Notes on hitchhiking from a forward-time perspective

Benjamin H Good

In this document, we explore how hitchhiking influences genetic diversity at a linked, neutral locus from a forward-time perspective. (I'm sure that much of this has already been done previously, but I've been too lazy to look up the references.)

I. SIMPLEST CASE: A SINGLE "CLASSIC" SWEEP

We'll start by considering a two-locus model with one neutral locus and one selected locus, separated by a map length $r\ell$. There is a segregating mutation at the neutral locus with allele frequency f, which drifts on a timescale $T_{\mathrm{drift}} \sim Nf(1-f)$. We then imagine that a new mutation occurs at the selected locus in a random individual, and provides fitness benefit $s \gg 1/N$. With probability $\sim s$, this beneficial mutation will survive drift, and for times larger than $T_{\mathrm{sel}} \sim 1/s$, will start to grow deterministically as

$$g(t) = \frac{e^{st}}{(Ns-1)+e^{st}} \, , \tag{1}$$

and will fix on a timescale $T_{\mathrm{fix}} \sim \frac{1}{s}\log(Ns)$. How does this sweep influence the allele frequency at the neutral locus at some later time T, once the selected allele has fixed?

We start by considering the basic timescales of the problem. There are four single-locus quantities: the observation time T_{obs} and the drift timescale $T_{\mathrm{drift}} \sim N$ at the neutral locus, as well as the selection timescale $T_{\mathrm{sel}} \sim 1/s$ and the fixation time $T_{\mathrm{fix}} \sim \frac{1}{s}\log(Ns)$ at the selected locus. Given our assumptions, these timescales satisfy the hierarchy

$$T_{\mathrm{sel}} \ll T_{\mathrm{fix}} \lesssim T_{\mathrm{obs}} \ll T_{\mathrm{drift}} \tag{2}$$

Recombination introduces two additional timescales. The first, $T_{\mathrm{QLE}} \sim 1/r\ell$, is the time recquired for a *typical* individual in the population to experience a recombination event between the neutral and selected locus. We refer to this as the **QLE timescale**, since it is the time required for the neutral and selected loci to relax to (quasi) linkage equilibrium. If $T_{\mathrm{QLE}} \ll T_{\mathrm{sel}}$, i.e., if $\ell \gg s/r$, then linkage equilibrium will typically be attained long before selection has a time to act, and there will be a negligible impact on the neutral locus. We refer to selected loci with $\ell \gg s/r$ as **unlinked**, and we will generally ignore them here. In particular, for $\ell \ll s/r$, the important dynamics occur long after the selected locus reaches the deterministic dynamics in Eq. 1, which simplifies the analysis significantly.

In addition to T_{QLE}, there is also a second timescale, which we refer to as the **recombinant timescale**, T_{rec}, which is the time required for the initial adaptive clone to produce its first successful recombinant lineage. As we will show below, T_{rec} can be much smaller than $T_{\mathrm{QLE}} \sim 1/r\ell$, which can have important implications for the hitchhiking process.

If $T_{\mathrm{rec}} \gg T_{\mathrm{fix}}$, then the loci will typically remain completely linked throughout the sweep. We refer to these as **completely linked** loci, and the dynamics are rather simple. The frequency at the neutral locus will go to $f = 1$ if the selected mutation falls on the mutant background, which appens with probability f; alternatively, the neutral mutation will be driven to extinction ($f = 0$) if the selected mutation falls on the other background, which occurs with probability $1-f$. Thus, the average post-sweep frequency is still equal to f (as required by symmetry), but the sweep completely wipes out diversity at the neutral locus, which must be regenerated in the time window T_{obs} after the sweep completes.

If $T_{\mathrm{rec}} \ll T_{\mathrm{fix}}$, then the first successful recombinant will be generated while the adaptive lineage is still at low frequency ($g(t) \ll 1$), and can be self-consistently calculated from the condition

$$\int_0^{T_{\mathrm{rec}}} Nr\ell \cdot \frac{1}{Ns} e^{st} \cdot s \sim 1 \tag{3}$$

which yields

$$T_{\mathrm{rec}} \sim \frac{1}{s}\log\left(\frac{s}{r\ell}\right) \tag{4}$$

Thus, we see that T_{rec} is generally much larger than T_{sel} and much smaller than T_{QLE} in the region of interest ($\ell \ll s/r$), and that the completely linked dynamics require that $\ell \lesssim 1/Nr$.

Important: Since $Ns \gg 1$, there is also a large middle region ($1/Nr \ll \ell \ll s/r$) where neither completely asexual or completely unlinked dynamics apply. This is where all the interesting behavior occurs. Note that in this regime, we have $T_{\rm sel} \ll T_{\rm rec} \ll T_{\rm fix}$ and $T_{\rm sel} \ll T_{\rm rec} \ll T_{\rm QLE}$, so we actually have two different sub-regimes depending on whether $T_{\rm QLE}$ is large or small compared to $T_{\rm fix}$. If $T_{\rm QLE} \ll T_{\rm fix}$, then the neutral and selected loci will reach linkage equilibrium while the beneficial lineage is still rare, so that the neutral locus can only hitchhike by a small amount. We refer to these as **loosely linked**, and they fall in the range $\frac{s}{r}\frac{1}{\log(Ns)} \ll \ell \ll \frac{s}{r}$. On the other hand, when $T_{\rm QLE} \gg T_{\rm fix}$, linkage equibrium will not arise until much later, and the neutral mutation can hitchhike by very large (i.e., $\mathcal{O}(1)$) amounts. We refer to these as **tightly linked** loci, and they fall in the range $\frac{1}{Nr} \ll \ell \ll \frac{s}{r}\frac{1}{\log(Ns)}$. We now analyze the dynamics that arise in these two regimes.

We begin by considering the structure of the different recombinant clones at the end of the sweep with a simple heuristic argument. The kth established recombinant occurs when

$$\int_0^{t_k} Nr \cdot \frac{e^{st}}{Ns} \cdot dt \sim k \tag{5}$$

or

$$t_k = \frac{1}{s}\log\left(\frac{s}{r\ell}\right) + \frac{1}{s}\log(k) \tag{6}$$

and the sizes of these clone are given by

$$g_k(t) = \frac{1}{Ns}e^{(s-r)(t-t_k)} = g_0(t)\left(\frac{r\ell}{sk}\right)^{1-\frac{r\ell}{s}} \tag{7}$$

where $g_0(t) = \frac{1}{Ns}e^{(s-r\ell)t}$ is the size of the founding clone. Since these lineages have identical time-dependence, their relative sizes are essentially frozen at birth, and the frequencies at the end of the sweep are given by

$$g_0 = \frac{1}{1+\sum_{k=1}^{k_{\max}}\left(\frac{r\ell}{sk}\right)^{1-\frac{r\ell}{s}}}\,, \quad g_k = \frac{\left(\frac{r\ell}{sk}\right)^{1-\frac{r\ell}{s}}}{1+\sum_{k=1}^{k_{\max}}\left(\frac{r\ell}{sk}\right)^{1-\frac{r\ell}{s}}} \tag{8}$$

There is a slight pathology here, in that the recombinant lineages dominate the post-sweep population as $k_{\max} \to \infty$. This arises because we have approximated the recombination probability $rg(t)[1-g(t)]$ as $rg(t)$, which breaks down for $t \sim T_{\rm fix}$ when $g(t) \sim \mathcal{O}(1)$. Additional recombinants after this point are very unlikely, so we simpy truncate the sum when $t_{k_{\max}} = T_{\rm fix}$, or $k_{\max} = Nr\ell$. Since this number is large, we can approximate the sum by an integral to obtain

$$g_0 = e^{-\frac{r\ell}{s}\log(Ns)}\,, \quad g_k = \left(\frac{r\ell}{sk}\right)^{1-\frac{r\ell}{s}} e^{-\frac{r\ell}{s}\log(Ns)} \tag{9}$$

We are now in a position to analyze how these dynamics influence the allele frequencies at the neutral site. The inital clone occurs on the mutant or wildtype background with probability f and $1-f$ respectively, and successive recombinants are drawn with the same probabilities. This means that the post-sweep frequency f' is still equal to f on average, but there will be stochastic variation around this value. In particular, since the largest recombinant clone is smaller than the initial clone by a factor $r\ell/s \ll 1$, the bulk of this variation will be dominated by the background of the initial mutation. The post-sweep frequency can be approximated by

$$f' \approx \begin{cases} \Delta + (1-\Delta)f & \text{with probability } f \\ (1-\Delta)f & \text{with probability } 1-f \end{cases} \tag{10}$$

where we have defined

$$\Delta(\ell) = e^{-\frac{r\ell}{s}\log(Ns)} \tag{11}$$

The change in frequency, $\delta f \equiv f' - f$, is therefore given by

$$\delta f = \begin{cases} \Delta(1-f) & \text{with probability } f \\ -\Delta f & \text{with probability } 1-f \end{cases} \tag{12}$$

with mean and variance

$$\langle \delta f | f \rangle = 0 \tag{13}$$

$$\mathrm{Var}(\delta f | f) = \Delta^2 f(1-f) \tag{14}$$

II. RECURRENT CLASSIC SWEEPS

To model a scenario of recurrent sweeps, we'll suppose that sweeps occur at a range of loci throughout the genome at an average rate λ per site. In the simplest mutation-limited case, we might have

$$\lambda \sim N\epsilon_b \mu s \tag{15}$$

where μ is the mutation rate per locus and ϵ_b is the fraction of sites that are beneficial. When selected sites are scattered uniformly across the genome, we must now account for the fact that the genomic distance ℓ can fall in any and all of the various length ranges analyzed above. Changing variables from ℓ to $\Delta(\ell)$, these assumptions imply that sweeps with effect $\Delta \pm d\Delta$ establish at a per generate rate

$$R(\Delta)d\Delta = \frac{1}{T_c}\frac{d\Delta}{\Delta}\,, \quad 0 \leq \Delta \leq 1 \tag{16}$$

where we have defined a chacteristic timescale

$$T_c \equiv \frac{r\log(Ns)}{\lambda s} \sim \frac{r\log(Ns)}{N\epsilon_b\mu s^2}\,. \tag{17}$$

that encapsulates most of the dependence on the underlying parameters. Roughly speaking, T_c is the typical waiting time between sweeps effect $\Delta \gtrsim 1/2$; we will show below that this coincides with the coalescent timescale in the limit that $T_c \ll N$. For a given value of Δ, the instantaneous change in frequency at the neutral locus follows the same formula derived in Eq. (12) above. Formally speaking, the change in frequency in an infinitesimal time dt has a mean and variance,

$$\langle \delta f \rangle = 0 \tag{18}$$

$$\langle \delta f^2 \rangle = \int \Delta^2 f(1-f)\cdot R(\Delta)dt \cdot d\Delta = \frac{\delta t}{T_c} f(1-f) \tag{19}$$

similar to a neutral Wright-Fisher model with an effective population size $N_e = T_c$. However, this formal similarity to genetic drift is misleading. As we will see below, the changes in frequency due to genetic drift become highly skewed when $f \ll 1$ or $1-f \ll 1$, and this will lead to important differences in the typical mutation trajectories compared to the genetic drift case.

A. Typical trajectories

To understand how genetic draft influences mutation frequency trajectories, we'll start by considering a mutation at frequency $f \ll 1$ and consider the time required for f to change by $\mathcal{O}(f)$. In the absence of genetic draft, we know that genetic drift requires a time of order $T_{\text{drift}} \sim Nf$ to change the frequency of a mutation by $\mathcal{O}(f)$, after which point there is a significant probability that the mutation has gone extinct. We'll now consider the corresponding dynamics under genetic draft alone.

Since $f \ll 1$, most sweeps will occur on the wildtype background and will cause a reduction in the mutation frequency. As long as $f(t)$ remains within an order of magnitude of f, these sweeps will contribute additively, so that the total change in frequency is given by

$$|\delta f|_- \sim f\left[\sum_{\log\Delta=-\infty}^{0} e^{\log\Delta}\cdot n_-(\log\Delta, \delta t)\right] \tag{20}$$

where $n(\log\Delta, \delta t)$ is the total number of sweeps that occured in the wildtype background over that time period with effect size $\log\Delta \pm \mathcal{O}(1)$. The number of sweeps in each bin will be Poisson distributed with mean $\delta t/T_c$, so $\delta t{\sim}T_c$ generations are required for $|\delta f|_-{\sim}f$.

The probability that a sweep occurs in the mutant background over this same time period is smaller by a factor of f. The maximum typical effect size Δ^* for these sweeps is given by the condition

$$\frac{f\delta t}{T_c}\int_{\Delta^*}^{1}\frac{d\Delta}{\Delta} \sim 1 \quad \rightarrow \quad \Delta^* \sim e^{-1/f} \tag{21}$$

so the typical change in frequency in the positive direction is

$$|\delta f|_+ \sim \int_0^\Delta \Delta \cdot \frac{f \delta t}{T_c} \frac{d\Delta}{\Delta} \sim f e^{-1/f} \tag{22}$$

which is smaller than the negative contribution by a factor of $e^{-1/f} \ll 1$. Thus, for a rare mutation, the typical effects of genetic draft for to reduce f by an $\mathcal{O}(1)$ factor every T_c generations – this will look very much like a fictitious selective force with an effective selection coefficient $s_e \sim -1/T_c$. If this draft timescale is much shorter than the corresponding drift timescale [$f \gg T_c/N$], the genetic draft will be the dominant evolutionary force, and a typical trajectory will look like

$$f(t) \approx f(0) e^{-t/T_c} \tag{23}$$

on the other hand if $f \ll T_c/N$, then genetic drift is the dominant evolutinary force, and there is a significant probability that the mutation will go extinct within the next Nf generations. Thus, the extinction time for a mutation with frequency $f \gg T_c/N$ is roughly $T_{\mathrm{ext}} \sim \frac{1}{T_c} \log(Nf/T_c)$ generations. An entirely analogous process holds for nearly fixed mutations ($1-f \ll 1$), except that in this case sweeps will tend to occur in the mutant background. This changes the sign of the fictitious selective coefficient $s_e \sim +1/T_c$, and will tend to drive the mutant allele to fixation.

Of course, since this is a neutral process ($\langle \delta f \rangle = 0$), these biased typical trajectories must be exactly balanced by rare jumps in the opposite direction. E.g., for rare mutations ($f \ll 1$), jumps to frequencies $f' \gtrsim f$ occur at an instantaneous rate

$$p(f \to f'|f) \cdot dt = \frac{f}{T_c} \frac{df'}{f'} \,, \tag{24}$$

Based on the typical trajectories of mutations, the probability that a mutation at frequency $f \gg T_c/N$ gets a rare jump back to its initial frequency is

$$\int_0^{T_{\mathrm{ext}}} dt \int_f^1 \frac{f e^{-t/T_c}}{T_c} \frac{df'}{f'} \approx f \log\left(\frac{1}{f}\right) \ll 1 \tag{25}$$

This implies that a typical rare mutation does not experience any additional jumps before it goes extinct.

How do the mutations get to their initial positions? Mutations are produced at rate $N\mu$ per generation and drift neutrally up to frequency T_c/N. At this point, the fictitious selective force sets in and prevents the mutation from rising much higher, so that it will typically go extinct in another T_c generations. Thus, for a mutation to reach frequencies $f \gg T_c/N$, it must be lucky enough to have a sweep arise in its genetic background within the first $\sim T_c$ generations after it arises. Thus, mutations are effectively "introduced" at frequency $f \pm df$ at a per generation rate

$$S(f) df \sim \mu \frac{df}{f} \,, \qquad f \gg T_c/N \tag{26}$$

The fate of the mutation will then depend on the magnitude of the introduced f. For $f \ll 1$, the fictitious selective force will cause the frequency of the mutation to decline nearly deterministically as $f(t) = f e^{-t/T_c}$ before going extinct. On the otheer hand, for $1-f \ll 1$, the mutation will typically tend to increase in frequency as $1-f(t) = (1-f)e^{-t/T_c}$ before fixing. In between these two regimes, the mutation may need $\mathcal{O}(1)$ additional big sweeps before its fate is sealed [CAN WE CALCULATE THE TYPICAL NUMBER OF 50% CROSSINGS? DOES IT DIFFER FROM DRIFT CASE?]

Site frequency spectrum

This gives us all the information we need to calculate the site frequency spectrum, $p(f)$, in the intermediate frequency ranges $T_c/N \ll f \ll 1$ or $T_c/N \ll 1-f \ll 1$. For rare mutations ($T_c/N \ll f \ll 1$), we typically catch a mutation as it is declining from its initial jump frequency to extiction. As it does so, it spends an equal amount of time in every logit window, which would ordinarily lead to a $1/f$ frequency specturm. However, we must also account for the source term. In logit space, mutations are introduced at frequency $\log f \pm \mathcal{O}(1)$ at a constant rate μ. This means that the effective number of trajectories passing through point f on their way down to extinction will grow proportionally to $\log(1/f)$. In logit space, we therefore have

$$p(\log f) \cdot d\log f = T_c\mu \log(1/f) \cdot d\log f \tag{27}$$

or

$$p(f) \sim \frac{T_c\mu \log(1/f)}{f} \tag{28}$$

This is steeper than in Kingman coalescent, but not as steep as the Bolthausen-Sznitman case.

For nearly fixed mutations ($T_c/N \ll 1-f \ll 1$) the fictitious selective force is reversed, so that trajectories look like beneficial mutations with effect $+1/T_c$. However, the source term is not reversed: for $1-f \ll 1$, it is increasingly unlikely to catch a sweep all the way up to f. Instead, most trajectories correspond to mutations that hitchhiked to $\mathrm{logit}(f) \sim \mathcal{O}(1)$ and take a typical trajectory to fixation. This means that the site frequency spectrum diverges as $f \to 1$ at rate

$$p(f) \sim \frac{T_c\mu}{1-f} \tag{29}$$

(cannot be produced by any neutral drift process).

Technically speaking, this is all valid for $N\mu \ll 1$ limit. When $N\mu \gg 1$, the mutation frequency will tend to grow deterministically at rate μ in between big sweeps. This will lead to a ficticious mutation-selection balance at frequency $f^* \sim T_c\mu$, similar to what we found in the fluctuating environment case (Cvijovic et al 2015). If $T_c\mu \ll 1$ (a reasonable assumption), this balance will occur at very low frequencies, and most of the SFS will continue to look as above.

Deleterious mutations

We can now think about hitchhiking of a deleterious mutation of cost $-s_d$. Much of the interesting behavior will occur for $s_d T_{\text{fix}} \ll 1$. If $s_d T_c \ll 1$, then selection will be too weak to overcome the bias due to jackpot events. Thus, once the mutation reaches macroscopic frequencies, it will behave similarly to a neutral mutation. In this case, selection is also too weak to bias the frequency of the mutation before the first lucky jackpot occurs. This means that both the microscopic and macroscopic frequencies will behave like a neutral mutation.

For $s_d \gg T_c^{-1}$, the picture changes considerably. In this case, the probability that the mutation catches a lucky jackpot event while microscopic is reduced by a factor of

$$\int_0^\infty \langle f(t)\rangle \frac{1}{T_c} e^{-t/T_c} dt \sim \frac{1}{T_c s_d} \ll 1 \tag{30}$$

but otherwise the frequencies that it can jump to are roughly the same. However, if it does jump to a higher frequency, selection will now be powerful enough to overwhelm the hitchhiking bias term (even for $f \gg 1/2$). This means that he mutation will deterministically decline back down to low frequencies again (and will likely go extinct). In order to fix,a deleterious mutation must catch a lucky hitchhiking event all the way up to $1-f \sim 1/Ns_d$, at which point there is now an $\mathcal{O}(1/2)$ probability that the mutation will fix. Since the final frequency is of order $f \sim e^{-r\ell/s\log(Ns)}$, this jump will require a sweep to occur within $\ell^* \sim \frac{s}{r\log(Ns)} \frac{1}{Ns_d}$ of the focal site. (convinced myself that one jump is the most likely path). The probability that this happens is

$$p_{\text{fix}} \sim \int_0^\infty \frac{1}{N} e^{-s_d t} \lambda \ell^* \sim \frac{\lambda \ell^*}{Ns_d} \sim \frac{1}{N} \cdot \frac{T_c}{N} (T_c s_d)^{-2} \tag{31}$$

Note the slow power-law scaling with s_d (as opposed to the exponential decay in the drift-like case). However, the prefactor is by definition very small in this regime ($T_c \ll N$). Formally as $N \to \infty$, the fixation probability vanishes. Since we had $p_{\text{fix}} \sim 1/N$ for $s_d \ll 1/T_c$, this means that there must be an extremely sharp (boundary layer?) transition for $s_d \sim 1/T_c$ (which we won't investigate in more detail here).

Although the fixation probablity is extremely low, the site frequency spectrum will still contain these strongly deleterious mutations. The dominant contribution will be from deleterious mutations that are lucky enough to hitchhike to high frequencies [jump probabilities supressed from neutral mutations by a factor of $1/(T_c s_d)$] and then deterministically decline back down to extinction. This will (check) give a SFS that is a balance between the source term (jumps to freq f with probability $1/f$) and ordinary logit decline (at rate s_d). This will give a SFS that has very similar shape to neutral mutations (at low freqs) but with a different prefactor ($(T_c s_d)^{-2}$ again?). For high frequencies, the behavior is now very different. For neutral mutations, we had a net bias pushing toward $f = 1$, which produced the $1/(1-f)$ divergence. For deleterious mutations, the selection coefficient still overwhelms the bias and leads to something like (check)

$$p(f) \sim \log(1/f)/f(1-f) \tag{32}$$

which approaches a contant value as $f \to 1$. This will be for frequencies that are not too rare, but will break down for sufficiently small frequencies for which the time to the successful hitchhiking event is small compared to $1/s_d$. For frequencies that are this small, selection won't have a chance to depress the chance that the deleterious mutation hitchhikes, so it will effectively behave as a neutral mutation (check). This will occur for probabilities

$$\lambda \ell^* \frac{1}{s_d} \lesssim 1 \tag{33}$$

or

$$\frac{\lambda s}{r s_d \log(Ns)} \log(1/f) \lesssim 1 \tag{34}$$

or

$$\log(1/f) \lesssim T_c s_d \tag{35}$$

Note that this is much smaller than the $1/T_c s_d$ threshold for a single locus popluation with $N_e = T_c$. If true, this suggests that the SFS is quite different from the single locus case. In the single locus case, deleterious mutations were exponentially surpressed at frequencies above $1/N_e s_d$, and were essentially neutral for frequencies below this threshold. In this case, deleterious mutations are neutral below a much smaller threshold, but can make it to higher frequencies with shapes that are essentially neutral and only suppressed by a factor $1/T_c s_d$ (or squared?)

We can play the same game for beneficial mutations. If $s_b \ll 1/T_c$, then everything will look neutral. But for $s_b \gg 1/T_c$, we will have a very rapid transition to $p_{\text{fix}} \sim s_b$ which is much greater than $1/N$. Again, formally in limit that $T_c \ll N$, the fixation probability of beneficial mutations diverges.

[CAN WE FIGURE OUT DIFFERENCES BETWEEN SINGLE SITE TRAJECTORY AND TIME-VARYING $N_e(t)$? For two timepoint data, there is a big difference. Can we do it even for single site, single locus data if we allow s to vary?]

III. COALESCENT PICTURE

For neutral passenger mutations, we can derive an analogous set of results within a coalescent framework. (This is similar to the treatment in Coop and Ralph, Genetics 2012.) Two individuals sampled today will coalesce if they are sampled from the same mutant or recombinant lineage during a sweep. For a sweep at distance ℓ, this occurs with probability

$$p_c(2|\ell) = \sum_{k=0}^{k_{\max}} g_k(\infty)^2 = e^{-\frac{2r\ell\log(Ns)}{s}} \left[1 + \sum_{k=1}^{\infty} \left(\frac{r\ell}{sk}\right)^{2-\frac{2r\ell}{s}}\right] \approx e^{-\frac{2r\ell\log(Ns)}{s}} \tag{36}$$

which is dominated by the probability of descending from the initial mutant lineage. Integrating over different possible sweeps, this gives a per generation coalescence rate due to sweeps of

$$p_c(2) = \int_0^\infty e^{-\frac{2r\ell\log(Ns)}{s}} \lambda d\ell \sim \frac{\lambda s}{r\log(Ns)} \equiv \frac{1}{T_c} \tag{37}$$

This implies that the typical coalescent time for a pair of individuals is of order T_c, in agreement with the forward time picture above. Note that the pairwise coalescence probability is dominated by anomalously close sweeps that occur within a typical "linkage block", $\ell^* = \frac{s}{r\log(Ns)}$.

The same approach can be used to calculate the probability of coalescence between larger numbers of lineages. E.g. the probability that all n lineages coalesce in a single sweep is

$$p_c(n|\ell) = \sum_{k=0}^{k_{\max}} g_k(\infty)^n \approx e^{-\frac{nr\ell\log(Ns)}{s}} \tag{38}$$

Integrated over different sweeps, this yields a per generation coalescence rate for multiple merger events,

$$p_c(n) \approx \int_0^\infty e^{-\frac{nr\ell\log(Ns)}{s}} \lambda \, d\ell = \frac{2}{n}\frac{1}{T_c} \tag{39}$$

which is on the same order as the pairwise rate, T_c^{-1}. This is a key departure from the Kingman coalescent, which can't be captured by any time-varying $N_e(t)$.

E.g. while a bottleneck could explain an elevated rate of coalescence in a given time interval, it cannot explain a multiple merger coalescence of 4 out of n lineages (which statistically speaking, are more likely to coalesce in disjoint pairs under Kingman statistics). [WHAT IS THE ALLELE FREQUENCY DUAL OF THIS PATTERN? CAN WE USE IT TO SEE EVIDENCE OF MULTIPLE MERGERS WITHOUT HAVING TO POLARIZE THE SFS? FOR SINGLE TIMEPOINT DATA, THIS MUST BE A MULTI-LOCUS OBSERVABLE]

IV. MULTI-LOCUS STATISTICS

[MOSTLY PIECED TOGETHER FROM CONVERSATIONS WITH JAMIE AND DANIEL]. Basic idea is that neutral sites within $\ell^* \sim s/r\log(Ns)$ of each other will hitchhike to an initial frequency f on the same big sweep. [THE INITIAL LD STATISTICS OF THESE MUTATIONS WILL FOLLOW THE RESULTS IN NEUTRAL LD NOTES. PERHAPS INTERESTING INTERPLAY BETWEEN TYPICAL FREQUENCIES THAT HITCHHIKE AND TYPICAL LINKAGE SCALES AT THOSE FREQUENCIES.] There is then a broad range of length scales where $T_c r\ell^* \gg 1$, so that the mutations will approach linkage equilibrium long before the next sweep, with haplotype frequencies

$$f_{AB} = f^2 f_{Ab} = f(1-f) f_{aB} = f(1-f) f_{ab} = (1-f)^2 \tag{40}$$

The future dynamics will then depend on the magnitude of f. If $f \ll 1$, then future sweeps will preferentially occur in the f_{ab} background, and will effectively reduce the frequencies at both loci by the same factor $1-\Delta$. Thus, ficticious selection will drive the two mutations to extinction with approximately the same frequency trajectories (despite the fact that they will almost always be observed to be unlinked).

Conversely, an initial sweep that takes the mutations to frequency $1 - f \ll 1$ will allow the ficticious selection force to drive them to fixation in concert. [PREDICTION FOR NON-MONOTONICITY OF SFS THAT IS MORE ROBUST TO POLARIZATION ERRORS?]

In between, sweeps that drive the mutations to $\mathcal{O}(1/2)$ frequencies will lead to roughly equal haplotype frequencies before the next sweep occurs. In this case, there is now a significant probability that the next sweep will occur in one of the off-diagonal haplotypes and separate the fates of the mutations at future times. [THE RELEVANCE OF THIS CASE WILL DEPEND ON HOW OFTEN MUTATIONS LEAPFROG OVER THIS FREQUENCY RANGE ON WAY TO FIXATION.]

[WHAT ABOUT LONGER HAPLOTYPE TRACTS? FICTICIOUS SELECTION FORCE SHOULD GET WEAK WHEN $(1-f)^n \sim \mathcal{O}(1/2)$]