Assignment 1

May 8, 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.

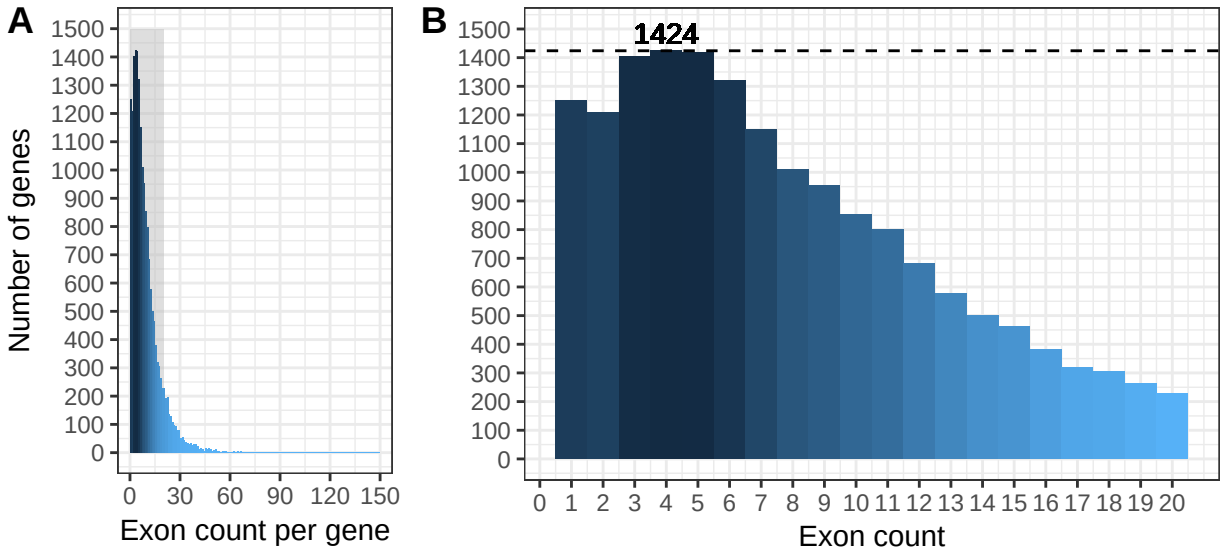

Figure 1. **A** *Histogram showing the distribution of the exon counts. Even though most of the genes contain less than 60 exons, as many as 150 may be found in some of them.* **B** *Detail for genes with max. 20 exons. The mode can be visualized at 3-5 exons per gene (max found at 4). The number of exons per gene decreases steadily beyond it.*

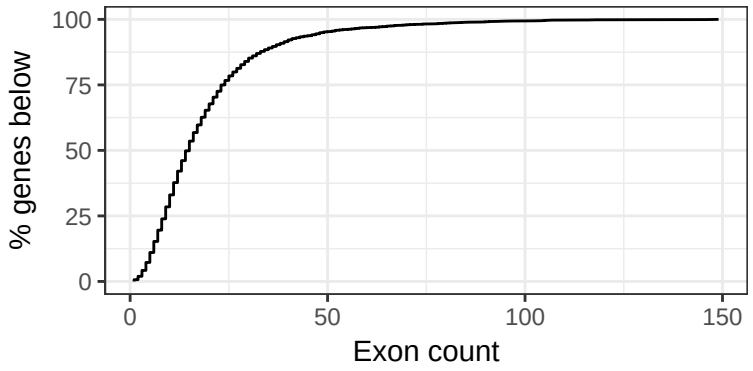

Figure 2. *Cumulative distribution of exon count per gene*

The majority of genes tend to be formed by a relatively low number of exons (Figure 1). 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
head(df)
```

```
##       name mrna_length genome_length exon_count intron_length
## 1   PP8961        2596          2596          1             0
## 2 FLJ00038         794          2615          6          1821
## 3    OR4F5         918           918          1             0
## 4    OR4F3         937           937          1             0
## 5   OR4F16         937           937          1             0
## 6   SAMD11        2555         18842         14         16287
```

Basically, the total length of introns for each gene is obtained by substracting the length of the mRNA (the exon length in this case) from the entire genome 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?

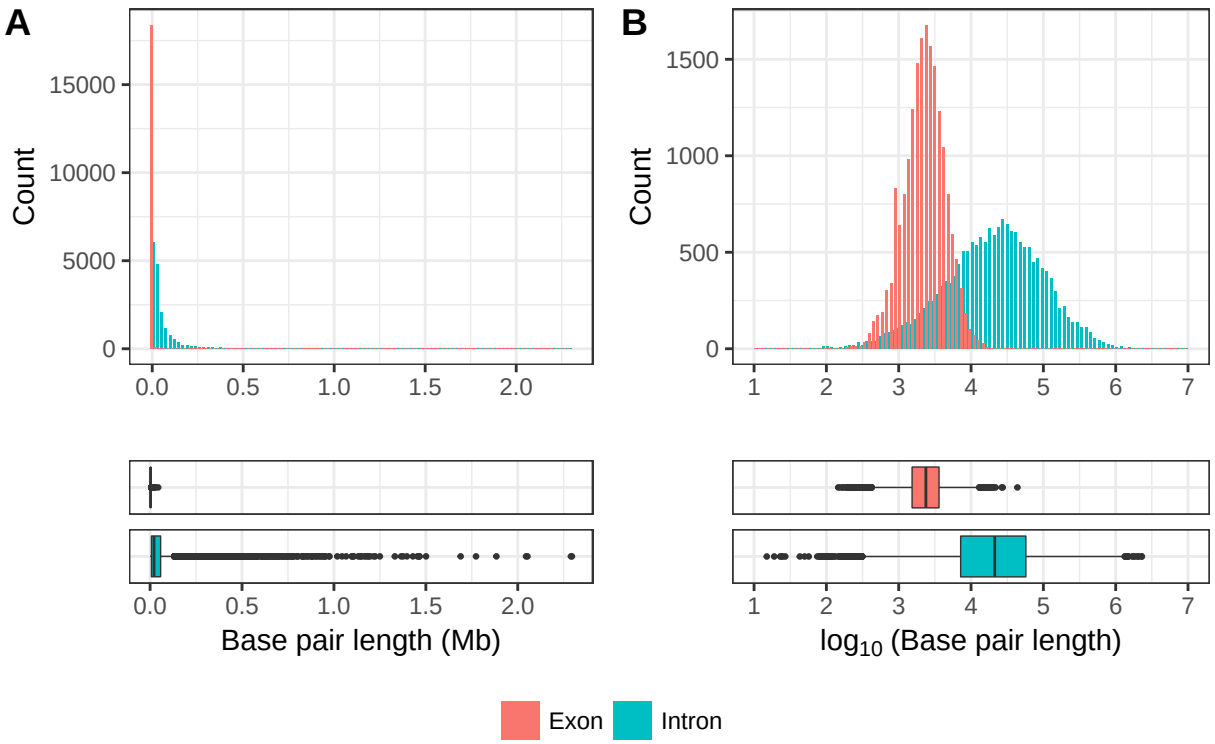

Figure 3. Distribution of intron and exons lengths in linear (A) and log10 (B) scale *The exon median length lies around 2kb whereas the intron median length lies around a value 10 times bigger than the intron's, around 20kb, as shown in the histograms (top) and boxplots (bottom).*

The histograms and box-plots in Figure 3.B are presented with a logarithmic scale in x-axis in order to clearly appreciate the differences between the two subsets. They show that while most of exons (red) tend to have shorter lengths -with a peak around 12 kB-, the introns (blue) have a more right-tailed distribution, with generally longer lengths, covering an extremly wide span.

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.

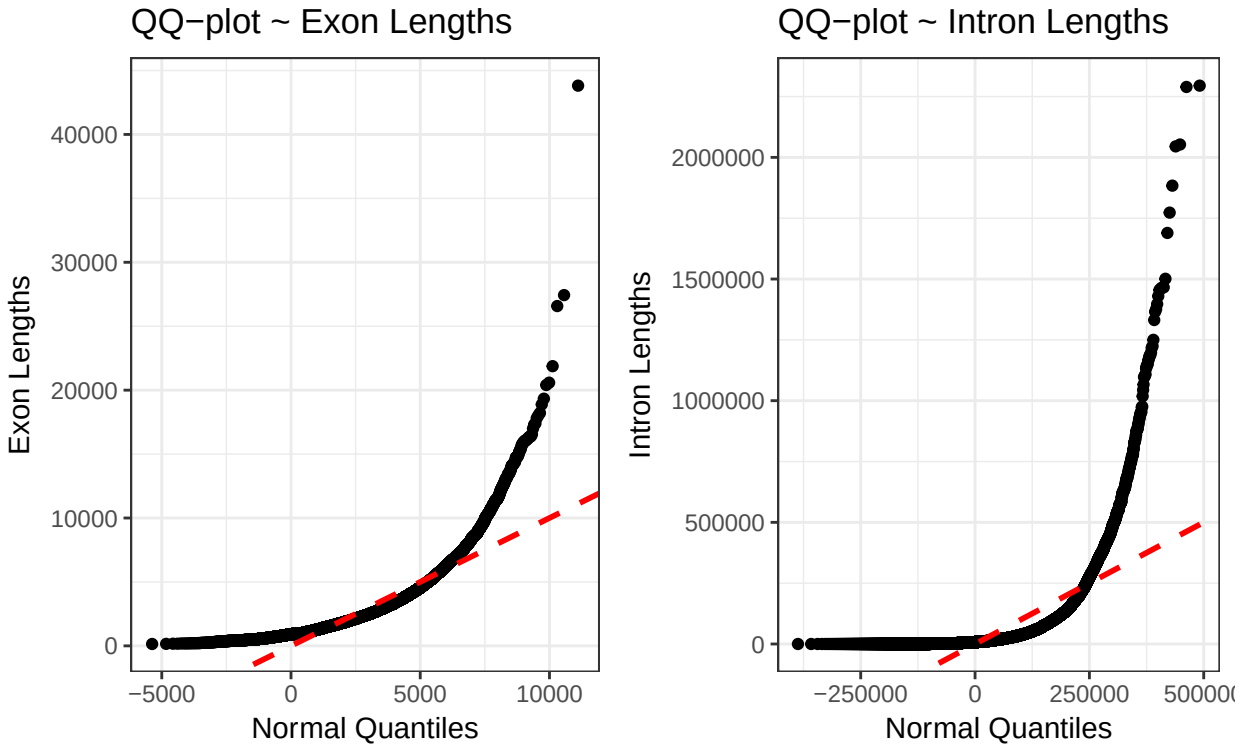

Figure 4 *QQplots of exon and intron length against normal distribution. Each point in the two plots above carries an observed component of intron or exon length (x-axis), and an expected component drawn from a normal distribution (y-axis) with a mean and standard deviation equal to the that of the intron or exon length sets. Both plots show significant deviations from the abline which maps evenly increasing values of a normal distribution, to itself.*

As we are comparing two data subsets that are not following a normal distribution (as seen in Figure 4), 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 of the data set.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df$mrna_length and df$intron_length
## W = 58458000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

As our p-value is $<$ 2.2e-16, it is below the significance threshold of 0.05, we can reject the null hypothesis and accept the alternative hypothesis, that is, there is a significant difference in the U-statistic between the length of the exons and introns. In addition to the previous observations of the lengths of introns and exons, we can conclude that the introns are significantly longer than the exons of the data set.

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

In order to determine whether total exons length is more correlated to the total intron length than the number of exons, we have calculated the Spearman correlation coefficient for both cases.

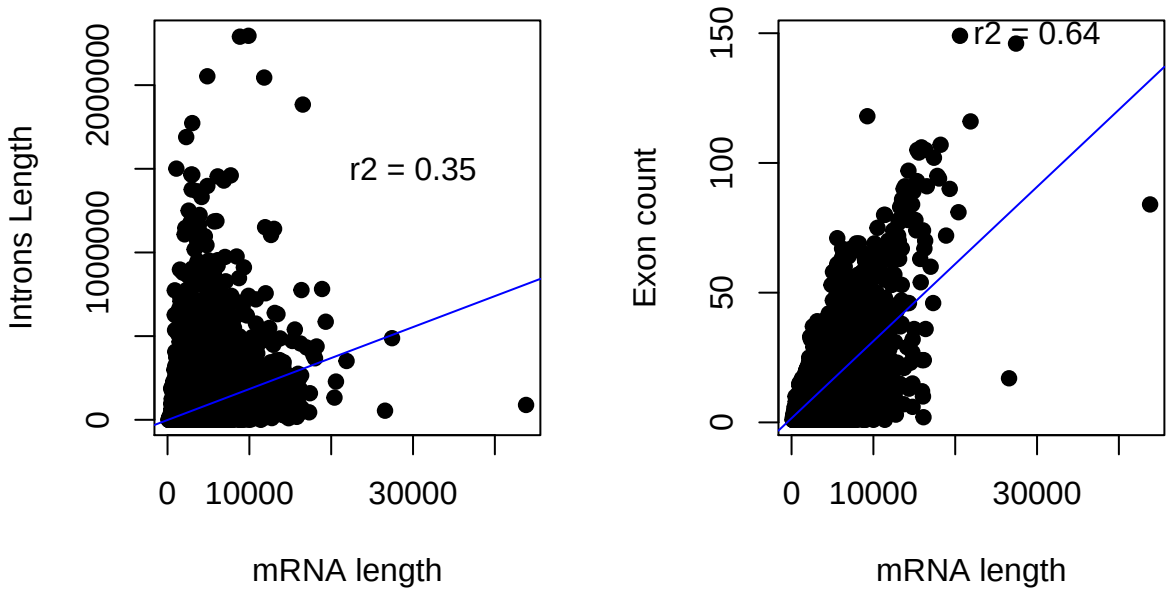

Figure 5 *Scatterplot of the studied variables showing regression lines and Pearson correlation coefficient.*

Based on the correlation scores of 0.35 for exon VS intron length and 0.64 for exon length VS number of exons and the scatter plots, it can be concluded that whereas there is a positive correlation in both cases, they are not very strong. It is proven though that the exon´s length is more correlated with the number of exons than with the length of the introns of the genes belonging our data set.

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
```

As can be seen above, the gene that has the longest total exon length is MUC16, with a length of 43815 base pairs and 84 exons.

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)
}
```

Test this function with the mRNA lengths using the the five settings below:

- Using the default of x1 and x2;
- Using the default of x2 and set x1=10000;

- x1=1000 and x2=10000;
- x1=100 and x2=1000;
- x1=0 and x2=100.

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

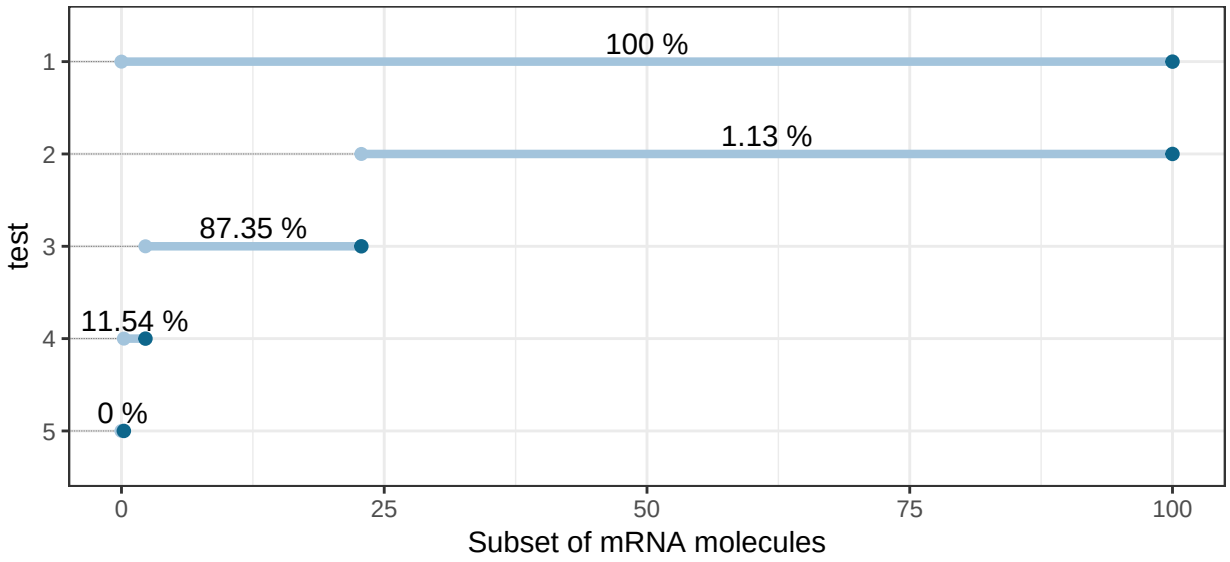

Figure 6 *Dumbbell plot of the intervals selected by the x1 and x2 pairs and the resulting percentage of the dataset within that interval.*