C S 509 HW3

Author: Indronil Bhattacharjee

Submitted on: September 19, 2023

==========

Task 1: Removing the version number from an Ensembl ID

```
cut.version.number <- function(ensembl.id) {
  truncated_ids <- gsub("\\.\\d+", "", ensembl.id)
  return(truncated_ids)
}

ensembl.id <- c("ENST00000621592.8", "ENST00000377970.6", "ENST00000259523.10")
truncated_ids <- cut.version.number(ensembl.id)
print(truncated_ids)
```

```
"ENST00000621592" "ENST00000377970" "ENST00000259523"
```

Task 2: Reading genome annotation in GTF format

```
library(dplyr)
library(ggplot2)

# Read the GTF file
gtf_file <- readLines("C:/Users/ibpri/Downloads/gencode.v44.annotation.gtf")
```

Task 2.1: Three differences between GTF and GFF3 formats

1. Field Structure and Order:

GTF: GTF has a fixed and well-defined structure with a specific order of fields. These fields include the sequence name, source, feature type, start position, end position, score, strand, frame, and attribute. The field order is consistent across GTF files.

GFF3: GFF3, on the other hand, is more flexible in terms of field structure and order. It uses column headers to specify the meaning of each column, allowing users to define and include additional attributes as needed. This flexibility makes GFF3 suitable for a wide range of annotations beyond gene-centric data.

2. Attribute Format:

GTF: GTF typically uses a simplified attribute format with predefined attributes like "gene_id" and "transcript_id." These attributes are represented as key-value pairs and are well-suited for gene-related annotations. For example, "gene_id "ENSG12345"; transcript_id "ENST67890";".

GFF3: GFF3 allows for more general attribute representations. It uses a "key=value" format for attributes, which provides greater flexibility. Users can define custom attributes specific to their annotation data. For example, "ID=gene123;Name=MyGene;Note=This is a custom annotation;".

3. Comments and Directives:

GTF: GTF does not have a dedicated structure for comments or directives within the file. It primarily focuses on the feature data.

GFF3: GFF3 includes support for comments and directives. Lines that start with "#" or "##" are used for comments and directives, respectively. Comments can provide additional information, while directives can specify data sources, versions, and other metadata. This makes GFF3 more versatile for including metadata and additional context.

Task 2.2: Extract gene names for every feature in the GTF file using Regular Expression and their runtimes

Method 1:
The words which match the pattern are extracted and printed immediately without storing this.

Method 2:
The words which match the pattern are extracted and stored in a list as a vector., which takes a longer amount of time since adding elements in a list in R creates a list of n+1 each time when something is added at the end of a list with length n.

```r
# Method 1
pattern <- "gene_name \"(.*?)\";"

# Measure the runtime using system.time
runtime1 <- system.time({
    for (line in gtf_file) {
        if (grepl("gene_name", line)) {
            gene_name <- regmatches(line, regexec(pattern, line))[[1]][2]
        }
        print(gene_name)
    }
})
```

```
"DDX11L2"
"DDX11L2"
"DDX11L2"
"DDX11L2"
"DDX11L2"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"DDX11L1"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"WASH7P"
"MIR6859-1"
"MIR6859-1"
"MIR6859-1"
"MIR1302-2HG"
"MIR1302-2HG"
"MIR1302-2HG"
"MIR1302-2HG"
"MIR1302-2HG"
```

"MIR1302-2HG"
"MIR1302-2HG"
"MIR1302-2HG"
"MIR1302-2"
"MIR1302-2"
"MIR1302-2"
"FAM138A"
"FAM138A"
"FAM138A"
"FAM138A"
"FAM138A"
"FAM138A"
"FAM138A"
"FAM138A"
"OR4G4P"
"OR4G4P"
"OR4G4P"
"ENSG00000290826"
"ENSG00000290826"
"ENSG00000290826"
"ENSG00000290826"
"ENSG00000290826"
"OR4G11P"
"OR4G11P"
"OR4G11P"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"OR4F5"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"

"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000238009"
"ENSG00000239945"
"ENSG00000239945"
"ENSG00000239945"
"ENSG00000239945"
"CICP27"
"CICP27"
"CICP27"
"ENSG00000268903"
"ENSG00000268903"
"ENSG00000268903"
"ENSG00000269981"
"ENSG00000269981"
"ENSG00000269981"
"ENSG00000239906"
"ENSG00000239906"
"ENSG00000239906"
"ENSG00000239906"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"

"ENSG00000241860"
"ENSG00000241860"
"ENSG00000241860"
"RNU6-1100P"
"RNU6-1100P"
"RNU6-1100P"
"ENSG00000241599"
"ENSG00000241599"
"ENSG00000241599"
"ENSG00000241599"
"DDX11L17"
"DDX11L17"
"DDX11L17"
"DDX11L17"
"DDX11L17"
"DDX11L17"
"DDX11L17"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"WASH9P"
"MIR6859-2"
"MIR6859-2"
"MIR6859-2"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"

"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000228463"
"ENSG00000286448"
"ENSG00000286448"
"ENSG00000286448"
"ENSG00000286448"
"RPL23AP24"
"RPL23AP24"
"RPL23AP24"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000236601"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"

"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"

"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"ENSG00000290385"
"WBP1LP7"
"WBP1LP7"
"WBP1LP7"
"OR4F29"
"OR4F29"
"OR4F29"
"OR4F29"
"OR4F29"
"OR4F29"
"OR4F29"
"ENSG00000237094"
"ENSG00000237094"
"ENSG00000237094"
"CICP7"
"CICP7"
"CICP7"
"CICP7"
"ENSG00000250575"
"ENSG00000250575"
"ENSG00000250575"
"ENSG00000250575"
"U6"
"U6"
"U6"

"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"

"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000230021"
"ENSG00000235146"
"ENSG00000235146"
"ENSG00000235146"
"ENSG00000235146"
"ENSG00000235146"
"ENSG00000235146"
"ENSG00000235146"
"MTND1P23"
"MTND1P23"
"MTND1P23"
"MTND2P28"
"MTND2P28"
"MTND2P28"
"MTCO1P12"
"MTCO1P12"
"MTCO1P12"
"ENSG00000278791"
"ENSG00000278791"
"ENSG00000278791"
"MTCO2P12"
"MTCO2P12"
"MTCO2P12"
"MTATP8P1"
"MTATP8P1"
"MTATP8P1"
"MTATP6P1"
"MTATP6P1"

"MTATP6P1"
"MTCO3P12"
"MTCO3P12"
"MTCO3P12"
"WBP1LP6"
"WBP1LP6"
"WBP1LP6"
"OR4F16"
"OR4F16"
"OR4F16"
"OR4F16"
"OR4F16"
"OR4F16"
"OR4F16"
"CICP3"
"CICP3"
"CICP3"
"CICP3"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"ENSG00000291215"
"RNU6-1199P"
"RNU6-1199P"
"RNU6-1199P"
"ENSG00000229905"
"ENSG00000229905"
"ENSG00000229905"
"ENSG00000229905"
"ENSG00000228327"
"ENSG00000228327"
"ENSG00000228327"
"ENSG00000228327"
"ENSG00000228327"
"ENSG00000228327"

```
"ENSG00000228327"
"LINC01409"
"LINC01409"
"LINC01409"
"LINC01409"
"LINC01409"
"LINC01409"

[Output truncated]
```

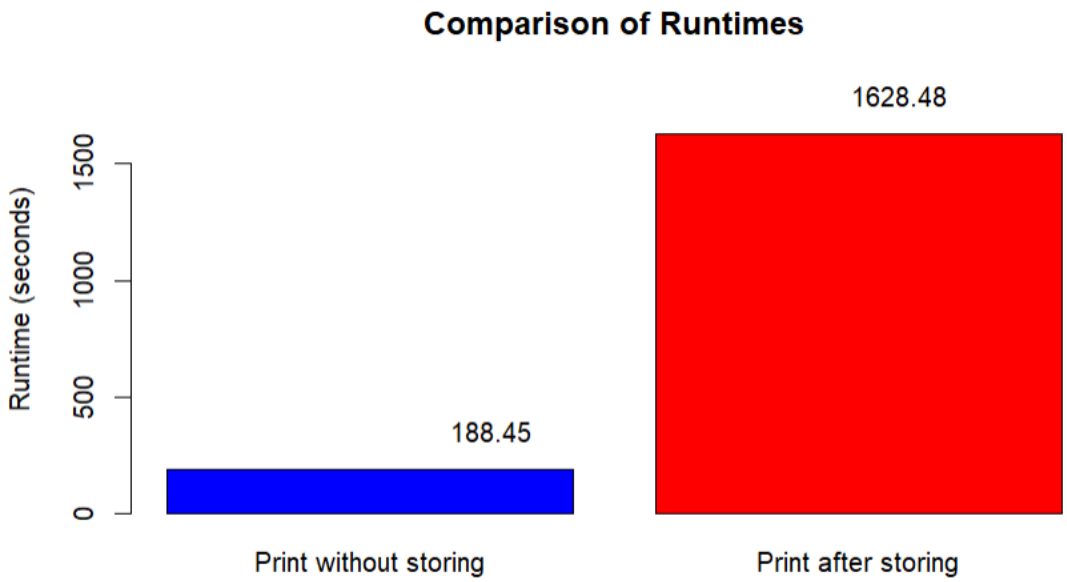
```
print(paste("Runtime:", runtime1[1], "seconds"))
```

```
"Runtime: 188.45 seconds"
```

```
# Method 2
gene_names <- character(0)
pattern <- "gene_name \"(.*?)\";"

# Measure the runtime using system.time
runtime2 <- system.time({
    for (line in gtf_file) {
        if (grepl("gene_name", line)) {
            gene_name <- regmatches(line, regexec(pattern, line))[[1]][2]
    }
    gene_names <- c(gene_names, gene_name)
  }
})

# Print the extracted gene names
# print(gene_names) [Print escaped for very long output]
print(paste("Runtime:", runtime2[1], "seconds"))
```

```
"Runtime: 1628.48 seconds"
```

```r
# Create a vector of runtimes
runtimes <- c(runtime1, runtime2)

# Create a bar chart
barplot(runtimes,
        names.arg = c("Print without storing", "Print after storing"),
        col = c("blue", "red"),  # Colors for the bars
        ylab = "Runtime (seconds)",  # Label for the y-axis
        main = "Comparison of Runtimes",  # Title of the plot
        ylim = c(0, max(runtimes) + 200))  # Adjust the y-axis limits

# Add values on top of the bars
text(x = 1:2, y = runtimes + 50, labels = runtimes, pos = 3, col = "black")
```



References

https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html