Short Problems:

1. We survey an intergenic region (expected to be neutral) and an enzyme coding locus in two related species of fish, finding the following data:

Region	Intergenic	Enzyme Coding
Sites polymorphic within species	200	84
Sites different between species	50	78

Note that the intergenic region and coding locus are *not* necessarily the same length.

   (a) (2 pts) Does this suggest selection on the coding locus, and if so, what kind? Explain briefly. (You don't need to perform a statistical test; just describe which direction the results are tending.) *There is not enough polymorphism, or too much divergence, in the enzyme coding locus compared to the intergenic region. This suggests that a disproportionately large number of the variants that arise in the population (and thus contribute to polymorphism) go on to be fixed (and thus contribute to divergence), which is the signature of directional selection.*

   I do not recommend trying to memorize which way this goes: it is more reliable to reason it out each time.

   (b) (3 pts) Would this test still work if our interesting gene were an RNA locus that is not translated? Why/why not? *It should work fine. Nothing in the test relies on the locus being protein-coding.*

   (c) (3 pts) Is this test likely to be disrupted by a difference in generation time between the two species? Why/why not? *Several students pointed out that a faster generation time may mean more mutations and chance the polymorphism/divergence ratio, which is true. However, it should change the ratio genome-wide and our comparison between two regions should remain valid. Thus, the test is expected to still work.*

2. The human-specific locus HLA-H (previously known as HLA-AR) is similar to functional MHC loci, but has fixed mutations that seem likely to disrupt its protein structure. However, it is known to be transcribed, leading to the possibility that it still has some function.

   (a) (2 pts) We measure dN/dS for this gene and find that it is close to 1. Is this likely to be a functional gene or a pseudogene? Explain briefly. *It looks like a pseudogene, but we should check that our dN/dS=1 is not an average of a region with* $>> 1$ *and a region with* $<< 1$.

   (b) (3 pts) We can't use between-species tests as the gene is not found in chimpanzees or other apes (it is likely a recent gene duplication of HLA-A). Can you suggest another way to determine if this locus is under selection? (Bear in mind that testing this experimentally will likely upset your IRB.) *In a non-human organism we'd knock the gene out and look for a phenotype, but you really should not do this in humans; and cell culture is unlikely to work for an immune system gene, as cell cultures don't have much of an immune system. Tajima's D would be a good statistic to try. You could compare the pattern of variation at HLA-H with HLA-A and other HLA loci, but if it is under completely different selection than the others you might have trouble interpreting your results.*

3. (3 pts) Selection which periodically reverses direction–for example selecting for larger individuals in winter and smaller in summer–can maintain polymorphism, but only if it reverses often enough. Can you suggest a rule of thumb for a timescale that is definitely *too long* for a given species, so that selection will remove one set of alleles or the other rather than maintaining both? *In* $4N_e$ *generations on average, all but one current variant will be lost by drift. Certainly if "seasons" are longer than this, the summer alleles will be lost during the winter and vice versa, and polymorphism will not be retained. If there is selection, loss will be faster than neutral, so* $4N_e$ *is a rather conservative upper bound. Some students tried PopG and found that* $2N_e$ *might be a better bound.*

4. (3 pts) Theory presented in class suggests that the chance a newly arisen allele with a heterozygote advantage s will fix is approximately $2s$, which does not depend on population size. None the less, if I have two populations of flies, one of size 10,000 and one of size 100, and I treat them with a new insecticide, the large population will become resistant more rapidly. What is missing from the theory? *Two things. First is that if we are waiting for a rare mutation, though its chance to survive will be the same in both populations, the mutation will arise much more often in the large population so it will get lucky sooner. Second is that the large population maintains higher neutral variability. If a previously neutral variant can help with the insecticide, the large population is more likely to have it already on hand and not have to wait for a new mutation.*

Long Problem:

This problem uses the PopG simulator, found at:

http://evolution.gs.washington.edu/popgen/popg.html

You may use another simulator if you prefer, but it should be a forward simulator, not a coalescent-based backwards simulator. Please do not turn in screenshots of PopG or other simulator runs; just tabulated data and explanations.

The approximation above gives a fixation chance of $2s$ for a newly arisen allele with a heterozygote advantage of s. We have no such handy rule for the fixation chance of a newly arisen allele with an advantage only in the homozygote–a new favorable recessive. All we know is that it should be greater than the $\frac{1}{2N}$ fixation chance for a newly arisen neutral allele and less, for a given s, than the chance for an allele that is not recessive.

1. (3 pts) Using the PopG simulator, check our $2s$ approximation for values of s around 0.1, using an allele which has an advantage s in both the heterozygote and the homozygote (i.e. fitnesses are 1, 1+s, 1+s). Is there a population size so low that it breaks down? (Don't forget that when you change the population size, you must change the starting frequency of the new allele so that it still exists as exactly one copy.) *Almost everyone found that the approximation is a slight overestimate, but is still fairly good except for ridiculously small populations.*

2. (4 pts) Again using PopG, estimate the chance of fixation of a newly arisen *favorable recessive* (fitnesses are 1, 1, 1+s) for at least three values of population size and four values of selection coefficient, and tabulate your results. (I recommend at least 1000 replicate populations in each run. Population size will have to be fairly small as PopG is a forward simulator.)

3. (3 pts) Draw any conclusions you can from your results. Does the chance that a favorable recessive will fix depend on population size? How big does s have to be before the chance is noticably above $\frac{1}{2N}$? *(Joint answer to 2 and 3) Students found three things: (1) It doesn't take much selection for the chance to be higher than the neutral chance. (2) The rare recessive has a better chance in a smaller population–probably it can make homozygotes sooner. (3) The chance is definitely lower than for a mutation with an effect in the heterozygote, but no simple rule of thumb works because it does depend quite strongly on population size.*

4. (2 pts) Why would a backwards-time coalescent simulator be a bad choice for these questions? *We are interested in how a very rare allele behaves, but the coalescent is an approximation for large numbers; it may not be reliable for very rare alleles. Also, it is hard to backwards-simulate selection–how do we pick the allele frequencies at the current time?*

   All existing coalescent-selection simulators work by trying to forward-simulate the allele frequency trajectory, then backward-simulate the coalescent tree depending on those allele frequencies.