names
  Age = c( "20months", "20months", "20months", "20months", "20months", "20months",
           "4months", "4months", "4months", "4months", "4months", "4months"),  # Age infor
  Sex = c("Male", "Male", "Male", "Female", "Female", "Female",
          "Male", "Male", "Male", "Female", "Female", "Female"),  # Sex information
  LibrarySize = size  # Library size information
)

# Combine PCA results with sample information
pca_data <- cbind(pca_data, sample_info)

# Calculate the variance explained by each principal component
variance <- round(100 * pca_result$sdev^2 / sum(pca_result$sdev^2), 2)

# Create a PCA plot
pca_plot <- ggplot(pca_data, aes(x = PC1, y = PC2, color = Age, shape = Sex, size = Librar
  geom_point() +
  scale_color_manual(values = c("4months" = "blue", "20months" = "red")) +
  scale_shape_manual(values = c("Female" = 16, "Male" = 17)) +
  labs(title = "PCA Plot of Raw Read Count Data",
       x = paste("PC1 (Variance:", variance[1], "%)"),
       y = paste("PC2 (Variance:", variance[2], "%)"))

# Display the PCA plot
print(pca_plot)
})})
```

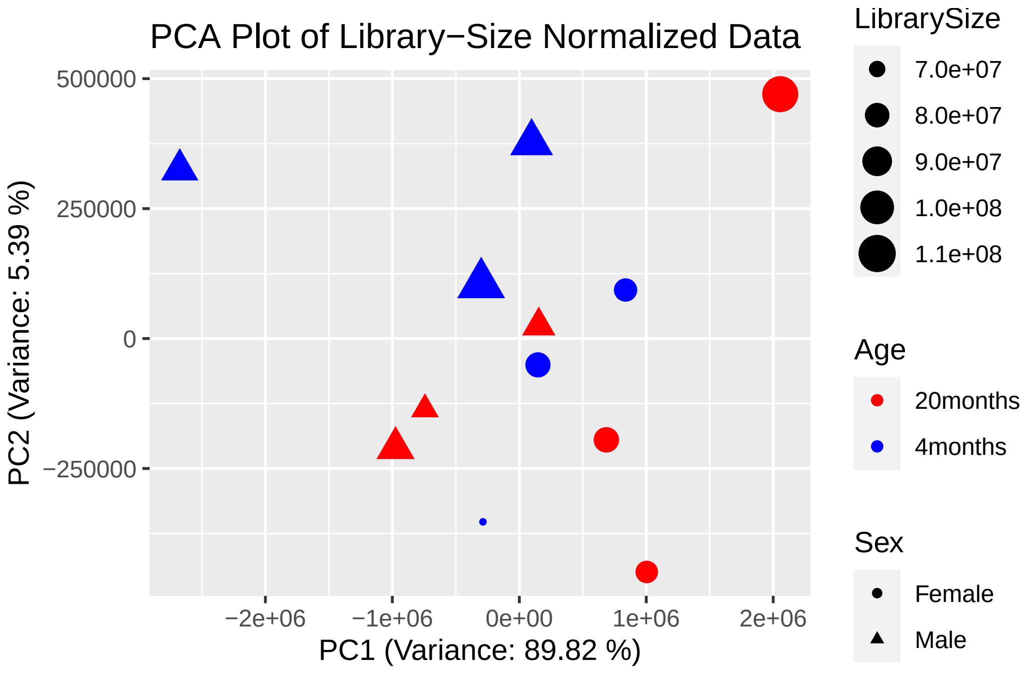


If we look along the PC1 axis, which accounts for 86.64% of the variance, we see that the main effect is due to sex. The majority of the male samples grouped together near the midline and negative quadrant and the female samples are more skewed to the left. Looking along the PC2 axis, which accounts for 3.75% of the variance, there is a slight effect of library size. Some of the larger libraries are grouped towards the negative axis and the more middle size libraries are near the positive axis. Overall, the male samples seem to be less variable than the female samples.

Task 3. PCA Plot on Normalized Counts

```
suppressWarnings({ suppressMessages({
# Load required libraries
library(DESeq2)
library(ggplot2)

# Create a DESeqDataSet
dds <- DESeqDataSetFromMatrix(countData = dplyr::select(raw_count_data0, -Gene), colData =
```

```r
# Normalize the data by library size
dds <- DESeq(dds)

# Extract the normalized count data
normalized_counts <- counts(dds, normalized = TRUE)

# Perform PCA on the normalized data
pca_result_normalized <- prcomp(t(normalized_counts), scale. = FALSE)

# Create a data frame with PCA results and sample information
pca_data_normalized <- data.frame(PC1 = pca_result_normalized$x[, 1], PC2 = pca_result_nor

# Combine PCA results with sample information
pca_data_normalized <- cbind(pca_data_normalized, sample_info)

# Calculate the variance explained by each principal component
variance_explained <- round(100 * pca_result_normalized$sdev^2 / sum(pca_result_normalized

# Create a PCA plot for the library-size normalized data with variance in axis labels
pca_plot_normalized <- ggplot(pca_data_normalized, aes(x = PC1, y = PC2, color = Age, shap
  geom_point() +
  scale_color_manual(values = c("4months" = "blue", "20months" = "red")) +
  scale_shape_manual(values = c("Female" = 16, "Male" = 17)) +
  labs(title = "PCA Plot of Library-Size Normalized Data",
       x = paste("PC1 (Variance:", variance_explained[1], "%)"),
       y = paste("PC2 (Variance:", variance_explained[2], "%)"))

# Display the PCA plot for normalized data
print(pca_plot_normalized)
})})
```



For the second PCA plot, samples were first normalized to library size. If we look along the PC1 axis, which accounts for 89.92% of the variance, we see a separation again in sex, with most of the male samples clustered together on the negative side and the female samples on the positive side. Along the PC2 axis (5.39% variance), we see a stronger effect of age than in the raw read count PCA plot. where the 20-month old samples are more clustered together and the 4-month old samples are closer together. There are, however two outliers in the top right quadrant and top left quadrant, particularly the female 20-month old sample which is farther away on the PC2 axis from the other samples in that same age/sex group.

The normalization of library size did seem to help group samples from one of the 4 groups closer together. Even when we ignore the outliers, there does seem to be a bit of overlap towards the middle of the plot. This may indicate some more complex interactions between age and sex that need to be investigated further.

Task 4. Differentially Expressed Genes

```
suppressWarnings({ suppressMessages({
# Assuming you have already loaded the DESeq2 library and created the DESeqDataSet object
library(DESeq2)
# Define the GLM model
design(dds) <- ~ Age + Sex + Age:Sex
```

```r
dds <- DESeq(dds)

# To extract results for each effect, you can use the following contrasts:

# 1. Aging effect
results_age0 <- results(dds, contrast=c("Age", "20months", "4months"))

# 2. Sex effect
results_sex0 <- results(dds, contrast=c("Sex", "Male", "Female"))

# 3. Age-Sex intersection
results_interaction0 <- results(dds, name = "Age4months.SexMale")
})})
```

```r
results_age0$Gene <- raw_count_data0$Gene
results_sex0$Gene <- raw_count_data0$Gene
results_interaction0$Gene <- raw_count_data0$Gene

results_age0
```

```
log2 fold change (MLE): Age 20months vs 4months
Wald test p-value: Age 20months vs 4months
DataFrame with 56941 rows and 7 columns
        baseMean log2FoldChange     lfcSE       stat    pvalue      padj
       <numeric>      <numeric> <numeric>  <numeric> <numeric> <numeric>
1     2480.01099     0.00821378 0.0914856  0.0897823 0.9284602  0.978919
2        0.24489     2.30183356 4.3224718  0.5325271 0.5943609        NA
3      261.51434    -0.03804055 0.1593583 -0.2387107 0.8113299  0.938462
4     1960.37691     0.45308644 0.5487639  0.8256492 0.4090031  0.731146
5       60.17451     0.51302724 0.2797859  1.8336419 0.0667072  0.320532
...          ...            ...       ...        ...       ...       ...
56937    0.00000             NA        NA         NA        NA        NA
56938    0.35343       0.705173   4.40729   0.160001   0.87288        NA
56939    0.00000             NA        NA         NA        NA        NA
56940    0.00000             NA        NA         NA        NA        NA
56941    0.00000             NA        NA         NA        NA        NA
                       Gene
                <character>
1      ENSMUSG00000000001.5
2     ENSMUSG00000000003.16
3     ENSMUSG00000000028.16
```

```
4     ENSMUSG00000000031.19
5     ENSMUSG00000000037.18
...                     ...
56937  ENSMUSG00002076988.1
56938  ENSMUSG00002076989.1
56939  ENSMUSG00002076990.1
56940  ENSMUSG00002076991.1
56941  ENSMUSG00002076992.1
```

```
results_sex0
```

```
log2 fold change (MLE): Sex Male vs Female
Wald test p-value: Sex Male vs Female
DataFrame with 56941 rows and 7 columns
        baseMean log2FoldChange     lfcSE      stat      pvalue      padj
       <numeric>      <numeric> <numeric> <numeric>   <numeric> <numeric>
1     2480.01099     -0.0336455 0.0914548 -0.367893 0.712953257 0.9933527
2        0.24489     -2.3538311 4.3224718 -0.544557 0.586058416        NA
3      261.51434     -0.0195886 0.1593915 -0.122896 0.902189138 0.9983566
4     1960.37691      0.2880053 0.5485793  0.525002 0.599581677 0.9814042
5       60.17451     -1.0688830 0.2897637 -3.688810 0.000225306 0.0258233
...          ...            ...       ...       ...         ...       ...
56937    0.00000             NA        NA        NA          NA        NA
56938    0.35343       -0.75717   4.40729 -0.171799    0.863595        NA
56939    0.00000             NA        NA        NA          NA        NA
56940    0.00000             NA        NA        NA          NA        NA
56941    0.00000             NA        NA        NA          NA        NA
                       Gene
                <character>
1      ENSMUSG00000000001.5
2     ENSMUSG00000000003.16
3     ENSMUSG00000000028.16
4     ENSMUSG00000000031.19
5     ENSMUSG00000000037.18
...                     ...
56937  ENSMUSG00002076988.1
56938  ENSMUSG00002076989.1
56939  ENSMUSG00002076990.1
56940  ENSMUSG00002076991.1
56941  ENSMUSG00002076992.1
```

```
results_interaction0
```

```
log2 fold change (MLE): Age4months.SexMale
Wald test p-value: Age4months.SexMale
DataFrame with 56941 rows and 7 columns
        baseMean log2FoldChange     lfcSE      stat      pvalue      padj
       <numeric>      <numeric> <numeric> <numeric>   <numeric> <numeric>
1     2480.01099     -0.0192955  0.129398 -0.149118 0.881460773 0.9739582
2        0.24489      2.0996859  6.173097  0.340135 0.733754925        NA
3      261.51434     -0.0539806  0.225514 -0.239367 0.810821328 0.9552149
4     1960.37691     -0.4750037  0.776044 -0.612084 0.540482429 0.8620503
5       60.17451      1.3469929  0.404531  3.329767 0.000869186 0.0682513
...          ...            ...       ...       ...         ...       ...
56937    0.00000             NA        NA        NA          NA        NA
56938    0.35343        3.14743   6.16982  0.510133    0.609958        NA
56939    0.00000             NA        NA        NA          NA        NA
56940    0.00000             NA        NA        NA          NA        NA
56941    0.00000             NA        NA        NA          NA        NA
                       Gene
                <character>
1      ENSMUSG00000000001.5
2     ENSMUSG00000000003.16
3     ENSMUSG00000000028.16
4     ENSMUSG00000000031.19
5     ENSMUSG00000000037.18
...                     ...
56937  ENSMUSG00002076988.1
56938  ENSMUSG00002076989.1
56939  ENSMUSG00002076990.1
56940  ENSMUSG00002076991.1
56941  ENSMUSG00002076992.1
```

Task 5. Top Differentially Expressed Genes for Each Effect

Differentially expressed genes (DEGs) were determined for each of the three effects:

- aging effect
- sex effect
- aging and sex interaction effect

DEGs were filtered at a specific FDR threshold of 0.05, in order to capture the most significant genes while minimizing the possibility of false positives, and report the total number of DEGs for each effect. After filtering, we report and visualize the top 5 DEGs for each effect according to padj values in addition to the top 5 DEGs according to log fold change.

```r
# Perform differential expression analysis for aging effect
results_age <- results(dds, contrast = c("Age", "20months", "4months"))

# Perform differential expression analysis for sex effect
results_sex <- results(dds, contrast = c("Sex", "Male", "Female"))

# Perform differential expression analysis for aging and sex interaction effect
results_interaction <- results(dds, name = "Age4months.SexMale")

# Filter DEGs at a specified FDR threshold (e.g., 0.05)
results_age <- results_age0[which(results_age$padj < 0.05), ]
results_sex <- results_sex0[which(results_sex$padj < 0.05), ]
results_interaction <- results_interaction0[which(results_interaction$padj < 0.05), ]
```

```r
# Total number of DEGs for each effect
total_DEGs_age <- nrow(results_age)
total_DEGs_sex <- nrow(results_sex)
total_DEGs_interaction <- nrow(results_interaction)

cat("Total DEGs for Aging Effect (FDR < 0.05):", total_DEGs_age, "\n")
```

```
Total DEGs for Aging Effect (FDR < 0.05): 1123
```

```r
cat("Total DEGs for Sex Effect (FDR < 0.05):", total_DEGs_sex, "\n")
```

```
Total DEGs for Sex Effect (FDR < 0.05): 226
```

```r
cat("Total DEGs for Interaction Effect (FDR < 0.05):", total_DEGs_interaction, "\n")
```

```
Total DEGs for Interaction Effect (FDR < 0.05): 169
```

```r
# Visualize the top five genes for each effect
top_genes_age <- head(results_age[order(results_age$padj), ], 5)
top_genes_sex <- head(results_sex[order(results_sex$padj), ], 5)
top_genes_interaction <- head(results_interaction[order(results_interaction$padj), ], 5)

# Print the top genes
cat("\nTop 5 DEGs for Aging Effect by Padj Value:\n")
```

```
Top 5 DEGs for Aging Effect by Padj Value:
```

```r
print(top_genes_age)
```

```
log2 fold change (MLE): Age 20months vs 4months
Wald test p-value: Age 20months vs 4months
DataFrame with 5 rows and 7 columns
   baseMean log2FoldChange     lfcSE      stat      pvalue        padj
  <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
1  4111.019        1.84080 0.1233708   14.9208 2.41217e-50 4.58410e-46
2  1087.596        1.65658 0.1164067   14.2309 5.88951e-46 5.59621e-42
3   810.023        2.55106 0.1825740   13.9727 2.28696e-44 1.44871e-40
4   885.180        2.58078 0.2135463   12.0853 1.26276e-33 5.99939e-30
5  1402.395        1.08500 0.0984462   11.0212 3.01848e-28 1.14726e-24
                   Gene
            <character>
1 ENSMUSG00000030862.14
2  ENSMUSG00000029330.9
3  ENSMUSG00000076612.9
4  ENSMUSG00000055489.9
5 ENSMUSG00000020333.18
```

```r
cat("\nTop 5 DEGs for Sex Effect by Padj Value:\n")
```

```
Top 5 DEGs for Sex Effect by Padj Value:
```

```r
print(top_genes_sex)
```

```
log2 fold change (MLE): Sex Male vs Female
Wald test p-value: Sex Male vs Female
DataFrame with 5 rows and 7 columns
   baseMean log2FoldChange     lfcSE      stat       pvalue         padj
  <numeric>      <numeric> <numeric> <numeric>    <numeric>    <numeric>
1  3901.163      10.604692 0.3693884   28.7088 2.96397e-181 5.19762e-177
2 17037.623      -9.274049 0.5394103  -17.1929  2.99931e-66  2.62979e-62
3   978.931      10.264820 0.6540370   15.6946  1.64791e-55  9.63257e-52
4   889.610      10.203241 0.6525122   15.6369  4.08376e-55  1.79032e-51
5 20213.566      -0.648096 0.0589987  -10.9849  4.51610e-28  1.58389e-24
                  Gene
           <character>
1 ENSMUSG00000069045.12
2  ENSMUSG00000086503.5
3 ENSMUSG00000068457.15
4 ENSMUSG00000056673.15
5 ENSMUSG00000000787.13
```

```
cat("\nTop 5 DEGs for Interaction Effect by Padj Value:\n")
```

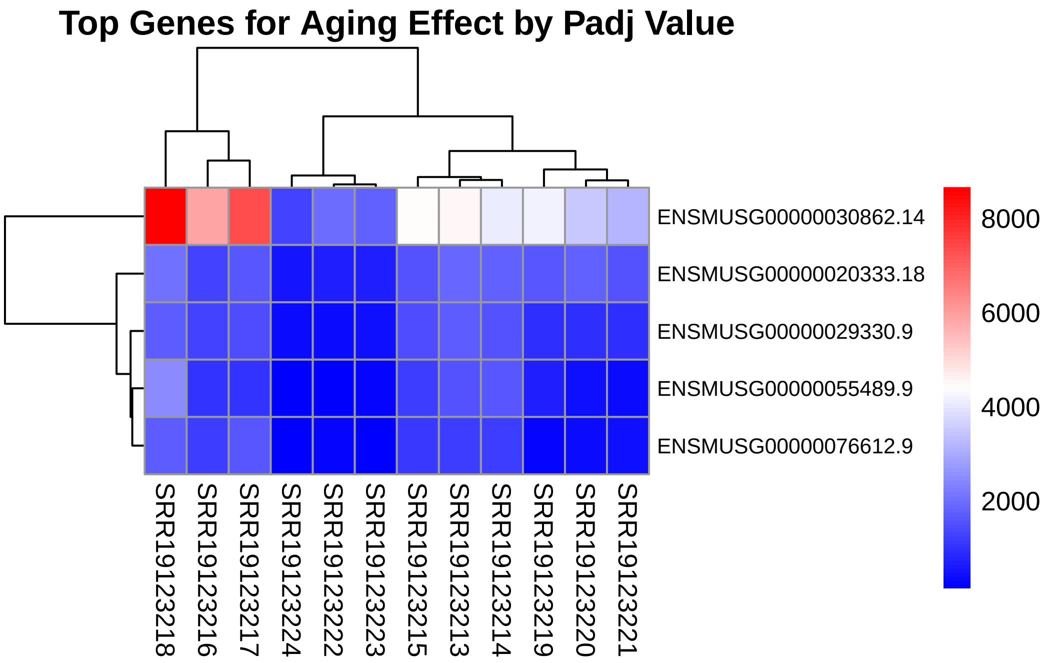
```
Top 5 DEGs for Interaction Effect by Padj Value:
```

```
print(top_genes_interaction)
```

```
log2 fold change (MLE): Age4months.SexMale
Wald test p-value: Age4months.SexMale
DataFrame with 5 rows and 7 columns
   baseMean log2FoldChange     lfcSE      stat      pvalue        padj
  <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
1   6379.92       1.761479  0.199260   8.84008 9.56468e-19 1.60715e-14
2   4111.02       1.365050  0.174020   7.84423 4.35627e-15 3.65992e-11
3   1776.96      -1.067360  0.173955  -6.13585 8.47029e-10 3.55816e-06
4   8675.64      -0.872903  0.141782  -6.15667 7.42887e-10 3.55816e-06
5   1402.40       0.822589  0.136732   6.01605 1.78722e-09 5.90241e-06
                  Gene
           <character>
1 ENSMUSG00000022797.17
2 ENSMUSG00000030862.14
3 ENSMUSG00000030889.15
```

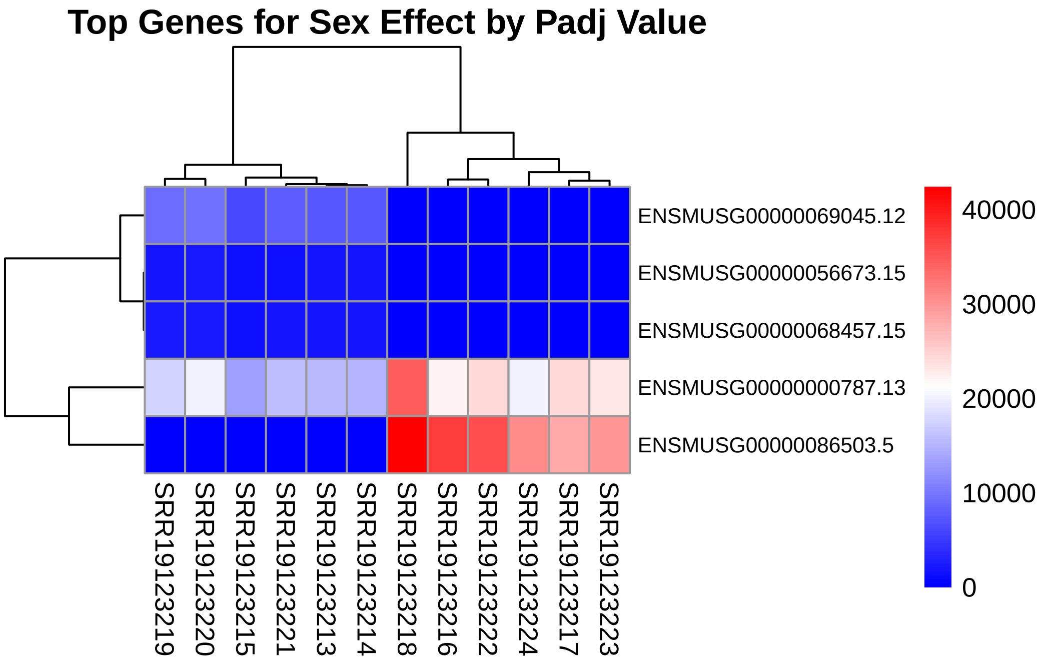
```
4 ENSMUSG00000040283.15
5 ENSMUSG00000020333.18
```

```r
# Load required libraries
library(pheatmap)

# Define the top gene names for each effect
top_gene_names_age <- top_genes_age$Gene
top_gene_names_sex <- top_genes_sex$Gene
top_gene_names_interaction <- top_genes_interaction$Gene

# Filter the raw count data for the top genes
top_gene_counts_age <- raw_count_data0[raw_count_data0$Gene %in% top_gene_names_age, ]
top_gene_counts_sex <- raw_count_data0[raw_count_data0$Gene %in% top_gene_names_sex, ]
top_gene_counts_interaction <- raw_count_data0[raw_count_data0$Gene %in% top_gene_names_in

plot_heatmap <- function(top_gene_counts, effect_name) {
  gene_names <- top_gene_counts$Gene
  top_gene_counts <- top_gene_counts[, -1]  # Exclude the 'Gene' column

  pheatmap(as.matrix(top_gene_counts),
           main = paste("Top Genes for", effect_name),
           color = colorRampPalette(c("blue", "white", "red"))(100),
           fontsize_row = 8,
           labels_row = gene_names)
}

# Plot the heatmaps for each effect
plot_heatmap(top_gene_counts_age, "Aging Effect by Padj Value")
```



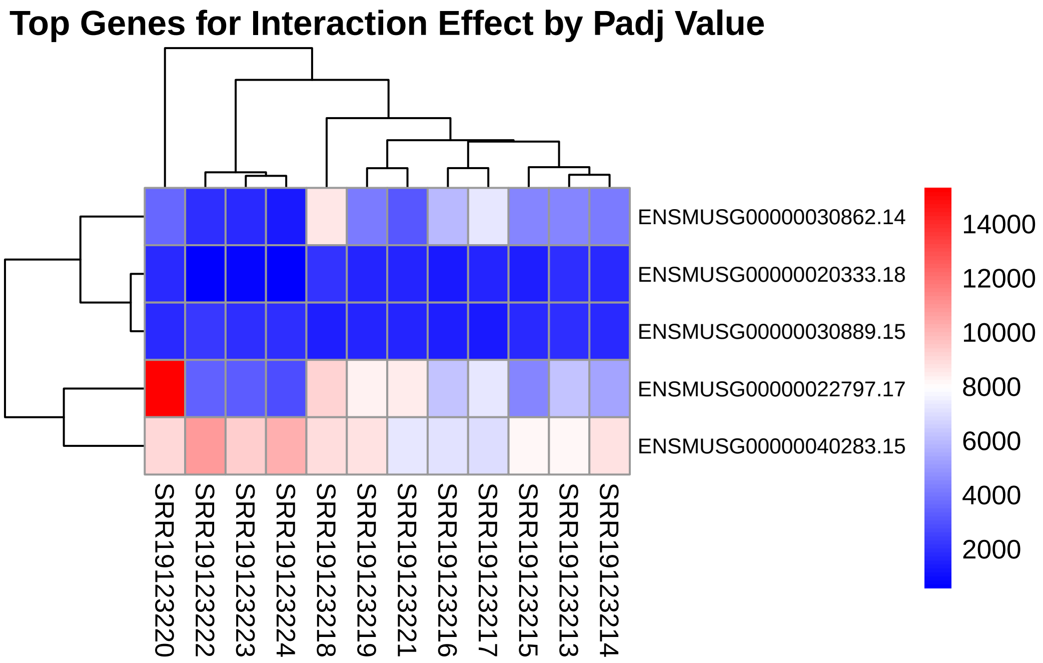
Top Genes for Aging Effect by Padj Value

```
plot_heatmap(top_gene_counts_sex, "Sex Effect by Padj Value")
```



Top Genes for Sex Effect by Padj Value

```
plot_heatmap(top_gene_counts_interaction, "Interaction Effect by Padj Value")
```



```
# Visualize the top five genes for each effect
top_genes_age_by_lfc <- tail(results_age[order(abs(results_age$log2FoldChange)), ], 5)
top_genes_sex_by_lfc <- tail(results_sex[order(abs(results_sex$log2FoldChange)), ], 5)
top_genes_interaction_by_lfc <- tail(results_interaction[order(abs(results_interaction$log

# Print the top genes
cat("\nTop 5 DEGs for Aging Effect by LogFoldChange:\n")
```

```
Top 5 DEGs for Aging Effect by LogFoldChange:
```

```
print(top_genes_age_by_lfc)
```

```
log2 fold change (MLE): Age 20months vs 4months
Wald test p-value: Age 20months vs 4months
DataFrame with 5 rows and 7 columns
         baseMean log2FoldChange     lfcSE      stat      pvalue        padj
```

```
        <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
[1119,]   19.0642        4.12404  1.085598   3.79887 1.45358e-04 6.16604e-03
[1120,]   21.8223        4.19087  0.818920   5.11756 3.09512e-07 4.20141e-05
[1121,]   45.0573        4.32299  0.626819   6.89671 5.32206e-12 2.46684e-09
[1122,]   11.4374        5.33078  1.144668   4.65705 3.20773e-06 3.17499e-04
[1123,]   59.4314        5.59113  0.753113   7.42402 1.13614e-13 6.98023e-11
                         Gene
                  <character>
[1119,]  ENSMUSG00000030046.7
[1120,]  ENSMUSG00000035186.7
[1121,]  ENSMUSG00000031495.9
[1122,]  ENSMUSG00000079190.4
[1123,] ENSMUSG00000045967.12
```

```
cat("\nTop 5 DEGs for Sex Effect by LogFoldChange:\n")
```

```

Top 5 DEGs for Sex Effect by LogFoldChange:
```

```
print(top_genes_sex_by_lfc)
```

```
log2 fold change (MLE): Sex Male vs Female
Wald test p-value: Sex Male vs Female
DataFrame with 5 rows and 7 columns
        baseMean log2FoldChange     lfcSE      stat       pvalue         padj
       <numeric>      <numeric> <numeric> <numeric>    <numeric>    <numeric>
[222,] 17037.623       -9.27405  0.539410  -17.1929  2.99931e-66  2.62979e-62
[223,]  2770.231        9.60106  0.903105   10.6312  2.13416e-26  5.34638e-23
[224,]   889.610       10.20324  0.652512   15.6369  4.08376e-55  1.79032e-51
[225,]   978.931       10.26482  0.654037   15.6946  1.64791e-55  9.63257e-52
[226,]  3901.163       10.60469  0.369388   28.7088 2.96397e-181 5.19762e-177
                        Gene
                 <character>
[222,]  ENSMUSG00000086503.5
[223,] ENSMUSG00000069049.12
[224,] ENSMUSG00000056673.15
[225,] ENSMUSG00000068457.15
[226,] ENSMUSG00000069045.12
```

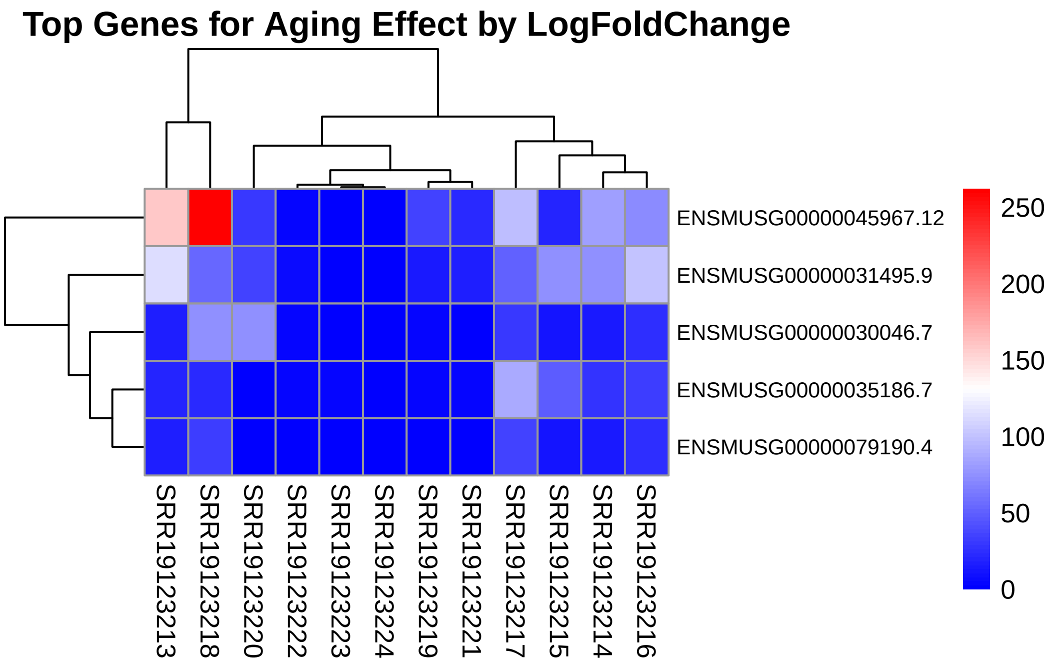
```r
cat("\nTop 5 DEGs for Interaction Effect by LogFoldChange:\n")
```

```
Top 5 DEGs for Interaction Effect by LogFoldChange:
```

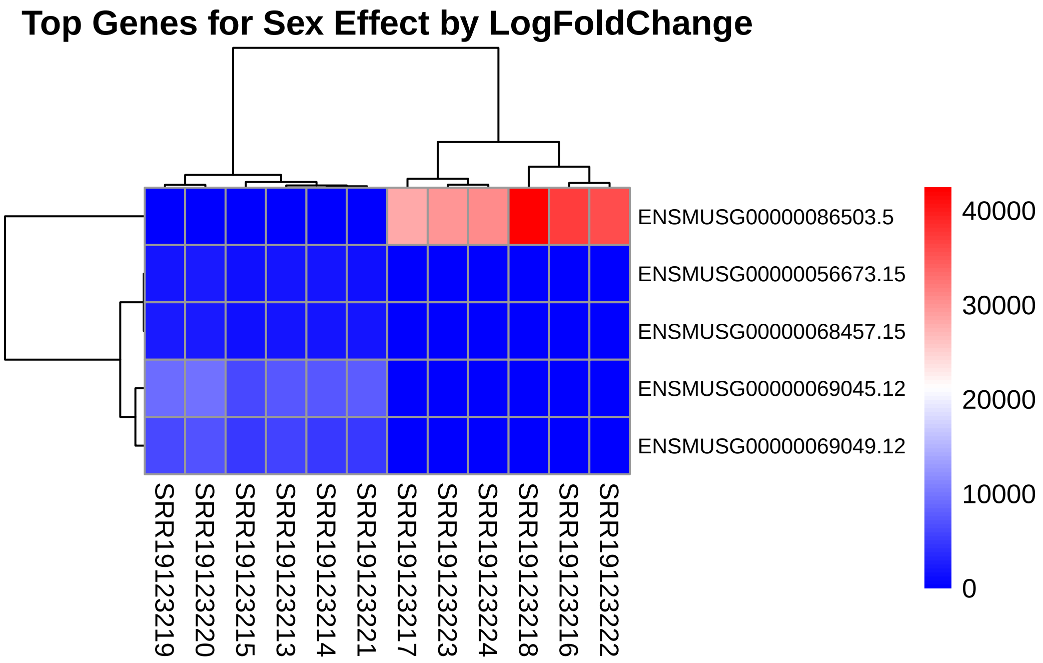
```r
print(top_genes_interaction_by_lfc)
```

```
log2 fold change (MLE): Age4months.SexMale
Wald test p-value: Age4months.SexMale
DataFrame with 5 rows and 7 columns
        baseMean log2FoldChange     lfcSE      stat      pvalue       padj
       <numeric>      <numeric> <numeric> <numeric>   <numeric>  <numeric>
[165,]   27.6958       -2.33410  0.603821  -3.86556 1.10836e-04 0.02165547
[166,]   24.4364       -2.38619  0.671209  -3.55507 3.77882e-04 0.04338215
[167,]   39.1245        2.77873  0.797271   3.48530 4.91590e-04 0.04946224
[168,]   55.2375        2.94676  0.707904   4.16266 3.14565e-05 0.00980887
[169,]   59.4314        3.84820  0.928571   4.14422 3.40979e-05 0.00980887
                        Gene
                 <character>
[165,]  ENSMUSG00000050359.8
[166,]  ENSMUSG00000114771.2
[167,] ENSMUSG00000055333.15
[168,] ENSMUSG00000039913.13
[169,] ENSMUSG00000045967.12
```

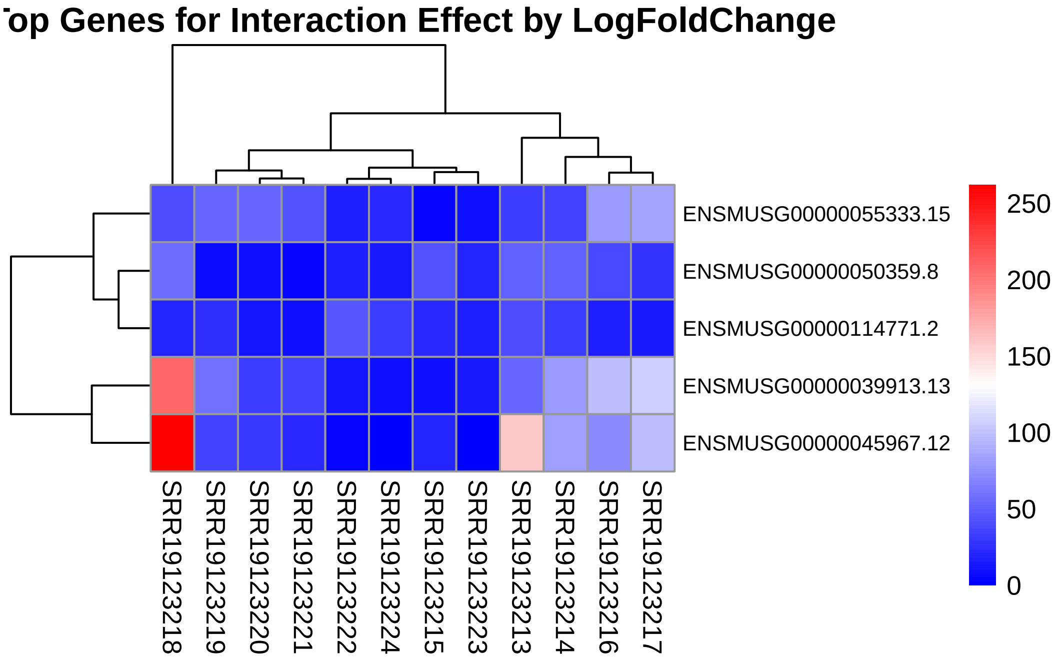
```r
# Load required libraries
library(pheatmap)

# Define the top gene names for each effect
top_gene_names_age_by_lfc <- top_genes_age_by_lfc$Gene
top_gene_names_sex_by_lfc <- top_genes_sex_by_lfc$Gene
top_gene_names_interaction_by_lfc <- top_genes_interaction_by_lfc$Gene

# Filter the raw count data for the top genes
top_gene_counts_age_by_lfc <- raw_count_data0[raw_count_data0$Gene %in% top_gene_names_age
top_gene_counts_sex_by_lfc <- raw_count_data0[raw_count_data0$Gene %in% top_gene_names_sex
top_gene_counts_interaction_by_lfc <- raw_count_data0[raw_count_data0$Gene %in% top_gene_n

plot_heatmap <- function(top_gene_counts_by_lfc, effect_name) {
  gene_names <- top_gene_counts_by_lfc$Gene
  top_gene_counts_by_lfc <- top_gene_counts_by_lfc[, -1]  # Exclude the 'Gene' column

  pheatmap(as.matrix(top_gene_counts_by_lfc),
           main = paste("Top Genes for", effect_name),
           color = colorRampPalette(c("blue", "white", "red"))(100),
           fontsize_row = 8,
           labels_row = gene_names)
}

# Plot the heatmaps for each effect
plot_heatmap(top_gene_counts_age_by_lfc, "Aging Effect by LogFoldChange")
```



```
plot_heatmap(top_gene_counts_sex_by_lfc, "Sex Effect by LogFoldChange")
```



```
plot_heatmap(top_gene_counts_interaction_by_lfc, "Interaction Effect by LogFoldChange")
```



Top DEGs for Each Effect

Here is a review of the results from the top DGEs and we would like to note that we go into more detail about the biological implications in the Extra Credit Task E1.

AGE

According to the dendrogram plotted along with the heatmap according to padj value, the male samples are clustered together, including a branch for the younger female samples. The older female sample shows very high expression of Cpxm2 (ENSMUSG00000030862.14) carboxypeptidase X, which is involved with peptide metabolic processes. Since we do not see the same level of expression in the older male samples, it seems there may be evidence of underlying sex-age interaction. Overall, many of the genes are involved in signaling, structure, and metabolic processes, which we would expect to be affected by age.

SEX

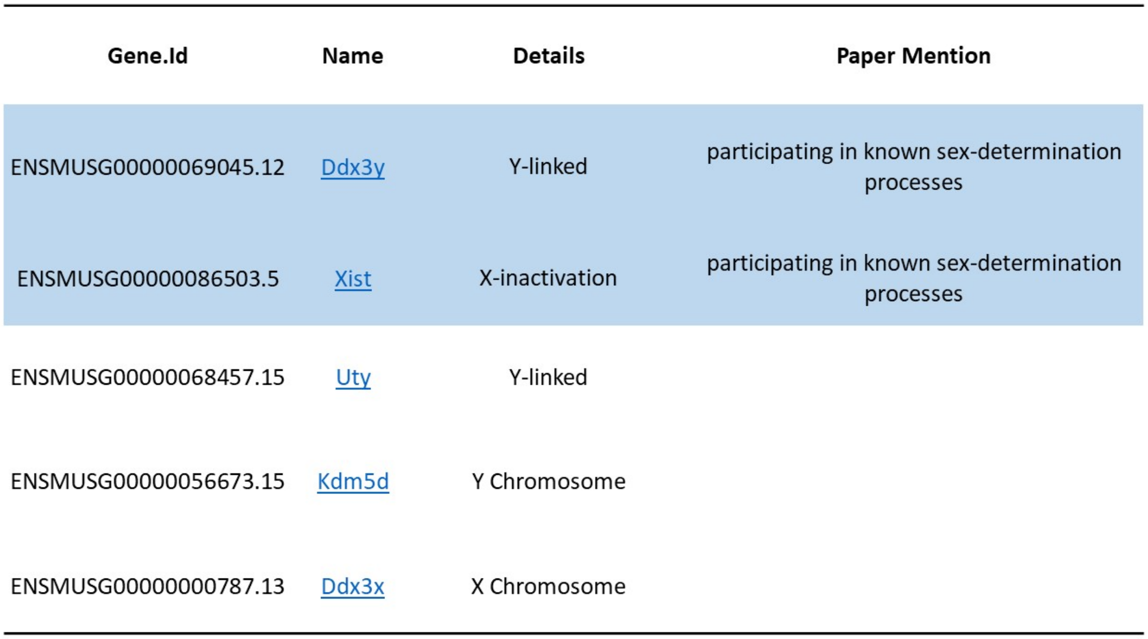
The heatmap clusters clearly into two groups, male and female, and as expected. All of these genes are found to be linked to either the X chromosome (Xist, Ddx3x) or Y chromosome (Ddx3y, Ulty, Kdm5d) according to the gene database and are expressed accordingly to sex in the data. In addition to 'participat[ing] in known sex-determination' according to the paper, many of these genes also function in the processes of transcription and chromatin remodeling (Han et al 2022). This indicates an underlying difference in transcription regulation between the sexes.

SEX-AGE INTERACTION

Similar to what was found by the original authors, we found several cardiac genes in our top DEGs that seem to demonstrate a sex-age interaction, where *Cpxm2* and *Tfrc* have a negative sex-age interaction (Han et al 2022). Some of these genes appear to also function in homestasis and immune processes, indicating an underlying difference in aging between the sexes.

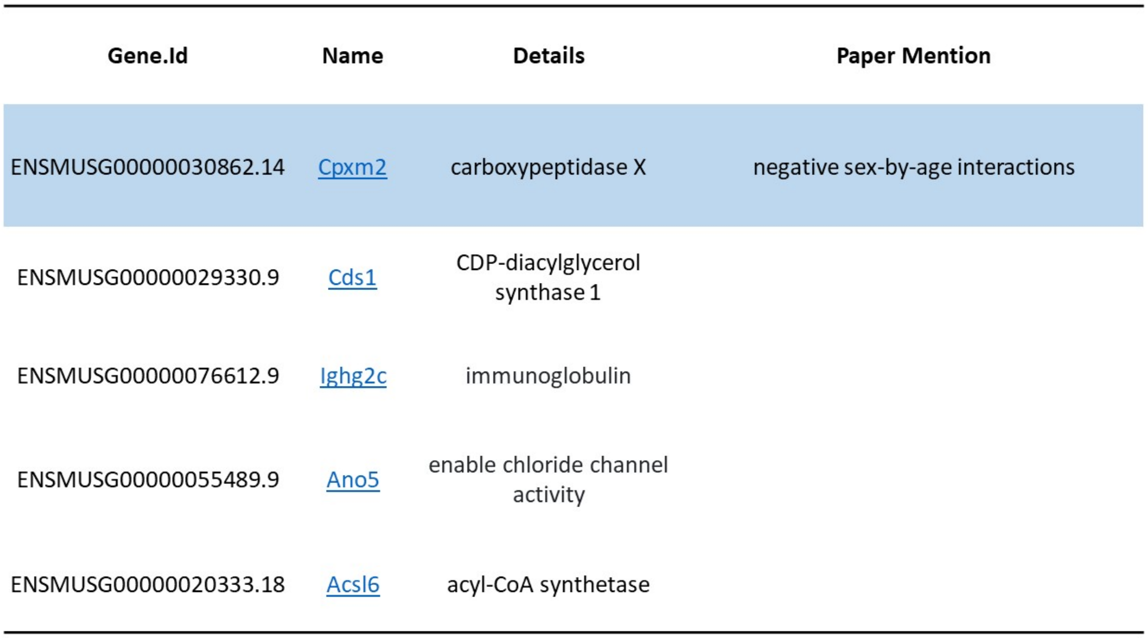
Extra Credit Task E1 - Interpretation of top genes

The tables below summarize our top five DEGs for each effect according to either the padj value or the log-fold change value. Common genes between our results and those discovered by the authors of the original project are highlighted in blue (Han et al 2022).



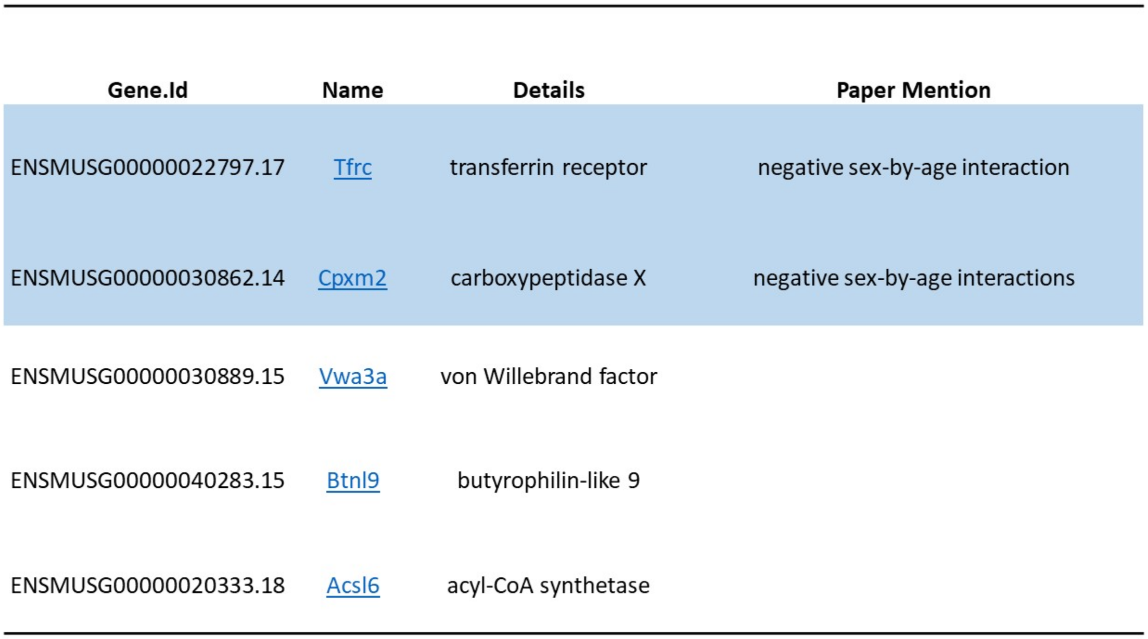
Gene.Id	Name	Details	Paper Mention
ENSMUSG00000069045.12	Ddx3y	Y-linked	participating in known sex-determination processes
ENSMUSG00000086503.5	Xist	X-inactivation	participating in known sex-determination processes
ENSMUSG00000068457.15	Uty	Y-linked	
ENSMUSG00000056673.15	Kdm5d	Y Chromosome	
ENSMUSG00000000787.13	Ddx3x	X Chromosome	

Figure 1: **Top 5 DEGs for Sex Effect by Padj Value**



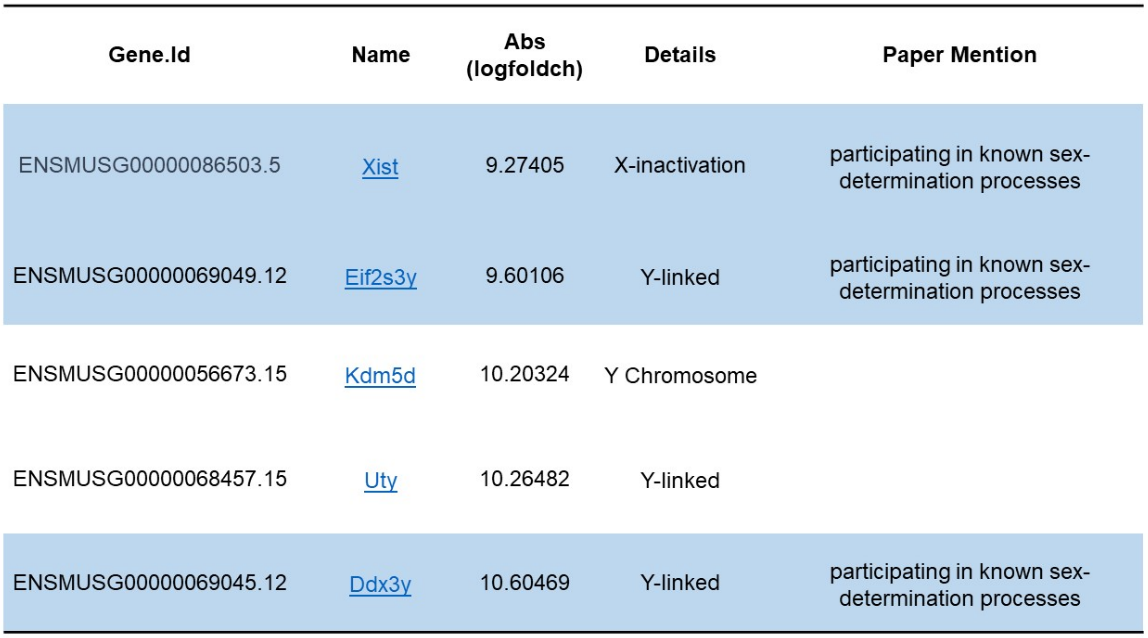
Gene.Id	Name	Details	Paper Mention
ENSMUSG00000030862.14	Cpxm2	carboxypeptidase X	negative sex-by-age interactions
ENSMUSG00000029330.9	Cds1	CDP-diacylglycerol synthase 1	
ENSMUSG00000076612.9	Ighg2c	immunoglobulin	
ENSMUSG00000055489.9	Ano5	enable chloride channel activity	
ENSMUSG00000020333.18	Acsl6	acyl-CoA synthetase	

Figure 2: **Top 5 DEGs for Aging Effect by Padj Value**



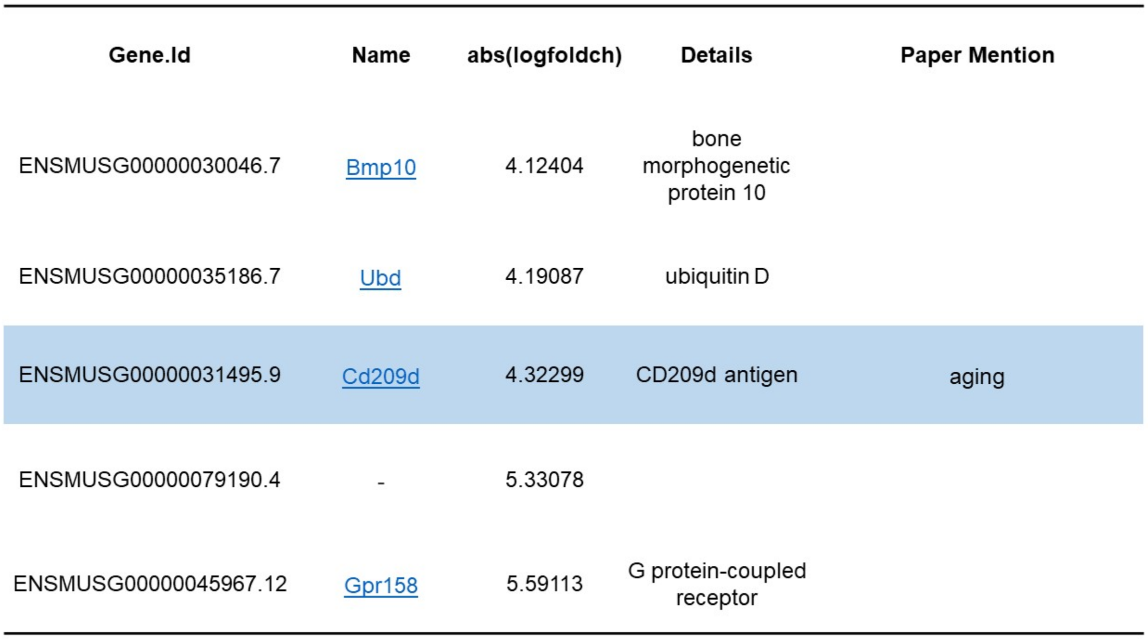
Gene.Id	Name	Details	Paper Mention
ENSMUSG00000022797.17	Tfrc	transferrin receptor	negative sex-by-age interaction
ENSMUSG00000030862.14	Cpxm2	carboxypeptidase X	negative sex-by-age interactions
ENSMUSG00000030889.15	Vwa3a	von Willebrand factor	
ENSMUSG00000040283.15	Btnl9	butyrophilin-like 9	
ENSMUSG00000020333.18	Acsl6	acyl-CoA synthetase	

Figure 3: **Top 5 DEGs for Age-Sex Interaction Effect by Padj Value**



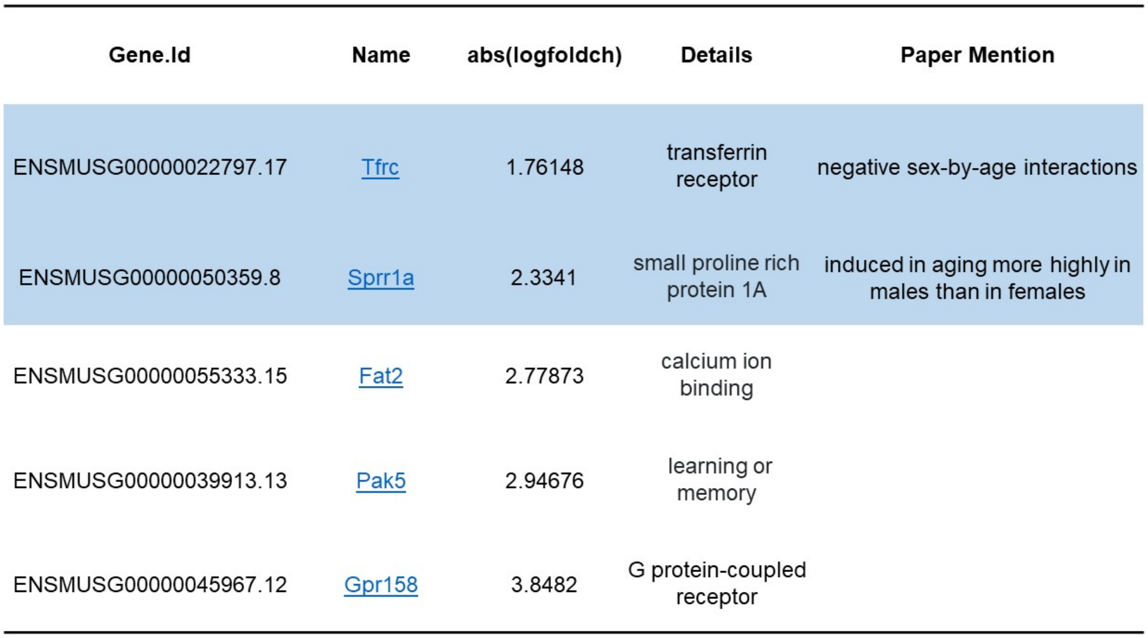
Gene.Id	Name	Abs (logfoldch)	Details	Paper Mention
ENSMUSG00000086503.5	Xist	9.27405	X-inactivation	participating in known sex-determination processes
ENSMUSG00000069049.12	Eif2s3y	9.60106	Y-linked	participating in known sex-determination processes
ENSMUSG00000056673.15	Kdm5d	10.20324	Y Chromosome	
ENSMUSG00000068457.15	Uty	10.26482	Y-linked	
ENSMUSG00000069045.12	Ddx3y	10.60469	Y-linked	participating in known sex-determination processes

Figure 4: **Top 5 DEGs for Sex Effect by Log-fold Change Value**



Gene.Id	Name	abs(logfoldch)	Details	Paper Mention
ENSMUSG00000030046.7	Bmp10	4.12404	bone morphogenetic protein 10	
ENSMUSG00000035186.7	Ubd	4.19087	ubiquitin D	
ENSMUSG00000031495.9	Cd209d	4.32299	CD209d antigen	aging
ENSMUSG00000079190.4	-	5.33078		
ENSMUSG00000045967.12	Gpr158	5.59113	G protein-coupled receptor	

Figure 5: **Top 5 DEGs for Aging Effect by Log-fold Change Value**



Gene.Id	Name	abs(logfoldch)	Details	Paper Mention
ENSMUSG00000022797.17	Tfrc	1.76148	transferrin receptor	negative sex-by-age interactions
ENSMUSG00000050359.8	Sprr1a	2.3341	small proline rich protein 1A	induced in aging more highly in males than in females
ENSMUSG00000055333.15	Fat2	2.77873	calcium ion binding	
ENSMUSG00000039913.13	Pak5	2.94676	learning or memory	
ENSMUSG00000045967.12	Gpr158	3.8482	G protein-coupled receptor	

Figure 6: **Top 5 DEGs for Sex-Age Interaction Effect by Log-fold Change Value**

E1.1 Comparison of Top DEGs to Paper

When we compare out top DEGs for each effect to the original project paper, we find that our results are consistent with the findings of the author. For each of the three effects, whether by padj or log-fold change, we had at least 1 top gene in common that was mentioned specifically in the paper.

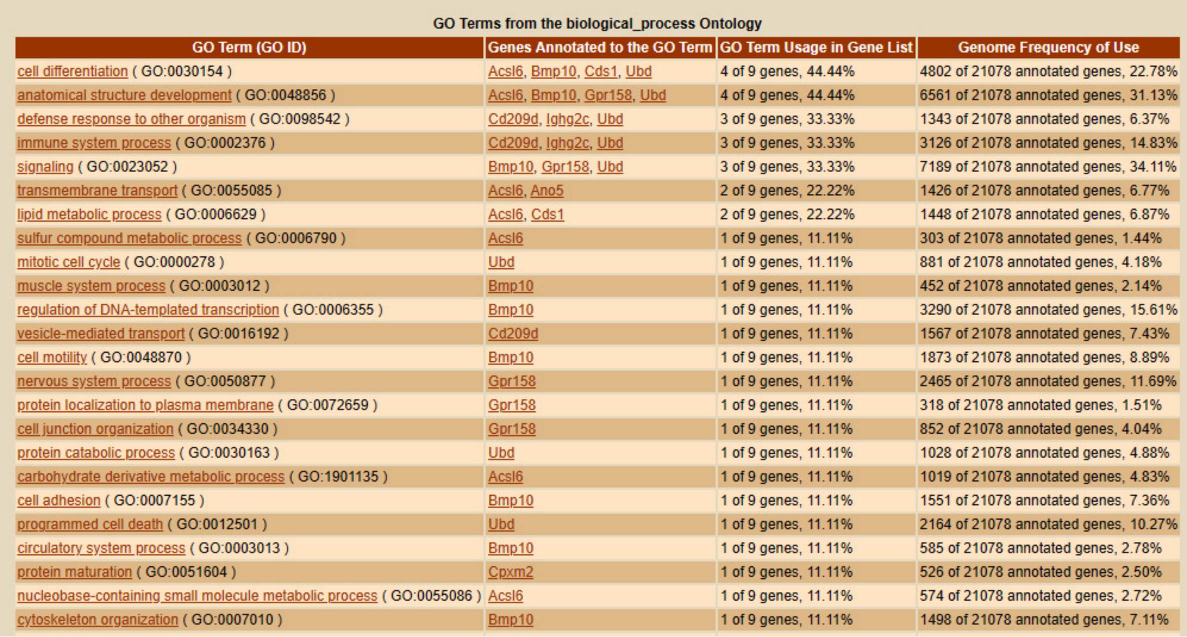
E1.2 Biological Interpretation

In order to determine a wider overview of the biological processes that our top DEGs are involved in, we determine the Gene Onotology (GO) terms using the Generic Gene Ontology Term Mapper (Lewis-Sigler Institute for Integrative Genomics, Princeton University). The top 5 DEGs sorted by padj value and log-fold change value were combined in order to capture more terms.

AGE

When we combine the top 5 DEGs by Age Effect by padj and logfold change values, the gene ontology (GO) terms include cell differentiation, anatomical structure development, defense response, immune response, signaling, transport, and metabolic processes. This is consistent with the paper, where they say “that sex-differential genes appear to primarily cluster around metabolic pathways” (Han et al 2022).

It is also important to consider that many genes function in multiple pathways, for example, Bmp10 (bone morphogenetic protein-10) sounds out of place in the heart. In fact, it has been found to be critical for cardiomyocyte proliferation (Sun et al 2014), a process that is necessary for cardiac repair after stress or injury and likely becomes less robust as age increases.



GO Terms from the biological_process Ontology

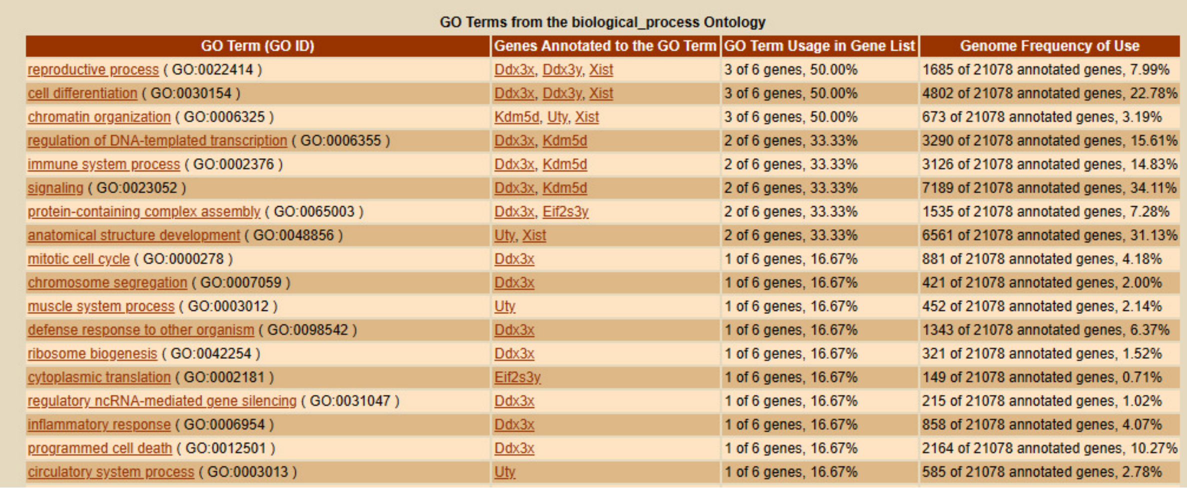
GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
cell differentiation (GO:0030154)	Acsl6, Bmp10, Cds1, Ubd	4 of 9 genes, 44.44%	4802 of 21078 annotated genes, 22.78%
anatomical structure development (GO:0048856)	Acsl6, Bmp10, Gpr158, Ubd	4 of 9 genes, 44.44%	6561 of 21078 annotated genes, 31.13%
defense response to other organism (GO:0098542)	Cd209d, Ighg2c, Ubd	3 of 9 genes, 33.33%	1343 of 21078 annotated genes, 6.37%
immune system process (GO:0002376)	Cd209d, Ighg2c, Ubd	3 of 9 genes, 33.33%	3126 of 21078 annotated genes, 14.83%
signaling (GO:0023052)	Bmp10, Gpr158, Ubd	3 of 9 genes, 33.33%	7189 of 21078 annotated genes, 34.11%
transmembrane transport (GO:0055085)	Acsl6, Ano5	2 of 9 genes, 22.22%	1426 of 21078 annotated genes, 6.77%
lipid metabolic process (GO:0006629)	Acsl6, Cds1	2 of 9 genes, 22.22%	1448 of 21078 annotated genes, 6.87%
sulfur compound metabolic process (GO:0006790)	Acsl6	1 of 9 genes, 11.11%	303 of 21078 annotated genes, 1.44%
mitotic cell cycle (GO:0000278)	Ubd	1 of 9 genes, 11.11%	881 of 21078 annotated genes, 4.18%
muscle system process (GO:0003012)	Bmp10	1 of 9 genes, 11.11%	452 of 21078 annotated genes, 2.14%
regulation of DNA-templated transcription (GO:0006355)	Bmp10	1 of 9 genes, 11.11%	3290 of 21078 annotated genes, 15.61%
vesicle-mediated transport (GO:0016192)	Cd209d	1 of 9 genes, 11.11%	1567 of 21078 annotated genes, 7.43%
cell motility (GO:0048870)	Bmp10	1 of 9 genes, 11.11%	1873 of 21078 annotated genes, 8.89%
nervous system process (GO:0050877)	Gpr158	1 of 9 genes, 11.11%	2465 of 21078 annotated genes, 11.69%
protein localization to plasma membrane (GO:0072659)	Gpr158	1 of 9 genes, 11.11%	318 of 21078 annotated genes, 1.51%
cell junction organization (GO:0034330)	Gpr158	1 of 9 genes, 11.11%	852 of 21078 annotated genes, 4.04%
protein catabolic process (GO:0030163)	Ubd	1 of 9 genes, 11.11%	1028 of 21078 annotated genes, 4.88%
carbohydrate derivative metabolic process (GO:1901135)	Acsl6	1 of 9 genes, 11.11%	1019 of 21078 annotated genes, 4.83%
cell adhesion (GO:0007155)	Bmp10	1 of 9 genes, 11.11%	1551 of 21078 annotated genes, 7.36%
programmed cell death (GO:0012501)	Ubd	1 of 9 genes, 11.11%	2164 of 21078 annotated genes, 10.27%
circulatory system process (GO:0003013)	Bmp10	1 of 9 genes, 11.11%	585 of 21078 annotated genes, 2.78%
protein maturation (GO:0051604)	Cpxm2	1 of 9 genes, 11.11%	526 of 21078 annotated genes, 2.50%
nucleobase-containing small molecule metabolic process (GO:0055086)	Acsl6	1 of 9 genes, 11.11%	574 of 21078 annotated genes, 2.72%
cytoskeleton organization (GO:0007010)	Bmp10	1 of 9 genes, 11.11%	1498 of 21078 annotated genes, 7.11%

Figure 7: GO Terms for Age Effect (Top Padj and Logfold Change

SEX

As expected, all of the DEGs determined for the effect of sex are linked to either the X or Y chromosome. The gene ontology (GO) terms include reproductive process, cell differentiation, chromatin organization, regulation of DNA-templeted transcription, immune processes, etc. These results indicate that the regulation of transcription differs between the sexes, which subsequently affects cell signaling pathways. There may be differences in the immune system reactions as well, which complicates the study of aging-related diseases.

In relation to heart disease, there are many sex-realted differences in the diagnosis and treatment of disease. One study found that “heart failure disproportionately contributes to coronary heart disease mortality in women, potentially due to undiagnosed ischaemic heart disease in women. The strength of the association with cardiovascular risk factors differ by sex.” (Snyder et al 2014). There are also differences in treatments for cardiac failure, where “evidence suggests that optimal survival in women occurs with lower doses of blockers, angiotensin receptor blockers, and angiotensin converting enzyme inhibitors than in men” (Santema et al 2019).



GO Terms from the biological_process Ontology

GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
reproductive process (GO:0022414)	Ddx3x, Ddx3y, Xist	3 of 6 genes, 50.00%	1685 of 21078 annotated genes, 7.99%
cell differentiation (GO:0030154)	Ddx3x, Ddx3y, Xist	3 of 6 genes, 50.00%	4802 of 21078 annotated genes, 22.78%
chromatin organization (GO:0006325)	Kdm5d, Uty, Xist	3 of 6 genes, 50.00%	673 of 21078 annotated genes, 3.19%
regulation of DNA-templated transcription (GO:0006355)	Ddx3x, Kdm5d	2 of 6 genes, 33.33%	3290 of 21078 annotated genes, 15.61%
immune system process (GO:0002376)	Ddx3x, Kdm5d	2 of 6 genes, 33.33%	3126 of 21078 annotated genes, 14.83%
signaling (GO:0023052)	Ddx3x, Kdm5d	2 of 6 genes, 33.33%	7189 of 21078 annotated genes, 34.11%
protein-containing complex assembly (GO:0065003)	Ddx3x, Eif2s3y	2 of 6 genes, 33.33%	1535 of 21078 annotated genes, 7.28%
anatomical structure development (GO:0048856)	Uty, Xist	2 of 6 genes, 33.33%	6561 of 21078 annotated genes, 31.13%
mitotic cell cycle (GO:0000278)	Ddx3x	1 of 6 genes, 16.67%	881 of 21078 annotated genes, 4.18%
chromosome segregation (GO:0007059)	Ddx3x	1 of 6 genes, 16.67%	421 of 21078 annotated genes, 2.00%
muscle system process (GO:0003012)	Uty	1 of 6 genes, 16.67%	452 of 21078 annotated genes, 2.14%
defense response to other organism (GO:0098542)	Ddx3x	1 of 6 genes, 16.67%	1343 of 21078 annotated genes, 6.37%
ribosome biogenesis (GO:0042254)	Ddx3x	1 of 6 genes, 16.67%	321 of 21078 annotated genes, 1.52%
cytoplasmic translation (GO:0002181)	Eif2s3y	1 of 6 genes, 16.67%	149 of 21078 annotated genes, 0.71%
regulatory ncRNA-mediated gene silencing (GO:0031047)	Ddx3x	1 of 6 genes, 16.67%	215 of 21078 annotated genes, 1.02%
inflammatory response (GO:0006954)	Ddx3x	1 of 6 genes, 16.67%	858 of 21078 annotated genes, 4.07%
programmed cell death (GO:0012501)	Ddx3x	1 of 6 genes, 16.67%	2164 of 21078 annotated genes, 10.27%
circulatory system process (GO:0003013)	Uty	1 of 6 genes, 16.67%	585 of 21078 annotated genes, 2.78%

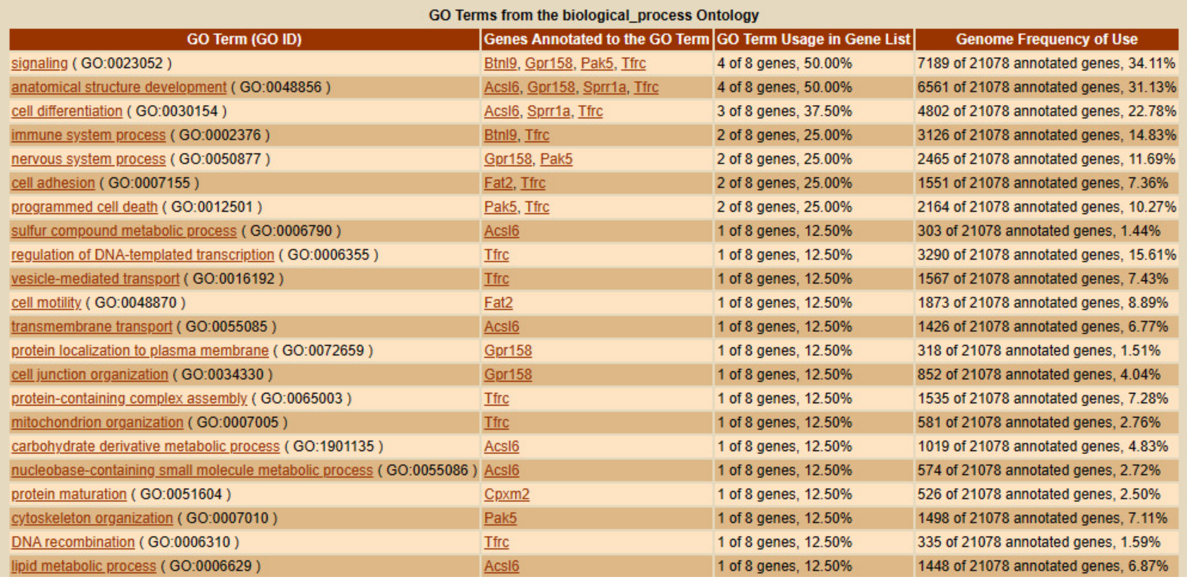
Figure 8: **Go Terms for Sex Effect (Top Padj and Logfold Changes)**

AGE-SEX INTERACTION

Our model explores not only the effects of age and sex on cardiac gene expression but also the possible interactions of age and sex. Due to the underlying differences between the transcriptomes of the two sexes, the expression of genes related to aging seems to be affected by sex.

The gene ontology (GO) terms include signaling, anatomical structure development, cell differentiation, immune system process, nervous system processes cell adhesion, programmed cell death, etc.

Interestingly, the cardiac gene *Tfrc* is critical to heart function by promoting iron uptake. In a recent paper, the researchers found that this gene also participates in immune processes by promoting macrophage infiltration (Pan et al 2023). This example demonstrates the hidden complexity of gene function and interactions, which may not be apparent without more research into specific genes.



GO Terms from the biological_process Ontology

GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
signaling (GO:0023052)	Btnl9, Gpr158, Pak5, Tfrc	4 of 8 genes, 50.00%	7189 of 21078 annotated genes, 34.11%
anatomical structure development (GO:0048856)	Acsl6, Gpr158, Sprr1a, Tfrc	4 of 8 genes, 50.00%	6561 of 21078 annotated genes, 31.13%
cell differentiation (GO:0030154)	Acsl6, Sprr1a, Tfrc	3 of 8 genes, 37.50%	4802 of 21078 annotated genes, 22.78%
immune system process (GO:0002376)	Btnl9, Tfrc	2 of 8 genes, 25.00%	3126 of 21078 annotated genes, 14.83%
nervous system process (GO:0050877)	Gpr158, Pak5	2 of 8 genes, 25.00%	2465 of 21078 annotated genes, 11.69%
cell adhesion (GO:0007155)	Fat2, Tfrc	2 of 8 genes, 25.00%	1551 of 21078 annotated genes, 7.36%
programmed cell death (GO:0012501)	Pak5, Tfrc	2 of 8 genes, 25.00%	2164 of 21078 annotated genes, 10.27%
sulfur compound metabolic process (GO:0006790)	Acsl6	1 of 8 genes, 12.50%	303 of 21078 annotated genes, 1.44%
regulation of DNA-templated transcription (GO:0006355)	Tfrc	1 of 8 genes, 12.50%	3290 of 21078 annotated genes, 15.61%
vesicle-mediated transport (GO:0016192)	Tfrc	1 of 8 genes, 12.50%	1567 of 21078 annotated genes, 7.43%
cell motility (GO:0048870)	Fat2	1 of 8 genes, 12.50%	1873 of 21078 annotated genes, 8.89%
transmembrane transport (GO:0055085)	Acsl6	1 of 8 genes, 12.50%	1426 of 21078 annotated genes, 6.77%
protein localization to plasma membrane (GO:0072659)	Gpr158	1 of 8 genes, 12.50%	318 of 21078 annotated genes, 1.51%
cell junction organization (GO:0034330)	Gpr158	1 of 8 genes, 12.50%	852 of 21078 annotated genes, 4.04%
protein-containing complex assembly (GO:0065003)	Tfrc	1 of 8 genes, 12.50%	1535 of 21078 annotated genes, 7.28%
mitochondrion organization (GO:0007005)	Tfrc	1 of 8 genes, 12.50%	581 of 21078 annotated genes, 2.76%
carbohydrate derivative metabolic process (GO:1901135)	Acsl6	1 of 8 genes, 12.50%	1019 of 21078 annotated genes, 4.83%
nucleobase-containing small molecule metabolic process (GO:0055086)	Acsl6	1 of 8 genes, 12.50%	574 of 21078 annotated genes, 2.72%
protein maturation (GO:0051604)	Cpxm2	1 of 8 genes, 12.50%	526 of 21078 annotated genes, 2.50%
cytoskeleton organization (GO:0007010)	Pak5	1 of 8 genes, 12.50%	1498 of 21078 annotated genes, 7.11%
DNA recombination (GO:0006310)	Tfrc	1 of 8 genes, 12.50%	335 of 21078 annotated genes, 1.59%
lipid metabolic process (GO:0006629)	Acsl6	1 of 8 genes, 12.50%	1448 of 21078 annotated genes, 6.87%

Figure 9: **GO Terms for Age-Sex Interaction (Top Padj and Log-Fold Change)**

Data Source

Mouse Reference Genome - GRCm39

https://www.gencodegenes.org/mouse/

GENCODE Mouse Genome Annotation (release M33)

https://www.gencodegenes.org/mouse/

NCBI GEO Sequencing Files - GSE202384

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE202384

Methods

Software

R Studio Version 4.2.2

https://www.r-project.org/

SRA Toolkit

https://hpc.nih.gov/apps/sratoolkit.html

STAR (Spliced Transcript Alignment to a Reference) Version 2.7.11a https://github.com/alexdobin/STAR/releases/download/2.7.11a/STAR_2.7.11a.zip

HTSeq (High-throughput sequence analysis in Python) Version 2.0.4

https://htseq.readthedocs.io/en/latest/

RNA-seq data download, assembly, and read counting

The initial steps of this project were accomplished via the command line using Linux. First, the SRA toolkit was used to download the fastq sequence data from the experiment from NCBI GEO. STAR was then used to build an index from the mouse genome and annotation files to generate the BAM and junction files. HTseq was then used to obtain the raw read counts for each annotated gene. We have included the code below.

Step 1.1 We installed the SRA Toolkit to download sequence data from NIH Sequence Read Archive (SRA).

Step 1.2 We downloaded the meta file *PRJNA835826-meta.csv*providing all information about the 12 libraries.

Step 1.3

We download the data from NCBI SRA as sra files with the following commands.

```
prefetch SRR19123213 --verbose
prefetch SRR19123214 --verbose
prefetch SRR19123215 --verbose
prefetch SRR19123216 --verbose
prefetch SRR19123217 --verbose
prefetch SRR19123218 --verbose
prefetch SRR19123219 --verbose
prefetch SRR19123220 --verbose
prefetch SRR19123221 --verbose
prefetch SRR19123222 --verbose
prefetch SRR19123223 --verbose
prefetch SRR19123224 --verbose
```

Then, we converted the sra files to two fastq files for each sample with the following commands.

```
fasterq-dump --threads 12 --verbose SRR19123213

fasterq-dump --threads 12 --verbose SRR19123214

fasterq-dump --threads 12 --verbose SRR19123215

fasterq-dump --threads 12 --verbose SRR19123216

fasterq-dump --threads 12 --verbose SRR19123217

fasterq-dump --threads 12 --verbose SRR19123218
```

```
fasterq-dump --threads 12 --verbose SRR19123219

fasterq-dump --threads 12 --verbose SRR19123220

fasterq-dump --threads 12 --verbose SRR19123221

fasterq-dump --threads 12 --verbose SRR19123222

fasterq-dump --threads 12 --verbose SRR19123223

fasterq-dump --threads 12 --verbose SRR19123224
```

Step 1.4 We installed STAR aligner.

Step 1.5 Then, we built STAR genome index for the mouse reference genome and annotation (GRCm39) with the following command.

```
STAR --runThreadN 12 --runMode genomeGenerate --genomeDir /home/student_no_backup/ibhat/P2/STAR_
--genomeFastaFiles /home/student_no_backup/ibhat/P2/GRCm39.primary_assembly.genome.fa
--sjdbGTFfile /home/student_no_backup/ibhat/P2/gencode.vM33.primary_assembly.basic.annotation.gtf
```

Step 1.6 We run STAR on the libraries to generate BAM and junction usage files with the following commands.

```
STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123213_1.fastq SRR19123213_2.fastq --outFileNamePrefix SRR19123213_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123214_1.fastq SRR19123214_2.fastq --outFileNamePrefix SRR19123214_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123215_1.fastq SRR19123215_2.fastq --outFileNamePrefix SRR19123215_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123216_1.fastq SRR19123216_2.fastq --outFileNamePrefix SRR19123216_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123217_1.fastq SRR19123217_2.fastq --outFileNamePrefix SRR19123217_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123218_1.fastq SRR19123218_2.fastq --outFileNamePrefix SRR19123218_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --
readFilesIn SRR19123219_1.fastq SRR19123219_2.fastq --outFileNamePrefix SRR19123219_
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts
```

```
STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --readFilesIn SRR19123220_1.fastq SRR19123220_2.fastq --outFileNamePrefix SRR19123220_ --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --readFilesIn SRR19123221_1.fastq SRR19123221_2.fastq --outFileNamePrefix SRR19123221_ --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --readFilesIn SRR19123222_1.fastq SRR19123222_2.fastq --outFileNamePrefix SRR19123222_ --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --readFilesIn SRR19123223_1.fastq SRR19123223_2.fastq --outFileNamePrefix SRR19123223_ --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts

STAR --runThreadN 12 --genomeDir /home/student_no_backup/ibhat/P2/STAR_index/ --readFilesIn SRR19123224_1.fastq SRR19123224_2.fastq --outFileNamePrefix SRR19123224_ --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts
```

Step 1.7 Next, we installed HTSeq.

Step 1.8 Finally, using htseq-count from HTSeq, we obtained read counts for each annotated gene with the following commands.

```
htseq-count -f bam -r pos -s no -i gene_id SRR19123213_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123213.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123214_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123214.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123215_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123215.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123216_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123216.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123217_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123217.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123218_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123218.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123219_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123219.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123220_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123220.txt
```

```
htseq-count -f bam -r pos -s no -i gene_id SRR19123221_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123221.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123222_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123222.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123223_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123223.txt

htseq-count -f bam -r pos -s no -i gene_id SRR19123224_Aligned.sortedByCoord.out.bam gencode.vM33.primary_assembly.basic.annotation.gtf > raw_read_counts_SRR19123224.txt
```

PCA Plots: Raw Reads and Normalized to Library

The raw read counts generated from the previous step were then organized into a count matrix and used to make create the ‘Raw Read PCA Plot’. Using DESeq2, the reads were then normalized to library-size and plotted again for comparison.

Differentially Expressed Genes

Using DESeq2, we apply a GLM model with a negative binomial distribution to the data in order to determine differentially expressed genes (DEGs) in the context of our experimental factors. We chose model 3, in order to determine the effects of age or sex on gene expression, as well as the effects of the interaction between age and sex.

~ Age + Sex + Age:Sex

The Top 5 DEgs, by either padj value or log-fold change value) were plotted as heatmaps with accompanying dendrograms to show expression levels of the top genes, as well as the overall similarity in expression profiles.

Discussion

RESULT SUMMARY

Sex represents an intrinsic biological factor that has been shown to be critical to our understanding of disease and aging. The underlying baseline gene expression profiles differ due to the inheritance of different sex chromosomes in males and females. Sex represents a confounding factor that must be accounted for in experimental design, even if that is not the variable of interest. In order to address questions of the effect of age and sex on gene expression in the mouse heart, we utilize the public data set published by the original authors of the study (Han et al 2022).

We calculated the total number of DEGs for each effect (FDR < 0.05) as follows:

Table 2: **Total Number DEGs for each effect**

Effect	Total DEGs
Age	1123
Sex	223
Age-Sex Interaction	169

Our overall results concur with the original paper, that sexual dimorphism and the underlying genetic differences influence subsequent molecular interactions and biological processes. While the top DEGs between age groups were associated with sex chromosomes (X or Y) and reproductive processes as expected, they also participated in transcription regulation and chromatin organization. These results reflect inherent differences in transcriptional processes between the sexes that produce different molecular landscapes for other biological processes, such as age, to act upon.

In terms of aging effect, many of the top DEGs were involved in cell homeostasis, metabolism, structure development, transport and signaling. Both the original authors results and ours confirm *Cd209d* as a DEG in older hearts, which functions as a receptor on macrophages and dendritic cells to recognize infectious agents. Another study shows that many of the DEGs in aging hearts are related to immune reactions and often upgregulated due to high protein turnover due to cellular damage (Bartling et al 2019).

While the effects of age or sex on gene expression seem clear, there appears to be a more complex effect of the interaction of these two factors. In terms of log-fold change, we see cardiac genes *Tfrc* and *Sprr1a* differentially expressed within the two age groups and the two sexes. Pak5 and Gpr158 were found to be upregulated in older hearts, which is interesting as members of the PAK family have been implicated in many age-related diseases (Amirthalingam et al 2021). Gpr158 represents a g-coupled protein receptor that is involved in pathways related to age-related memory loss (Kosmidis et al 2018).

Sexual dimorphisms are well known to be present in many cardiac diseases, in fact a study comparing mRNA microarray data of mouse and human heart tissue found that "sexually dimorphic genes overrepresented gene ontologies (GOs) important for cardiac homeostasis" (Tsuji et al 2020). These results highlight how crucially important sex-bias studies are to the progression of medical research. Treatments for disease that may be successful in one sex, may not be as effective in the other due to the fundamental differences in their gene expression profiles. A variety of biological factors can influence treatment efficacy, such as weight. Many drug doses are not standard per person, but determined based on patient weight to ensure efficacy.

In the field of cancer therapy, tumor biopsy samples from patients are often collected to determine their expression profile through molecular techniques, such as sequencing or visualization with known markers, which can influence the treatments that they may be amenable to. In the same way, understanding the molecular differences due to any biological factor, such as sex, can better direct research to develop therapies that are effective in as many people as possible.

CHALLENGES

One of the significant challenges of this work was storage space, as downloading the RNA-seq files and processing them through STAR requires an enormous amount of storage space. To solve this problem, we decided to utilize the computers on campus through the Computer Science Department. Storage was still an issue, which resulted in a slight change in workflow, downloading and processing individual files and then deleting them before starting on the next file.

FURTHER QUESTIONS

In order to delve further into the question of aging, it may be interesting to look specifically into energy metabolism and mitochondrial genes, as well as those involved in oxidative stress. Mitochondria are the main contributor of Reactive Oxygen Species (ROS), through the use of the electron transport chain, which can cause physical DNA damage (Cui et al 2012). ROS is produced through normal cellular functions, but can also be induced by environmental stress, such as radiation or UV. We may gain some more insight into the aging process between the sexes by narrowing focus to these pathways.

Distribution of Work

Indronil Bhattacharjee (IB) and Erica Flores (EF) both contributed to the project. IB provided the coding for data analysis and produced the outputs and figures. EF wrote the report, providing some more biological context for the results.

References

Han, Yu, Sara A Wennersten, Julianna M Wright, R W Ludwig, Edward Lau, and Maggie P Y Lam. 2022. "Proteogenomics Reveals Sex-Biased Aging Genes and Coordinated Splicing in Cardiac Aging." *Am J Physiol Heart Circ Physiol* 323 (3): H538–58. https://doi.org/10.1152/ajpheart.00244.2022.

Tower J. Sex-Specific Gene Expression and Life Span Regulation. Trends Endocrinol Metab. 2017 Oct;28(10):735-747. doi: 10.1016/j.tem.2017.07.002. Epub 2017 Aug 2. PMID: 28780002; PMCID: PMC5667568. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667568/

National Center for Health Statistics. Multiple Cause of Death 2018–2021 on CDC WONDER Database. Accessed February 2, 2023.

Pan Y, Yang J, Dai J, Xu X, Zhou X, Mao W. TFRC in cardiomyocytes promotes macrophage infiltration and activation during the process of heart failure through regulating Ccl2 expression mediated by hypoxia inducible factor-1 . Immun Inflamm Dis. 2023 Aug;11(8):e835. doi: 10.1002/iid3.835. PMID: 37647427; PMCID: PMC10461419. https://pubmed.ncbi.nlm.nih.gov/37647427/

Cui H, Kong Y, Zhang H. Oxidative stress, mitochondrial dysfunction, and aging. J Signal Transduct. 2012;2012:646354. doi: 10.1155/2012/646354. Epub 2011 Oct 2. PMID: 21977319; PMCID: PMC3184498. https://pubmed.ncbi.nlm.nih.gov/21977319/

21 Feb 2020: Tsuji M, Kawasaki T, Matsuda T, Arai T, Gojo S, et al. (2020) Correction: Sexual dimorphisms of mRNA and miRNA in human/murine heart disease. PLOS ONE 15(2): e0229750. https://doi.org/10.1371/journal.pone.0229750

Babett Bartling, Katja Niemann, Rainer U. Pliquett, Hendrik Treede, Andreas Simm, Altered gene expression pattern indicates the differential regulation of the immune response system as an important factor in cardiac aging, Experimental Gerontology, Volume 117, 2019, Pages 13-20, ISSN 0531-5565, https://doi.org/10.1016/j.exger.2018.05.001. (https://www.sciencedirect.com/science/article/pii/S0531556518300950)

Mohankumar Amirthalingam, Sundararaj Palanisamy, Shinkichi Tawata, p21-Activated kinase 1 (PAK1) in aging and longevity: An overview, Ageing Research Reviews, Volume 71, 2021, 101443, ISSN 1568-1637, https://doi.org/10.1016/j.arr.2021.101443. (https://www.sciencedirect.com/science/article/pii/S1568163721001902)

Kosmidis S, Polyzos A, Harvey L, Youssef M, Denny CA, Dranovsky A, Kandel ER. RbAp48 Protein Is a Critical Component of GPR158/OCN Signaling and Ameliorates Age-Related Memory Loss. Cell Rep. 2018 Oct 23;25(4):959-973.e6. doi: 10.1016/j.celrep.2018.09.077. PMID: 30355501; PMCID: PMC7725275.

Sun L, Yu J, Qi S, Hao Y, Liu Y, Li Z. Bone morphogenetic protein-10 induces cardiomyocyte proliferation and improves cardiac function after myocardial infarction. J Cell Biochem. 2014 Nov;115(11):1868-76. doi: 10.1002/jcb.24856. PMID: 24906204. https://pubmed.ncbi.nlm.nih.gov/24906204/

Snyder ML , Love S-A , Sorlie PD et al. **Redistribution of heart failure as the cause of death: the Atherosclerosis Risk in Communities Study.** *Popul Health Metr.* 2014; **12**: 10 10.1186/1478-7954-12-10

Santema BT Ouwerkerk W Tromp J et al. Identifying optimal doses of heart failure medications in men compared with women: a prospective, observational, cohort study. Lancet. 2019; 394: 1254-1263 https://doi.org/10.1016/S0140-6736(19)31792-1