BIO3PG3 PROBLEM SET

Due Date: Friday February 1st. 5:00 PM.

Question 1. Using the associated R script called `ProblemSet2.R` on "http://helix.mcmaster.ca/courses.html/" place your student number in the setseed function. Then run the function called `SequenceGenerator`. Using the sample of 5 sequences it has generated for you, calculate the number of segregating sites (S), the value of π (average number of pairwise differences), the number of haplotypes and $\bar{H}_{exp}$ for these sequences.

Question 2. Using the code in the same R script for question 2 it will generate some allele numbers for you. Either using a function, such as that developed in tutorial, or by hand, estimate allele frequencies, expected genotypic frequencies and whether this deviates from Hardy Weinberg Equilibrium using a χ^2 test. Please show your work (R script or written).

Question 3. A population survey for a microsatellite locus (with 3 alleles) in a human population resulted in the following numbers of different genotypes:

$A_1A_1 = 13$
$A_1A_2 = 48$
$A_1A_3 = 108$
$A_2A_2 = 22$
$A_2A_3 = 258$
$A_3A_3 = 391$

Please calculate allele frequencies and determine expected genotypic frequencies. You can do this by hand or you can modify the function we wrote in R. Either way, show your work (or function).

Question 4. Part A: Below are some blood type data for the Navaho (US) and Koori (Aust) populations.

	MM	MN	NN	Sum	f(M)	f(N)
Navaho (US)	305	52	4	361	0.92	0.08
Koori (Aus.)	22	216	492	730	0.18	0.82
Total	327	268	496	1091	0.42	0.58

Often we are interested in departures from Hardy-Weinberg equilibrium to attempt to find interesting effects such as natural selection operating. Are these populations in HWE?

Is the total in HWE? Show calculations and/or code that proves this. For the departures from HWE does this necessarily indicate selection is operating?

Part B: Perhaps it is more useful to separate synonymous and non-synonymous changes. Below is the data from the Adh gene of *Drosophila melanogaster* that we discussed previously from Kreitman's work. In addition, to separating the data into syn and nonsyn it also distinguishes if the changes are fixed (monomorphic) or polymorphic. They did this by looking at the fixed differences between *D. melanogaster* and either *D.simulans* or *D.yakuba*.

	Fixed	Polymorphic
Nonsyn	7	2
Syn	17	42

If there is no selection on these changes what would you expect the ratio of Syn/NonSyn to be? But most genes code for proteins that carry out some useful function and hence there should be selection against nonsyn change. Do you see an indication of this here (in what way)? If a few of the nonsyn changes are under strong positive selection (selection acting quickly and strongly to favour a new amino acid) what do you expect to see?

Use a 2x2 contingency chi-square test to see if the ratio of Syn/NonSyn in columns are equal (remember to use numbers, not frequencies or ratios; show you work and/or R script). What do you conclude?

> An aside: In reality this test is not very statistically accurate and instead of the chi-square test, most population geneticists use a simulation to give an idea about the significance of the test.

Question 5. Also on the `helix` website, you have a second R file, `HWE_HAPMAP.R` and three associated files (ending in `.gz`). These files contain the observed genotypic data, allele frequencies and some other data for 10000 SNPs in the human genome. The YRI is for a sample from a human population from Sub-Saharan Africa. The CEU is from a sample of humans from Europe while the file with `CEU_YRI` has both sets of population samples, but with genotypic counts and frequencies counted as if they were from a single population.

Download these files and open the R script. Make sure to set the working directory in R so that the program knows where to find the files ending in `.gz`. Run all of the lines for the function `plot.geno.vs.HW()`. Then run each line in turn to generate the three plots (African, European, and both populations together). Each point represents a single SNP in each population comparing genotypic frequency and allele frequency. The solid line (mean) represents a sort of average of those points (smoothed through all points), while the dashed line represents the expected values under HWE.

Spend some time evaluating each of these plots (you can have R save them all for you so you can look at them at the same time). Describe how well the data conforms to HWE for each of the three data sets (CEU, YRI, CRI and YRI pooled), and give your reasoning explaining the patterns you observe for the pooled (YRI and CEU together) sample. Is this

pattern true for both the African and European populations individually? If not, please provide some explanations for why they might not conform to our expectations under HWE? (for any or all of the three figures).