Session 5: Population Structure Inference

Inferring Genetic Ancestry with Principal Components Analysis in R

Before you begin:

- Make sure that the latest version of R is installed on your computer:

  http://cran.r-project.org/mirrors.html

- Install the Bioconducter R packages "gdsfmt" and "SNPRelate" on your laptop by first loggin on to the internet. Then open an R session and enter the following;

```
source("http://bioconductor.org/biocLite.R")
biocLite("gdsfmt")
biocLite("SNPRelate")
```

The file "YRI_CEU_ASW_MEX_NAM.bed" is a binary file in PLINK format (as well as corresponding ".bim" and ".fam" files) that is available on the short course website. The file contains genotype data at autosomal SNPs for Native American samples from the Human Genome Diversity Panel (HGDP) and 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), and African Americans from the south-western United States (ASW).

1. Perform a PCA in R for the sample using all of the SNPs (Note: use the R script provided on the short course website). Make a scatterplot of the first two principal components (PCs) with each point colored according to population membership. Interpret the first two PCs? What ancestries are the PCs reflecting?

2. Now redo question 1 above using a subset of SNPs that are selected based on having a correlation (or linkage disequilibrium [LD]) between SNPs that is no greater than 0.2 (see the commands provided in the R script).

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

4. Now make a barplot of the proportional ancestry estimates from question 3. Compare your plot to the supervised model-based ancestry barplot for the MXL provided on the short course website.