Population Structure using PCA

We are going to use a small subset of HashMap project data: 279 samples of 9026 SNPs taken from 4 different populations—Centreal European (CEU), African (YRI), Japanese (JPT) and Chinese (HCB). We're going to see if PCA can identify population structure.

```
library(ade4)
```

```
## Warning: package 'ade4' was built under R version 3.2.4
```

```
library(adegenet)
```

```
##
##    /// adegenet 2.0.1 is loaded ////////////
##
##    > overview: '?adegenet'
##    > tutorials/doc/questions: 'adegenetWeb()'
##    > bug reports/feature requests: adegenetIssues()
```

```
e <- read.csv("data/genotype_population.csv", row.names=1)
tab <- read.csv("data/population_info.csv", row.names=1)
head(t(e))[,1:5] # rows represent genes
```

```
##            NA19152 NA19139 NA18912 NA19160 NA07034
## rs1695824        2       1       2       2       0
## rs13328662       1       1       1       1       0
## rs4654497        0       0       1       1       0
## rs10915489       1       1       2       2       2
## rs12132314       2       2       2       2       2
## rs12042555       1       2       2       1       2
```

Standardizing data

We begin by standardizing each SNP to have 0 mean and standard deviation of 1.

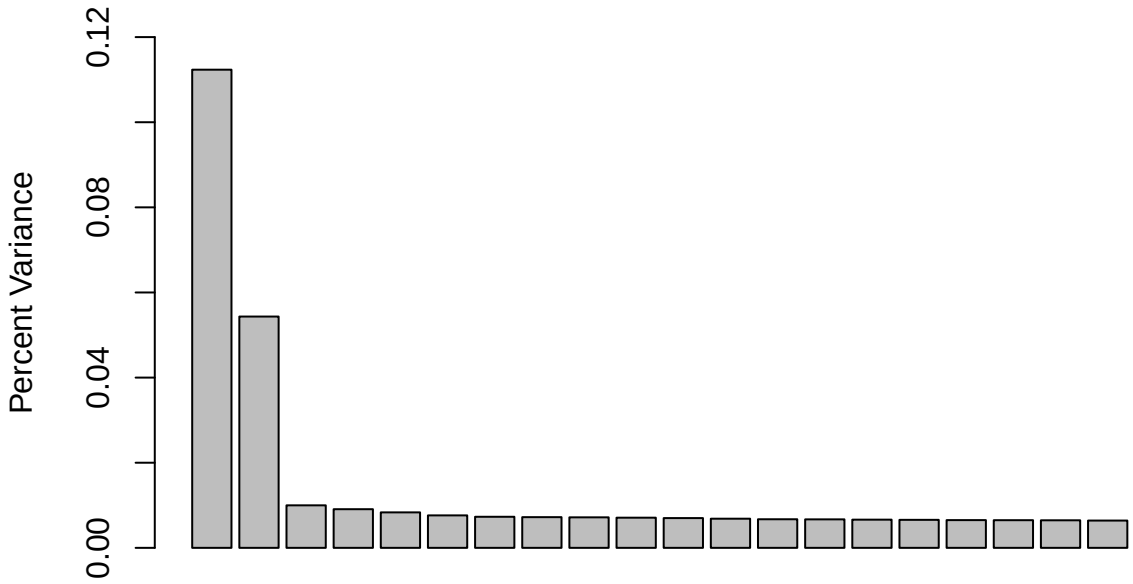
```
E <- scale(e) # standardize
```

Dimensionality reduction

Dimension reduction is performed to better visualize data. Say that the original data is represented by n data points. The goal is to reduce this to a subspace of d points while mainting the variability in data. The new subspace is represented by d orthogonal vectors called principal components, and $d < n$.

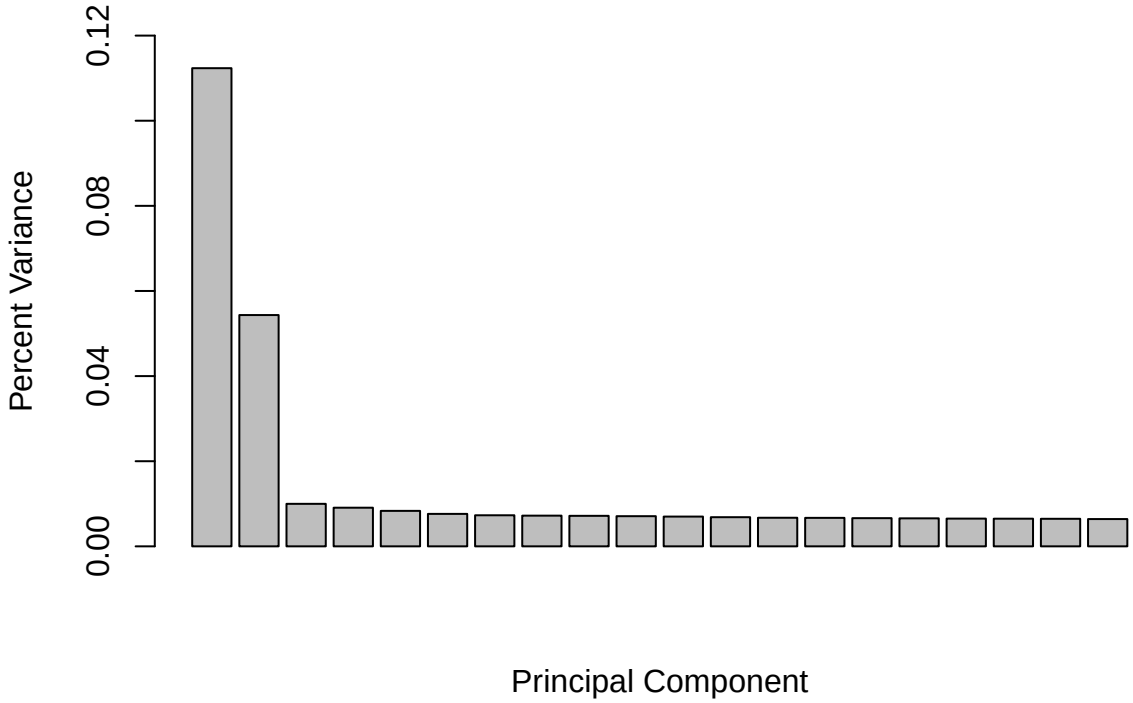
Let us see dimension reduction over feature space.

```
pca <- dudi.pca(t(E), center=FALSE, scale=FALSE, scannf=FALSE, nf=20) # Feature space
eig.perc <- pca$eig/sum(pca$eig) # Percent of variance
barplot(eig.perc[1:20], ylim=c(0,0.12), ylab="Percent Variance", xlab="Principal Component")
```



As you can see, most of the variance in the data is expressed by the first two components. Similarly, we can perform dimension reduction over sample space.

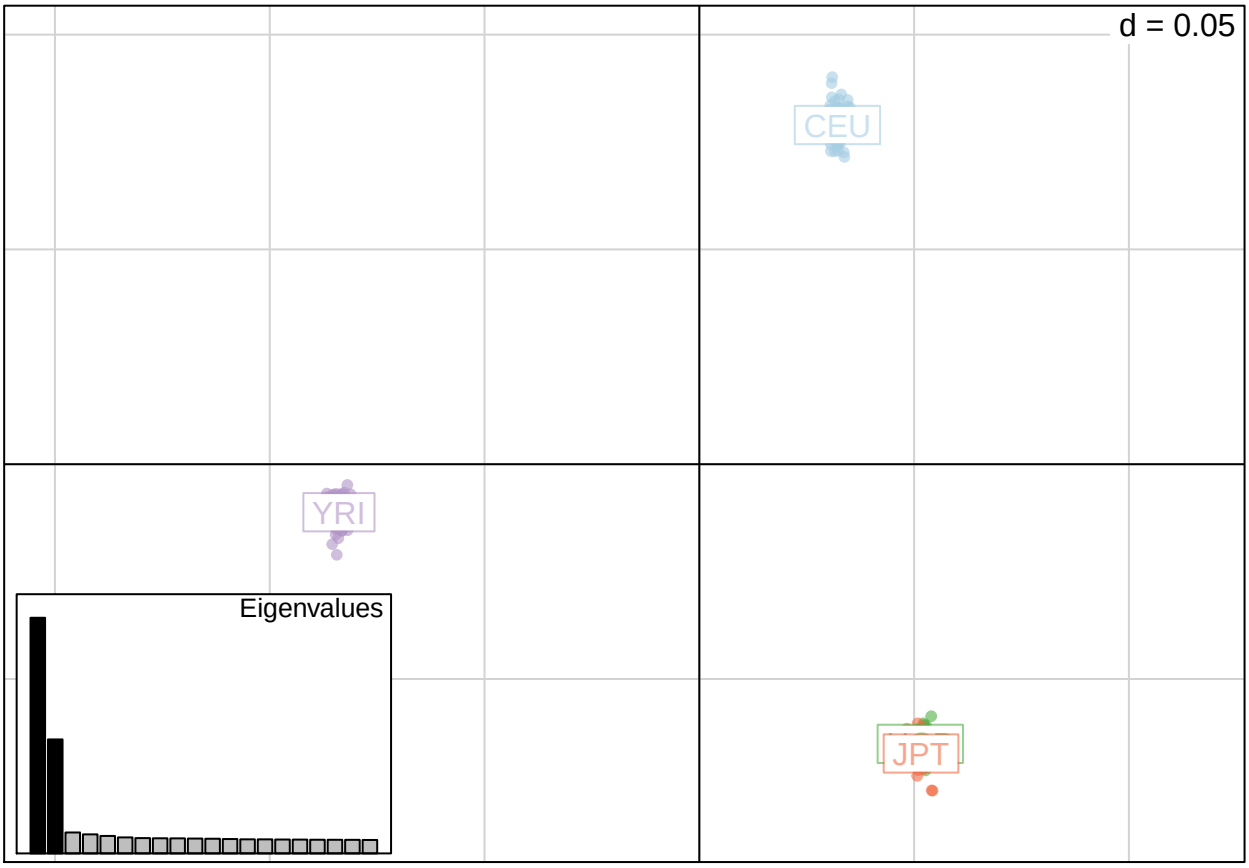
```
pca2 <- dudi.pca(E, center=FALSE, scale=FALSE, scannf=FALSE, nf=20) # Sample space
eig.perc2 <- pca2$eig/sum(pca2$eig) # Percent of variance
barplot(eig.perc2[1:20], ylim=c(0,0.12), ylab="Percent Variance", xlab="Principal Component")
```



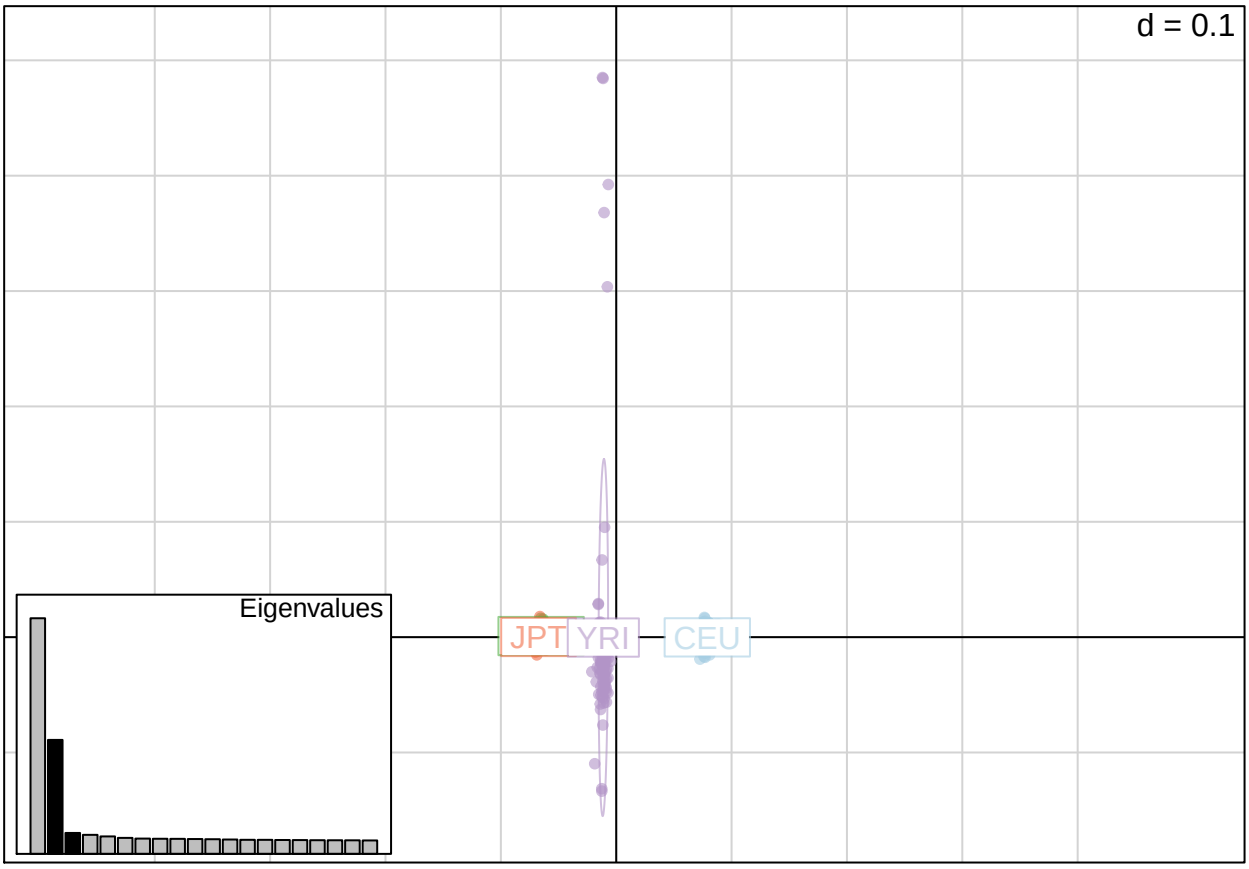
Scatter plots

In the first two plots (*PC1 vs PC2* and *PC2 vs PC3*), we see **three** distinct population clusters. The last two plots do not show any population structure.

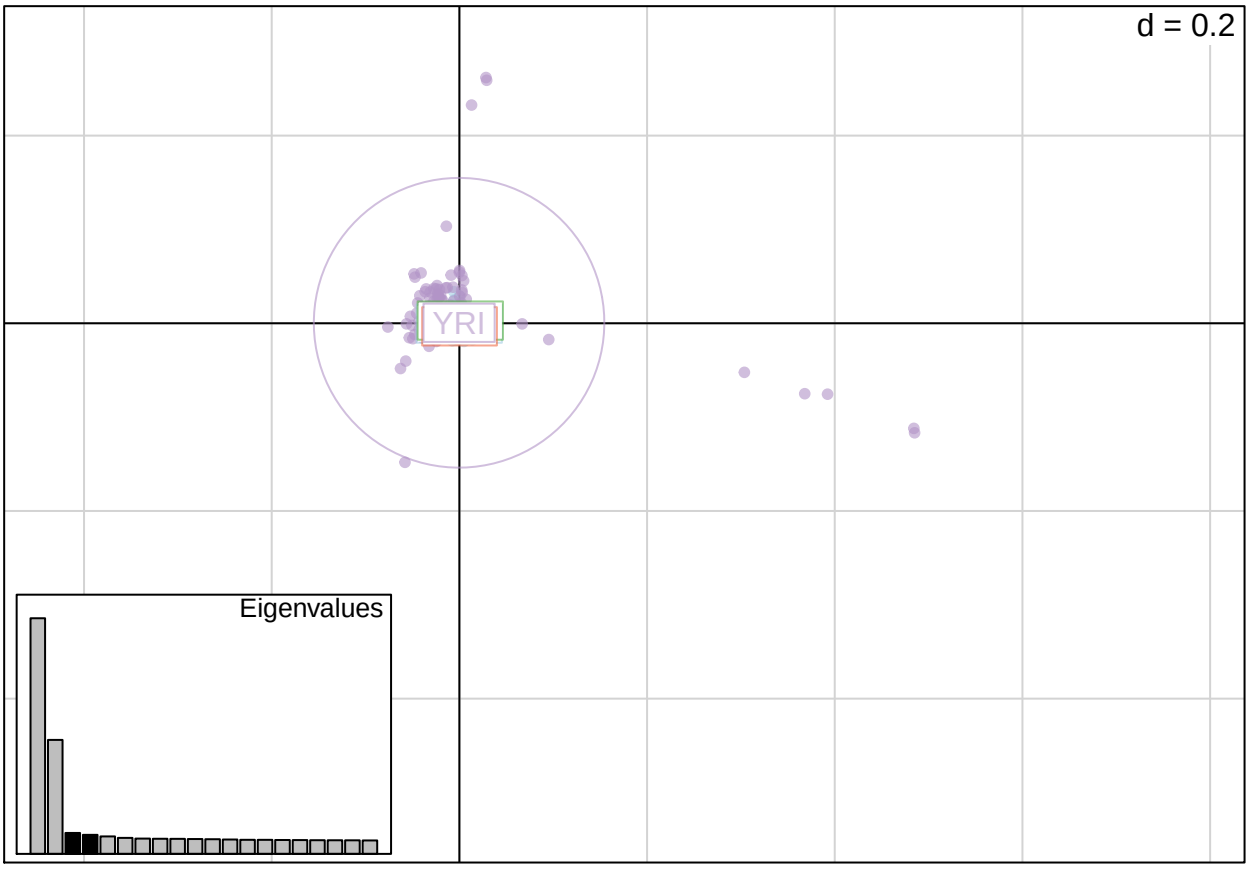
```
s.class(pca$c1[,1:2], fac=tab$V2, col=transp(funky(5),.6),
        axesell=FALSE, cstar=0, cpoint=1)
add.scatter.eig(pca$eig[1:20],20,1,2, ratio=.3)
```



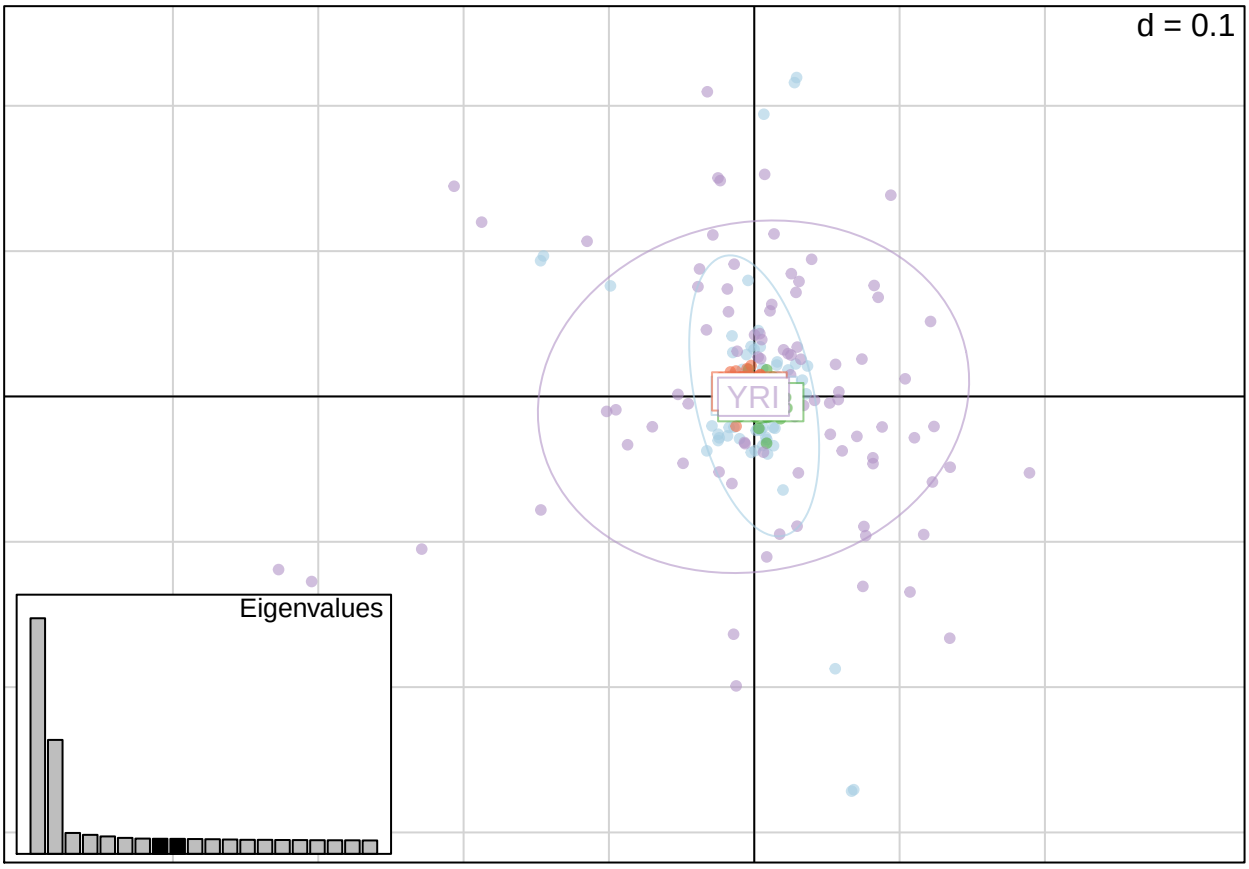

```
s.class(pca$c1[,2:3], fac=tab$V2, col=transp(funky(5),.6),
        axesell=FALSE, cstar=0, cpoint=1)
add.scatter.eig(pca$eig[1:20],20,2,3, ratio=.3)
```




```
s.class(pca$c1[,3:4], fac=tab$V2, col=transp(funky(5),.6),
        axesell=FALSE, cstar=0, cpoint=1)
add.scatter.eig(pca$eig[1:20],20,3,4, ratio=.3)
```




```
s.class(pca$c1[,8:9], fac=tab$V2, col=transp(funky(5),.6),
        axesell=FALSE, cstar=0, cpoint=1)
add.scatter.eig(pca$eig[1:20],20,8,9, ratio=.3)
```


d = 0.1
YRI
Eigenvalues