Assignment 1

BOHTA group 6

May 4, 2017

0. Load the dataset, featuring 4 features:

- Gene name
- mRNA molecule length (base pairs)
- Genome length
- Exon count

```
df <- read.table("gene_lengths_v2.txt", header = T)
```

```
##       name mrna_length genome_length exon_count
## 1   PP8961        2596          2596          1
## 2 FLJ00038         794          2615          6
## 3    OR4F5         918           918          1
## 4    OR4F3         937           937          1
## 5   OR4F16         937           937          1
## 6   SAMD11        2555         18842         14
```

1. Make a histogram that shows what the typical number of exons is. Adjust the bins so that we can pinpoint exactly what number of exons that is the most common. Comment the plot.

Full distribution

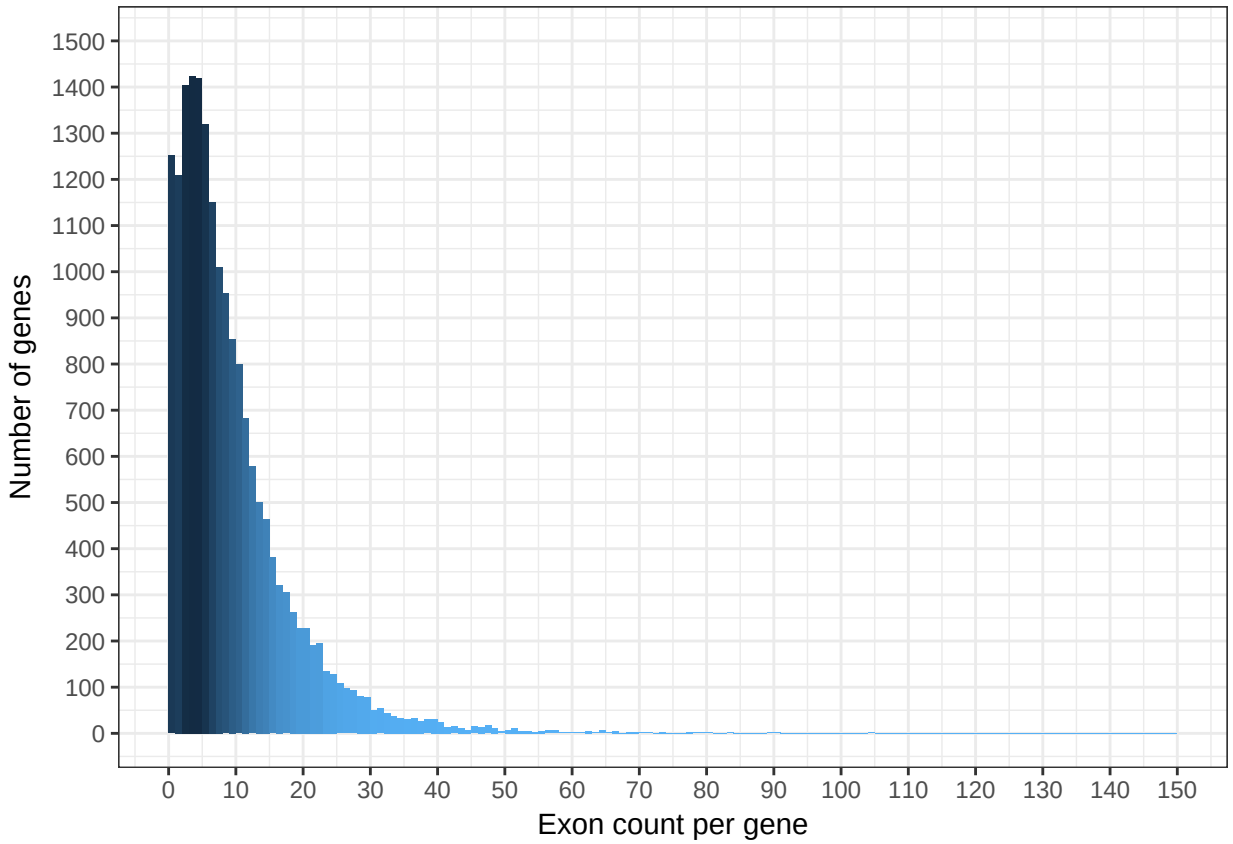


Figure 1 *Histogram showing the distribution of the exon counts. Our data set tends to concentrate between 3 and 5.*

First 20 bins

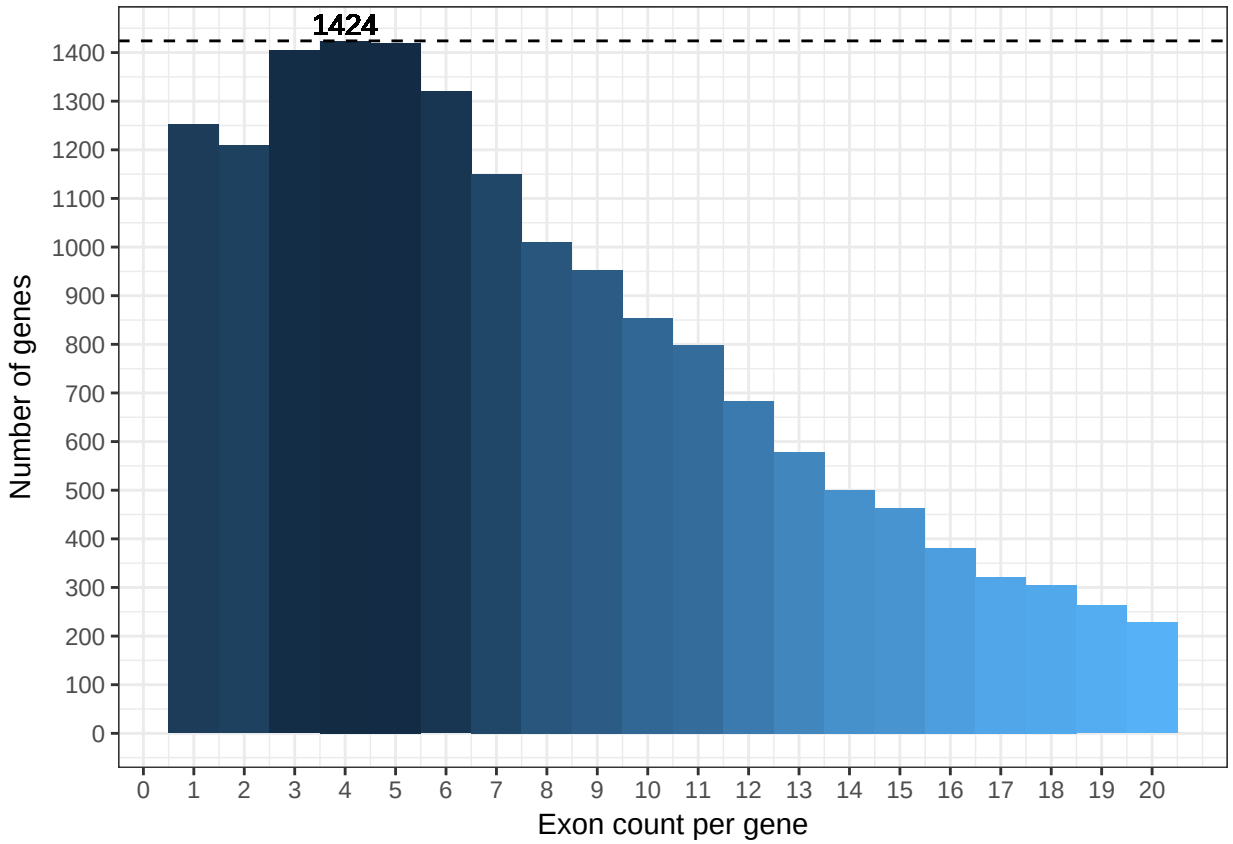


Figure 2 *Inset of the above histogram for exon counts below 20. The mode of the distribution at 4 exons per gene becomes clearer.*

It seems that majority of the genes tend to be formed by a low number of exons. It can be seen that most of the genes (1424) are formed by 4 exons.

2. Add an additional column to the dataframe that contains the total length of introns for each gene

```
df$intron_length <- df$genome_length - df$mrna_length
```

3. Make histograms and boxplots showing the distribution of total exon and total intron lengths, all as subplots in the same larger plot, where each dataset have a different color.

On the histograms, the number of bins should be exactly the same, and the x-axis should have the same scale. Comment the plot – are exons larger than introns or vice versa?

```
## No id variables; using all as measure variables
```

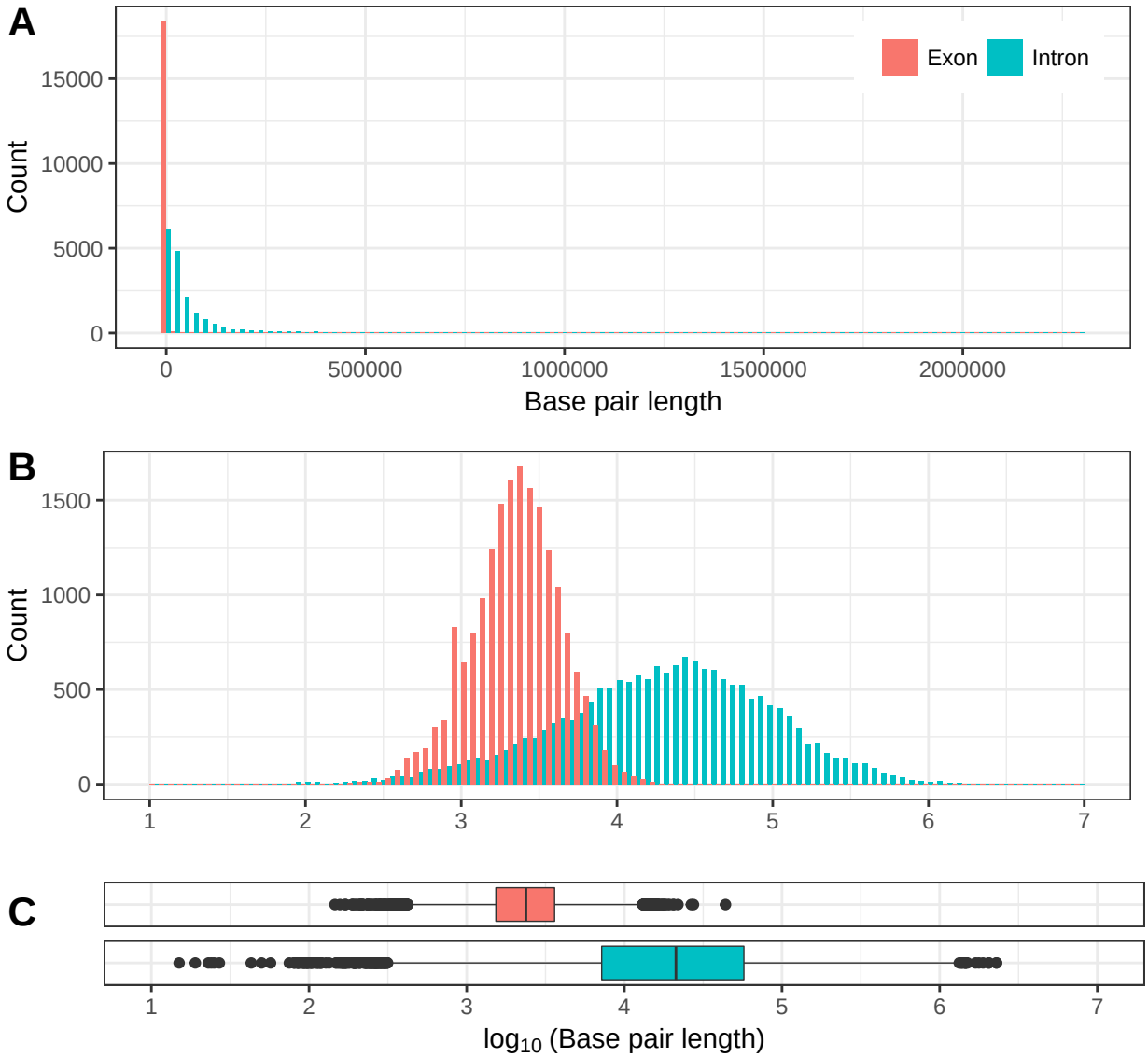
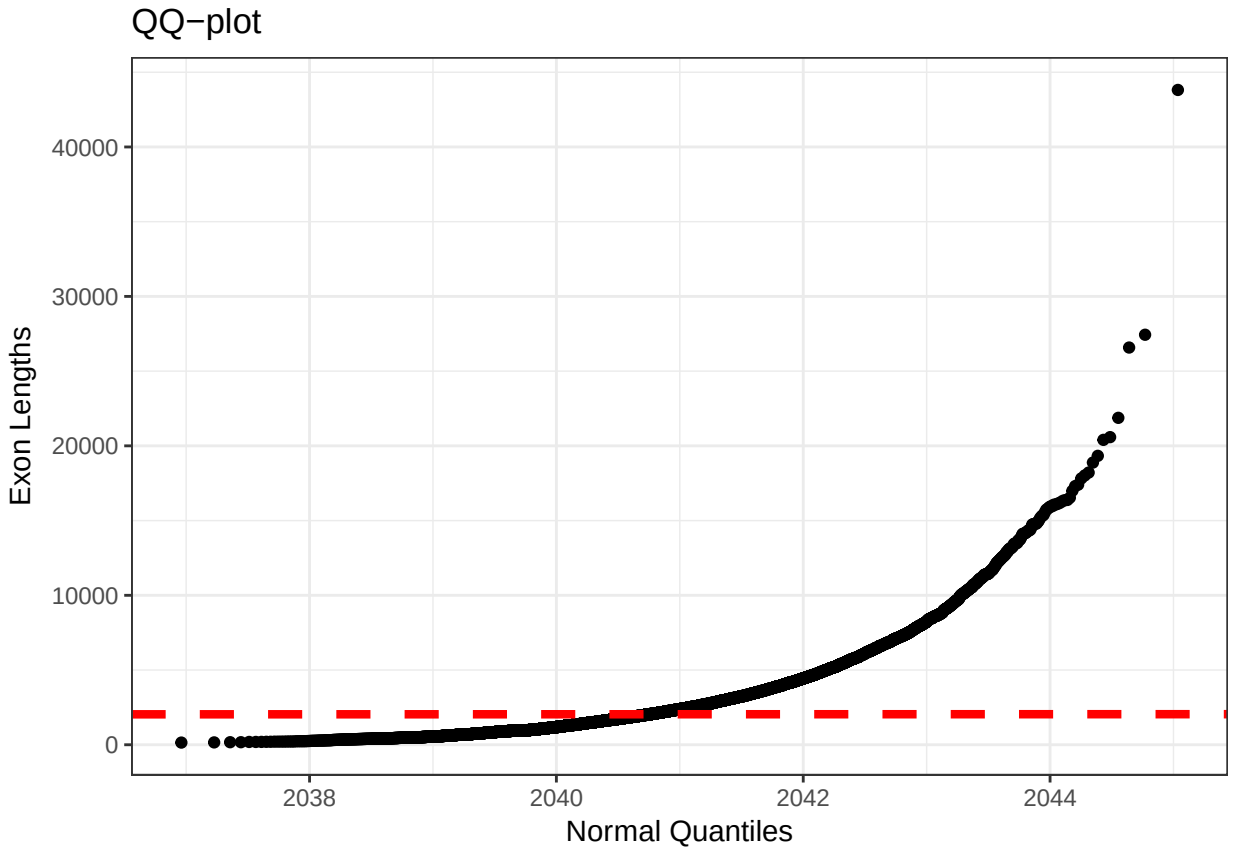


Figure 3. Distribution of intron and exons lengths *Histograms and boxplots showing the intron and exon length distributions. The exon median length lies around 2kb whereas the intron median length lies around a value 10 times bigger than the intron's, around 20kb.*

The histograms and box-plots are presented with a logarithmic scale in x-axis in order to clearly appreciate the differences between the two subsets. In them it can be seen that while most of exons (blue) tend to have shorter lengths -with a peak around 12 kB-, the introns (red) have a more right-tailed distribution, with generally longer lengths.

4. Are the mRNA lengths significantly longer than the total intron lengths, or is it the other way around?

We need to test the difference of the means of both distributions. The Student's T test may be used if a normal distribution can be assumed. Otherwise, only the corresponding non parametric test ought to be used (Wilcoxon test). In order to test normality, a Q-Q plot between the observed lengths and the normal distribution was drawn.



As we are comparing two data subsets that are not following a normal distribution (as seen in Figure 3), we have chosen to perform a Wilcoxon test to investigate if there is a significant difference in the lengths of the introns and exons of the genes of our data set. Our null hypothesis is that there is no significant difference in the U-statistic between the length of the exons and introns.

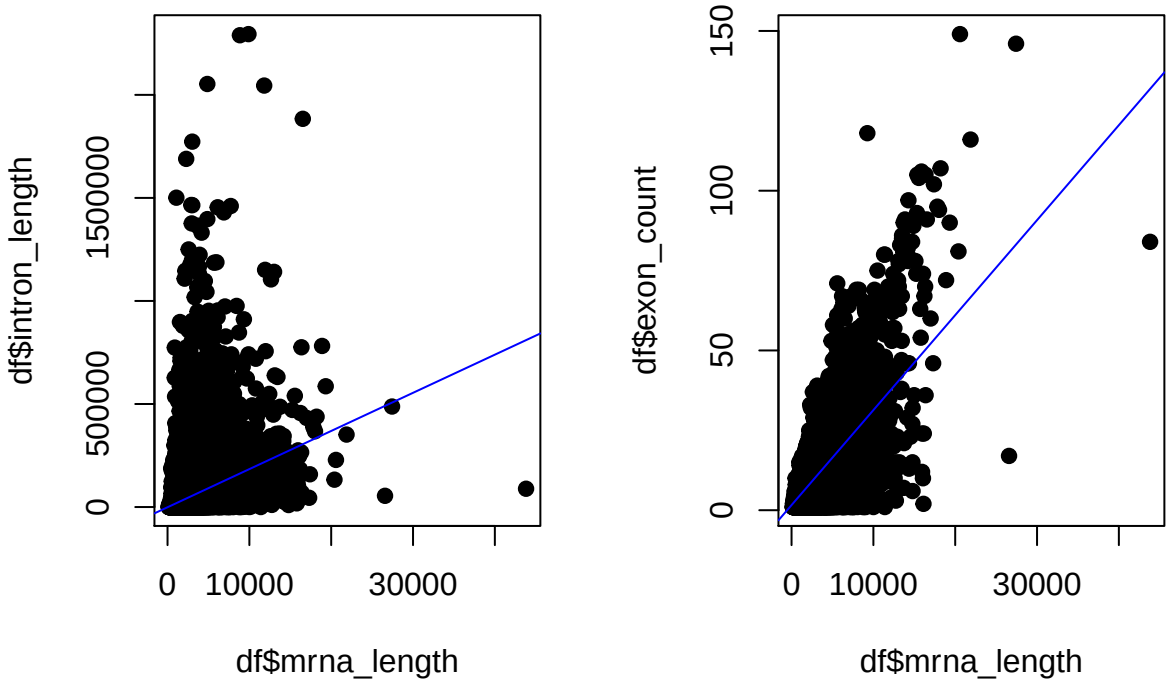
```
## [1] 0
```

5. Continuing on the same question: is the total exon length more correlated to the total intron length than the number of exons? Show this both with a plot and with correlation scores. Comment on your result

```
##
##  Pearson's product-moment correlation
##
## data:  df$mrna_length and df$intron_length
## t = 50.356, df = 18487, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3345642 0.3599163
## sample estimates:
##       cor
## 0.3473037
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$mrna_length and df$exon_count
```

```
## t = 112.96, df = 18487, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6304305 0.6474880
## sample estimates:
##       cor
## 0.6390378
```



6. What gene has the longest (total) exon length? How long is this mRNA and how many exons does it have? Do this in a single line of R (without using ";").

```
print(df[which.max(df$mrna_length), c(1,2,4)], row.names = FALSE)
```

```
##   name mrna_length exon_count
##  MUC16       43815         84
```

7. In genomics, we often want to fish out extreme examples – like all very short genes, or all very long genes. It is often helpful to make a function to do these tasks – it saves time in the long run.

```
count_genes <- function( df, x1 = 0, x2 = max(df$mrna_length))
{
 total.mrna <- length(df$name)
 mrna.interval <- sum(df$mrna_length >= x1 & df$mrna_length <= x2)
 mrna.fraction = mrna.interval / total.mrna
 return ( mrna.fraction * 100)
}
```

```
## [1] 100.000000   1.130402  87.349235  11.541998   0.000000
```