PCA: Genomic

Import data and library

```r
library(ggplot2)

data <- read.table("p4dataset2023.txt", header = FALSE, stringsAsFactors = FALSE)
```

Define data matrix X

```r
meta <- data[,c(1:3)] # The first three columns are metadata
raw_data <- data[,-c(1:3)]

# Define a function to find the mode of each column
find_modes <- function(col){
  count <- table(col)
  max_count <- max(count)
  mode <- names(count[count ==max_count])
  return(mode)
}

modes <- sapply(raw_data,find_modes)

# Create a binary matrix X
X <- matrix(0, nrow = nrow(raw_data), ncol = ncol(raw_data))
for(i in 1:ncol(raw_data)){
  X[,i] <- ifelse(raw_data[,i] == modes[i], 0, 1)
}
X <- as.data.frame(X)
```

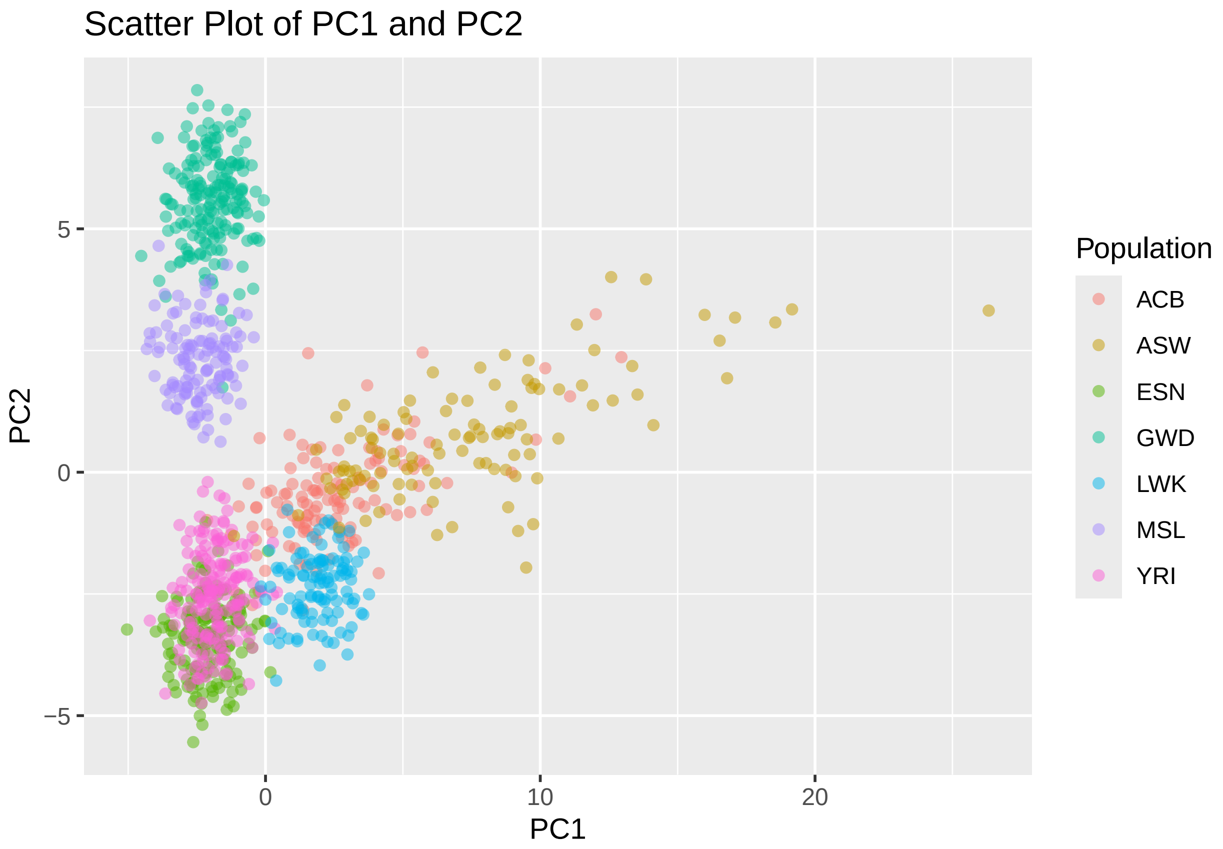
Perform PCA on sample covariance matrix of X

```
pca <- prcomp(X, center = TRUE, scale = FALSE)
```

Plot of PC1 and PC2

```
pc1_score <- pca$x[,1]
pc2_score <- pca$x[,2]
scatter <- data.frame(pc1 = pc1_score, pc2 = pc2_score, pop = meta$V3)

fig1 <- ggplot(scatter, aes(x=pc1, y=pc2, color = pop)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatter Plot of PC1 and PC2",
       x = "PC1",
       y = "PC2") +
  scale_color_discrete(name = "Population")
print(fig1)
```



Interpretation

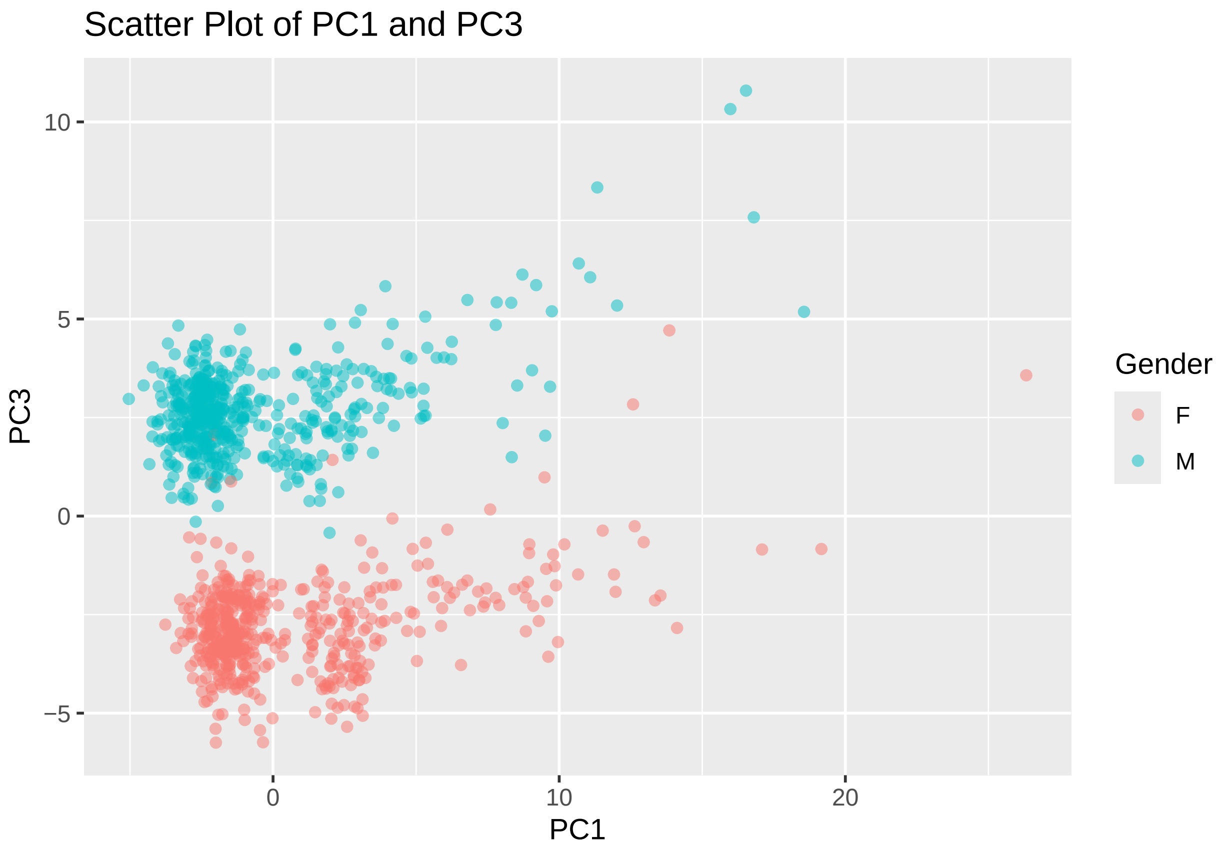
- From the above scatter plot, the first two principal components PC1 and PC2 preserve the information of geographical location and the proximity of the populations.

Plot of PC1 and PC3

```r
pc3_score <- pca$x[,3]
scatter <- data.frame(pc1 = pc1_score, pc3 = pc3_score, gender = meta$V2)

fig2 <- ggplot(scatter, aes(x=pc1, y=pc3, color = gender)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatter Plot of PC1 and PC3",
       x = "PC1",
       y = "PC3") +
  scale_color_discrete(name = "Gender")

print(fig2)
```



Interpretation

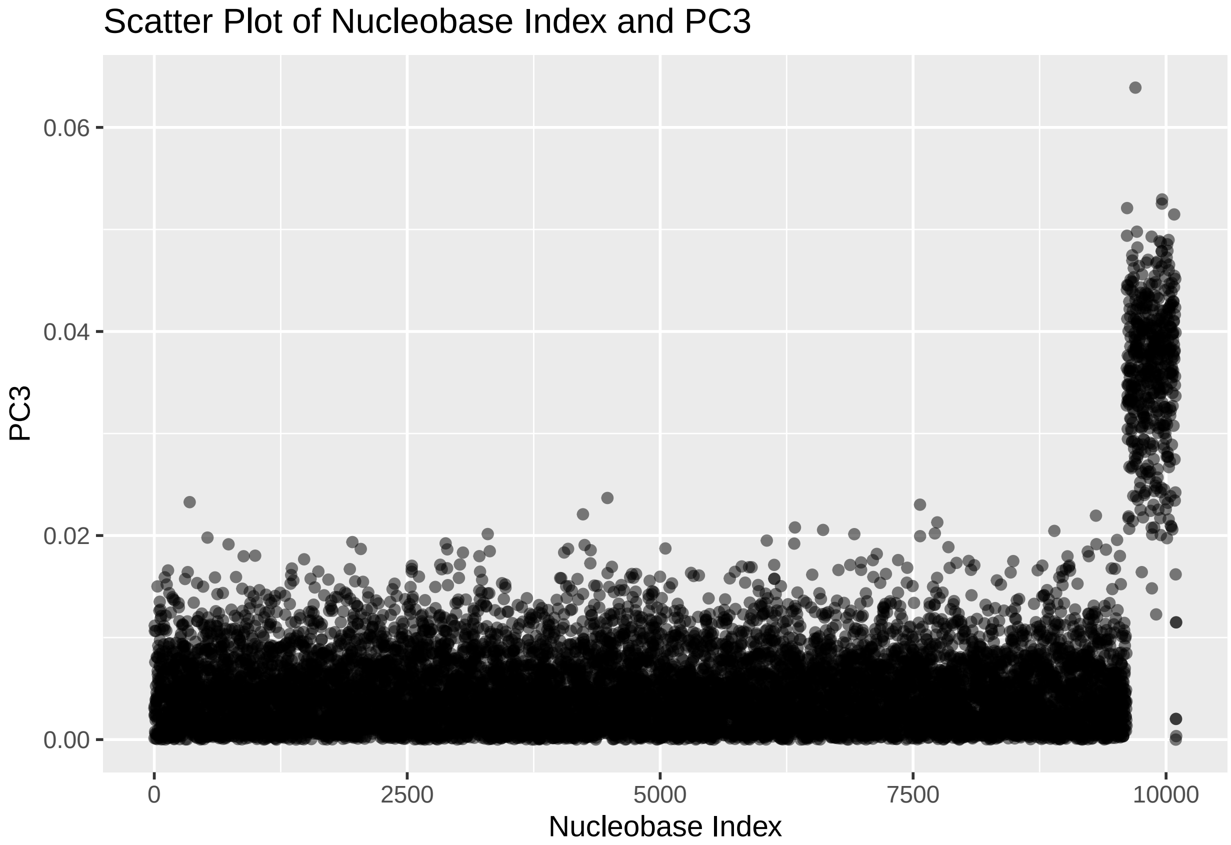
- From the above scatter plot, the third principal component PC3 preserve the information related to gender.

Plot of nucleobase index and PC3

```r
index <- c(1:ncol(raw_data))
pc3_abs <- abs(pca$rotation[,3])
scatter <- data.frame(index = index, pc3 = pc3_abs)

fig3 <- ggplot(scatter, aes(x=index, y=pc3)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatter Plot of Nucleobase Index and PC3",
       x = "Nucleobase Index",
       y = "PC3")

print(fig3)
```



Interpretation

- The absolute value of the third principal component PC3 is significantly larger in the latter part of the nucleobase index
- It is possibly because of differences in the number and type of genes on the X and Y chromosomes, and PC3 captures this difference