I519: Homework 1

Will Shoemaker

08 September, 2015

Overview

Set working directory

```
rm(list=ls())
getwd()
```

```
## [1] "/Users/WRShoemaker/github/PopGen/I519/HW1"
```

```
setwd("~/github/PopGen/I519/HW1/")
```

Import packages

```
library("quantmod")
```

```
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: TTR
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

```
library("ggplot2")
library("reshape2")
library("wesanderson")
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following object is masked from 'package:xts':
##
##     last
```

Import the data and add headers

```
draft <- read.table("./genesize-draft.txt",header=F)
complete <- read.table("./genesize-complete.txt",header=T)

colnames(draft) <- c("Strain", "Genes")
colnames(complete) <- c("Strain", "Genes")
```

Mean gene number

```
mean(draft[,2])
```

```
## [1] 2678.527
```

```
mean(complete[,2])
```

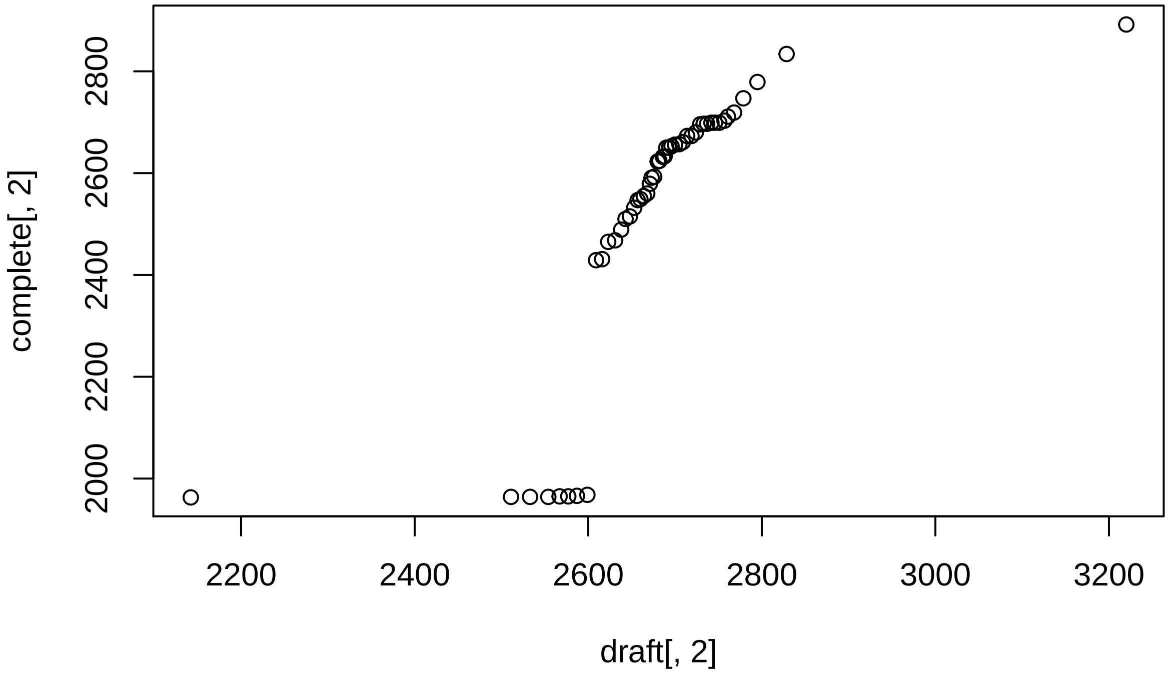
```
## [1] 2521.816
```

So the draft genomes have a mean gene number of ~2,679 and the complete genomes have a mean gene number of ~2522. Let's look if there's a significant difference.

Examine the quantile-quantile plots for the samples

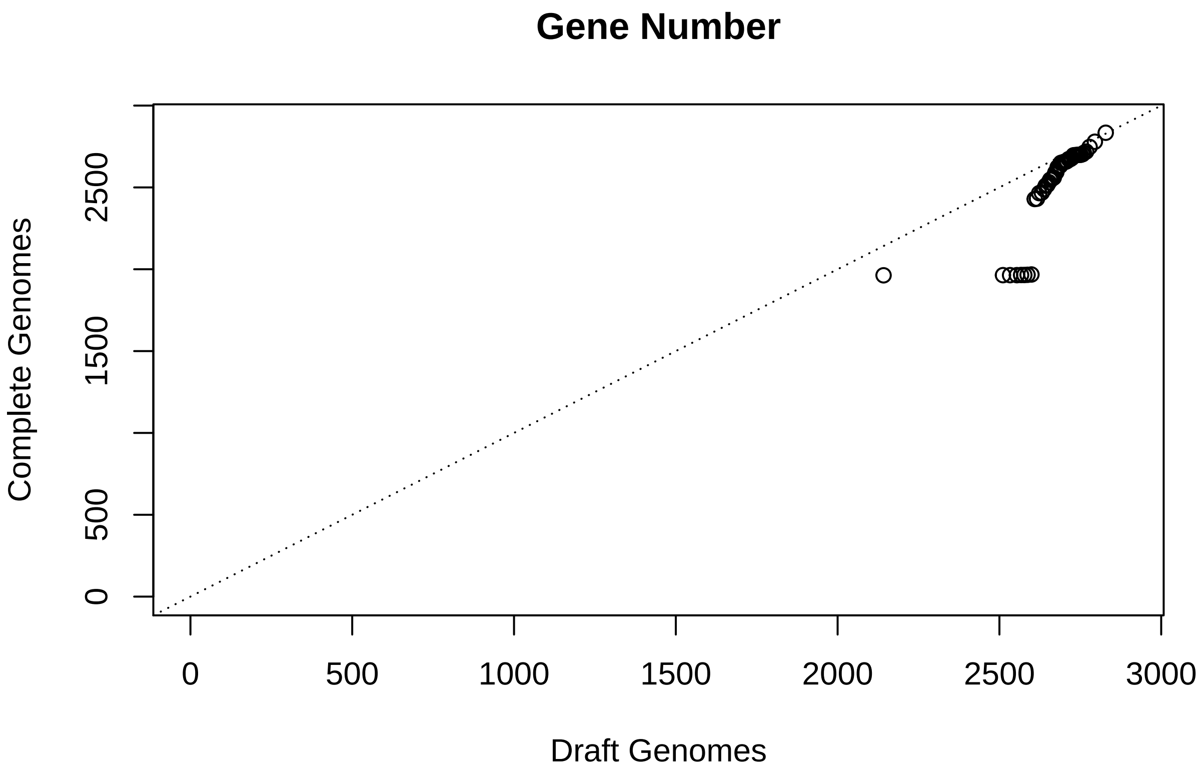
First we make a Q-Q plot that examines both samples.

```
qqplot(draft[,2], complete[,2])
```



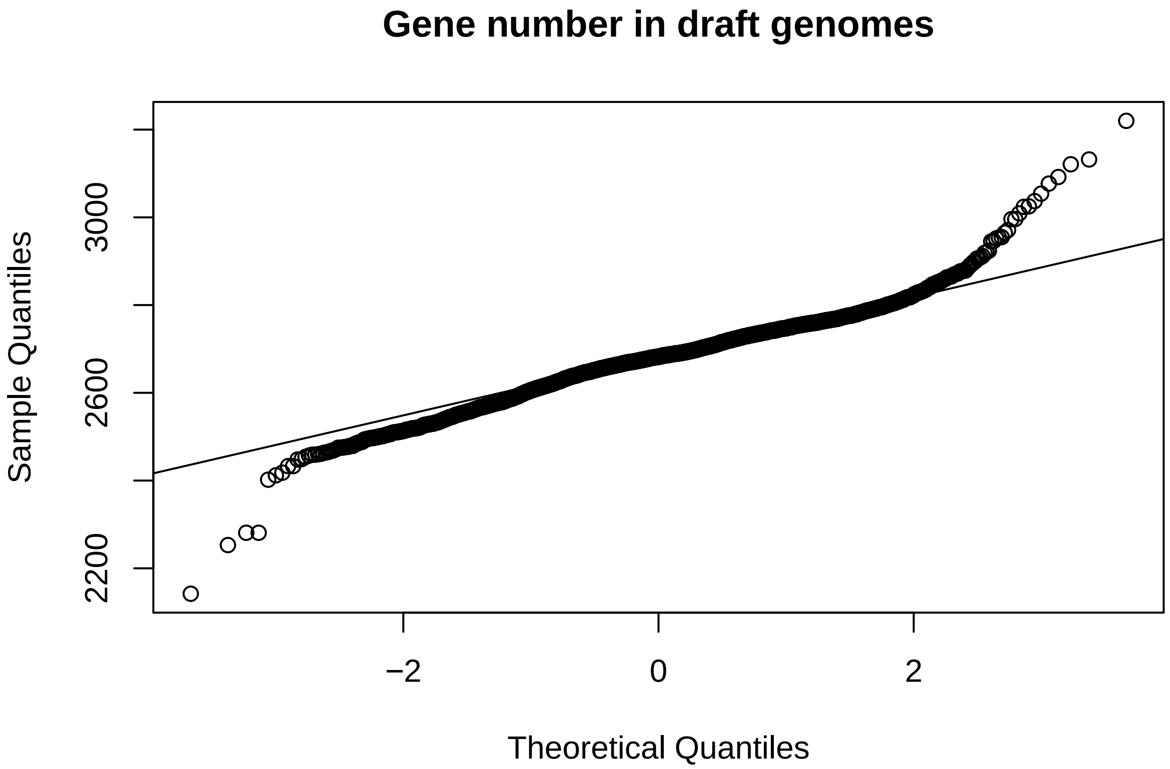
```r
xlim = range(1800, draft)
ylim = range(1800, complete)

qqplot(draft[,2], complete[,2],
xlim = ylim, ylim = ylim,
xlab = "Draft Genomes",
ylab = "Complete Genomes",
main = "Gene Number")
abline(a=0, b=1, lty="dotted")
```

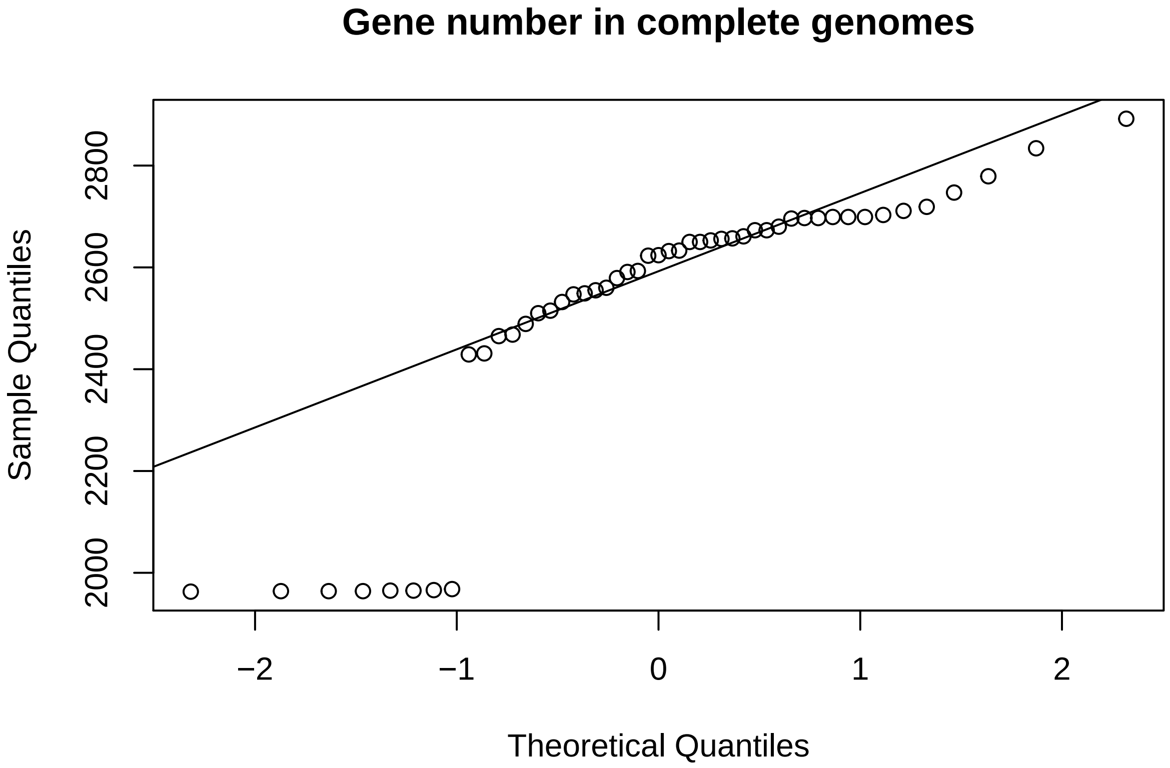


Then we make a Q-Q plot for each sample that compares the theoretical quantiles with sample quantiles.

```r
qqnorm(draft[,2],
main = "Gene number in draft genomes")
qqline(draft[,2])
```



```
qqnorm(complete[,2],
main = "Gene number in complete genomes")
qqline(complete[,2])
```



We see from above that there is a clear deviation from the theoretical expectation for each sample, suggesting that it would not be a good idea to assume normality. We can then proceed with a Wilcoxon rank-sum test.

```
wilcox.test(draft[,2], complete[,2], alternative = c("two.sided"), paired = FALSE, conf.level = 0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  draft[, 2] and complete[, 2]
## W = 140410, p-value = 6.078e-07
## alternative hypothesis: true location shift is not equal to 0
```

There is a significant difference in the number of genes between draft and complete genomes.

Now we plot the data.

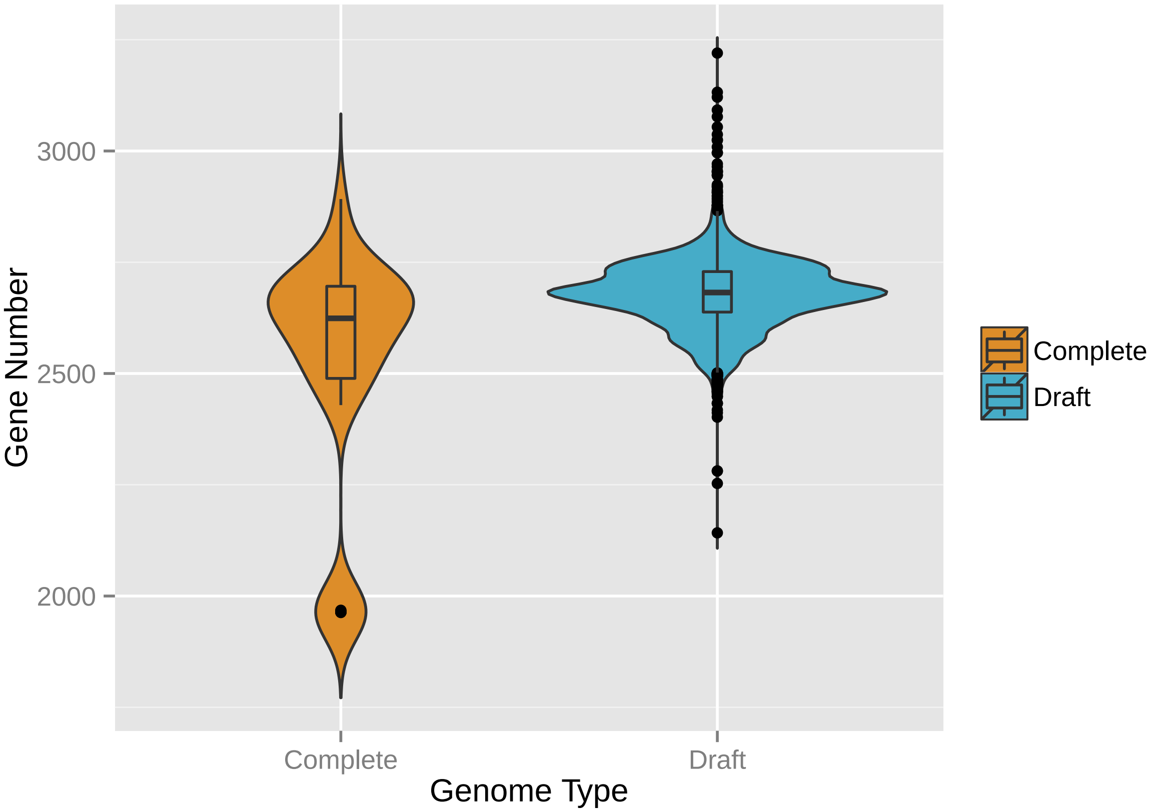
```
draft$Type <- rep("Draft",nrow(draft))
complete$Type <- rep("Complete",nrow(complete))
mergeData <- rbind(draft, complete)
mergeData$logGenes <- log(mergeData[,2])
class(mergeData)
```

```
## [1] "data.frame"
```

```
head(mergeData)
```

```
##                                      Strain Genes  Type logGenes
## 1 Staphylococcus_aureus_06BA18369_uid170654  2804 Draft 7.938802
## 2  Staphylococcus_aureus_07_03450_uid226303  2476 Draft 7.814400
## 3  Staphylococcus_aureus_07_03451_uid226304  2484 Draft 7.817625
## 4  Staphylococcus_aureus_08_01059_uid226297  2566 Draft 7.850104
## 5  Staphylococcus_aureus_08_01062_uid226298  2566 Draft 7.850104
## 6  Staphylococcus_aureus_08_01084_uid226305  2574 Draft 7.853216
```

```
palette <- wes_palette(5, name = "FantasticFox", type = "discrete")
ggplot(mergeData, aes(x=Type, y=Genes, fill = Type)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width=0.1) +
  scale_fill_manual(values=palette[-2]) +
  xlab("Genome Type") +
  scale_size_area("Genome Type") +
  ylab("Gene Number") +
  guides(fill=guide_legend(title=NULL))
```



Summary

From a Wilcoxon rank-sum test we see that there is a significant difference in the mean ranks of our complete and draft genome datasets (p-value = 6.078e-07). This suggests that draft and complete genomes have a different number of genes. However, there are several factors we must consider. The first one being that we do not know what software was used to annotate the genomes. If different annotation software was used across samples, this would make it difficult to distinguish whether gene number differences are due to the fragmented nature of the draft genomes or certain biases in the annotation process of the software. For example, if a draft genome is fairly complete, but highly fragmented, and the annotation software counts genes towards the end of contigs that do not contain a stop codon as genes, it is possible that you may double count some genes. It is unlikely that the differences here are biological in the organismal sense, as the text file labels suggest that all the genomes given are sequenced within the species *Staphylococcus aureus*.