HW #6 due Thursday, Feb. 21

Note: I made a mistake. Discard the file we used in discussion on Thursday. Download the file "Corrected Data for HW 6" from Content.

This file is a subset of the 1,000 Genome data set. The format is the same as we discussed in class (rows are SNPs, columns are individuals, see links from Lecture 10). Columns 1-9 (R starts with 1) are details about the SNPs, columns 10 to 90 are unrelated individuals from Europe, columns 91 to 179 are unrelated individuals from Africa.

Each individual has two haplotypes. Write an R function that takes as input two individuals and outputs the pairwise nucleotide diversity (number of sites that differ) between one haplotype from each of these individuals (from each individual you can choose the one haplotype randomly). Write another R function that takes as input a set of individuals and a number *n* and calculates the average pairwise diversity for *n* randomly chosen pairs from the set.

Using *n=100*, calculate the average pairwise diversity within the European population. Do this 10 times and report the mean and standard deviation of these 10 runs. Repeat for the African population. Finally, repeat when the pairs are such that one individual is European and the other individual is African.

Is average pairwise nucleotide diversity greater in the European population or the African population? Turn in your results and your code.