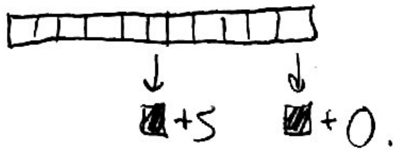
Linked selection from "classic selective sweeps"

Last time, we saw that the appealing & commonly used QLE/ "independent sites" approx breaks down when $s \gtrsim r\Delta\ell$, which is not so uncommon for strongly selected mutations ($s \gtrsim 10^{-3}$) & realistic recombination rates ($r \sim \mu \sim 10^{-8} - 10^{-10}$).

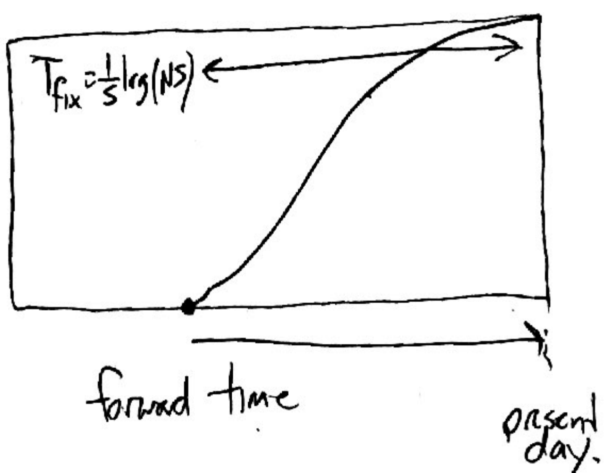
Today we'll discuss some methods for treating cases where $s \gg r\Delta\ell$.

$\Rightarrow$ as before, the simplest scenario to start with is where we have a single strongly beneficial mutation ($s \gg 1/N$,) on a genome w/ other neutral sites:



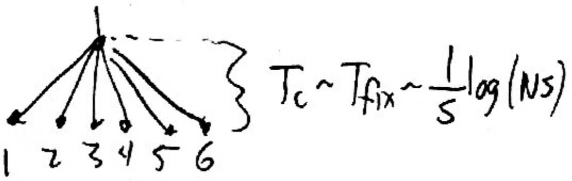
$\Rightarrow$ in this case, we know exactly how the beneficial mutation behaves: it will establish & grow as

$$f(t) = \frac{\frac{1}{Ns} e^{st}}{1 + \frac{1}{Ns} e^{st}}$$



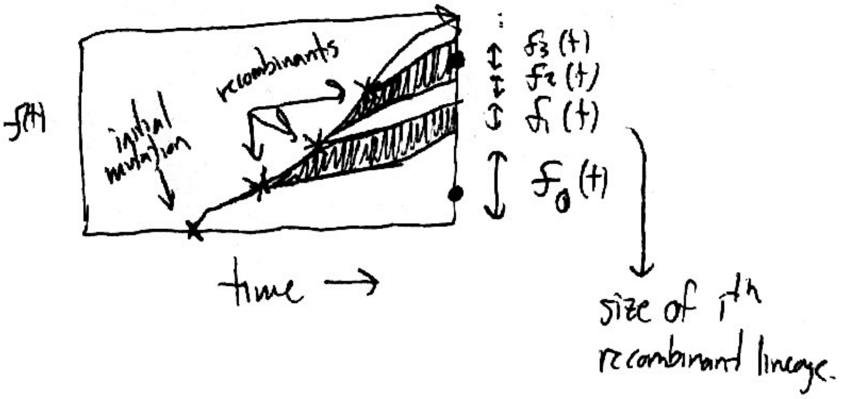
we can then think about the genealogy of a sample of 2 (or more indindvals) sampled at the present day, conditioned on this trajectory.

⇒ in Lecture 15 (p. 6-8), we worked out genealogy at selected locus: everyone descended from original mulation event w/ star-like tree:



⇒ for other loci $(r>0)$, we have to account for the fact that some lineages may not have descended from the original mutation event, but a recombinant lineage produced by recombination between the selected genetic background + wildtype population.

i.e., lineage structure of population might look like:



⇒ 2 lineages coalesce during sweep if drawn from same lineage $\left(P_c = \sum_{i=0}^{n} (f_i)^2 \right)$, $T_c \leq \frac{1}{s}\log(Ns)$

if the individuals aren't from the same sweep lineage, then their ancestors will remain as uncoalesced lineages before origin of sweep. Since no more selection, will coalesce neutrally from here on back ($T_c \sim N \gg T_{fix}$). thus, the coalescence process looks like:

1. if same $\overset{\text{sweep}}{\uparrow}$ lineage, coalesce "immediately" ($T_c \ll N$)
2. otherwise, coalesce neutrally ($T_c \sim N$)

$\Rightarrow$ crucial step is to determine the sweep lineage sizes, $f_i(t)$

$\Rightarrow$ we will do this now using a heuristic analysis.

1. at short times ($t \ll T_{fix} \sim \frac{1}{s}\log(Ns)$), the selected mutation will still be at low frequency ($f_s(t) \ll 1$)

$\Rightarrow$ most recombinations will occur between a selected haplotype and the wildtype [illegible] haplotype, at rate $\rho f_{sel}(t)$, producing new recombinant lineages.

$\Rightarrow$ each of these recombinant lineages will satisfy the SDE ($t \ll T_{fix}$)

$$\boxed{\frac{\partial f_i}{\partial t} = s f_i(t) - r f_i(t) + \sqrt{\frac{f_i}{N}}\eta_i}$$

we know exactly how these recombinant lineages behave:

⇒ w/ probability $p_{est} \sim (s-r) \overset{\swarrow \text{ since } \rho \ll s}{\sim} s$, the lineage will establish and grow as $f_k(t) \sim \frac{1}{Ns} e^{(s-\rho)(t-\tau_k)}$

where τ_k is the establishment time of k^{th} recombinant

⇒ since all recombinants grow at the same rate, their relative sizes, $f_k(t)/f_{k-1}(t) = e^{(s-\rho)(\tau_{k-1}-\tau_k)}$, will only depend on the difference in establishment times and will be "frozen in" for remainder of the sweep, even when $f_{sel}(t)$ is no longer rare.

⇒ thus, understanding what happens during the early phase (when $f_{sel}(t)$ is rare) will be useful for understanding the lineage sizes @ end of the sweep: $f_k(\infty) = \frac{f_k(t)}{\sum_{k'} f_{k'}(t)} = \frac{e^{-s\tau_k}}{\sum_{k'} e^{-s\tau_{k'}}}$

⇒ crucial step is to calculate the establishment times, τ_k, of the k^{th} recombinant lineage. By convention, will take $\tau_0 = 0$ (establishment time of original mutant lineage) from our discussion above, <u>successful</u> recombinants are produced at rate

$$\underbrace{N\rho f_1(t)}_{\text{total \# recombinants produced @ gen t.}} \cdot \underbrace{s}_{\text{prob that recombinant survives genetic drift.}} = N\rho\cdot\frac{1}{Ns}e^{st}\cdot s = \underbrace{\rho e^{st}}_{\text{very similar to Luria delbrück problem from PS1}}$$

$\Rightarrow$ avg # successful recombinants produced by time t $= \int_0^t dt'\,\rho e^{st'}$

$\Rightarrow$ heuristically, time to <u>first</u> successful recombinant (τ_1) occurs when $\int_0^{\tau_1}\rho e^{st}dt \sim 1$

$$\Rightarrow 1 \sim \frac{\rho}{s}(e^{s\tau_1}-1) \Rightarrow \tau_1 \approx \frac{1}{s}\log\left(\frac{s}{\rho}\right) \quad (s \gg \rho)$$

$\gg \frac{1}{s}$ (i.e., must wait quite a while for first recombinant to occur)

$\Rightarrow$ similarly, time to k^{th} successful recombinant (τ_k) occurs when $\int_0^{\tau_k}\rho e^{st}dt \sim k$ →

$$\tau_k = \frac{1}{s}\log\left(\frac{sk}{\rho}\right) \approx \tau_1 + \frac{1}{s}\log(k)$$

Since recombinant lineages grow as $f_i(t) = \frac{1}{Ns} e^{(s-\rho)(t-\tau_k)}$

we have: $f_0(t) = \frac{1}{Ns} e^{(s-\rho)t}$, $f_k(t) = \frac{1}{Ns} e^{(s-r)t} \left(\frac{\rho}{sk}\right)^{1-\frac{\rho}{s}} = f_0(t) \left(\frac{\rho}{sk}\right)^{1-\frac{\rho}{s}}$

extra tiny bit will be important in a bit.

$\Rightarrow$ recombinant lineages are all smaller than f_0 by a factor of $\rho/s \ll 1$.

(hint that linkage to $f_0(t)$ is important for dynamics)

How many recombinant lineages do we have to keep track of?

(since $f_k \propto 1/k$, $\sum_{k=0}^{\infty} f_k$ diverges!)

$\Rightarrow$ once selected mutations become common, many recombination events will occur between 2 selected haplotypes. $\Rightarrow$ doesn't produce recombinant lineage (from perspective of coalescence during sweep)

$\Rightarrow$ rate of recombinants becomes $\rho f_{sel}(t)\left[1-f_{sel}(t)\right]$

$\Rightarrow$ since $f_{sel}(t)$ grows logistically, know that this rate rapidly drops to zero w/ in $\mathcal{O}(\frac{1}{s})$ generations of $f_{sel}(t) \approx \frac{1}{2}$

$\Rightarrow$ $T_{max} \approx \frac{1}{s}\log(Ns) + \mathcal{O}\left(\frac{1}{s}\right) \approx \frac{1}{s}\log(Ns)$

$\Rightarrow$ last recombinants occur when $\tau_{K_{max}} \approx \frac{1}{s}\log\left(\frac{sK_{max}}{\rho}\right) \simeq \frac{1}{s}\log(Ns)$

$\Rightarrow K_{max} \approx N\rho$

Can now see that there are 2 regimes depending on value of $N\rho$:

(1) If $N\rho \ll 1 \Rightarrow K_{max} \ll 1 \Rightarrow$ typically no recombinants before mutation sweeps

$\Rightarrow$ "effectively asexual"

(2) If $N\rho \gg 1 \Rightarrow K_{max} \gg 1 \Rightarrow$ many recombinant lineages contribute @ end of sweep.

$\Rightarrow$ total size is

$$\sum_{k=0}^{K_{max}} f_k(t) = f_0(t)\left[1 + \sum_{k=1}^{K_{max}} \left(\frac{\rho}{sk}\right)^{1-\frac{\rho}{s}}\right]$$

$$\approx f_0(t)\left[1 + \int_1^{K_{max}} \left(\frac{\rho k}{s}\right)^{-1+\frac{\rho}{s}} dk\right]$$

$$\simeq f_0(t)\exp\left[+\frac{\rho}{s}\log(Ns)\right]$$

$\Rightarrow f_0(\infty) = e^{-\frac{\rho}{s}\log(Ns)}$, $f_k(\infty) = \left(\frac{\rho}{sk}\right)^{1-\frac{\rho}{s}} e^{-\frac{\rho}{s}\log(Ns)}$

Probability that 2 individuals share the same sweep lineage:

$$p_c = \sum_{k=0}^{K_{max}} f_k(\infty)^2 = e^{-\frac{2e}{s}\log(Ns)}\left[1+\sum_{k=1}^{K_{max}}\left(\frac{e}{sk}\right)^{2(1-\frac{e}{s})}\right] \simeq e^{-\frac{2e}{s}\log(Ns)}$$

$\to \mathcal{O}\left(\frac{e}{s}\right) \ll 1$

$\to$ dominated by probability of descending from initial mutant lineage.

similarly, probability that n individuals share the same lineage:

$$p_c(n) = \sum_{k=0}^{K_{max}} f_k(\infty)^n = e^{-\frac{ne}{s}\log(Ns)}\left[1+\sum_{k=1}^{K_{max}}\left(\frac{e}{sk}\right)^{n(1-\frac{e}{s})}\right] \simeq e^{-\frac{ne}{s}\log(Ns)}$$

$\Rightarrow$ can use this to calculate pairwise coalescence time:
(conditioned on sweep just fixing now)

$$\langle T_2\rangle = \underbrace{\frac{2}{s}\log(Ns)\,p_c(2)}_{\text{coalesce during sweep}} + \underbrace{2N\left(1-p_c(2)\right)}_{\text{coalesce afterwards}} \approx \underbrace{2N\left(1-e^{-\frac{2e}{s}\log(Ns)}\right)}_{\text{since } \frac{2}{s}\log(Ns) \ll N,\ Ne \gg 1}$$

if sweep hasn't just fixed, also need to integrate over establishment time $T_{est} \sim \text{Exp}\left(\frac{1}{2N\mu s}\right)$

2 regimes as usual: ① if $T_{est} \gg N$ $(N_\mu N s \ll 1)$

$\Rightarrow$ neutral coalescence before sweep, $T_2 \sim N$

② if $T_{est} \ll N$ $(N_\mu N s \gg 1)$

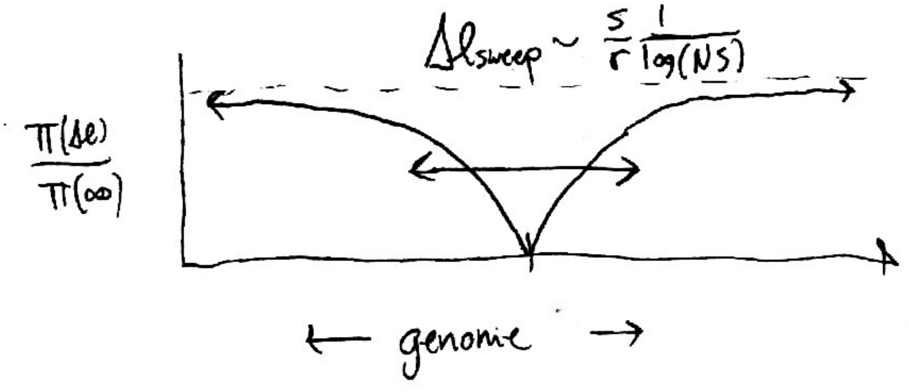
$\Rightarrow$ little chance of coalescence until sweep happens

$\Rightarrow$ $\langle T_2 \rangle$ same as before.

$\Rightarrow$ leads to reduction in pairwise heterozygosity (π) near sweep.

e.g. for neutral site distance $\Delta\ell$ from sweep, $\rho = r\Delta\ell$

$$\frac{\pi(\Delta\ell)}{\underbrace{\pi(\infty)}_{\text{neutral}}} = \left(1 - e^{-\frac{2r\Delta\ell}{s}\log(Ns)}\right)$$

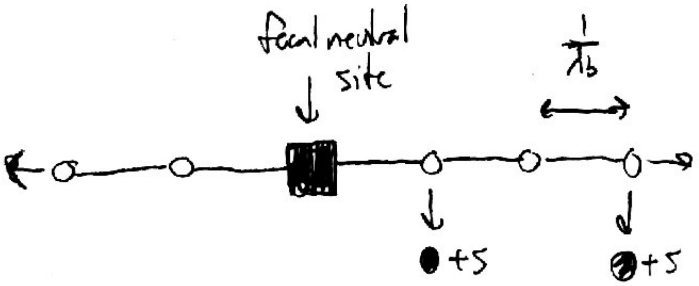


this is a common signal that people scan for in genomes from present day.

"Recurrent sweeps": so far we have focused on single selected site. can extend this picture to lots of selected sites, provided that they act like SSWM sweeps like this one.

$\Rightarrow$ again, can analyze w/ self consistency argument:

Focus on neutral site surrounded by selected sites @ density λ_b:



$\hookrightarrow$ fraction of sites that are strongly beneficial.

$\Rightarrow$ per generation probability of generating a sweep that leads to pairwise coalescence:

$$p_c = \underbrace{\int_0^\infty e^{-\frac{2r\Delta\ell}{s}\log(Ns)} \, 2N\mu\lambda_b \cdot s \, d\Delta\ell}_{} \approx \frac{N\mu\lambda_b s^2}{r\log(Ns)}$$

$\hookrightarrow$ dominated by probability of having a really close sweep w/ $p_c(z) \approx \mathcal{O}(1)$.

$$\Delta\ell \lesssim \ell^* = \frac{s}{r}\frac{1}{\log(NS)}$$

$\Rightarrow$ again, if $p_c \gg \frac{1}{N}$ $\Rightarrow$ $\langle T_2 \rangle \approx \frac{r\log(Ns)}{N\mu\lambda_b s^2} \approx \frac{1}{N U_{eff} s}$

"linkage block, $\ell^* \approx \frac{s}{r}\frac{1}{\log(NS)}$" $\leftarrow$ like asexual case w/ $U_{b,eff} = \mu\lambda_b\ell^*$

Note, however, that coalescent process looks very different from neutral coalescence w/ $N_e \approx \frac{1}{N U_{eff} s}$ (or any $N_e(t)$)

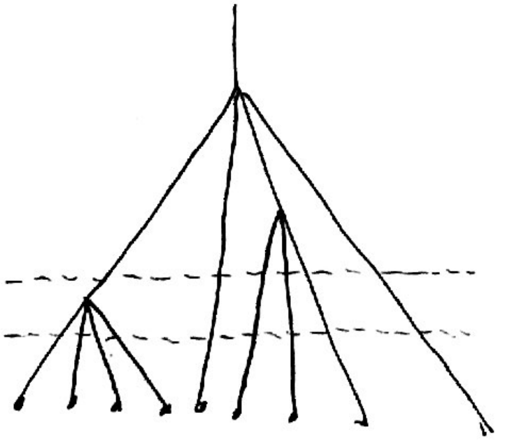
↳ common misconception in pop-gen. world @ moment

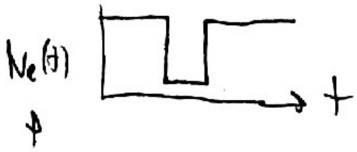
$\Rightarrow$ per generation probability of generating a sweep that leads to coalescence of all n lineages:

$$P_c(n) = \int_0^\infty e^{-\frac{n r \Delta\ell}{s}\log(Ns)}\, 2N\mu\lambda_b s \, d\Delta\ell \approx \frac{P_c(2)}{n} \rightarrow \text{decays very slowly in } n.$$

[compare to $P_c(n) = P_c(2)^n$ for neutral coalescent]

$\Rightarrow$ this means that conditioned on 2 lineages coalescing, very likely to have several coalescing @ once:





} even for transient bottleneck, rare for all k lineages to coalsce into 1, rather than $\binom{k}{2}$ disparate pairs.

Finally, when is this picture of recurrent non-Interfering sweeps a good approximation? ⇒ need to check self-consistency.

① 2 sweeps cannot occur w/in ℓ^* of each other w/in a single fixation time, $\frac{1}{s}\log(Ns)$: (i.e, SSWM w/in ℓ^*)

$$\Rightarrow N\mu\lambda_b\ell^*\cdot s\cdot\frac{1}{s}\log(Ns) \ll 1 \Rightarrow \frac{\mu}{r}\cdot\lambda_b\cdot Ns \ll 1$$

② when 2 sweeps do occur in same sweep time, should have $s \ll r\hat{\ell}$. (unliked even before establishment time.)

$$\Rightarrow N\mu\lambda_b\hat{\ell}\cdot s\cdot\frac{1}{s}\log(Ns) \sim 1 \Rightarrow \hat{\ell} = \frac{1}{N\mu\lambda_b\log(Ns)}$$

$$\Rightarrow s \ll \frac{r}{N\mu\lambda_b\log(Ns)} \Rightarrow \boxed{\frac{\mu}{r}\cdot\lambda_b\cdot Ns\log(Ns) \ll 1}$$

↳ slightly more stringent.

⇒ when this condition is <u>not</u> met, multiple beneficial mutations will interfere w/ each other. ⇒ <u>generic</u> for large enough N!