Visualization of the global genome structure

By

Olagunju Grace Boluwatife

For: HackBio Genomics Workshop

July,2022

Task: Represent visually the genome diversity of 4 continents (Africa, America, Asia and Europe) using their Chromosome 1 data.

Objective: The results should depict the diversity of the world's genome and show the need for more inclusion in sequencing projects.

1.0 Introduction

Genetic diversity is defined as the range of different inherited traits within a species. It is also referred to as the genetic variability present within species.

The Human Genome Project completed in 2003 elucidated the genetic blueprints of human species. However, a lack of diversity in populations involved in the project places a restriction on its reach to certain populations and resulted in a growing disparity in healthcare.

A number of factors contribute to variation observed among species and they include mutations, recombination of genes through sexual reproduction, natural selection and human migration patterns.

This project focuses on the visual representation of the diversity of the human genome and the need for more inclusion in the genome sequencing projects.

2.0 Methods and softwares

2.1 Datasets

The datasets that were used are already prepared files available in a public GitHub repository. The datasets were downloaded using the `wget` command.

The dataset includes the Complete 1000 genomes sample list (Tab delimited file containing the ID of each sample and the population code), binary plink files and the processed plink files.

The processed plink files include the ped (pedigree), map (ancestry information) and info (vcf to ped) files while the binary plink files include the bed (binary pedigree), bim (binary ancestry file) and fam (binary family data for each data) files.

The following commands were used:

```
wget https://github.com/HackBio-Internship/public_datasets/blob/main/Global_genome_structure_project/complete_1000_genomes_sample_list_.tsv?raw=true -O complete_1000_genomes_sample_list_tsv
```

```
wget https://github.com/HackBio-Internship/public_datasets/blob/main/Global_genome_structure_project/binary_plink_files/1_1-150000.bed?raw=true -O sample.bed
wget https://github.com/HackBio-Internship/public_datasets/blob/main/Global_genome_structure_project/binary_plink_files/1_1-150000.bim?raw=true -O sample.bim
```

```
wget https://github.com/HackBio-Internship/public_datasets/blob/main/Global_genome_structure_project/binary_plink_files/1_1-150000.fam?raw=true -O sample.fam
```

2.2 PLINK

PLINK (PuTTY Link) is a free, commonly used, open-source whole genome association analysis toolset designed to perform a wide range of genetic analyses which include data management, basic statistics, linkage equilibrium (LD), Identity by descent (IBD), Identity by State (IBS), Population stratification, such as Principal Component Analysis and genome-wide association study.

It is important to note that all commands involving the use of plink starts with typing "plink" followed by the options to specify the data files (all starting with - -option).

In this project, PLINK was used in conjunction with the R statistical tool for Principal Component Analysis.

2.2.1 Preparing files in PLINK and generation of eigenvalues.

First, the plink software should be installed and then loaded on the server.

#to load plink

```
module load plink
```

All files to be analyzed with plink were put in the same directory and the following command was used to generate the eigenvalues (eigenvec and eigenval).

```
plink --bed sample.bed - -bim sample.bim --fam sample.fam –pca
```

Note that, the bed, bim and fam files were saved with the prefix "sample". The - -pca means generate eigenvalues. Also, the number of chromosomes were not specified because plink was designed for human data.

All files, including the eigenvalues generated were downloaded from the remote server using WinSCP (a free and open-source SSH File Transfer Protocol (SFTP), File Transfer Protocol (FTP), WebDAV, Amazon S3 and secure copy protocol client for Microsoft Windows). It is used to secure file transfer between a local computer and a remote server.

2.2.2 Generating a Principal Component Analysis (PCA) plot in R.

The files downloaded from the remote server (eigenval, eigenvec and Complete 1000 genomes files) were uploaded to the RStudio for the generation of PCA plot.

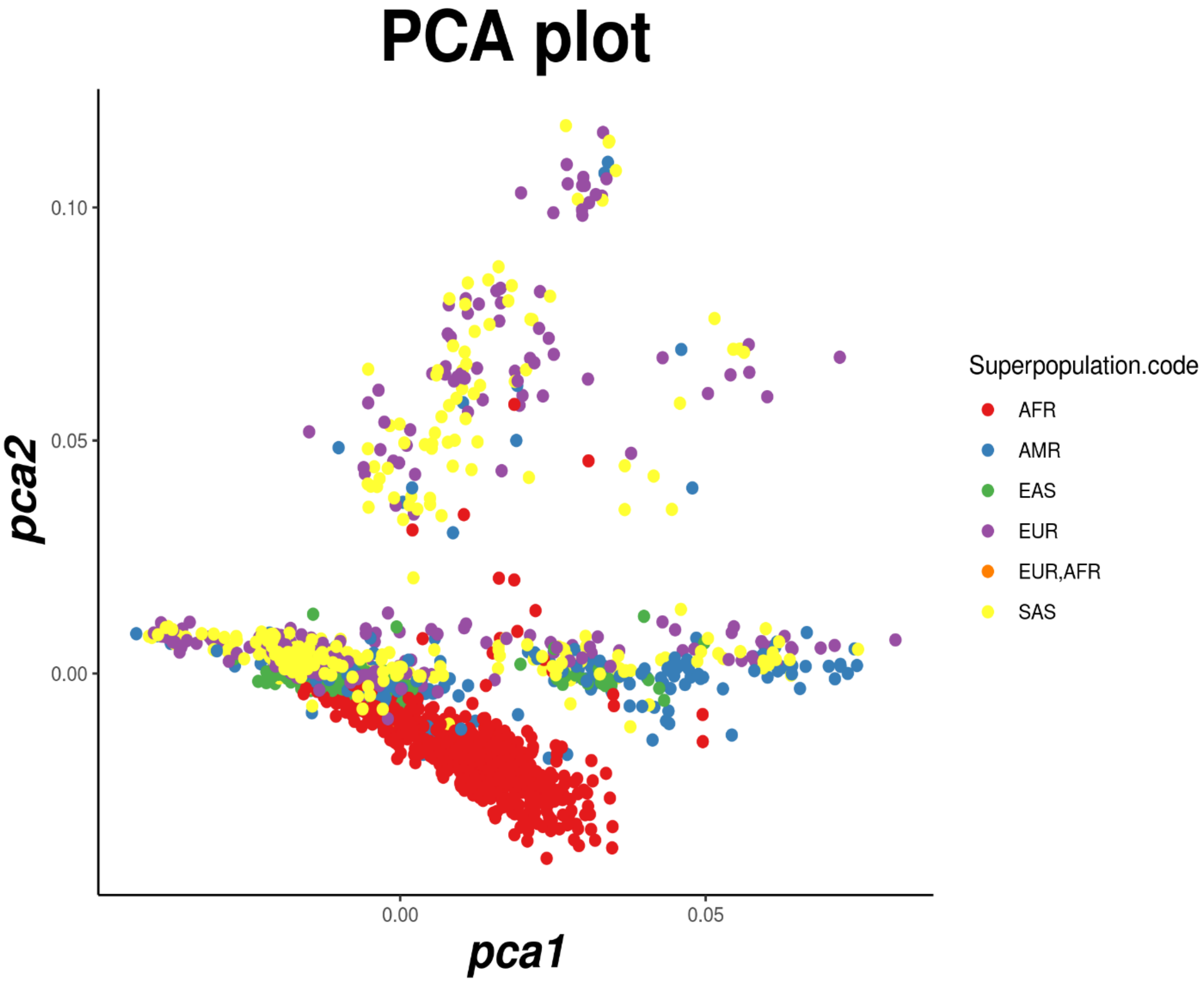
Find the R script in the link below:

https://raw.githubusercontent.com/Grace1-g/HackBio-Project/main/project.R

3.0 Result

The image below is a PCA plot which shows clusters in the samples based on their similarity as well as differences in the population considered.

The plot was generated by plotting the maximum variance (PCA1) measured in the eigenvalue against the second most important variance (PCA2).



where AFR stands for African Ancestry

AMR stands for American Ancestry

EAS stands for Eastern Asian Ancestry

EUR stands for European Ancestry

SAS stands for South Asian Ancestry

4.0 Discussion

The clusters in the PCA plot above shows how distinct each population is. This shows how inappropriate it is to represent the world's population with few sequenced genomes.

Inclusion of different population in the genome sequencing project is important as it can help to reveal the structure of genetic variation among human, and the history and relationships among different populations.

5.0 Conclusion

Variation in the world's genome as seen among the populations considered calls for inclusion of participants from diverse ancestries in genome sequencing projects. This will help to improve disease risk prevention, identify genetic variants, as well as predict, diagnosis and develop treatments for diseases across all populations thus bridging the gap in health disparities.

Skills developed

The project contributed to the development of the following skills:

1. Use of plink for Principal Component Analysis
2. Interpretation of PCA plots
3. Introduction to R