Project 1: Analysis of a yeast genome using R

Probability and statistics for modelling 1

Gautham Ganesh

14 October 2021

This report and a separate R file can be found on the public Github repository https://github.com/gsgautham98/probastat1

2. Creating and setting the working directory:

```
dir.home <- Sys.getenv("HOME")
dir.results <- file.path(dir.home, "Documents", "cmb", "stat1",
    "tp1")
dir.create(path = dir.results, showWarnings = FALSE, recursive = TRUE)
```

4. Downloading the genome file:

```
# setwd(dir.results)
download.file("http://ftp.ensembl.org/pub/release-104/gtf/saccharomyces_cerevisiae/Saccharomyces_cerevis
    "genome.gtf.gz")
download.file("https://raw.githubusercontent.com/gsgautham98/probastat1/main/chrom_sizes.tsv",
    "chrom_sizes.tsv")
download.file("https://raw.githubusercontent.com/gsgautham98/probastat1/main/codon_frequencies.tab",
    "codon_frequencies.tab")
```

5. Loading the data:

```
feature.table <- read.table("genome.gtf", comment.char = "#",
    sep = "\t", header = FALSE, row.names = NULL)
names(feature.table) <- c("seqname", "source", "feature", "start",
    "end", "score", "strand", "frame", "attribute")
```

6. Computing the length of coding genes:

```
feature.table$length <- feature.table$end - feature.table$start
knitr::kable(table(feature.table$feature), col.names = c("Feature",
    "Freq"))
```

Feature	Freq
CDS	6913

Feature	Freq
exon	7507
five_prime_utr	4
gene	7127
start_codon	6601
stop_codon	6600
transcript	7127

```
cds <- subset(feature.table, feature == "CDS")
cds.count <- table(cds$seqname)
knitr::kable(cds.count, col.names = c("Chromosome", "CDS Freq"))
```

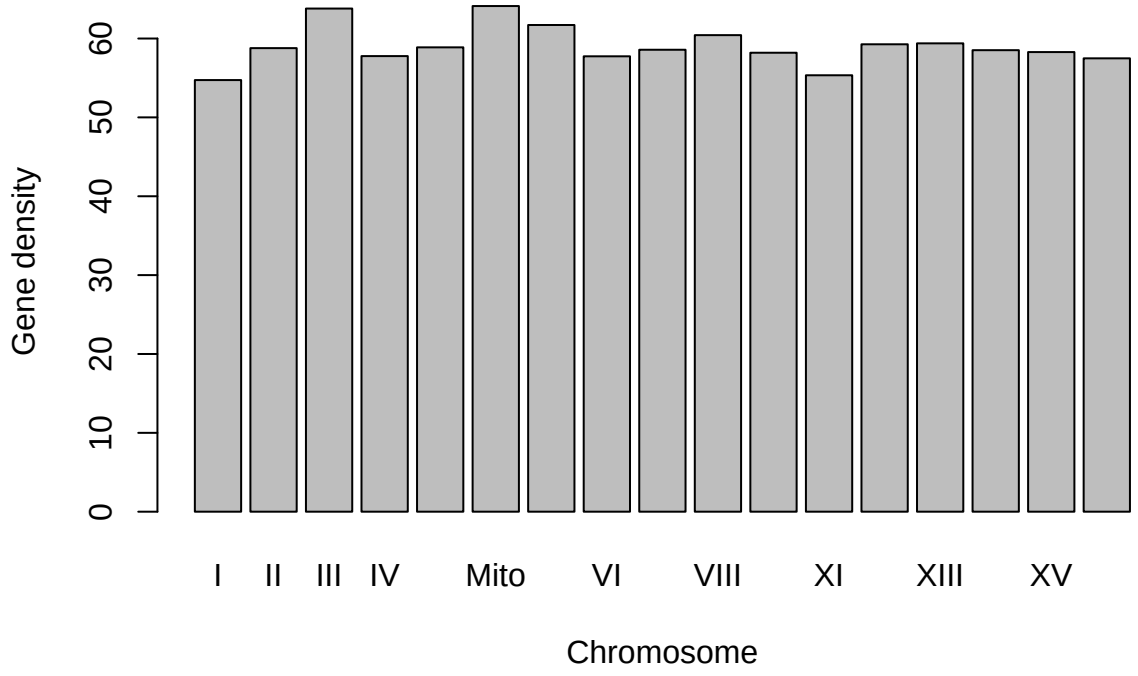
Chromosome	CDS Freq
I	120
II	483
III	192
IV	870
IX	252
Mito	59
V	338
VI	146
VII	605
VIII	340
X	412
XI	361
XII	604
XIII	531
XIV	454
XV	609
XVI	537

```
chromosomes <- read.delim("chrom_sizes.tsv", sep = "\t")
chromosomes <- chromosomes[order(chromosomes$chrom), ]
genes <- subset(feature.table, feature == "gene")
genes.count <- table(genes$seqname)
knitr::kable(genes.count, col.names = c("Chromosome", "Gene Freq"))
```

Chromosome	Gene Freq
I	126
II	478
III	202
IV	885
IX	259
Mito	55
V	356
VI	156
VII	639
VIII	340
X	434

Chromosome	Gene Freq
XI	369
XII	639
XIII	549
XIV	459
XV	636
XVI	545

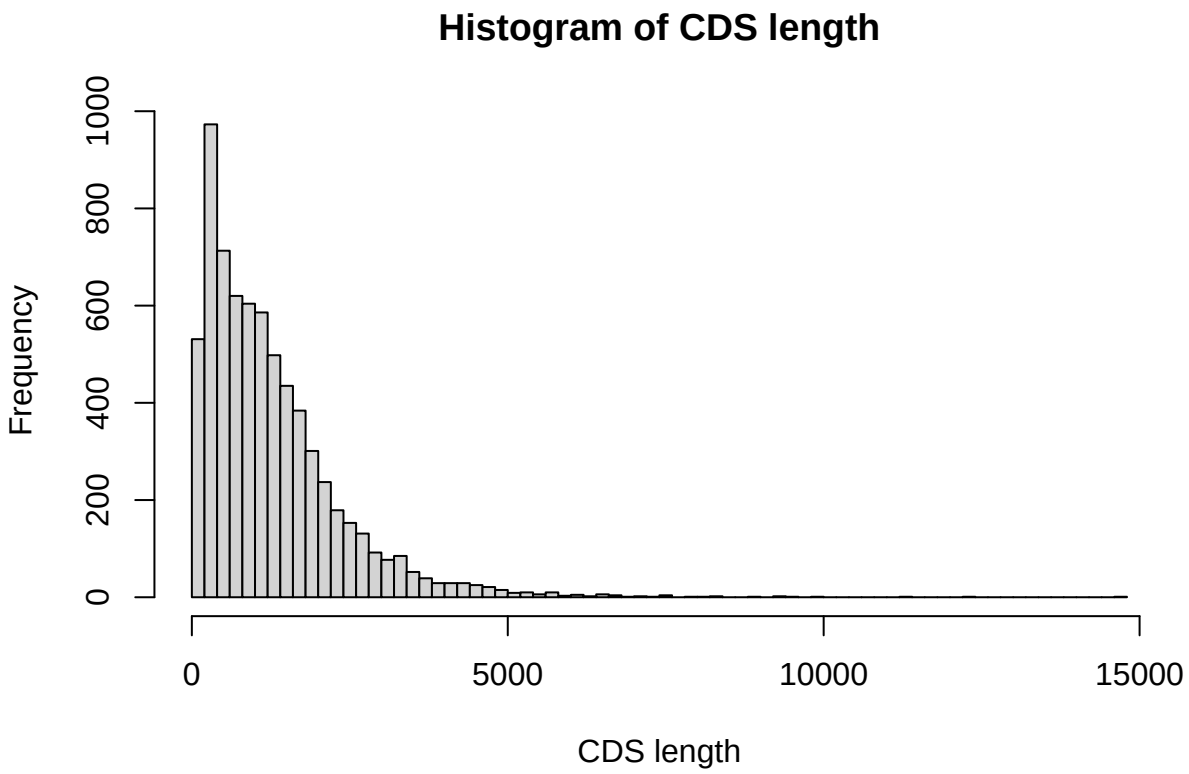
```
genes.per.mb <- genes.count/(chromosomes$size * 1e-05)
barplot(genes.per.mb, xlab = "Chromosome", ylab = "Gene density")
```



Here, an interesting observation is the nearly-uniform distribution of genes among these chromosomes in yeast, which is in contrast to some other more complex eukaryotes (such as humans). Surprisingly, even the mito-genome shares a similar gene density to that of other chromosomes despite its distinct properties and functions.

7. Histogram of gene length:

```
cds.length.hist <- hist(cds$length, breaks = 74, xlab = "CDS length",
    main = "Histogram of CDS length")
```

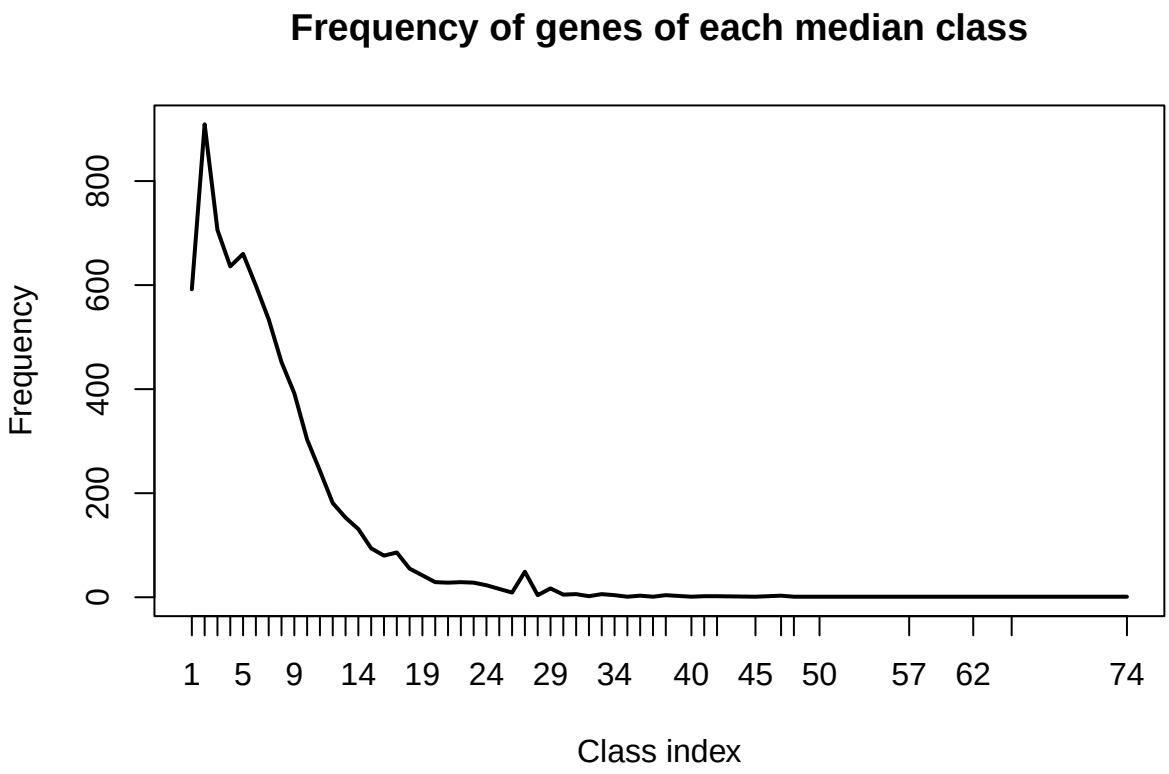



```
print(cds.length.hist)
```

```
## $breaks
##  [1]     0   200   400   600   800  1000  1200  1400  1600  1800  2000  2200
## [13]  2400  2600  2800  3000  3200  3400  3600  3800  4000  4200  4400  4600
## [25]  4800  5000  5200  5400  5600  5800  6000  6200  6400  6600  6800  7000
## [37]  7200  7400  7600  7800  8000  8200  8400  8600  8800  9000  9200  9400
## [49]  9600  9800 10000 10200 10400 10600 10800 11000 11200 11400 11600 11800
## [61] 12000 12200 12400 12600 12800 13000 13200 13400 13600 13800 14000 14200
## [73] 14400 14600 14800
##
## $counts
##  [1] 531 973 713 620 604 586 498 435 384 301 237 179 153 131  92  77  85  52  39
## [20]  29  29  29  25  21  15   9  10   6  10   3   5   2   6   4   1   2   1   4
## [39]   0   1   1   2   0   0   1   0   2   1   0   1   0   0   0   0   0   0   1
## [58]   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   1
##
## $density
##  [1] 3.840590e-04 7.037466e-04 5.156951e-04 4.484305e-04 4.368581e-04
##  [6] 4.238391e-04 3.601909e-04 3.146246e-04 2.777376e-04 2.177058e-04
## [11] 1.714162e-04 1.294662e-04 1.106611e-04 9.474902e-05 6.654130e-05
## [16] 5.569217e-05 6.147837e-05 3.761030e-05 2.820772e-05 2.097497e-05
## [21] 2.097497e-05 2.097497e-05 1.808187e-05 1.518877e-05 1.084912e-05
## [26] 6.509475e-06 7.232750e-06 4.339650e-06 7.232750e-06 2.169825e-06
## [31] 3.616375e-06 1.446550e-06 4.339650e-06 2.893100e-06 7.232750e-07
```

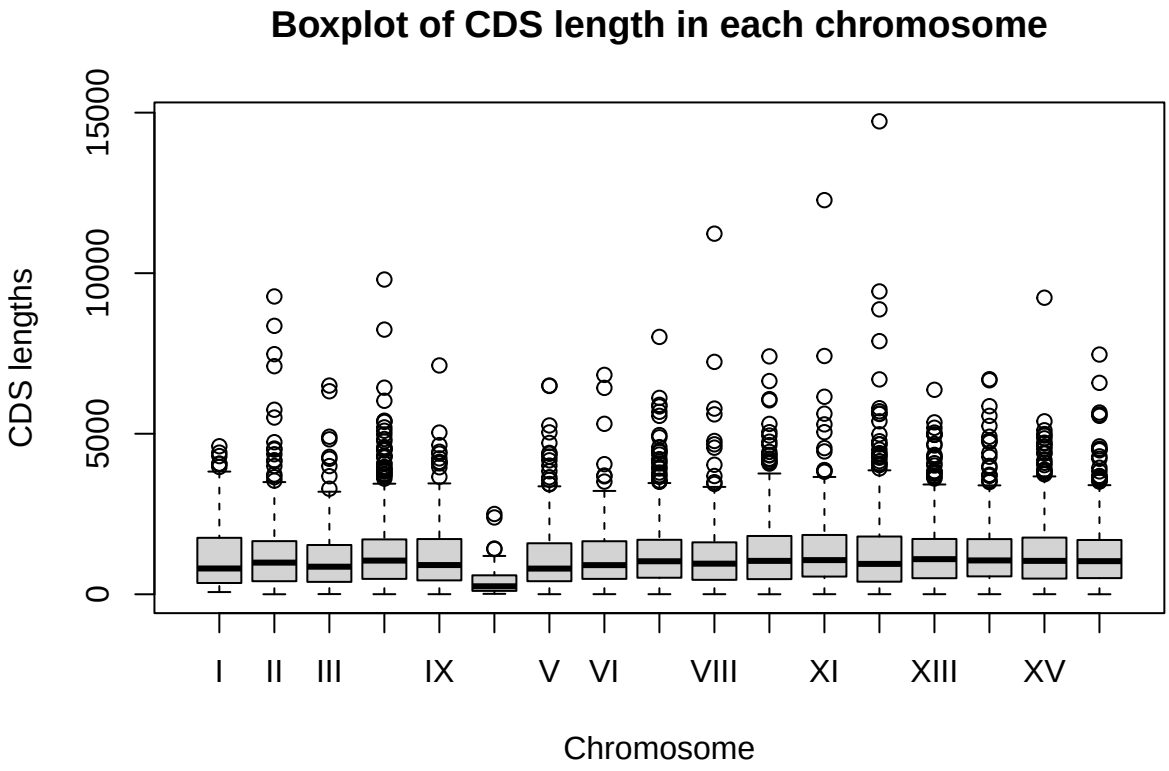
```
## [36] 1.446550e-06 7.232750e-07 2.893100e-06 0.000000e+00 7.232750e-07
## [41] 7.232750e-07 1.446550e-06 0.000000e+00 0.000000e+00 7.232750e-07
## [46] 0.000000e+00 1.446550e-06 7.232750e-07 0.000000e+00 7.232750e-07
## [51] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## [56] 0.000000e+00 7.232750e-07 0.000000e+00 0.000000e+00 0.000000e+00
## [61] 0.000000e+00 7.232750e-07 0.000000e+00 0.000000e+00 0.000000e+00
## [66] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## [71] 0.000000e+00 0.000000e+00 0.000000e+00 7.232750e-07
##
## $mids
##  [1]   100   300   500   700   900  1100  1300  1500  1700  1900  2100  2300
## [13]  2500  2700  2900  3100  3300  3500  3700  3900  4100  4300  4500  4700
## [25]  4900  5100  5300  5500  5700  5900  6100  6300  6500  6700  6900  7100
## [37]  7300  7500  7700  7900  8100  8300  8500  8700  8900  9100  9300  9500
## [49]  9700  9900 10100 10300 10500 10700 10900 11100 11300 11500 11700 11900
## [61] 12100 12300 12500 12700 12900 13100 13300 13500 13700 13900 14100 14300
## [73] 14500 14700
##
## $xname
## [1] "cds$length"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
genes.length.break <- table(findInterval(genes$length, cds.length.hist$breaks))
plot(genes.length.break, type = "l", xlab = "Class index", ylab = "Frequency",
    main = "Frequency of genes of each median class")
```



The frequency of genes of each median class correctly resembles its CDS counterpart because genes consist of coding sequences and with yeast being a simple eukaryote, non-coding sequences (introns) are often short or even non-existent for most genes.

```
boxplot(cds$length ~ cds$seqname, xlab = "Chromosome", ylab = "CDS lengths",
    main = "Boxplot of CDS length in each chromosome")
```



The boxplot of CDS lengths in each chromosome also indicates the conservation of average CDS length, which could correspond to a decade-old study of the conservation of average gene lengths in eukaryotes (Luo et.al., 2006). In this study, gene length was found to be conserved within the prokaryotic and eukaryotic domains (although they were different from each other), implying a potential link between gene length and organism complexity. This length could be associated with factors such as evolutionary conservation and gene function. Although irrelevant in the case of unicellular eukaryotes such as yeast, gene length is also hypothesized to be associated with stage of gene expression and function in higher organisms. (Magalhaes et. al., 2021) Also, the mito-genome has a contrasting, short average gene length because of its size, although the gene density is similar to that of chromosomes.

8. Descriptive parameters:

```
mode.finder <- function(a) {
    a.unique <- unique(a)
    a.unique[which.max(tabulate(match(a, a.unique)))]
}

chrom3 <- subset(genes, seqname == "III")
chrom3.mean <- mean(chrom3$length)
chrom3.median <- median(chrom3$length)
chrom3.mode <- mode.finder(chrom3$length)
sprintf("For chromosome III gene lengths, mean = %f, median = %f, mode = %f",
    chrom3.mean, chrom3.median, chrom3.mode)
```

```
## [1] "For chromosome III gene lengths, mean = 1143.420792, median = 839.000000, mode = 71.000000"
```

```
chrom3.var <- var(chrom3$length)
chrom3.sd <- sd(chrom3$length) * sqrt((length(chrom3$length) -
    1)/length(chrom3$length))
chrom3.iqr <- IQR(chrom3$length)
sprintf("Variance = %f, std deviation = %f, interquartile range = %f",
    chrom3.var, chrom3.sd, chrom3.iqr)
```

```
## [1] "Variance = 1228376.145436, std deviation = 1105.574545, interquartile range = 1107.250000"
```

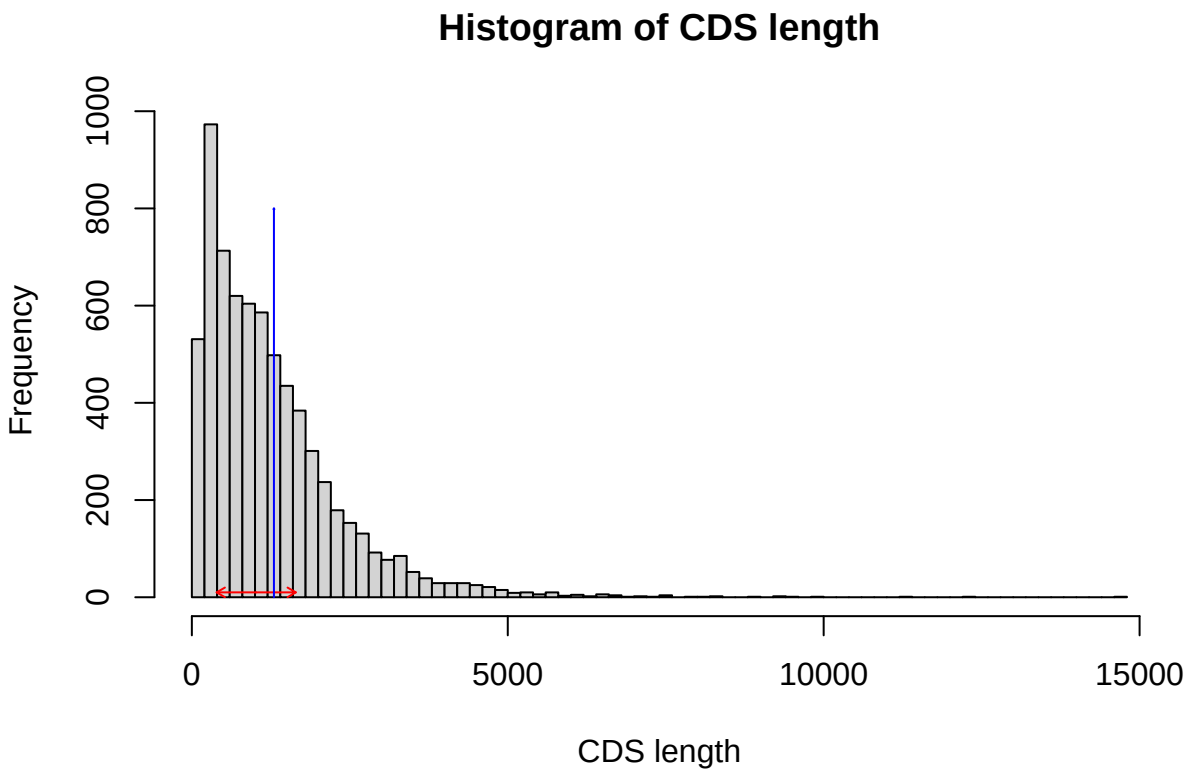
```
genes.mean <- mean(genes$length)
genes.median <- median(genes$length)
genes.mode <- mode.finder(genes$length)
sprintf("For all gene lengths, mean = %f, median = %f, mode = %f",
    genes.mean, genes.median, genes.mode)
```

```
## [1] "For all gene lengths, mean = 1299.310650, median = 1019.000000, mode = 71.000000"
```

```
genes.var <- var(genes$length)
genes.sd <- sd(genes$length)
genes.iqr <- IQR(genes$length)
sprintf("Variance = %f, std deviation = %f, interquartile range = %f",
    genes.var, genes.sd, genes.iqr)
```

```
## [1] "Variance = 1358014.938005, std deviation = 1165.338980, interquartile range = 1254.000000"
```

```
hist(cds$length, breaks = 74, xlab = "CDS length", main = "Histogram of CDS length")
arrows(genes.median, 10, genes.median - (genes.iqr/2), 10, 0.05,
    col = "red")
arrows(genes.median, 10, genes.median + (genes.iqr/2), 10, 0.05,
    col = "red")
arrows(genes.mean, 0, genes.mean, 800, 0.005, col = "blue")
```



The interquartile range is shown in the above histogram using arrows (red) that move away from the median in either direction. The blue arrow indicates the mean.

9. Confidence intervals

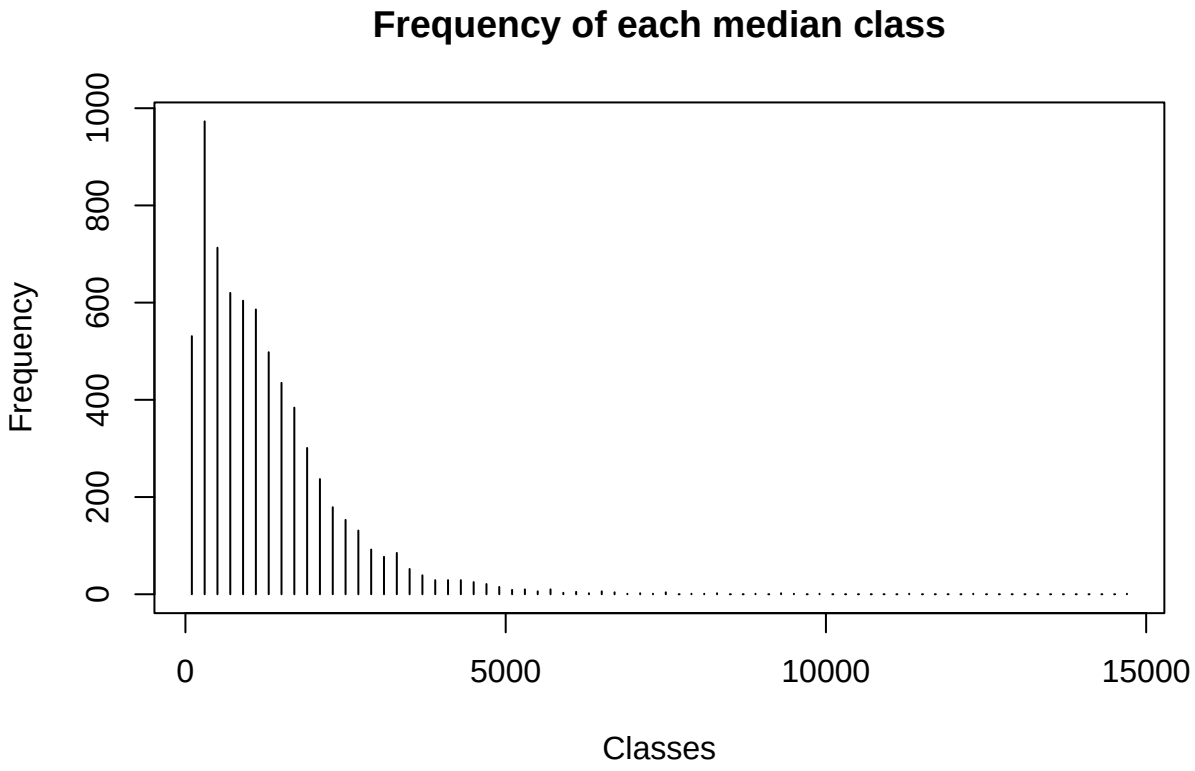
```
n <- length(chrom3$length)
margin <- qt(0.975, df = n - 1) * chrom3.sd/sqrt(n)
lower.end <- chrom3.mean - margin
upper.end <- chrom3.mean + margin
sprintf("The confidence interval for chromosome III around the mean is %f to %f",
    lower.end, upper.end)
```

```
## [1] "The confidence interval for chromosome III around the mean is 990.035665 to 1296.805919"
```

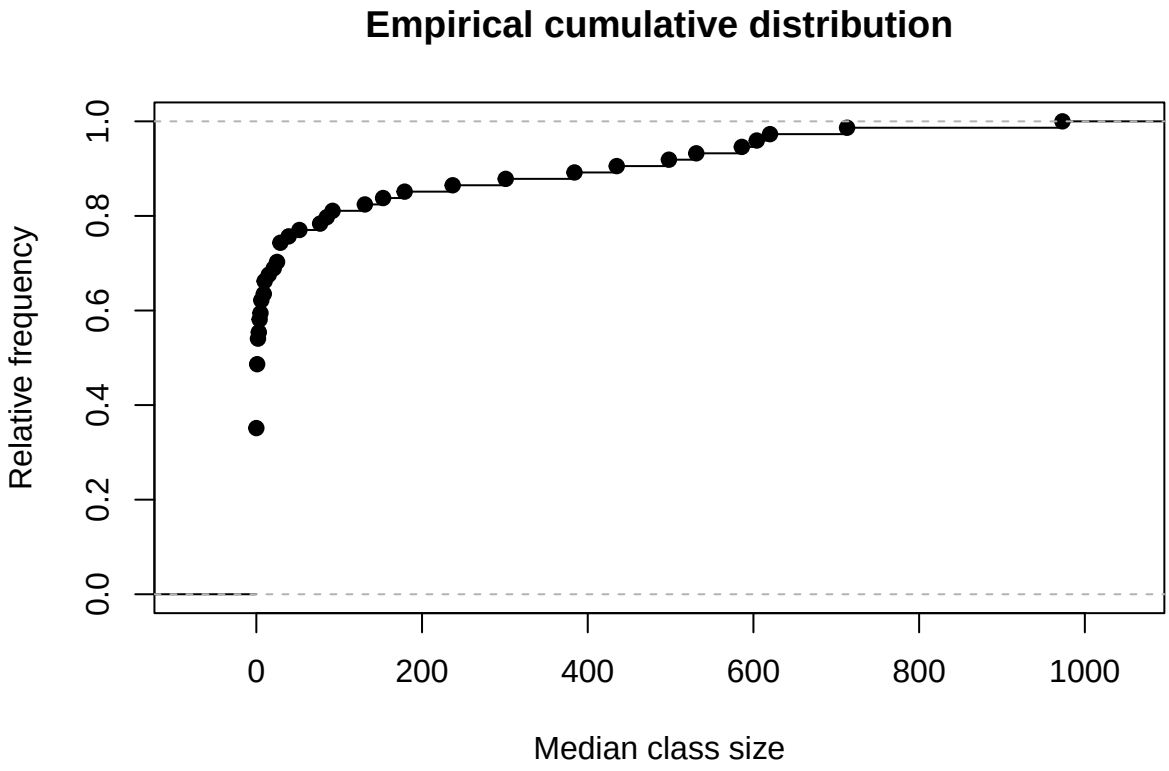
The mean correctly occurs in the middle of this confidence interval.

10. Visualising distribution of gene length

```
lengths.mids <- data.frame(cds.length.hist$mids, cds.length.hist$counts)
names(lengths.mids) <- c("mids", "counts")
relative.freq <- lengths.mids$counts/sum(lengths.mids$counts)
lengths.mids <- cbind(lengths.mids, relative.freq)
lengths.mids <- cbind(lengths.mids, cumsum(lengths.mids$counts))
empirical.freq <- ecdf(lengths.mids$counts)
plot(lengths.mids$mids, lengths.mids$counts, type = "h", xlab = "Classes",
    ylab = "Frequency", main = "Frequency of each median class")
```



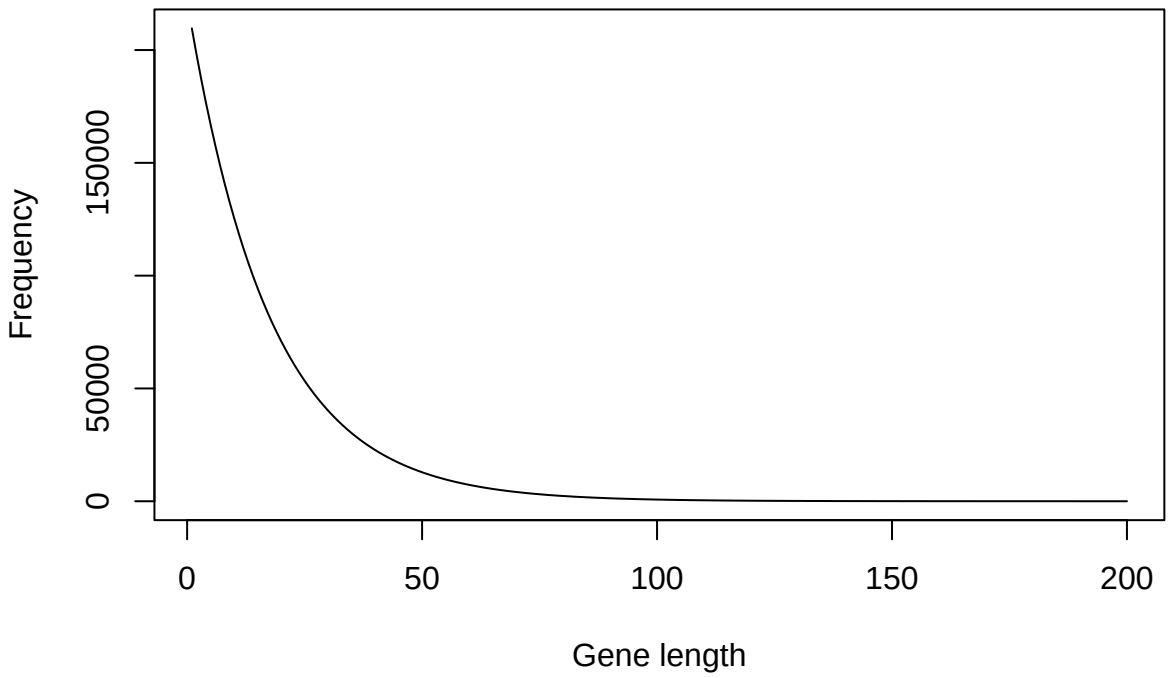

```
plot(empirical.freq, xlab = "Median class size", ylab = "Relative frequency",
    main = "Empirical cumulative distribution")
```



11. Expected distribution of gene lengths

```
prob.3nt <- c(0.0182536851094, 0.0223781252573, 0.0128875284527,
    0.0201230463324)
gene.lengths <- c()
freq.genes <- c()
for (gene.length in 1:200) {
    freq.gene <- 12156679 * (prob.3nt[1] * (1 - sum(prob.3nt[2:4]))^gene.length)
    freq.genes <- append(freq.genes, freq.gene)
    gene.lengths <- append(gene.lengths, gene.length)
}

plot(gene.lengths, freq.genes, type = "l", xlab = "Gene length",
    ylab = "Frequency")
```



This expected distribution shares striking similarities with a plot of the frequency distribution of CDS based on their lengths (Ex. 7) because the length of a coding sequence only depends upon the probabilities of occurrence of its start and stop codons. However, unlike the expected distribution, the peak frequency of a CDS is observed when its length is between 200 and 400 nucleotides.