HST.508 PROBLEM SET 1 **Due Oct 19,2005**

1. Effect of mutations and selection on allele frequency. Consider Fisher-Wright population with selection and mutation. Consider two alleles A1 and A2 mutating at the rate or u1 for mutation A1 -> A2, and u2 for A2 -> A1. Also, model selection by assuming fitness of 1 for A1A1, 1-s/2 for A1A2, and 1-s for A2A2.

a) obtain an expression for mean allele frequency after one generation M(p) due to mutations;
b) obtain M(p) due to selection;
c) obtain expression for the variance of allele frequency V(p) after one generation. Keep only leading terms.
d) If there were no drift, what would be the steady state due to mutations alone? – and for selection only? Under what conditions can we disregard the effect of genetic drift?

2. Mean time to fixation or loss.
Obtain an expression for mean time of fixation or loss $\bar{t}(x)$ of an allele with initial frequency *x* in Fisher-Wright population of N diploids without selection or mutations. Start with a recursive relationship obtained in the class:

$$\bar{t}(x) = \sum_{\Delta x} t(x+\Delta x)\Pr(\Delta x) + 1$$

(a) by expanding and keeping two first terms you can obtain a differential equation for $\bar{t}(x)$.
(b) Write down boundary conditions for $t(x)$ and solve this equation.
(c) Interpret the different terms in your expression for $t(x)$.

3. Effective population size. Derive a formula for the expected population size, Ne, if the number of males (Nm) and number of females (Nf) are not equal. (Hint: use the same method we used to derive Ne for the case of a population that is fluctuating in size.)

To fix a concrete case, consider the following example. Imagine a zoo population of primates with 20 males and 20 females. Due to the dominance hierarchy, only one of the males actually breeds. What is the relevant population size that informs us about the strength of drift in this system? Is it 40? 21? To get the answer, compute the probability that two genes drawn at random are alike

(identical by descent) in this new situation, depending on random draws of genes from the males (actually just one male) and 20 females.

---

3. Emulating Marty Kreitman, you sequenced 1 kb from two different anonymous regions in seven different strains of *Drosophila melanogaster*, a reference strain and six more. The figures showing only segregating sites are shown below. The numbers on top show positions of the segregating sites in your sequences. Do you see any evidence of selection affecting the observed pattern in either of the two regions? If yes, describe a test that you can use to provide evidence for that? Give as many plausible interpretations of the data as you can.

Region A

	117	294	663
REF	A	G	G
SEQ1 -	T	A	
SEQ2	-	-	-
SEQ3	-	-	-
SEQ4	C	-	-
SEQ5	-	-	-
SEQ6	-	-	-
SEQ7	-	-	-

Region B:

	112	245	366	416	518	655	792	811	913
REF	C	T	T	T	A	G	A	C	C
SEQ1	-	-	-	-	C	-	-	-	-
SEQ2	-	-	-	-	-	T	-	-	-
SEQ3	-	-	-	-	-	T	-	-	G
SEQ4	G	A	-	-	C	T	T	-	G
SEQ5	-	-	A	-	-	-	-	-	G
SEQ6	G	-	-	G	-	-	-	-	-
SEQ7	-	A	-	-	-	-	-	G	-

(a) What simple evidence indicates that region (A) might be a protein coding region, as opposed to region (B)?
(b) In order to test for selection, you compute Tajima's D score for region (A) and for region (B). Why would the Tajima's D score be an appropriate metric in this context?
(c) What do you conclude from your comparison of the scores for region (A) vs. region (B)?

4. Maize and Rice. You study two anonymous regions in the maize and rice genomes. You find that in the first region (region A) 25% of the nucleotide positions are different in

the two species and in the other one (region B) only 5% are different.

a) This result is entirely consistent with the neutral theory. Explain why.
b) You proceed to study nucleotide polymorphism in these two regions. In a sample of 10 Maize alleles from locus A (2000bp in length) you find 20 segregating sites. On the assumptions of neutral theory, how many segregating sites do you expect to observe in a sample of 5 alleles of the same length (2000bp) in the region?

5. Testing neutrality of human SNPs
Download a database of SNPs and their frequencies obtained for 24 African-American subjects and 23 European subjects. The database is available from the class web-site. Write a program to calculate Tajimas'D and its P-value (described in lecture notes) for the whole database.

(a) Compare obtained scores for African and European populations. Interpret your results taking into account that Tajima's D tests neutrality in a population of constant size.
(b) Analyze in the same way SNPs with non-synonymous amino acid substitutions. Interpret obtained scores.
(c) Same for synonymous substitutions.