Kristina Black

Midterm exam, Part B
Population Genetics, Spring 2020

This part of the midterm is due at **5:00 pm on Friday, March 13** (tomorrow). You can bring it to my office or email me a PDF. You are free to use your notes and calculator, but no internet. You cannot communicate about the exam with anyone else. **For questions 1 and 2, show your calculations, and be sure to put a box around your answer for each question.**

1. The relative fitnesses at locus *A* are:

W_{11}	W_{12}	W_{22}
1	1.1	1.21

If allele A_1 is currently at a frequency of $p_1 = 0.4$, what will be its frequency in 10 generations? (Two significant digits is fine.) Assume that selection is the only force at work. [12 points]

	A_1A_1	A_1A_2	A_2A_2
W	1	1.1	1.21
	1	1+s	$(1+s)^2$

p_1=p=0.4 therefore p_2=q=0.6. Also, s= 0.1.

$p(10)= pe^{10s}/(q+pe^{10s})$
$= 0.4e^{10*0.1}/(0.6+0.4e^{10*0.1})$
$=0.4e^{1}/(0.6+0.4e^{1})$
~ 0.64
Therefore, in 10 generations, p_1 will spread to 0.64.

2. Consider a mutation under multiplicative selection that has a fitness effect (selection coefficient) of *s*. What the ratio of fixation probabilities in two populations, one of which has size *n* and the other has size 4*n*? In which population is the mutation more likely to fix? Why does population size have that effect on the fixation probability? [13 points]

For two populations of different sizes (n and 4n) experiencing the same fitness effect (s), alleles will be more likely to evolve as if neutral in the smaller population and selection will be more effective in the larger population.

Fixation probabilities:

U (for pop w/ 4n) ~ 2s because alleles under positive selection will increase toward fixation.

U (for pop w/ n) ~ 1/2n because alleles in a smaller population are more prone to drift.

3. What affects genetic diversity (π) in a region of a genome? Name at least three factors, explain how and why they affect π. [25 points]

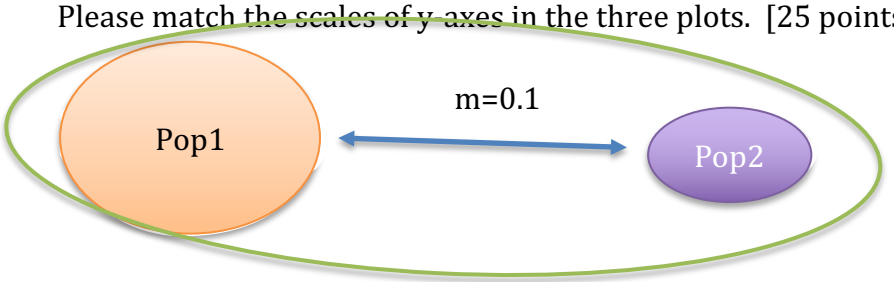
1. *High linkage disequilibrium*- When alleles are more likely to be linked together, this can cause neutral genes to hitchhike with genes under selection toward fixation. A selective sweep can therefore reduce neutral variation.

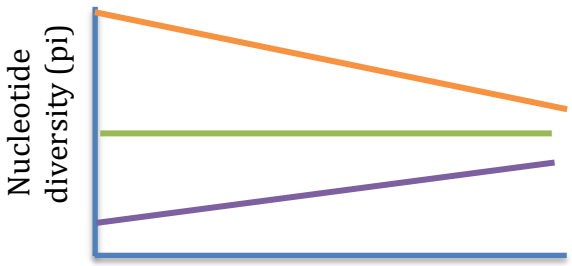
2. *Low Fst*- Fst is a measure of genetic distance between 0 and 1. When Fst- between QTLs for example- is close to 0, then those genomic regions contain similar sequences. Therefore, the genetic diversity of that region is low.

3. *Negative Tajima's D*- Tajima's D is negative when the gene tree contains long branches, and the gene tree contains long branches when there are few singletons in the site frequency spectrum. Many singletons implies high genetic diversity in growing populations, so fewer singletons would imply relatively lower genetic diversity in a stable or shrinking population.

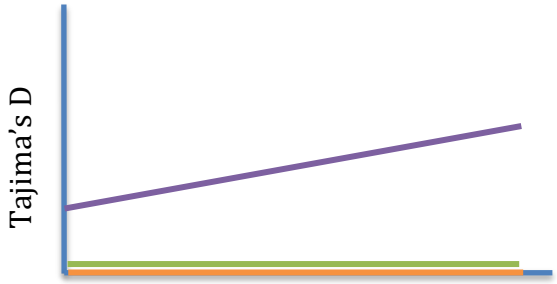
4. Assume there are two populations, pop1 and pop2, adapting to different environments. Pop2 is much smaller than pop1 and split from pop1 a long time ago. Migration rate is ~0.1 migrants/generation. Draw profiles of π, Tajima's *D*, and LD around a site that is under divergent selection between pops 1 and 2 (i.e., selected in both pops, but for different haplotypes), for three data situations:
   a) Sample of *N* individuals from pop1
   b) Sample of *N* individual from pop2
   c) Situation where we sampled *N*/2 individuals from each of the two populations (perhaps not being aware of their subdivision).

Please match the scales of y-axes in the three plots. [25 points]

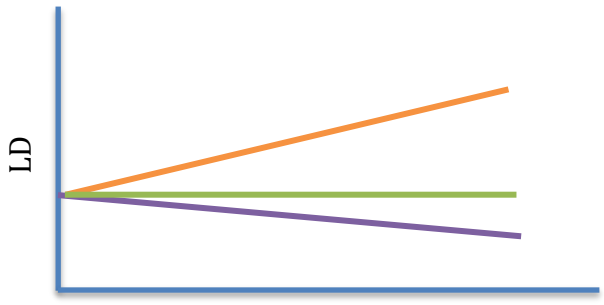




Pi is largest in Pop1 because pi is proportional to N but it is under selection so pi is decreasing over time. Pi is smaller in Pop 2 because of smaller N, but it is increasing because drift will overpower selection, and drift increases variation. In both populations, pi is in between because the amount of variation is being exchanged between the two pops and is not changing.



Tajima's D is more positive in Pop2 because it is starting at a lower N and therefore migration of singletons into the population can make it appear like it is growing. But D is about 0 for Pop 2 and both pops together, because the larger N won't appear to be growing or shrinking.



LD increases for Pop1 because neutral variation surrounding the site under selection will become fixed. LD decreases for Pop2 because gene flow from immigrants will cause large allele frequency shifts due to drift. LD for both pops, uncontrolled for population structure, will be mid range.

5. You have performed GWAS and found that nearly half of the genome appears to be significantly associated with your trait of interest. What is going on? Think of multiple possible explanations, indicate under which conditions they would be likely, and how we can test them. [25 points]

First, if I didn't sequence at least a few individuals at high coverage, then my analysis may not have detected linkage disequilibrium. In this case, some neutral variation may wrongly appear to be significantly associated with the trait of interest. It's important to use the polygenic score to conduct "LD clumping" in order to remove SNPs that are not significant.

Second, if I didn't sequence enough individuals, then I probably didn't collect enough data for the GWAS to impute missing data (like between RAD sites). This could lead to incorrect guesses or bias the results. Especially if my sample size was small (like n<1000) then I would try sequencing more individuals and adding to the GWAS data to see if it clears up uncertainties.

Third, it's possible that I didn't control for population structure in my sampling scheme. If there is cryptic population structure, then individuals can appear drastically different for reasons unrelated to the trait of interest. This also goes back to the sample design, making sure to sample equal representatives from predicted subpopulations across the study area.