

VISUALIZATION OF THE GLOBAL GENOME STRUCTURE

By: Ankita Murmu

1. INTRODUCTION

The 1000 Genomes Project is the most detailed catalogue of human genetic variation from at least one thousand anonymous individuals from different ethnic groups [1]. It was an international research effort using newly developed technologies which were faster and less expensive. Until now, the research communities have extensively used the reference data resources generated from this project. Although human beings are 99.9% identical, the 0.1% difference holds important clues about the diversity of the genome and the causes of diseases. Hence, understanding the genome diversity in different populations is crucial.

In this project, the samples from four different continents (Asia, Africa, Europe, and America) in the 1000 genomes database were analysed using their chromosome 1 data.

2. OBJECTIVES

- To investigate correlations between the genomic data points of the populations from the 4 continents (Africa, America, Asia and Europe).
- To generate a low dimensional representation of the genomic data of the 4 continents using PCA in order to preserve as much information as possible (maximizing variance)
- To depict (using graphs) the diversity of the world's genome and show why there is a need for more inclusion in sequencing projects.

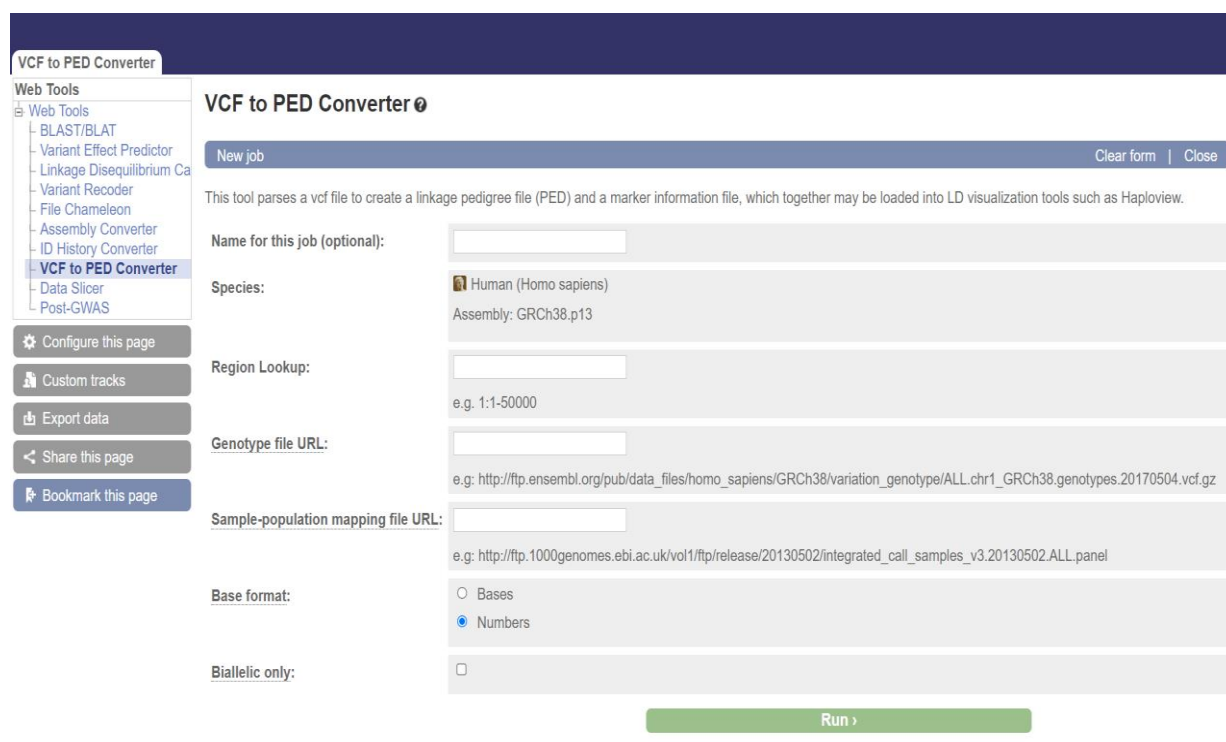
3. METHODOLOGY

i. Downloading the datasets

The complete sample list of all the populations was downloaded from the 1000 genomes database of the GRCh38 assembly [2]. The metadata can be found in the [GitHub](#) repository.

ii. Converting VCF to PED

The conversion can be performed using the online tool or the API script. In this project, the online [VCF to PED converter tool](#) was used (**Fig 1**). The online converter tool requires the region of the chromosome (eg. chr:chr_start-chr_end), [genotype file URL](#) (which lists chrom pos, id, ref, alt, qual, filter, info, format) [sample-population mapping file URL](#) (which lists the sample names, populations and gender) and base format (either in numbers or bases) [3]. All the required fields were filled and the base format was chosen as the bases.



The screenshot shows the 'VCF to PED Converter' web interface. On the left, a sidebar lists 'Web Tools' including BLAST/BLAT, Variant Effect Predictor, Linkage Disequilibrium Calculator, Variant Recoder, File Chameleon, Assembly Converter, ID History Converter, **VCF to PED Converter** (highlighted), Data Slicer, and Post-GWAS. Below the sidebar are buttons for 'Configure this page', 'Custom tracks', 'Export data', 'Share this page', and 'Bookmark this page'. The main panel is titled 'VCF to PED Converter' and contains a 'New job' button, 'Clear form', and 'Close' links. A description states: 'This tool parses a vcf file to create a linkage pedigree file (PED) and a marker information file, which together may be loaded into LD visualization tools such as Haploview.' The form includes fields for: 'Name for this job (optional):', 'Species:' (set to 'Human (Homo sapiens)' with 'Assembly: GRCh38.p13'), 'Region Lookup:' (with example 'e.g. 1:1-50000'), 'Genotype file URL:' (with example 'e.g. http://ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh38/variation_genotype/ALL.chr1_GRCh38.genotypes.20170504.vcf.gz'), 'Sample-population mapping file URL:' (with example 'e.g. http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel'), 'Base format:' (radio buttons for 'Bases' and 'Numbers', with 'Numbers' selected), and 'Biallelic only:' (checkbox). A green 'Run >' button is at the bottom right.

Fig1. VCF to PED converter webpage

Converting the VCF to PED creates a linkage pedigree (PED file) and marker information file (INFO file) of the VCF which can be visualized using [Haploview](#).

iii. Generating MAP file from the INFO file

In the next step, a [Perl script](#) was created using nano which was run to generate the MAP file from the INFO file [4].

```
#creating info_to_map.pl
```

```
nano info_to_map.pl
```

```
#Converting the info file to map using info_to_map.pl
```

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

iv. Generating Binary versions of the MAP and PED files using PLINK

Plink is a free and open-source toolset for analysing genotype or phenotype data in a computationally efficient manner [5]. Stable Linux version of Plink was downloaded using the wget command and installed successfully.

```
#downloading and installing plink
```

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

```
unzip plink_linux_x86_64_20220402.zip
```

```
#making PLINK accessible from the command line anywhere in the file system,
```

```
sudo cp plink /usr/local/bin
```

```
sudo chmod 755 /usr/local/bin/plink
```

```
#adding PLINK to PATH
```

```
sudo nano ~/.bashrc
```

```
# adding the following line at the bottom of the text editor:
```

```
export PATH=/usr/local/bin:$PATH
```

```
#save and exit the file.
```

```
#testing to check if plink is successfully installed, run
```

```
plink
```

```
#Using plink to generate the binary versions of ped and map files
```

```
plink --file 1_1-150000 --make-bed --out 1_1-150000
```

When using the `--make-bed` option, the threshold filters for missing rates and allele frequency were automatically set to exclude nobody. Although these filters can be specified manually (using `--mind`, `--geno` and `--maf`) to exclude people, this default tends to be wanted when creating a new PED or binary PED file. The commands `--extract` / `--exclude` and `--keep` / `--remove` can also be applied at this stage.

v. Principal component analysis

Principal components analysis (PCA) is performed to reduce the dimensionality of the datasets. In this project, PCA was used to identify the population structure in the samples analysed. This analysis computes principal components (PCs) that explain the differences between individuals in the genetic data [6]. In brief, PCA works by:

1. Subtracting the mean
2. Calculating the covariance matrix
3. Calculating the eigenvectors and eigenvalues of the covariance matrix
4. Choosing components and forming a feature vector

In this project, Plink was used to generate the eigen values and eigen vectors to plot them directly in R. The files with the suffix `bed`, `bim` and `fam` were used. The `-pca` part of the command generates the eigenvalues.

#generating eigenvalues

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

vi. PCA plotting with R

Plotting was performed using the `ggplot2` package in RStudio. The complete 1000 genomes sample list and the eigenvector file were used as input files and merged to create a data frame. This data frame was then used and the first and second principal components were plotted using `ggplot2`. Because there are no headers in the file eigenvector file, R arbitrarily gives these as V numbers, so the first two columns are called V1 and V2 which contain the sample names and V3 corresponds with PC1, V4 with PC2 and so on. `ggplot` package in R produced colourful plots which were easy to visualize and comprehend. The R script for the PCA plot can be found in the [GitHub](#) repository.

4. RESULTS AND DISCUSSION

i. VCF to PED conversion

The VCF to PED conversion produced two files with suffix PED and INFO. The first five columns of the PED file generated from the VCF to PED online conversion were:

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)

However, the PED file contained zeroes for paternal and maternal ID and sex despite having that information in the panel file. The INFO file contained two columns: Generic ID and location of the ID which was used to create the MAP file.

ii. MAP file creation

Each line of the MAP file describes a single marker and contained exactly 4 columns:

1. Chromosome (1-22, X, Y or 0 if unplaced)
2. rs# or snp identifier
3. Genetic distance (morgans)
4. Base-pair position (bp units)

iii. Binary versions of PED and Map files

The generation of the binary versions of PED and MAP files using PLINK produced three output files. Three files created are -- the binary file BED which contains the raw genotype data, but also a revised map file BIM which contains two extra columns that give the allele names for each SNP, and FAM file which is just the first six columns of the PED file. The .bim and .fam files can be viewed and none of these three files should be manually edited.

iv. Principal Component Analysis

The command using Plink generated two output files with the suffix eigenval and eigenvec. The eigenvalue file tells the order of each PC (so PC1, PC2....) and the percentage each eigenvalue contributes to the variance. The highest variance is listed at the top and the lowest at the bottom. The eigenvector file contains the coordinates for each sample. This file has no headers, is tab-separated and contains the sample name in columns one and two, and then

subsequently the eigenvalue for each PC. These values were then outputted and used to generate a plot in R.

v. PCA plots with R

The eigenvector with the highest eigenvalue is the principal component of the dataset. PC1 has the largest sample variance (first principal component) followed by PC2, PC3 and so on. The plot (**Fig. 2**) shows clusters of the superpopulation of different ancestries: African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). It is interesting to note that a sample (HG0173) of mixed ancestries of EUR and AFR is found.

The first principal component PC1 has the largest contribution to the genetic variance and explains 41.9411% of the variance and the second principal component PC2 explains 40.8488% of the variance. The populations of AFR are the most genetically distinct compared to the populations of EUR, SAS, AMR and EAS according to the clusters formed. However, the clustering of the populations from the 4 continents also depicts the genetic similarity among the samples.

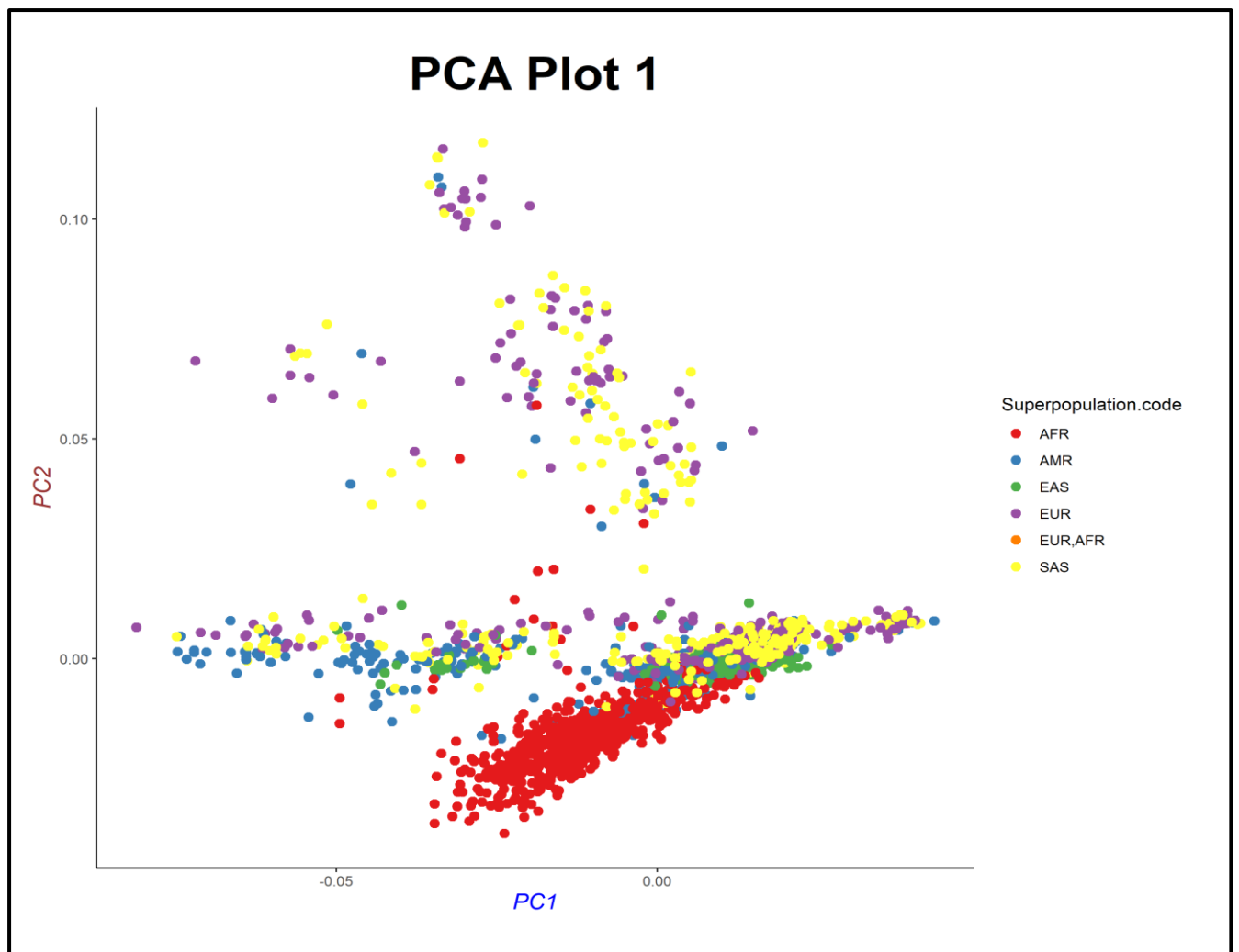


Fig 2. PCA plot between PC1 and PC2 using ggplot

AFR: "African Ancestry" (red); AMR: "American Ancestry" (blue); EAS: "East Asian Ancestry" (green); EUR: "European Ancestry" (purple); AFR: "African Ancestry", EUR: "European Ancestry" (orange); SAS: "South Asian Ancestry" (yellow)

Similarly, in **Fig. 3**, the populations of AFR are found to be distinct from the populations of EUR, SAS, AMR and EAS according to the clusters formed. In this case, the third principal component PC3 explains 34.8714% of the variance which is less than the first and second principal components.

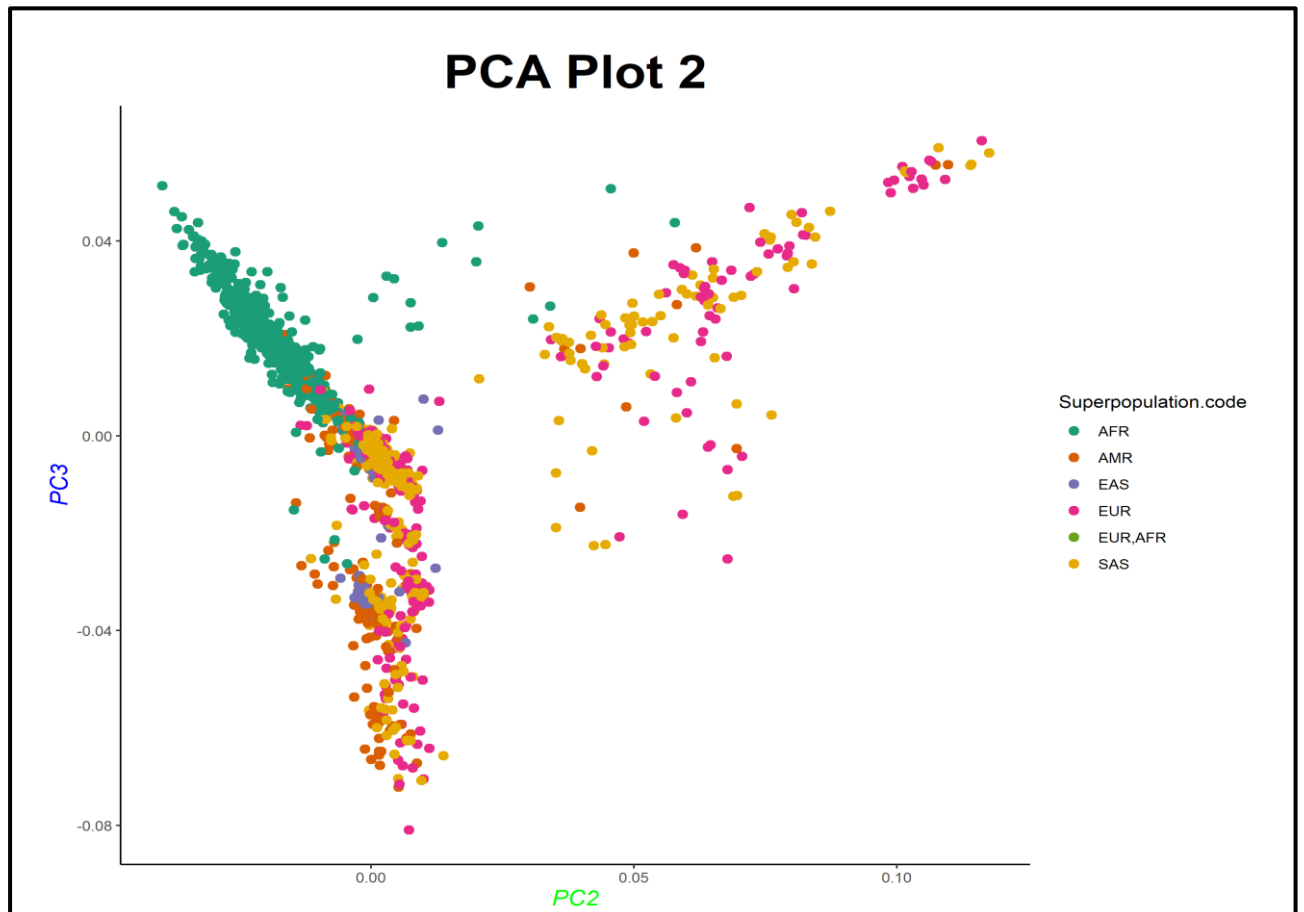


Fig 3. PCA plot between PC2 and PC3 using ggplot

AFR: "African Ancestry" (dark green); AMR: "American Ancestry" (orange); EAS: "East Asian Ancestry" (blue); EUR: "European Ancestry" (pink); AFR: "African Ancestry", EUR: "European Ancestry" (light green); SAS: "South Asian Ancestry" (golden)

Lastly, the PCA plot generated between PC19 and PC20 is compared with the previously generated plots as these principal components has the lowest contribution to the genetic variance. PC19 explains 11.1564% of the variance and PC20 explains 11.1043% of the variance. Surprisingly, **Fig. 4** showed only one cluster containing all the populations from all the ancestries indicating their high genetic similarity and low genetic diversity.

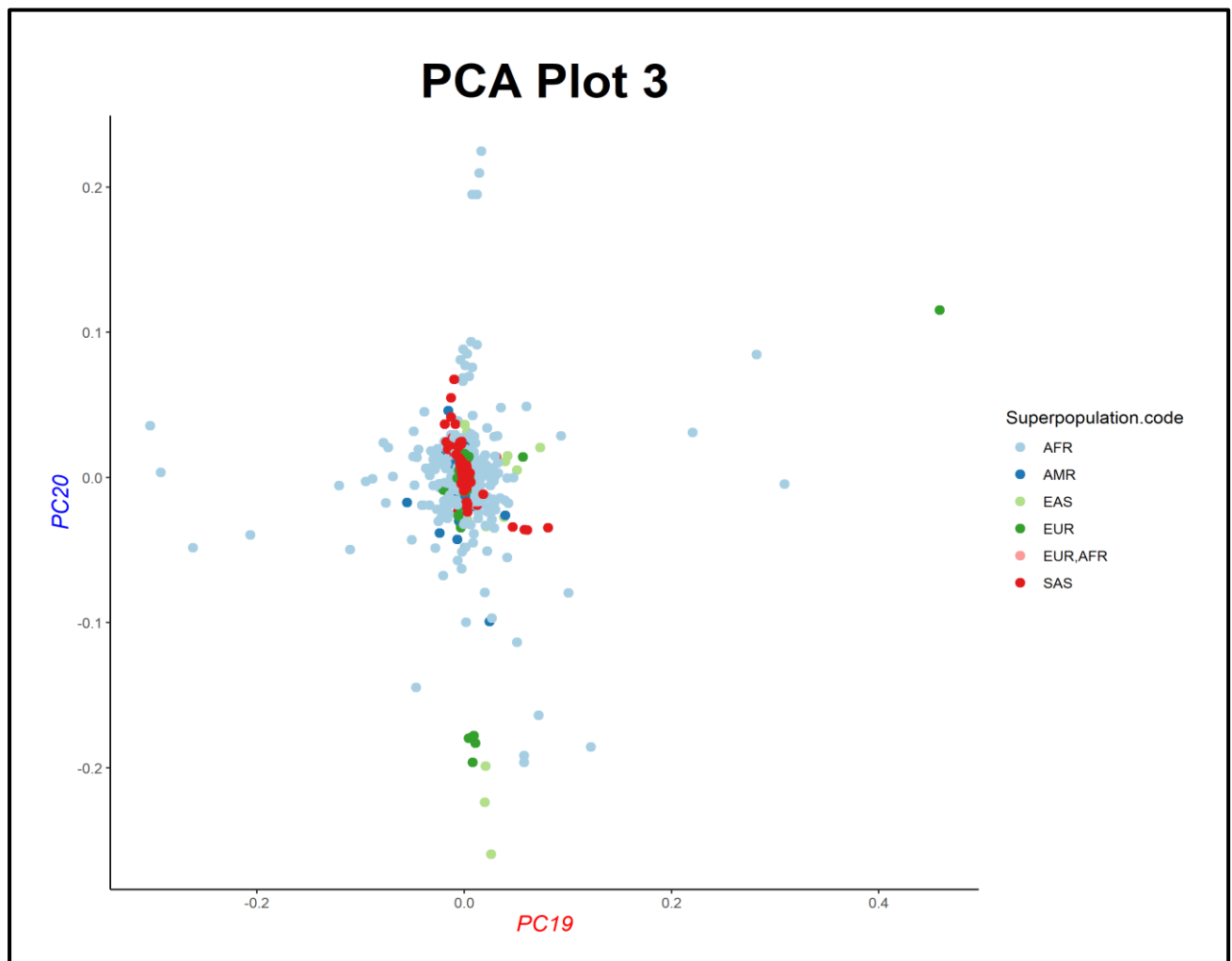


Fig 4. PCA plot between PC19 and PC20 using ggplot

AFR: "African Ancestry" (light blue); AMR: "American Ancestry" (dark blue); EAS: "East Asian Ancestry" (light green); EUR: "European Ancestry" (dark green); AFR: "African Ancestry", EUR: "European Ancestry" (pink); SAS: "South Asian Ancestry" (red)

Hence, from the plots generated, it can be inferred that the populations from the four different continents also have genome similarities among them despite their distinct ancestries and contributions to variance.

5. CONCLUSION

The principal components explained the differences between individuals in the genetic data from chromosome 1. Hence, it can be concluded that the proportion of human genetic variation due to differences between populations is modest. Yet sufficient genetic data from high through-put sequencing studies can permit accurate classification of individuals into populations and understanding the genetic variations comprehensively. This project can be further extended to other pairs of chromosomes and studies of linkage disequilibrium.

6. REFERENCES

1. <https://www.internationalgenome.org/>
2. <https://www.internationalgenome.org/data-portal/data-collection/grch38>
3. http://grch37.ensembl.org/Homo_sapiens/Tools/VcfToPed
4. <https://davetang.org/muse/2016/07/28/vcf-to-ped/>
5. <https://knowledgebase.aridhia.io/article/installing-plink-on-your-virtual-machine/>
6. http://hpc.ilri.cgiar.org/beca/training/data_mgt_2017/BackgroundMaterial/PlinkTutorial.pdf

ACKNOWLEDGEMENTS

The author acknowledges the team of HackBio for providing the opportunity to implement this project. Special thanks to Adewale Ogunleye for helping in producing the PCA plots. Lastly, the author heartily acknowledges Opeoluwa Adewale Fasoro for providing the “Data Science Training Grant for Women” in order to be part of the Genomics workshop offered by HackBio.