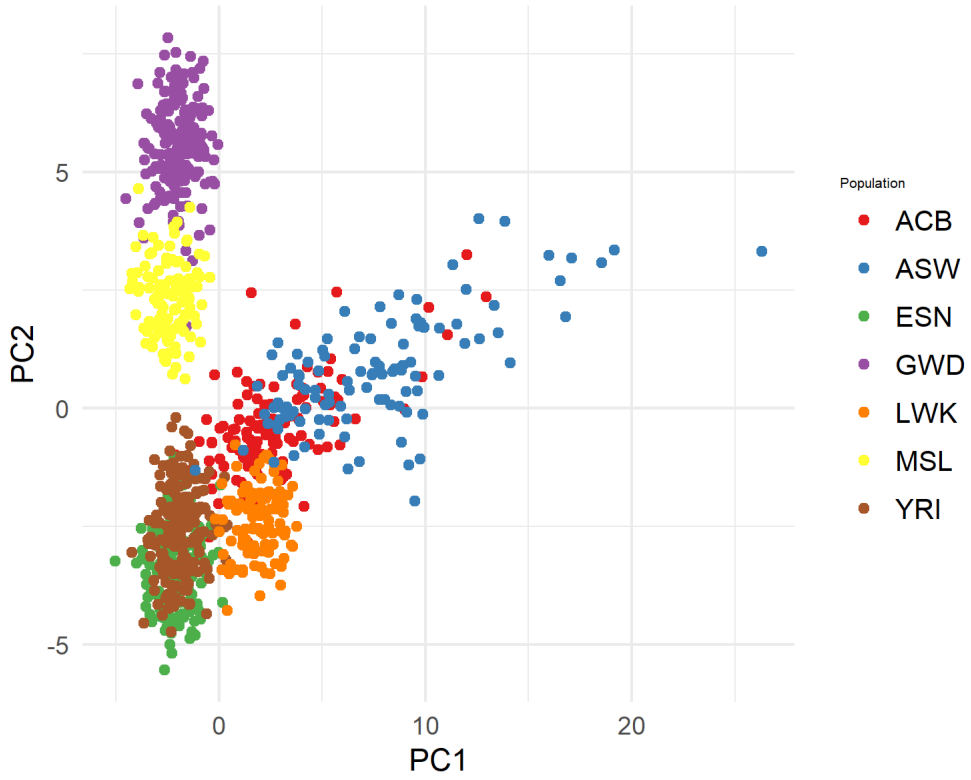
Math308 A2 Coding

Jiahang Wang (261011319)

Part a

The Dimension of each Eigenvector is 10101×1

Part b

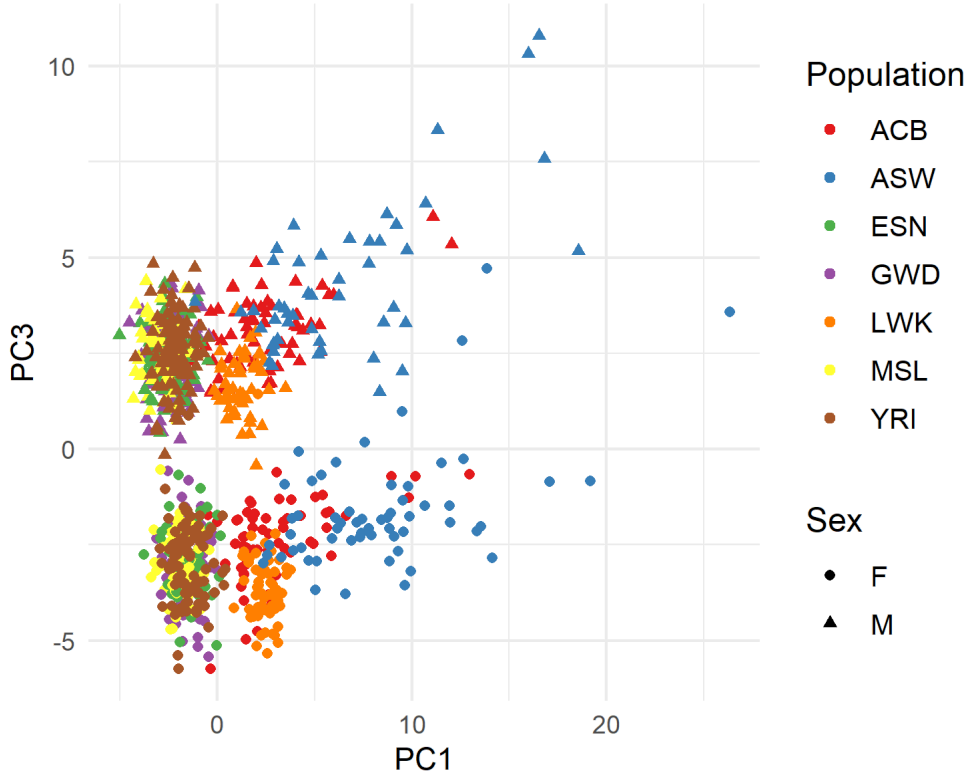


PCA score b1 vs. b2 by population

Part c

The plot shows a clear clustering by population, which suggests that the first two components capture significant genetic variation that corresponds to some separations among the them. PC1 accounts for the most variation and approximately represent differences that align with the latitude and longitude of the region of each population. PC2 may reflect the diverse historical migrations and socio-cultural distinctions among groups, approximately categorized by transatlantic movements, intra-African migrations from West and East Africa, and unique, localized cultural developments within Africa.

Part d

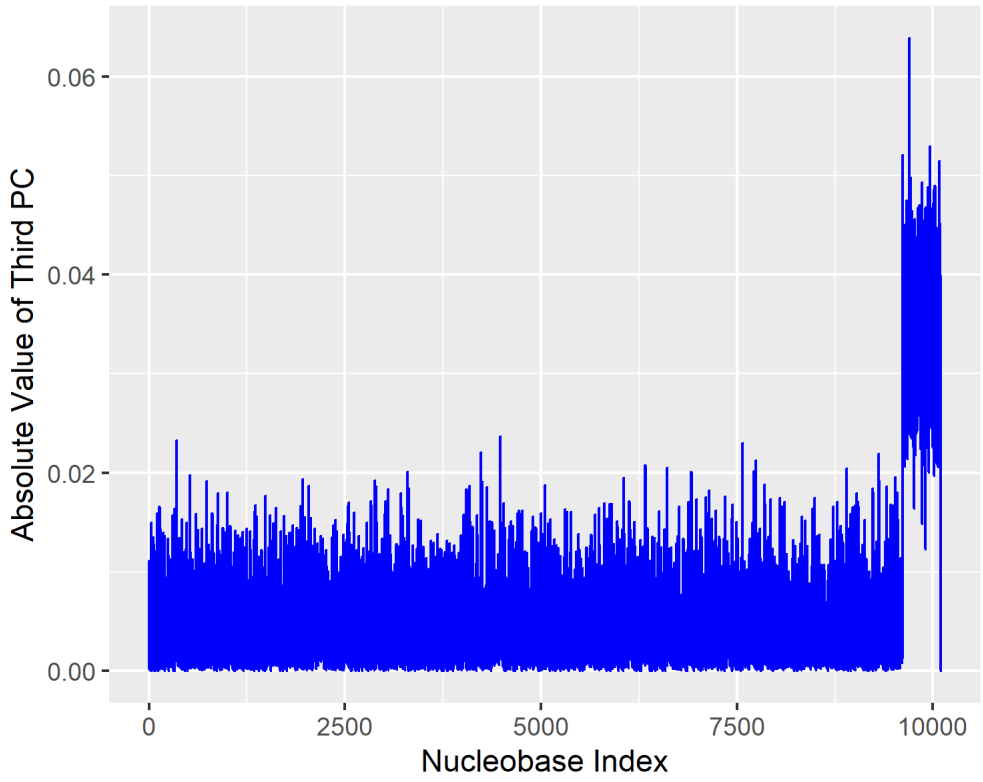


PCA score b1 vs. b3 by population and sex

Part e

After labeling the data by sex we can see a strong separation between clusters, thus the third principal component captures the gender information or difference in different population.

Part f



Nucleobase Index vs Absolute Value of PC3

Since PC3 captures the information of sex, the spike in the PCA3 graph's tail indicates a pronounced variance in genetic data that correlates with the sex of the individuals, likely reflecting the distinctive genetic information carried by the sex chromosomes(the last pair of chromosomes), which are known to differ markedly between males and females.

Code

```
library(ggplot2)
library(dplyr)
```

Preprocessing

Load data

```
# read
data <- read.table("p4dataset2023.txt", header = FALSE, sep = " ")
colnames(data)[1:3] <- c("Id", "Sex", "Population")

# view
# data[1:5,]
```

Delete first 3 columns

```
data1 <- data[,4:ncol(data)]
# view
# data1[1:5,]
```

Convert the data

```
# Function to calculate mode
getMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Initialize the nucleobases dataframe with zeros
nucleobases <- data.frame(matrix(ncol = ncol(data1), nrow = nrow(data1)))
colnames(nucleobases) <- colnames(data1)

# Compute mode for each column and assign 0 or 1 to nucleobases
for (j in 1:ncol(data1)) {
  mode_nucleobase <- getMode(data1[[j]])
  nucleobases[[j]] <- ifelse(data1[[j]] == mode_nucleobase, 0, 1)
}

nucleobases <- as.data.frame(lapply(nucleobases, as.integer))

# nucleobases[1:5,]
```

Centering the data

```
# Center each column without scaling
centered_nucleobases <- as.data.frame(lapply(nucleobases, function(x) x - mean(x)))

# centered_nucleobases[1:5,]
```

PCA

a

```
# Perform PCA using prcomp
pca_result <- prcomp(centered_nucleobases, scale. = FALSE, center = FALSE)



# Eigenvalues
eigenvalues <- pca_result$sdev^2

# Eigenvectors
eigenvectors <- pca_result$rotation

# Print
cat("Number of Eigenvalues:", length(eigenvalues), "\n")
cat("First 5 Eigenvalues:\n")
print(eigenvalues[1:5])
cat("Number of Eigenvectors:", ncol(eigenvectors), "\n")
cat("Dimension of all Eigenvector:", dim(eigenvectors), "\n")
# cat("First 5 Eigenvectors:\n")
# print(eigenvectors[, 1:5])
```

b

```
# Extract Eigenvectors
pc1_pc2_pc3_vectors <- pca_result$rotation[, 1:3]
# Project
projection <- as.matrix(centered_nucleobases) %*% pc1_pc2_pc3_vectors
# Combine
data_b123 <- cbind(data[, c("Id", "Sex", "Population")], projection)
colnames(data_b123)[4:6] <- c("PC1", "PC2", "PC3")
# View
head(data_b123)
```

```
# scatter plot
plot1 <- ggplot(data_b123, aes(x = PC1, y = PC2, color = Population)) +
  geom_point() +
  theme_minimal() +
  scale_color_brewer(type = "qual", palette = "Set1") +
  theme(legend.title = element_text(size = 5), legend.text = element_text(size = 10))

plot1
ggsave("./image/b.png", plot = plot1, width = 5, height = 4, units = "in")
```

d

```
# scatter plot
plot2 <- ggplot(data_b123, aes(x = PC1, y = PC3, color = Population, shape = Sex)) +
  geom_point() +
  theme_minimal() +
  scale_color_brewer(type = "qual", palette = "Set1") +
  theme(legend.title = element_text(size = 12), legend.text = element_text(size = 10))

plot2
ggsave("./image/d.png", plot = plot2, width = 5, height = 4, units = "in")
```

f

```r
# Extract
third_pc <- pca_result$rotation[, 3]

# Calculate the absolute values
abs_third_pc <- abs(third_pc)

# x-axis
nucleobase_index <- 1:length(abs_third_pc)

# Create a dataframe for plotting
df_for_plot <- data.frame(NucleobaseIndex = nucleobase_index, AbsValueThirdPC = abs_third_pc)

# Use ggplot2 for plotting
plot3 <- ggplot(df_for_plot, aes(x = NucleobaseIndex, y = AbsValueThirdPC)) +
  geom_line(color = "blue") +
  # ggtitle("Nucleobase Index vs Absolute Value of Third Principal Component") +
  xlab("Nucleobase Index") + ylab("Absolute Value of Third PC")

plot3
ggsave("./image/e.png", plot = plot3, width = 5, height = 4, units = "in")
```