VISUALIZATION OF THE GLOBAL GENOME STRUCTURE

Ayano Temitope Ayanfunke.

INTRODUCTION

The 1000 Genomes Project uses whole-genome sequencing on a varied group of individuals from various populations to provide a thorough description of common human genetic diversity. An understanding of genetic diversity can be gained by population structure analysis, which also makes subsequent association mapping research easier to conduct. The variety of various inherited features within a species is referred to as genetic diversity. In a species such as humans, with high genetic diversity, there are many individuals with a wide variety of different traits. Even though human differences are barely 0.1%, it has been observed that this disparity offers important clues concerning genomic diversity.

AIM

This project aims to visually represent the genomic diversity of the human genome on the four continents using their chromosomes.

The Github repository contains the script used for the execution of this project.

DATASET SOURCE

The datasets used were obtained from the 1000 genomes project database. The 1000 Genomes database is a catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy. This GitHub repository contains the metadata.

i. Chromosome 1 data

```
wget
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr1.p
hase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz
```

ii. Panel file

```
wget
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated
_call_samples_v3.20130502.ALL.panel
```

iii. Sample list

```
wget
https://github.com/AyanTemi/Visualization-of-the-global-genome-structure/blob/main/complete_1000_genomes_sample_list_.tsv.txt
```

METHODOLOGY

For population structure analysis, a non-parametric approach (Principal Component Analysis) was used. For analyzing large datasets, this approach is said to be more viable than parametric approaches. (Alhusain and Hafez, 2018). PCA is a standard method for correcting for population stratification in ancestry-specific genome-wide association studies (GWASs) and is used to group individuals by ancestry. (Gaspar ad Breen 2019).

STEPS

1. Convert the variant call format (vcf) files to pedigree (ped) files.

The VCF files were converted to pedigree files following these steps (Davetang, 2016).

This can also be done using the Ensembl VCF to PED converter.

i. The vcf_to_ped_convert.pl API script was downloaded from the 1000 Genomes Project's FTP site.

```
wget
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/vcf_to_ped_converter/version_1.1/vcf_to_ped_convert.pl
```

The vcf_to_ped_convert.pl script makes use of the VCF, the sample panel file, the region of interest (-region), and the population (-population).

ii. Index the vcf files with tabix

```
tabix -p
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz
```

iii. The perl script was run using the four required parameters.

```
perl vcf_to_ped_convert.pl -vcf
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vc
f.gz -sample_panel_file integrated_call_samples_v3.20130502.ALL.panel
-region 1:1-150000 -population GBR ... -population ITU  -base_format
letter
```

The ped and info files generated can be found here.

2. Convert the info file to map

The info file generated above is converted into a map file using a Perl script (info_to_map.pl) gotten from here and run using:

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

The map file generated can be found here.

3. Generate the binary versions of ped and map files.

The binary versions of the ped and map files were generated using plink.

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

The processed plink file can be found here

4. Performing principal component analysis

PCA (Principal Component Analysis) is a multivariate analysis that reduces data's dimensionality while preserving its covariance. It summarizes the major axes of variation in allele frequencies and then produces the coordinates of individuals along these axes.

PCA requires the following steps:

1. Subtract the mean.
2. Compute the covariance matrix.
3. Compute the eigenvectors and eigenvalues of the covariance matrix.
4. Choose components and form a feature vector.

Plink can perform these four steps using the generated bed, fam, and bim files to generate the

eigenvalues and eigenvectors of the covariance matrix of allele frequencies using the command below.

```
plink --bfile 1_1-150000 --pca
```

The outputted eigenvalues and eigenvectors can be found here.

5. Generating PCA Plot in R Studio

Visualisation plays an essential role in genomics research by making it possible to observe correlations and trends in large datasets as well as communicate findings to others. Individual genotypes can be projected onto the space spanned by the PC axes, which allows visualizing the samples and their distances from one another in a colourful scatter plot. The Percentage Variance Explained (PVE) was also calculated.

The R script used to generate the plots can be found in this GitHub repository.

RESULTS

Principal components (PC) can be used as the axes of variations to provide a graphical overview of the population structure, which can help to highlight outlier individuals or those who seem to lie farther out than others. A set of significant principal components can be used to cluster individuals into genetically homogeneous subpopulations. The top principal components (PCs) of SNP data have been shown to map well to geographic distances between human populations (Novembre *et al*., 2008), thus capturing the coarse-grain allelic variation between these groups (Abraham and Inouye, 2014). PCA's most attractive property for population geneticists is that the distances between clusters reflect genetic and geographic distances between them. Some of the plots generated using some of the PCs are shown below.

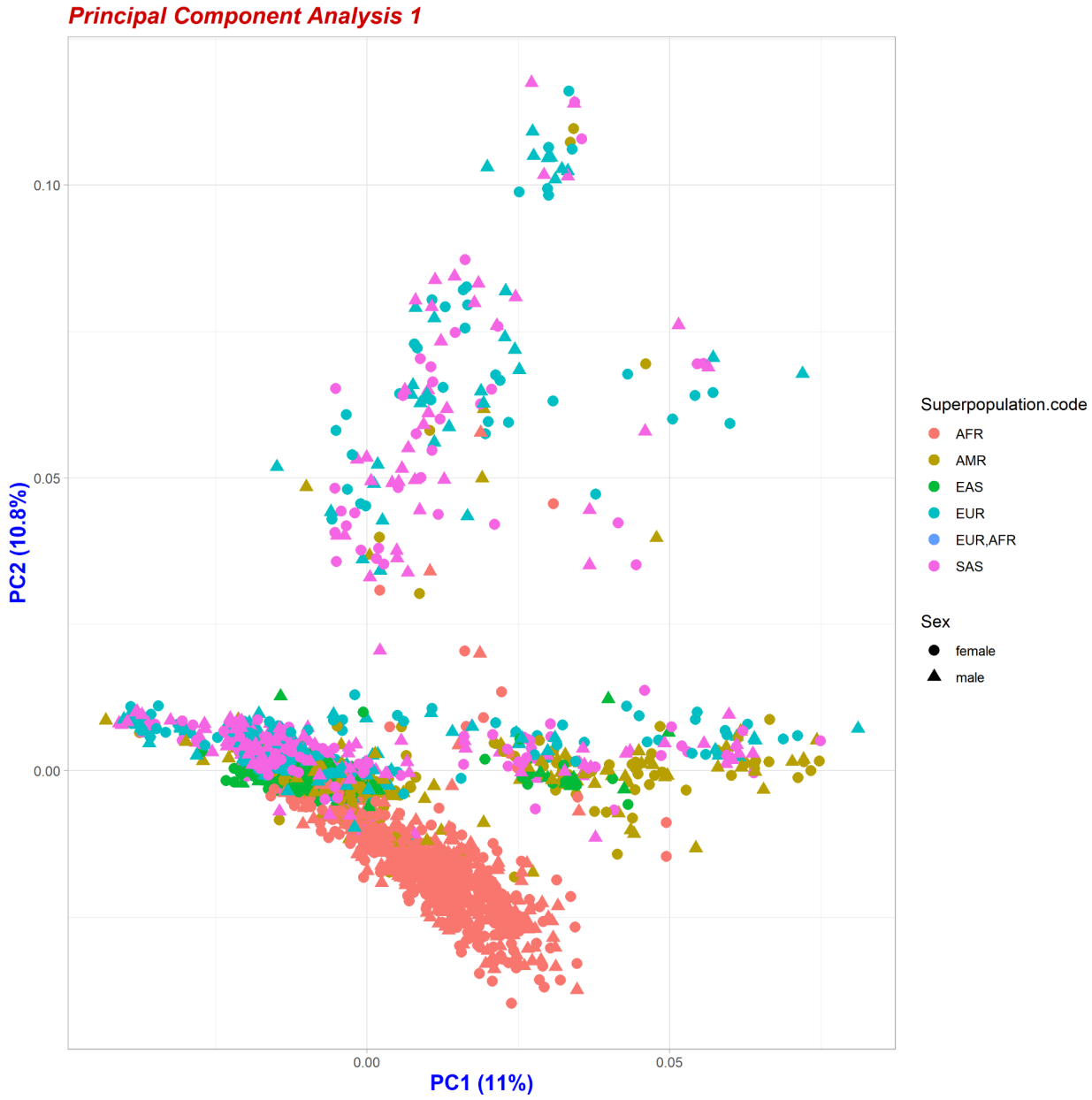

Fig. 1: Principal Component Analysis of PC1 and PC2

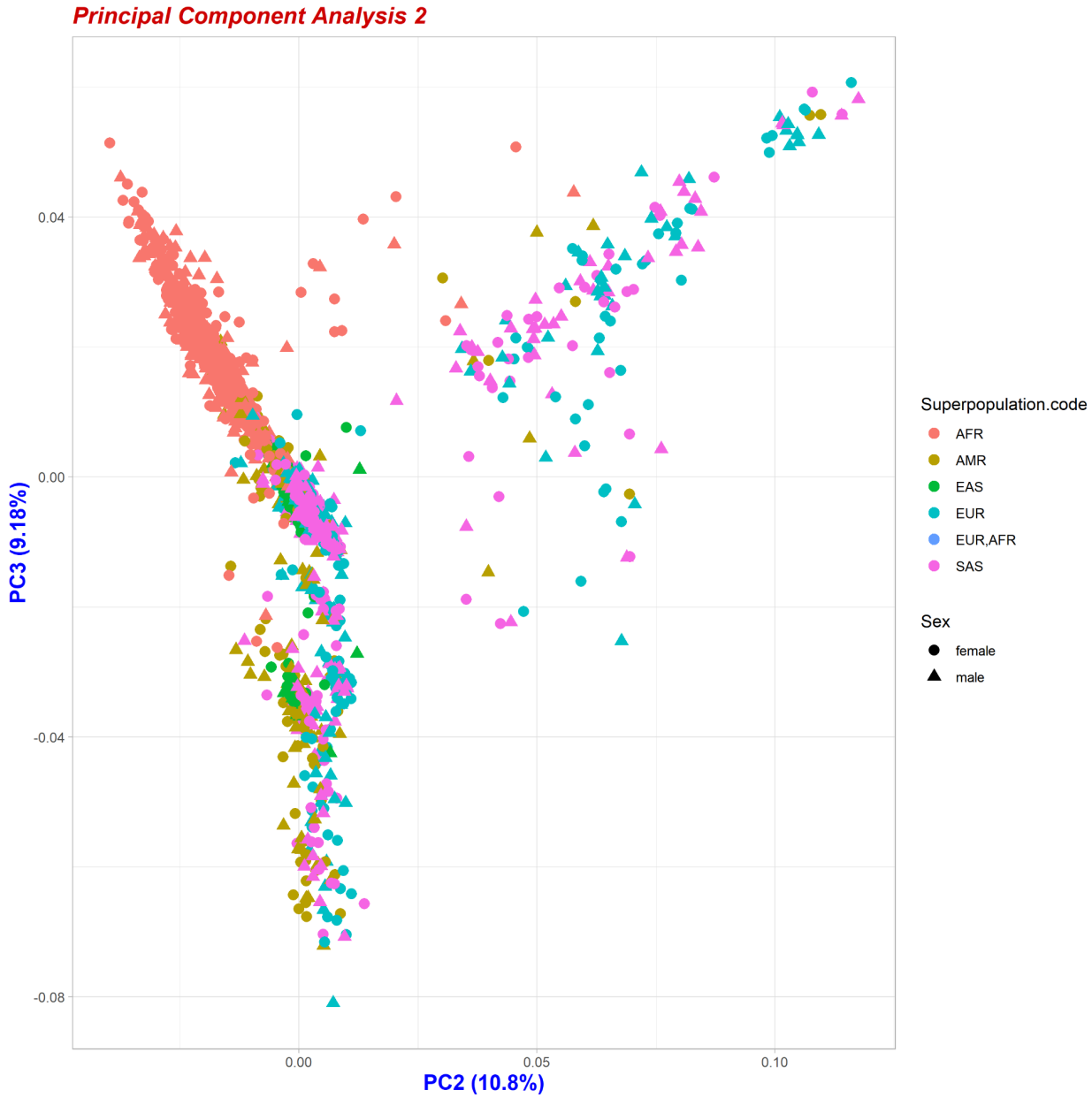

Fig. 2: Principal Component Analysis of PC2 and PC3

Fig. 3: Principal Component Analysis of PC3 and PC4

Keys:

AFR: Africa

AMR: America

EAS: East Asia

EUR: Europe

SAS: South Asia

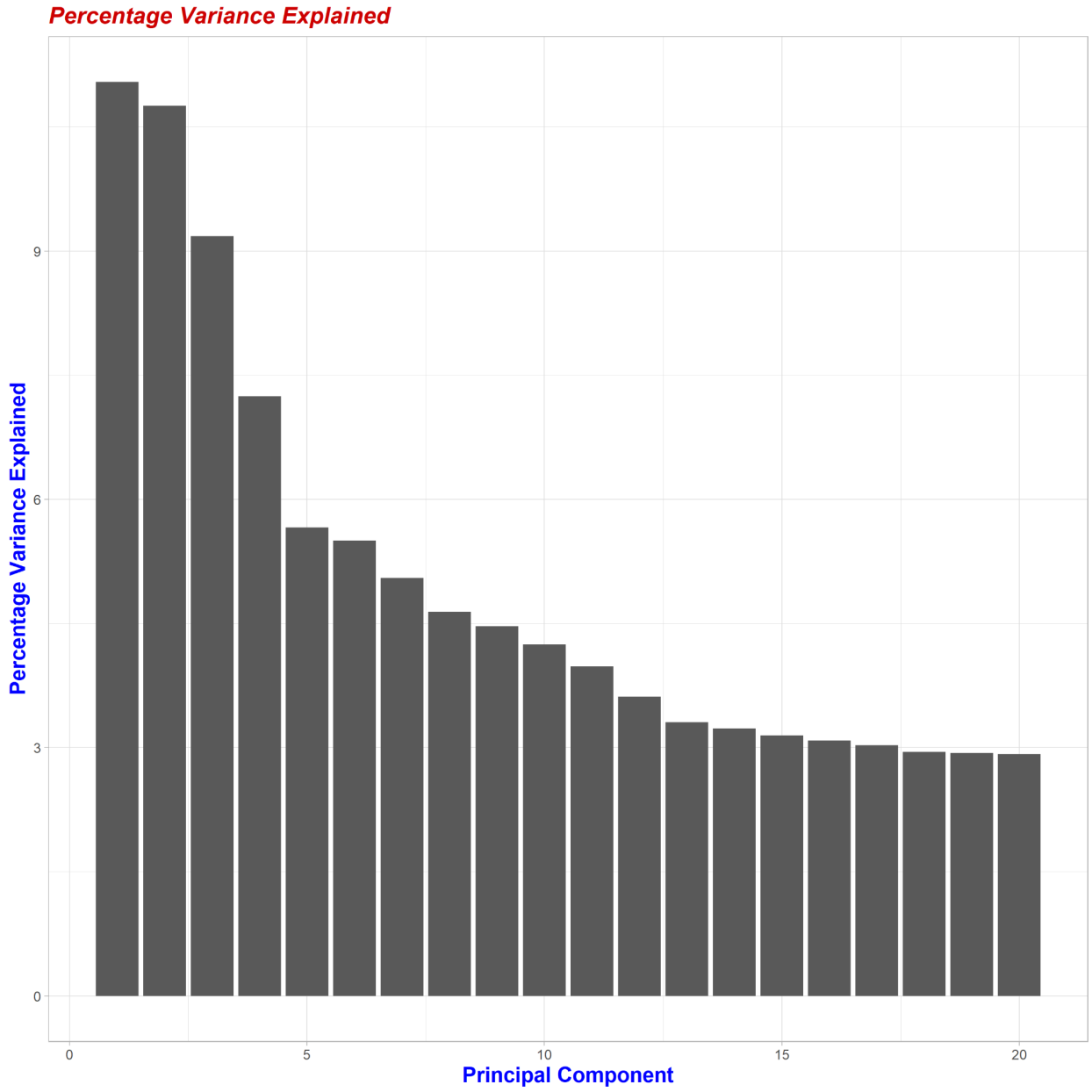

Fig. 4: Percentage Variance Explained

DISCUSSION

Since our data has 20 dimensions, we have 20 principal components, but PCA tries to put the maximum possible information in the first component, then the maximum remaining information in the second, and so on, until we have something like shown in the plot in Fig. 4. The percentage variance was plotted using the percentage eigenvalue (PCA2). As seen from the plot, PC1 accounts for 11%, PC2 10.8%, PC3 9.18%, PC4 7.25%, and it progresses in descending order, with PC20 accounting for the lowest percentage of 2.9%.

The analysis of genome structure provides a clear insight into the underlying genetic population substructure and can serve as a crucial prerequisite for analysing genetic data. PCA's scatterplots were plotted to visualise population structure, where the most genetically isolated subpopulations appear as distinct clusters of individuals.

Fig. 1-3 shows the plot generated from the top four PCs. It can be inferred from the plots that the African(AFR) continent is more genetically diverse than other populations. This degree of genomic diversity can be attributed to population history. Individuals with African ancestry are not receiving the same level of care as individuals of European ancestry due to limitations in available data (Bentley et al., 2017). As a result of the diversity, a greater number of variants are required to tag the same amount of variation as in European ancestry populations.

Some South Asian (SAS), East Asian (EAS), American (AMR), and European (EUR) population samples are overlapped, which reflects common ancestry (Price *et al.*, 2006) or mixed populations due to migration and other factors. Also, there is a mixed sample of Africa and Europe. This further substantiates that human genetic variation is found within populations, not between them, and that individuals are frequently more similar to members of other populations than to members of their population.

CONCLUSION

The inference of population structure from genetic markers is very helpful in different applications, such as genome-wide association studies (GWAS). In addition, determining population structure is required for association mapping studies to avoid making spurious correlations or missing genuine correlations, which would eventually reduce false-positive rates. (Alhusain and Hafez, 2018). Population structure also helps to improve diversity and inclusion in genomic research (Wang et al., 2022), which can provide needed data and guidance to physicians to help them make better decisions for patients of diverse ancestry at the point of

clinical care (Bentley *et al.,* 2017). The diversity of the genomes shows the need for more inclusion in sequencing projects.

ACKNOWLEDGMENTS

To the HackBio team, thank you for the great opportunity. I also acknowledge Fasoro Opeoluwa Adewale for providing the grant for me to be a part of this genomics workshop.

REFERENCES

1. Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS One*, 9(4), e93766. https://doi.org/10.1371/journal.pone.0093766.

2. Al-Husain, L., and Hafez, A.M. (2018). Nonparametric approaches for population structure analysis. *Hum Genomics*, 12, 25. https://doi.org/10.1186/s40246-018-0156-4

3. Bentley AR, Callier S, Rotimi CN. (2017). Diversity and inclusion in genomic research: Why the uneven progress? *J Community Genet*., 8(4):255-266. doi:10.1007/s12687-017-0316-6.

4. Gaspar, H.A., Breen, G. (2019). Probabilistic ancestry maps: a method to assess and visualise population substructures in genetics. *BMC Bioinformatics*, 20, 116. https://doi.org/10.1186/s12859-019-2680-1

5. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.,* 40, 646–649.

6. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M.E. (2006). Principal component analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38, 904–909.

7. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

8. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74.
https://doi.org/10.1038/nature15393
https://www.internationalgenome.org/data-portal/data-collection/grch38

9. Wang, T., Antonacci-Fulton, L., and Howe, K. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature,* 604, 437–446.

https://doi.org/10.1038/s41586-022-04601-8

10. Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A., Jorde, L.B. (2007). Genetic similarities within and between human populations. *Genetics*, 176(1), 351-359. https://doi.org/10.1534/genetics.106.067355

11. https://knowledgebase.aridhia.io/article/installing-plink-on-your-virtual-machine

12. https://davetang.org/muse/2016/07/28/vcf-to-ped/

13. https://www-users.york.ac.uk/~dj757/popgenomics/workshop6.html

14. https://zzz.bwh.harvard.edu/plink/

15. https://www.cog-genomics.org/plink/