92.651: Homework 2

Alexander Frieden

March 24, 2016

1. Suppose we have a sample of 2000 people from the UK who were typed for the alleles at two different blood group loci: (i) the MN blood group and (ii) the Ss blood group. Let's use p and q to denote the frequencies of the alleles at the MN locus and x and y for the frequencies of alleles at the Ss locus. The frequencies are as follows:

MN blood group

fM = p = 0.5425
fN = q = 0.4575

Ss blood group

fS = x = 0.3080
fs = y = 0.6920

Gamete frequencies

MS = 474/4000 = 0.1185
Ms = 611/4000 = 0.15275
NS = 142/4000 = 0.0355
Ns = 773/4000 = 0.19325

What is the likage disequilibrium of these alleles?

2. Given the following observed genotype frequencies:

MN Locus	Ss locus
MM=298	SS=483
MN=489	Ss=418
NN=213	ss=99

show that the MN and Ss loci are in Hardy-Weinberg equilibrium.

3. Explain what it means for the first principal component "explains 20% of the variation"

4. Generate a simulated data set with 20 observations in each of the three classes (60 observations total) and 50 variables. *Hint: In R you can use rnorm() or runif() to generate data for their corresponding distributions. Be sure to add a mean shift to the observations in each class so there are three distinct classes*

5. Perform PCA on the 60 observations and plot the first two prinicpal component score vectors. Use a different color to indicate the observations in each of the three classes. Do this until you see a seperation in the three classes. Explain how you were able to get this seperation.

6. From the data from problem (5), what is the percent of the variance observed seen by the first principal component? How about the second? From these numbers, what conclusion can you draw about the effectiveness of PCA on your dataset?

BONUS: Use Google Genomics to Compute Hardy Weinberg on variants in the CFTR gene. What do you observe? Can the results be broken down by ethnicity? If so, what ethnicity?