Task 1

Pranathi Limmala

2024-03-29

Function to download the GTF file

```
# Function to download the GTF file
downloadGTF <- function(url, destfile) {
  download.file(url, des_file, method="libcurl")
}
```

Data loading and exploration

Mapping transcripts to genes and saving .rds file

```
# Get mappings of transcripts to genes
transcripts_mapped <- transcriptsBy(txdb, by = "gene")

# Compute the number of transcripts for each gene
transcript_counts <- sapply(transcripts_mapped, length)

# Compute mean, minimum, and maximum number of transcripts per gene
mean_transcripts <- mean(transcript_counts)
min_transcripts <- min(transcript_counts)
max_transcripts <- max(transcript_counts)

# Print the statistics
cat("Mean number of transcripts per gene:", mean_transcripts, "\n")
```

```
## Mean number of transcripts per gene: 4.000395
```

```
cat("Minimum number of transcripts per gene:", min_transcripts, "\n")
```

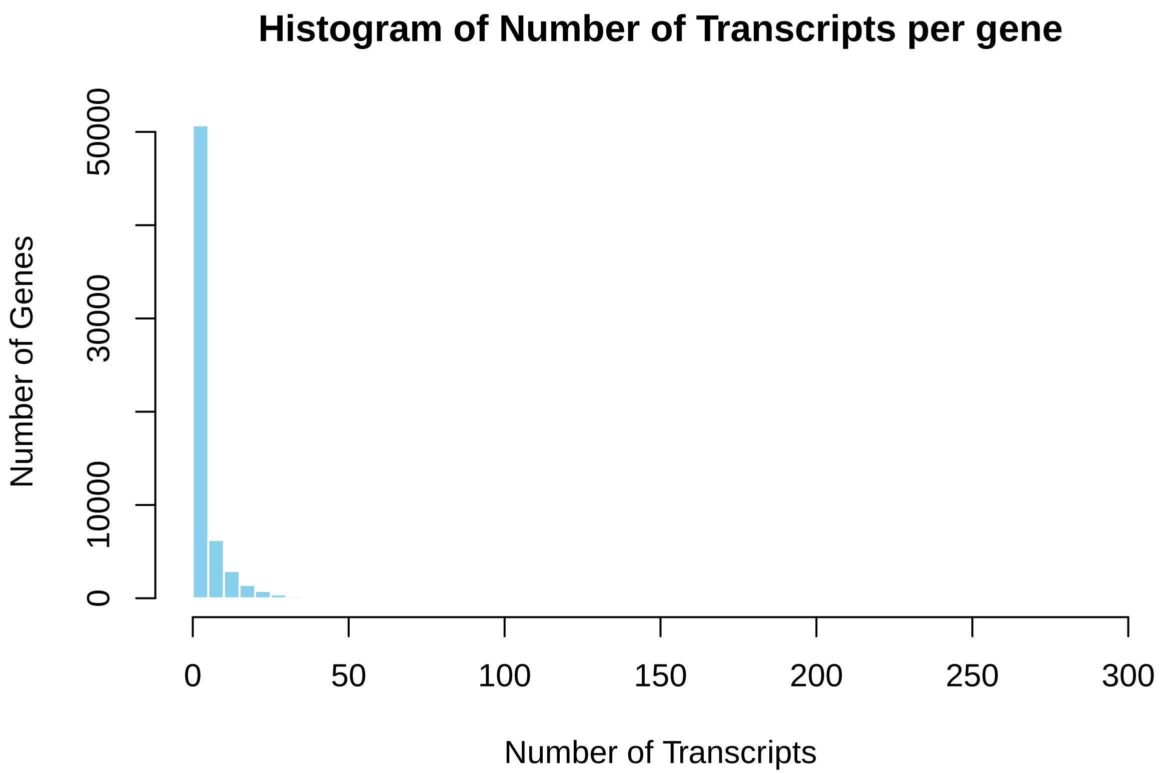
```
## Minimum number of transcripts per gene: 1
```

```
cat("Maximum number of transcripts per gene:", max_transcripts, "\n")
```

```
## Maximum number of transcripts per gene: 296
```

```
# Create the histogram
hist(transcript_counts, breaks = 50, main = "Histogram of Number of Transcripts per gene",
     xlab = "Number of Transcripts", ylab = "Number of Genes", col = "skyblue", border = "white")

# Add labels for clarity
mtext("Distribution of Transcripts Across Genes", side = 3, line = 0.5, outer = TRUE, cex = 1.2)
```



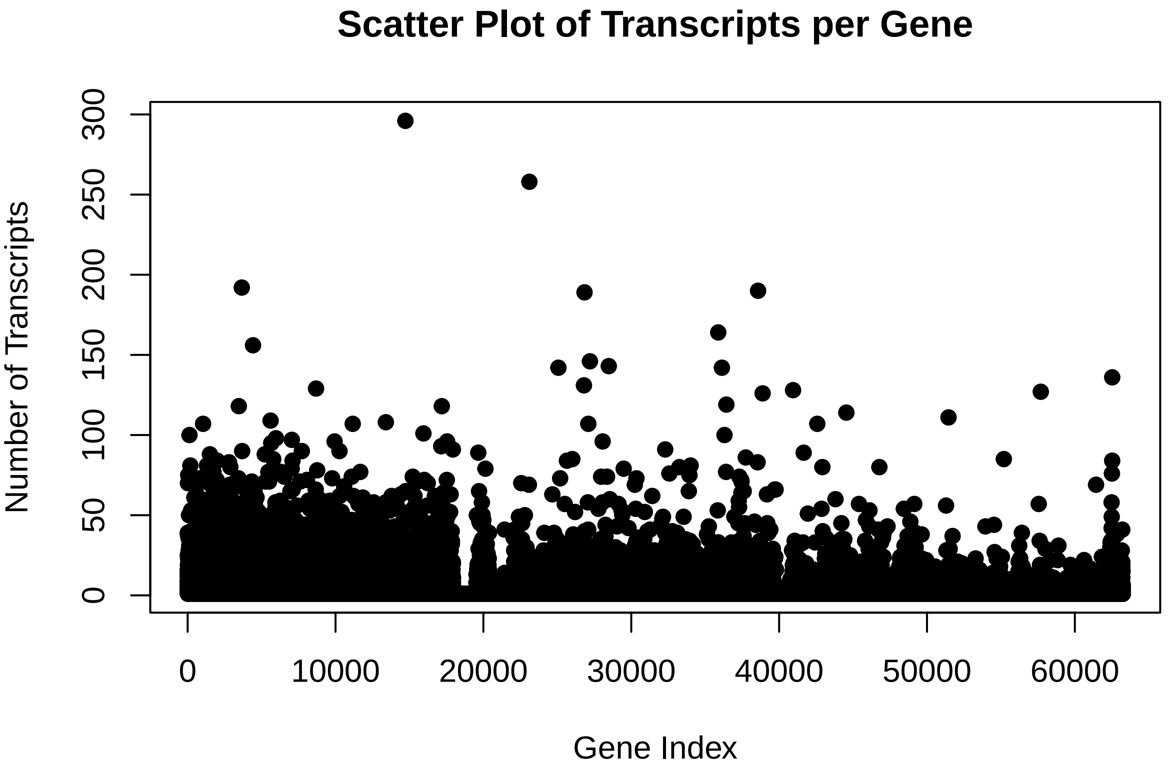
```
# Extract gene information
gene_info <- genes(txdb)

# Get the number of genes
num_genes <- length(gene_info)

# Print the number of genes
cat("Number of genes in the dataset:", num_genes, "\n")
```

```
## Number of genes in the dataset: 63241
```

```
# Create a bar plot (scatter plot is plotted for large datasets generally and hence including here)
if (length(transcript_counts) <= 100) {
  barplot(transcript_counts, main = "Number of Transcripts per gene",
          ylab = "Number of Transcripts", xlab = "Genes",
          cex.names = 0.5, las = 2)
} else {
  # For large datasets, a scatter plot is more practical
  plot(transcript_counts, pch = 19, xlab = "Gene Index", ylab = "Number of Transcripts",
       main = "Scatter Plot of Transcripts per Gene")
}
```



```r
# Save the S4 object to an .rds file
saveRDS(transcript_counts, file = "transcripts_to_genes.rds")
```