CS509 - Project 1

Fruit Fly Genome and Differential Expression in Reproductive Tissues

Indronil Bhattacharjee (IB) and Erica Flores (EF)

Abstract

Scientific research continues to trend toward experimentation that produces high-throughput data, necessitating a partnership between research and computer programming, especially in the field of molecular biology. RNA sequencing is a common technique in moleculr studies, providing transcript expression data from biological samples. Using a statistical analysis program/language such as R, large data frames containing millions of reads can be interpreted to understand differential gene expression between multiple samples. In this study, we utilize several public databases to download and interpret gene expression in the ovaries and testis of the fruit fly.

Introduction

Bioinformatics is an emerging interdisciplinary field, which lies at the intersection of biology and computational analysis. With the modern advancements in molecular biology, it is now more feasible for researchers to conduct experiments that generate high-throughput data sets, resulting in a need for both an understanding of computer programming and an understanding of the biological processes involved.

RNA sequencing is an example of such an experiment, generating thousands or millions of sequence reads that need to be interpreted in a biological context. By mapping these reads to an annotated genome and generating read counts, we can calculate the difference in gene expression between different biological samples. You could determine how the expression profile of tumor cells changes in response to a cancer drug or compare the expression of specific genes in different tissues.

For this study, we are particularly interested in the differential expression of genes and transcripts in the reproductive tissues, ovary and testis, of the fruit fly. The fruit fly, *drosophila melanogaster*, has played a critical role in the advancement of our understanding of genetics, neuroscience, and disease (1). The extensive use of this model has resulted in an abundance of scientific databases, including a well-annotated genome.

Results

1. Transcriptome Assembly

For each reproductive tissue, ovary and testis, two replicates were individually aligned to the genome using the FR and RF parameters for strand specificity, resulting in 8 outputs (Figure 1). As expected, the overall alignment rate varied between the replicates. When we compare the alignment for individual replicates using either setting (FR or RF) for strand specificity, the outputs are identical.

For the testis samples, the overall alignment rate was 88.08% for replicate 1 and 90.29% for replicate 2 (Fig 1-2). In the ovary samples, the overall alignment rate was 90.01% in replicate 1 and 91.03% in replicate 2 (Fig 3-4).

Figures 1-4: HISAT2 alignment outputs - Testis 1, Testis 2, Ovary 1, Ovary 2

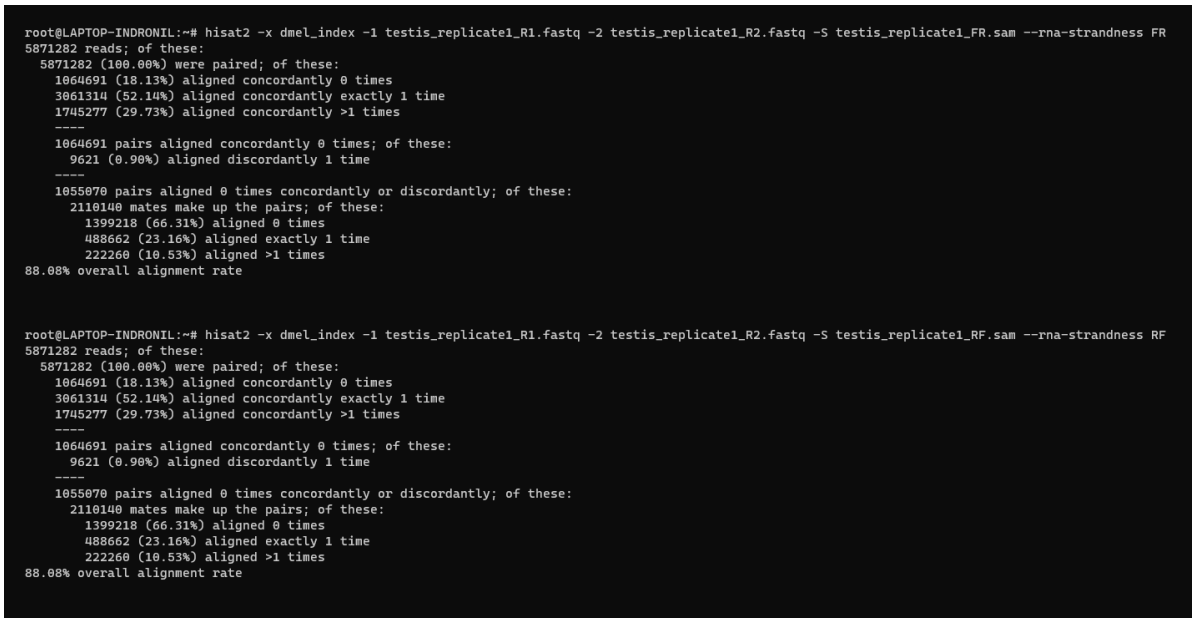


Figure 1: HISAT2 - Testis 1

```
root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 testis_replicate2_R1.fastq -2 testis_replicate2_R2.fastq -S testis_replicate2_FR.sam --rna-strandness FR
4663928 reads; of these:
  4663928 (100.00%) were paired; of these:
    775484 (16.63%) aligned concordantly 0 times
    2441565 (52.35%) aligned concordantly exactly 1 time
    1446879 (31.02%) aligned concordantly >1 times
    ----
    775484 pairs aligned concordantly 0 times; of these:
      8212 (1.06%) aligned discordantly 1 time
    ----
    767272 pairs aligned 0 times concordantly or discordantly; of these:
      1534544 mates make up the pairs; of these:
        905896 (59.03%) aligned 0 times
        443075 (28.87%) aligned exactly 1 time
        185573 (12.09%) aligned >1 times
90.29% overall alignment rate


root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 testis_replicate2_R1.fastq -2 testis_replicate2_R2.fastq -S testis_replicate2_RF.sam --rna-strandness RF
4663928 reads; of these:
  4663928 (100.00%) were paired; of these:
    775484 (16.63%) aligned concordantly 0 times
    2441565 (52.35%) aligned concordantly exactly 1 time
    1446879 (31.02%) aligned concordantly >1 times
    ----
    775484 pairs aligned concordantly 0 times; of these:
      8212 (1.06%) aligned discordantly 1 time
    ----
    767272 pairs aligned 0 times concordantly or discordantly; of these:
      1534544 mates make up the pairs; of these:
        905896 (59.03%) aligned 0 times
        443075 (28.87%) aligned exactly 1 time
        185573 (12.09%) aligned >1 times
90.29% overall alignment rate
```

Figure 2: 1A Testis 1

```
root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 ovary_replicate1_R1.fastq -2 ovary_replicate1_R2.fastq -S ovary_replicate1_FR.sam --rna-strandness FR
8341619 reads; of these:
  8341619 (100.00%) were paired; of these:
    1287027 (15.43%) aligned concordantly 0 times
    4637515 (55.59%) aligned concordantly exactly 1 time
    2417077 (28.98%) aligned concordantly >1 times
    ----
    1287027 pairs aligned concordantly 0 times; of these:
      22283 (1.73%) aligned discordantly 1 time
    ----
    1264744 pairs aligned 0 times concordantly or discordantly; of these:
      2529488 mates make up the pairs; of these:
        1500056 (59.30%) aligned 0 times
        742385 (29.35%) aligned exactly 1 time
        287047 (11.35%) aligned >1 times
91.01% overall alignment rate


root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 ovary_replicate1_R1.fastq -2 ovary_replicate1_R2.fastq -S ovary_replicate1_RF.sam --rna-strandness RF
8341619 reads; of these:
  8341619 (100.00%) were paired; of these:
    1287027 (15.43%) aligned concordantly 0 times
    4637515 (55.59%) aligned concordantly exactly 1 time
    2417077 (28.98%) aligned concordantly >1 times
    ----
    1287027 pairs aligned concordantly 0 times; of these:
      22283 (1.73%) aligned discordantly 1 time
    ----
    1264744 pairs aligned 0 times concordantly or discordantly; of these:
      2529488 mates make up the pairs; of these:
        1500056 (59.30%) aligned 0 times
        742385 (29.35%) aligned exactly 1 time
        287047 (11.35%) aligned >1 times
91.01% overall alignment rate
```

Figure 3: HISAT2 - Ovary 1

```
root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 ovary_replicate2_R1.fastq -2 ovary_replicate2_R2.fastq -S ovary_replicate2_FR.sam --rna-strandness FR
5823083 reads; of these:
  5823083 (100.00%) were paired; of these:
    916788 (15.74%) aligned concordantly 0 times
    3151778 (54.13%) aligned concordantly exactly 1 time
    1754517 (30.13%) aligned concordantly >1 times
    ----
    916788 pairs aligned concordantly 0 times; of these:
      15062 (1.64%) aligned discordantly 1 time
    ----
    901726 pairs aligned 0 times concordantly or discordantly; of these:
      1803452 mates make up the pairs; of these:
        1044948 (57.94%) aligned 0 times
        543465 (30.13%) aligned exactly 1 time
        215039 (11.92%) aligned >1 times
91.03% overall alignment rate

root@LAPTOP-INDRONIL:~# hisat2 -x dmel_index -1 ovary_replicate2_R1.fastq -2 ovary_replicate2_R2.fastq -S ovary_replicate2_RF.sam --rna-strandness RF
5823083 reads; of these:
  5823083 (100.00%) were paired; of these:
    916788 (15.74%) aligned concordantly 0 times
    3151778 (54.13%) aligned concordantly exactly 1 time
    1754517 (30.13%) aligned concordantly >1 times
    ----
    916788 pairs aligned concordantly 0 times; of these:
      15062 (1.64%) aligned discordantly 1 time
    ----
    901726 pairs aligned 0 times concordantly or discordantly; of these:
      1803452 mates make up the pairs; of these:
        1044948 (57.94%) aligned 0 times
        543465 (30.13%) aligned exactly 1 time
        215039 (11.92%) aligned >1 times
91.03% overall alignment rate
```

Figure 4: HISAT2 - Ovary 2

2. Transcriptome Quantification

The output of StringTie is a GTF of all the aligned reads for each sample of genes and transcripts, which includes abundance data, FPKM and TPM, as well. These files will be included with the submission but, for the purpose of the report, we show the first few rows of each sample for genes and transcripts.

```
# Load necessary libraries
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.2.3

Warning: package 'ggplot2' was built under R version 4.2.3

Warning: package 'tibble' was built under R version 4.2.3

Warning: package 'tidyr' was built under R version 4.2.3

Warning: package 'readr' was built under R version 4.2.3

Warning: package 'purrr' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3

Warning: package 'stringr' was built under R version 4.2.3

Warning: package 'forcats' was built under R version 4.2.3

Warning: package 'lubridate' was built under R version 4.2.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
library(dplyr)
library(ggplot2)

# Load the transcript abundance data for both ovary and testis replicates
ovary_replicate1_transcript <- read.table("transcript_abundance_ovary_replicate1.tab", hea
ovary_replicate2_transcript <- read.table("transcript_abundance_ovary_replicate2.tab", hea
testis_replicate1_transcript <- read.table("transcript_abundance_testis_replicate1.tab", h
testis_replicate2_transcript <- read.table("transcript_abundance_testis_replicate1.tab", h

# Load the gene abundance data for both ovary and testis replicates
ovary_replicate1_gene <- read.table("gene_abundance_ovary_replicate1.tab", header = TRUE,
ovary_replicate2_gene <- read.table("gene_abundance_ovary_replicate2.tab", header = TRUE,
testis_replicate1_gene <- read.table("gene_abundance_testis_replicate1.tab", header = TRUE
testis_replicate2_gene <- read.table("gene_abundance_testis_replicate2.tab", header = TRUE


#print rows to report gene abundance

head(ovary_replicate1_gene)
```

```
      Gene.ID Gene.Name Reference Strand Start   End  Coverage      FPKM
1 FBgn0031208         -        2L      +  7529  9484  0.462234  0.307535
2 FBgn0002121         -        2L      -  9839 21376 41.767136 31.712936
3 FBgn0051973         -        2L      - 25402 65404  0.503833  0.524640
4 FBgn0267987         -        2L      + 54817 55767  0.000000  0.000000
5 FBgn0266879         -        2L      + 66318 66524  0.000000  0.000000
6 FBgn0067779         -        2L      + 66482 71390 32.008511 24.743692
       TPM
1  0.766298
2 79.020485
3  1.307268
4  0.000000
5  0.000000
6 61.654926
```

```
head(ovary_replicate2_gene)
```

```
      Gene.ID Gene.Name Reference Strand Start   End  Coverage      FPKM
1 FBgn0031208         -        2L      +  7529  9484  0.179787  0.171182
2 FBgn0002121         -        2L      -  9839 21376 26.942322 29.362421
3 FBgn0031209         -        2L      - 21823 25155  0.000000  0.000000
4 FBgn0263584         -        2L      + 21952 24237  0.232735  0.221595
5 FBgn0051973         -        2L      - 25402 65404  0.338708  0.504750
6 FBgn0267987         -        2L      + 54817 55767  0.161935  0.154184
       TPM
1  0.394298
2 67.633194
3  0.000000
4  0.510421
5  1.162637
6  0.355146
```

```
head(testis_replicate1_gene)
```

```
      Gene.ID Gene.Name Reference Strand Start   End  Coverage      FPKM
1 FBgn0031208         -        2L      +  7529  9484 51.843086 50.786350
2 FBgn0002121         -        2L      -  9839 21376  1.240430  1.336045
3 FBgn0031209         -        2L      - 21823 25155  0.000000  0.000000
4 FBgn0263584         -        2L      + 21952 24237  0.408072  0.399754
5 FBgn0051973         -        2L      - 25402 65404  0.545158  0.835253
```

```
6 FBgn0267987         -        2L      + 54817 55767  0.957939  0.938413
       TPM
1 75.254364
2  1.979730
3  0.000000
4  0.592349
5  1.237663
6  1.390525
```

```
head(testis_replicate2_gene)
```

```
     Gene.ID Gene.Name Reference Strand Start   End  Coverage      FPKM
1 FBgn0031208         -        2L      +  7529  9484 51.432980 61.989746
2 FBgn0002121         -        2L      -  9839 21376  1.116486  1.479531
3 FBgn0031209         -        2L      - 21823 25155  0.000000  0.000000
4 FBgn0263584         -        2L      + 21952 24237  0.852915  1.027978
5 FBgn0051973         -        2L      - 25402 65404  0.912768  1.790706
6 FBgn0267987         -        2L      + 54817 55767  0.000000  0.000000
        TPM
1 119.639702
2   2.855483
3   0.000000
4   1.983990
5   3.456047
6   0.000000
```

```
#print rows o report transcript abundance

head(ovary_replicate1_transcript)
```

```
  Transcript.ID Transcript.Name Reference Strand Start   End Coverage     FPKM
1   FBtr0475186               -        2L      +  7529  9484 0.462234 0.307535
2   FBtr0078166               -        2L      -  9839 21376 0.000000 0.000000
3   FBtr0078167               -        2L      -  9839 21376 0.000000 0.000000
4   FBtr0078169               -        2L      -  9839 21376 0.000000 0.000000
5   FBtr0306589               -        2L      -  9839 21376 0.000000 0.000000
6   FBtr0306590               -        2L      -  9839 21376 0.000000 0.000000
      TPM
1 0.766298
2 0.000000
```

```
3 0.000000
4 0.000000
5 0.000000
6 0.000000
```

```
head(ovary_replicate2_transcript)
```

```
  Transcript.ID Transcript.Name Reference Strand Start   End Coverage     FPKM
1   FBtr0475186               -        2L      +  7529  9484 0.179787 0.171182
2   FBtr0078171               -        2L      -  9839 18570 0.000000 0.000000
3   FBtr0078166               -        2L      -  9839 21376 0.000000 0.000000
4   FBtr0078167               -        2L      -  9839 21376 0.000000 0.000000
5   FBtr0078168               -        2L      -  9839 21376 0.000000 0.000000
6   FBtr0078169               -        2L      -  9839 21376 0.000000 0.000000
       TPM
1 0.394298
2 0.000000
3 0.000000
4 0.000000
5 0.000000
6 0.000000
```

```
head(testis_replicate1_transcript)
```

```
  Transcript.ID Transcript.Name Reference Strand Start   End Coverage     FPKM
1   FBtr0475186               -        2L      +  7529  9484 51.84309 50.78635
2   FBtr0078170               -        2L      -  9839 18570  0.00000  0.00000
3   FBtr0078171               -        2L      -  9839 18570  0.00000  0.00000
4   FBtr0078166               -        2L      -  9839 21376  0.00000  0.00000
5   FBtr0078167               -        2L      -  9839 21376  0.00000  0.00000
6   FBtr0078168               -        2L      -  9839 21376  0.00000  0.00000
       TPM
1 75.25436
2  0.00000
3  0.00000
4  0.00000
5  0.00000
6  0.00000
```

```
head(testis_replicate2_transcript)
```

```
  Transcript.ID Transcript.Name Reference Strand Start   End Coverage     FPKM
1   FBtr0475186               -        2L      +  7529  9484 51.84309 50.78635
2   FBtr0078170               -        2L      -  9839 18570  0.00000  0.00000
3   FBtr0078171               -        2L      -  9839 18570  0.00000  0.00000
4   FBtr0078166               -        2L      -  9839 21376  0.00000  0.00000
5   FBtr0078167               -        2L      -  9839 21376  0.00000  0.00000
6   FBtr0078168               -        2L      -  9839 21376  0.00000  0.00000
       TPM
1 75.25436
2  0.00000
3  0.00000
4  0.00000
5  0.00000
6  0.00000
```

3. Genes and Transcripts of High Fold-Change

Replicate data was merged, in order to create a single file for each tissue and the log fold change for each gene or transcript. This data is visualized as heatmaps for the top genes or transcripts with the highest and lowest fold changes. Abundance metrics include TPM and FPKM, as well as coverage.

```
testis_merged_gene <- bind_rows(testis_replicate1_gene, testis_replicate2_gene)
ovary_merged_gene <- bind_rows(ovary_replicate1_gene, ovary_replicate2_gene)

# Calculate the average gene abundance for ovary and testis
avg_abundance_ovary <- rowMeans(ovary_merged_gene[, 7:9])
avg_abundance_testis <- rowMeans(testis_merged_gene[, 7:9])

# Calculate the log fold change (r_g)
log_fold_change_gene <- log2((1 + avg_abundance_ovary) / (1 + avg_abundance_testis))


testis_merged_transcript <- bind_rows(testis_replicate1_transcript, testis_replicate2_tran
ovary_merged_transcript <- bind_rows(ovary_replicate1_transcript, ovary_replicate2_transcr

# Calculate the average transcript abundance for ovary and testis
avg_abundance_ovary <- rowMeans(ovary_merged_transcript[, 7:9])
avg_abundance_testis <- rowMeans(testis_merged_transcript[, 7:9])
```

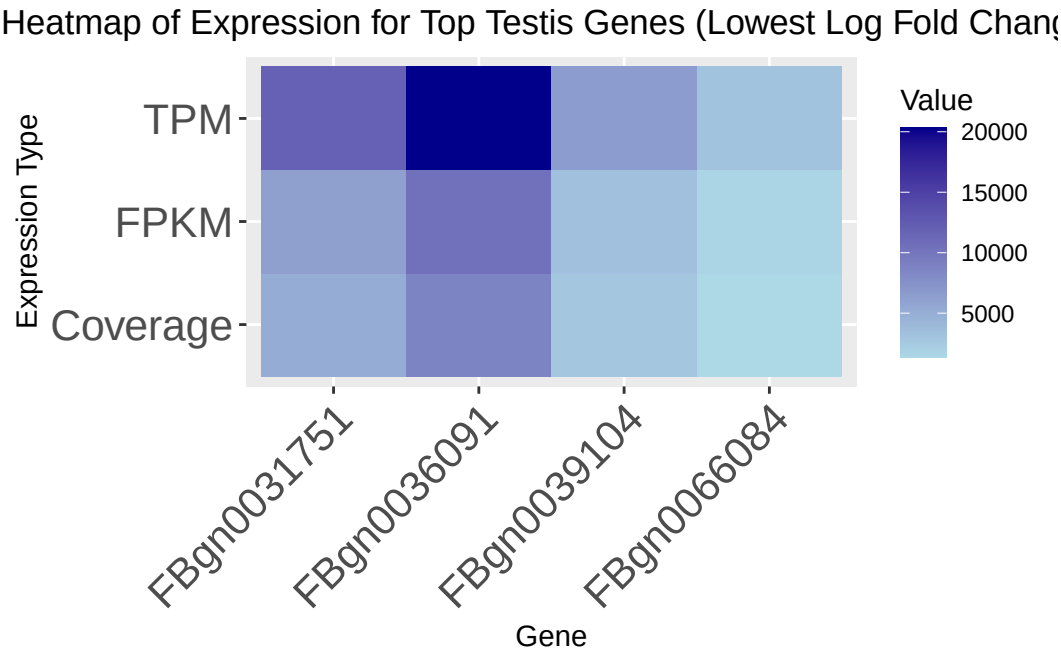
```r
# Calculate the log fold change (r_g)
log_fold_change_transcript <- log2((1 + avg_abundance_ovary) / (1 + avg_abundance_testis))
```

Genes - Testis Highest and Lowest Fold Change

```r
# Add the log fold change values to the data
testis_merged_gene$log_fold_change_gene <- log_fold_change_gene

#HIGHEST
# Find the top genes with the greatest and lowest log fold change values
top_genes_greatest <- testis_merged_gene %>% arrange(desc(log_fold_change_gene)) %>% head(


# Visualize the expression of top genes using a boxplot
# Assuming the columns "Coverage" to "TPM" represent expression values
top_genes_names_greatest <- c(top_genes_greatest$Gene.ID)
top_testis_genes_highest <- top_genes_names_greatest

# Filter data for topgenes
testis_merged_gene_filtered <- testis_merged_gene %>%
  filter(Gene.ID %in% top_genes_names_greatest)

# Pivot the data for heatmap visualization
heatmap_data <- testis_merged_gene_filtered %>%
  select(Gene.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Gene.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Gene.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Testis Genes (Highest Log Fold Change)",
       x = "Gene",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```


```
#LOWEST
top_genes_lowest <- testis_merged_gene %>% arrange(log_fold_change_gene) %>% head(5)
top_genes_names_lowest <- c(top_genes_lowest$Gene.ID)
top_testis_genes_lowest <- top_genes_names_lowest

# Filter data for top genes
testis_merged_gene_filtered <- testis_merged_gene %>%
  filter(Gene.ID %in% top_genes_names_lowest)

# Pivot the data for heatmap visualization
heatmap_data <- testis_merged_gene_filtered %>%
  select(Gene.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Gene.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Gene.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Testis Genes (Lowest Log Fold Change)",
       x = "Gene",
       y = "Expression Type") +
```

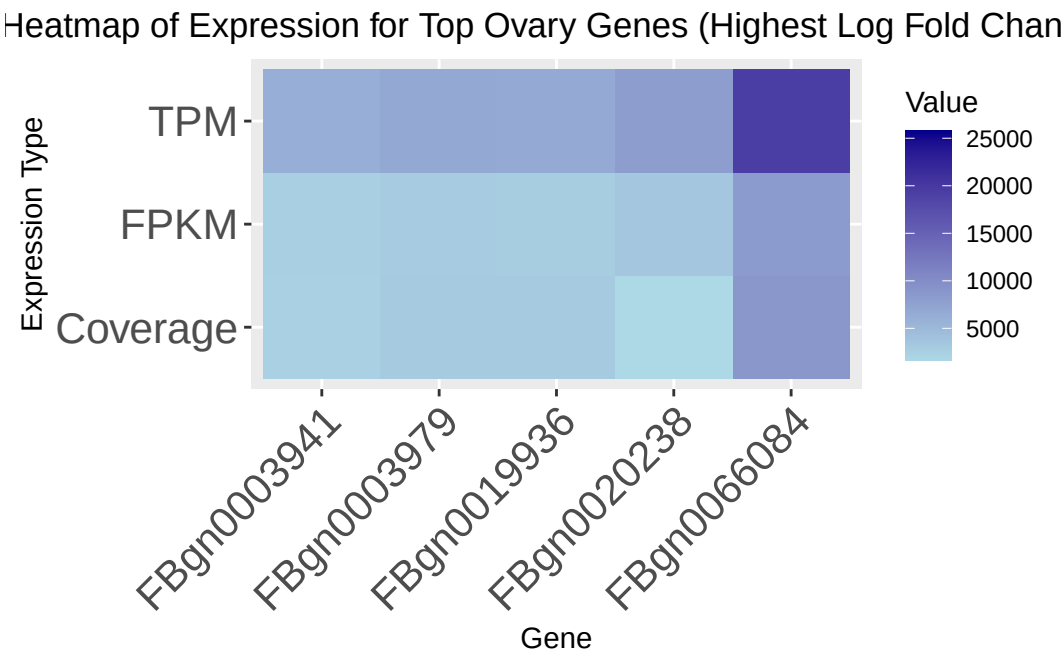
```
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```



Genes - Ovary Highest and Lowest Fold Change

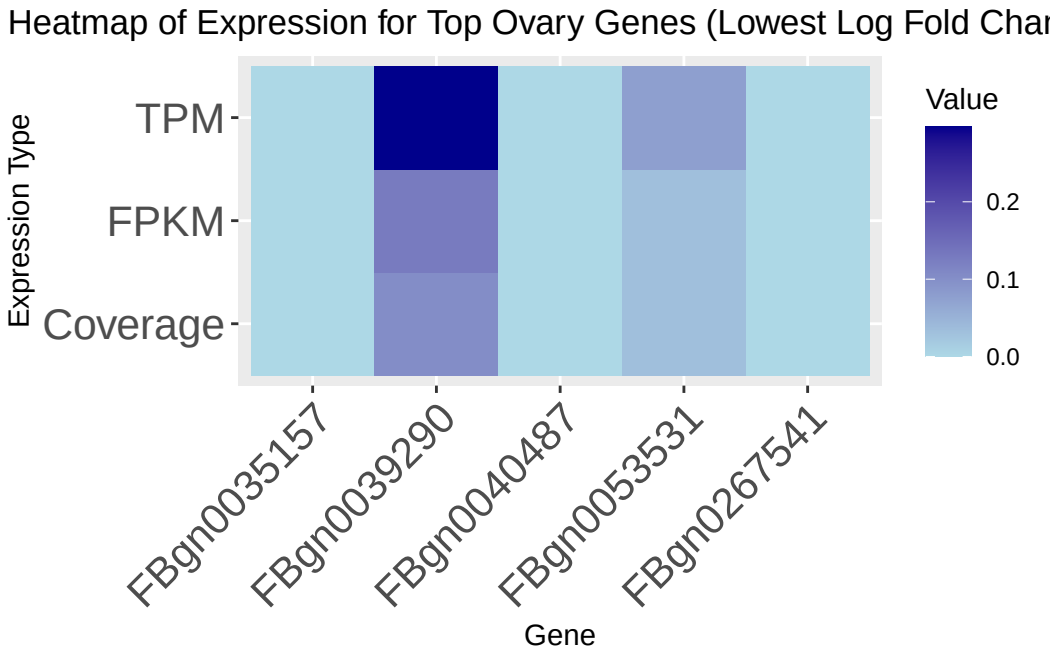
```
# Add the log fold change values to the data
ovary_merged_gene$log_fold_change_gene <- log_fold_change_gene

#HIGHEST
# Find the top genes with the greatest and lowest log fold change values
top_genes_greatest <- ovary_merged_gene %>% arrange(desc(log_fold_change_gene)) %>% head(5


# Visualize the expression of top genes using a boxplot
# Assuming the columns "Coverage" to "TPM" represent expression values
top_genes_names_greatest <- c(top_genes_greatest$Gene.ID)
top_ovary_genes_highest <- top_genes_names_greatest

# Filter data for top genes
ovary_merged_gene_filtered <- ovary_merged_gene %>%
  filter(Gene.ID %in% top_genes_names_greatest)
```

```r
# Pivot the data for heatmap visualization
heatmap_data <- ovary_merged_gene_filtered %>%
  select(Gene.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Gene.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Gene.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Ovary Genes (Highest Log Fold Change)",
       x = "Gene",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```



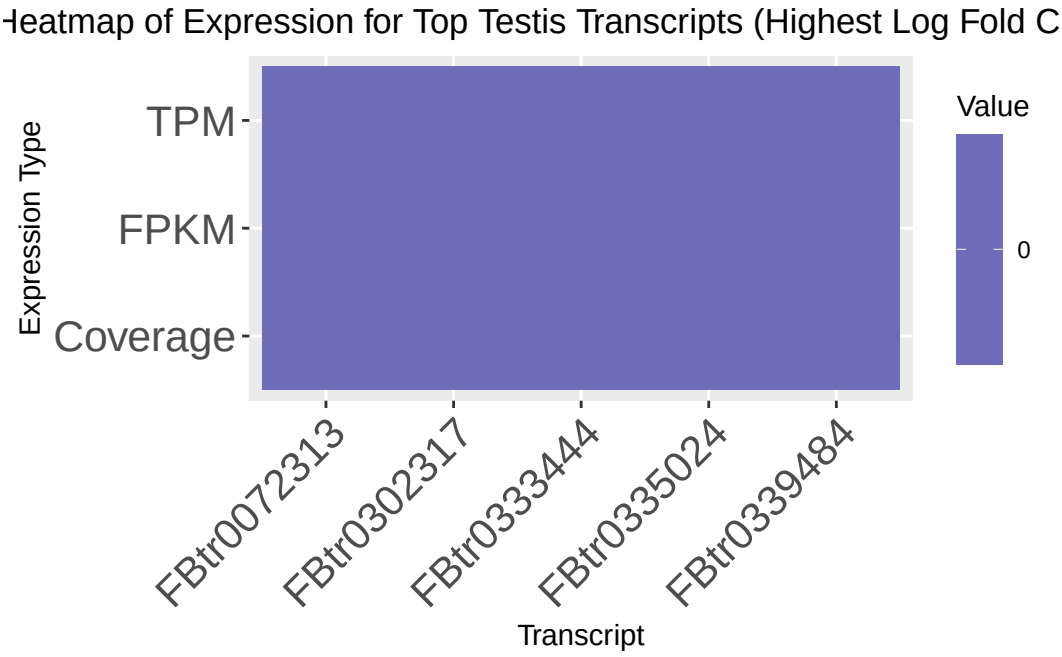

```r
#LOWESTgene
top_genes_lowest <- ovary_merged_gene %>% arrange(log_fold_change_gene) %>% head(5)
top_genes_names_lowest <- c(top_genes_lowest$Gene.ID)
top_ovary_genes_lowest <- top_genes_names_lowest
```

```r
# Filter data for top genes
ovary_merged_gene_filtered <- ovary_merged_gene %>%
  filter(Gene.ID %in% top_genes_names_lowest)

# Pivot the data for heatmap visualization
heatmap_data <- ovary_merged_gene_filtered %>%
  select(Gene.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Gene.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Gene.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Ovary Genes (Lowest Log Fold Change)",
       x = "Gene",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```



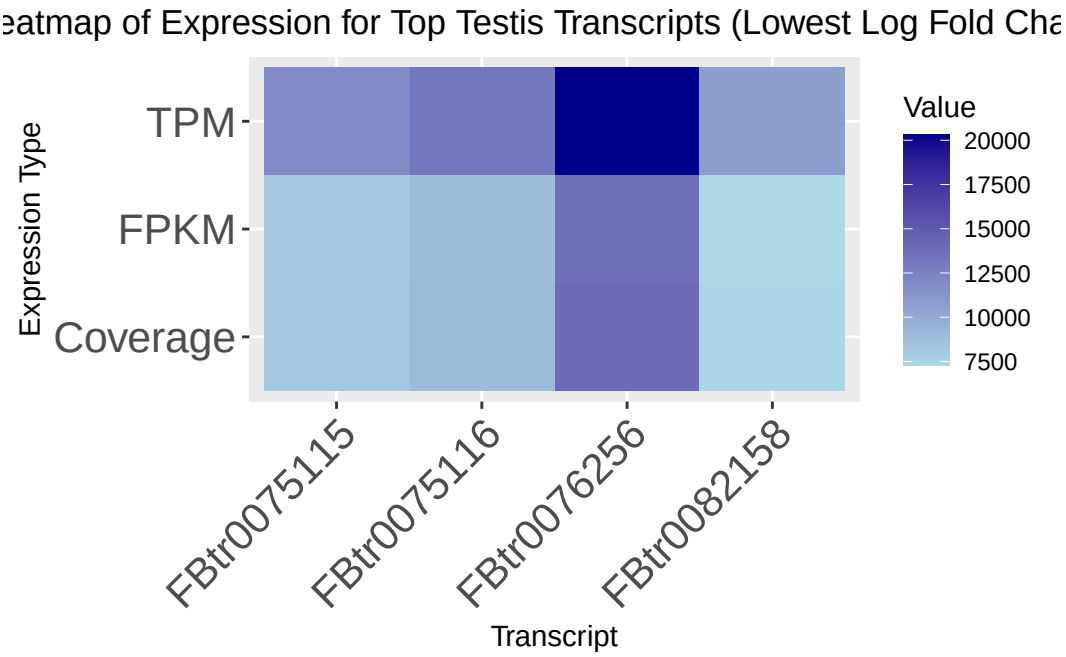
Transcripts - Testis Highest and Lowest Fold Change

```r
# Add the log fold change values to the data
testis_merged_transcript$log_fold_change_transcript <- log_fold_change_transcript

#HIGHEST
# Find the top transcripts with the greatest and lowest log fold change values
top_transcripts_greatest <- testis_merged_transcript %>% arrange(desc(log_fold_change_tran


# Visualize the expression of top transcripts using a boxplot
# Assuming the columns "Coverage" to "TPM" represent expression values
top_transcripts_names_greatest <- c(top_transcripts_greatest$Transcript.ID)
top_testis_transcripts_highest <- top_transcripts_names_greatest

# Filter data for top transcripts
testis_merged_transcript_filtered <- testis_merged_transcript %>%
  filter(Transcript.ID %in% top_transcripts_names_greatest)

# Pivot the data for heatmap visualization
heatmap_data <- testis_merged_transcript_filtered %>%
  select(Transcript.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Transcript.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Transcript.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Testis Transcripts (Highest Log Fold Change)
       x = "Transcript",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```




```r
#LOWEST
top_transcripts_lowest <- testis_merged_transcript %>% arrange(log_fold_change_transcript)
top_transcripts_names_lowest <- c(top_transcripts_lowest$Transcript.ID)
top_testis_transcripts_lowest <- top_transcripts_names_lowest

# Filter data for top transcripts
testis_merged_transcript_filtered <- testis_merged_transcript %>%
  filter(Transcript.ID %in% top_transcripts_names_lowest)

# Pivot the data for heatmap visualization
heatmap_data <- testis_merged_transcript_filtered %>%
  select(Transcript.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Transcript.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Transcript.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Testis Transcripts (Lowest Log Fold Change)"
       x = "Transcript",
       y = "Expression Type") +
```

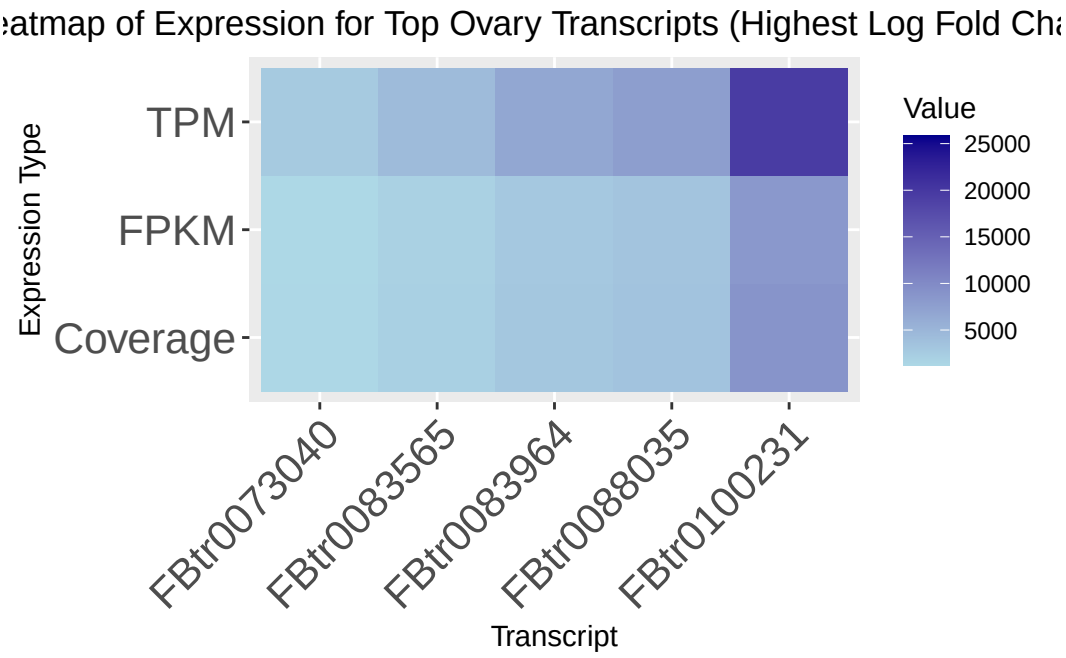
```
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```



Transcripts - Ovary Highest and Lowest Fold Change

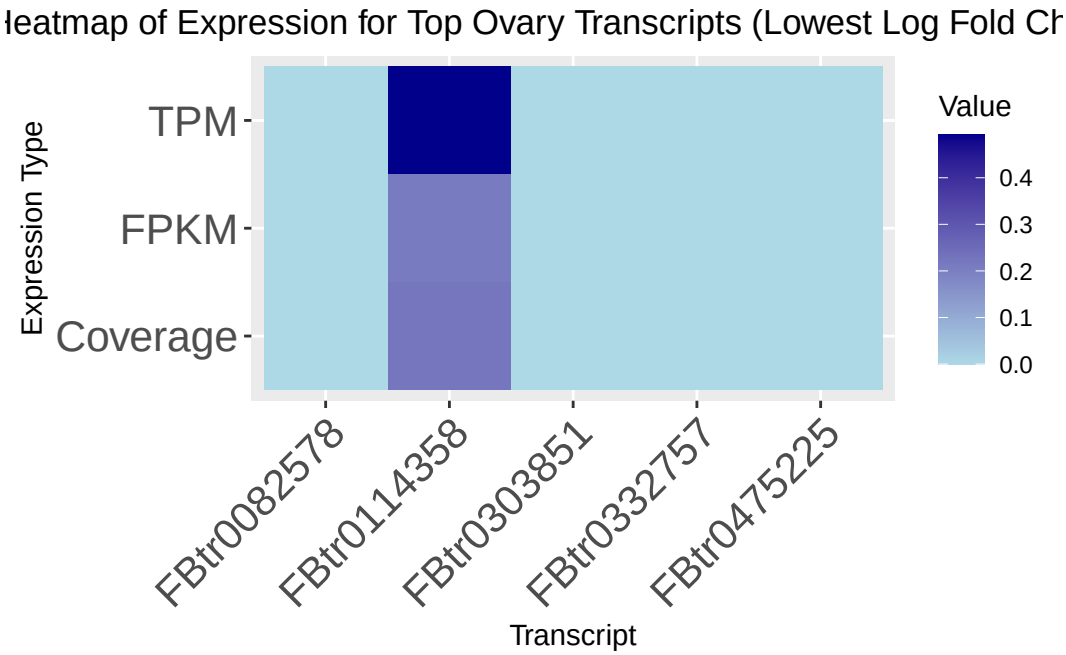
```
# Add the log fold change values to the data
ovary_merged_transcript$log_fold_change_transcript <- log_fold_change_transcript

#HIGHEST
# Find the top transcripts with the greatest and lowest log fold change values
top_transcripts_greatest <- ovary_merged_transcript %>% arrange(desc(log_fold_change_trans


# Visualize the expression of top transcripts using a boxplot
# Assuming the columns "Coverage" to "TPM" represent expression values
top_transcripts_names_greatest <- c(top_transcripts_greatest$Transcript.ID)
top_ovary_transcripts_highest <- top_transcripts_names_greatest

# Filter data for top transcripts
ovary_merged_transcript_filtered <- ovary_merged_transcript %>%
  filter(Transcript.ID %in% top_transcripts_names_greatest)
```

```r
# Pivot the data for heatmap visualization
heatmap_data <- ovary_merged_transcript_filtered %>%
  select(Transcript.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Transcript.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Transcript.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Ovary Transcripts (Highest Log Fold Change)"
       x = "Transcript",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```




```r
#LOWEST
top_transcripts_lowest <- ovary_merged_transcript %>% arrange(log_fold_change_transcript)
top_transcripts_names_lowest <- c(top_transcripts_lowest$Transcript.ID)
top_ovary_transcripts_lowest <- top_transcripts_names_lowest
```

```r
# Filter data for top transcripts
ovary_merged_transcript_filtered <- ovary_merged_transcript %>%
  filter(Transcript.ID %in% top_transcripts_names_lowest)

# Pivot the data for heatmap visualization
heatmap_data <- ovary_merged_transcript_filtered %>%
  select(Transcript.ID, Coverage:TPM) %>%
  pivot_longer(cols = -Transcript.ID, names_to = "Expression", values_to = "Value")

# Create a heatmap
options(repr.plot.width = 10, repr.plot.height = 8)
ggplot(heatmap_data, aes(x = Transcript.ID, y = Expression, fill = Value)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Expression for Top Ovary Transcripts (Lowest Log Fold Change)",
       x = "Transcript",
       y = "Expression Type") +
  theme(axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16),plot.title = element_text(hjust = 0.5))
```



Summary Table for Gene/Transcript Functions of Top Hits

	Flybase ID	Gene Name
Ovary Genes Highest Fold Change	FBgn0003941	Ribosomal protein L40
	FBgn0003979	Vitelline membrane 26Aa
	FBgn0019936	Ribosomal protein S20
	FBgn0020238	14-3-3ε embryonic hatching, germ cell migration, gonad formation, wing venation and eye development.
	FBgn0066084	Ribosomal protein L41
Ovary Genes Lowest Fold Change	FBgn0035157	CG13894 presumptive embryonic/larval nervous syste
	FBgn0039290	CG13654 Orthologous to human PGAP6 (post-GPI attachment to proteins 6)
	FBgn0040487	Brother of Bearded A negative regulation of Notch signaling pathway; and sensory organ precursor cell fate determination.
	FBgn0053531	Discoidin domain receptor egulation of neuron projection development
	FBgn0053531	long non-coding RNA:CR45881 The **biological processes** in which it is involved are not known

Figure 5: Ovary Genes

	Flybase ID	Gene Name
Testis Genes Highest Fold Change	FBgn0010438	mtSSB (mitochondrial single stranded DNA-binding protein) mtDNA replication
	FBgn0020429	Glutamate receptor IIB muscle glutamate receptor
	FBgn0034769	Odorant-binding protein 58c sensory perception of chemical stimulus.
	FBgn0038250	CG3505 proteolysis. embryonic/larval fat body.
	FBgn0259954	CG42464 Predicted to enable serine-type endopeptidase inhibitor activity.
Testis Genes Lowest Fold Change	FBgn0031751	CG9016 The **biological processes** in which it is involved are not known
	FBgn0036091	CG18628 Involved in sexual reproduction.
	FBgn0039104	CG10252 spermatozoon; and testis.
	FBgn0066084	Ribosomal protein L41

Figure 6: Testis Genes

	Flybase ID	Gene Name
Ovary Transcripts Highest Fold Change	FBtr0073040	Hsp83-RA Molecular chaperone
	FBtr0083565	14-3-3ε-RA embryonic hatching, germ cell migration, gonad formation, wing venation and eye development.
	FBtr0083964	RpS20-RA A structural constituent of ribosome
	FBtr0088035	eEF1α1-RA protein biosynthesis
	FBtr0100231	RpL41-RA Ribosomal protein L41
Ovary Transcripts Lowest Fold Change	FBtr0082578	CG33098-RD locomotion and post-embryonic development
	FBtr0114358	ncrna chromosome:BDGP6:3L:20453733:20453821
	FBtr0303851	CG42832-RA **biological process** described with: .
	FBtr0332757	asRNA:CR43885-RA non-coding RNA
	FBtr0475225	CG46434-RB **biological processes** in which it is involved are not known.

Figure 7: Ovary Transcripts

	Flybase ID	Gene Name
Testis Transcripts Highest Fold Change	FBtr0072313	piopio (pio) encodes a zona pellucida (ZP) domain protein
	FBtr0302317	Brf RNA polymerase III subunit
	FBtr0333444	Myocardin-related transcription factor (Mrtf)
	FBtr0335024	Osa BAP chromatin remodeling, Wnt signaling
	FBtr0339484	menage a trois (metro) expansion of larval neuromuscular junctions (NMJs)
Testis Transcripts Lowest Fold Change	FBtr0075115	CG32197-RA (Met75Ca) Involved in sexual reproduction.
	FBtr0075116	CG18064-RA Met75Cb Involved in sexual reproduction.
	FBtr0076256	CG18628-RA Involved in sexual reproduction.
	FBtr0082158	Metallothionein A (MtnA)

Figure 8: Testis Transcripts

Top Transcripts located within genomic region of top genes

```
cat("Top Testis Genes and Transcripts with lowest Log Fold Change\n")
```

```
Top Testis Genes and Transcripts with lowest Log Fold Change
```

```
cat("============================================================\n")
```

```
============================================================
```

```
filtered_gene_data <- testis_merged_gene %>%
  filter(Gene.ID %in% top_testis_genes_lowest)

# Extract the start and end positions for the selected genes
gene_ids <- unique(filtered_gene_data$Gene.ID)
gene_start_positions <- unique(filtered_gene_data$Start)
gene_end_positions <- unique(filtered_gene_data$End)


filtered_transcript_data <- testis_merged_transcript %>%
  filter(Transcript.ID %in% top_testis_transcripts_lowest)

# Extract the start and end positions for the selected transcripts
transcript_ids <- unique(filtered_transcript_data$Transcript.ID)
transcripts_start_positions <- unique(filtered_transcript_data$Start)
transcripts_end_positions <- unique(filtered_transcript_data$End)

for (x in 1:length(transcripts_start_positions)) {
    for (y in 1:length(gene_start_positions)) {
      if(gene_end_positions[y] >= transcripts_end_positions[x] & gene_start_positions[y] <
      {
          cat("Transcript", transcript_ids[x], " is in genomic region of ", gene_ids[y], "
      }
  }
}
```

```
Transcript FBtr0076256  is in genomic region of  FBgn0036091 gene
```

```
  cat("Top Ovary Genes and Transcripts with highest Log Fold Change\n")
```

```
Top Ovary Genes and Transcripts with highest Log Fold Change
```

```
  cat("============================================================\n")
```

```
============================================================
```

```
  filtered_gene_data <- ovary_merged_gene %>%
    filter(Gene.ID %in% top_ovary_genes_highest)

  # Extract the start and end positions for the selected genes
  gene_ids <- unique(filtered_gene_data$Gene.ID)
  gene_start_positions <- unique(filtered_gene_data$Start)
  gene_end_positions <- unique(filtered_gene_data$End)


  filtered_transcript_data <- ovary_merged_transcript %>%
    filter(Transcript.ID %in% top_ovary_transcripts_highest)

  # Extract the start and end positions for the selected transcripts
  transcript_ids <- unique(filtered_transcript_data$Transcript.ID)
  transcripts_start_positions <- unique(filtered_transcript_data$Start)
  transcripts_end_positions <- unique(filtered_transcript_data$End)

  for (x in 1:length(transcripts_start_positions)) {
      for (y in 1:length(gene_start_positions)) {
        if(gene_end_positions[y] >= transcripts_end_positions[x] & gene_start_positions[y] <
        {
            cat("Transcript", transcript_ids[x], " is in genomic region of ", gene_ids[y], "
        }
    }
  }
```

```
Transcript FBtr0100231  is in genomic region of  FBgn0066084 gene
Transcript FBtr0083565  is in genomic region of  FBgn0020238 gene
Transcript FBtr0083964  is in genomic region of  FBgn0019936 gene
```

Data Source

Genome sequence and annotation files were acquired from FlyBase.

Genome Sequence File - FlyBase (FB2023_04)

http://ftp.flybase.net/releases/FB2023_04/dmel_r6.53/fasta/dmel-all-chromosome-r6.53.fasta.gz

Genome Annotation File - FlyBase

http://ftp.flybase.net/releases/FB2023_04/dmel_r6.53/gtf/dmel-all-r6.53.gtf.gz

RNA-seq for Fly Testis and Ovary were acquired from the Encode Project.

RNAseq - Testis - Encode Project

https://www.encodeproject.org/experiments/ENCSR254JFC/

RNAseq - Ovary - Encode Project

https://www.encodeproject.org/experiments/ENCSR272DXE/

Methods

Software

R Studio - R version 4.2.2 (2022-10-31 urct)

Jupyter Notebook

HISAT2 - HISAT2 2.2.1

Bowtie2-2.5.1

Transcriptome Assembly

Two replicates of ovary sequence reads and two replicates of testis sequence reads were individually aligned to the fly genome using HISAT2. Both the FR and RF parameters were used initially to determine the best method. The links to these database files can be found in the 'Data Source' section. Samples were take at Day 4 after synchronization at occlusion.

HISAT

1. Download the Drosophila melanogaster reference genome from the provided URL.

2. Index the reference genome using HISAT2:

hisat2-build dmel-all-chromosome-r6.53.fasta dmel_index

3. Download the RNA-seq data for the testis and ovary from the provided URLs.

4. Map the reads to the reference genome for both testis and ovary samples using HISAT2 with both FR and RF strand specificity options:

For FR strand specificity:

hisat2 -x dmel_index -1 testis_replicate1_R1.fastq.gz -2 testis_replicate1_R2.fastq.gz -S testis_replicate1_FR.sam --rna-strandness FR

hisat2 -x dmel_index -1 testis_replicate2_R1.fastq.gz -2 testis_replicate2_R2.fastq.gz -S testis_replicate2_FR.sam --rna-strandness FR

For RF strand specificity:

hisat2 -x dmel_index -1 testis_replicate1_R1.fastq.gz -2 testis_replicate1_R2.fastq.gz -S testis_replicate1_RF.sam --rna-strandness RF

hisat2 -x dmel_index -1 testis_replicate2_R1.fastq.gz -2 testis_replicate2_R2.fastq.gz -S testis_replicate2_RF.sam --rna-strandness RF

5.Convert the SAM files to BAM format using samtools view:

samtools view -b -o testis_replicate1_FR.bam testis_replicate1_FR.sam

samtools view -b -o testis_replicate2_FR.bam testis_replicate2_FR.sam

samtools view -b -o testis_replicate1_RF.bam testis_replicate1_RF.sam

samtools view -b -o testis_replicate2_RF.bam testis_replicate2_RF.sam

6. Choose the strand specificity option (FR or RF) that resulted in the largest number of paired alignments for further analysis.

Transcript Quantification

Quantification

HISAT2 was used to also output a SAM file with gene counts, which was then passed to StringTie. StringTie was used to compile a GTF file including details of all the aligned reads, such as chromosome location, start/end position, gene id, etc. This file also contains columns 7-9, which calculate coverage, TPM (transcripts per million), and FPKM (fragements per kilobase of transcript per million reads mapped).

1. Download the genome annotation file in GTF format from the provided URL.
2. Sort the bam files.

samtools sort -o testis_replicate1_FR_sorted.bam testis_replicate1_FR.bam

samtools sort -o testis_replicate2_FR_sorted.bam testis_replicate2_FR.bam

samtools sort -o ovary_replicate1_FR_sorted.bam ovary_replicate1_FR.bam

samtools sort -o ovary_replicate2_FR_sorted.bam ovary_replicate2_FR.bam

3. Quantify read counts per genes and transcripts for both testis and ovary samples using StringTie. Make sure to specify the strand specificity option (FR or RF) that you chose in Task 1. Use the -G option to provide the genome annotation file to flag known genes or transcripts.

For transcript abundance,

stringtie testis_replicate1_FR_sorted.bam -G dmel-all-r6.53.gtf -o testis_replicate1_FR.gtf -e -A transcript_abundance_testis_replicate1.tab

stringtie testis_replicate2_FR_sorted.bam -G dmel-all-r6.53.gtf -o testis_replicate2_FR.gtf -e -A transcript_abundance_testis_replicate2.tab

stringtie ovary_replicate1_FR_sorted.bam -G dmel-all-r6.53.gtf -o ovary_replicate1_FR.gtf -e -A transcript_abundance_ovary_replicate1.tab

stringtie ovary_replicate2_FR_sorted.bam -G dmel-all-r6.53.gtf -o ovary_replicate2_FR.gtf -e -A transcript_abundance_ovary_replicate2.tab

4. For filtering the genes from the output gtf,

awk '$3 == "transcript"' testis_replicate1_FR.gtf > filtered_testis_replicate1_FR.gtf

awk '$3 == "transcript"' testis_replicate2_FR.gtf > filtered_testis_replicate2_FR.gtf

awk '$3 == "transcript"' ovary_replicate1_FR.gtf > filtered_ovary_replicate1_FR.gtf

awk '$3 == "transcript"' ovary_replicate2_FR.gtf > filtered_ovary_replicate2_FR.gtf

For gene abundance,

```
awk -F"\t" 'BEGIN { OFS = "\t" } {
split($9, attrs, /;/);
transcript_id = gensub(/.*transcript_id "([^;]+)".*/, "\\1", "g", $9);
transcript_name = "-";
reference = $1;
strand = $7;
start = $4;
```

end = $5;

coverage = gensub(/.*cov “([^;]+)”.*/, “\\1”, “g”, $9);

fpkm = gensub(/.*FPKM “([^;]+)”.*/, “\\1”, “g”, $9);

tpm = gensub(/.*TPM “([^;]+)”.*/, “\\1”, “g”, $9);

print transcript_id, transcript_name, reference, strand, start, end, coverage, fpkm, tpm;

}’ filtered_testis_replicate1_FR.gtf > gene_abundance_testis_replicate1.tab

Repeat this process for other samples too.

Genes and Transcripts of high fold-change

GTF files of genes and transcripts were imported into R, where the remaining calculations were done. Replicate files were merged and average read counts for each gene in the ovary and testis were calculated. These were used to calculate log fold change in expression as follows:

$$r_g = \log_2 \frac{1 + \bar{g}_{\text{ovary}}}{1 + \bar{g}_{\text{testis}}}$$

The results are displayed as heatmaps to visualize the log fold change in genes and transcripts between the two tissues.

Discussion

Results

Focusing first on the genes of highest and lowest fold change, we observe that a majority of the high-fold change genes in the ovary are ribosomal, and the other two are involved in the vitelline membrane (26Aa) and gonad formation (14-3-3). The lowest fold-change genes in the ovary were typically related to developmental processes of sensory organs, and one was a long non-coding RNA.

In the testis tissues, the highest-fold change (HFC) genes are involved in mtDNA replication, proteolysis, and muscle glutamate receptors. The lowest fold-change (LFC) genes show another ribosomal protien, and a few genes associated with sexual reproduction (FBgn0036091) and testis (FBgn0039104).

Next, the HFC transcripts in the ovary are nearly all associated with protien biosynthesis and include ribosomal proteins and a molecular chaperone. Most of the LFC transcripts were either non-coding RNAs or the biological process it was involved in was unknown.

In the testis tissue, the HFC transcripts function in chromatin remodeling, neuromuscular junctions expansion, and form an RNA polymerase III subunit. A majority of the LFC transcripts are involved in sexual reproduction.

When he check to see if top transcripts are located in the same region of the genome where the top genes are, we found that 3 of the top 5 ovary transcripts with the HFC did. In the testis, only one of the LFC transcripts mapped to the same region of the genome.

Significance of Results

I thought it was very interesting that a majority of the top genes and transcripts did not seem to be sex-specific, but mostly comprised of transcription factors, ribosome components, and developmental processes. In the testis tissue, genes/transcripts that are involved in sexual reproduction fall into the LFC category, which seems counterintuitive. However, transcript expression relies heavily on the temporal dimension, so it may be possible that these transcripts do not become abundant until the fly actively engages in sexual reproduction. We can conclude however, that many of the components involved in transcription are highly upregulated in the reproductive tissues.

Challenges

One of the challenges of working with the data was using the Linux command prompts to run some of the programs, as most of our previous experience is with Python and R programming languages. Unfortunately, this presented a major barrier to EF, as they were unable to run HISAT2.

Future Work

Since this study focuses on gene expression in the reproductive tissues of male and female fruit flies, it may be interesting to analyze other tissue samples to determine if those differences extend beyond the testis and ovaries. In EF's own research on cancer therapies, we observe a sex bias in particular tumor models, both in the growth rate of tumors and tumor response to therapy. Since the fruit fly is a popular model for human disease, it may be important to determine if there is any differential expression of genes involved in the immune response as it may play a role in the efficacy of treatment.

Conclusions

The ability to analyze large data sets is critical to the continued progression of scientific research as the technology to generate the data improves. In this study, we were able to obtain not only the read counts for genes and transcripts in the ovary and testis tissues, but

also go a step further and determine the genes/transcripts with the highest and lowest fold changes between the different samples.

Our work was able to determine that transcripts that produce different components of transcription or aid the process were found to have the highest fold change in both tissues. The testis samples interestingly also contain some factors associated with muscle tissue, Myocardin related transcription factor (Mrtf) and metro (neuromuscular junctions).

The significance of this work is the ability to extract biological information from thousands of short RNA sequence reads. These types of studies play an important role in genetic research, greatly amplifying the scale at which we can compare biological conditions. In many cases, these studies are used to determine target genes for further study by focusing the research questions.

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 provided some biology background to help shape the code, compiled the Gene/Transcript Function Summary Tables, and wrote the report.

References

1. Stephenson R, Metcalfe NH. Drosophila melanogaster: a fly through its history and current use. J R Coll Physicians Edinb. 2013;43(1):70-5. doi: 10.4997/JRCPE.2013.116. PMID: 23516695.