Assignment 6

Due Wednesday, October 7 before midnight (California time)

On Blackboard you can find a file named: “abbgen1k.csv”. This file is a subset of the 1,000 Genomes Project for chromosome 22. The format is the same as we discussed in class: rows are SNPs, columns 1 to 9 (R starts with 1) are details about the SNPs, columns 10 to 90 are unrelated individuals from Europe, and columns 91 to 179 are unrelated individuals from Africa.

Compute the average pairwise diversity within the European samples. Also, compute this within the African samples, and between the European and African samples. Due to the out-of-Africa hypothesis, we expect that the average pairwise diversity (average number of different SNPs between two individuals) within Europe will be less than that within African.

We expect you to complete the following functions in R (you can design your functions without following the instructions to achieve the same task):

1. Write an R function that inputs a vector representing the genotype of an individual and a logical indicating “left” or “right” haplotype. This function returns either the “left” or “right” haplotype as a vector. [2pt].

2. Write an R function that computes the average pairwise diversity for n randomly chosen pairs of haplotypes (use the built-in R function ”sample”). These pairs of haplotypes can both be chosen from one sample, or they can be chosen from two different samples. The inputs to this function are a dataframe with the format in “abbgen1k.csv”, a vector of column numbers for individuals in the first sample, a vector of column numbers for individuals in the second sample, and a number n [3pt].

3. Run Function #3 for three comparisons including within the European samples (with the input the dataframe columns 10:90 and 10:90), the African samples (with the input the dataframe columns 91:179 and 91:179), and between the European and African samples (with the input dataframe columns 10:90 and 91:179). Set the parameter n equal to 1000. Tell us what you find from the results and explain whether the results can support the out-of-Africa hypothesis [3pt].

Turn in the code for the aforementioned R functions and the answers into one file in Jupyter Notebook format (.ipynb). Use the “Turnitin” link on Blackboard/Assignments/Assignment 6 to submit this file.