

Infection Timing Project Thoughts and Questions

This is what we believe to be true:

About the biology:

1. We define time to be the age of the individual.
2. Time zero is birth.
3. Individuals are infected in-utero, at birth, or through breastfeeding.
4. Individuals could have been infected at different times.
5. "Training data" individuals were infected in-utero
6. Infection time is defined as the time (age) at which average pairwise sequence diversity is zero.
7. Estimated infection time (age) for each individual can not be greater than the first time point at which we have an average pairwise sequence diversity measure greater than zero.
8. Average pairwise diversity of the sequence increases linearly with time.
9. The rate of average pairwise diversity change over time is different for each individual.
10. The rate of average pairwise diversity change over time is different for each sequence fragment.
11. Each "sequence run" (for each individual) was sequenced independently.
12. APD1 is defined to be the measure of average pairwise diversity at the third codon position using only sites at which the sum of all minor variants is greater than 0.01.

About the model:

1. We are using APD1 measurements (cutoff 0.01) from all runs for each individual (not an average).
2. We model the function (average pairwise diversity) \rightarrow (time since infection) as a linear map.
3. We model the baseline slope of this function as a uniform distribution.
4. We model the variation in slope between individuals of this function as a normal distribution.
5. We model the variation in slope between sequence fragments of this function as a normal distribution.
6. We model the total slope of this function to be baseline slope (#3) + individual variable slope (#4) + fragment variable slope (#5).
7. We model the baseline intercept of this function (i.e. time when APD is zero OR "estimated infection time") as a uniform distribution from 0 to -0.75 years (9 months before birth).
 - see question on this below
8. We model the variation in intercept of this function between individuals as a normal distribution.
9. We model the total intercept of this function to be the baseline intercept (#7) + individual variable intercept (#8).

Some questions to think about:

1. How do we want to limit the y-intercept (time (age) for APD = 0)?
 - Should we restrict the y-intercept to be between 0 to -0.75 years (9 months before birth)?
 - OR should we be more restrictive (i.e. y-intercept in the third trimester)?
 - OR should we be less restrictive (i.e. y-intercept between -0.75 to 0.75)?
 - OR something else entirely...