PCA on masked data

GOAL

Investigate relationships between Khoisan groups while ignoring differences that are due to recent admixture with Bantu-speaking farmers and East African pastoralists.

DATASET

Human Origins genotype data - subset of southern African Khoisan populations + Yoruba (West African) + Somali (East African). Download the data here: https://share.eva.mpg.de/index.php/s/rcxeQW6EDHaCxe7

MASKING

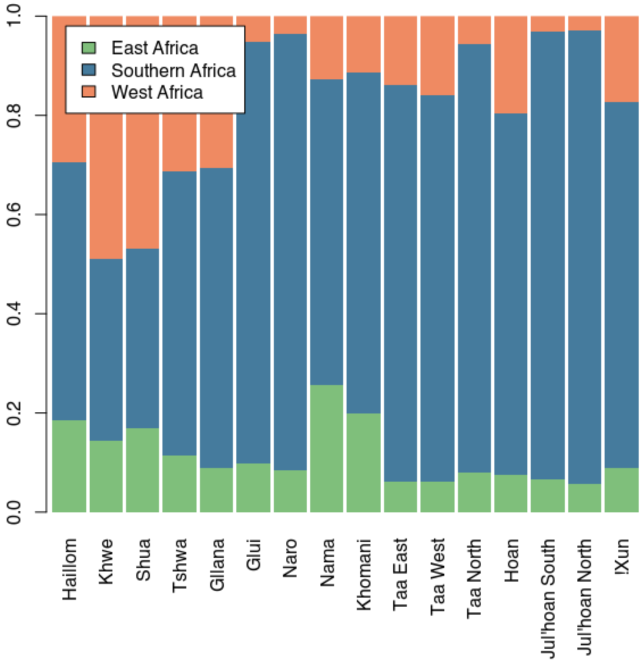
1. Local ancestry was performed in RFMIX v2 using 3 source groups:

- 13 West Africans
- 13 East Africans
- 13 Southern Africans (the least admixed Khoisan)

All other Khoisan were used as target.

2. We kept all sites belonging to one specific ancestry (with a marginal probablity > 1) and masked all other sites by converting them to missing. Diploids (ind1) converted to haploids (ind1.0 and ind1.1) to allow file manipulation with plink or other tools.

The following plot shows the inferred global ancestry proportions



EXERCISE 1 – Confirm that the local ancestry returned reasonable results

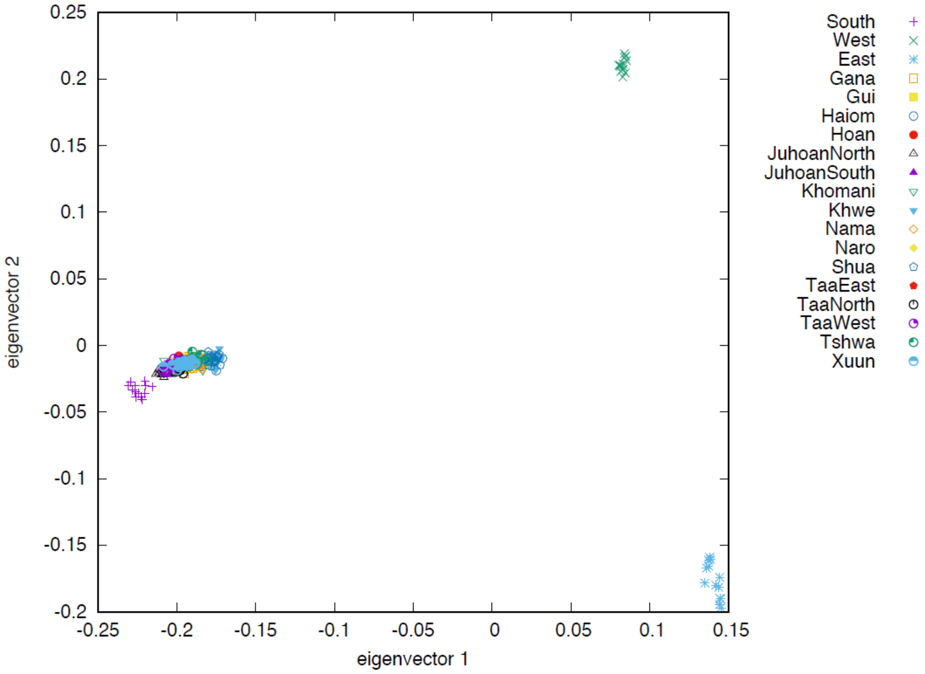
The data from each masked file was previously merged with non-masked data from the 3 sources. Compute a PCA on the 3 sources and project the target populations for each ancestry.

```
$ less parfile_smartpca_south_lsq
```

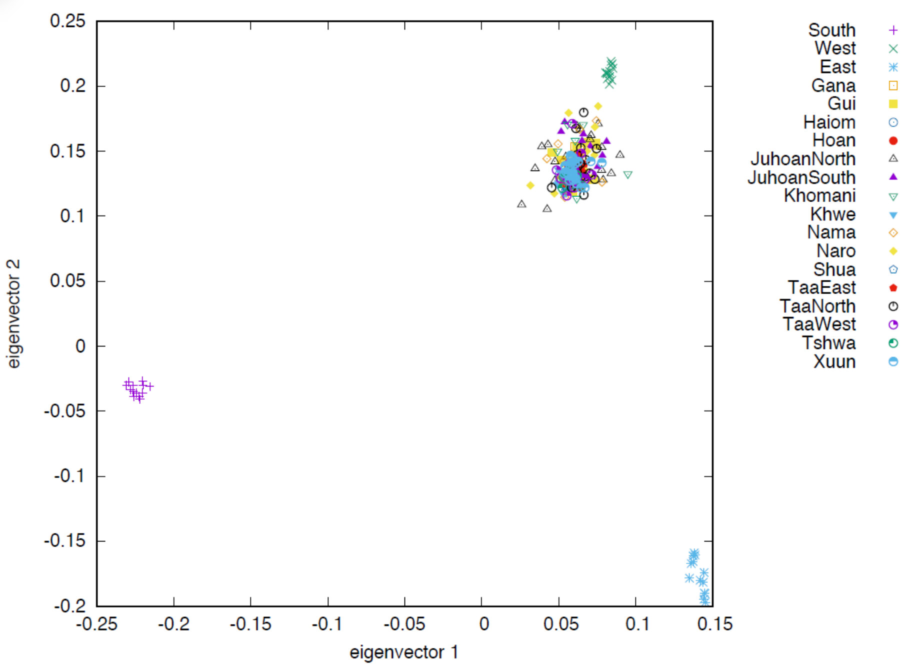
```
genotypename: south_africa_ml1_3sourcetest.geno
snpname: south_africa_ml1_3sourcetest.snp
indivname: south_africa_ml1_3sourcetest.ind
evecoutname:  south_africa_ml1_3sourcetest_evec
evaloutname:  south_africa_ml1_3sourcetest_eval
poplistname: 3sourcetest.pop
numoutevec: 4
numoutlieriter: 0
lsqproject:  YES
```

```
$ smartpca -p parfile_smartpca_south_lsq
```

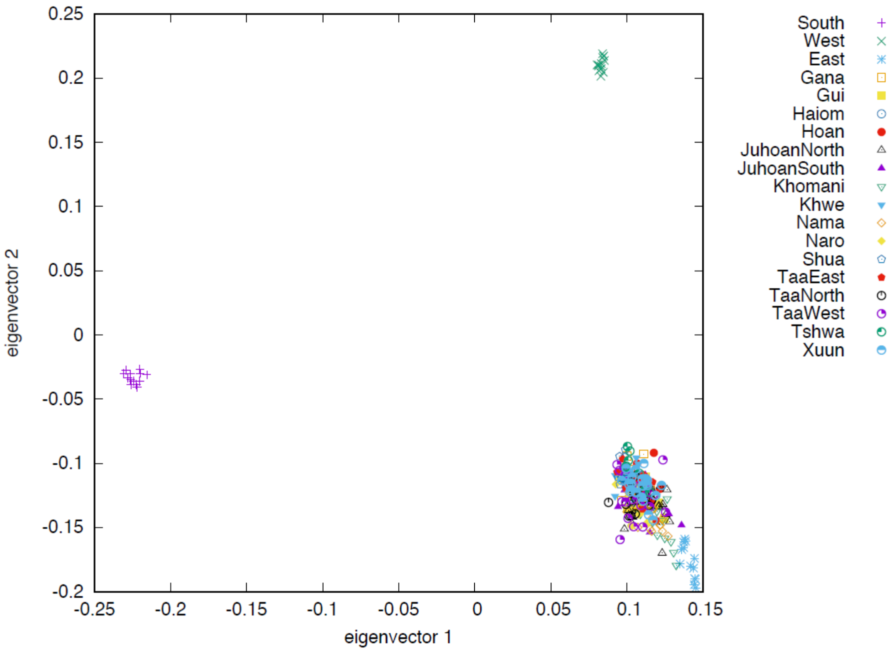
```
$ ploteig -i south_africa_ml1_3sourcetest_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o south_africa_ml1_3sourcetest.xtxt
```



```
$ smartpca -p parfile_smartpca_west_lsq
$ ploteig -i west_africa_ml1_3sourcetest_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o west_africa_ml1_3sourcetest.xtxt
```

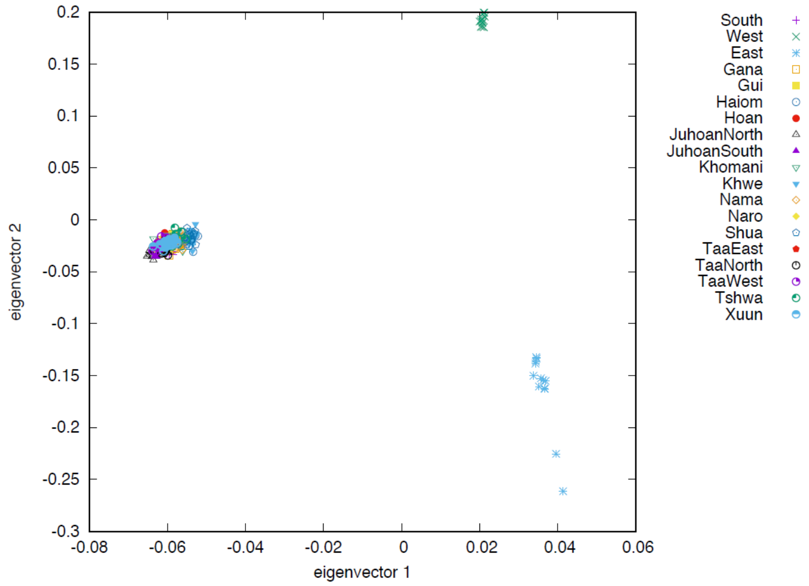


```
$ smartpca -p parfile_smartpca_east_lsq
$ ploteig -i east_africa_ml1_3sourcetest_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o east_africa_ml1_3sourcetest.xtxt
```

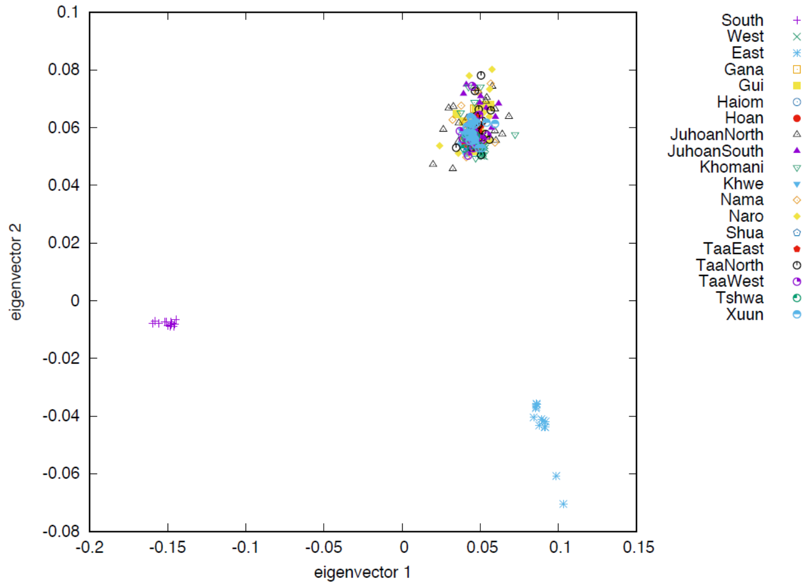


Try the option “shrinkmode: YES”!

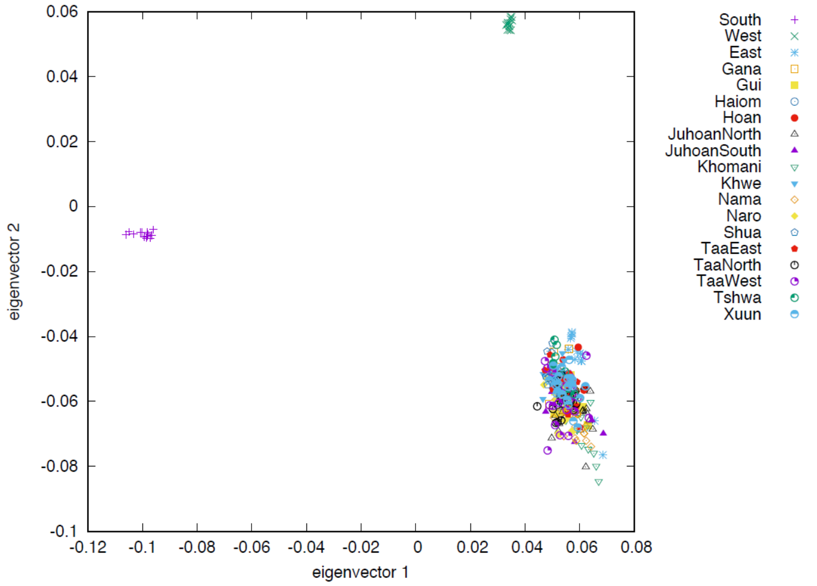
```
$ smartpca -p parfile_smartpca_south_lsq_shrink
$ ploteig -i south_africa_ml1_3sourcetest_shrink_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o south_africa_ml1_3sourcetest_shrink.xtxt
```



```
$ smartpca -p parfile_smartpca_west_lsq_shrink
$ ploteig -i west_africa_ml1_3sourcetest_shrink_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o west_africa_ml1_3sourcetest_shrink.xtxt
```



```
$ smartpca -p parfile_smartpca_east_lsq_shrink
$ ploteig -i east_africa_ml1_3sourcetest_shrink_evec -c 1:2 -p \
South:West:East:Gana:Gui:Haiom:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Khwe:Nama:Naro:Shua:TaaEast:\
TaaNorth:TaaWest:Tshwa:Xuun -x -o east_africa_ml1_3sourcetest_shrink.xtxt
```

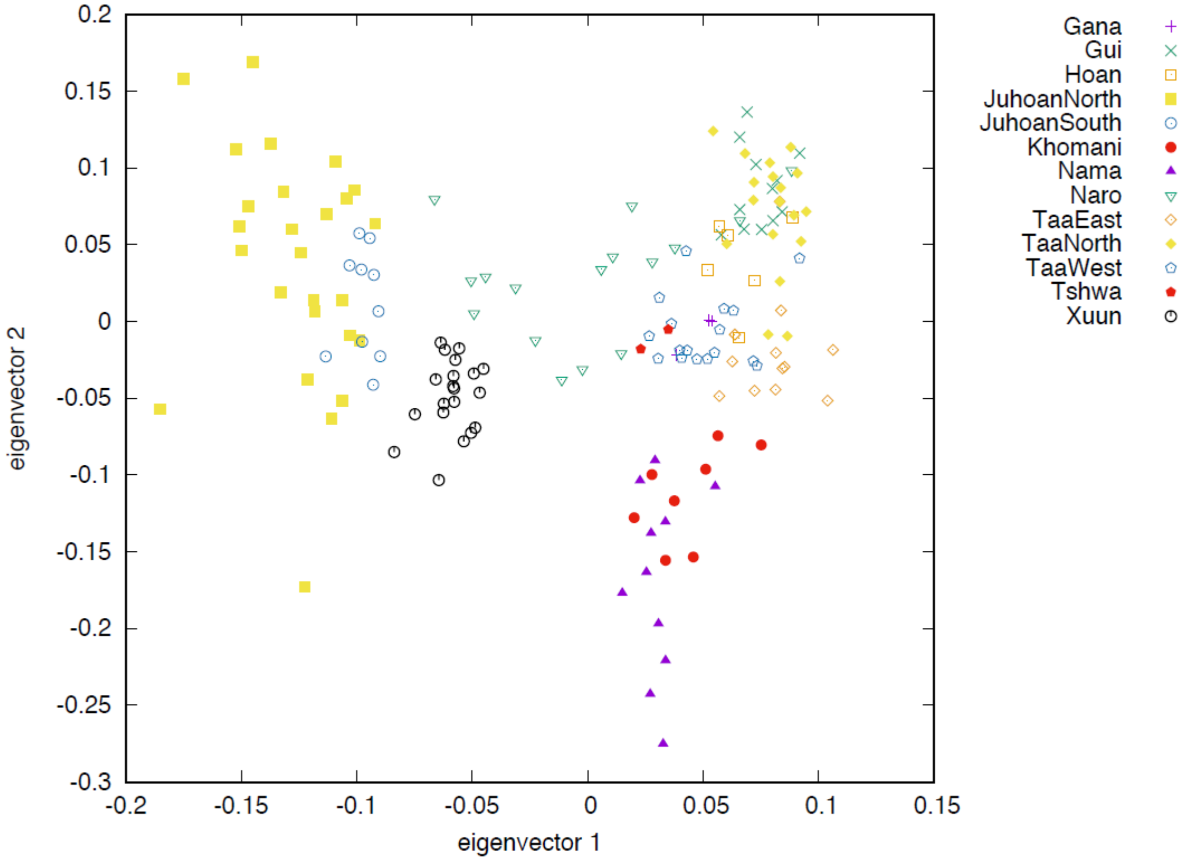


EXERCISE 2 – Do a PCA for the Khoisan-specific ancestry

The file with Khoisan ancestry was filtered for missingness at the individual level (ind >35% missing removed) and at the SNP level (SNPs with more than 15% missing removed).

```
$ smartpca -p parfile_smartpca_south_keepmiss35
```

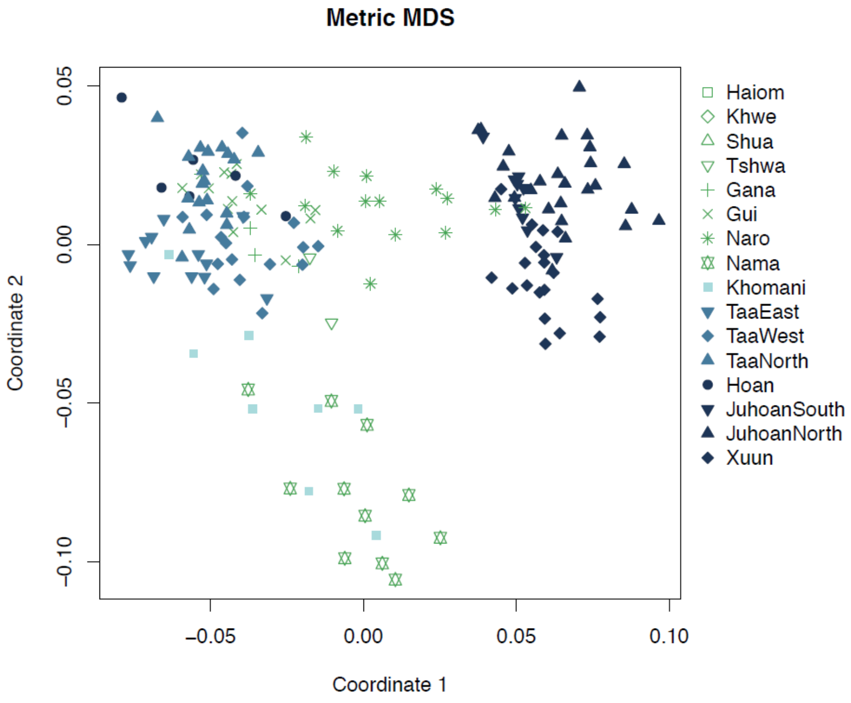
```
$ ploteig -i south_africa_ml1_keepmiss35_evec -c 1:2 -p \
Gui:Hoan:JuhoanNorth:JuhoanSouth:Khomani:Nama:Naro:TaaEast:TaaNorth:TaaWest:Tshwa:Xuun -x -o \
south_africa_ml1_keepmiss35.xtxt
```



EXERCISE 3 – Do a MDS for the Khoisan-specific ancestry

```
#R code

#The matrix of pairwise differences between all Khoisan individuals is provided
ind_mat=read.table("pwdiff_khoisan_ind_mat.csv", sep=",", head=T, row.names = 1)

#color and pch per population are already defined
info=read.table("info_southernafrica.csv", sep=",", head=T, as.is=T,quote = "\"", comment.char = "")

#extract population names from the ids
pop_list=unlist(lapply(colnames(ind_mat), function(x){strsplit(x, split="_")[[1]][1]}))
colvec=vector()
pchvec=vector()
for(i in 1:length(pop_list)){
  colvec[i]=info[which(info[,1]==pop_list[i]),4]
  pchvec[i]=info[which(info[,1]==pop_list[i]),5]
}

#compute MDS
d <- dist(ind_mat) # euclidean distances
```

```
fit <- cmdscale(d,eig=TRUE, k=2) # k is the number of dim
x <- fit$points[,1]
y <- fit$points[,2]

#plot results
pdf("Khoisan_MDS.pdf",height= 7,width=8)
par(oma=c(2,2,2,8), xpd=NA)
plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2", main="Metric MDS", pch=pchvec, col=colvec, bg=colvec)
legend("topright", inset=c(-0.3,0), info[,1], col=info[,4], pch=info[,5], pt.bg=info[,4],  bty="n")
dev.off()
```


Metric MDS
Coordinate 2
Coordinate 1
□ Haiom
◇ Khwe
△ Shua
▽ Tshwa
+ Gana
× Gui
✳ Naro
✡ Nama
■ Khomani
▼ TaaEast
◆ TaaWest
▲ TaaNorth
● Hoan
▼ JuhoanSouth
▲ JuhoanNorth
◆ Xuun