Session 02 - Exercises

Deepika Dokuru

☰ workflowr ❗

Before you begin:

- Make sure that R is installed on your computer
- For this lab, we will use the following R libraries:

Set your working directory to your home directory using in R*

The data files are in the folder `/data/SISG2022M15/data/`.

Population Structure Inference

Introduction

We will be working with a subset of the genotype data from the Human Genome Diversity Panel (HGDP) and HapMap.

The file “YRI_CEU_ASW_MEX_NAM.bed (https://github.com/joellembatchou/SISG2022_Association_Mapping/tree/master/data)” is a binary file in PLINK BED format with accompanying BIM and FAM files. It contains genotype data at autosomal SNPs for:

- Native American samples from HGDP
- Four population samples from HapMap:
  - Yoruba in Ibadan, Nigeria (YRI)
  - Utah residents with ancestry from Northern and Western Europe (CEU)
  - Mexican Americans in Los Angeles, California (MXL)
  - African Americans from the south-western United States (ASW)

File with ancestry labels assignment for each sample: Population_Sample_Info.txt (https://raw.githubusercontent.com/joellembatchou/SISG2022_Association_Mapping/master/data/Population_Sample_Info.txt)

Exercises

Here are some things to look at:

1. Examine the dataset:

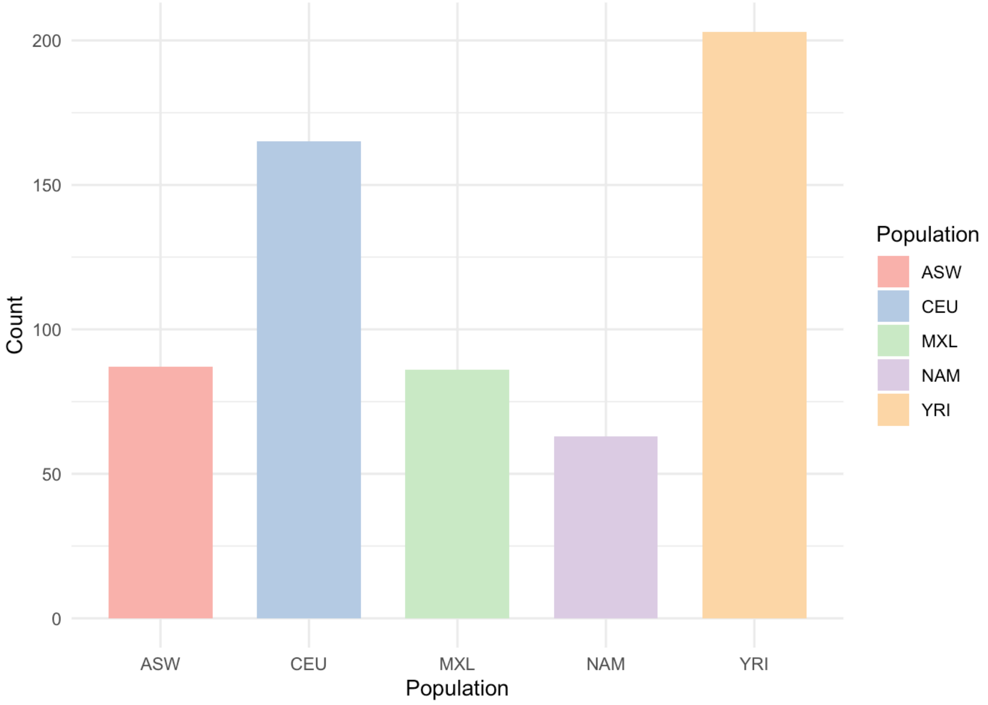
- How many samples are present?

```
[1] 604
```

- How many SNPs?

```
[1] 150872
```

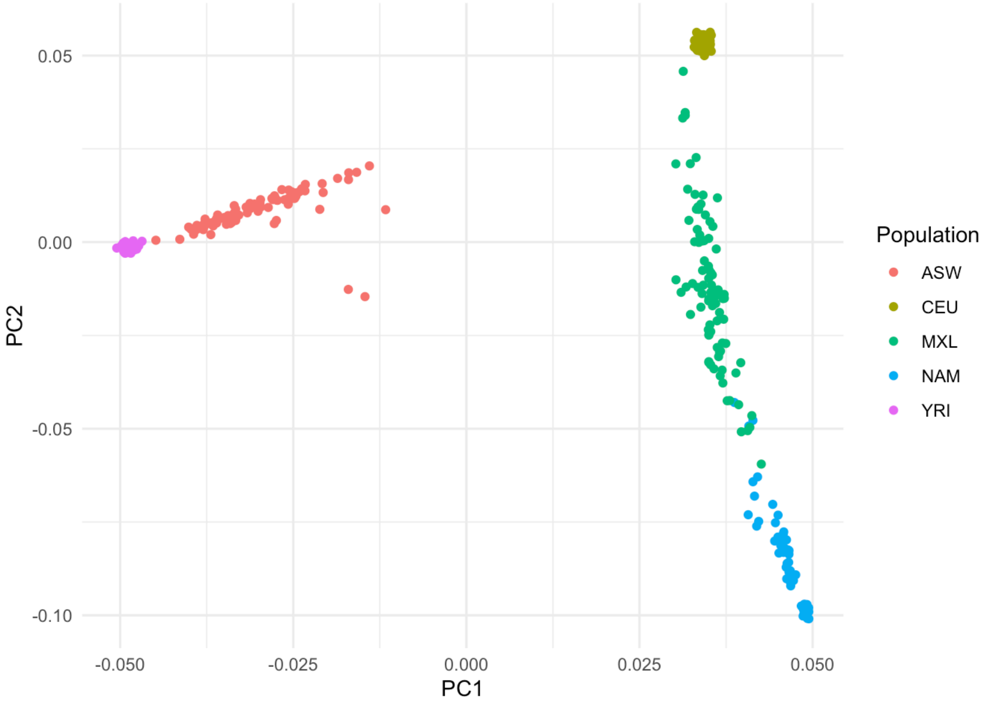
- What is the number of samples in each population?



2. Get the first 10 principal components (PCs) in PLINK using all SNPs. The basic command would look like

This generates a file `<output_prefix>.eigenvec` containing the PCs (eigenvectors) as well as another file `<output_prefix>.eigenval` containing the top eigenvalues.

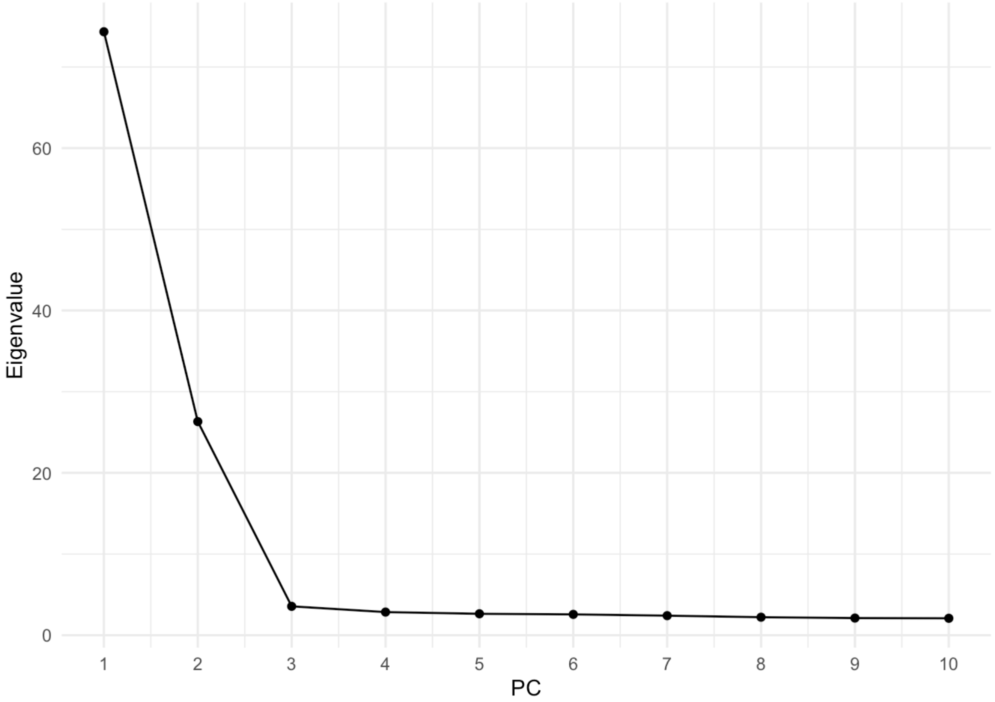
- Make a scatterplot of the first two PCs with each point colored by population membership.



* Interpret the first two PCs, what ancestries are they reflecting?

```
[1] "African ancestry vs non−African ancestry"
```

- Make a scree plot of the eigenvalues for the first 10 PCs.



Approximate the proportion of variance explained by the first two PCs.

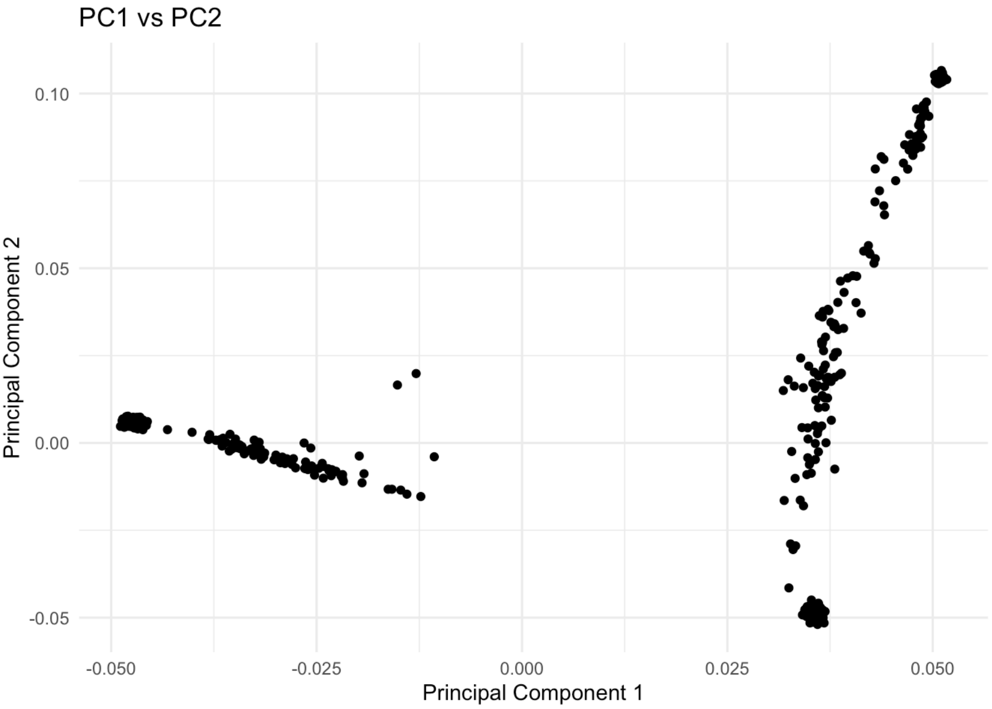
```
[1] 0.8308752
```

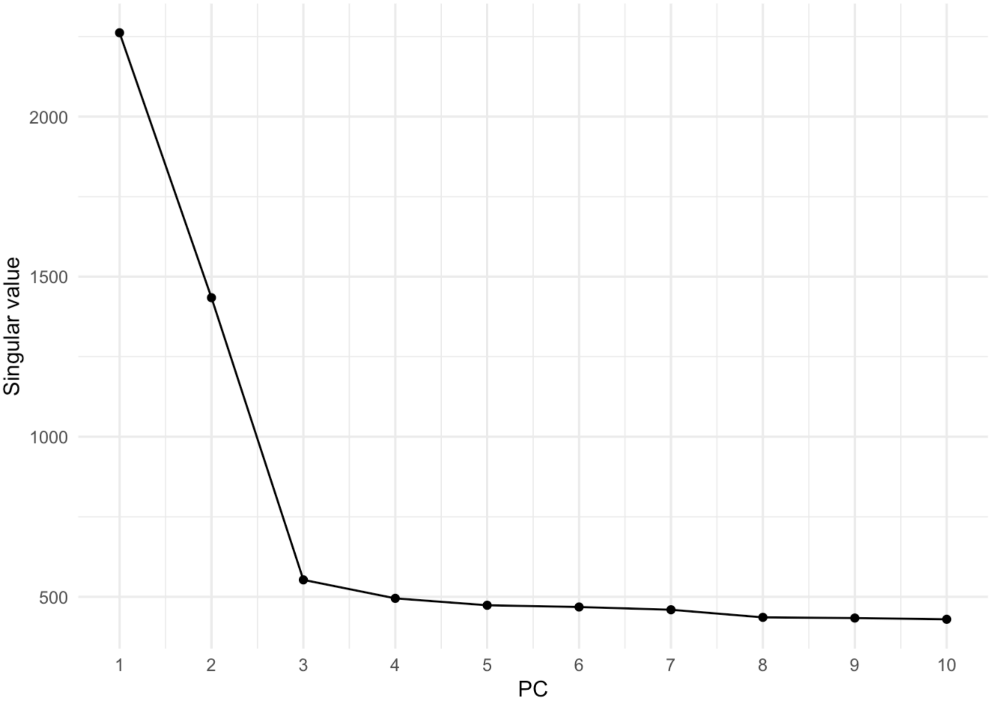
3. Now redo Question 2 above using the `bigsnpr` R package (https://privefl.github.io/bigsnpr/reference/index.html) specifying a r^2 threshold of 0.2 (i.e. LD pruning) as well as a minimum minor allele count (MAC) of 20. The basic command would look like

```

Phase of clumping (on MAC) at r^2 > 0.2.. keep 87127 variants.
Discarding 48 variants with MAC < 20.

Iteration 1:
Computing SVD..
```

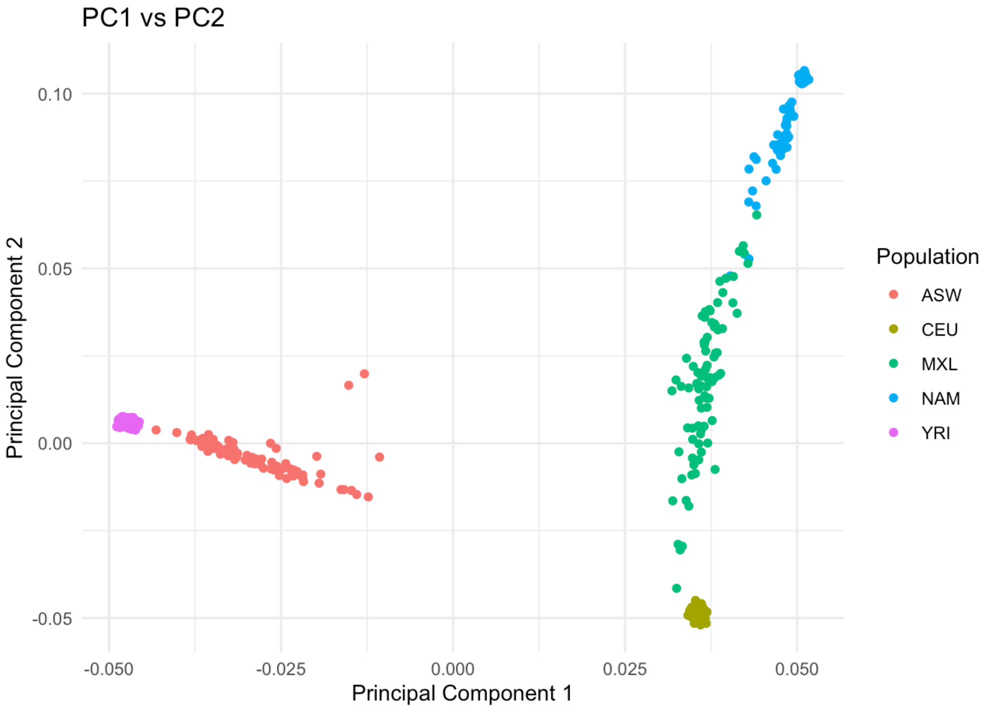
```
The default of 'doScale' is FALSE now for stability;
  set options(mc_doScale_quiet=TRUE) to suppress this (once per session) message
```

```
0 outlier variant detected..

Converged!
```





- Run PCA and make a scatter plot of the first two principal components (PCs) with each point colored according to population membership.

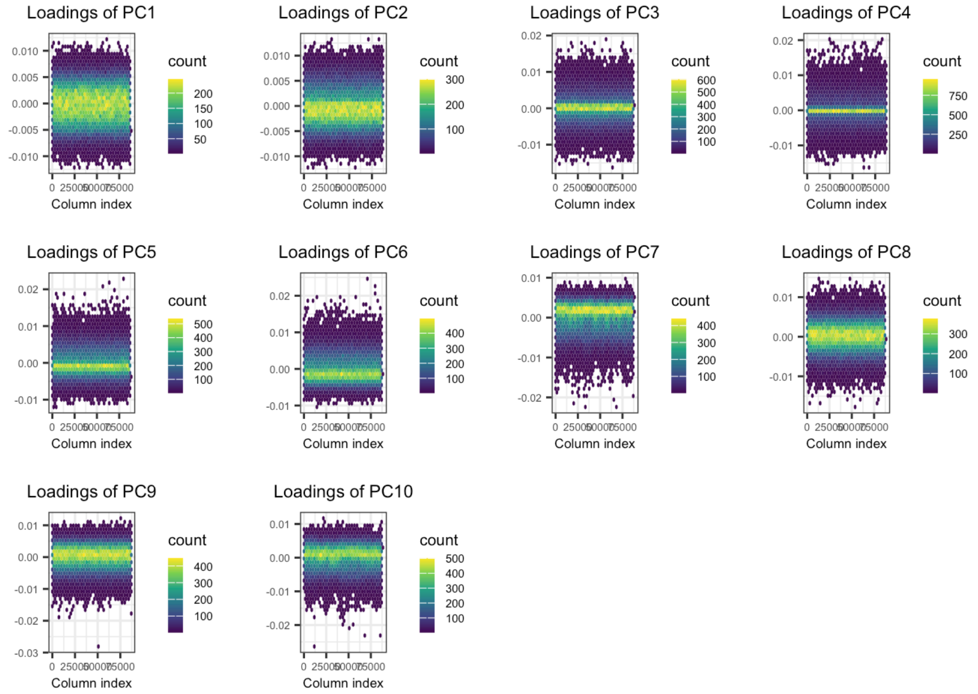


- Does the plot change from the one in Question 2?

```
[1] "No"
```

- Check the SNP loadings for the first 10 PCs.

```
$mfrow
[1] 1 1
```


Loadings of PC1

Loadings of PC2

Loadings of PC3

Loadings of PC4

Loadings of PC5

Loadings of PC6

Loadings of PC7

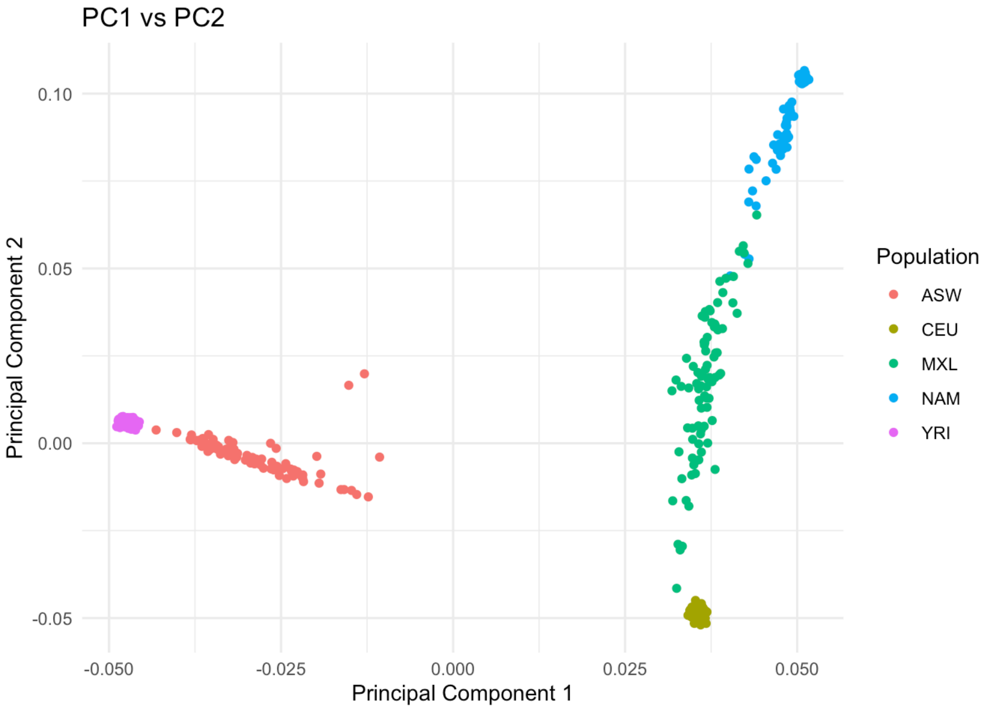
Loadings of PC8

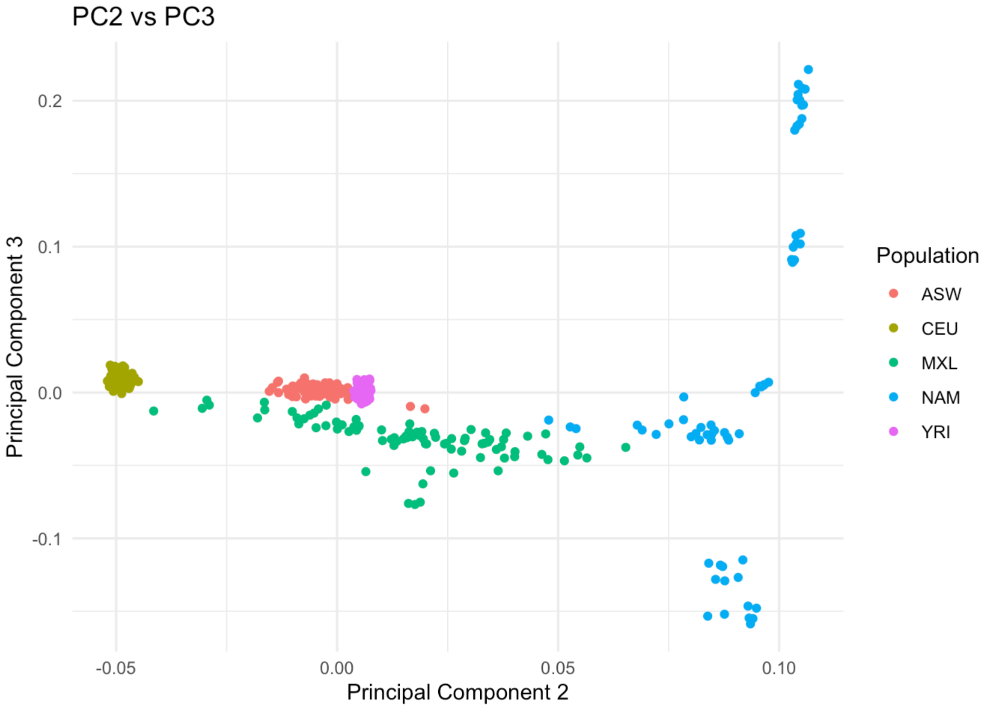
Loadings of PC9

Loadings of PC10

(Hint: This tutorial document (https://privefl.github.io/bigsnpr/articles/bedpca.html) from `bigsnpr` *might be helpful)*

4. Predict the proportional Native American and European Ancestry for the HapMap MXL from the PCA output in Question 3 *using one of the principal components*. (Which PC is most appropriate for this analysis?) Assume that the HapMap MXL have negligible African Ancestry.



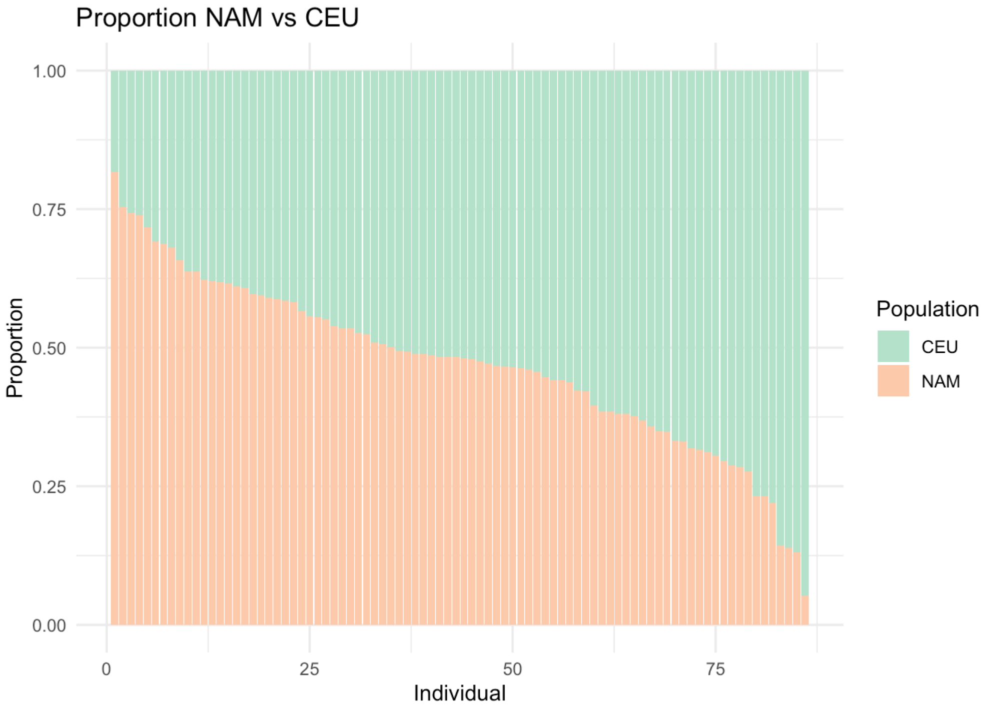


Best component is Principal Component 2 as MXL spans between Native American and European Ancestry on that component

```
            Mean
CEU -0.04885356
NAM  0.09084845
```

```
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 86 0.47 0.15   0.48    0.48 0.15 0.05 0.82  0.76 -0.3    -0.06 0.02
```

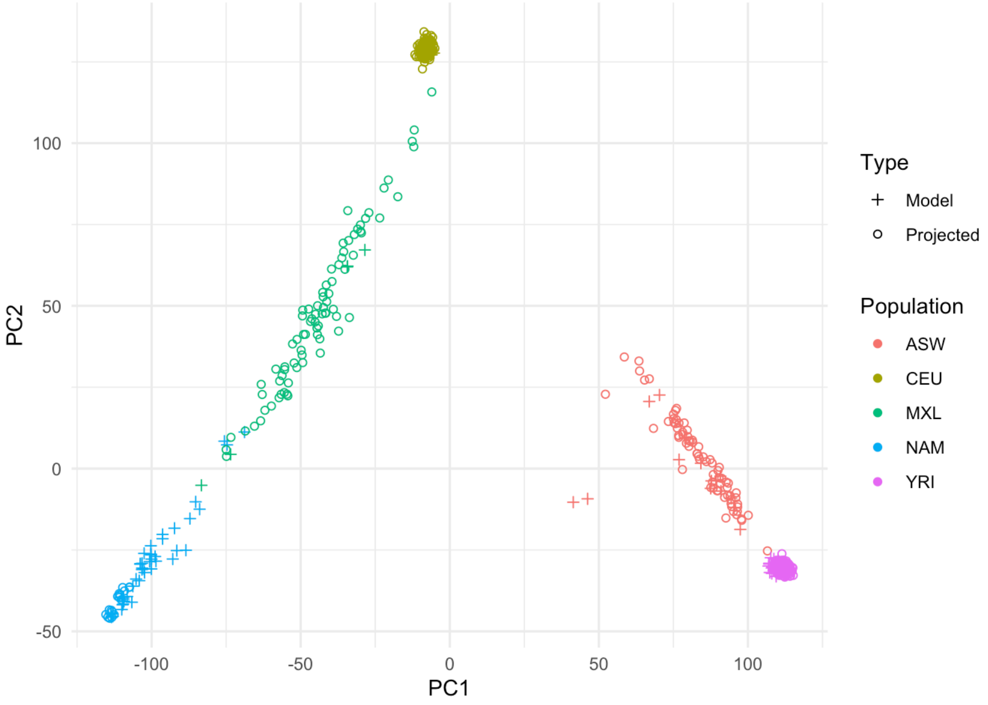
5. Make a barplot of the proportional ancestry estimates from question 4.



Extra: 6. Check if there are samples related 2nd degree or closer. If so, run PCA as in Question 3 removing these samples then project the remaining samples onto the PC space. The basic command would look like

```
int [1:116] 1 2 3 4 5 6 8 12 15 16 ...
```

```

Phase of clumping (on MAC) at r^2 > 0.2.. keep 77173 variants.
Discarding 12226 variants with MAC < 20.

Iteration 1:
Computing SVD..
0 outlier variant detected..

Converged!
```



Session information