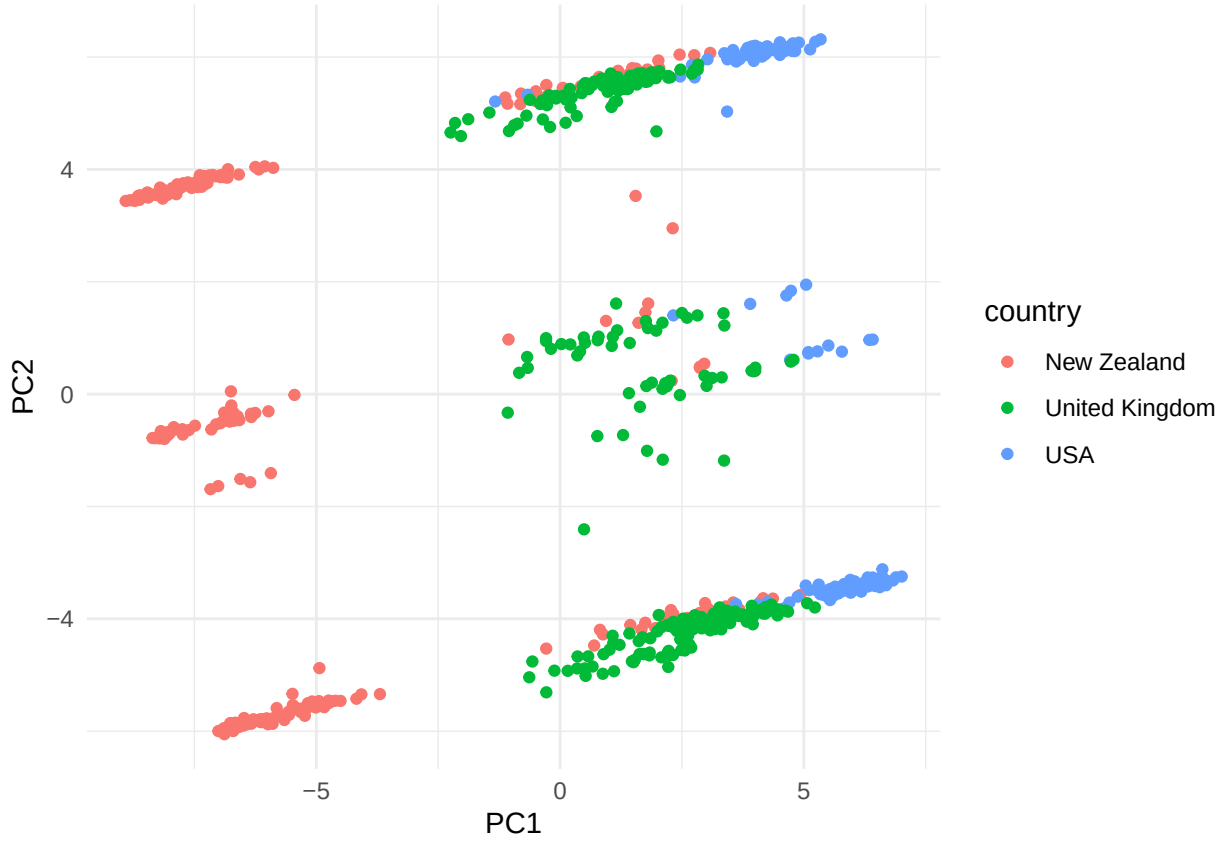
Investigating Clustering of PC1 and PC2

Noah Legall

10/8/2020

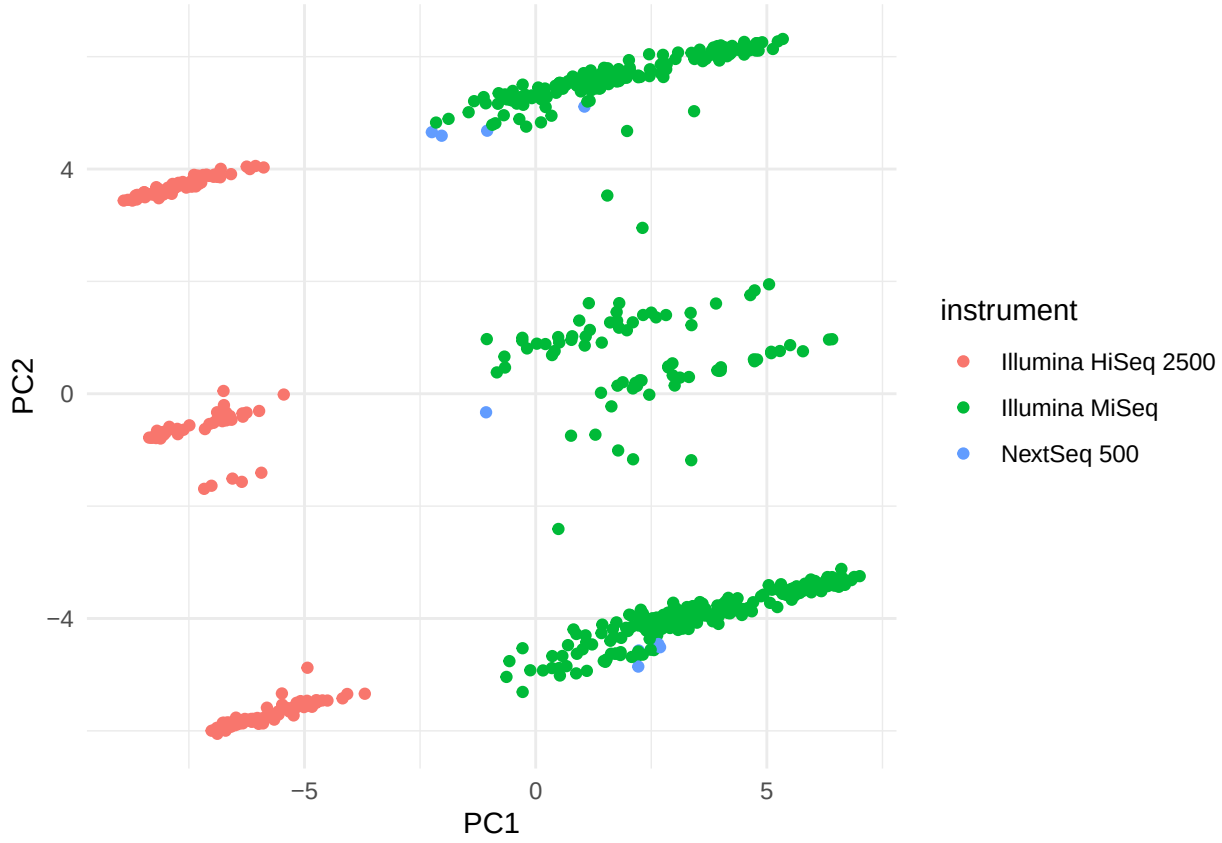
I wanted to dig into what could possibly lead to the clustering that is observed when we look at PC1 & PC2. As a reminder let's observe PC's 1 & 2 score plot:

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



For the first principal component PC1, we can see that the clustering might be explained by the method used to sequence the M. bovis isolates. PC2 indicates that some isolates are more like each other across countries than within countries, and there are three different classes of isolates. This is the clustering we can observe if we project the points onto PC2.

```
ggplot(as.data.frame(scores),aes(x=PC1,y=PC2, fill = instrument, col = instrument)) +
  geom_point() +
  theme_minimal()
```



In trying to figure out how this clustering could be occuring, the closest possible solution I could find is from "Population Structure, Stratification, and Introgression of Human Structural Variation" where the 1st PC's are probably due to genome assembly differences.

To test is this is happening in my data, I will merge the assembly stats with the PCA data and do correlation testing on PC1 and PC2. I'll run a for loop to test every numeric column in the genome assembly statistics dataset.

```
pc1 <- scores$PC1
#quick test
#cor.test(pc2,scores$N50,method = "pearson")

correlation_pvalue <- c()
genome_quality_stat <- c()
for(i in names(scores)){
  if(is.numeric(scores[,i])){
    PC1_corr <- cor.test(pc1,scores[,i])
```

```
    correlation_pvalue <- append(correlation_pvalue,PC1_corr$p.value)
    genome_quality_stat <- append(genome_quality_stat,i)
  }
  else{
    next()
  }

}
```

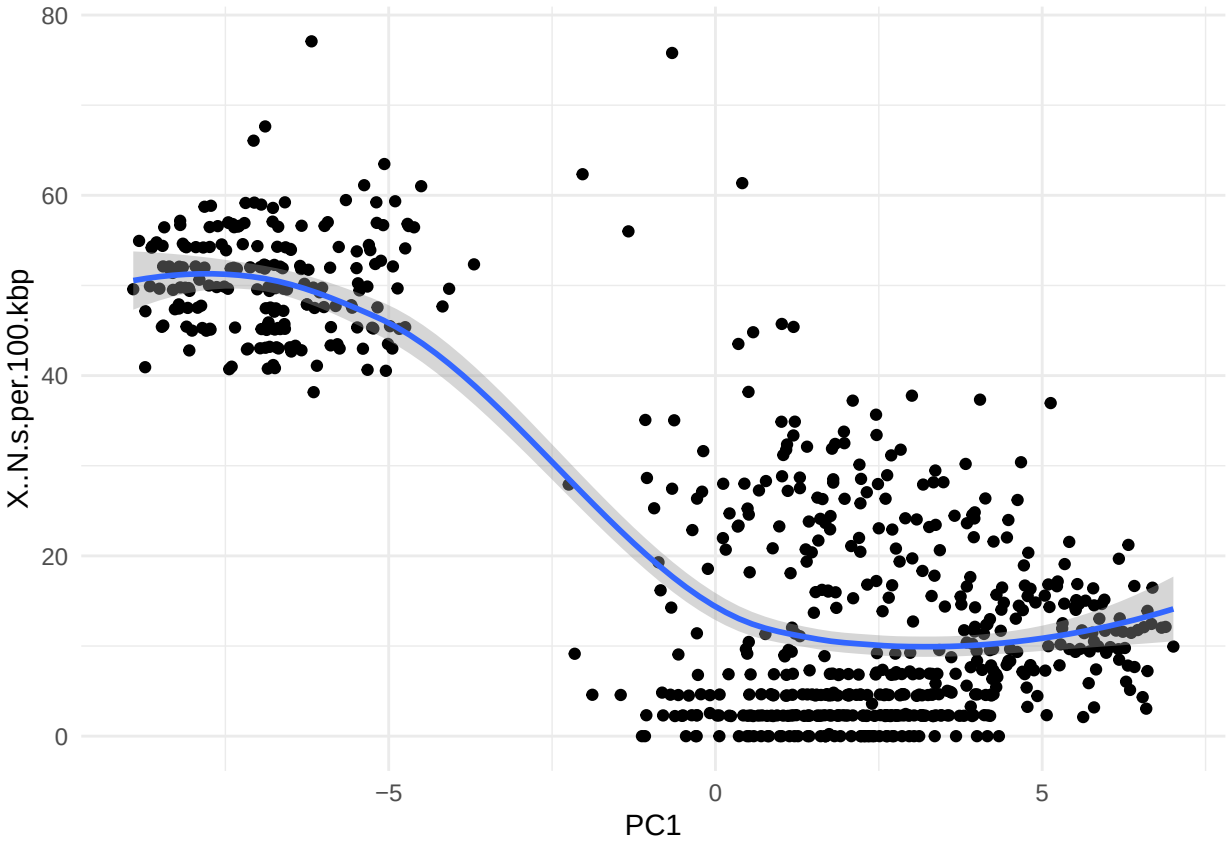
```
## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero
```

```
corr_values <- data.frame(genome_quality_stat,correlation_pvalue)


#sort from lowest to highest
#order by the lowest p - value
head(corr_values[order(corr_values$correlation_pvalue),])
```

```
##    genome_quality_stat correlation_pvalue
## 1                  PC1       0.000000e+00
## 42  X..N.s.per.100.kbp      4.743733e-169
## 22              GC....      3.465260e-109
## 27                NG75       2.494758e-80
## 25                NG50       1.035167e-79
## 26                 N75       4.771922e-78
```

PC1 seems to be strongly correlated with genome quality based on the correlation pvalues, the highest being the number of N's per kilobase for the genome assemblies (p = 4.743733e-169).This can be visualized by plotting the values of PC1 and X..N.s.per.100.kbp:

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
pc2 <- scores$PC2
#quick test
#cor.test(pc2,scores$N50,method = "pearson")

correlation_pvalue <- c()
genome_quality_stat <- c()
for(i in names(scores)){
  if(is.numeric(scores[,i])){
    PC2_corr <- cor.test(pc2,scores[,i])
    correlation_pvalue <- append(correlation_pvalue,PC2_corr$p.value)
    genome_quality_stat <- append(genome_quality_stat,i)
  }
  else{
    next()
  }

}
```

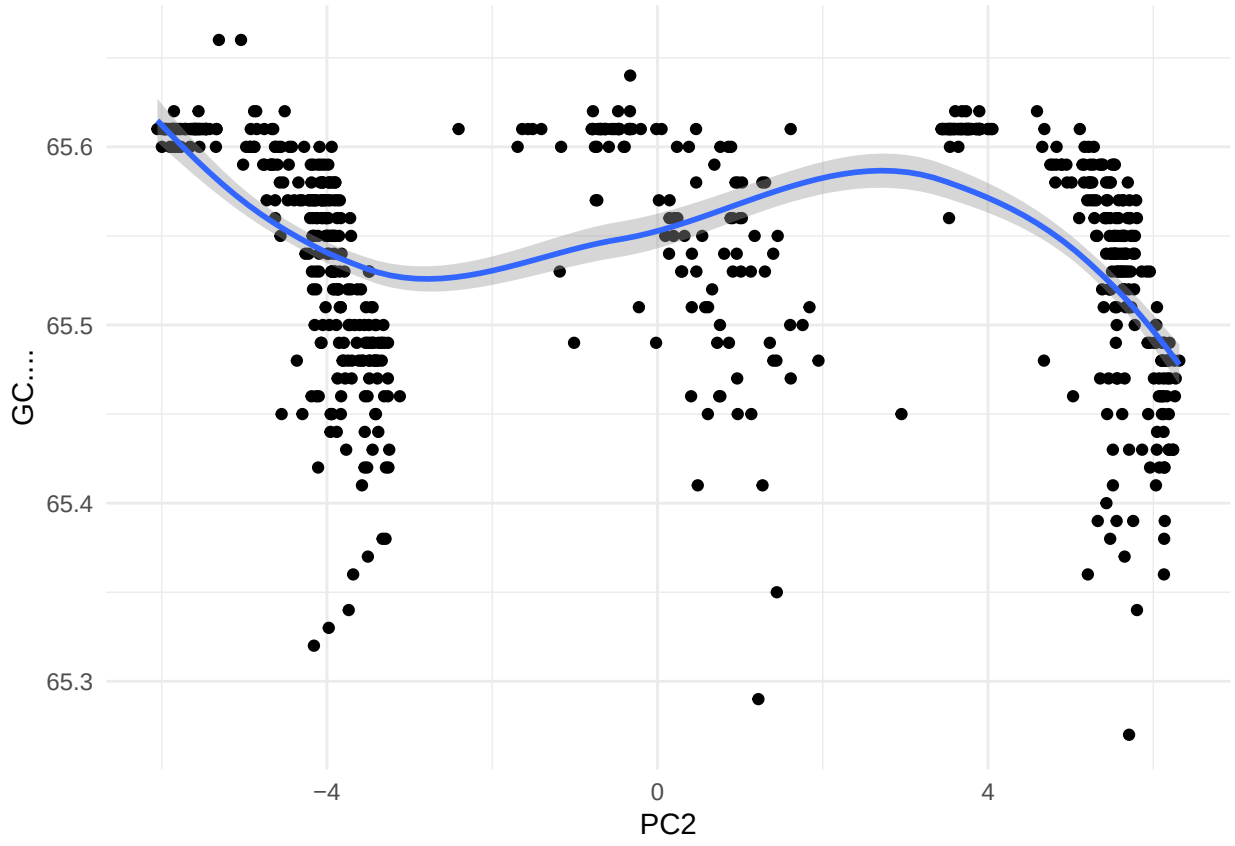
```
## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero
```

```r
corr_values <- data.frame(genome_quality_stat,correlation_pvalue)


#sort from lowest to highest
#order by the lowest p - value
head(corr_values[order(corr_values$correlation_pvalue),])
```

```
##    genome_quality_stat correlation_pvalue
## 2                  PC2       0.000000e+00
## 22              GC....       5.259854e-07
## 40 Genome.fraction....       7.459569e-05
## 42  X..N.s.per.100.kbp       3.585962e-04
## 48               NGA75       1.434025e-03
## 32    X..misassemblies       3.987008e-03
```

PC2 seems to also have some type of correlation with genome quality but it isn't as convincing as the correlations for PC1. GC content was the lowest pvalue for PC2, so let's look at the scatterplot of GC content to PC2:

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The GC content doesn't seem to really segregate well the PC2 values (we still see 3 distinct classes of clusters). This suggests that the genome quality isn't what explains the clustering we see along this axis.

Summary

Genome assembly can't particularly explain the patterns we see in PC2. PC1 can be explained by Genome Quality based on the correlations that match with PC1. We still do not understand fully the clustering along PC2, but this informs my next intuitions that maybe the variation we see in PC2 might be through technical noise.