Chapter 14

Genetic hitchhiking from classic selective sweeps

Last time: **Linkage equilibrium approx ("independent sites")**

$$\frac{\partial \bar{s}(\vec{g})}{\partial t} = \sim(x-\bar{x}) + \sim L\cdot\mu + \sim \epsilon + \sim \frac{z}{\sqrt{N}}$$

$\epsilon \to \infty$ ⟶ $\prod_{\ell=1}^{L}$

$$\frac{\partial f_\ell}{\partial t} = \sim s_\ell + \sim \mu_\ell + \sim \frac{z_\ell}{\sqrt{N}}$$

Selection on genotypes

Selection on alleles

(~ the "ideal gas" of evolutionary dynamics)

⇒ A victory for reductionism?

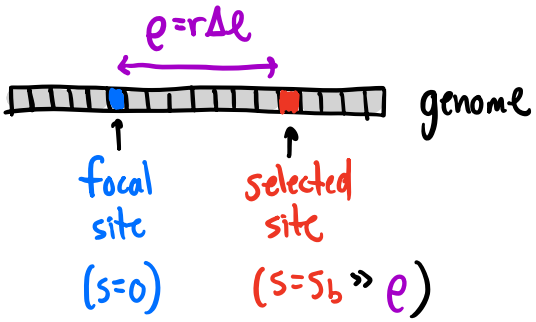
⇒ Requires $\rho_{eff} = r\Delta\ell \gg \frac{1}{N}, s$ ⇒ $\frac{r}{\mu} \gg 1, Ns$

⇒ empirically, $\frac{r}{\mu} \sim \mathcal{O}(1)$ ⇒ breaks down for strong beneficial mut'ns!

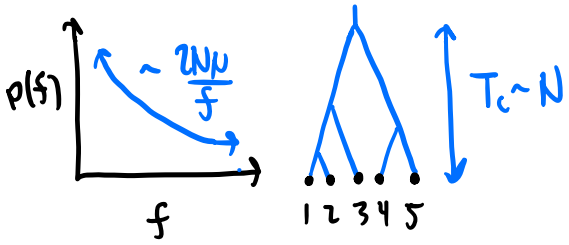
Today: what happens when this condition breaks down?

"genetic hitchhiking"

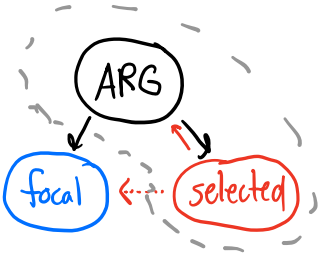
Consider simplest scenario:



⇒ when $\varrho \to \infty$ focal site looks like neutral model:

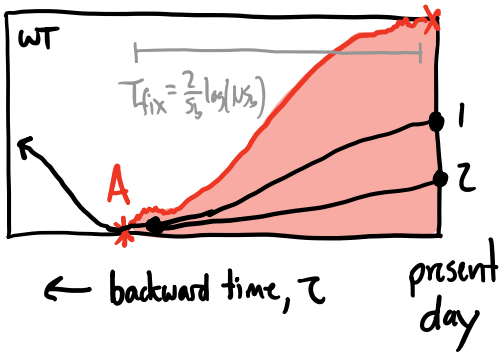


⇒ these patterns can change when $\varrho < \infty$ ...

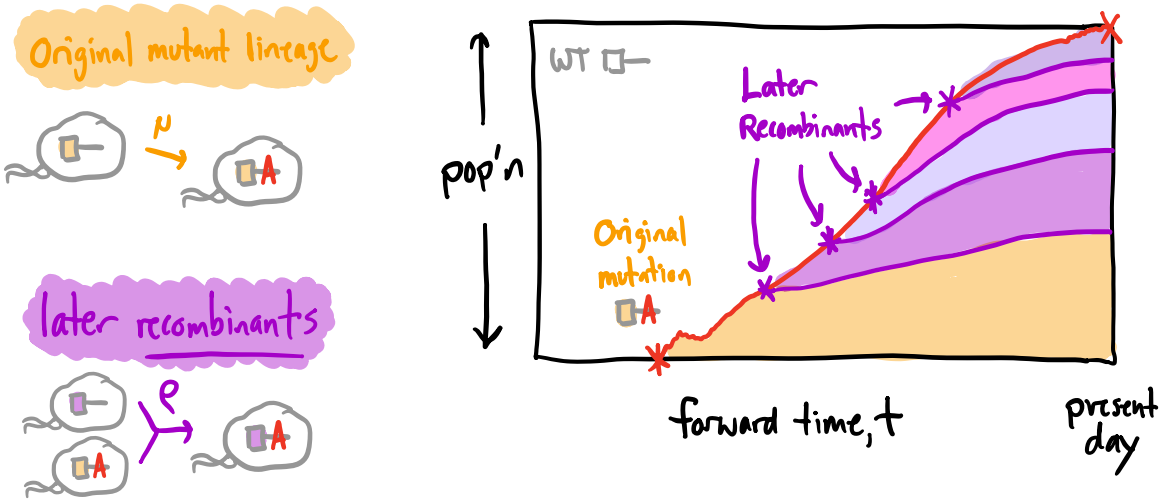


"linked selection" -or- "genetic hitchhiking"

⇒ behavior @ selected site is easy:

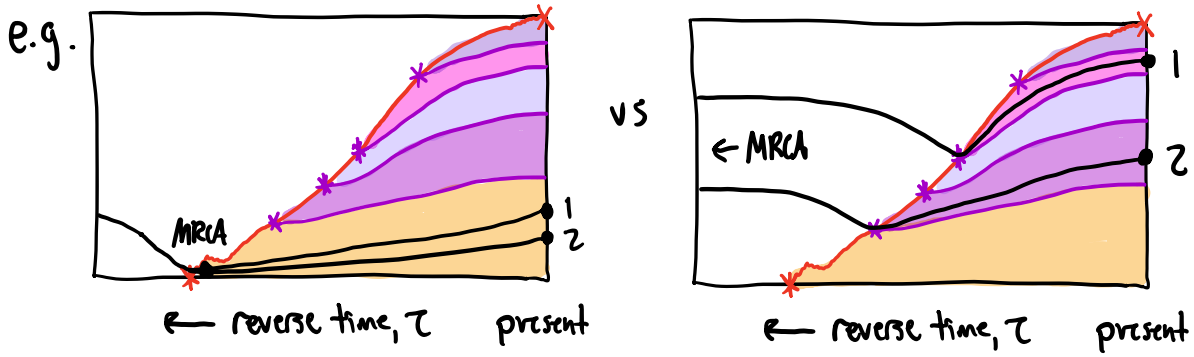


⇒ @ linked neutral site, must now distinguish between:



Why is this important?

⇒ individuals only coalesce during sweep if drawn from same lineage!

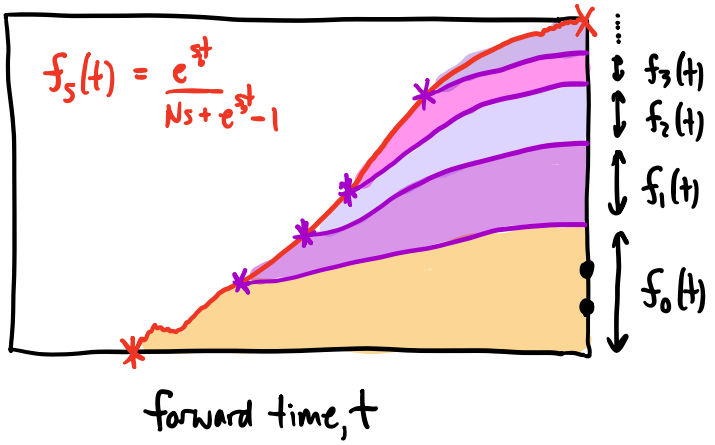


⇒ otherwise, trace back to different pre-sweep ancestors

⇒ neutral coalescence ($T_{MRCA} \sim N \gg T_{fix}$)

⇒ Total probability that 2 individuals coalesce during sweep:

$$p_c \equiv \sum_{k=0}^{k_{max}} f_k(t)^2$$



where $f_k(t)$ = size of k^{th} recombinant lineage.

⇒ How do we predict $f_k(t)$?

⇒ can learn a lot by focusing on short times

$\frac{1}{s_b} \ll t \ll T_{fix} = \frac{2}{s_b}\log(Ns_b)$ when A is still rare.

$f_s \sim \frac{1}{Ns_b} e^{s_b t} \ll 1$

⇒ recombinant lineages are founded @ total rate:

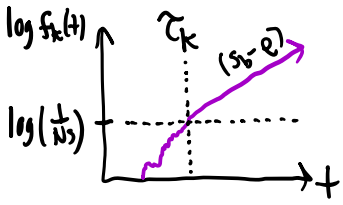
$$\Theta_r(t) \sim N\rho f_s(t) \cdot \underbrace{(1 - f_s(t))}_{\approx 1} \approx \frac{\rho}{s_b} e^{s_b t}$$

⇒ each recombinant lineage satisfies:

$$\frac{df_k}{dt} = s_b f_k - e f_k \cdot (1) + \sqrt{\frac{f_k}{N}}\,\eta_k(t)$$

- $s_b f_k$: selection
- $e f_k \cdot (1)$: outflow due to recomb. (w/ WT)
- $\sqrt{\frac{f_k}{N}}\,\eta_k(t)$: genetic drift.

⇒ we know how these behave:

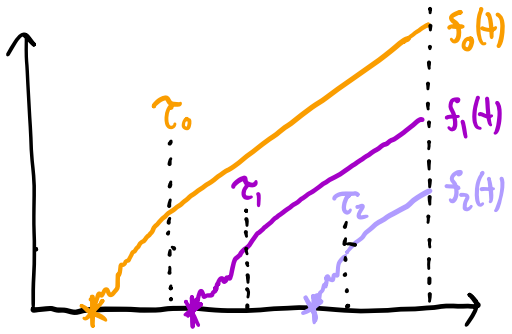


w/ probability $p_{est} \sim s_b - e \sim s_b$

lineage <u>establishes</u> & grows as $f_k(t) \sim \frac{1}{Ns} e^{(s_b - e)(t - \tau_k)}$

where τ_k = establishment time of lineage k.

⇒ <u>key insight</u>: all recombinants grow @ same rate $(s_b - e)$
so <u>relative</u> sizes independent of time!



$$\frac{f_k(t)}{f_0(t)} = e^{-(s_b - e)(\tau_k - \tau_0)}$$

⇒ holds even for $t \gtrsim T_{fix}$!

⇒ crucial step: How do we find τ_k?

① By convention, set $\tau_0 \equiv 0$ (i.e. t = time since start of sweep)

② Successful recombinants are produced @ rate

$$\theta_{r,est}(t) = \theta_r(t) \times p_{est} \sim \frac{\rho}{s_b} e^{s_b t} \cdot s_b \sim \rho e^{s_b t}$$

$\theta_r(t)$: total # recombinants produced @ gen t

p_{est}: prob. that each survives drift

⇒ avg # of successful recombinants by time t:

$$\langle k \rangle = \int_0^t \theta_{r,est}(t')\,dt'$$

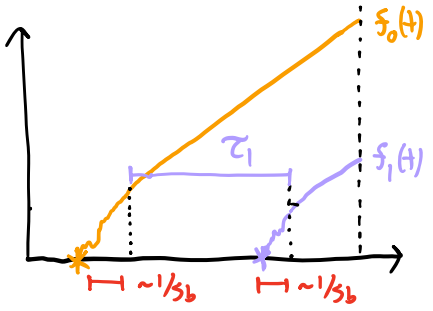
③ Heuristically, time to first successful recomb occurs when:

$$\int_0^{\tau_1} \theta_{r,est}(t')\,dt' \sim \mathcal{O}(1)$$

$$\Rightarrow 1 \sim \int_0^{\tau_1} \rho e^{s_b t'}\,dt' = \frac{\rho}{s_b}\left(e^{s_b \tau_1} - 1\right) \Rightarrow \tau_1 \approx \frac{1}{s}\log\left(\frac{s_b}{\rho} + 1\right)$$

$$\approx \frac{1}{s}\log\left(\frac{s_b}{\rho}\right) \quad [\text{when } s_b \gg \rho]$$

Note: $\tau_1 \gg \frac{1}{s_b}$ when $s_b \gg e$:



(4) Similarly, k^{th} successful recombinant typically occurs when:

$$\int_0^{\tau_k} \theta_{r,est}(t)\,dt \sim k \quad \Rightarrow \quad \tau_k \approx \frac{1}{s_b}\log\left(\frac{s_b k}{e}\right)$$

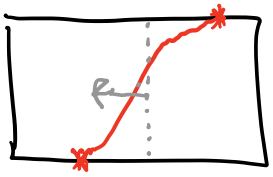
$$\Rightarrow \quad \frac{f_k(t)}{f_0(t)} = e^{(s_b - e)(\tau_0 - \tau_k)} = e^{-(s_b-e)\frac{1}{s_b}\cdot\log\left(\frac{s_b k}{e}\right)} = \left(\frac{e}{s_b k}\right)^{1-e/s_b}$$

$\Rightarrow$ @ end of sweep, size of k^{th} lineage is given by

$$f_k(\infty) = \frac{f_k(t)}{f_0(t) + \sum_{j=1}^{K_{max}} f_j(t)} = \frac{f_k(t)/f_0(t)}{1 + \sum_{j=1}^{K_{max}} f_j(t)/f_0(t)}$$

$$\Rightarrow \quad f_k(\infty) = \begin{cases} \left[1 + \sum_{j=1}^{K_{max}} \left(\frac{\rho}{s_b j}\right)^{1-\rho/s_b}\right]^{-1} & \text{if } k=0 \\ f_0(\infty)\left(\frac{\rho}{s_b k}\right)^{1-\rho/s_b} & \text{if } k \geq 1 \end{cases}$$

$\Rightarrow$ what is K_{max}?



$$f_s = \frac{e^{s_b t}}{N s_b + e^{s_b t} - 1}$$

$\Rightarrow$ total # of successful recombinants:

$$\leq \int_0^\infty N\rho\, f_s(t)(1-f_s(t)) \cdot s_b\, dt \sim N\rho$$

$\Rightarrow$ decent approx to take $K_{max} \sim N\rho$

$\Rightarrow$ Two regimes:

(1) $N\rho \ll 1 \Rightarrow K_{max} \ll 1 \Rightarrow$ typically no recombinants
$\Rightarrow$ like asexual case

(2) $N\rho \gg 1 \Rightarrow$ many recombinants contribute!

$$\Rightarrow \frac{1}{f_0(\infty)} = 1 + \sum_{j=1}^{K_{max}} \left(\frac{e}{s_b j}\right)^{1-e/s_b} \simeq 1 + \int_1^{Ne} \left(\frac{e}{s_b j}\right)^{1-e/s_b} dj$$

$$= 1 + \left(\frac{e}{s_b}\right)^{1-e/s_b} \frac{s_b}{e} \left(j\right)^{e/s_b} \Big|_1^{Ne} = 1 + \frac{e}{s_b}^{-e/s_b} \left[Ne^{e/s_b} - 1\right]$$

$$\approx \exp\left[+\frac{e}{s_b} \log\left(Ne \cdot \frac{s_b}{e}\right)\right] = \exp\left[+\frac{e}{s_b} \log(N s_b)\right]$$

Finally, probability that 2 individuals coalesce during sweep:

$$P_c = \sum_{k=0}^{K_{max}} f_k(\infty)^2 = f_0(\infty)^2 \left[1 + \sum_{k=1}^{K_{max}} \left(\frac{e}{s_b k}\right)^{2(1-e/s_b)}\right]$$

$2(1-e/s_b) \to 0$

→ dominated by initial mutant lineage!

$$= \exp\left(-\frac{2e}{s_b} \log(N s_b)\right)$$

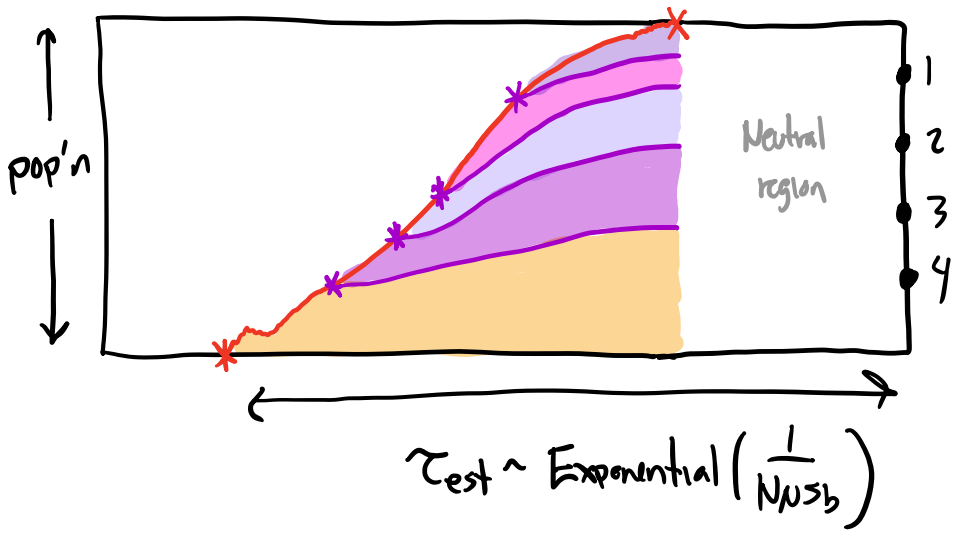
$$\Rightarrow \langle T_{MRCA} \rangle = T_{fix} \cdot P_c + N \cdot (1 - P_c) \approx N(1 - P_c) \quad \text{(when } Ne \gg 1\text{)}$$

$$\approx N\left(1 - e^{-\frac{2e}{s_b} \log(N s_b)}\right) \simeq \begin{cases} N & \text{if } e^{T_{fix}} \gg 1 \\ \frac{2Ne}{s_b} \log(N s_b) & \text{if } e^{T_{fix}} \ll 1 \end{cases}$$

⇒ works for larger sample sizes:

$$p_c(n) = \sum_{k=0}^{K_{max}} f_k(\infty)^n \approx e^{-\frac{ne}{s_b}\log(Ns_b)}$$

⇒ what happens if sweep fixed earlier?



Two regimes: ↗ ("after"?)

① $N \ll \tau_{est}$ ⇒ neutral coalescence before sweep!

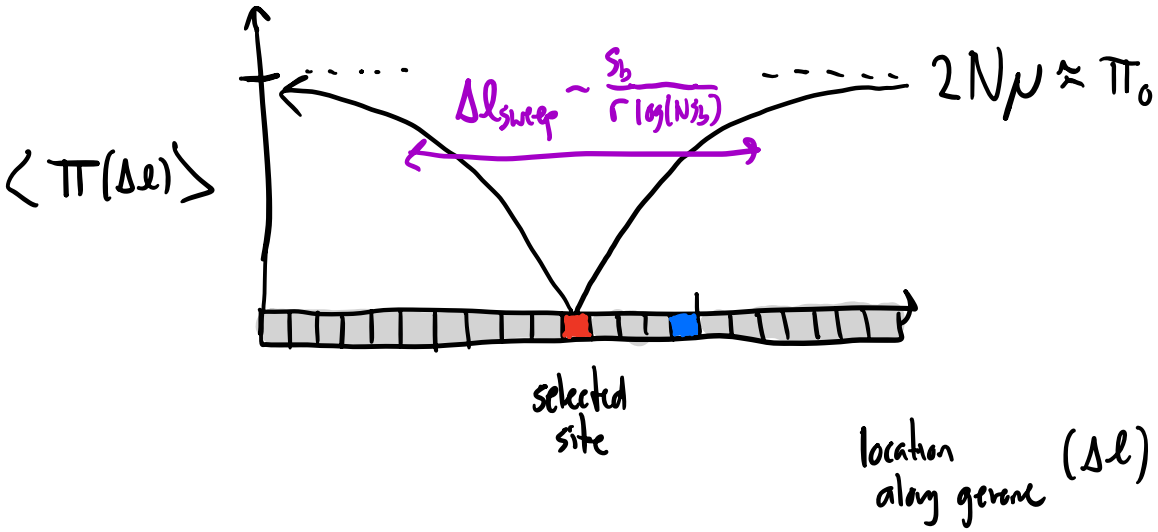
② $N \gg \tau_{est}$ ⇒ coalesce like before!

$$\langle T_{MRCA} \rangle \simeq N\left(1 - e^{-\frac{2\rho}{s_b}\log(Ns_b)}\right)$$

$\Rightarrow$ since $\rho = r\Delta\ell$:

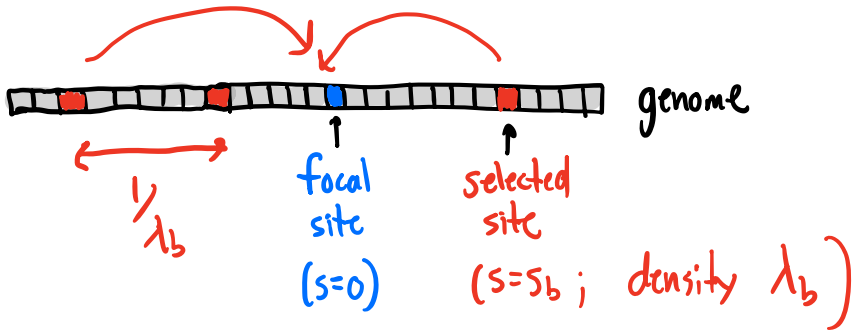
$$\langle T_{MRCA} \rangle \simeq N\left(1 - e^{-\Delta\ell \cdot \frac{2r}{s_b}\log(Ns_b)}\right)$$

$\Rightarrow$ can visualize as distance from selected site:



$\Rightarrow$ major signal that people try to look for in data!
("selection scans")

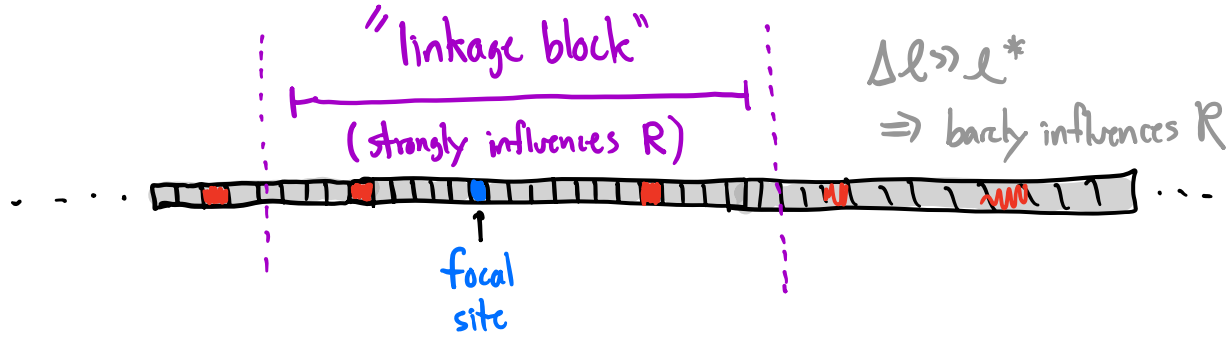
Recurrent sweeps: can extend to multiple selected sites as long as they don't interfere...



$\Rightarrow$ per generation rate of sweeps that lead to coalescence:

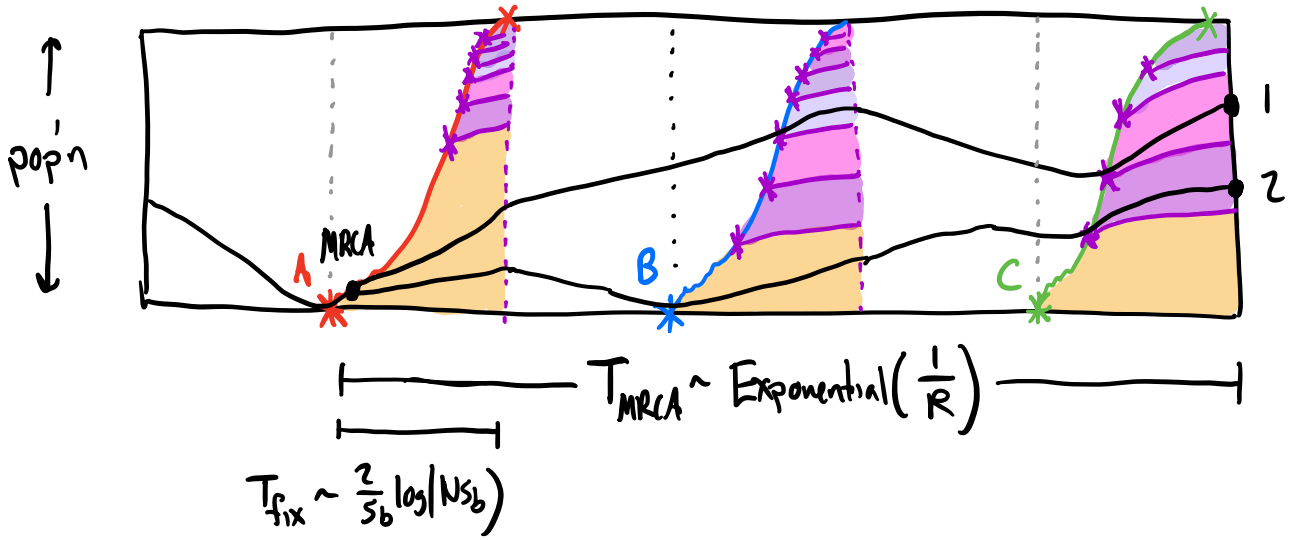
$$R = \int_0^\infty \underbrace{e^{-\frac{2r\Delta\ell}{s_b}\cdot\log(Ns_b)} \cdot 2N\mu\lambda_b \cdot s_b \cdot d\Delta\ell}_{\text{dominated by probability of really close sweep:}} = \frac{N\mu\lambda_b s_b^2}{r\log(Ns_b)}$$

$$\left[\Delta\ell \lesssim \ell^* \equiv \frac{s_b}{r\log(Ns_b)}\right] \quad \text{when} \quad p_c(2) \sim \mathcal{O}(1)$$



$\Rightarrow$ if time between sweeps $\left(\frac{1}{R}\right)$ is $\gg T_{fix}$ but $\ll N$

$$\Rightarrow \langle T_{MRCA} \rangle = \frac{1}{R} = \frac{r \log(Ns)}{N \mu \lambda_b s_b^2}$$



$\Rightarrow$ Using definition of ℓ^* can also write as:

$$\langle T_{MRCA} \rangle = \frac{1}{N \cdot \ell^* \mu \lambda_b \cdot s_b} = \frac{1}{N U_{b,eff} s_b}$$

$\Rightarrow$ looks like asexual case w/ $U_{b,eff} \equiv \ell^* \mu \lambda_b$

$\Rightarrow$ differences emerge in larger samples...