A TECHNICAL REPORT ON THE VISUALIZATION OF THE HUMAN GENOME STRUCTURE

INTRODUCTION

Human genetic diversity are the differences and similarities in the genomic structure among various populations or ancestries. The combined differences in the DNA of all individuals makes up the genetic diversity of that specie. It ranges broadly from the number of species to variance within species, and can be attributed to the span of survival for species. Genetic diversities can arise due to various reasons such as mutations (insertion& deletions), natural selections, Single Nucleotide Polymorphism and Copy Number Variants.

This project centers on the visualization of the diversity in the genomic structure of Chromosome 1 among five ancestries namely: African Ancestry, East Asian Ancestry, American Ancestry, South Asian Ancestry, South Asian Ancestry and European Ancestries. Although a small number of genetic variants are found more frequently in certain geographic regions or in people with ancestry from those regions this study is important in understanding genetic diversity and how it can be beneficial to sciences.

.

METHOD

A total of 2709 samples made up of various individuals from six ancestries were collected namely: African Ancestry, East Asian Ancestry, American Ancestry, South Asian Ancestry, South Asian Ancestry and European Ancestries. Data was collected from 1000 genome project database. https://www.internationalgenome.org/data-portal/data-collection/grch38. Chromosome number one was used for the analysis. File was converted from vcf to ped using the vcf to ped converter http://grch37.ensembl.org/Homo_sapiens/Tools/VcftoPed . Info file and Ped Files were gotten from the conversion and used in the linux terminal for analysis.

Linux Terminal Analysis

#Gunzip Ped File

gunzip 1_1-150000.ped.gz

 # Open a Script

nano 1_1-150000.ped

#Convert Info to Map

perl info_to_map.pl 1_1-150000.info > 1_1-150000.map

 # Download Plink

wget https://s3.amazonaws.com/plink1-assets/plink_linux_x86_64_20220402.zip

#Unzip Plink

unzip plink_linux_x86_64_20220402.zip

#Install Plink

```
sudo cp plink /usr/local/bin
sudo chmod 755 /usr/local/bin/plink
sudo cp plink /usr/local/bin
sudo chmod 755 /usr/local/bin/plink
sudo nano ~/.bashrcstall Plink
```

#Generate binary version of ped and map file using plink

plink --bed 1_1-150000.bed --bim 1_1-150000.bim --fam 1_1-150000.fam –pca

#Download Eigenvalues and every file used on the analysis to the local storage and use R for the plotting

R Analysis

 # Set Directory

```
setwd("C:/Users/user/OneDrive/Hackbio Stage 3")

getwd()
```

#install ggplot

```
library ("ggplot2")

metadata <- read.table("C:/Users/user/OneDrive/Hackbio Stage 3/sample list 1000 genomes on grch38.tsv.tsv", sep = "\t", header = TRUE)

head(metadata)

#pcal
```

#Read the eigenvec file

```
pcal <- read.table("C:/Users/user/OneDrive/Hackbio Stage 3/plink.eigenvec", sep =" ", header = F)
```

#Merge the data from pcal and metadata

```
merge_data <- merge(x= pcal,y=metadata,by.x = "V2", by.y="Sample.name", all = F)

library ("ggplot2")
```

#Plot with ggplot

```
ggplot(data = merge_data, aes(V3,V4,color = Superpopulation.code)) + geom_point(size = 2.5)+
 scale_color_brewer(palette = "Set1") + theme_classic() + labs (title = "PCA Plot 1")+
 xlab("Pca1") + ylab("pca2")+
 theme(plot.title = element_text(hjust = 0.5, face = "bold", size =30),
axis.title.x = element_text(face = "italic", color="green", size =14),
axis.title.y = element_text(face = "italic", color= "#33993d", size = 14))
  #annotate("text", label = "cluster 1", x = 0.01, y = -0.02, size = 8, color = "green")+
  #annotate("text", label = "cluster 2", x = -0.05, y = 0.02, size = 8, color = "blue")+
  #annotate("text", label = "cluster 3", x = 0.01, y = -0.08, size = 8, color = "black")+
  #annotate("text", label = "cluster 4", x = 0.00, y = -0.12, size = 8, color = "red")+
 ggsave("plot.png", width = 25, height = 20, units = "cm", limitsize = FALSE, path = "C:/Users/user/OneDrive/Hackbio Stage 3", dpi =300 )
```

PCA PLOT ANALYSIS

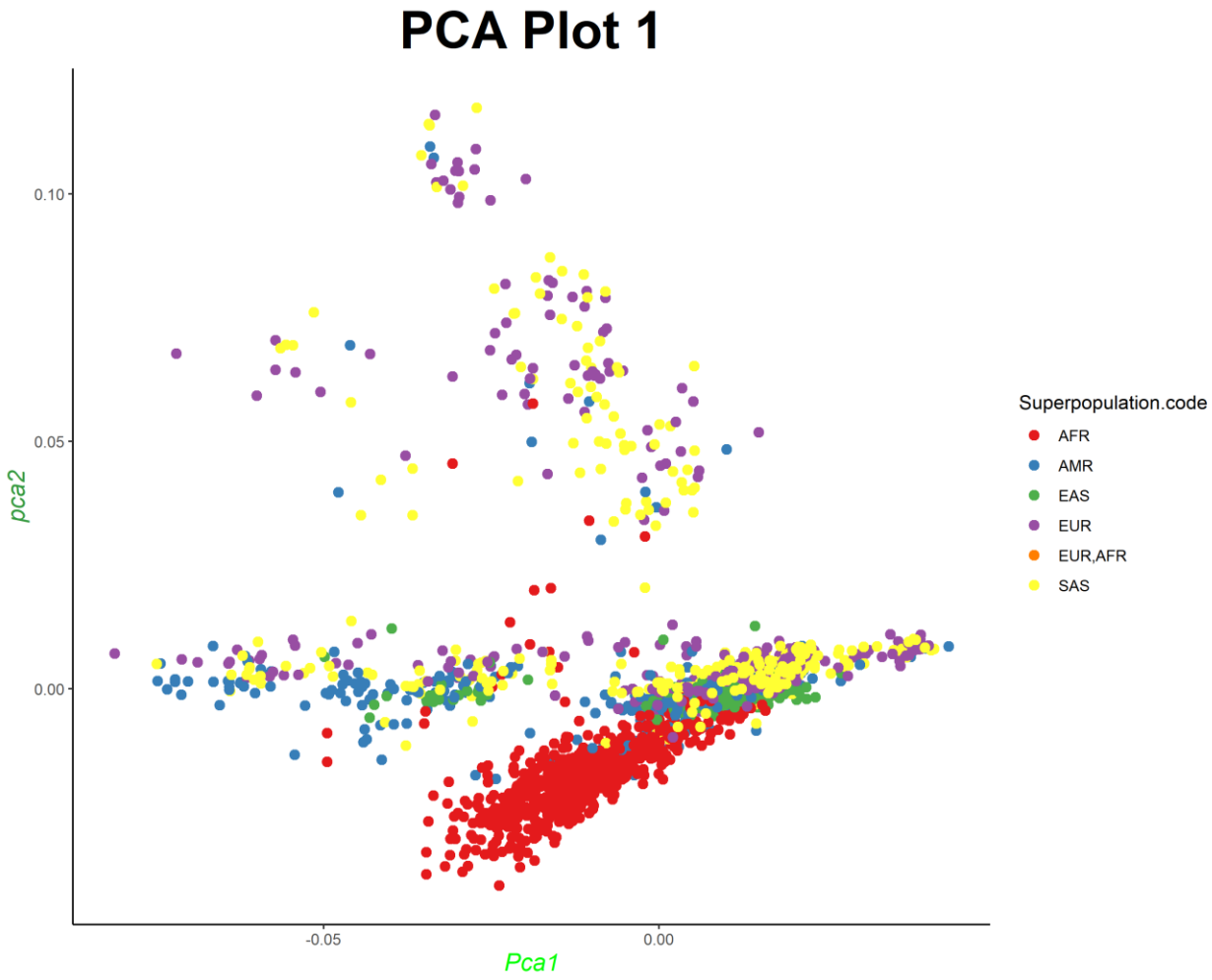

Fig 1: Principal components analysis (PCA) showing the genetic diversity of six populations. For ease of reading, the figure populations are divided by color into six groups based on ancestry: red, African ancestry; blue, American Ancestry; green, East Asian Ancestry; purple, European Ancestry; orange, European Ancestry & African Ancestry; yellow, South Asian Ancestry. The PCA is presented as a plot that shows the PCA scores of individual ancestry

DISCUSSION

The PCA plot shows the distinctiveness and similarities among the ancestries. This research contributes to the understanding of genetics of each population and the distinctiveness among these populations. This is important as it contributes knowledge to genomic sciences in several ways, like in pharmacogenomics, the use of precision treatment where the variability of genetics can be taking into consideration in predicting accurate treatment for a particular population. It can also serve as a powerful tool in understanding the evolution of humans among others. More individual genes should be sequenced so as to have more genetic information, which will be of benefit to genomic science and molecular biology. This calls for more inclusion/diversity in the genomic datasets that are being studied for the therapeutic or diagnostic purposes.