Chapter 14

Genetic hitchhiking from classic selective sweeps

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

L-site genome $\xrightarrow[\text{recomb. is "fast"}]{}$ collection of L single locus problems

$$\frac{\partial f(\vec{g})}{\partial t} = \sim(x-\bar{x}) + \sim L\times\mu + \sim \rho + \sim\frac{Z}{\sqrt{N}} \quad \xrightarrow{\rho\to\infty} \quad \prod_{\ell=1}^{L} \boxed{\frac{\partial f_\ell}{\partial t} = \sim s_\ell + \sim \mu_\ell + \sim \frac{Z_\ell}{\sqrt{N}}}$$

Selection on genotypes $\qquad$ Selection on alleles

(~ the "ideal gas" of evolutionary dynamics)

⇒ A victory for reductionism?

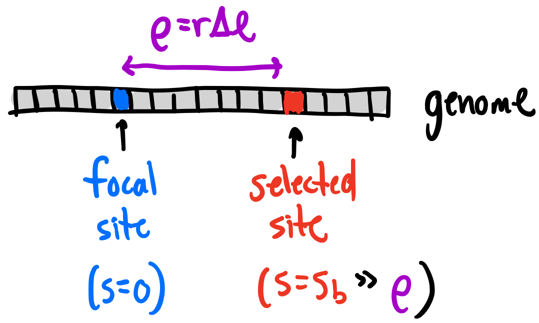
⇒ Requires $\rho_{eff} = r\Delta\ell \gg \frac{1}{N}, s$ ⇒ $\frac{r}{\mu} \gg 1, Ns$

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

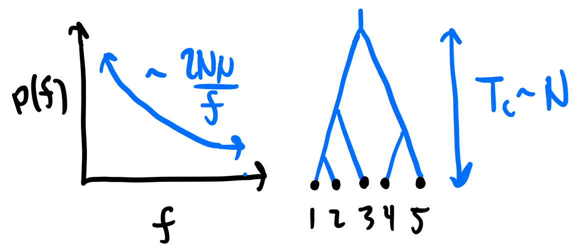
Today: what happens when this condition breaks down?

"genetic hitchhiking"

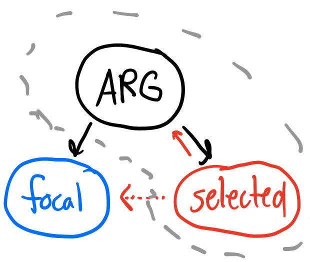
Consider simplest scenario:



$\Rightarrow$ when $\rho \to \infty$ focal site looks like neutral model:

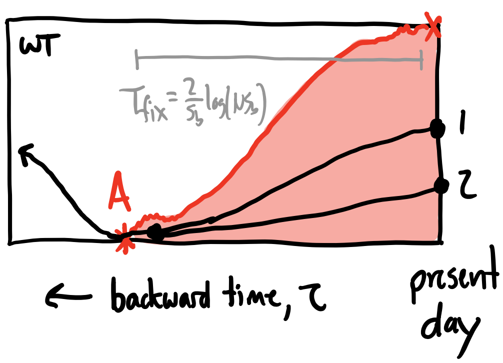


$\Rightarrow$ these patterns can change when $\rho < \infty$...

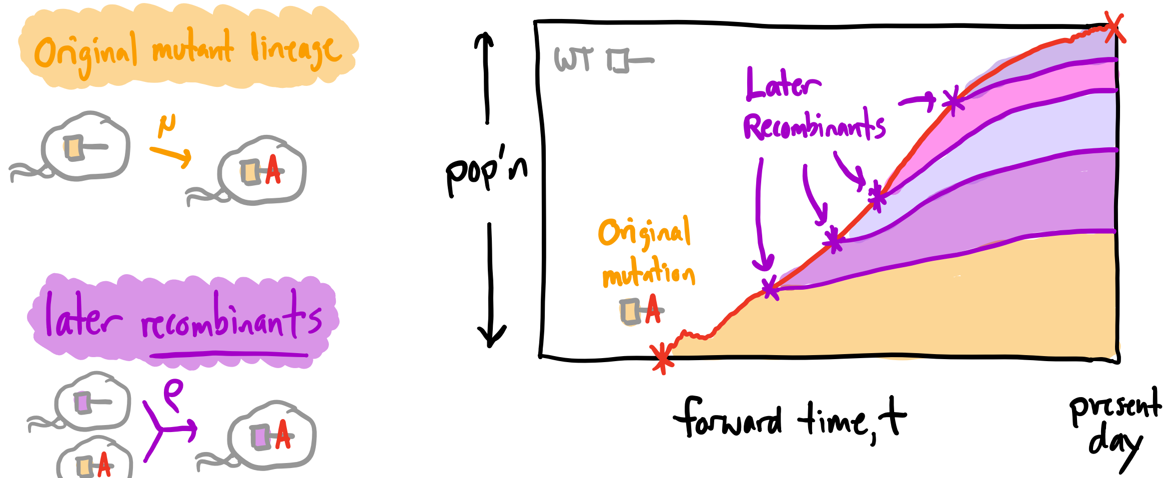


"linked selection" -or- "genetic hitchhiking"

$\Rightarrow$ behavior @ selected site is easy:



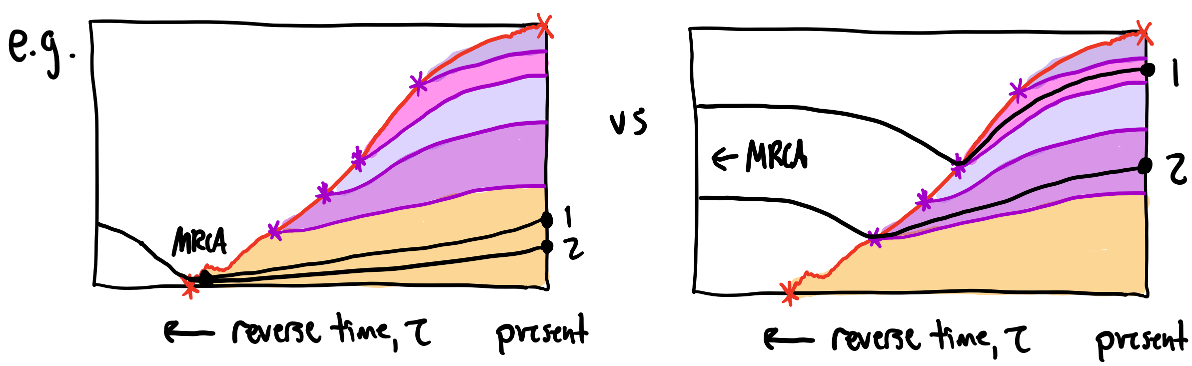
⇒ @ linked neutral site, must now distinguish between:



Why is this important?

⇒ individuals <u>only</u> coalesce during sweep if drawn from <u>same</u> <u>lineage</u>!

e.g.

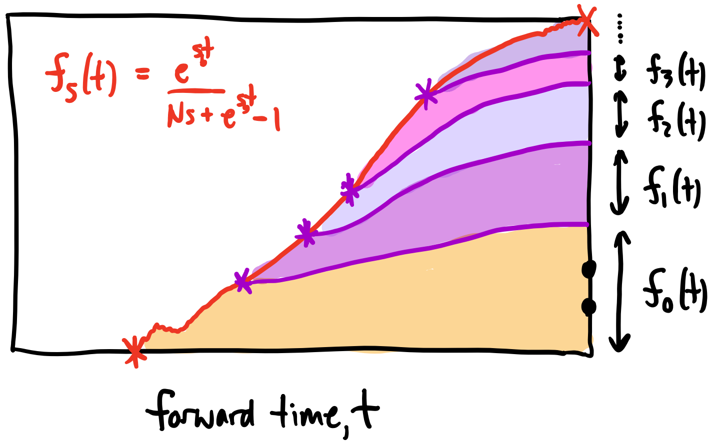


⇒ otherwise, trace back to <u>different</u> pre-sweep ancestors

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

$\Rightarrow$ 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

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

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

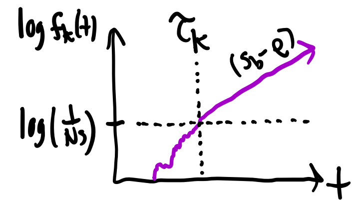
$\Rightarrow$ recombinant lineages are founded @ total rate:

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

$\Rightarrow$ each recombinant lineage satisfies:

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

$\Rightarrow$ we know how these behave:



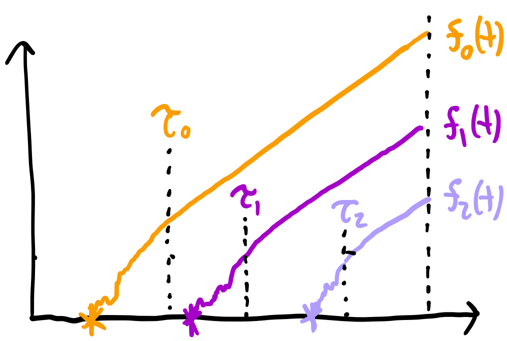
w/ probability $p_{est} \sim s_b - e \sim s_b$

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

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 $t \gtrsim T_{fix}$!

$\Rightarrow$ 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

$$\Theta_{r,est}(t) = \underbrace{\Theta_r(t)}_{\text{total \# recombinants produced @ gen } t} \times \underbrace{p_{est}}_{\text{prob. that each survives drift}} \sim \frac{\rho}{s_b} e^{s_b t} \cdot s_b \sim \rho e^{s_b t}$$

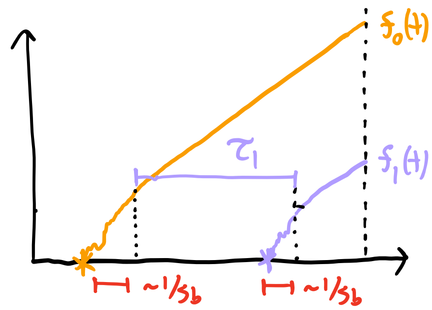
$\Rightarrow$ avg # of successful recombinants by time t: $\langle k \rangle = \int_0^t \Theta_{r,est}(t')\,dt'$

③ Heuristically, time to first successful recomb occurs when: $\int_0^{\tau_1} \Theta_{r,est}(t')\,dt' \sim \mathcal{O}(1)$

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

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

<u>Note</u>: $\tau_1 \gg \frac{1}{s_b}$ when $s_b \gg e$:



④ Similarly, <u>k^{th}</u> successful recombinant typically occurs when:

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

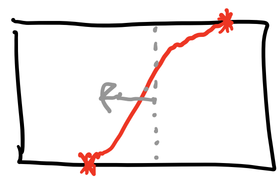
$$\Rightarrow \quad \frac{f_k(t)}{f_0(t)} = e^{(s_b - e)(\tau_0 - \tau_k)} = e^{-(s_b - e)\frac{1}{s_b} \cdot \log\left(\frac{s_b k}{e}\right)} = \left(\frac{e}{s_b k}\right)^{1 - e/s_b}$$

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

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

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

$\Rightarrow$ what is K_{max}?



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

$\Rightarrow$ total # of successful recombinants:

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

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

$\Rightarrow$ Two regimes:

① $Ne \ll 1 \Rightarrow K_{max} \ll 1 \Rightarrow$ typically no recombinants

$\Rightarrow$ like asexual case

② $Ne \gg 1 \Rightarrow$ many recombinants contribute!

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

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

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

Finally, probability that 2 individuals coalesce during sweep:

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

dominated by initial mutant lineage!

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

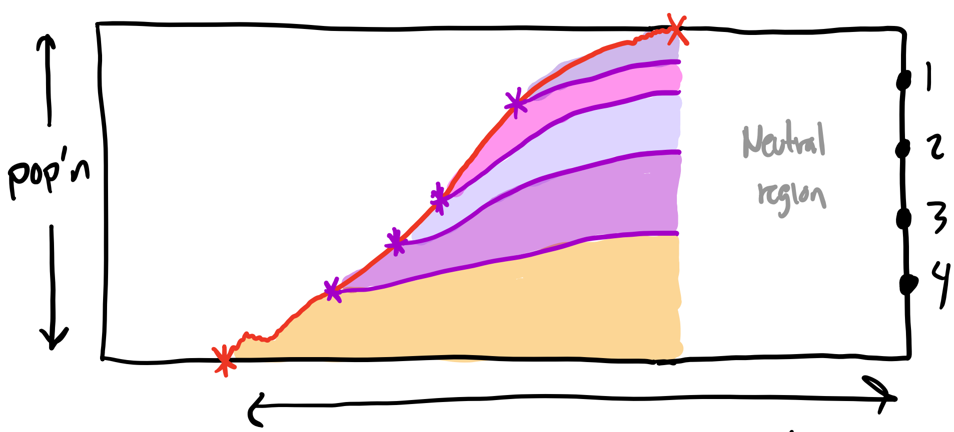
$$\Rightarrow \langle T_{MRCA} \rangle = T_{fix} \cdot p_c + N \cdot (1 - p_c) \overset{\text{when } Ne \gg 1}{\approx} N(1 - p_c)$$

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

$\Rightarrow$ works for larger sample sizes:

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

$\Rightarrow$ what happens if sweep fixed earlier?



$\tau_{est} \sim \text{Exponential}\left(\frac{1}{N\mu s_b}\right)$

Two regimes:

① $N \ll \tau_{est} \Rightarrow$ neutral coalescence before sweep! ("after"?)

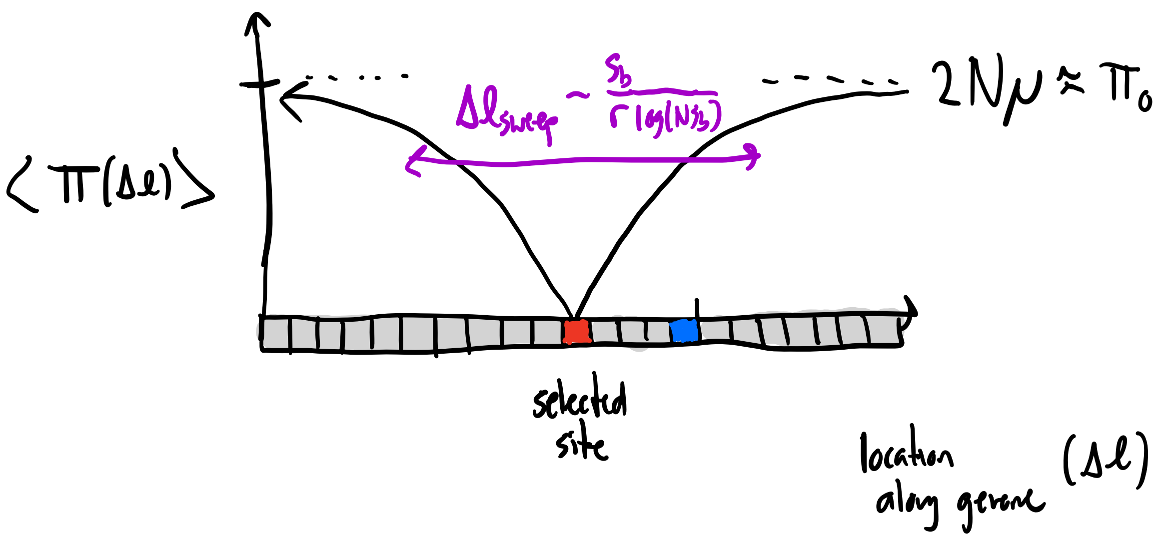
② $N \gg \tau_{est} \Rightarrow$ coalesce like before!

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

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

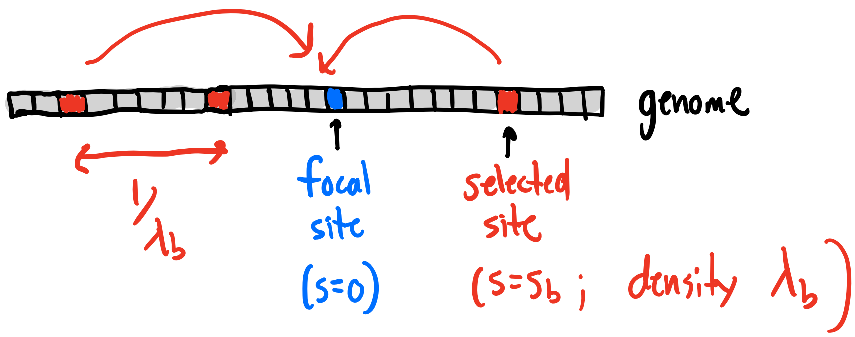
$$\langle T_{MRCA}\rangle \approx N\left(1-e^{-\Delta\ell\cdot\frac{2r}{s_b}\log(Ns_b)}\right)$$

$\Rightarrow$ 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...

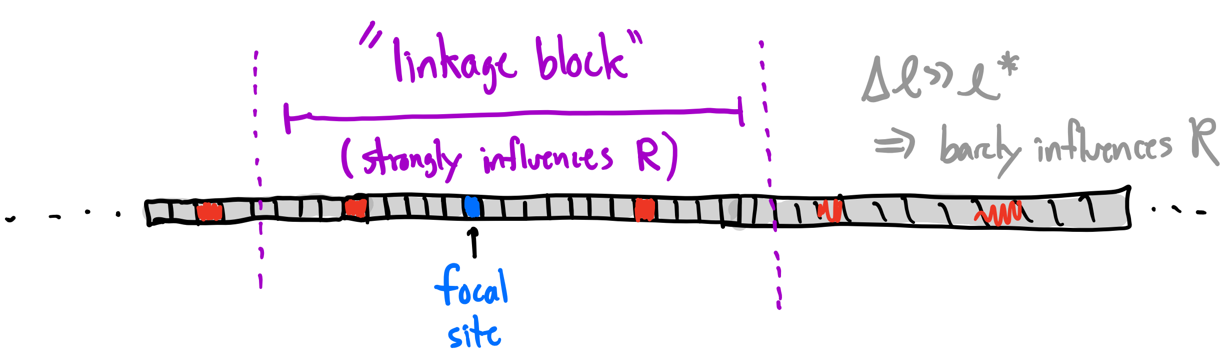


$\Rightarrow$ per generation rate of sweeps that lead to coalescence:

$$R = \int_0^\infty \underbrace{e^{-\frac{2r\Delta\ell}{s_b}\cdot\log(Ns_b)} \cdot 2N\mu\lambda_b\cdot s_b\cdot d\Delta\ell}_{} = \frac{N\mu\lambda_b s_b^2}{r\log(Ns_b)}$$

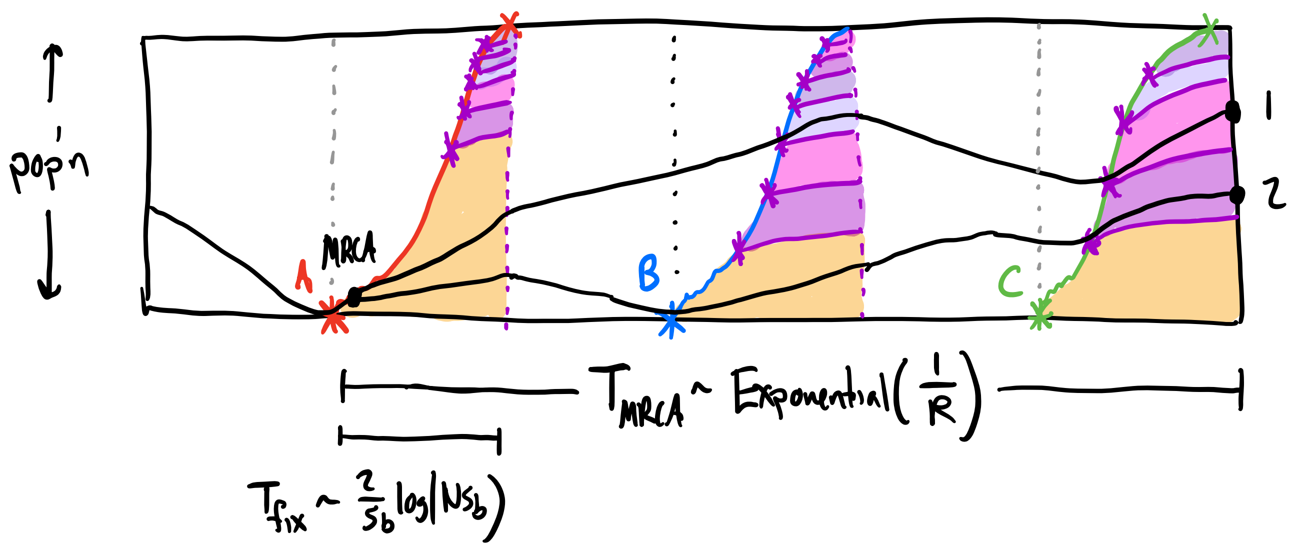
dominated by probability of really close sweep:

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



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

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



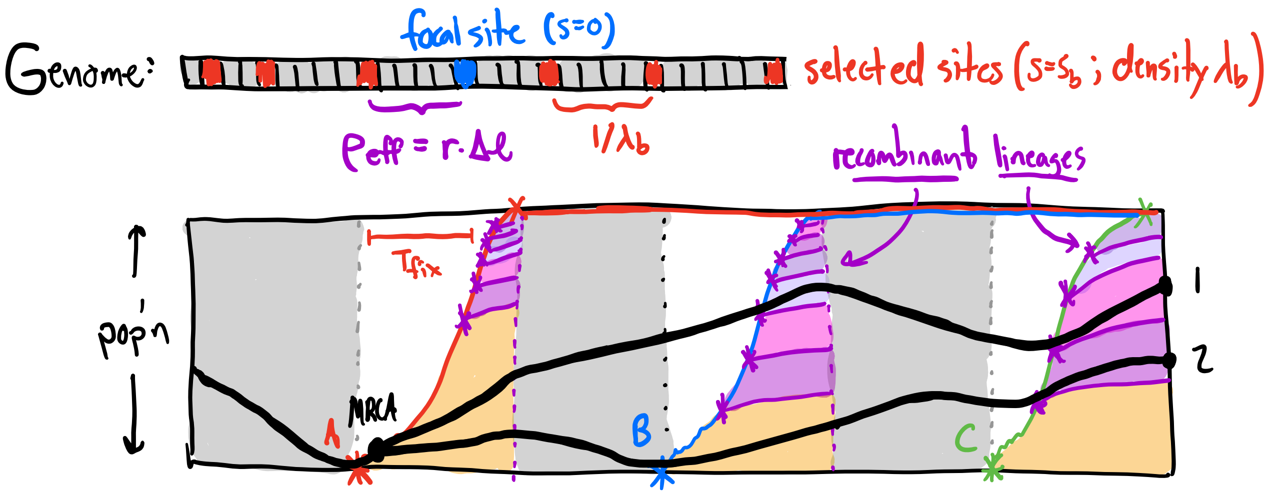
$\Rightarrow$ Using definition of ℓ^* can also write as:

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

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

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

Recap: Linked selection via "classic selective sweeps"



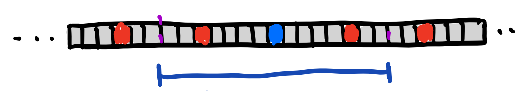
Coalescence Prob Per Sweep:

$$p_c(n,\Delta\ell) = \exp\left[-n\cdot\Delta\ell\cdot\frac{r}{s_b}\cdot\log(Ns_b)\right]$$

Total coalescence Rate from sweeps

$$R = \int_{-\infty}^{\infty} p_c(2,|\Delta\ell|)\cdot 2N\mu\lambda_b s_b\cdot d\Delta\ell$$

$$\approx \int_{-\ell^*/2}^{\ell^*/2} \mathcal{O}(1)\cdot 2N\mu\lambda s_b\cdot d\Delta\ell$$



$$\ell^* \sim s_b/r\log(Ns_b)$$

When $N \gg \frac{1}{R} \gg T_{fix}$:

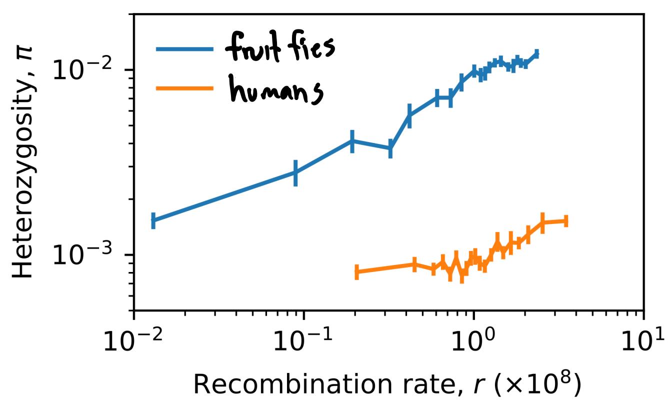
$$\langle T_{MRCA}\rangle \approx \frac{1}{R} = \frac{r\log(Ns_b)}{2N\mu\lambda_b s_b^2}$$

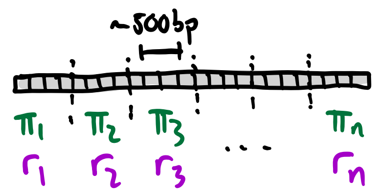
⇒ **Key prediction**: genetic diversity (e.g. π) @ neutral (e.g. syn) sites depends on local recombination rate r!

(since controls linkage to selected sites)

$$\langle \pi \rangle \approx \frac{r \log(N s_b)}{s_b \cdot N s_b \cdot \lambda_b}$$

⇒ can test using natural variation in r along genome:





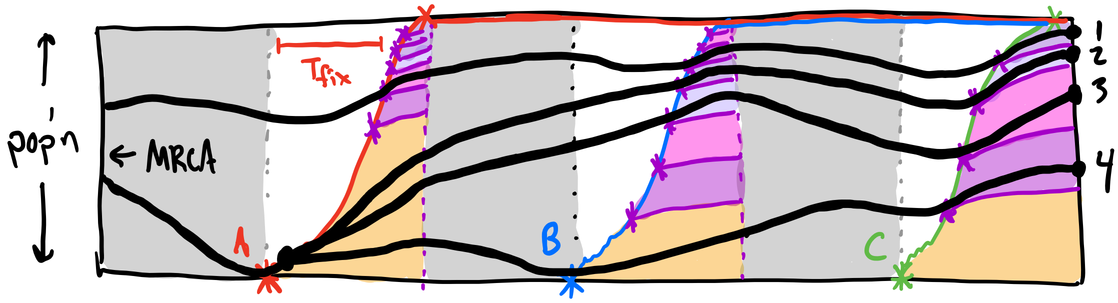
(e.g. hot + cold spots)

⇒ ~~sometimes spun as "linked selection = local reductions in Ne"~~

~~e.g. $\pi_\ell \approx 2N_e(\ell)\mu$~~

⇒ WRONG!

⇒ can see by examining larger sample sizes:



Recall: Coalescence Prob Per Sweep: $p_c(n,\Delta\ell) = \exp\left[-n\cdot\Delta\ell\cdot\frac{r}{s_b}\cdot\log(Ns_b)\right]$

⇒ Total rate of sweeps w/ n lineages coalescing:

$$R(n\to 1) = \int_0^\infty e^{-\frac{n\Delta\ell r}{s_b}\log(Ns_b)}\cdot 4N\nu\lambda_b s_b\, d\Delta\ell = \frac{4N\nu\lambda_b s_b}{\frac{nr}{s_b}\log(Ns_b)}$$

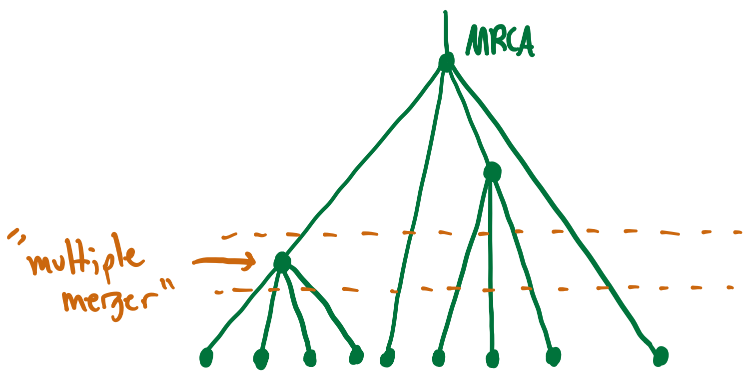
⇒ $\boxed{R(n\to 1) = \frac{2}{n}R}$ ⇒ Decays very slowly w/ n!

[compare to $N\cdot\left(\frac{1}{N}\right)^n$ for neutral (Kingman) coalescent]

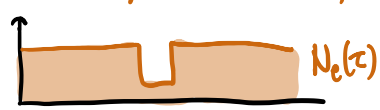
Upshot: if 2 lineages coalesce in a given timestep,

$\Rightarrow$ likely **multiple** lineages coalesce into same block!

$\Rightarrow$ can produce genealogies like:



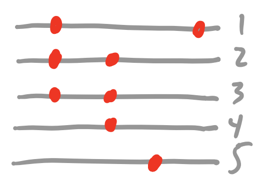
even w/ a bottleneck,



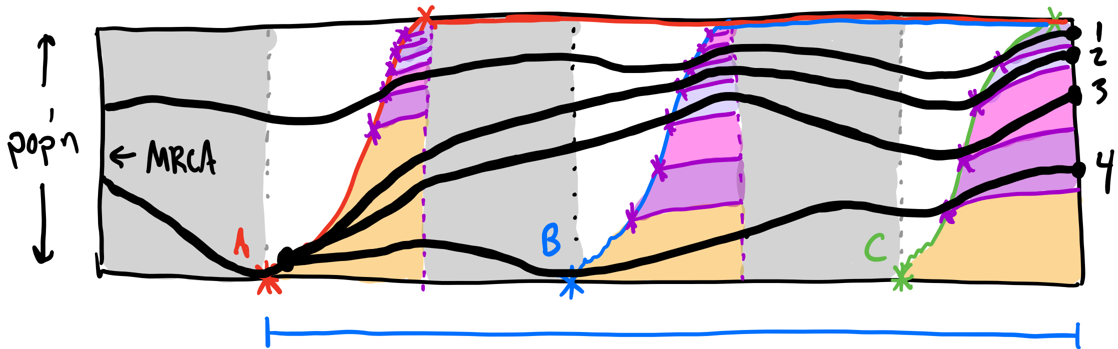
$k \to 1$ less likely than $\binom{k}{2}$ pairs!

$\Rightarrow$ **Important open question**:

How can we detect these effects from mutation data?



$\Rightarrow$ when is this successive mutations - like picture a good approx?



$T_{fix} \sim \frac{1}{s_b}\log(Ns_b)$

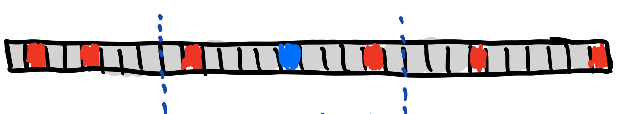
$1/R \sim \frac{r\log(Ns_b)}{N\mu\lambda_b s_b^2}$

$\Rightarrow$ check self consistency:

Each coalescence - causing sweep should fix before next one occurs

$$\Rightarrow RT_{fix} \ll 1 \Rightarrow \frac{N\mu\lambda_b s_b^2}{r\log(Ns_b)}\cdot\frac{1}{s_b}\cdot\log(Ns_b) = \frac{\mu}{r}\cdot\lambda_b\cdot Ns_b \ll 1$$

Alternative interpretation: multiple sweeps cannot occur w/in ℓ^* of each other in a single fixation time:



Linkage block

$\ell^* \sim \frac{r}{s}\log(Ns)$

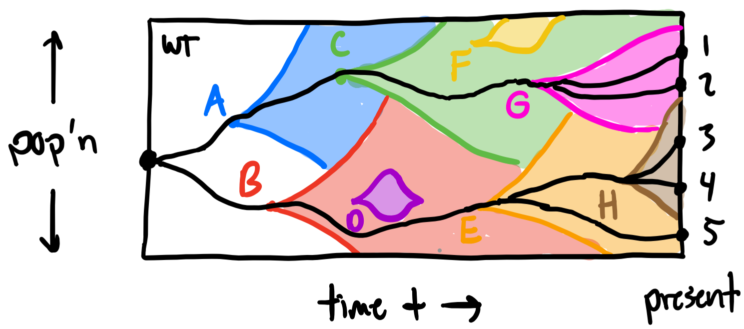
$$N\cdot\mu\lambda_b\ell^*\cdot s_b\cdot T_{fix} \ll 1$$

$$\Rightarrow \frac{\mu}{r}\cdot\lambda_b\cdot Ns \ll 1$$

$\Rightarrow$ if $\frac{\mu}{r} \sim \mathcal{O}(1) \Rightarrow$ need $\lambda_b \ll 1/Ns_b \ll 1$

$\Rightarrow$ will always break down in sufficiently large pop'ns!

$\Rightarrow$ Back to clonal interference regime!



$\Rightarrow$ Next: Finally time to consider in detail...