Problem 2

Maryam Gholampour

2023-12-11

```
#list out the top five genes that are mutated in various human disease.

download.file("https://github.com/cb2edu/CB2-101-2023-assignment/raw/main/data/humsavar.tsv.gz", destfil

data <- read.table(gzfile("humsavar.tsv.gz"), header = TRUE, sep = "\t", quote = "", comment.char = "",

colnames(data) <- c("gene_names", "swiss_prot_ac", "ftid", "aa_changes", "variant", "dbsnp")

gene_names <- data$gene_names

gene_counts <- table(gene_names)

gene_counts_df <- as.data.frame(table(gene_names))

gene_names <- data.frame(Gene = character(), Mutation_Count = numeric())

colnames(gene_counts_df) <- c("Gene", "Mutation_Count")

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
```

```
gene_counts_df <- arrange(gene_counts_df, desc(gene_counts_df$Mutation_Count))
top_five_genes <- head(gene_counts_df, 5)
print(top_five_genes
      )
```

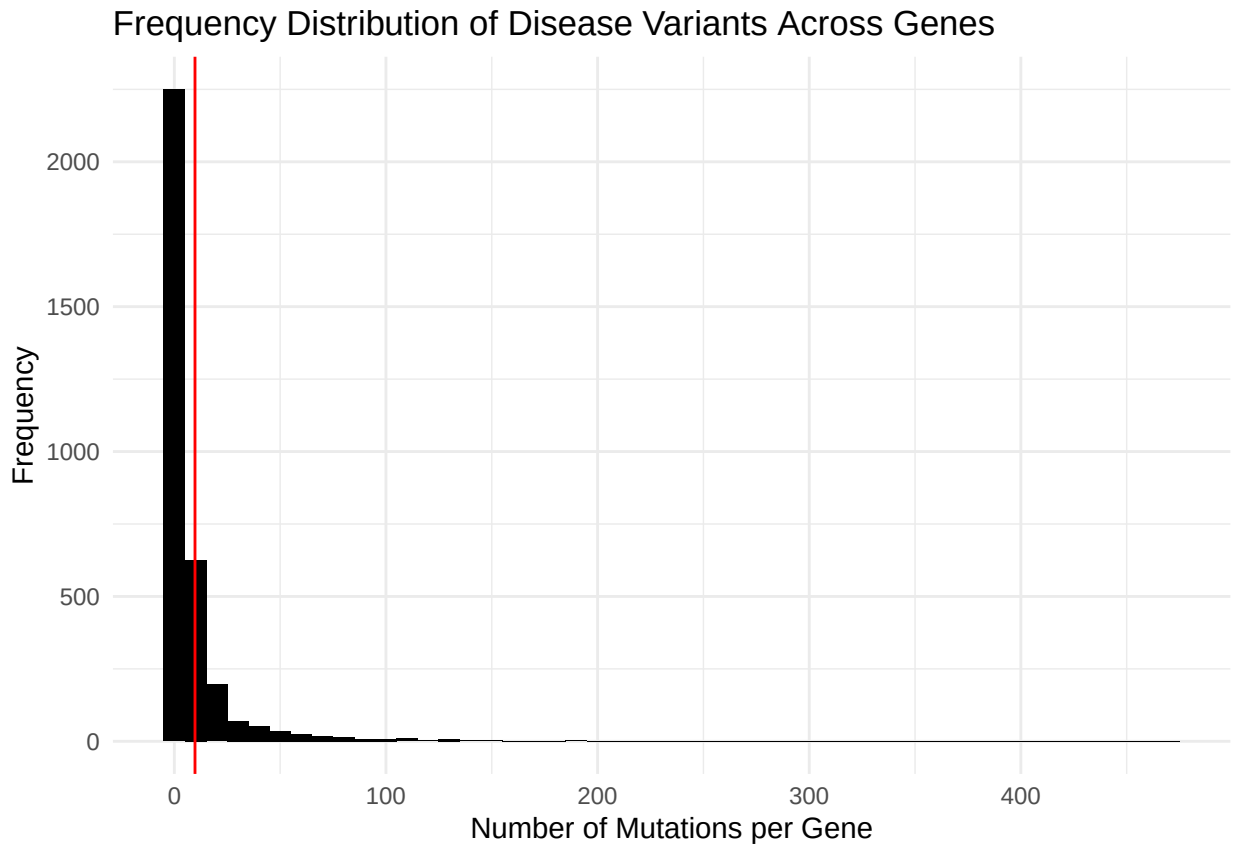
```
##    Gene Mutation_Count
## 1  TP53           1338
## 2    F8            477
## 3 SCN5A            459
## 4 SCN1A            437
## 5  FBN1            414
```

```r
#plot the frequency distribution of disease variants in human genome across all the genes in the file.
#calculate the average number disease causing mutations across all genes in human genome and mark this
```

```r
library(ggplot2)

data <- read.table(gzfile("humsavar.tsv.gz"), header = TRUE, sep = "\t", quote = "", comment.char = "",

colnames(data) <- c("gene_names", "swiss_prot_ac", "ftid", "aa_changes", "variant", "dbsnp")

disease_variant <- data %>%
  filter(variant %in% "LP/P") %>%
  group_by(gene_names, swiss_prot_ac) %>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'gene_names'. You can override using the
## `.groups` argument.
```

```r
ggplot(disease_variant, aes(x = n)) +
  geom_histogram(binwidth = 10, fill = "black") +
  geom_vline(xintercept = mean(disease_variant$n), color = "red") +
  labs(x = "Number of Mutations per Gene", y = "Frequency", title = "Frequency Distribution of Disease V
  theme_minimal()
```



```r
#Plot a graph showing the fraction of mutations affecting each 20 amino acid on the x-axis. Which amino

amino_acid <- sub(".*p\\.([A-Za-z]{3})\\d+.*", "\\1", data$aa_change)

aa_counts <- table(amino_acid)
aa_probabilities <- as.data.frame(table(amino_acid))
aa_probabilities$freq_new <- aa_counts / sum(aa_counts)

library(ggplot2)

ggplot(aa_probabilities, aes(x = amino_acid, y = freq_new)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Fraction of Mutations Affecting Each Amino Acid",
       x = "Amino Acid", y = "Frequency") +
  theme_minimal()
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```