**Molecular Population Genetics 2 lab**

*This lab follows up on previous lectures and a lab that introduced coalescence simulation and the estimation of Pi and ThetaW for a toy dataset from Hartl and Clark 2003 and for HCov-19 on Feb 28th. A couple of days before this March 6th lab, students read Trevor Bedford’s blogpost on the phylodynamic estimation of incidence and prevalence of ncov infections through time at a mirror site (with follow-up discussion):*

[*http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391/4*](http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391/4)

*And Andrew Rambaut’s blog estimating the MRCA:*

[*http://virological.org/t/phylodynamic-analysis-129-genomes-24-feb-2020/356*](http://virological.org/t/phylodynamic-analysis-129-genomes-24-feb-2020/356)

*The lab allowed 50 minutes for 16 graduate students to work through the exercises (working together in teams if they like), followed up by a “hive mind” discussion of the answers.*

The HCoV-19 pandemic is spreading across the globe, and today (March 6th 2020) the number of cases of infection passed 100,000 (<https://www.nytimes.com/2020/03/06/world/coronavirus-news.html#link-5f7cc0ce>). One of the ways in which researchers estimate the number of current cases is by genome sequence analysis. Important tools for these DNA sequence analysis are phylogenetics and the application of coalescent theory. This type of analysis focuses on estimating the current and recent effective population size (Ne) of the HCoV and is different than the extremely useful ecological models that can predict the effects of changing behavior (check out the UGA-based site: http://2019-coronavirus-tracker.com).

**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.

The suggested reading for this week was a blog on Trevor Bedford’s use of a phylodynamics analysis to estimate the number of active cases of HCoV-19 on February 8th and the total number of infections since the start of the epidemic. The link that includes that post with all the data he was permitted to share is on github: <https://github.com/blab/ncov-phylodynamics>. In this lab we estimate some of the same parameters from this analysis **in a much cruder way** that includes violation of assumptions. Unfortunately, there is no time to develop a more sophisticated lab that would require analysis of the raw DNA sequences.

In last week’s lab you estimated Watterson’s Theta (ThetaW) using using the number of segregating sites (S) in the dataset and the number of sequences (n). In addition you estimated Theta as , the average number of pairwise differences. These measures assume constant effective population size and in a haploid Theta = 2 Ne. We know that Ne for HCoV is not constant, but that the mutation rate () probably is. Which of the two estimators (ThetaW or) is likely closer to the true current enlarged population size?

We will use ThetaW () for the estimates in this lab because we do not have access to the sequences for estimating . For our data we will use a freeze of the summarized nCoV genome data from Feb 7th with 81 genomes (<https://nextstrain.org/ncov/2020-02-10?gmax=5512>) because this is close to the set of sequences used for Bedford’s analysis from 24 Dec 2019 to 4 Feb 2019.

1. **Count the number of SNPs in these 81 genomes.** Scroll down to the plot labelled “Diversity” that shows the segregating sites for these sequences positioned along the map of the HCoV-19 genome. Zoom into the Diversity plot using the small arrows at the foot of the plot, and select “NT” for nucleotide to count segregating sites, then select “Events” to count the number of SNPs: at positions with multiple mutations (e.g. 3) count all 3 mutations

0-5kb = ?

5-10kb = ?

10-15kb = ?

15-20kb = ?

20-25kb = ?

25-30kb = ?

Total NT = ? segregating sites

Note: a quick and dirty way to compare synonymous to non-synonymous changes in the ORFs is to compare the AA track to the NT track.

Bedford does not mention the number of SNPs he used for his maximum likelihood and BEAST phylogenetic analyses to estimate Ne, but he does mention the much lower number of genomes: 53 genomes not 81. Bedford uses fewer genome because “coalescent models assume that infections are sampled randomly from the infected population”. This is clearly not true for the actual sampled data: people are tested in clusters. Our sample size (n) is therefore an overestimate. Will our overestimated sample size lead to overestimation or underestimation of Ne and the number of cases? Remember that

where

1. **Estimate parameters for comparison to Bedford’s analysis.**

Treat your SNP count as S and use n=81 to estimate ThetaW for Feb 7th using the above equation. Remember that for 81 sequences you can estimate a in R with “sum(1/(1:80))”.

Next, use your estimate of ThetaW to estimate Ne the timescale of coalescence (Ne) and the point prevalence (I) for comparison to Beford’s estimates from Feb 4th. To estimate Ne use which rearranges to . You can use the same estimate of used by Bedford: 0.9x10-3 mutations per site per year. But note that our use of assumes that the total tree length is longer than it would be for a growing population. HCov definitely has a rapidly growing population that violates that assumption.

For , use the same estimate of generation time as Bedford (7.5 days) and remember that you need to convert to generation time measured in years.

For use E[R0] = 2.3 and k=0.3 to get

Warning: this conversion for I might not be a correct! (I have not had time to check).

Ne =

Ne =

I =

Compare your estimates to those of Bedford for Feb 4th. From the blog

= 100 - from the Ne(t) figure

I = 28,500 people currently infected - from the point prevalence figure and the text

Note whether your estimates are inside his confidence intervals. Which broken assumptions could explain any difference.

1. **Repeat all of the above to get parameter estimates for today (March 6th 2020)**

Number of genome (n) =

SNPs =

ThetaW =

Ne =

Ne =

I =

How does your estimate of current cases compare to today’s estimate of 100,000 infections since the epidemic began? Given what you know about the broken assumptions, do you expect your estimate to be an overestimate or an underestimate? Explain your answer.