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(rafalib)
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.

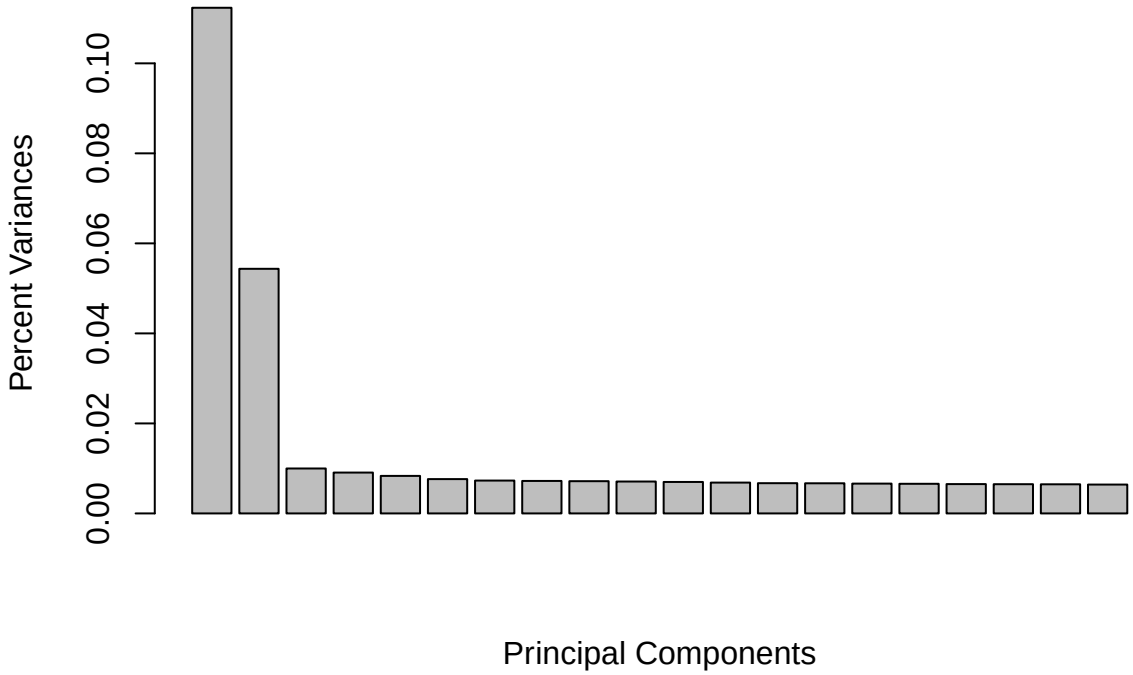
```
E <- scale(e) # standardize
```

Dimension reduction

Dimension reduction is performed to better visualize data. We can use principal components analysis (PCA) for this. 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 <- prcomp(t(E), center=FALSE, scale=FALSE)
pca.var <- pca$sdev^2
pca.percent_var <- pca.var/sum(pca.var)
barplot(pca.percent_var[1:20], ylab="Percent Variances", xlab="Principal Components") # Feature space
```



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

Scatter plots