

# LINKED SELECTION NOTEBOOK

JEREMY BERG

## RUNNING NOTES

November 20, 2012

A sweep occurs yesterday  $l$  base pairs away from a neutral site of interest. The probability that two randomly chosen lineages are forced to coalesce by the sweep is

$$\frac{1}{1 + \theta_b} e^{-\frac{1}{2} l r_{bp} T_{fix}}$$

where  $r_{bp}$  is the recombination rate per base pair,  $\theta_b$  is the population scaled mutation rate of beneficial mutations at the selected site, and  $T_{fix}$  is the duration of the sweep in generations<sup>?</sup>. The  $\frac{1}{1 + \theta_b}$  term is the probability neither lineage mutates off the selected background before coalescing, while  $e^{-\frac{1}{2} l r_{bp} T_{fix}}$  is the probability that neither lineage recombines off the selected background during the course of the sweep.

Now, imagine that sweeps can occur at any base pair along the chromosome, and that they occur homogeneously along the chromosome, and through time, at a rate  $\nu_{BP}$ . Once again examining a single neutral locus, the rate at which recurrent, full, homogeneous soft sweeps force two lineages at the locus to coalesce is equal to

$$\lambda = \nu_{BP} \frac{1}{1 + \theta_b} \int_0^\infty e^{-\frac{1}{2} l r_{bp} T_{fix}} dl \quad (1)$$

$$= \nu_{BP} \frac{1}{1 + \theta_b} \left[ \frac{1}{-\frac{1}{2} r_{bp} T_{fix}} e^{-\frac{1}{2} l r_{bp} T_{fix}} \right]_0^\infty \quad (2)$$

$$= \frac{2\nu_{BP}}{(1 + \theta_b) r_{bp} T_{fix}} \quad (3)$$

The expectation of  $\pi$  under this model is

$$\mathbb{E}[\pi] = \frac{2\mu_n}{\frac{1}{2N} + \frac{2\nu_{BP}}{(1 + \theta_b) r_{bp} T_{fix}}} \quad (4)$$

$$= \frac{\theta_n}{1 + \frac{4N\nu_{BP}}{(1 + \theta_b) r_{bp} T_{fix}}} \quad (5)$$

$$= \frac{r_{bp} \theta_n}{r_{bp} + \frac{4N\nu_{BP}}{(1 + \theta_b) T_{fix}}}. \quad (6)$$

So  $\frac{4N\nu_{BP}}{(1 + \theta_b) T_{fix}}$  appears to be a big, messy, difficult to interpret compound parameter.

## SOFT SWEEPS FROM MUTATION-SELECTION-DRIFT BALANCE.

**November 22, 2012**

To start, we'll just consider standing variation in a deterministic limit. We're interested in diversity at some neutral locus, which is some distance  $l$  away from a selectively significant locus. Prior to some environmental change, mutations are deleterious with selection coefficient  $s_d$ , occur at rate  $\mu_d$ . Assuming  $s_d \gg \mu_d$ , the frequency of deleterious alleles in the population is  $\sim \frac{\mu_d}{s_d}$ . As some point in time, a change in the environment occurs, causing this previously deleterious locus to become advantageous, and to sweep through the population. We are interested in the probability that this sweep forces two randomly chosen lineages at our neutral locus,  $l$  base pairs away, to coalesce.

As with the recurrent mutation case above, the probability that the two lineages are forced to coalesce by a sweep that happened yesterday is

$$Pr(\text{no recombination off sweep})Pr(\text{choose the same mutation class}).$$

**November 24, 2012**

Just as in the recurrent beneficial mutation example,

$$Pr(\text{no recombination off sweep}) = e^{-\frac{1}{2}lr_{bp}T_{fix}} \quad (7)$$

where here I assume that the difference in  $T_{fix}$  that results from the fact that the sweep begins at  $\frac{\mu_d}{s_d}$  instead of  $\frac{1}{2N}$  is negligible (and of course, sweep-like behavior doesn't actually begin at  $\frac{1}{2N}$ , but only once it gets to higher enough frequency to begin stochastic loss; probably, this all doesn't matter too much, but it'll be worth coming back to before publication). It conveniently turns out that, the sampling distribution for deleterious mutations at mutation-selection-drift balance is the Ewens sampling distribution,<sup>?</sup> where the  $N$  in  $\theta$  is replaced by  $Nq$ , where  $q$  is the frequency of deleterious mutations in the population<sup>???</sup> (in this case  $\frac{\mu_d}{s_d}$ ). Thus,

$$Pr(\text{choose the same mutation class}) = \frac{1}{1 + \theta_{d,q}} \quad (8)$$

where  $\theta_{d,q} = 4N\frac{\mu_d}{s_d}\mu_d$ . As such,

$$Pr(\text{sweep forces coalescence}) = \frac{1}{1 + \theta_{d,q}} e^{-\frac{1}{2}lr_{bp}T_{fix}}. \quad (9)$$

Once again, envisioning a scenario where sweeps occur homogeneously across a chromosome of infinite length at rate  $\nu_{BP}$ , the total rate of coalescence at a site is

$$\lambda = \frac{1}{2N} + \frac{\nu_{BP}}{1 + \theta_{d,q}} \int_0^\infty e^{-\frac{1}{2}lr_{bp}T_{fix}} dl \quad (10)$$

$$= \frac{1}{2N} + \frac{2\nu_{BP}}{(1 + \theta_{d,q})r_{bp}T_{fix}}. \quad (11)$$

The expectation of  $\pi$  is thus

$$\mathbb{E}[\pi] = \frac{r_{bp}\theta_n}{r_{bp} + \frac{4N\nu_{BP}}{(1 + \theta_{d,q})T_{fix}}}. \quad (12)$$

The important insight is that, at least under certain assumptions, the effect of sweeps from standing, deleterious mutation on linked neutral diversity is essentially the same as that for sweeps from new, recurrent mutations, with the only different being the magnitude of the population scaled mutation rate parameter at the linked, non-neutral sites.

**November 25, 2012**

It seems like  $\theta_{d,q}$  ought to be smaller than  $\theta_b$  in most cases, and thus sweeps from recurrent beneficial mutation during the progress of the sweep ought to have a smaller effect on diversity than sweeps from standing deleterious variation.

**December 3, 2012**

It's been too long since my last entry. Need to get better about that.

First, it seems that my initial reading of the Hartl and Campbell paper<sup>?</sup> was a bit off. Their equations (2) and (9) together suggest that if one samples two lineages in the deleterious class, the probability that they are of the same type is

$$Pr(k = 1 \mid n = 2) = \frac{1}{1 + \theta_a} \quad (13)$$

where  $\theta_a = 4N\mu_a q_{eq}$ . What I neglected was that their  $\mu_a$  is not the same as the general mutation rate to a alleles. They define  $\mu_a$  as follows:

$$\mu_a = \frac{p\mu_{Aa} + q\mu_{aa}}{p\mu_{Aa} + q(1 - \mu_{aA})} \quad (14)$$

where  $\mu_{ij}$  is the mutation rate from state i to state j. Roughly in parallel to the argument they put forward, if

$$\mu_{Aa} = \mu_{aa} = \mu \quad (15)$$

$$\mu_{aA} = 0 \quad (16)$$

and,

$$q = \frac{\mu}{s} \quad (17)$$

then,

$$\mu_a = \frac{1}{1 + \frac{1}{s}}. \quad (18)$$

As

$$\theta_a = 4N\frac{\mu}{s} \frac{1}{1 + \frac{1}{s}} = 4N\frac{\mu}{1 + s}. \quad (19)$$

Equations (13) and (19) together thus give us our probability of sampling two of the same type under mutation-selection drift balance. Of course, it turns out that for small  $s$  this yields essentially the same value as the Hermisson and Pennings result<sup>?</sup>. As  $s$  grows larger, the Hartl and Campbell result returns a lower probability than the Hermisson and Pennings result. However, for most biologically plausible parameters, it seems like the results should be fairly close.

In the mean time, I've managed to write up a single locus Wright-Fisher simulation of an infinite alleles model for arbitrary mutation, selection, and population size parameters. It produces figures like this

This is a cumulative frequency plot, where each partition is a different independent mutation. Blue denotes the ancestral, neutral allele. Red and green mutations are selectively equivalent, the only difference being that the frequency trajectories of mutations that end up being members of the

fixing class are colored green, while those that are lost before fixation are colored red. The dashed line denotes the point at which the mutation switches from being deleterious to beneficial.

These figures are pretty, but the real purpose is to use this simulation as a method of double checking my theoretical predictions.

**December 12, 2012**

In the past week or so we've managed to sort through some of this stuff and discover that my initial intuition, (which didn't even make it into the notebook because we got sidetracked by the Hartl and Campbell stuff) was correct. That is, as long as mutation continues during the course of the sweep (which is really the only biologically reasonable case, as there's no reason for mutation to stop once the sweep starts), then the Hermisson and Pennings result holds even if adaptation occurs from recurrent mutation to standing variation instead.

I had a chance to present the basic recurrent model at CoopIbarra this past Monday (and to chat with Michael yesterday morning), and folks who hadn't seen it yet seemed fairly receptive, and also pointed to working out the effect on linkage disequilibrium, or some other summary of haplotype structure. These results already exist for individual hard<sup>?</sup> <sup>?</sup> <sup>?</sup> and soft<sup>?</sup> sweeps, so it seems like putting it into the recurrent model shouldn't be too hard.

#### THE RELATIONSHIP BETWEEN DIVERSITY AND POPULATION SIZE

**September 4, 2013**

We'll begin from the recurrent soft sweep result above, namely the observation that

$$\lambda = \frac{1}{2N} + \frac{2\nu}{(1 + \theta_b) r_{bp} T_{fix}}. \quad (20)$$

Next, assume  $N$  is very large, such that  $\theta_b = 4N\mu_b \gg 1$ , and thus we can safely ignore the 1 and let

$$\lambda = \frac{1}{2N} + \frac{2\nu}{4N\mu_b r_{bp} T_{fix}}. \quad (21)$$

We want to consider the effect of recurrent soft sweeps on neutral genetic diversity, but we don't want to land in the Desai/Neher (fitness class coalescent) world<sup>?</sup> <sup>?</sup> <sup>?</sup>, where new beneficial mutations occur on the background of the last beneficial mutation before it has managed to fix. We need the rate of sweeps to be slow enough that the previous sweep has managed to fix before the next comes along. We can ensure that we are in this parameter range by considering that the waiting time until the next sweep (starting from the beginning of the last one) can be written as

$$\frac{1}{\nu} = \frac{1}{\alpha} + \frac{1}{2N\mu_b s} \quad (22)$$

where  $\alpha$  describes the rate of ecological necessity (i.e. the rate at which the environment changes such that new mutations will be beneficial; they are assumed deleterious or neutral in the intervening time). We can then solve for

$$\nu = \frac{2N\mu_b s \alpha}{2N\mu_b s + \alpha} \quad (23)$$

Because the population size is large enough to be generating predominantly soft sweeps, but the environment changes rather infrequently, we can assume  $2N\mu_b s \gg \alpha$ , and thus simply take  $\nu = \alpha$ .

It is also worth noting that because  $s = \frac{\log(2Ns)}{T}$ , we can rewrite our assumption as

$$\frac{2N\mu_b \log(2Ns)}{T} \gg \alpha \quad (24)$$

and then multiple both sides by  $\frac{1}{T}$  to get

$$\frac{2N\mu_b \log(2Ns)}{T^2} \gg \frac{\alpha}{T}. \quad (25)$$

This sets an upper bound on the value of  $\frac{\alpha}{T}$ , but we can do one better. To make things easy, assume that the rate of ecological change is sufficiently slow relative to the duration of the sweeps, such that the next sweep cannot begin until the previous one has fixed

$$\frac{1}{\alpha} > T \quad (26)$$

which we can rewrite as

$$\frac{1}{T^2} > \frac{\alpha}{T} \quad (27)$$

thus setting an upper bound on  $\frac{\alpha}{T}$ .

Now, we rewrite (21) as

$$\lambda = \frac{1}{2N} + \frac{2\alpha}{4N\mu_b r_{bp} T} \quad (28)$$

and therefore note that

$$\pi = \frac{2\mu_n}{\frac{1}{2N} + \frac{2\alpha}{4N\mu_b r_{bp} T}} \quad (29)$$

$$= \frac{4N\mu_n}{1 + \frac{\alpha}{\mu_b r T}} \quad (30)$$

$$= \frac{4N\mu_n}{1 + \frac{\alpha}{T} \frac{1}{\mu_b r}}. \quad (31)$$

We can restate our assumption  $\frac{1}{T^2} > \frac{\alpha}{T}$  as

$$\frac{T}{\alpha} > T^2. \quad (32)$$

We find that coalescence is dominated by drift when  $\frac{1}{\mu_b r} \ll \frac{T}{\alpha} \approx T^2$  (resulting in  $\pi \approx 4N\mu_n$ ), and by sweeps when  $\frac{1}{\mu_b r} \gg \frac{T}{\alpha} \approx T^2$  in which case

$$\pi \approx \frac{4NT\mu_n\mu_b r_{bp}}{\alpha} \quad (33)$$

and the transition between the two parameter regimes occurs roughly when  $\frac{1}{\mu_b r} \approx T^2$ .

Just to get a rough sense of where this puts us, if sweeps are occurring in a population of  $N = 1,000,000$  with  $s = 10^{-5}$  ( $2Ns = 20$ ), and  $r_{bp} = 10^{-8}$ , then drift dominates if  $\mu_b \gg 10^{-3}$  (because sweeps are so soft that they hardly force any coalescence), while draft dominates if  $\mu_b \ll 10^{-3}$ . If instead  $s = 10^{-4}$ , then  $\mu_b \gg 10^{-2}$  in order for drift to dominate, while at  $s = 10^{-3}$  the mutation rate would have to be greater than one in order for drift to dominate.

These calculations assume that  $\alpha$  is near its maximum value of  $\sim \frac{1}{T}$ . As  $\alpha$  decreases, so too must  $\mu_b$  in order for sweeps to continue to dominate coalescence. Eventually, of course, this approximation breaks down because the sweeps start to become predominantly hard (i.e.  $4N\mu_b < 1$ ). Many things

seem to happen at once here, as we also reach the point at which  $\nu = \alpha$  is no longer a valid approximation either (because the waiting time for a new mutation becomes non-negligible). It seems like these should happen at roughly the same time, but I haven't worked through that yet.

## SWEEPS FROM IBD STANDING VARIATION

**December 12, 2012**

It also seems like we might relatively easily be able to make progress on the hard sweep from IBD standing variation case, and that once this is done we should be able to include these sort of sweeps in a general recurrent model. Graham sent me a few papers about the conditional genealogy of a sample in which a subset of chromosomes are observed to carry a particular mutation, notably one from Griffiths and Tavaré<sup>?</sup>, and this seems significant enough, with enough of a clear way forward that I'm going to create a separate entry here.

Another really useful thing to do might be to work out the sampling formula for the number of haplotypes as a function of the recombination distance from the sweeping locus (which we suspect should have a roughly Ewens like distribution, with the  $\theta$  parameter given by something like  $2Nrf$ , where  $r$  is the recombination rate, while  $f$  is the frequency from which the sweep started), and then develop a likelihood framework to infer  $f$  from the haplotype structure on either side of the sweep.

In addition to this analytical work, another objective is going to be writing out a structured coalescent simulation on a sweeping background, probably using Graham and Molly's trick from 2005<sup>?</sup> to simulate the trajectory of the sweeping allele, and then just run coalescent simulations under that background.

**February 4, 2012**

The simulation is pretty much written. More on that later hopefully. At present, the problem is this:

Assuming a sample of three lineages, what is the probability that there are 2 different haplotypes at a distance from the selected site of  $l_1$ , given that there is only one haplotype at a distance of  $l_1 - 1$ . If  $b_1$ ,  $b_2$  and  $b_3$  are the external branches of the tree, and  $b_4$  is the one internal branch, then

$$P(2 \text{ haplotypes at } l_1 \mid 1 \text{ haplotype at } l_1 - 1) = \tag{34}$$

$$P(\text{rec before } \tau \text{ and no rec on other two lineages before they coalesce}) + \tag{35}$$

$$P(\text{no rec before } \tau, \text{ then rec after } \tau \text{ but before coalescence}) \tag{36}$$

where  $\tau$  is the point in time at which the other two lineages coalesce with one another (conditional on previous base assumed and dropped from here out).

First recognize that

$$P(\text{rec before } \tau, \text{ then no rec on other two lineages before they coalesce}) = \tag{37}$$

$$P(\text{rec before } \tau)P(\text{no rec on other two lineages before they coalesce}) \tag{38}$$

$$P(\text{rec before } \tau) = \int_0^\infty 3r(1-f)e^{-3tr(1-f)}e^{-\frac{t\binom{3}{2}}{2N_e f}} dt \quad (39)$$

$$= \frac{3r(1-f)}{3r(1-f) + \frac{3}{2N_e f}} \quad (40)$$

$$= \frac{6N_e r(1-f)}{6N_e r f(1-f) + 3} \quad (41)$$

$$= \frac{4N_e r f(1-f)}{4N_e r f(1-f) + 2} \quad (42)$$

$$P(\text{no rec on other two lineages before they coalesce}) = \int_0^\infty \frac{1}{2N_e f} e^{-2tr(1-f)} e^{-\frac{t}{2N_e f}} dt \quad (43)$$

$$= \frac{\frac{1}{2N_e f}}{2r(1-f) + \frac{1}{2N_e f}} \quad (44)$$

$$= \frac{1}{4N_e r f(1-f) + 1} \quad (45)$$

So,

$$P(\text{rec before } \tau, \text{ then no rec on other two lineages before they coalesce}) = \quad (46)$$

$$\left( \frac{1}{4N_e r f(1-f) + 1} \right) \left( \frac{4N_e r f(1-f)}{4N_e r f(1-f) + 2} \right) \quad (47)$$

Next

$$P(\text{no rec before } \tau, \text{ then rec after } \tau \text{ but before coalescence}) = \quad (48)$$

$$P(\text{rec after } \tau \text{ but before coalescence} \mid \text{no rec before } \tau) P(\text{no rec before } \tau) \quad (49)$$

$$P(\text{no rec before } \tau) = \int_0^\infty \frac{\binom{3}{2}}{2N_e f} e^{-3r(1-f)} e^{-\frac{t\binom{3}{2}}{2N_e f}} dt \quad (50)$$

$$= \frac{\frac{3}{2N_e f}}{3r(1-f) + \frac{3}{2N_e f}} \quad (51)$$

$$= \frac{3}{6N_e r f(1-f) + 3} \quad (52)$$

$$= \frac{2}{4N_e r f(1-f) + 2} \quad (53)$$

$$P(\text{rec after } \tau \text{ but before coalescence} \mid \text{no rec before } \tau) = \quad (54)$$

$$= \int_0^\infty 2r(1-f)e^{-2r(1-f)}e^{-\frac{t}{2N_e f}} dt \quad (55)$$

$$= \frac{2r(1-f)}{2r(1-f) + \frac{1}{2N_e f}} \quad (56)$$

$$= \frac{4N_e r f(1-f)}{4N_e r f(1-f) + 1} \quad (57)$$

So,

$$P(\text{no rec before } \tau, \text{ then rec after } \tau \text{ but before coalescence}) = \quad (58)$$

$$\left( \frac{2}{4N_e r f(1-f) + 2} \right) \left( \frac{4N_e r f(1-f)}{4N_e r f(1-f) + 1} \right) \quad (59)$$

Letting  $R_f = 4N_e r f(1-f)$ , then

$$P(2 \text{ haplotypes at } l_1 \mid 1 \text{ haplotype at } l_1 - 1) = \quad (60)$$

$$\frac{R_f}{R_f + 2} \frac{1}{R_f + 1} + \frac{2}{R_f + 2} \frac{R_f}{R_f + 1} \quad (61)$$

$$= \frac{3R_f}{(R_f + 2)(R_f + 1)} \quad (62)$$

which is the Ewens Sampling result for  $P(k = 2 \mid n = 3, R_f)$ , where  $R_f$  is filling the role of  $\theta$ .



**February 5, 2013**

Now I'm going to move on to thinking about the probability of having a recombination event that increased the number of haplotypes from 1 to 2 at  $l_1$ , followed by a recombination event that increases the number of haplotypes from 2 to 3 at  $l_2$ , where  $l_1 < l_2$  are distances in base pairs from the selected site. In other words

$$P(\{3\} \rightarrow \{2, 1\} \text{ at } l_1 \text{ and } \{2, 1\} \rightarrow \{1, 1, 1\} \text{ at } l_2) = \quad (63)$$

$$P(\{l_1 b_3\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) + \quad (64)$$

$$P(\{l_1 b_4\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) + \quad (65)$$

$$P(\{l_1 b_5\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) + \quad (66)$$

$$P(\{l_2 b_3\} \text{ and } \{\{l_1 b_1\} \text{ or } \{l_1 b_2\}\}) + \quad (67)$$

$$P(\{l_2 b_4\} \text{ and } \{\{l_1 b_1\} \text{ or } \{l_1 b_2\}\}) + \quad (68)$$

$$P(\{l_2 b_5\} \text{ and } \{\{l_1 b_1\} \text{ or } \{l_1 b_2\}\}) \quad (69)$$

where  $b_1$  and  $b_2$  are the branches connecting two of the tips to the first coalescent event, which occurs at  $\tau_2$  (the first moment when there are two ancestors of the sample),  $b_3$  is the branch from the third tip leading back to time  $\tau_2$ ,  $b_4$  is contiguous with  $b_3$  and leads from  $\tau_2$  back to  $\tau_1$ , where it coalesces with  $b_5$ , which is the branch subtending  $b_1$  and  $b_2$ ;  $l_i b_j$  is a recombination event occurring at sequence location  $i$  on branch  $j$ . Moving forward

$$P(\{l_1 b_3\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) = \quad (70)$$

$$P(\{l_1 b_3\}, \text{ then } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) + \quad (71)$$

$$P(\{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{l_1 b_3\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) \quad (72)$$

These two events have equal probability and thus

$$P(\{l_1 b_3\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) = \quad (73)$$

$$2P(\{l_1 b_3\}, \text{ then } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) \quad (74)$$

$$P(\{l_1 b_3\}, \text{ then } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) = \quad (75)$$

$$r(1-f) \int_0^\infty e^{-tr l_1(1-f)} e^{-2tr l_2(1-f)} e^{-\frac{t \binom{3}{2}}{2N_e f}} dt \quad (76)$$

$$2r(1-f) \int_0^\infty e^{-tr l_1(1-f)} e^{-2tr l_2(1-f)} e^{-\frac{t \binom{3}{2}}{2N_e f}} dt \quad (77)$$

$$\frac{1}{2N_e f} \int_0^\infty e^{-tr l_1(1-f)} e^{-2tr l_2(1-f)} e^{-\frac{t \binom{3}{2}}{2N_e f}} dt \quad (78)$$

$$\frac{1}{2N_e f} \int_0^\infty e^{-tr l_1(1-f)} e^{-tr l_2(1-f)} e^{-\frac{t}{2N_e f}} dt = \quad (79)$$

$$\frac{[2N_e r f(1-f)] [4N_e r f(1-f)] [1] [1]}{[2N_e r f(1-f)l_1 + 2N_e r f(1-f)l_2 + 3]^3 [2N_e r f(1-f)l_1 + 2N_e r f(1-f)l_2 + 1]} \quad (80)$$

Letting  $R_f = 2N_e r f(1 - f)$ , we can write down

$$P(l_1 b_3, l_2 b_1 \text{ or } l_2 b_2) = \frac{4R_f^2}{(R_f l_1 + 2R_f l_2 + 3)^3 (R_f (l_1 + l_2) + 1)} \quad (81)$$

Next

$$P(\{l_1 b_4\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) = \quad (82)$$

$$P(\{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{l_1 b_4\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) = \quad (83)$$

$$2r(1 - f) \int_0^\infty e^{-tr l_1(1-f)} e^{-2tr l_2(1-f)} e^{-\frac{t}{2N_e f} \binom{3}{2}} dt \quad (84)$$

$$\frac{1}{2N_e f} \int_0^\infty e^{-tr l_1(1-f)} e^{-2tr l_2(1-f)} e^{-\frac{t}{2N_e f} \binom{3}{2}} dt \quad (85)$$

$$r(1 - f) \int_0^\infty e^{-tr l_1(1-f)} e^{-tr l_2(1-f)} e^{-\frac{t}{2N_e f}} dt \quad (86)$$

$$\frac{1}{2N_e f} \int_0^\infty e^{-tr l_1(1-f)} e^{-tr l_2(1-f)} e^{-\frac{t}{2N_e f}} dt = \quad (87)$$

$$\frac{[2r(1 - f)] \left[ \frac{1}{2N_e f} \right]^2 [r(1 - f)]}{\left[ r l_1(1 - f) + 2r l_2(1 - f) + \frac{3}{2N_e f} \right]^2 \left[ r l_1(1 - f) + r l_2(1 - f) + \frac{1}{2N_e f} \right]^2} \quad (88)$$

Again, multiplying top and bottom by  $(2N_e f)^4$ , and letting  $R_f = 2N_e r f(1 - f)$

$$P(\{l_1 b_4\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) = \frac{2R_f^2}{(R_f l_1 + 2R_f l_2 + 3)^2 (R_f l_1 + R_f l_2 + 1)^2} \quad (89)$$

Next (now omitting the integrals for brevity's sake)

$$P(\{l_1 b_5\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}) = \quad (90)$$

$$P(\{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\}, \text{ then } \{b_1 \text{ and } b_2 \text{ coalesce}\}, \text{ then } \{l_1 b_5\}, \text{ then } \{b_4 \text{ and } b_5 \text{ coalesce}\}) = \quad (91)$$

$$\left[ \frac{2r(1 - f)}{3r l_2(1 - f) + \frac{3}{2N_e f}} \right] \left[ \frac{\frac{1}{2N_e f}}{3r l_2(1 - f) + \frac{3}{2N_e f}} \right] \left[ \frac{r(1 - f)}{2r l_2(1 - f) + \frac{1}{2N_e f}} \right] \left[ \frac{\frac{1}{2N_e f}}{2r l_2(1 - f) + \frac{1}{2N_e f}} \right] = \quad (92)$$

$$\frac{2R_f^2}{(3R_f l_2 + 3)^2 (2R_f l_2 + 1)^2} \quad (93)$$

and if we carry out the same calculation for  $P(\{\{l_1 b_1\} \text{ or } \{l_1 b_2\}\} \text{ and } \{l_2 b_5\})$ , we find that it is equivalent to the calculation we just carried out for  $P(\{l_1 b_5\} \text{ and } \{\{l_2 b_1\} \text{ or } \{l_2 b_2\}\})$  (which makes sense if you squint at the tree a bit).

**September 3, 2013**

Ok! Been a while since I've written anything here. Polygenic selection paper is up on arXiv and submitted to PLoS Genetics, so time to get back to sweep work. While variations of the above can in principle be used to infer the frequency parameter, we now live in a world where direct inference of the distribution of ARGs from sequence data seems to be possible<sup>?</sup>, and thus it may be worth considering how to do this inference problem if we get to know the coalescent times at the selected site.

Imagine we have an allele that begins at frequency  $f$  (having been neutral before that), and sweeps with strength  $s$  in a population of size  $N$ . Consider then sampling  $n$  lineages at  $t = 0$  (the present), and following them backward in time. To a first approximation, no coalescence occurs during the sweep, and thus we might imagine that we could write down a likelihood function that looks like

$$\mathcal{L}(s, f \mid \vec{\tau}) = \left( \frac{1}{Nf} \right)^{n-1} e^{-\frac{\binom{n}{2}}{Nf}(\tau_1 - \tau_s)} \prod_{i=2}^{n-i+1} e^{-\frac{\binom{n-i+1}{2}}{Nf}(\tau_i - \tau_{i-1})} \quad (94)$$

where  $\tau_s = \frac{\ln(N \frac{1-f}{f})}{s}$  is the number of generations (backward in time) that it take for the sweep to go from fixation down to frequency  $f$ .

Unfortunately, coalescence can occur during the sweep, and thus by completely prohibiting it in our likelihood function, we could find ourselves in situations where we have no choice but to infer totally ridiculous selection coefficients, simply because that's the only way that we can have  $\tau_s < \tau_1$ .

Alternatively, if we allow coalescence during the sweep, we can make the assumption that the frequency path is deterministic, and thus use that to get coalescence probabilities in each generation. We take the number of beneficial alleles present in generation  $t$  (before the present) to be a deterministic function of  $s$  and  $t$

$$N(t) = N - \frac{Ne^{st}}{N-1+e^{st}} \quad (95)$$

where  $N$  is the population size. If, for example, we have a single coalescent event that occurs at  $t = \tau_1 < \tau_s$ , then we can calculate the probability of that coalescent event occurring during the sweep as

$$\frac{1}{N(\tau_1)} e^{-\binom{n}{2} \int_0^{\tau_1} \frac{1}{N(t)} dt} \quad (96)$$

where

$$\int \frac{1}{N(t)} dt = \int \frac{1}{N - \frac{Ne^{st}}{N-1+e^{st}}} dt = \frac{e^{st}}{(N-1)Ns} + \frac{t}{N} + C. \quad (97)$$

We then need to prohibit coalescence from  $\tau_1$  to  $\tau_s$ , and then start running the standard coalescent at rate  $\frac{1}{Nf}$

$$\left( \frac{1}{Nf} \right)^{n-2} e^{-\binom{n-1}{2} \left( \int_{\tau_1}^{\tau_s} \frac{1}{N(t)} dt - \frac{1}{Nf}(\tau_2 - \tau_s) \right)} \prod_{i=3}^{n-i+1} e^{-\frac{\binom{n-i+1}{2}}{Nf}(\tau_i - \tau_{i-1})} \quad (98)$$

and thus the full likelihood can be written as

$$\mathcal{L}(s, f \mid \vec{\tau}) = \frac{1}{N(\tau_1)} e^{-\binom{n}{2} \int_0^{\tau_1} \frac{1}{N(t)} dt} \left( \frac{1}{Nf} \right)^{n-2} e^{-\binom{n-1}{2} \left( \int_{\tau_1}^{\tau_s} \frac{1}{N(t)} dt - \frac{1}{Nf}(\tau_2 - \tau_s) \right)} \prod_{i=3}^{n-i+1} e^{-\frac{\binom{n-i+1}{2}}{Nf}(\tau_i - \tau_{i-1})}. \quad (99)$$

In this example, we've imagined that only one coalescent event occurs during the sweep, but in practice it is possible that more than one could (particularly if we are evaluating the likelihood for a fairly small  $s$ ), in which case we simply include terms such as  $\frac{1}{N(\tau_2)} e^{-\binom{n-1}{2} \int_{\tau_1}^{\tau_2} \frac{1}{N(t)} dt}$ , and take the product for the neutral phase starting with  $i = 4$ , etc.

**September 13, 2013**

Consider now a case where a selectively neutral allele arises in some ancestral population and drifts up to some frequency  $f$ . Individuals from this ancestral population then go on to found multiple independent populations in some new environment. A bottleneck occurs for some number of generations during the founding event, after which the previously neutral mutation becomes beneficial in all of the new environments. The biological model one should have in mind is something like Sticklebacks colonizing lakes and streams in British Columbia, or *Mimulus* on heavy metal soils in California.

We'll generally focus on pairwise quantities here, although thinking about the joint AFS later on may be interesting. One useful thing to consider is the expectation of  $F_{ST}$  between two of the founder populations as we move away from the selected site. Consider that  $F_{ST}$  can be expressed as

$$F_{ST} = \frac{T_T - T_S}{T_T} \quad (100)$$

where  $T_T$  is the average pairwise coalescence time of a pair of lineages sampled randomly from the two founder populations, and  $T_S$  is the average pairwise coalescence time of two lineages sampled from the same population.

If  $T_D$  is the average pairwise coalescence time of two lineages sampled from different populations, then because sampling two lineages randomly from the total population results in a 50% chance of sampling lineages from the same population, and 50% chance of sampling from different populations,  $T_T = \frac{T_D + T_S}{2}$ . Therefore

$$F_{ST} = \frac{\frac{T_D + T_S}{2} - T_S}{\frac{T_D + T_S}{2}} = \frac{T_D + T_S - 2T_S}{T_D + T_S} \quad (101)$$

$$= \frac{T_D - T_S}{T_D + T_S} \quad (102)$$

and so if we can simply write down expectations for  $T_D$  and  $T_S$  in terms of relevant parameters, then we can get an expression describing the expected value of  $F_{ST}$  as a function of distance from the recombination site and other relevant population genetic parameters.

**October 2, 2013**

$$T_S = P(\text{rec out during sweep})\mathbb{E}[T_2] + \quad (103)$$

$$P(\text{no rec out during sweep})P(\text{rec out during bottleneck before coal})\mathbb{E}[T_2] \quad (104)$$

$$P(\text{no rec out during sweep})P(\text{no rec out or coal during bottleneck}) \quad (105)$$

$$P(\text{ancestral rec before coal})\mathbb{E}[T_2] \quad (106)$$

or, for shorthand, if  $P_{S,S}$  is the probability of a recombination event happening during the sweep,  $P_{B,S}$  the probability of recombining before coalescing during the bottleneck,  $P_{EB,S}$  is the probability of escaping the bottleneck without recombining or coalescing, and  $P_{A,S}$  the probability of recombining before coalescing during the ancestral phase, then

$$T_S = P_S\mathbb{E}[T_2] + (1 - P_S)P_{B,S}\mathbb{E}[T_2] + (1 - P_S)P_{EB,S}P_A\mathbb{E}[T_2]. \quad (107)$$

In contrast, when considering expected coalescent time of two sequences sampled from different populations

$$T_D = P(\text{rec during sweep})\mathbb{E}[T_2] + P(\text{no rec in sweep})P(\text{rec in bottleneck})\mathbb{E}[T_2] + \quad (108)$$

$$P(\text{no rec in sweep})P(\text{no rec bottleneck})P(\text{rec ancestral})\mathbb{E}[T_2] \quad (109)$$

or

$$T_D = P_S\mathbb{E}[T_2] + (1 - P_S)P_{B,D}\mathbb{E}[T_2] + (1 - P_S)(1 - P_{B,D})P_A\mathbb{E}[T_2]. \quad (110)$$

Plugging in to our expression above, we find that the numerator can be written as

$$(1 - P_S)[P_{B,D} - P_{B,S} + P_A(1 - P_{B,D} - P_{EB,S})]\mathbb{E}[T_2], \quad (111)$$

and the denominator as

$$2P_S + (1 - P_S)[(1 - P_A)P_{B,D} + P_{B,S} + P_A(1 + P_{EB,S})]\mathbb{E}[T_2]. \quad (112)$$

Our expression for  $F_{ST}$  is thus

$$F_{ST} = \frac{(1 - P_S)[P_{B,D} - P_{B,S} + P_A(1 - P_{B,D} - P_{EB,S})]}{2P_S + (1 - P_S)[(1 - P_A)P_{B,D} + P_{B,S} + P_A(1 + P_{EB,S})]}. \quad (113)$$

From previous work above, we have

$$P_A = \frac{R_f}{1 + R_f} \quad (114)$$

$$P_S = 1 - e^{-2r \int_0^{T_{\text{fix}}} 1 - X(t) dt} \quad (115)$$

$$\text{while} \quad (116)$$

$$P_{B,D} = 1 - e^{-2rT_{\text{bot}}(1-f)} \quad (117)$$

$$P_{B,S} = r(1-f) \int_0^{T_{\text{bot}}} e^{-2tr(1-f)} e^{-\frac{t}{2NfB}} dt \quad (118)$$

$$P_{EB,S} = e^{-2T_{\text{bot}}r(1-f) - \frac{T_{\text{bot}}}{2NfB}} \quad (119)$$