Homework 1

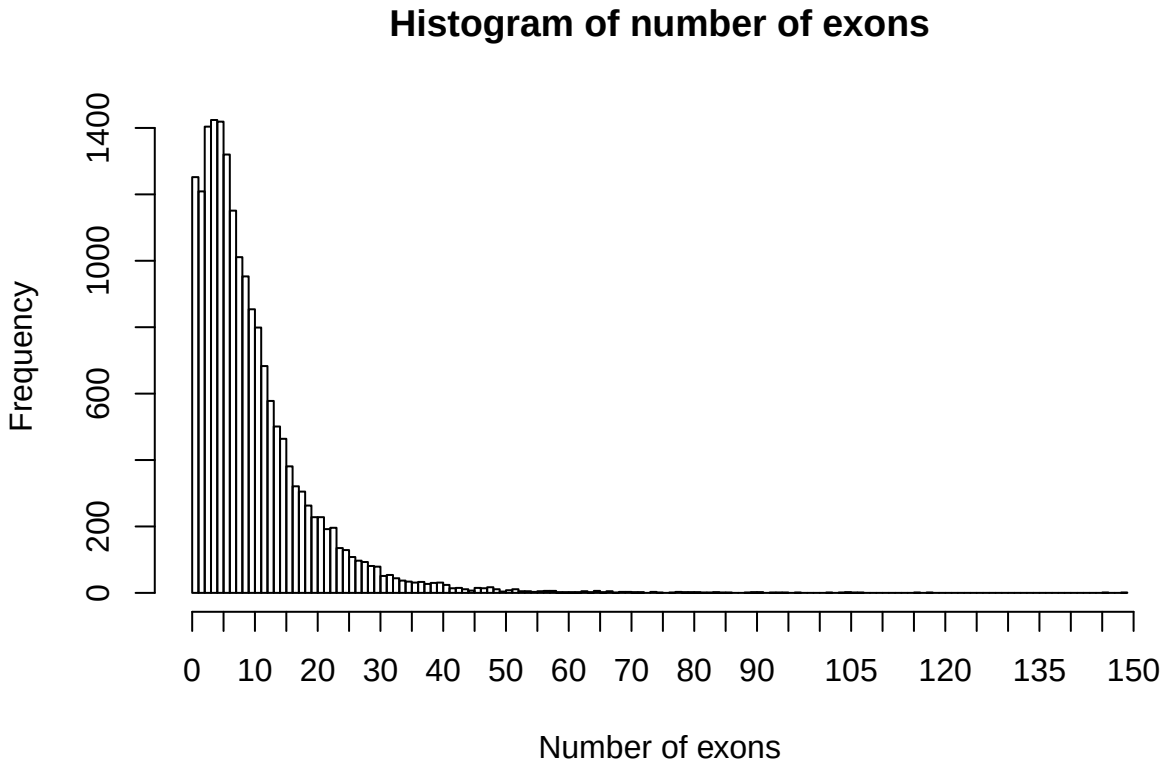
Group 7: Nynne Nymann, Peter Horskjr, Carlotta Porcelli, Max Tomlinson & Ke Zhai

5/5/2017

```
# clearing R-history and reading data
rm(list=ls())
gene_lengths<-read.table("gene_lengths_v2.txt", header=T)
```

Question 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.

```
hist(gene_lengths$exon_count, breaks=seq(0,max(gene_lengths$exon_count),by=1), xaxt="n",
     main="Histogram of number of exons", xlab="Number of exons")
axis(side=1, at=seq(0,150,5), labels=T)
```



The histogram shows the exon count distribution, where the most common number of exon lies between 3 and 5 with the most typical number of exons being equal to 4 (narrow margin). The number of genes with a high exon number decreases a lot after 5 exon.

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

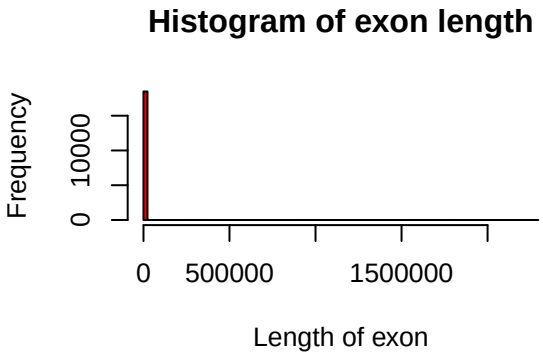
The intron length can be calculated as gene length minus its mRNA length. The column for intron length is added at the end of the data frame.

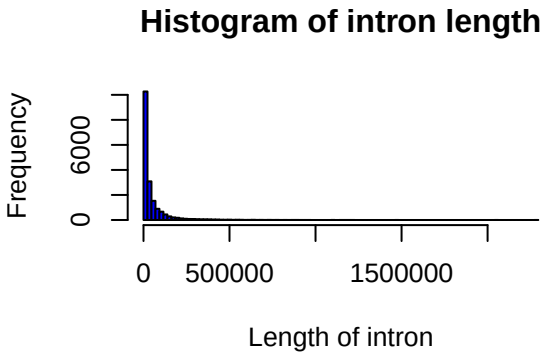
```
gene_lengths$intron_length<-gene_lengths$genome_length-gene_lengths$mrna_length
head(gene_lengths, n=3)
```

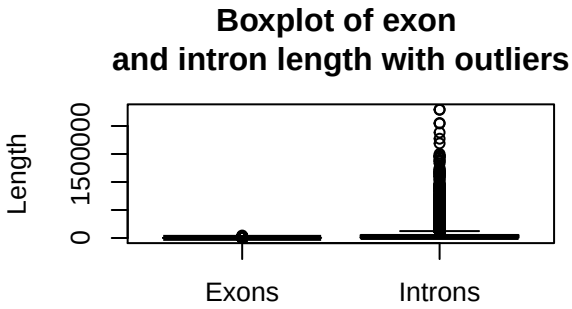
```
##       name mrna_length genome_length exon_count intron_length
## 1   PP8961        2596          2596          1             0
## 2 FLJ00038         794          2615          6          1821
## 3    OR4F5         918           918          1             0
```

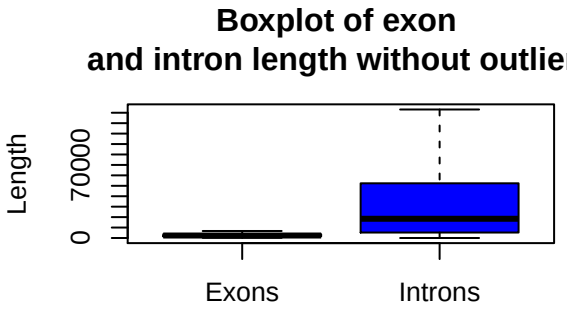
Question 3: Make histograms and boxplots showing the distribution of total exon and total intron lengths. Are exons larger than introns or vice versa?

```
par(mfrow=c(2,2))
seq_break_points = seq(0, max(gene_lengths$intron_length), l=100)
# Exon length histogram
hist(gene_lengths$mrna_length, col="red", breaks=seq_break_points,
     xlim=c(0,max(gene_lengths$intron_length)), main="Histogram of exon length",
     ylab="Frequency", xlab="Length of exon")

# Intron_length histogram:
hist(gene_lengths$intron_length,col="blue",
     breaks=seq_break_points, xlim=c(0,max(gene_lengths$intron_length)),
     main="Histogram of intron length", ylab="Frequency", xlab="Length of intron")

# Boxplot for exon and intron length:
boxplot(gene_lengths$mrna_length, gene_lengths$intron_length,
col=c("red","blue"),names=c("Exons","Introns"), main="Boxplot of exon
and intron length with outliers", ylab="Length", outline = T)

boxplot(gene_lengths$mrna_length, gene_lengths$intron_length,
col=c("red","blue"),names=c("Exons","Introns"), main="Boxplot of exon
and intron length without outliers", ylab="Length", outline = F, yaxt="n")
axis(side=2, at=seq(0,120000,10000), labels=T)
```







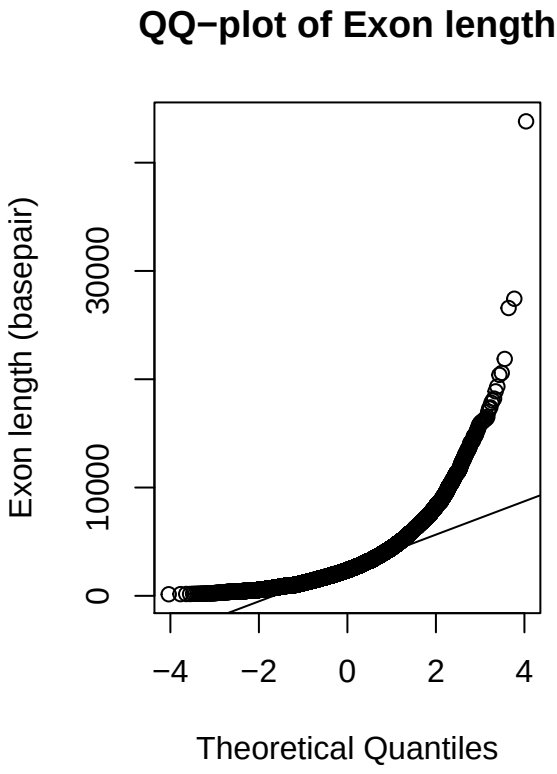


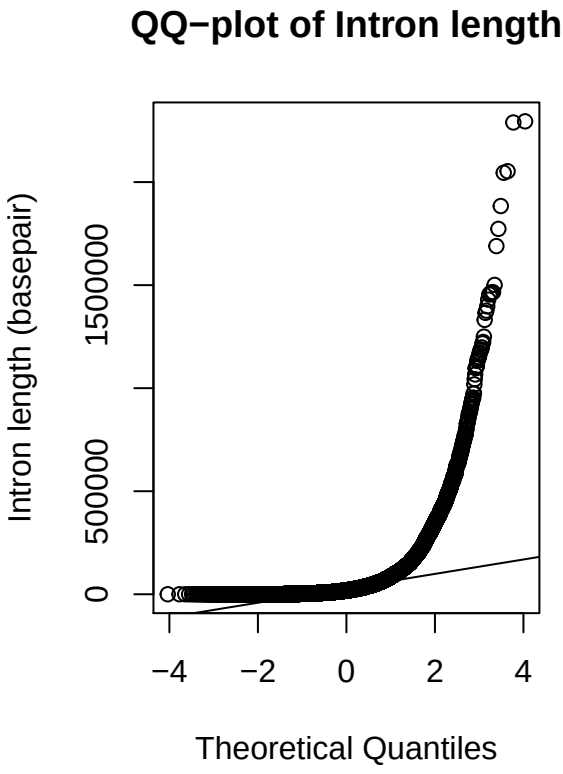
It is clear by looking at both the histograms and the box plot that the intron length is longer than the exon length. Also, there is greater variability in intron length when compared to exon. The boxplots show the intron and exon lenght with and without outliers. The first boxplot shows a valuable presence of outlier in the introns and a minor amount in the exons. The second version has been made to allow a better view of the results setting the outline to False.

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

In order to choose the appropriate statistical test we used the qqnorm function in R to determine if our data satisfied the assumption of normality. Neither the intron and exon lengths follow a normal distribution so we used the non-parametric Wilcoxon test.

```
par(mfrow=c(1,2))
#QQ_plot for exon length
qqnorm(gene_lengths$mrna_length,ylab="Exon length (basepair)",
       main="QQ-plot of Exon length")
qqline(gene_lengths$mrna_length)
#QQ_plot for intron length
qqnorm(gene_lengths$intron_length,ylab="Intron length (basepair)",
       main="QQ-plot of Intron length")
qqline(gene_lengths$intron_length)
```





The Wilcoxon test does not look at the difference in medians but uses the difference in ranked U-statistics. Measuring U-statistics is advantageous because it takes all values into consideration not just the median. We are using a non-paired test because we can't assume that the data is paired. Our hypotheses for the Wilcoxon test are as follows:
H0 = The U-statistic between the length of mRNAs and introns is the same
HA = The U-statistic between the length of mRNAs and introns is not the same

```
wilcox.test(gene_lengths$mrna_length, gene_lengths$intron_length, paired=F)
```

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

Our p-value $<$ 2.2e-16, which is below our significance threshold of 0.05, therefore we can reject the null hypothesis that there is no difference between the mRNA and intron length. In summary, the results from the Wilcoxon test suggests that the intron lengths are significantly different from the exon lengths, which is concurrent with our observations in question 3.

Question 5: Is the total exon length more correlated to the total intron length than the number of exons?

The Pearson correlation coefficient goes from -1 to +1, where a score equals one of these values indicate perfect correlation. As we can see, both plots exhibit a positive correlation score. The score for exon count is closer to +1, indicating that there is a higher correlation between the exon count and exon length than the interplay between intron - and exon length.

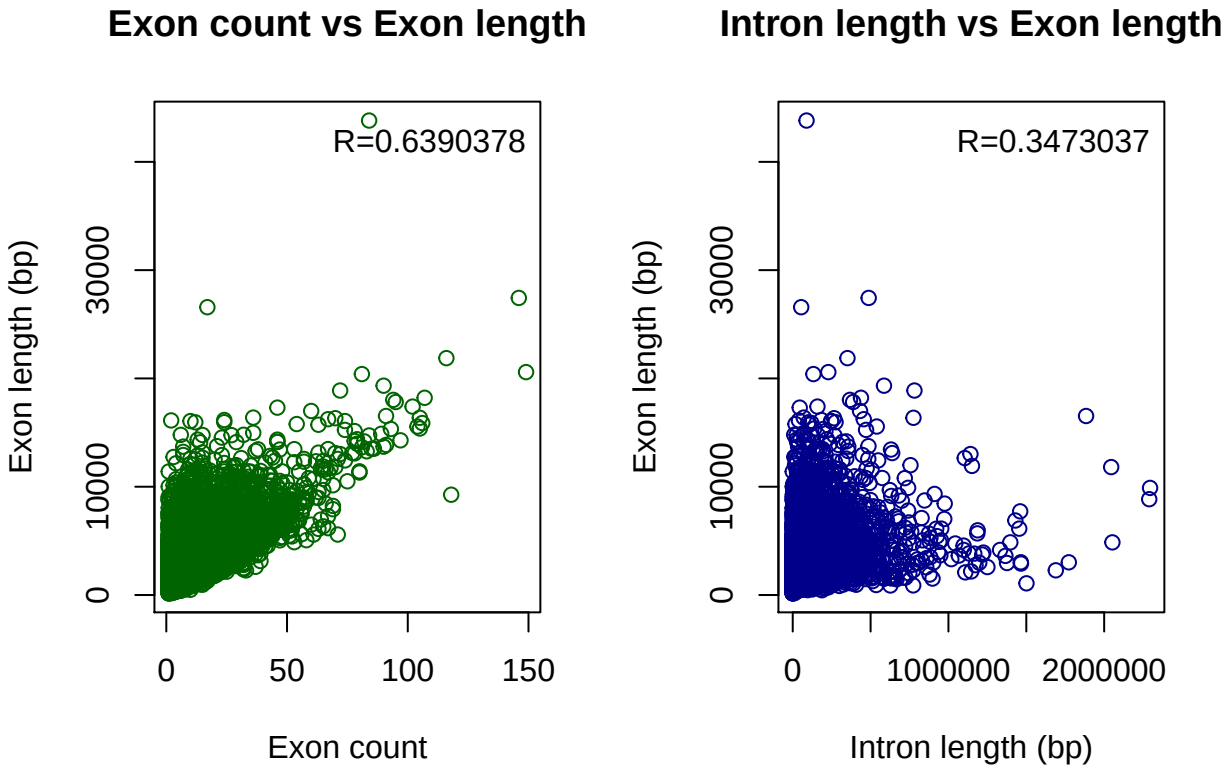
```
cor(gene_lengths$mrna_length,gene_lengths$exon_count,method="pearson")
```

```
## [1] 0.6390378
```

```
cor(gene_lengths$mrna_length,gene_lengths$intron_length,method="pearson")
```

```
## [1] 0.3473037
```

```
par(mfrow=c(1,2))

lm_exons<-lm(gene_lengths$mrna_length~gene_lengths$exon_count)
plot(gene_lengths$exon_count, gene_lengths$mrna_length, col="dark green",
     main = "Exon count vs Exon length",
     xlab = "Exon count", ylab = "Exon length (bp)")
     legend("topright", "R=0.6390378", bty="n")

lm_intron_length<-lm(gene_lengths$mrna_length~gene_lengths$intron_length)
plot(gene_lengths$intron_length, gene_lengths$mrna_length, col=" dark blue",
     main = "Intron length vs Exon length",
     xlab = "Intron length (bp)", ylab = "Exon length (bp)")
     legend("topright", "R=0.3473037", bty="n")
```



Question 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 ???;???)

```
gene_lengths[which.max(gene_lengths$mrna_length),c(1,2,4)]
```

```
##       name mrna_length exon_count
```

```
## 8385 MUC16       43815         84
```

This gene is identified as 8385 MUC16 is 43815bp long with 84 exons.

Question 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(vector_l, x1=0, x2=max(vector_l)){
  count <- sum(vector_l > x1 & vector_l <= x2)
  total_mrna_count <- length(vector_l)
  return(count/total_mrna_count)
}
# Test the function with:
mrna_l <- gene_lengths$mrna_length
count_genes(mrna_l)
```

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

```
count_genes(mrna_l, x1=10000)
```

```
## [1] 0.01130402
```

```
count_genes(mrna_l, x1=1000, x2=10000)
```

```
## [1] 0.873276
```

```
count_genes(mrna_l, x1=100, x2=1000)
```

```
## [1] 0.11542
```

```
count_genes(mrna_l, x1=0, x2=100)
```

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