Short Problems:

1. We take equal-sized samples from two plant species: a Native species which has had constant population size, and an Invasive species which has expanded rapidly. Their mutation rates and current population sizes are the same. Briefly explain whether you expect Native or Invasive to have more of the following:
   (a) Sites that vary within the sample?
   (b) Variants found in just one individual in the sample, as a fraction of total variants in the sample?
   (c) Effective population size?
2. Consider one of your two copies of a particular genomic region. For this problem, assume only genetic drift, no selection; ignore population subdivision; and assume that reproduction is at random. The current human population size is roughly 7.7 billion ($7.7x10^9$) and the human generation time is roughly 25 years.
   (a) Eventually all copies in the human species will trace back to just one copy that exists today. What is the approximate chance that your specific gene copy will be the winner?
   (b) Does future population growth or shrinkage affect this estimate?
   (c) How many generations is the lucky winner expected to take to reach fixation?
   (d) If the population instead continues to grow, will the expected time to fixation increase or decrease?
3. Does genetic drift even matter in humans anymore now that there are so many of us? Briefly defend your answer.
4. Give a formula for the expected number of variable sites in a sequence of length L for a sample of size k from a haploid population of size N and mutation rate μ per base per generation.

Long Problem:

When a region of the genome contains far more variable positions than average, researchers often propose "interesting" explanations: a mutational hotspot, selection for diversity, or introgression (interbreeding) with another species. The null hypothesis to which these must be compared is stochastic variation in coalescent depth or mutation accumulation. This problem will focus on coalescent depth.

A fictional organism has a genome containing 10 segments. Assume that there is no linkage between segments and no recombination within segments, for simplicity.

- What is the probability that one or more of the 10 genomic segments will have a common ancestor at least twice as old as expected?

This problem can be solved by calculation, computer simulation, or a mix of the two. If you use calculation, please show your reasoning. If you use computer simulation, please attach at least the key calculation routine of your code.

Helpful hints:

In general. You do not need to track which lineages coalesce for this exercise, only the time it takes them to do so. Also, under the coalescent the answer doesn't depend on N or μ, so you can make life simpler by setting $2N\mu = 1$. (The actual answer does depend on these parameters, but if N is reasonably large and μ is reasonably small, the approximation is extremely good.)

Solving by calculation. Directly calculating the chance that one or more events will happen is often painful due to combinatorics– there are a lot of different ways that, say, 3 out of 10 segments could be high. Instead, note that the chance that something

will happen is 1 minus the chance that nothing will happen. In this and many similar problems, the chance that nothing will happen is relatively straightforward to calculate.

Solving by simulation. If your programming language doesn't have a routine to draw randomly from an exponential, but only from a uniform between 0 and 1, the following trick is useful:

exponential draw = -1.0 * log(uniform random draw)

These simulations should be fast, so don't skimp on how many you do. Also, please resist the temptation to use PopG as your simulation engine. (You could use PopG to confirm your results, though: verifying that two independent algorithms agree is a powerful test of software correctness.)