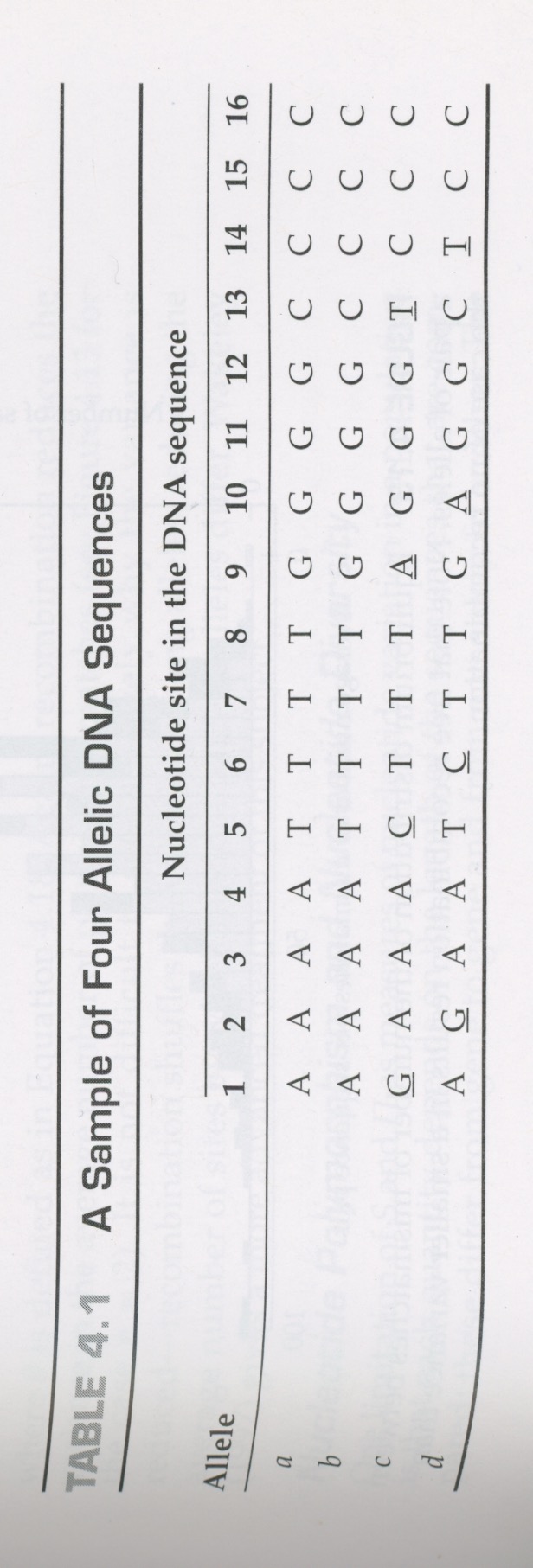
**Molecular Population Genetics lab**

According to the neutral theory of molecular evolution, most polymorphisms are neutral, therefore, levels of population polymorphism () are a function of the strength of genetic drift () and the rate of mutation ().

There are two widely-used ways to estimate from DNA sequences. One is a function of the number of segregating sites (S) in the dataset and the number of sequences (n):

where

The other way to estimate is to calculate , which is the average number of pairwise differences. In the example below (see Table), you compare every allele to every other (6 comparisons are possible: a/b, a/c, a/d, b/c, b/d, c/d) counting the number of differences each time, then calculate the average.



1. Estimate from the sequences in the Table in the 2 different ways (2 points):

and

Show your working for both. Theoretically, they should be the same and should be 0.

1. In a past study, the estimate of was 3.5. How different are your two estimates of based on and from the past estimate? Could the difference be statistically significant? You will use coalescent simulations to decide.

Hudson (2002) developed a program (ms) for generating random samples under a neutral coalescent model (<http://home.uchicago.edu/rhudson1/source/mksamples.html>). The full documentation is available on that website (with Hudson’s code) and on eLC in msdoc.pdf. You can use ms to decide whether the difference between your and is likely to be statistically significant:

* 1. Log in to teach (teach.gacrc.uga.edu). To get the basic usage instructions for ms type (it will look like an error message):

ml ms

ms

* 1. You will run a coalescent simulation to randomly simulate 4 sequences (n=4 with 100 replicates) that will show you how many segregating sites (S) to expect, if the true value of is 3.5. Therefore type:

ms 4 100 -t 3.5

To save the output to look at later, then use:

ms 4 100 -t 3.5 > saveforlater.txt

Note that every run of these ms coalescence simulations is different.

The results give you an indication of how much the number of segregating sites (segsites) and other parameters can vary under a random coalescent model, if the true value of is 3.5, with no natural selection and no change in population size. Read the manual (msdoc.pdf) to understand the output.

//

segsites: 7

positions: 0.0952 0.1029 0.2012 0.2516 0.4440 0.8214 0.9282

0100000

0011110

1000000

0100001

//

segsites: 2

positions: 0.1589 0.2023

10

10

10

01

This output is straightforward and easy for bioinformaticians. For each simulated replicate (out of the 100), ms reports the number of segregating sites, their positions along the length of the DNA sequence and a code for each of the 4 simulated sequences: 0 means the sequence remained in the ancestral state at that position and 1 is the derived state. And there are many simple options that are beyond the scope of this lab exercise to explore: Example 1, with the -L option you can get the time to the most recent common ancestor in units of or for haploids ( is generation time). Example 2, with the -T option you can also get the associated simulated genealogy. As a result, ms is still widely used even though it was one of the first coalescent simulators.

* 1. Next, you will use R to summarize the results from 100 coalescent simulations, and thus decide how much you can trust your estimates of . Use the unix command “grep” to extract only the counts of segregating sites from each replicate:

ms 4 100 -t 3.5 | grep segsites

Then print these counts as a simple list that you can copy and paste into R:

ms 4 100 -t 3.5 | grep segsites | awk '{ print $2; }'

In R, create a new variable with

sim4<-scan()

then paste in your list of segsites counts and press enter. Summarize your results and check whether the number of segregating sites you actually observe is outside the range you observe 95% of the time in the simulations:

summary(sim4)

hist(sim4,30)

quantile(sim4,c(0.025,0.975))

Show the results of your analysis of simulated numbers of segregating sites with =3.5; include a histogram, mean, median, 95% quantile range (1 point). How often are the number of segregating sites in the simulated data (segsites) equal to or higher than the number of segregating sites (S) you observe in the Table? Do you think that your estimates of from the above four sequences are likely to be significantly greater than =3.5, and do you think your two estimates of (for question 1) are likely to be significantly different? Explain your answer (1 point). Remember that for a significant difference (e.g. p<0.05), the observed S would need to be greater than simulated segsites at least 95% of the time.

1. Coalescent theory and simulations are important tools for epidemiologists. Consider the current novel coronavirus pandemic (nCoV-2019). Estimates of effective population size () from DNA sequence data are useful in this situation because not all cases will have symptoms. In addition, the effective population size is probably more useful than census population size because the viral population size we seek is not a count of the number of viral particles (including the many within each host).

**Disclaimer:** I work on a species with relatively constant Ne, do not work on viruses and am not a phylodynamics expert. The purpose of this lab is to learn more about coalescent theory and to think about the assumptions in its most basic form. The experts are using much more sophisticated analyses (with fewer known violations of assumptions). We are not aiming to make predictions about HCoV-19.

On Friday February 28th 2020 (today), Trevor Bedford (@trvrb) tweeted his count of “129 SNPs across the 124 genomes” of nCoV-2019 available. Estimate from this number of segregating sites (SNPs) and sequences (genomes) using the equation. Hint: to get Waterson’s constant quickly in R type:

sum(1/(1:123))

.. and check that this shortcut gives the same result when applied to question 1 above.

Show your estimate of , use R to summarize the results of 100 simulated replicates as in question 2 above and show your results (1 point). Compare this histogram of a large-dataset-simulation to the one from question 2: how different are they and are you surprised?

Ewen’s formula () assumes a Wright-Fisher model. What do you think is the most important departure from that model that makes this simple ms analysis of nCoV-2019 overly simplistic? (1 point).

For those of you that are interested, Bedford used the time to the most recent common ancestor and an extremely useful tool for phylodynamics (BEAST, Suchard et al 2018) to estimate (https://twitter.com/trvrb/status/1226739962160115712). You can run ms (as above) with the -L option to simulate this parameter.

1. The McDonald-Kreitman test is a classic test for detecting selection in DNA sequences. Egea et al (2008) developed a useful website that implements the McDonald-Kreitman test and many of its extensions here: <http://mkt.uab.es/mkt/>. They also provide example test data that you can use to familiarize yourself with the test options and possible results. Click on the “Standard MKT” tab and select “Example” from the left-hand MKT Menu. You will see data for 5 *tre1* gene sequences from *Drosophila melanogaster* and the outgroup *D. yakuba*.

* 1. Test for selection at this gene using a classic McDonald Kreitman test with synonymous changes set as neutral. Do not use the Jukes Cantor correction because that was not in the original McDonald-Kreitman test. Show your results (1 point). It is standard practice to show your statistic (e.g. chi-sq estimate) and your degrees of freedom alongside your p-value when discussing the result of any statistical test. How do you interpret your result? (1 point).
  2. There is also an option on this website to compare non-coding regions to nearby coding regions, and to use such comparisons to test for selection. Try this out using the test data provided by clicking on “Example” in the “MKT Menu”, then “Advanced MKT”, then see the Results. The test data allows you to compare levels of polymorphism and divergence for the intron and 4-fold degenerate sites of alcohol dehydrogenase (*Adh*) in two different *Drosophila* species. 4-fold degenerate sites are nucleotide sites where none of the possible nucleotide substitutions could result in a non-synonymous change. Are the number of polymorphisms in *Adh* introns higher or lower than expected under neutrality? (1 point). How do you interpret this result? (1 point)

References

Egea, R., Casillas, S., Barbadilla, A., 2008. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res 36, W157–W162. <https://doi.org/10.1093/nar/gkn337>

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18, 337–338. <https://doi.org/10.1093/bioinformatics/18.2.337>

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4. <https://doi.org/10.1093/ve/vey016>