# M. bovis Pangenome Analysis

Noah Legall

9/22/2020

### Introduction

### Figures

## 1. Data Description

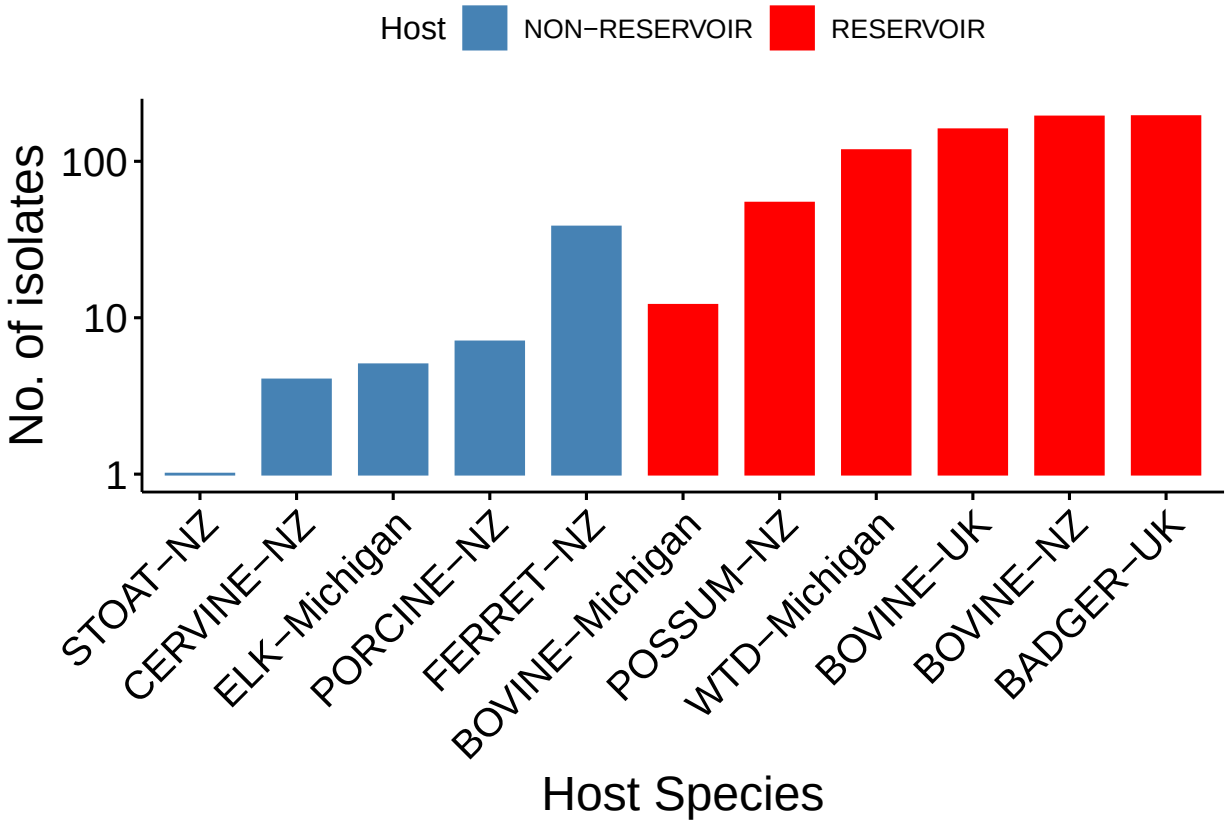
```
###1. This is the metadata we pulled from NCBI. Make sure your working directory is the Resource folder
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
library(dplyr)
```

```
##
## 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
```

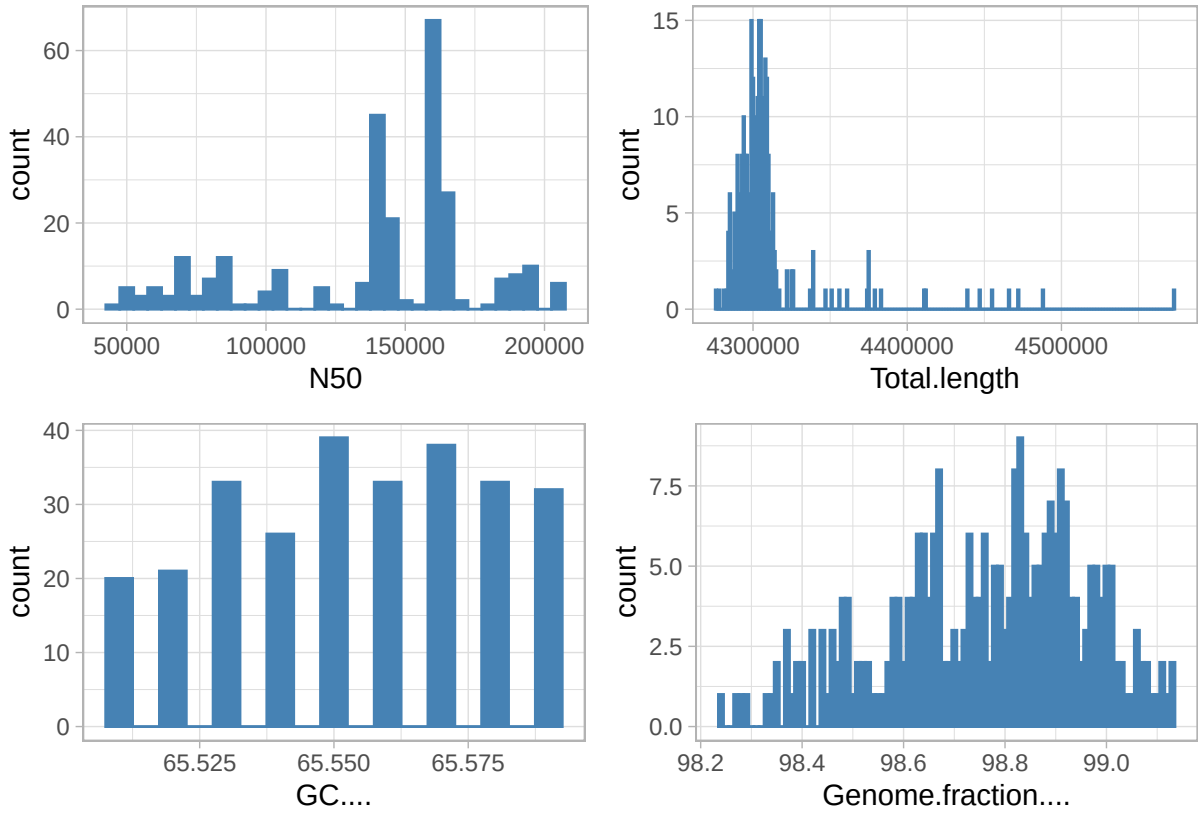
```
accession_meta <- read.csv("ProjectBovisReservoir_metadata.csv", header = TRUE, stringsAsFactors = FALSE

accession_bar <- accession_meta %>% group_by(Species, Host) %>% summarise(no_in_data = length(Species))

accession_bar <- accession_bar[order(accession_bar$Host,accession_bar$no_in_data),]

#A bar plot separated by species, colored by reservoir status.
ggbarplot(accession_bar,"Species","no_in_data", fill = "Host", color = "Host",palette = c("steelblue","r
  theme(axis.text.x = element_text(angle = 45, hjust = 1),axis.text=element_text(size = 15), axis.title
scale_y_log10()
```




## 2. Assembly Stat Distributions

```
library(patchwork)
library(magrittr)
setwd("/Users/noah_/Documents/Bovis-PangenomeOfReservoirs/Resources/")
assembly_stats <- read.csv("mbovis_transposed_report.csv",header = TRUE)
assembly_stats %<>%
  filter(N50 > 45000) %>%
  filter(Total.length < 5000000) %>%
  filter(GC.... > 65.50 & GC.... < 65.6) %>%
  filter(Genome.fraction.... > 97.0)

N50_hist <- assembly_stats %>% ggplot(aes(x=N50)) +
                  geom_histogram(binwidth = 5000, fill = "steelblue", color = "steelblue") +
                  theme_light()

genome_length_hist <- assembly_stats  %>% ggplot(aes(x=Total.length)) +
  geom_histogram(binwidth = 1000, fill = "steelblue", color = "steelblue") +
  theme_light()

GC_hist <- assembly_stats %>% ggplot(aes(x=GC....)) +
  geom_histogram(binwidth = 0.005, fill = "steelblue", color = "steelblue") +
  theme_light()
```

```
genome_frac_hist <- assembly_stats  %>% ggplot(aes(x=Genome.fraction....)) +
  geom_histogram(binwidth = 0.01, fill = "steelblue", color = "steelblue") +
  theme_light()

(N50_hist | genome_length_hist)/(GC_hist | genome_frac_hist)
```



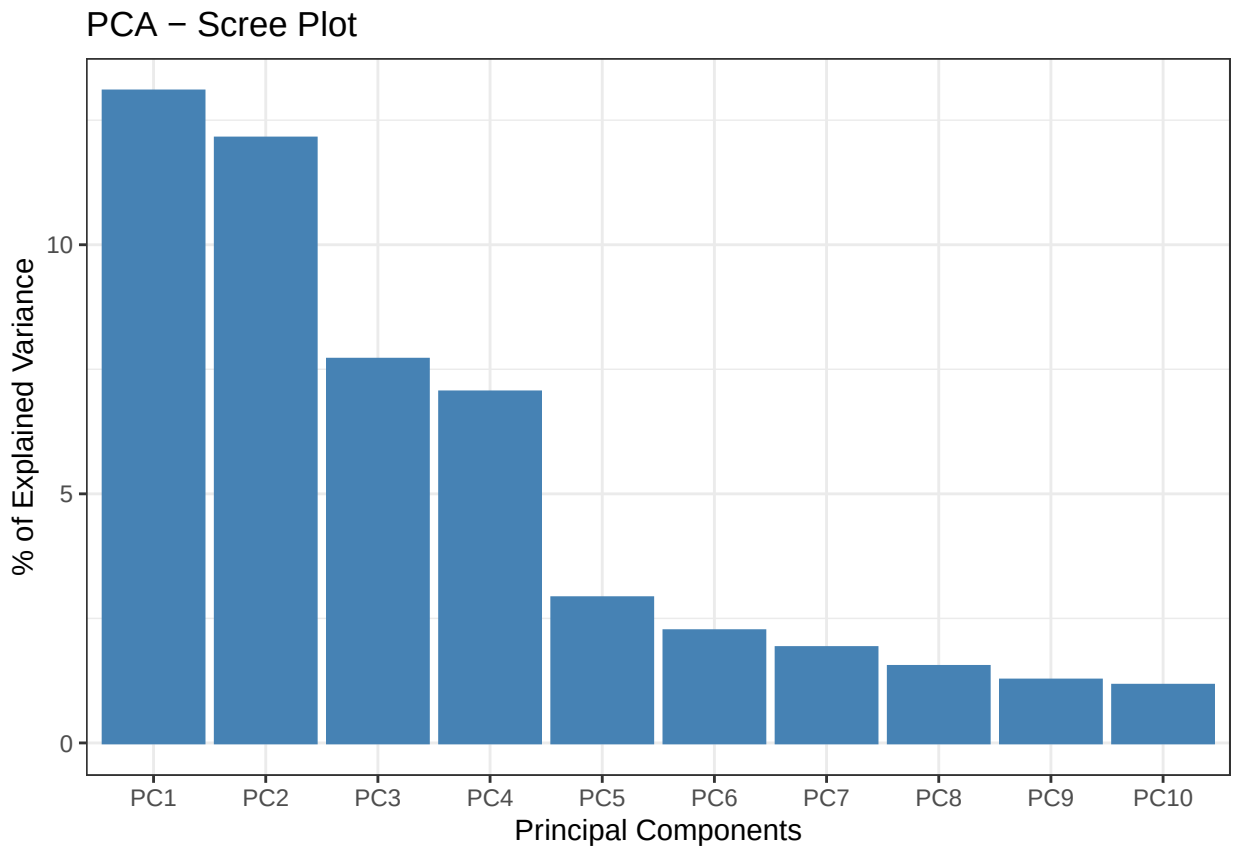

## 3. Accessory Genome Presence/Absence Matrix

```
library(dplyr)
gene_pres_abs <- read.csv("mbovis_prab.csv", header = TRUE, stringsAsFactors = FALSE, row.names = "Gene
accessory_genome <- gene_pres_abs[!(is.na(gene_pres_abs$Accessory.Fragment)),]
core_genome <- gene_pres_abs[is.na(gene_pres_abs$Accessory.Fragment),]

accessory_genome_non_unique <- accessory_genome
auxil <- gene_pres_abs %>% select(2:14)

accessory_pa <- accessory_genome_non_unique %>% select(14:(ncol(accessory_genome_non_unique)))
##edit the gene presence absence to be numeric.
accessory_pa[!(accessory_pa=="")] <- 1
accessory_pa[accessory_pa==""] <- 0
```

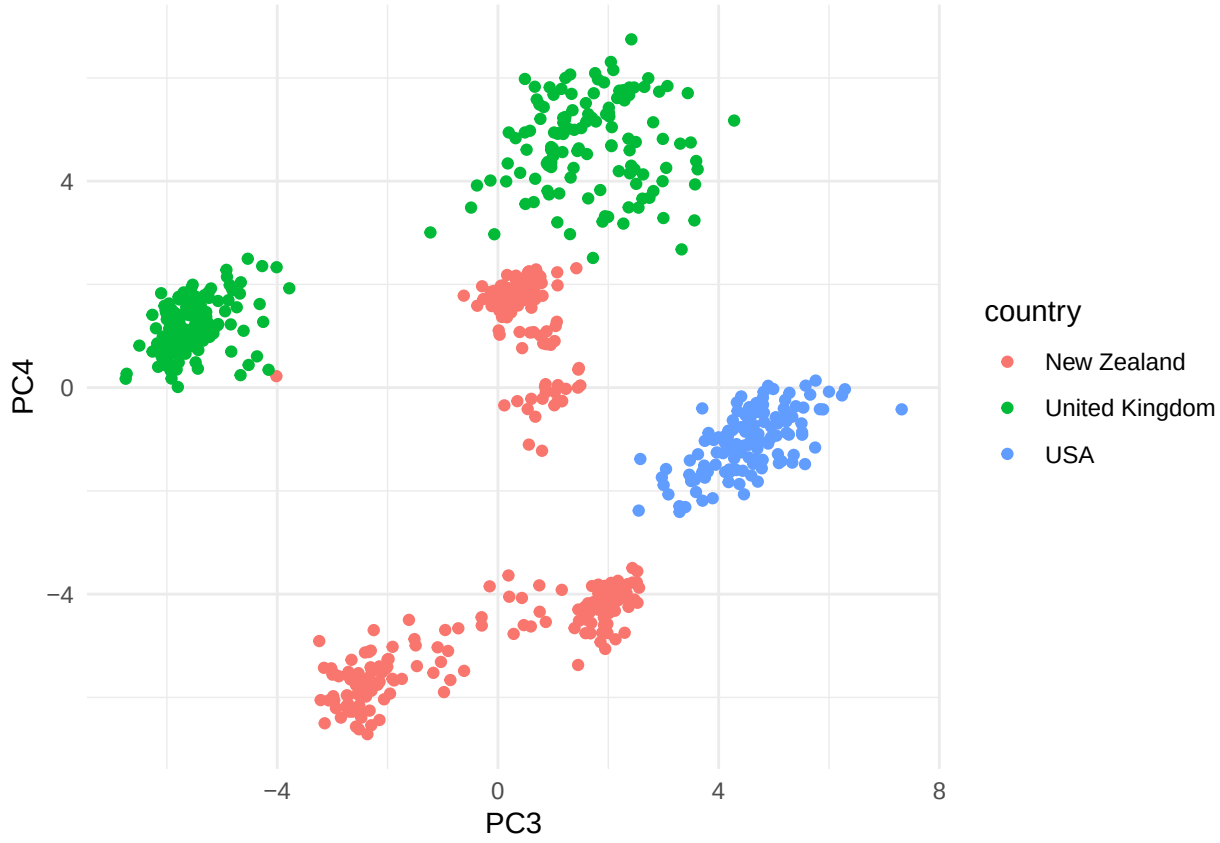
```
##let's transpose the dataframe by turning it into a matrix first.
pa_transpose <- t(data.matrix(accessory_pa))
```

## 4. Principal Component Analysis

```
# for this we just need to convert the full PRAB matrix into a PCA plot. I want to know about the clust
library("ggplot2")
prab_pca <- prcomp(pa_transpose)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(pa_transpose)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_bw()
```

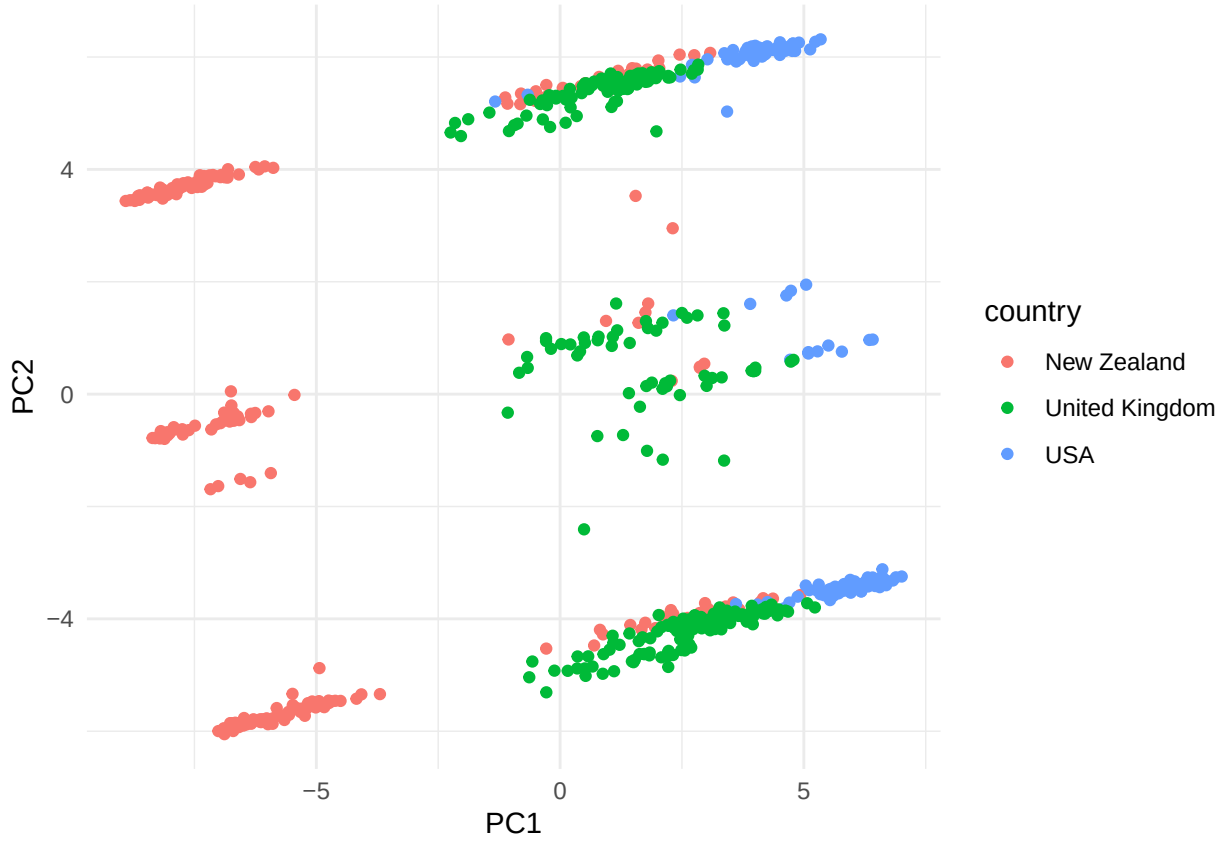


```r
# I am curious how the clustering looks for the top 4 PC's. For the PCoA, we notice some strange cluste

mbov_full_meta <- read.csv("../../../Documents/Mbovis_meta.csv", header = TRUE,stringsAsFactors = FALSE)
mbov_meta <- read.csv("/Users/noah_/Documents/filtered_isolate_list.csv",header = TRUE, stringsAsFactors
upd_mbov_meta <- mbov_meta %>% left_join(mbov_full_meta %>% select(Experiment,Instrument,Center.Name,Co
mbov_meta <- upd_mbov_meta

scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC3,y=PC4, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```



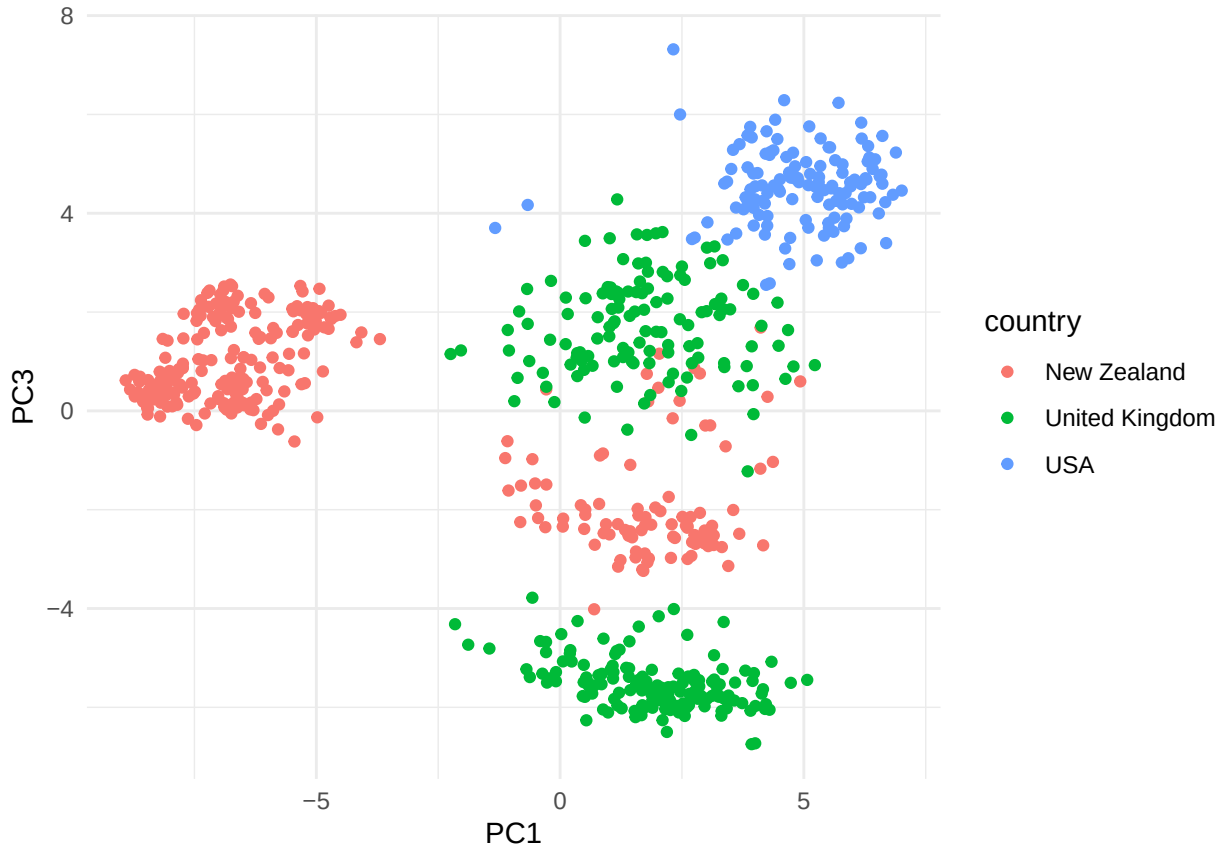

```r
#looks just like the PCoA! so I think I will just work with this hence forth.
```

```r
#Let's explore the 1st two PC's to see what can be done to resolve the clustering confusion.

# Let's relook at PC1 + PC2
ggplot(as.data.frame(scores),aes(x=PC1,y=PC2, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```

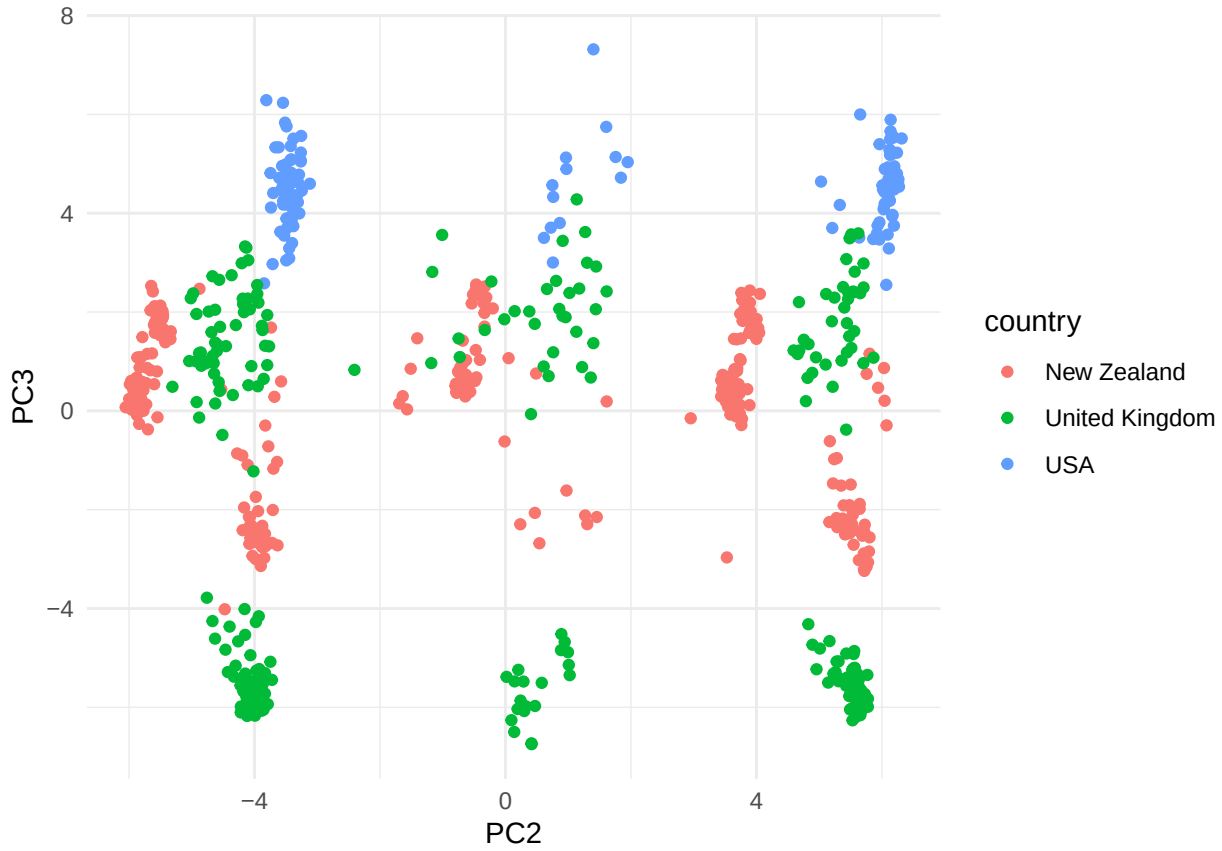



Why is this weird to me? It seems that by looking at the first 2 PC's, the separation of classes just don't make sense. PC1 isn't terrible, it can at least describe the some differences between the points based on country, but on PC2, its saying there are 3 different classes of isolates that are similar across all the countries, even with little to no overlap.

```
ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```

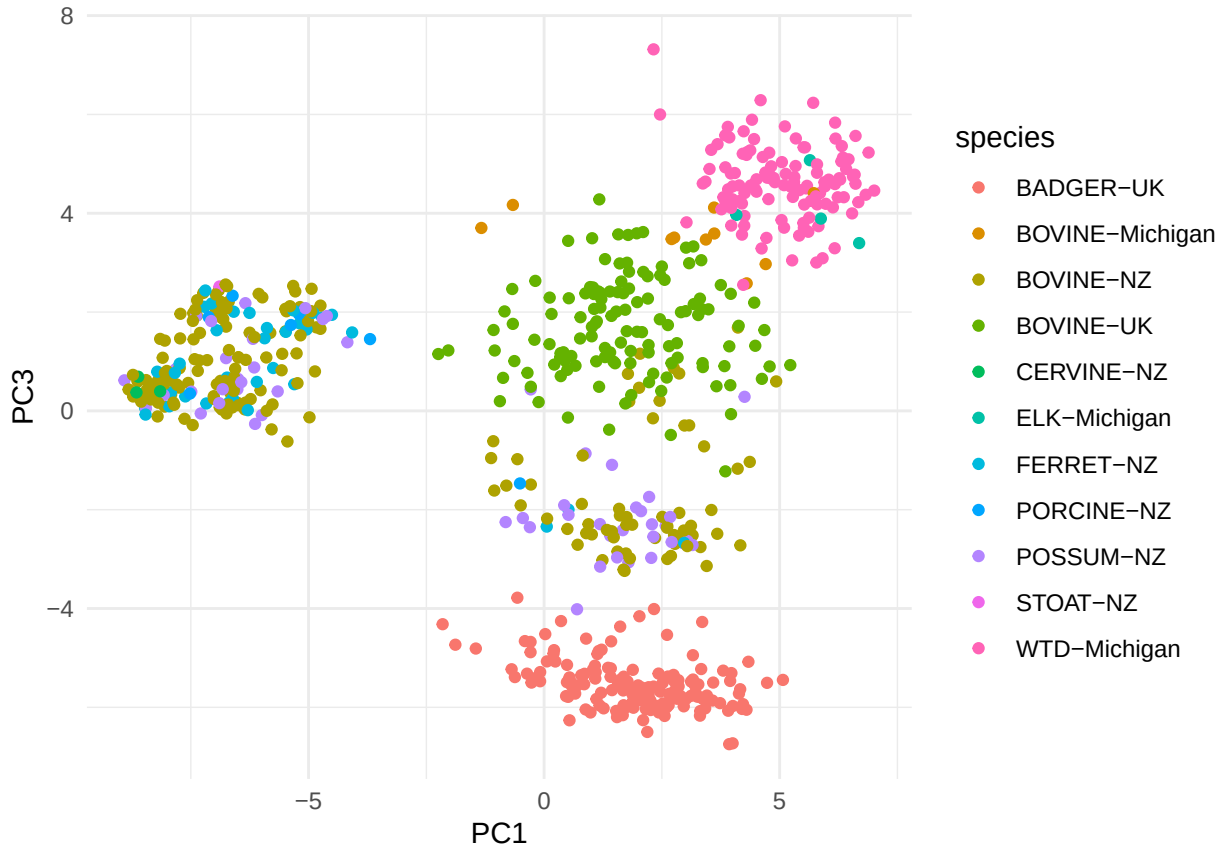



```
ggplot(as.data.frame(scores),aes(x=PC2,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```




It seems that PC2 is the strange actor in this analysis. I observed this during the summer and just opted to remove PC's 1 & 2. Without knowing too much however, I think it might be harder to explain removing both PC's so I will argue for keeping PC 1, and then analyzing PC3. However, the clustering seems to be amazing for PC 3 & 4. So what should be done? I will use Variable loadings with the PC's so this might be an important decision (but not one I want to belabor the point over)

```
ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = species, col = species)) +
  geom_point() +
  theme_minimal()
```



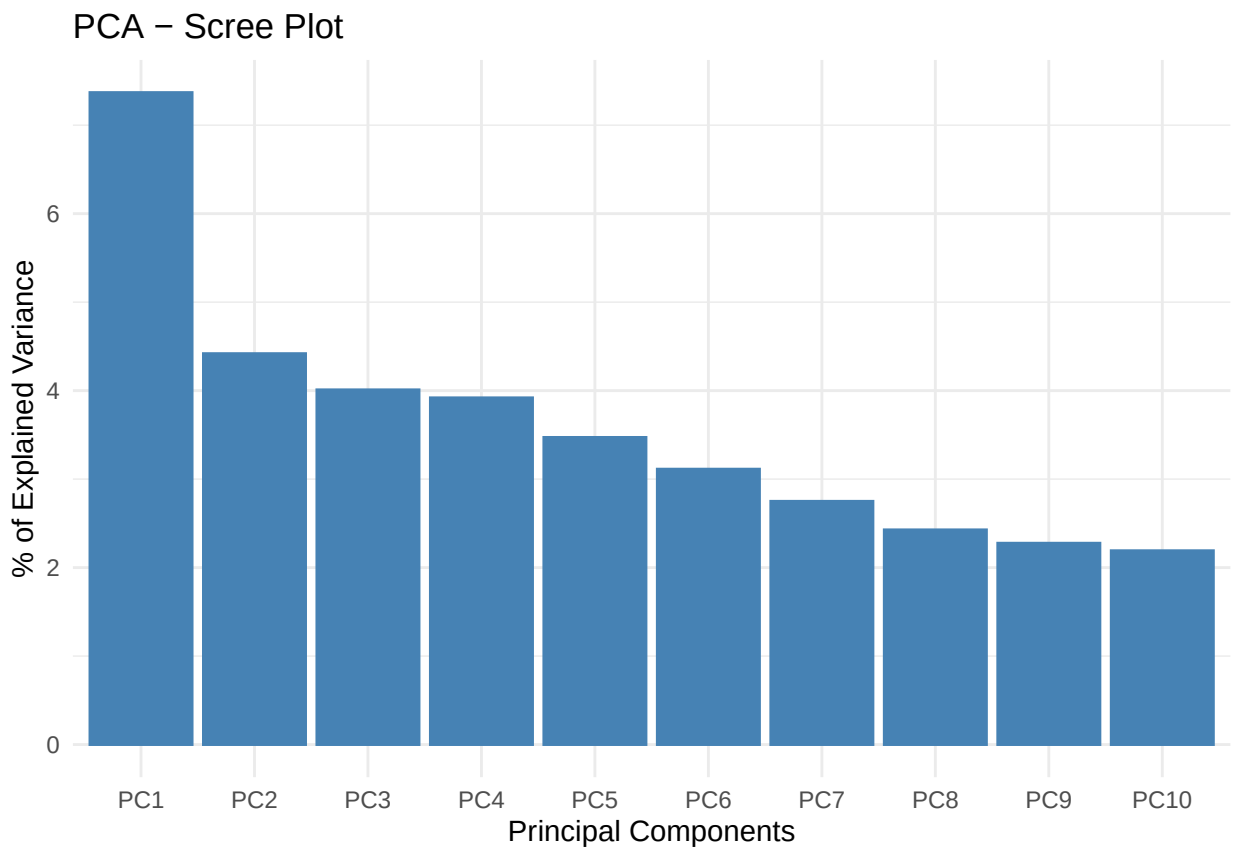

Next I wonder how the clustering looks given different thresholds of accessory genome presence. What part of the accessory genome explains the clustering that we see?

```
soft_core_genes <- subset(accessory_genome,No..isolates > 665 & No..isolates < 693) %>% select(14:(ncol
soft_core_genes[!(soft_core_genes=="")] <- 1
soft_core_genes[soft_core_genes==""] <- 0
soft_core_prab <- t(data.matrix(soft_core_genes))

shell_genes <- subset(accessory_genome,No..isolates > 105 & No..isolates < 665) %>% select(14:(ncol(acc
shell_genes[!(shell_genes=="")] <- 1
shell_genes[shell_genes==""] <- 0
shell_prab <- t(data.matrix(shell_genes))

cloud_genes <- subset(accessory_genome,No..isolates > 0 & No..isolates < 105) %>% select(14:(ncol(acces
cloud_genes[!(cloud_genes=="")] <- 1
cloud_genes[cloud_genes==""] <- 0
cloud_prab <- t(data.matrix(shell_genes))
```

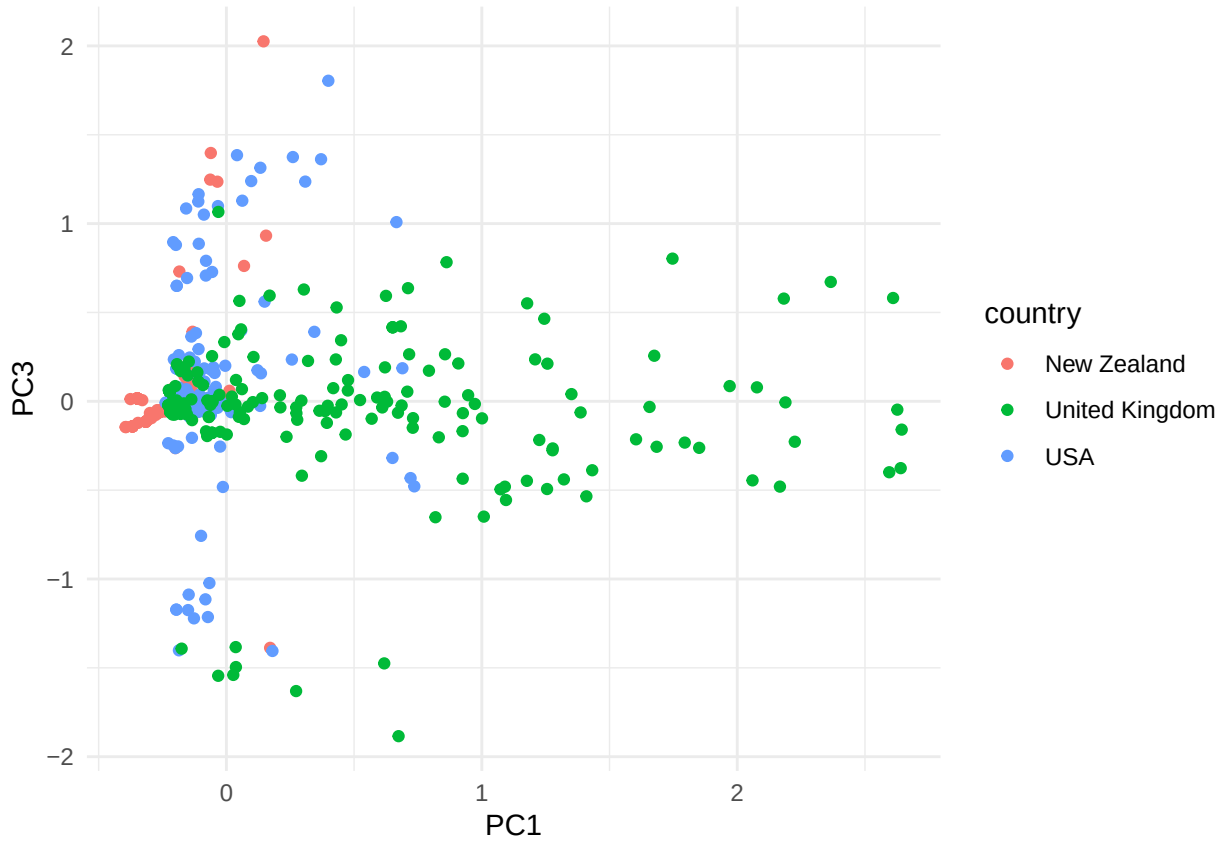
Now we will just redo the analysis we did with all the data:

```
prab_pca <- prcomp(soft_core_prab)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(soft_core_prab)
scores <- prab_pca$x
```

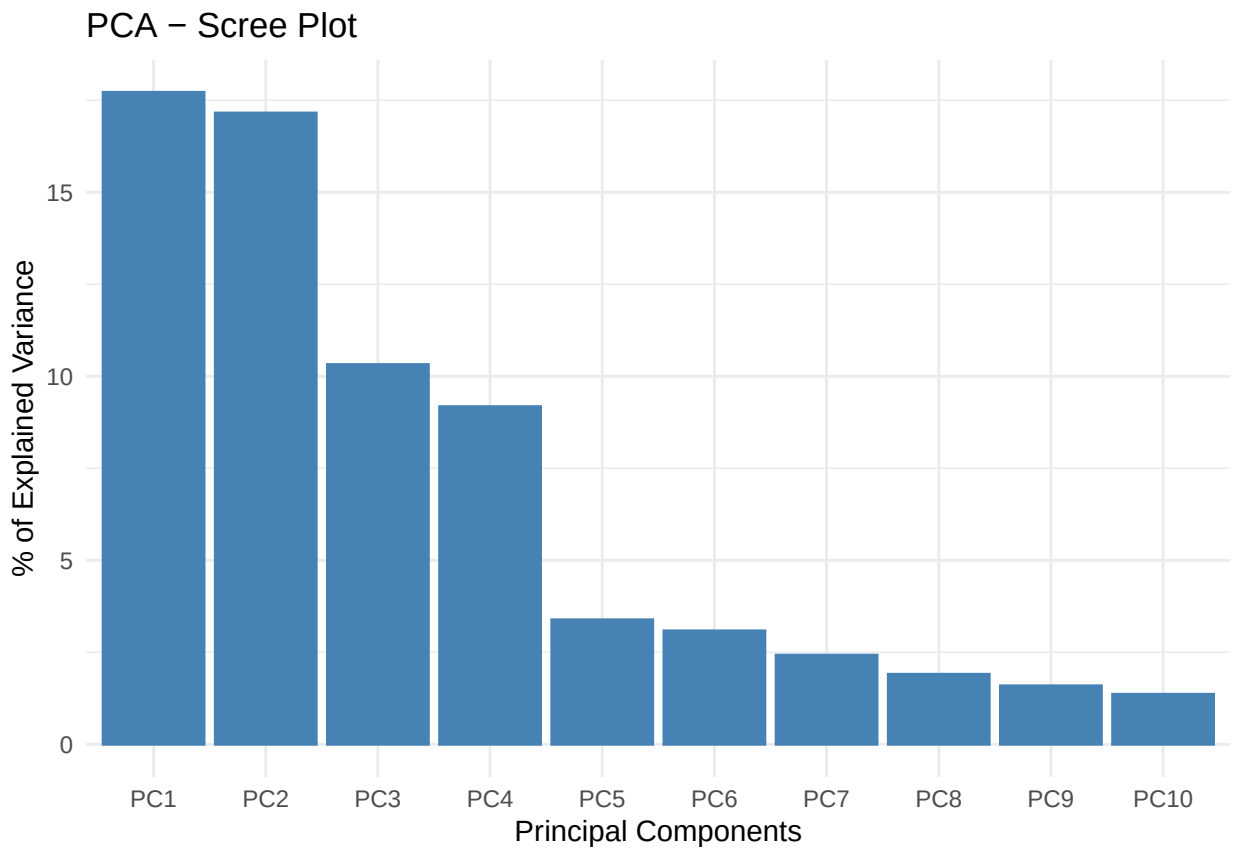
```
varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```



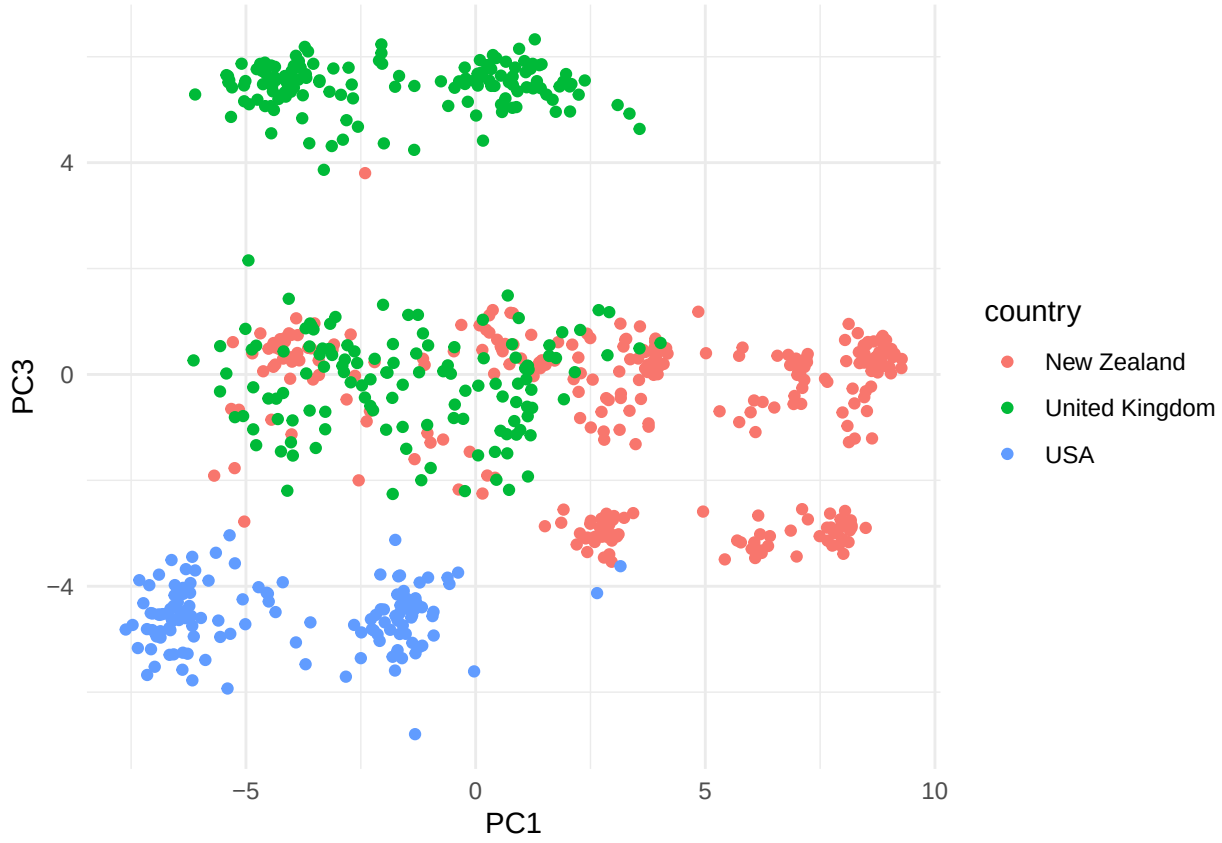

```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```



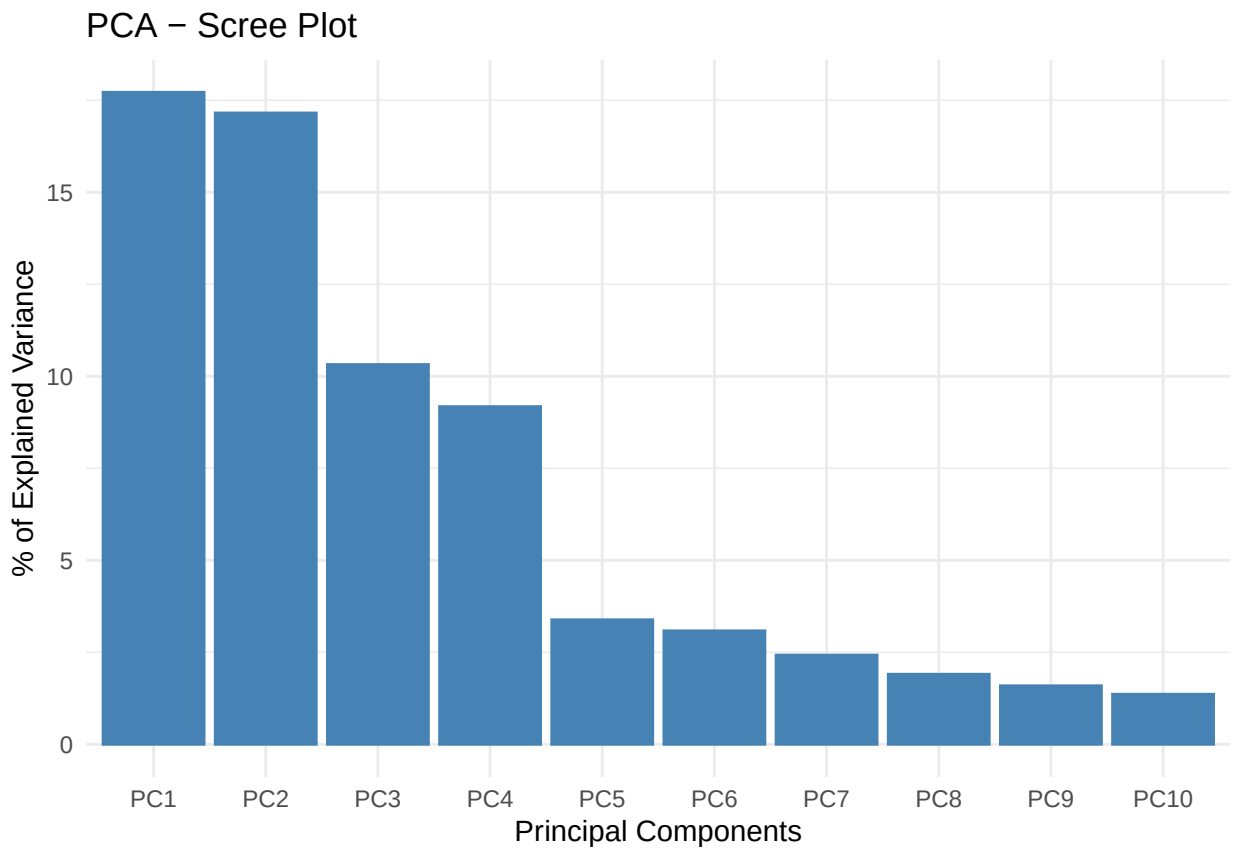

```
prab_pca <- prcomp(shell_prab)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(shell_prab)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```



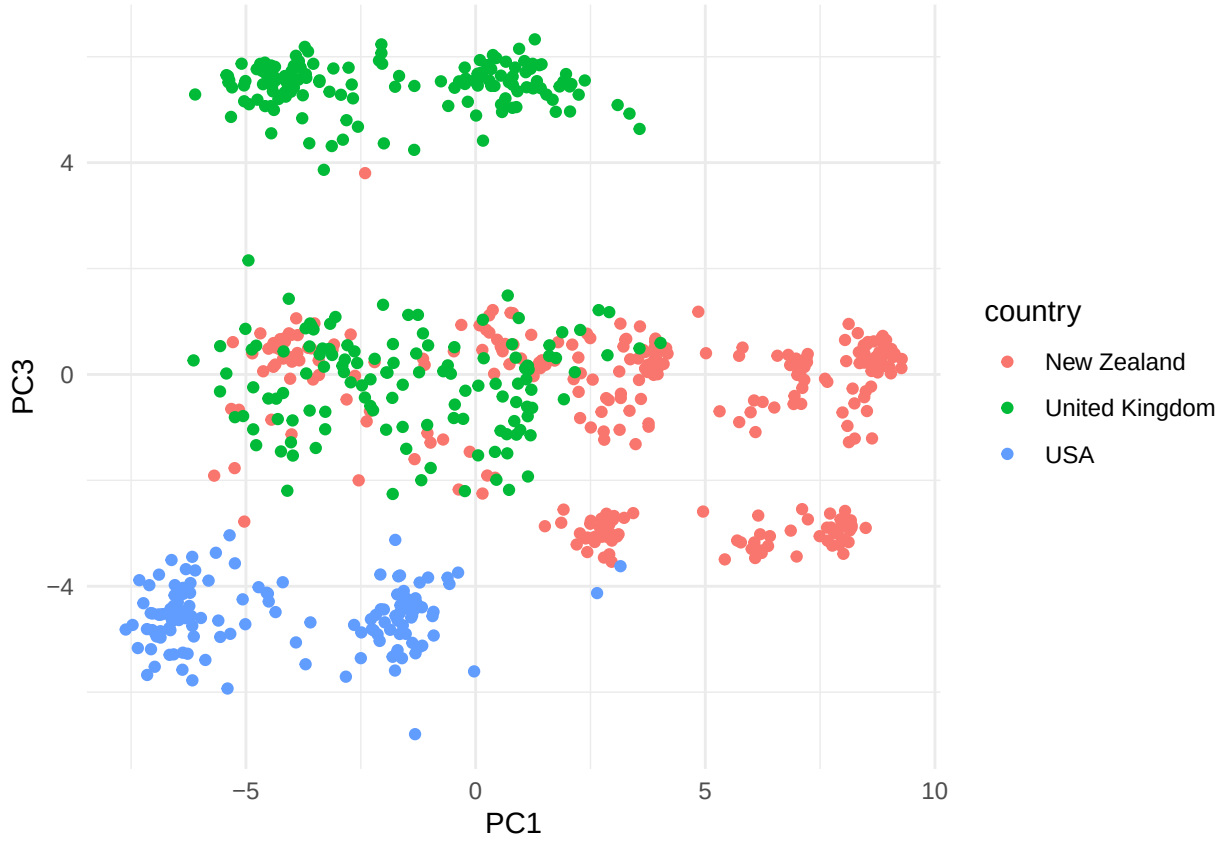

```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```




```
prab_pca <- prcomp(cloud_prab)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(cloud_prab)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```




```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```

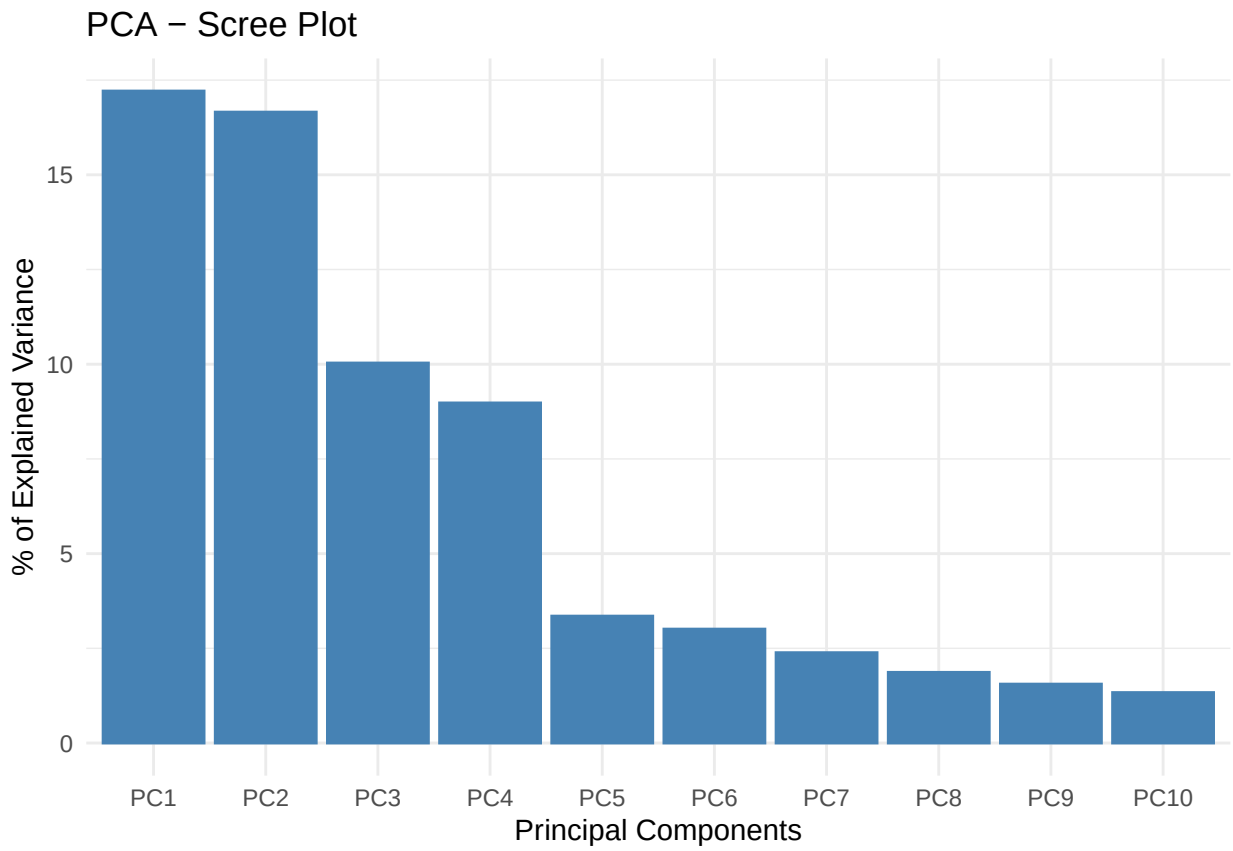



So the soft core genes don't give us enough information to make clusters, the shell and cloud genes seem to be better at making this distinction, and they are identical in their output. Maybe we should see what happens if we exlude one (cloud or shell) from the analysis.

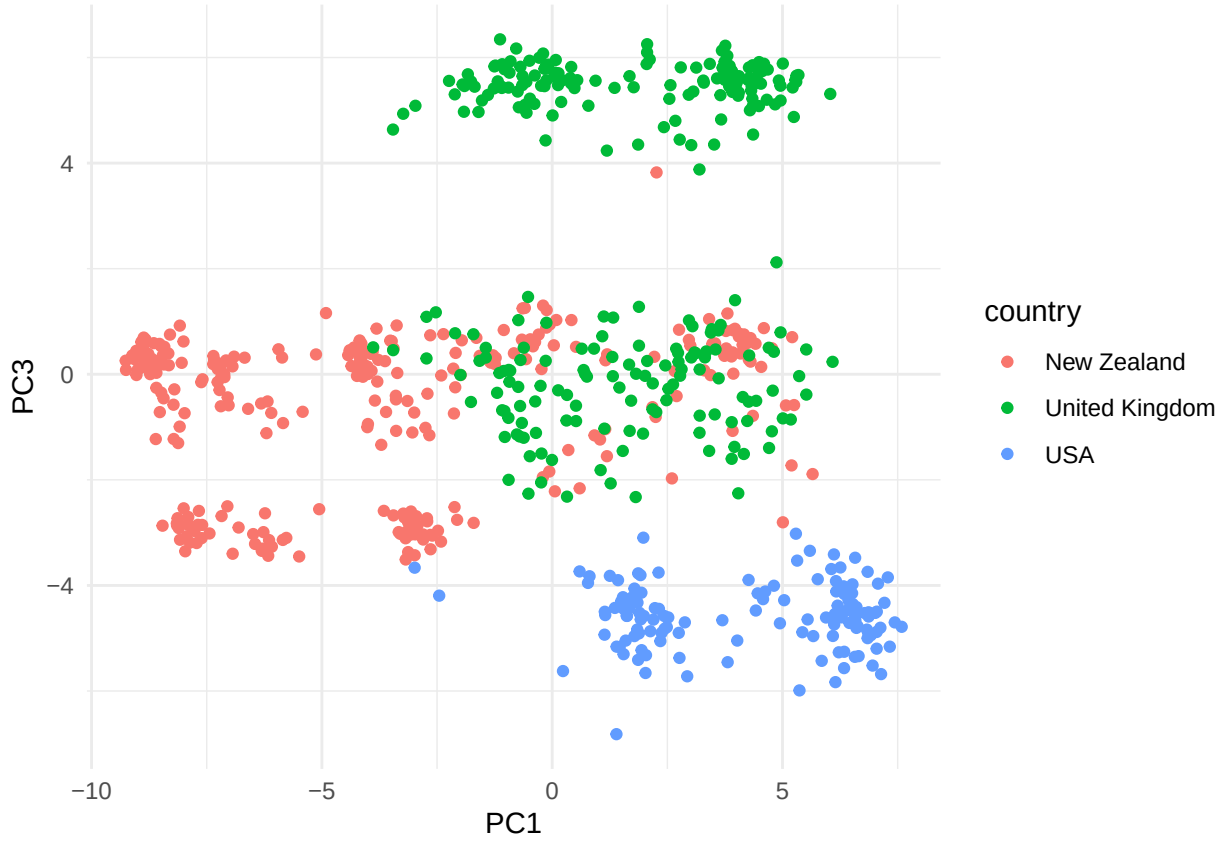
```
soft_shell <- subset(accessory_genome,No..isolates > 105 & No..isolates < 693) %>% select(14:(ncol(acces
soft_shell[!(soft_shell=="")] <- 1
soft_shell[soft_shell==""] <- 0
soft_shell <- t(data.matrix(soft_shell))
soft_cloud <- subset(accessory_genome,(No..isolates > 0 & No..isolates < 105) | (No..isolates > 665 & N
soft_cloud[!(soft_cloud=="")] <- 1
soft_cloud[soft_cloud==""] <- 0
soft_cloud <- t(data.matrix(soft_cloud))
shell_cloud <- subset(accessory_genome,No..isolates > 0 & No..isolates < 665) %>% select(14:(ncol(acces
shell_cloud[!(shell_cloud=="")] <- 1
shell_cloud[shell_cloud==""] <- 0
shell_cloud <- t(data.matrix(shell_cloud))
```

```
prab_pca <- prcomp(soft_shell)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(soft_shell)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
```

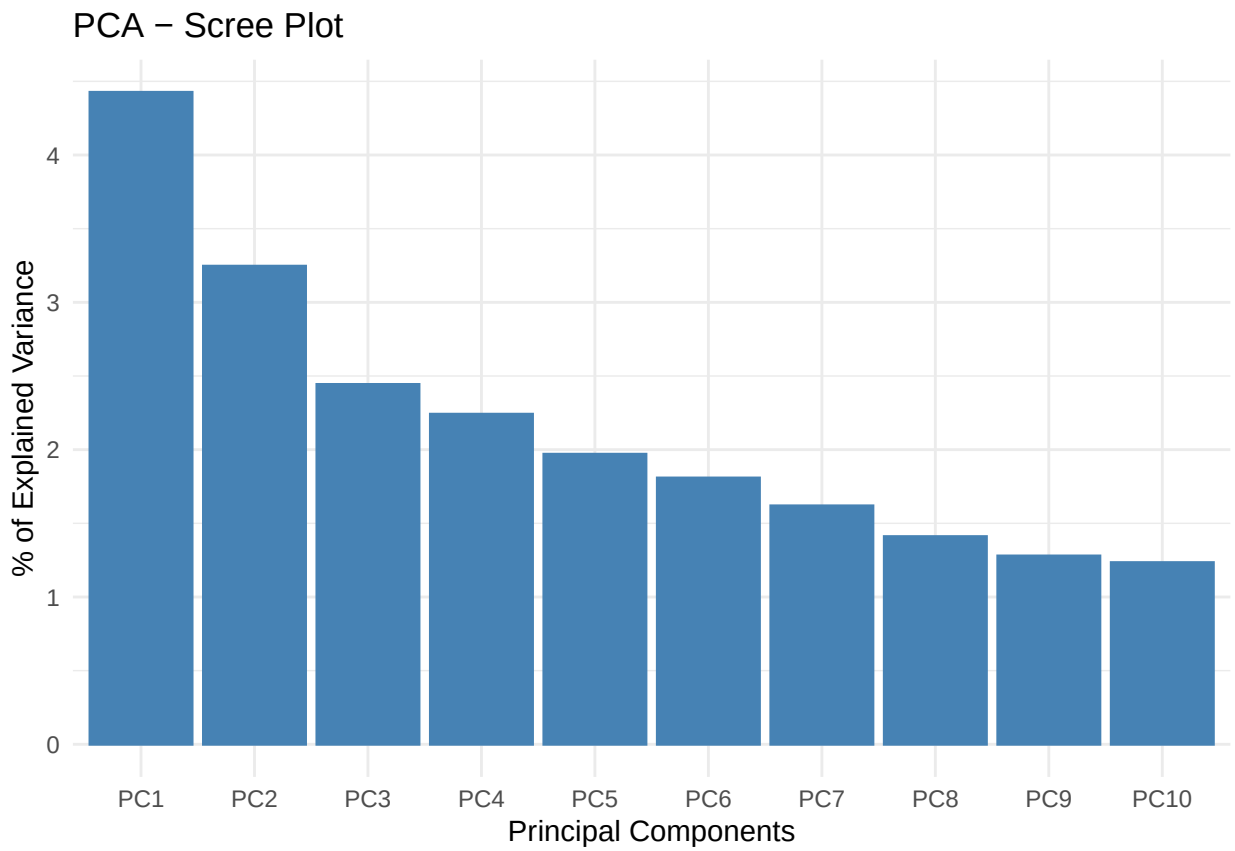
```
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```



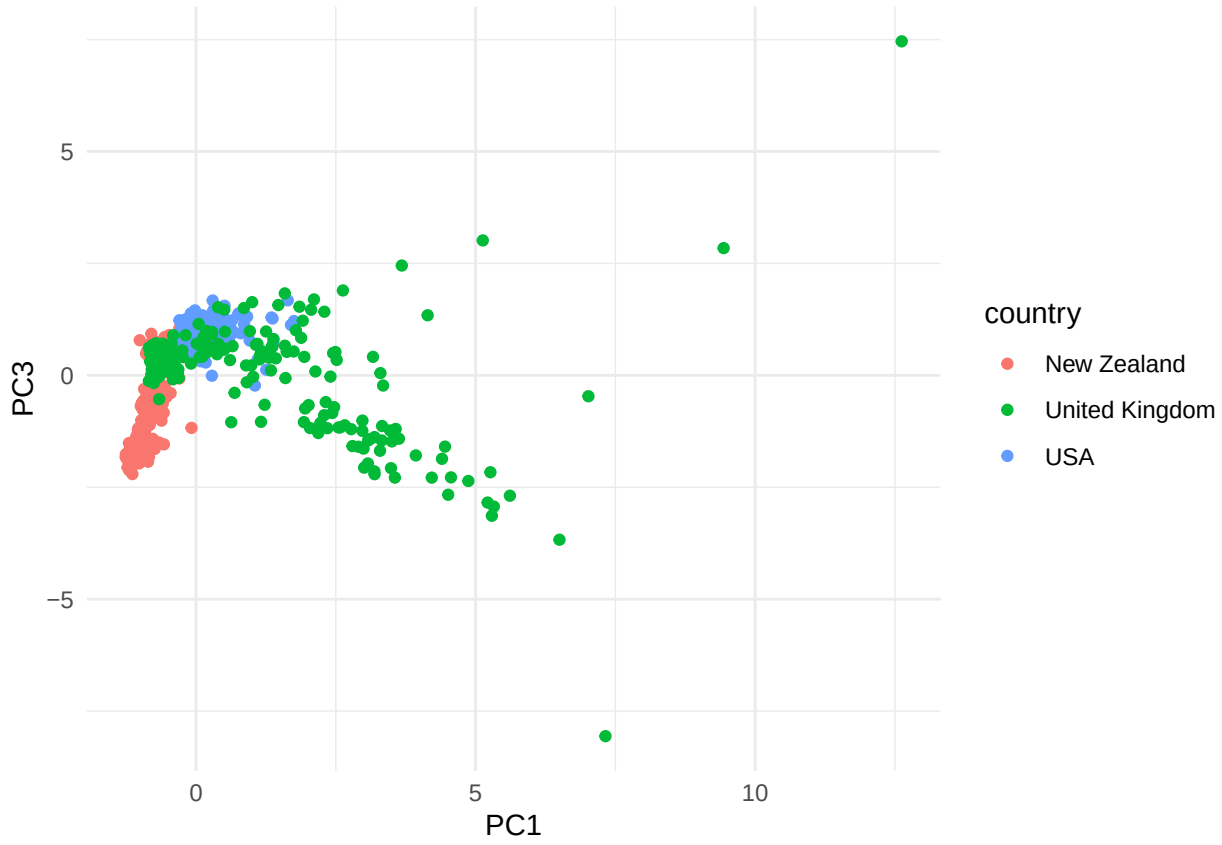

```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```



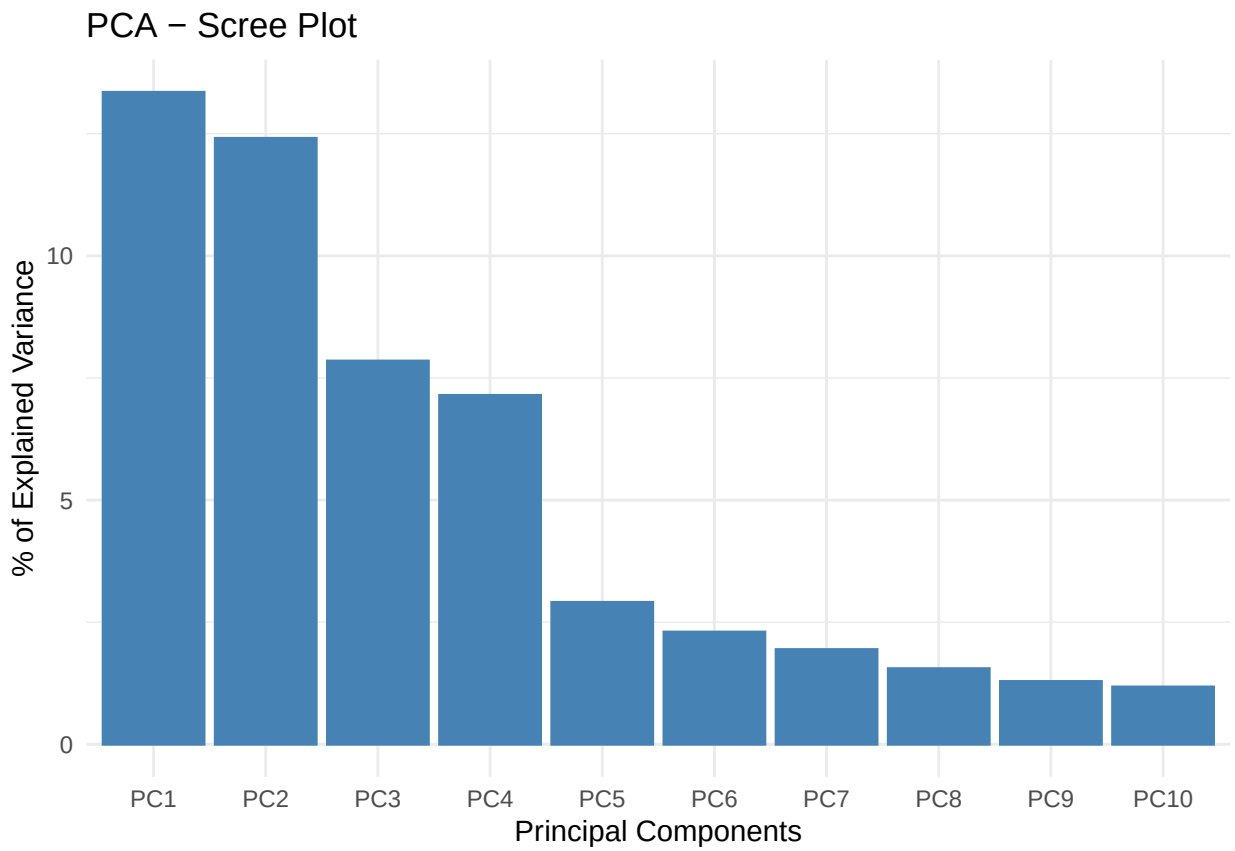

```
prab_pca <- prcomp(soft_cloud)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(soft_cloud)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```



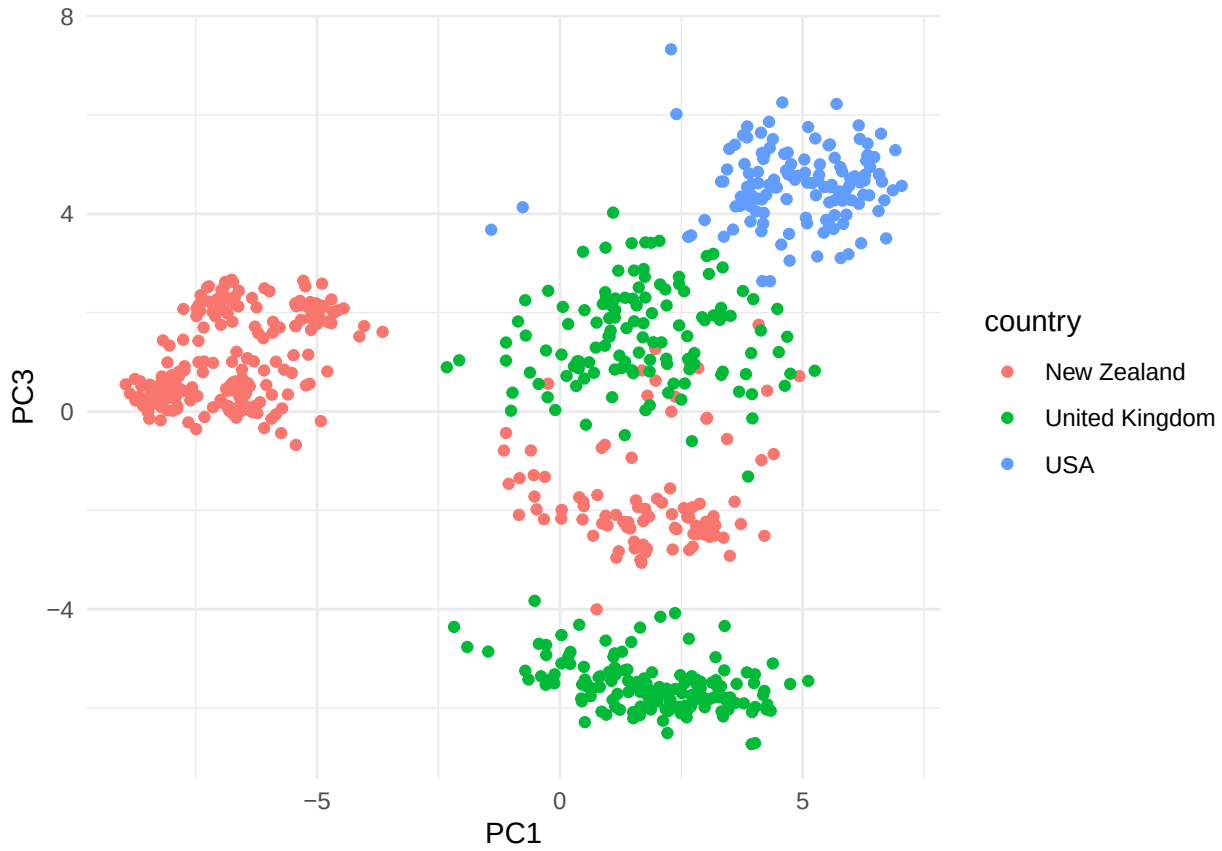

```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```




```
prab_pca <- prcomp(shell_cloud)
variance <- (prab_pca$sdev)^2
loadings <- prab_pca$rotation
rownames(loadings) <- colnames(shell_cloud)
scores <- prab_pca$x

varPercent <- variance/sum(variance) * 100
prin_comp <- paste0("PC",1:10)
data <- data.frame(prin_comp,varPercent[1:10])
ggplot(data, aes(x=reorder(prin_comp, -varPercent[1:10]), y=varPercent[1:10], )) +
  geom_bar(stat = "identity",fill='steelblue', color="steelblue") +
  labs(x = "Principal Components", y= "% of Explained Variance") +
  ggtitle("PCA - Scree Plot") +
  theme_minimal()
```

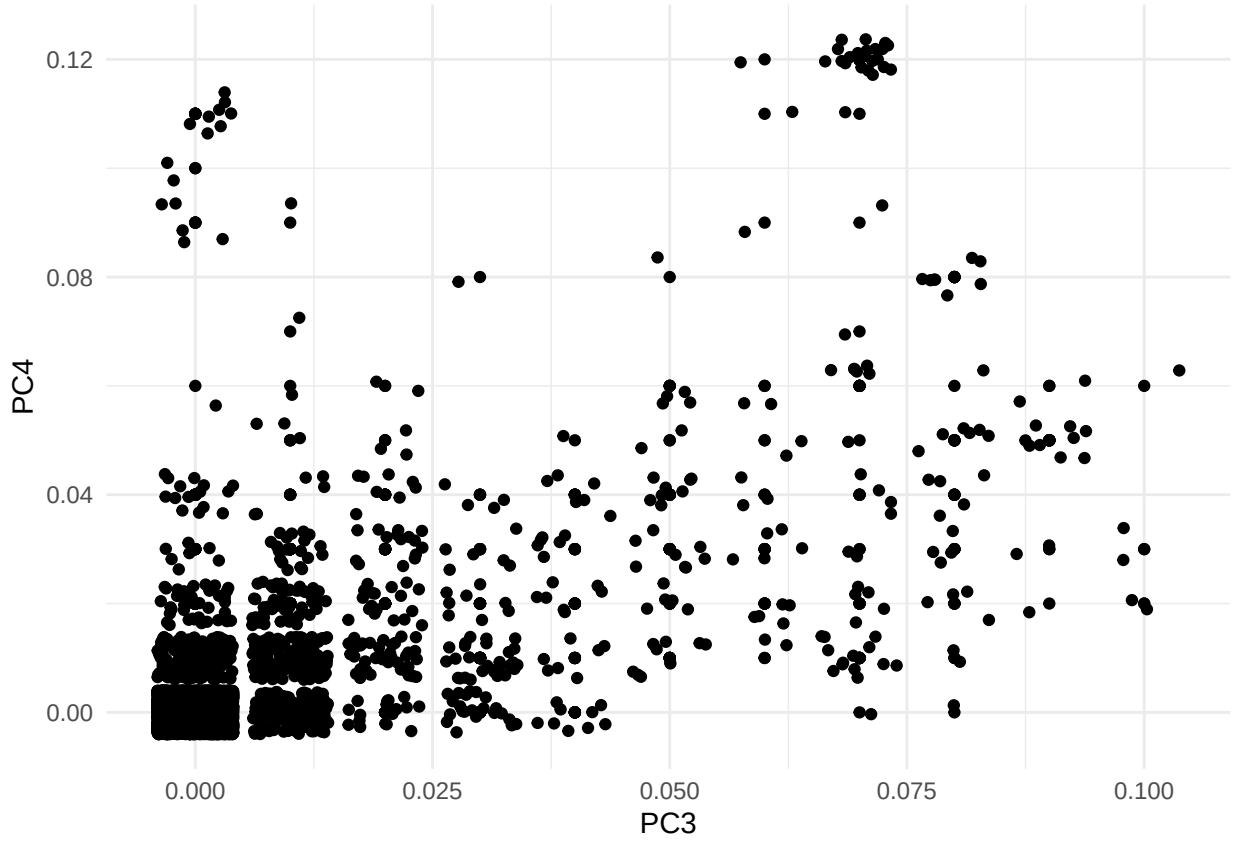



```
scores <- as.data.frame(scores[,1:4])
scores$country <- mbov_meta$Country
scores$species <- mbov_meta$Species

ggplot(as.data.frame(scores),aes(x=PC1,y=PC3, fill = country, col = country)) +
  geom_point() +
  theme_minimal()
```

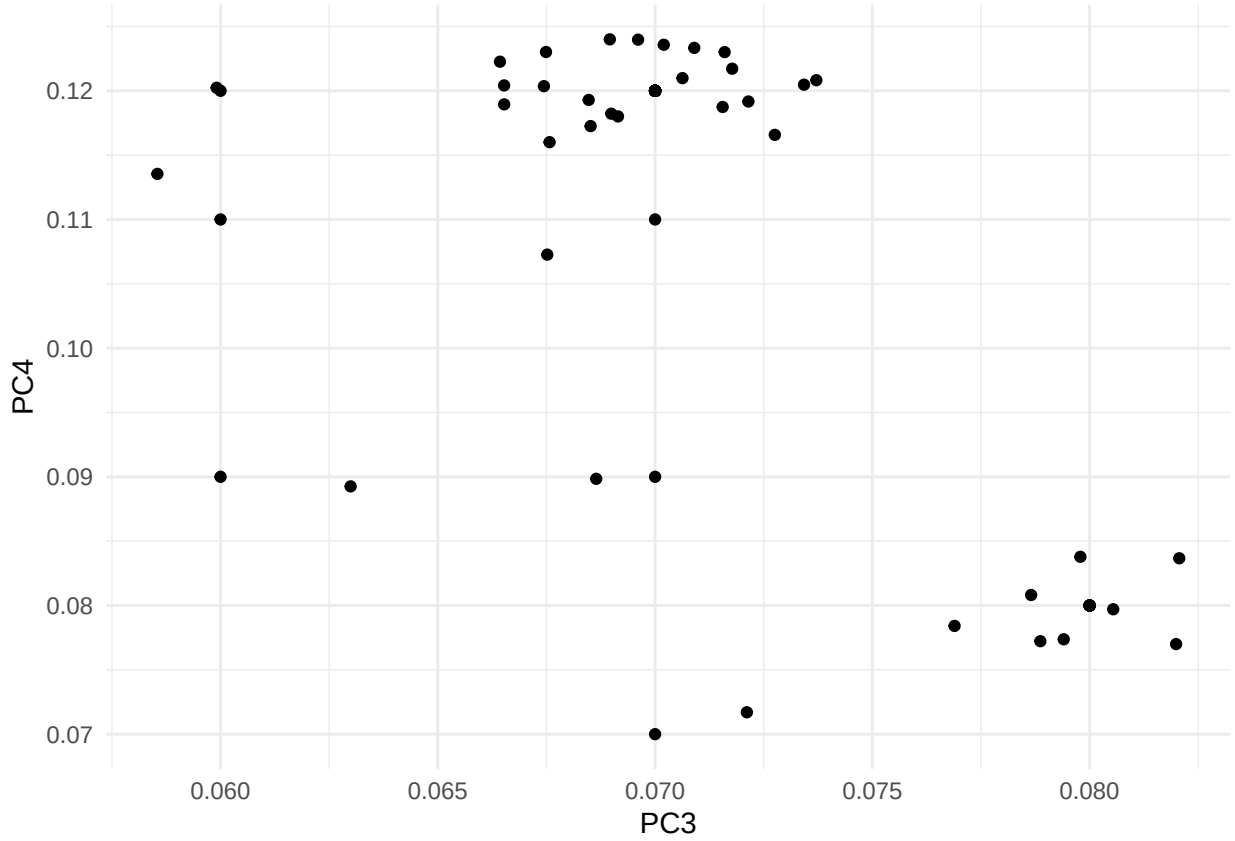



These plots are intriguing because the clustering we see with all the is closest when we include the shell genes! so for this type of analysis, shell genes provides the most information in order to cluster using PCA. My interpretation: The shell genes carry the most variation in presence/absence, so are the most useful for the analysis. Next, the cloud genes provide the necessary information to improve this clustering (the unique genes.)

```
loadings <- prab_pca$rotation
loadings <- data.frame(round(abs(loadings[,3:4]), 2))

ggplot(loadings,aes(x=PC3,y=PC4)) +
  geom_point() +
  geom_jitter() +
  theme_minimal()
```




```
#I'm searching for ways to conduct loading analysis, BUT FOR NOW, I will just get the higher values. To

filtered_loadings <- subset(loadings, PC3 > 0.05 & loadings$PC4 > 0.06)

ggplot(filtered_loadings,aes(x=PC3,y=PC4)) +
  geom_point() +
  geom_jitter() +
  theme_minimal()
```




```
key_COGs <- data.frame(filtered_loadings,stringsAsFactors = FALSE)
key_COGs$COG <- rownames(key_COGs)
```

```
library(dplyr)

#now, let's see how these COGs relate to genome annotations
auxil$COG <- rownames(auxil)
key_COGS <- key_COGs %>% left_join(auxil,by = "COG")
key_COGS$Annotation
```

```
##  [1] "[gene=fadE18] [locus_tag=BQ2027_MB1968C] [db_xref=GOA:A0A1R3XZR1,InterPro:IPR009075,InterPro:I
##  [2] "hypothetical protein"
##  [3] "[locus_tag=BQ2027_MB1966C] [db_xref=GOA:A0A1R3XZS8,InterPro:IPR002818,InterPro:IPR009057,Interl
##  [4] "[gene=fadE17] [locus_tag=BQ2027_MB1969C] [db_xref=GOA:A0A1R3XZU2,InterPro:IPR006091,InterPro:I
##  [5] "[gene=tpx] [locus_tag=BQ2027_MB1967] [db_xref=GOA:P66953,InterPro:IPR002065,InterPro:IPR013740
##  [6] "hypothetical protein"
##  [7] "[locus_tag=BQ2027_MB1963C] [db_xref=GOA:A0A1R3XZS7,InterPro:IPR002347,InterPro:IPR020904,Interl
##  [8] "hypothetical protein"
##  [9] "[locus_tag=BQ2027_MB1013] [db_xref=GOA:A0A1R3XZ45,InterPro:IPR003838,InterPro:IPR025857,UniPro
## [10] "hypothetical protein"
## [11] "hypothetical protein"
## [12] "[locus_tag=BQ2027_MB1603C] [db_xref=InterPro:IPR006433,UniProtKB/TrEMBL:A0A1R3XYR4] [protein=P
## [13] "[locus_tag=BQ2027_MB1607C] [db_xref=InterPro:IPR036869,UniProtKB/TrEMBL:A0A1R3XYP8] [protein=P
## [14] "[locus_tag=BQ2027_MB1602C] [db_xref=InterPro:IPR024455,UniProtKB/TrEMBL:A0A1R3XZ36] [protein=P
## [15] "[locus_tag=BQ2027_MB1609C] [db_xref=InterPro:IPR024384,UniProtKB/TrEMBL:A0A1R3Y0W3] [protein=P
```

```
## [16] "[locus_tag=BQ2027_MB1599] [db_xref=UniProtKB/TrEMBL:A0A1R3Y0V5] [protein=Probable phiRV1 phage
## [17] "[locus_tag=BQ2027_MB1604C] [db_xref=InterPro:IPR006448,InterPro:IPR022357,UniProtKB/TrEMBL:A0A
## [18] "[locus_tag=BQ2027_MB1605C] [db_xref=UniProtKB/TrEMBL:A0A1R3XYR0] [protein=Probable phiRv1 phag
## [19] "[locus_tag=BQ2027_MB1606C] [db_xref=UniProtKB/TrEMBL:A0A1R3XYR2] [protein=Probable phiRv1 phag
## [20] "[locus_tag=BQ2027_MB1608C] [db_xref=InterPro:IPR006500,InterPro:IPR014015,InterPro:IPR014818,In
## [21] "[locus_tag=BQ2027_MB1610C] [db_xref=UniProtKB/TrEMBL:A0A1R3XZN2] [protein=Possible phiRv1 phag
## [22] "[locus_tag=BQ2027_MB1611C] [db_xref=UniProtKB/TrEMBL:A0A1R3XYR1] [protein=Possible phage phiRv
## [23] "hypothetical protein"
## [24] "hypothetical protein"
## [25] "hypothetical protein"
## [26] "[locus_tag=BQ2027_MB1540] [db_xref=InterPro:IPR023296,UniProtKB/TrEMBL:A0A1R3Y0Q2] [protein=HYP
## [27] "[gene=glnQ] [locus_tag=BQ2027_MB2593] [db_xref=GOA:P63402,InterPro:IPR000595,InterPro:IPR003439
## [28] "Putative thiosulfate sulfurtransferase SseB"
## [29] "[gene=atsAa] [locus_tag=BQ2027_MB0731] [db_xref=GOA:A0A1R3XW61,InterPro:IPR000917,InterPro:IPR0
## [30] "hypothetical protein"
## [31] "hypothetical protein"
## [32] "[locus_tag=BQ2027_MB2836] [db_xref=GOA:A0A1R3Y486,InterPro:IPR001584,InterPro:IPR012337,InterPr
## [33] "hypothetical protein"
## [34] "hypothetical protein"
## [35] "[gene=wag22a] [locus_tag=BQ2027_MB1790C] [db_xref=InterPro:IPR000084,UniProtKB/Swiss-Prot:P0A68
## [36] "hypothetical protein"
```

# 5. Homologous Recombination

# 6. Positive Selection