R Notebook

Code

Mr. Geb

June 2020

Plink PCA of the HGDP data

First the HGDP text files were converted to `.ped` fomats and then the ped formats to `.bed`, `.bim` and `.fam` formats using a bash script below.

```
#!/bin/bash
set -x
dos2unix HGDP_FinalReport_Forward.txt
dos2unix HGDP_Map.txt
dos2unix SampleInformation.txt
head --lines=1 HGDP_FinalReport_Forward.txt > header.txt
awk '{for (i=1;i<=NF;i++) print "0",$i,"0","0"}' header.txt > hgdp_nosex.tfam
sed '1d' HGDP_FinalReport_Forward.txt > HGDP_Data_NoHeader.txt
sort -k 1b,1 HGDP_Data_NoHeader.txt > HGDP_Data_Sorted.txt
sort -k 1b,1 HGDP_Map.txt > HGDP_Map_Sorted.txt
join -j 1 HGDP_Map_Sorted.txt HGDP_Data_Sorted.txt > HGDP_compound.txt

awk '{if ($2=="M") $2="MT";printf("%s %s 0 %s ",$2,$1,$3);
    for (i=4;i<=NF;i++)
        printf("%s %s ",substr($i,1,1),substr($i,2,1));
    printf("\n")}' HGDP_compound.txt > hgdp.tped

# Add sex info
sed '1d' SampleInformation.txt > temp.txt
sed '$d' temp.txt > SampleInformation_noheader.txt
awk '{printf("HGDP%05d ",$1);
    if ($6=="m") print "1";
    else if ($6=="f") print "2";
    else print "0";}' SampleInformation_noheader.txt > Sample_sex.txt
awk 'BEGIN {
    while ((getline < "Sample_sex.txt") > 0)
        f2array[$1] = $2}
    {if (f2array[$2])
        print $1, $2, $3, $4, f2array[$2], "0"
    else
        print $2 "not listed in file2" > "unmatched"
    }' hgdp_nosex.tfam > hgdp.tfam

# convert to ped
plink --tfile hgdp --out hgdp --make-bed --missing-genotype - --output-missing-genotype 0

# Filter to 952 (or 940) people using the SampleInformation.txt file
awk '{if ($16=="1") printf("0 HGDP%05d\n",$1);}' SampleInformation_noheader.txt > Sample_keep.tx
t
plink --bfile hgdp --keep Sample_keep.txt --make-bed --out hgdp940
```

Then the following command line codes are used to produce egenvalues and eigenvectors of the HGDP data.

`plink --bfile hgdp --pca --out hgdp` generates the first 20 PCs (by default)

`plink --bfile hgdp --pca 10 --out hgdp` generates the first 10 PCs

The ***eigen vectors*** represent the directions in which the data has maximum variance while ***eigen values*** are the numbers that tell us how the data set is spread out on the eigen vector.

The ***Principal Component Analysis (PCA)*** represents the directions in which the data has maximum variance and also the directions in which the data is most spread out.

Therefore, we can say `eigen_vectors = direction of maximum variance (PCs)` and `eigen_values = magnitudes or percentage of variances (%tage of PCs)`

Reading in and cleaning data

Hide

```
library(tidyverse)
pops=read.csv("pops.csv", header = TRUE)
eigenvec <- as.matrix(read.table("hgdp/hgdp.eigenvec"))
eigenval <- scan("hgdp/hgdp.eigenval")
```

Hide

```
# Sort out the pca data
 pca <- eigenvec[, 3:12]
 names(pca)[1] <- "ind"
 names(pca)[1:ncol(pca)] <- paste0("PC", 1:(ncol(pca)))
```

Plotting the Percentage variance explained against the principal components

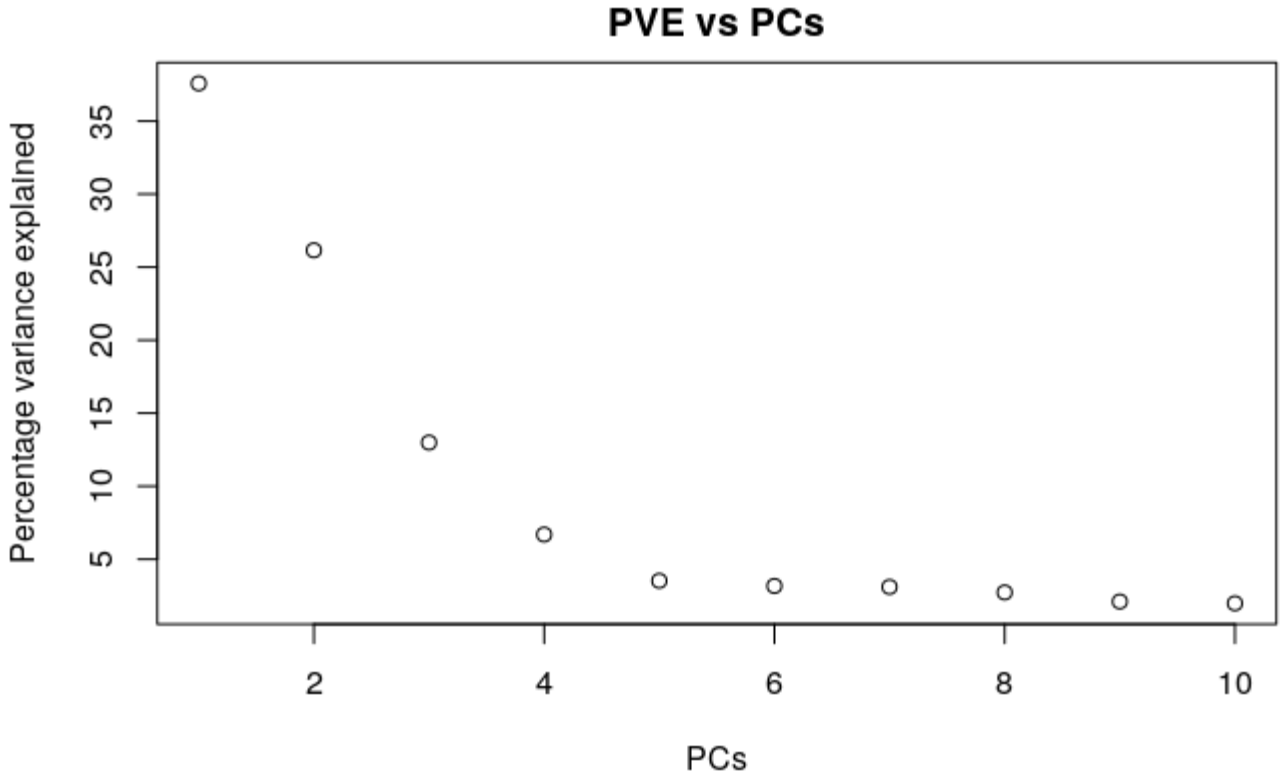
Hide

```
pve = eigenval/sum(eigenval)*100
# Rounding the pve values to two decimal digits
round(pve, 2)
```

```
 [1] 37.58 26.15 12.99  6.69  3.52  3.17  3.10  2.72  2.10  1.98
```

Hide

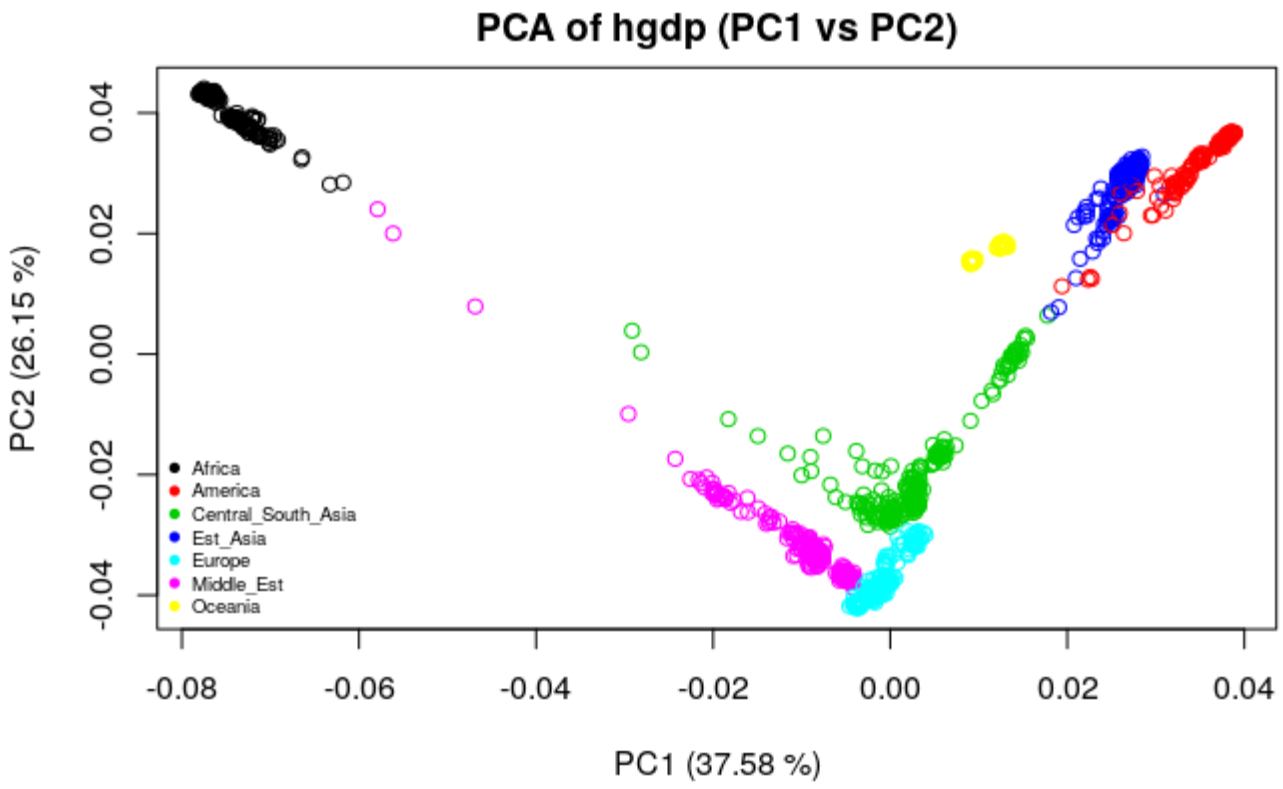
```
plot(pve, xlab = "PCs", ylab = "Percentage variance explained", main = "PVE vs PCs")
```



Plotting the first two principal components (PC1 versus PC2)

Hide

```
 plot(pca, col=pops$pop_group, xlab = paste("PC1 (37.58 %)"), ylab = paste("PC2 (26.15 %)"), mai
n ="PCA of hgdp (PC1 vs PC2)")
 legend("bottomleft", legend=levels(pops$pop_group),col = 1:7, pch = 19, bty = "n", cex = 0.6)
```

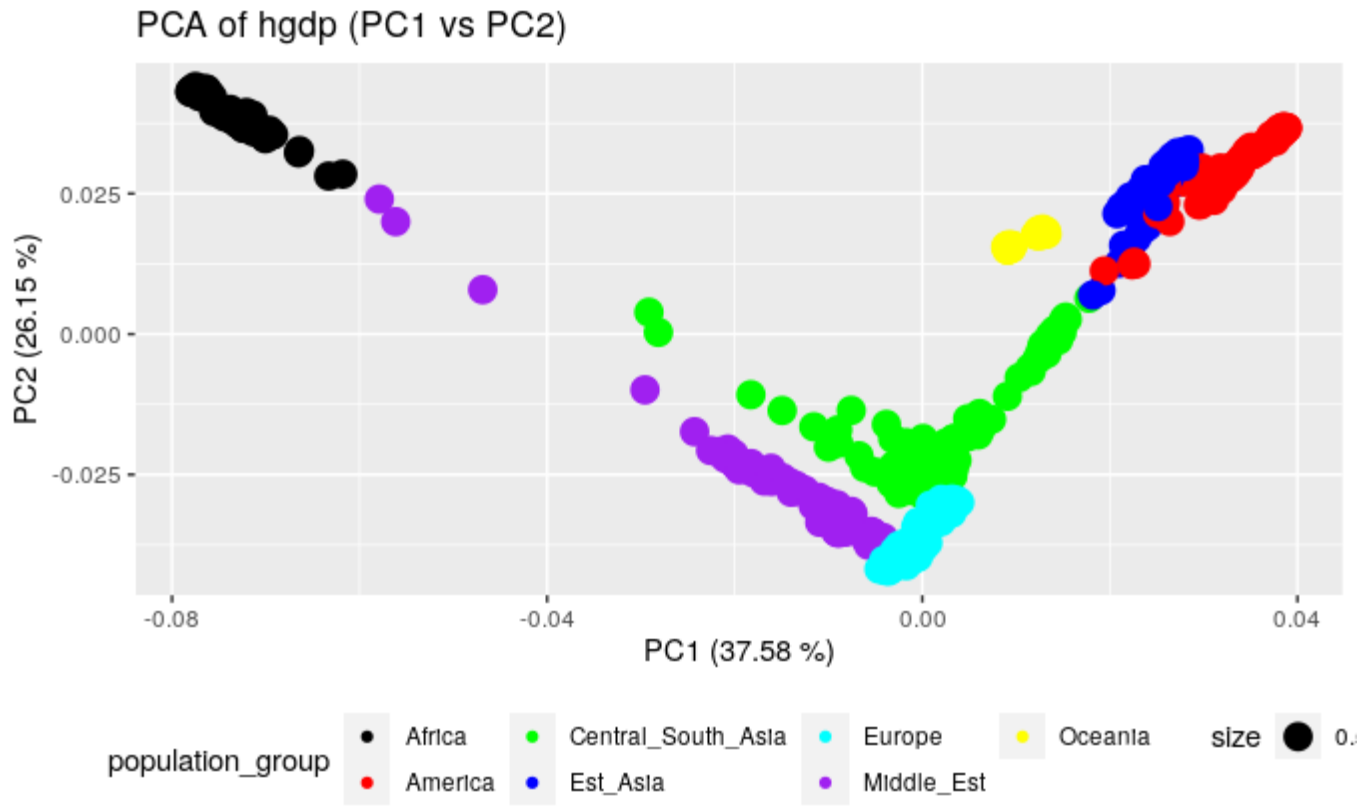


PC1 vs PC2 can also be plotted as follows:

Plotting the first and second principal components (PC1 versu PC2)

Hide

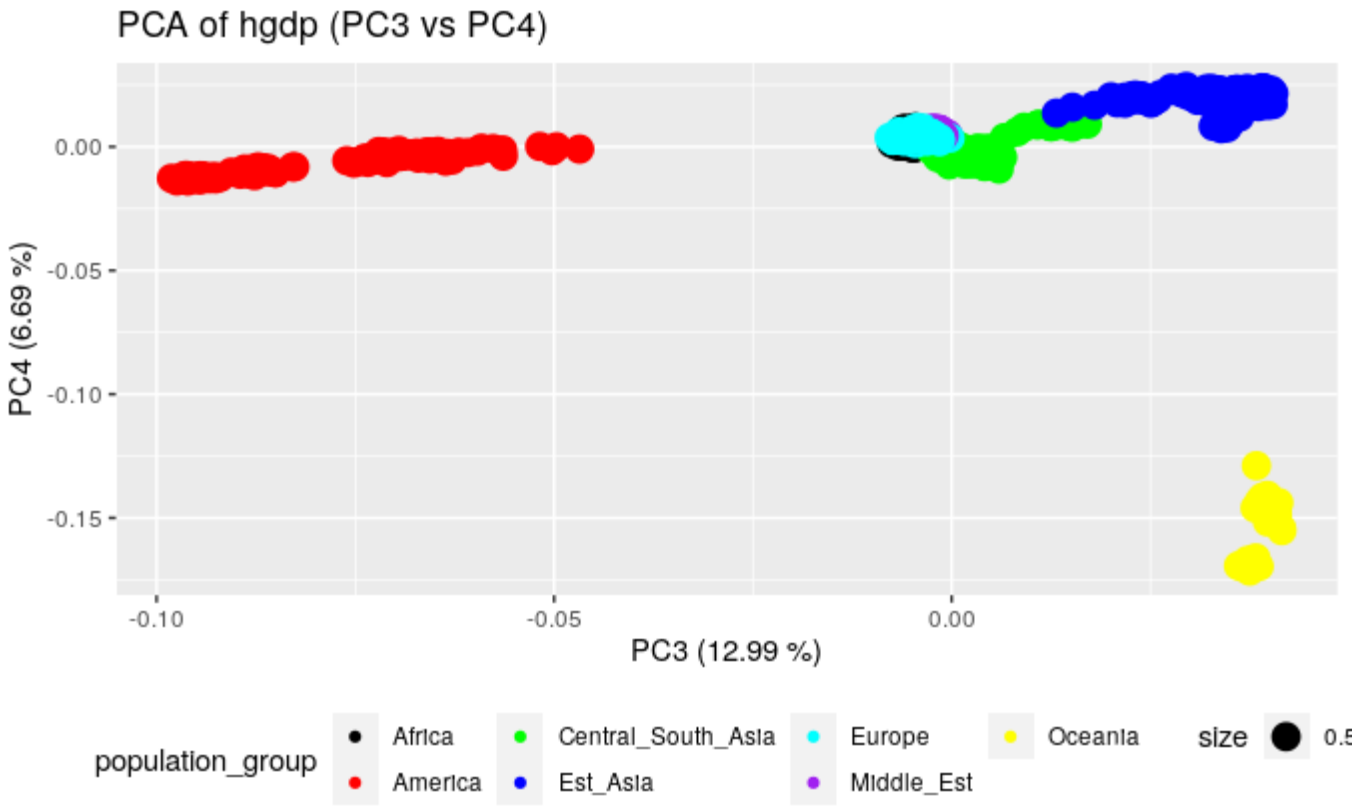
```
population_group = pops$pop_group
eigenvec<- read.table("hgdp/hgdp.eigenvec",sep=" ",header=F)
pca <- eigenvec[, 3:12]
# names(pca)[1:ncol(pca)] <- paste0("PC", 1:(ncol(pca)))
ggplot(data=pca, aes(V3,V4)) +
  geom_point()+
  geom_point(aes(col=population_group, size = 0.5))+
  labs(x = "PC1 (37.58 %)", y = "PC2 (26.15 %)")+
  scale_color_manual(values=c("black","red","green","blue","cyan","purple","yellow")) +
  theme(legend.position = "bottom")+
  ggtitle("PCA of hgdp (PC1 vs PC2)")
```



Ploting the third and fourth Principal Components (PC3 versu PC4)

Hide

```
population_group = pops$pop_group
eigenvec<- read.table("hgdp/hgdp.eigenvec",sep=" ",header=F)
pca <- eigenvec[, 3:12]
# names(pca)[1:ncol(pca)] <- paste0("PC", 1:(ncol(pca)))
ggplot(data=pca, aes(V5,V6)) +
  geom_point()+
  geom_point(aes(col=population_group, size = 0.5))+
  labs(x = "PC3 (12.99 %)", y = "PC4 (6.69 %)")+
  scale_color_manual(values=c("black","red","green","blue","cyan","purple","yellow"))+
  theme(legend.position = "bottom")+
  ggtitle("PCA of hgdp (PC3 vs PC4)")
```


PCA of hgdp (PC3 vs PC4)
0.00
-0.05
-0.10
-0.15
PC4 (6.69 %)
-0.10
-0.05
0.00
PC3 (12.99 %)
population_group
Africa
America
Central_South_Asia
Est_Asia
Europe
Middle_Est
Oceania
size
0.5