

Announcements: Office hours Today 12:30-2pm

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



$$\frac{\partial f(\vec{g})}{\partial t} = \sim(x-\bar{x}) + \sim L \times \mu + \sim e + \sim \frac{z}{\sqrt{N}}$$

$\xrightarrow{e \rightarrow \infty} \prod_{l=1}^L \frac{\partial f_l}{\partial t} = \sim S_l + \sim \mu_l + \sim \frac{z_l}{\sqrt{N}}$

Selection on genotypes

Selection on alleles

(\sim the "ideal gas" of evolutionary dynamics)

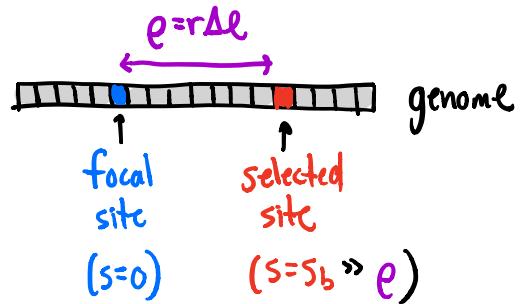
\Rightarrow A victory for reductionism?

$$\Rightarrow \text{Requires } \rho_{\text{eff}} = r \Delta l \gg \frac{1}{N}, S \Rightarrow \frac{r}{\mu} \gg 1, Ns$$

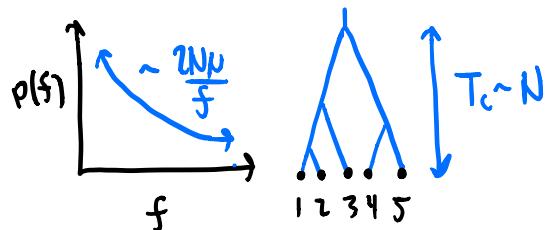
\Rightarrow empirically, $\frac{r}{\mu} \sim O(1) \Rightarrow$ breaks down for strong beneficial mut'n's!

Today: what happens when this condition breaks down?

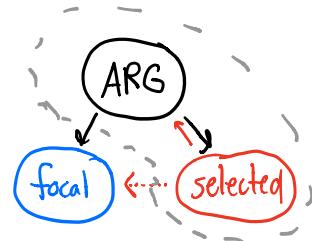
Consider simplest scenario:



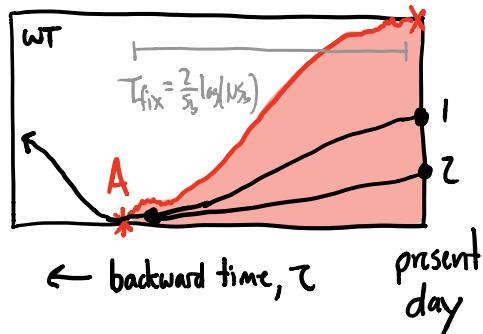
\Rightarrow when $e \rightarrow \infty$ focal site looks like neutral model:



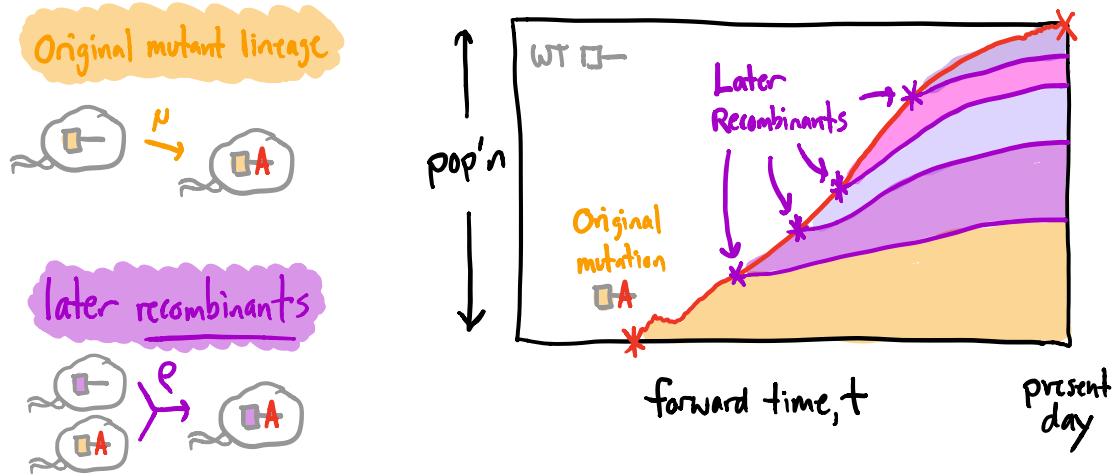
\Rightarrow these patterns can change when $e < \infty$ ("linked selection")



\Rightarrow behavior @ selected site is easy:

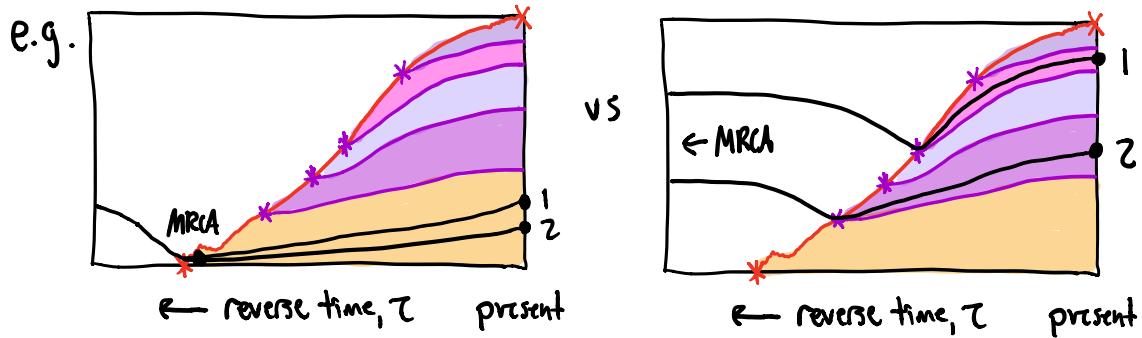


⇒ @ linked neutral site, must now distinguish between:



Why is this important?

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

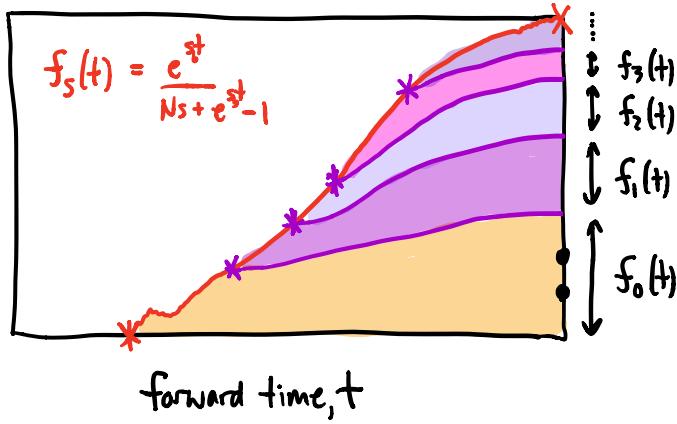


⇒ otherwise, trace back to different pre-sweep ancestors

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

\Rightarrow 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.

\Rightarrow How do we predict $f_k(t)$?

\Rightarrow can learn a lot by focusing on short times

$$\frac{1}{S_b} \ll T_{\text{fix}} = \frac{2}{S_b} \log(N_s) \quad \text{when } A \text{ is still rare.}$$

$$f_s \sim \frac{1}{N_s} e^{st} \ll 1$$

\Rightarrow recombinant lineages are founded @ total rate:

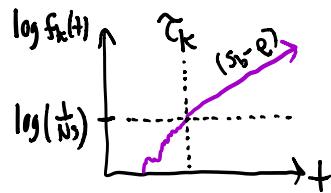
$$\Theta_r(t) \sim N_p f_s(t) \cdot \underbrace{(1-f_s(t))}_{\approx 1} = \frac{\rho}{S_b} e^{S_b t}$$

\Rightarrow each recombinant lineage satisfies:

$$\frac{df_k}{dt} = s_b f_k - \underbrace{\rho f_k(1)}_{\substack{\text{outflow} \\ \text{due to} \\ \text{recomb. } (\omega/\text{UT})}} + \underbrace{\sqrt{\frac{f_k}{N}} \eta_k(t)}_{\text{genetic drift.}}$$

\Rightarrow we know how these behave:

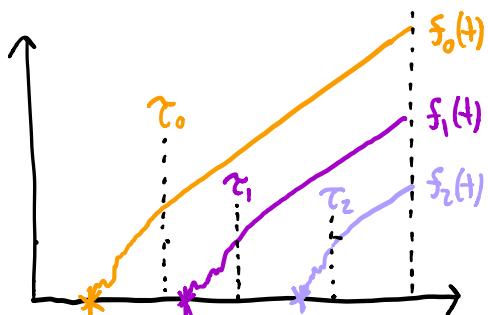
w/ probability $P_{\text{est}} \sim s_b - e \sim s_b$



lineage establishes + grows as $f_k(t) \sim \frac{1}{N_s} e^{(s_b - e)(t - \tau_k)}$

where τ_k = establishment time of lineage k.

\Rightarrow key insight: all recombinants grow @ same rate ($s_b - e$)
so relative sizes independent of time!



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

\Rightarrow holds even for $t \gtrsim T_{\text{fix}}$!

\Rightarrow crucial step: How do we find τ_k ?

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

② Successful recombinants are produced @ rate

$$\Theta_{r,\text{est}}(t) = \Theta_r(t) \times p_{\text{est}} \sim \frac{\ell}{s_b} e^{s_b t} \cdot s_b \sim \ell e^{s_b t}$$

total # recombinants produced @ gent
prob. that each survives drift

\Rightarrow avg # of successful recombinants by time t :

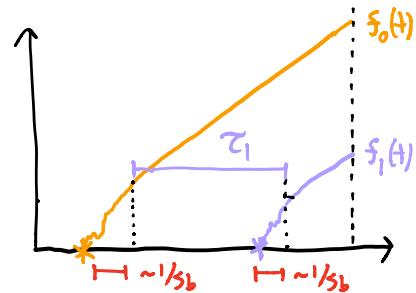
$$\langle k \rangle = \int_0^t \Theta_{r,\text{est}}(t') dt'$$

③ Heuristically, time to first successful recomb occurs when:

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

$$\Rightarrow 1 \sim \int_0^{\tau_1} \ell e^{s_b t'} dt' = \frac{\ell}{s_b} (e^{s_b \tau_1} - 1) \Rightarrow \tau_1 \approx \frac{1}{s_b} \log \left(\frac{s_b}{\ell} + 1 \right) \approx \frac{1}{s_b} \log \left(\frac{s_b}{\ell} \right) \quad [\text{when } s_b \gg \ell]$$

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



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

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

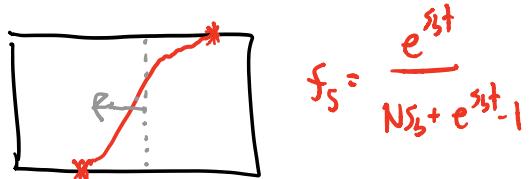
$$\Rightarrow \frac{f_k(t)}{f_0(t)} = e^{(s_b - \rho)(\tau_0 - \tau_k)} = e^{- (\rho - s_b) \frac{1}{s_b} \cdot \log\left(\frac{s_b k}{\rho}\right)} = \left(\frac{\rho}{s_b k}\right)^{1 - e^{s_b k}}$$

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

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

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

\Rightarrow what is K_{\max} ?



\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!

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

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

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

Finally, probability that 2 individuals coalesce during sweep:

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

dominated by initial mutant lineage!

$$= \exp \left(- \frac{2\epsilon}{s_b} \log (Ns_b) \right)$$

when $Ne \gg 1$

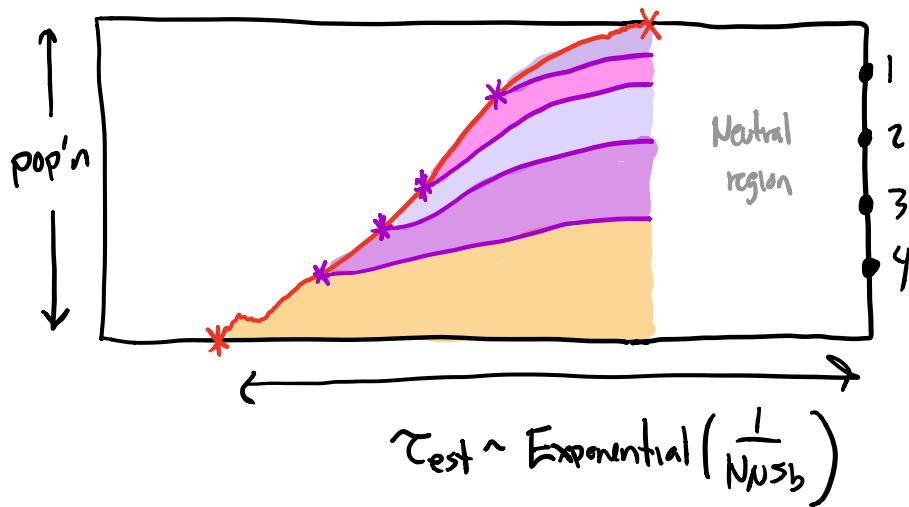
$$\Rightarrow \langle T_{\text{MRCA}} \rangle = T_{\text{fix}} \cdot p_c + N \cdot (1-p_c) \approx N(1-p_c)$$

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

\Rightarrow works for larger sample sizes:

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

\Rightarrow what happens if sweep fixed earlier?



Two regimes:

\rightarrow ("after"?)

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

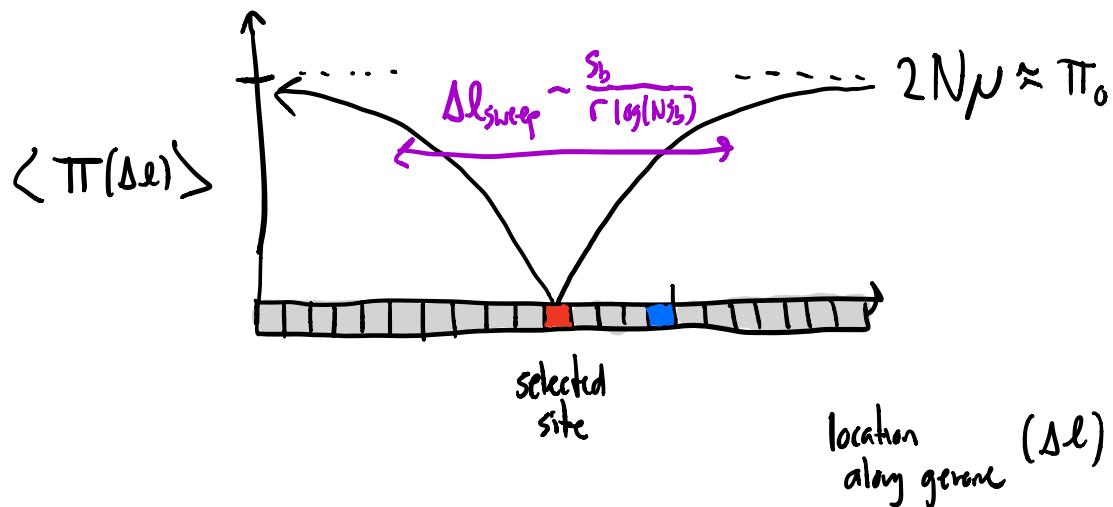
② $N \gg \tau_{est} \Rightarrow$ coalesce like before!

$$\leftarrow \langle T_{MRA} \rangle \approx N\left(1 - e^{-\frac{2\theta}{S_b} \log(Ns_b)}\right)$$

\Rightarrow since $\theta = r\Delta l$:

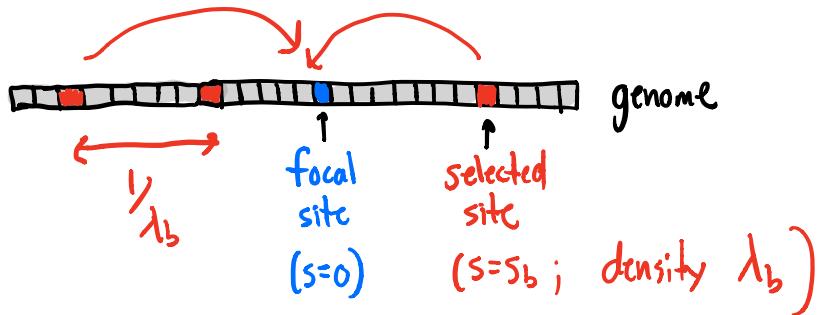
$$\langle T_{MRA} \rangle = N\left(1 - e^{-\Delta l \cdot \frac{2r}{S_b} \log(Ns_b)}\right)$$

\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...

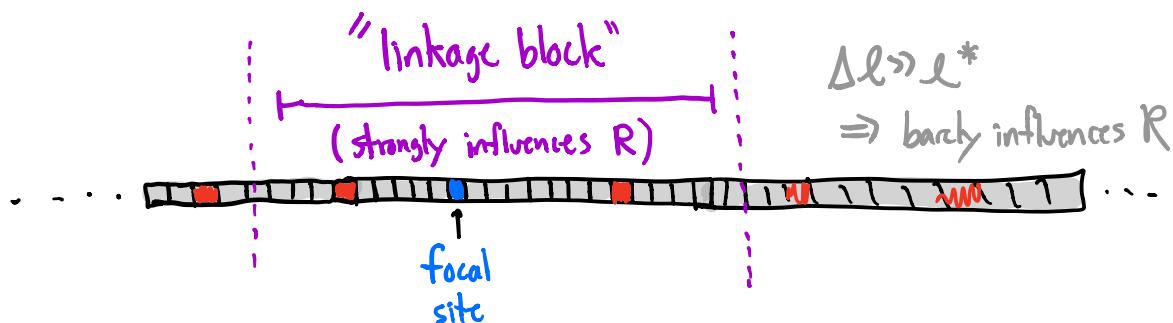


\Rightarrow per generation rate of sweeps that lead to coalescence:

$$R = \int_0^\infty e^{-\frac{2rsl}{s_b} \cdot \log(Ns_b)} \cdot 2N \lambda_b \cdot s_b \cdot dsl = \frac{N \lambda_b s_b^2}{r \log(Ns_b)}$$

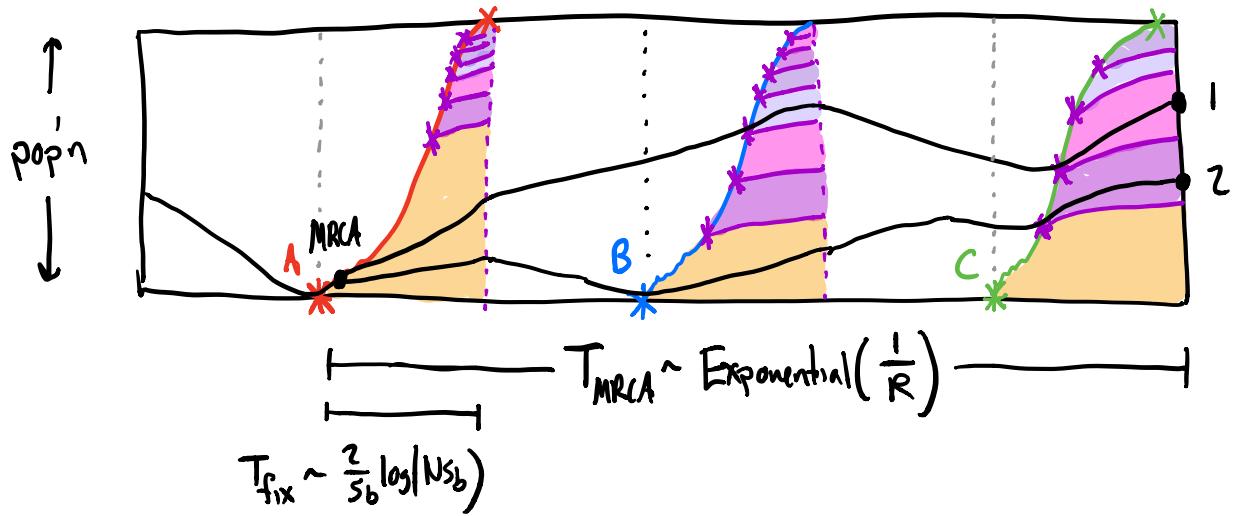
dominated by probability of really close sweep:

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



\Rightarrow if time between sweeps $(\frac{1}{R})$ 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...