A few recombination / hitchhiking project ideas (7/11/21)

(1) Stronger signatures of recombination in the recent past?

Background: our LD manuscript introduced tools for predicting LD patterns (e.g. r^2) across a range of frequency- (and therefore time-) scales. When piloting these predictions on a between-host population of human gut bacteria (E. rectale), we noticed an interesting trend: there seemed to be relatively *less* LD among progressively rarer mutations, even when they were separated by larger genetic distances. This is the opposite of the trend predicted by our simple model, in which rarer mutations tended to have *higher* relative values of LD, because they have had less time to recombine away from each other. This qualitative difference is curious, and suggests a few potential next steps:

(1A) One issue with these recombination measurements is that they're a bit backwards: i.e., they're inferring the presence of recombination from the relative depletion of linkage *dis*-equilibrium (which is itself a measure of deviation from the free recombination case). This approach is always a bit tricky because it requires us to know what the background values of linkage disequilibrium *should have been* (so that we can know when there is a depletion), and these background can vary quite a bit depending on the frequencies of mutations, demography, etc. It would be great to have some statistics that measure what we actually care about – recombination.

One approach for dealing with this is to define a so-called "linkage equilibrium" (LE) statistic rather than the LD statistics that we usually use. The original motivation for this comes from the classical 4-gamete test: for a system of 2 loci, the presence of all 4 haplotypes indicates that there must have been a recombination event, while 3 or fewer haplotypes are inconclusive. One such LE statistic could obviously be the fraction of sites "passing" the 4 gamete test, but this has some drawbacks. You might have a pair of loci where the 4th haplotype is present at a very tiny frequency (e.g. because it is a sequencing error), and this could count just as much as a site where all 4 haplotypes are very close to linkage equilibrium. It would be great to have a more quantitative version of the 4 gamete test that weighted these examples accordingly.

One idea is to define an LE statistic that looks like LE ~ fAB * fAb * faB * faa. This statistic is nonzero only if all 4 gametes are present (i.e. a recombination event occurred). In the limit of full linkage equilibrium, it approaches LE ~ fA^2 (1-fA)^2 fB^2 (1- fB)^2. Thus, the quantity LE ~ fAB * fAb * faB * faa / fA^2 (1-fA)^2 fB^2 (1- fB)^2 should approach zero for non-recombining loci and 1 for the limit of infinite recombination.

One project idea would be to flesh out this linkage equilibrium statistic along the lines of the original LD manuscript. This would involve (i) repeating as many of the analytical and heuristic calculations as possible and (ii) applying this new statistic to the gut microbiome data, to see if the signal of "more recombination in the recent past" persists under this more direct metric. We now have access to much larger sample sizes for many of these populations in the UHGG database, which should create more opportunities to probe these patterns across a broader range of frequency and time scales.

(1B) Another issue with our original finding is that the theory in our LD manuscript was derived only for the simple case where there is a constant effective population size. Many would argue that there could actually be some time-varying Ne(t), and that this is the true null model that one would need to compare against. I have a hunch that the qualitative signal above is not something that could be reproduced by an Ne(t), but we would have to somehow show this to know for sure. This is one of the central challenges in modern population genetics: how do we show that there is no choice of Ne(t) that could reproduce a given trend? So far, folks have mostly approached this task using inference + simulations: infer your best guess of Ne(t) using other features of the data, and then run some simulations to show that it can't reproduce your signal. While this is usually a good place to start, it is still hard to know whether there were not other choices of Ne(t) that could reproduce your signal with only a moderate decrease in likelihood.

Fortunately, I think we now have some new tools that will allow us to attack this problem theoretically. The notes I posted on Dec 16 show how one could potentially extend the "rare mutations" approach of the LD manuscript to situations where there is an arbitrary Ne(t). By pushing forward this calculation in the special case of the QLE regime, I obtained some preliminary evidence our original signal can't be produced by a time varying Ne(t), suggesting that it might be a "topological" deviation from neutrality (similar to the "uptick" of high frequency mutations in the site frequency spectrum).

Another project would be to push this idea toward completion. This would involve (i) verifying that the preliminary calculation is correct, (ii) extending it beyond the QLE regime if possible, using the same approach in the LD paper, and (iii) potentially extending it to the new LE statistic suggested above.

(1C) If demography can't cause the signal above, what can? One potential idea is hitchhiking via recurrent selective sweeps. The idea behind this is that recurrent hitchhiking could invert the standard relationship between frequency in time. In a neutral scenario, the most likely way to reach frequency f is through an upward triangle trajectory (Fig. 1 in LD paper). However, in the hitchhiking notes (posted Nov 20), I show that recurrent sweeps generate a fictitious selection force (similar to the one in Oskar Hallatschek's jackpot paper) that causes neutral mutations to decline in frequency over time. I *think* this means that the most likely way to reach a given frequency f is to hitchhike to a higher frequency, and then get sampled on the way down (i.e. a downward triangle trajectory). This would invert the frequency-age relationship, and provide a potential explanation for the LD patterns above.

I think this would be an interesting avenue for a project. It would involve (i) using the heuristic approaches in the hitchhiking notes to demonstrate the frequency-age idea outlined above and (ii) extending the calculations in the LD manuscript to this recurrent sweeps world. There are very few theoretical predictions in this regime to begin with, so I think this would be a useful thing to do regardless.

(2) How do we get around the drift barrier?

Background: as we saw in class, one of the most celebrated results in population genetics is the so-called "drift barrier": Pfix(mutation) / Pfix(back mutation) = exp(2 Ne s), which shows that natural selection is only effective for fitness effects that exceed the drift barrier, 1/Ne. But what is "Ne"? The classic version of this result was derived for the single locus case with a constant population size, so that Ne=N. In subsequent work (Good and Desai 2014; Rice et al 2015), we have also shown that similar result holds in rapidly adapting asexual populations, except that Ne must be replaced by the coalescent timescale Tc (which is typically << N).

This has always been a bit puzzling to me: if there are so many local sweeps, etc occurring in the background in bacterial populations, how do dN/dS and pN/pS manage to stay so low? E.g. if we take typical dN/dS values of 0.1 at face value, they suggest that something like 90% of all missense mutations are efficiently selected against (i.e. >> 1/Tc). For this reason, part of me has always curious about potential mechanisms that would explain the low values of dN/dS that seem to be universally observed among bacteria (including the gut bacterial populations that we tend to look at).

(2A) One interesting mechanism might be recurrent selective sweeps (in recombining populations). The reason for this is that, unlike in the asexual case above, the fixation time, Tc, and sweep time are asymptotically separated from each other. This suggests that deleterious mutations may need to get much luckier than simply hitchhiking on a single neutral sweep. This is borne out a bit in the hitchhiking notes above: although Tc is still a threshold for efficient selection, deleterious mutations that are *slightly* above Tc are much less likely to fix. Thus, while this doesn't get around the drift barrier entirely, it does suggest that recurrent sweeps could significantly enhance the drift barrier compared to our classical expectation. (This could potentially become important for more complex traits, where the tradeoff between target size and selective effect becomes much more important.)

Another project would be to push this idea to completion. This would involve (i) fleshing out the steps in the heuristic argument more carefully, both for fixation of deleterious mutations as well as the deleterious site frequency spectrum and (ii) verifying these heuristics with forward time simulations. (iii) it would also be interesting to see how these results extend to scenarios of recurrent partial sweeps (do we need hard sweeps to get the result? Or do partial sweeps work even better?)

(2B) Similar to the LD patterns above, one issue with our classical drift barrier prediction is that it is derived for a constant population size. Surprisingly, we don't really know the corresponding drift barrier that is produced by an Ne(t). Is it the pairwise coalescent time, Tc? Something else? The standard approach is to use inference + simulations as in the LD case above (see, e.g. the classic Boyko / Bustamante / Lohmueller / Gutenkunst et al stuff). This makes it difficult to know whether an Ne(t) produce the enhanced drift barrier expected in the recurrent sweeps scenario above.

I have a hunch that it can’t, based on the idea that (unlike a sweep), a demographic bottleneck that causes a mutation to reach frequency >> 50% is very likely to have that variant fix during the same bottleneck event. But we could need to show this to be sure.

This would be another interesting avenue for a project. This would involve (i) doing a little calculation to prove the bottleneck idea above and (ii) using the Ne(t) results in the Dec 16 notes to try to understand how the drift barrier emerges from an Ne(t). Can we prove that it is or isn’t given by the pairwise coalescent time? And if not, what controls it?