Untitled

3.1 Find out the top five genes with the greatest log fold change rg between ovary and testis and top five with the lowest rg Visualize the expression for these top genes from the raw count data. You can use box plTo calculate the log fold change for each gene, the given formula is:

To calculate the log fold change for each gene, the given formula is: $\mathrm{rg} = \log2((1+\mathrm{govary})/(1+\mathrm{gtestis}))$

```
# import gene_count_matrix.csv
gene_count_matrix <- read.csv("E:/Language/R/gene_count_matrix.csv")

# Calculate the log fold change for each gene
gene_count_matrix$log_fold_change <- log2(
  (1 + (gene_count_matrix$ENCLB117FKX + gene_count_matrix$ENCLB129EAK) / 2) /
  (1 + (gene_count_matrix$ENCLB597ZOR + gene_count_matrix$ENCLB925FOQ) / 2)
)

# Get the top and bottom five genes by fold change
top_genes_greatest <- head(gene_count_matrix[order(-gene_count_matrix$log_fold_change), ],
top_genes_lowest <- head(gene_count_matrix[order(gene_count_matrix$log_fold_change), ], 5)
print (top_genes_greatest)
```

```
              gene_id ENCLB117FKX ENCLB129EAK ENCLB597ZOR ENCLB925FOQ
2347  ENCLB117FKX.1238       14894        6837           0           0
24865      FBgn0261987       10408        8632           0           0
22512      FBgn0003015       39983       39419          10           0
7312       FBgn0032876        6275        4315           0           0
29915      FBgn0003028        5134        3799           0           0
      log_fold_change
2347         13.40760
24865        13.21690
22512        12.69196
```

```
7312         12.37069
29915        12.12525
```

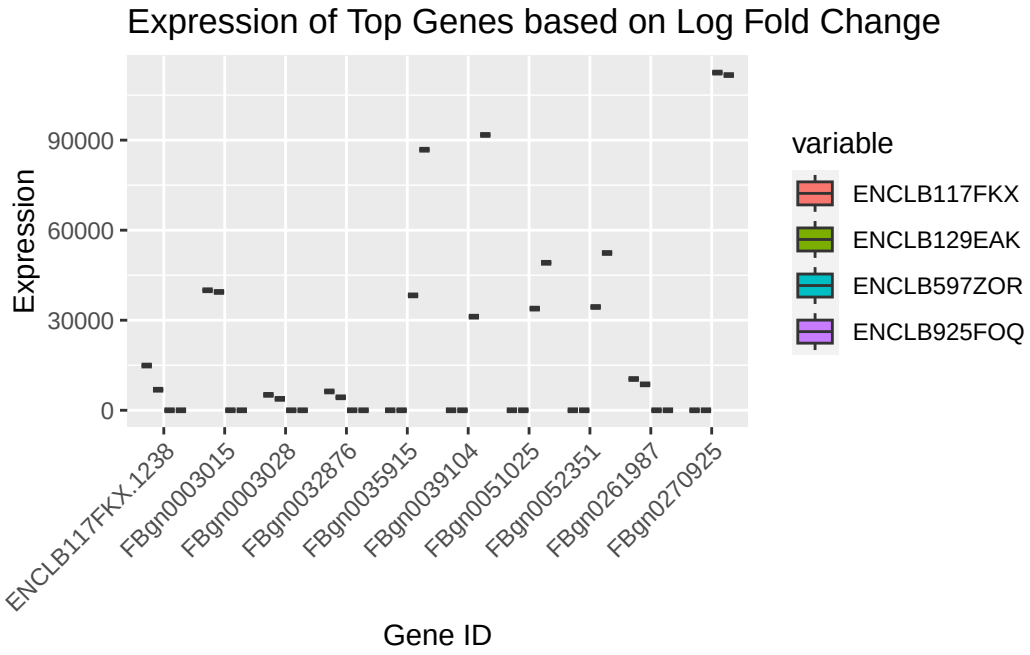
```
print(top_genes_lowest)
```

```
         gene_id ENCLB117FKX ENCLB129EAK ENCLB597ZOR ENCLB925FOQ
25360 FBgn0270925           0           0      112495      111685
17581 FBgn0035915           0           0       38278       86801
26284 FBgn0039104           0           0       31174       91738
17583 FBgn0052351           0           0       34410       52403
28068 FBgn0051025           0           0       33865       49135
      log_fold_change
25360       -16.77431
17581       -15.93250
26284       -15.90729
17583       -15.40566
28068       -15.34086
```

```
# Visualizing these genes using a boxplot
library(ggplot2)

selected_genes <- rbind(top_genes_greatest, top_genes_lowest)
melted_gene_data <- reshape2::melt(selected_genes, id.vars = c("gene_id", "log_fold_change

ggplot(melted_gene_data, aes(x = gene_id, y = value, fill = variable)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Expression of Top Genes based on Log Fold Change", y = "Expression", x = "
```



3.2 Repeat 3.1 on the transcript read counts.

```
# Import transcript_count_matrix.csv
transcript_count_matrix <- read.csv("E:/Language/R/transcript_count_matrix.csv")

# Calculate the log fold change for each transcript
transcript_count_matrix$log_fold_change <- log2(
  (1 + (transcript_count_matrix$ENCLB117FKX + transcript_count_matrix$ENCLB129EAK) / 2) /
  (1 + (transcript_count_matrix$ENCLB597ZOR + transcript_count_matrix$ENCLB925FOQ) / 2)
)

# Get the top and bottom five transcripts by fold change
top_transcripts_greatest <- head(transcript_count_matrix[order(-transcript_count_matrix$lo
top_transcripts_lowest <- head(transcript_count_matrix[order(transcript_count_matrix$log_f
print(top_transcripts_greatest)
```

```
      transcript_id ENCLB117FKX ENCLB129EAK ENCLB597ZOR ENCLB925FOQ
7983     FBtr0080804       28882        6467           0           0
34504    FBtr0301632       18293        8572           0           0
21252    FBtr0072258       12735       11030           0           0
26541    FBtr0332636        7600       14962           0           0
```

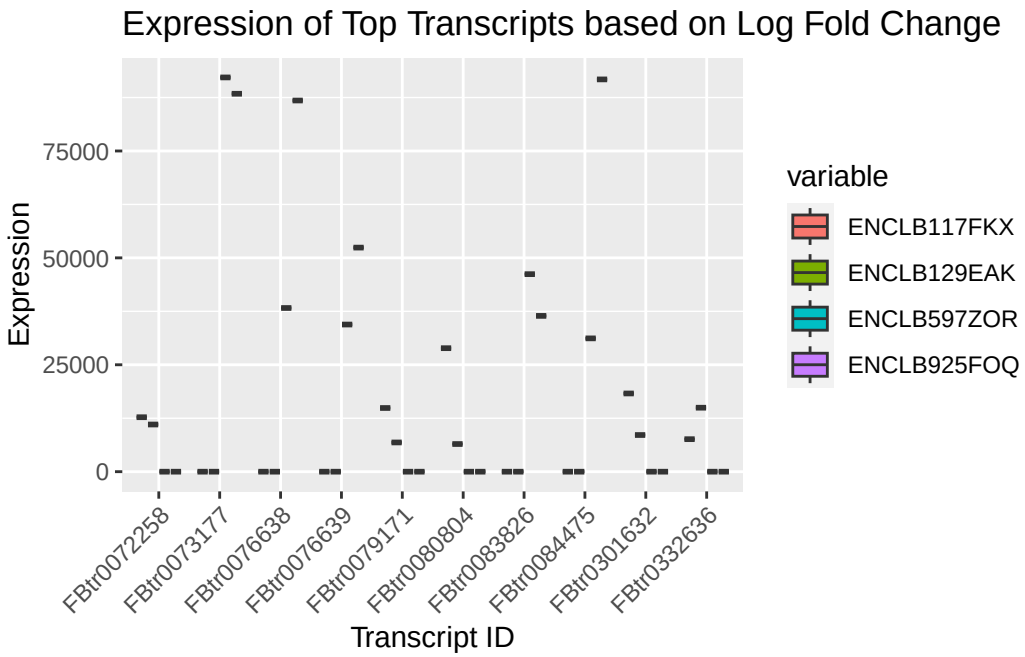
```
3180    FBtr0079171       14894        6837           0           0
      log_fold_change
7983         14.10946
34504        13.71355
21252        13.53667
26541        13.46174
3180         13.40760
```

```r
print(top_transcripts_lowest)
```

```
      transcript_id ENCLB117FKX ENCLB129EAK ENCLB597ZOR ENCLB925FOQ
23273   FBtr0073177           0           0       92195       88381
24790   FBtr0076638           0           0       38278       86801
38002   FBtr0084475           0           0       31174       91738
24792   FBtr0076639           0           0       34410       52403
36563   FBtr0083826           0           0       46193       36432
      log_fold_change
23273       -16.46226
24790       -15.93250
38002       -15.90729
24792       -15.40566
36563       -15.33433
```

```r
# Visualizing these transcripts using a boxplot
selected_transcripts <- rbind(top_transcripts_greatest, top_transcripts_lowest)
melted_transcript_data <- reshape2::melt(selected_transcripts, id.vars = c("transcript_id"

ggplot(melted_transcript_data, aes(x = transcript_id, y = value, fill = variable)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Expression of Top Transcripts based on Log Fold Change", y = "Expression",
```



3.3 Are all top transcripts located within the genomic region of the top genes?

import merged.gtf(this is ENCLB117FKX.gtf ENCLB129EAK.gtf ENCLB597ZOR.gtf ENCLB925FOQ.gtf merged, use stringtie :

stringtie –merge -p 8 -o E:/Language/R/merged.gtf E:/Language/R/mergelist.txt)

```
library(rtracklayer)
```

```
Loading required package: GenomicRanges
```

```
Loading required package: stats4
```

```
Loading required package: BiocGenerics
```

```

Attaching package: 'BiocGenerics'
```

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

    IQR, mad, sd, var, xtabs

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

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
    table, tapply, union, unique, unsplit, which.max, which.min

Loading required package: S4Vectors


Attaching package: 'S4Vectors'

The following object is masked from 'package:utils':

    findMatches

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

    expand.grid, I, unname

Loading required package: IRanges


Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

    windows

Loading required package: GenomeInfoDb
```

```r
gtf_path <- "E:/Language/R/merged.gtf"
merged_gtf <- import(gtf_path)

genes <- merged_gtf[merged_gtf$type == "gene"]
transcripts <- merged_gtf[merged_gtf$type == "transcript"]
```

```r
results <- sapply(top_transcripts_greatest$transcript_id, function(transcript_id) {
  transcript_region <- transcripts[transcripts$transcript_id == transcript_id, ]
  overlapping_genes <- genes[genes$seqnames == transcript_region$seqnames &
                             genes$start <= transcript_region$end &
                             genes$end >= transcript_region$start, ]

  any(overlapping_genes$gene_id %in% top_genes_greatest$gene_id)
})

names(results) <- top_transcripts_greatest$transcript_id
results
```

```
FBtr0080804 FBtr0301632 FBtr0072258 FBtr0332636 FBtr0079171
      FALSE       FALSE       FALSE       FALSE       FALSE
```