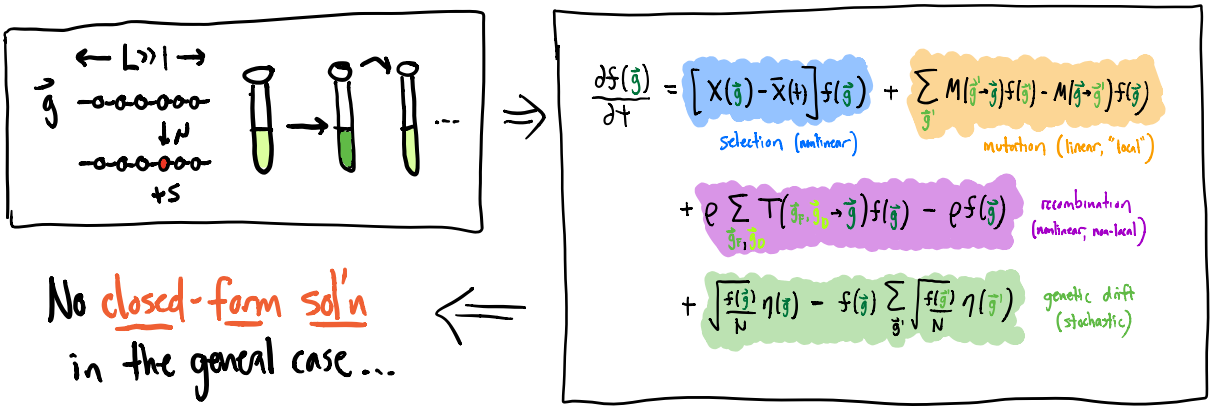
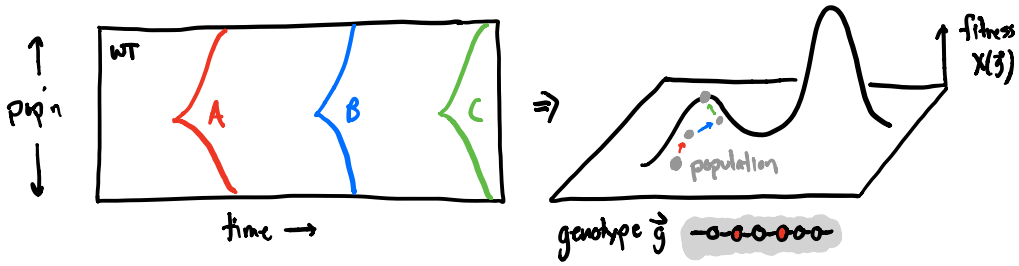
Announcements: PSET 4 DUE 3/9/21 ; Office Hrs ~~12:30~~ → 1pm Thurs

Last time: Multi-locus models of evolution ($L \gg 1$)



No closed-form soln in the general case...

① **Successive mutations regime** $[N \cdot L \cdot \mu \to 0]$



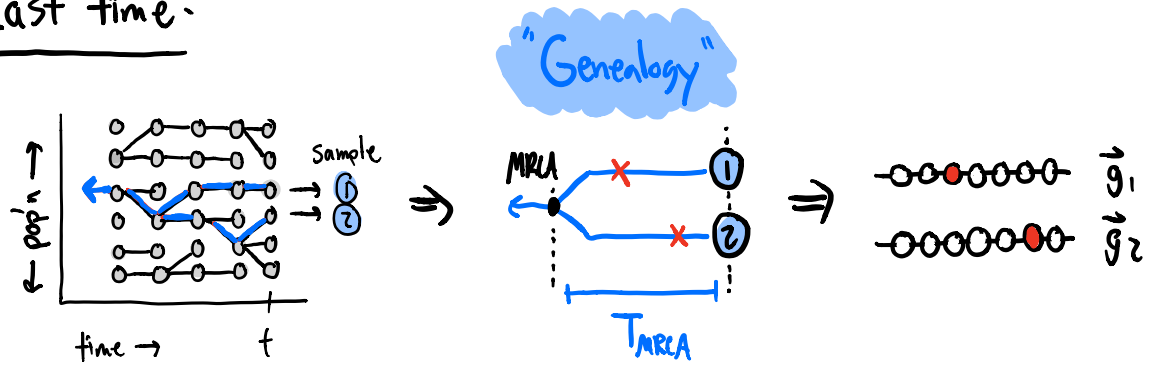
Today: ② **The Neutral Limit** (natural selection ≈ 0)

↳ new tool: **"Coalescent theory"**

↓

"backward-time approaches"

Last time:



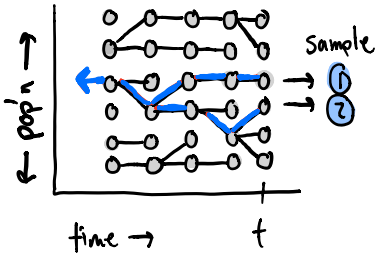
⇒ Given genealogy (T_{MRCA}), mutations occur as Poisson Process along each branch ("mutation painting")

$$\Pr\left[\begin{array}{c}\text{difference} \\ @\ \text{site } \ell\end{array} \,\middle|\, T_{MRCA}\right] \approx \begin{cases} 2\mu_\ell T_{MRCA} & \text{if } \mu T_{MRCA} \ll 1, \\ 1/2 & \text{else.} \end{cases}$$

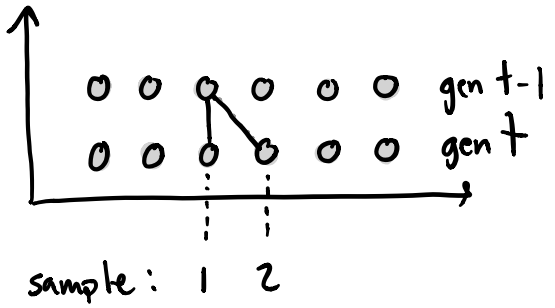
Today: what determines genealogy (T_{MRCA})?

⇒ Note: T_{MRCA} is random quantity

(genealogy will vary from sample-to-sample & simulation-to-simulation...)



⇒ key insight: start from present & work backward in time:



"coalesced"

⇒ Two individuals share ancestor in previous gen w/ probability:

total # of ancestors

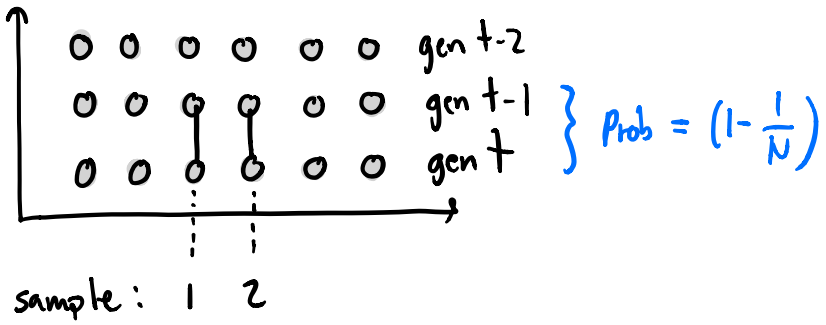
prob that 2 individuals draw it.

$$N \times \left(\frac{1}{N}\right) \times \left(\frac{1}{N}\right) = \frac{1}{N}$$

⇒ w/ probability $1/N$ ⇒ $T_{MRCA} = 1$

⇒ otherwise, diff ancestors in gen t-1 ⇒ repeat!

Process repeats itself w/ next gen:

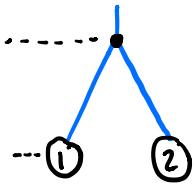


$\Rightarrow$ w/ prob $\frac{1}{N}(1-\frac{1}{N}) \Rightarrow T_{MRCA} = 2$

$\Rightarrow$ w/ prob $\frac{1}{N}(1-\frac{1}{N})^2 \Rightarrow T_{MRCA} = 3$

$\Rightarrow$ coalescence is also a Poisson Process w/ rate $\frac{1}{N}$!

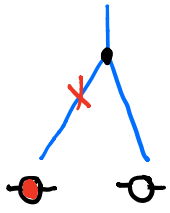
$\Rightarrow$ $T_{MRCA} \sim \text{Exponential}(N)$



$\Rightarrow \langle T_{MRCA} \rangle = N \quad \sqrt{\text{Var}(T_{MRCA})} = N$

$\Rightarrow$ total probability of mutation @ site ℓ is integral over T_{MRCA}:

$$Pr\left[\begin{matrix}\text{difference}\\ @\ \text{site } \ell\end{matrix}\right] = \int \underbrace{Pr\left(\begin{matrix}\text{diff}\\ @\ \ell\end{matrix} \middle| T_{MRCA}\right)}_{\text{mutation painting}} \underbrace{p(T_{MRCA})}_{\text{coalescent}} dT_{MRCA}$$



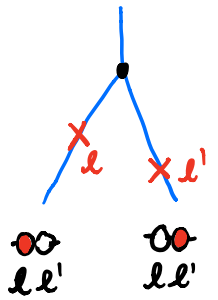
$$\underset{(\mu T \ll 1)}{\approx} \int 2N_\ell T_{MRCA}\, p(T_{MRCA})\, dT_{MRCA} = 2N_\ell \langle T_{MRCA} \rangle = 2N\mu_\ell$$

$\Rightarrow$ matches our previous result for $\langle \pi \rangle$, ✓.

Since $\langle \pi \rangle \equiv Pr\left(\begin{matrix}\text{difference}\\ @\ \text{site } \ell\end{matrix}\right)$

$\Rightarrow$ Distribution of T_{MRCA} becomes more important when considering mutations @ multiple sites, e.g.

$$Pr\left(\begin{matrix}\text{diff @ site}\\ \ell \text{ \& } \ell'\end{matrix}\right) = \int Pr\left[\pi_\ell = 1, \pi_{\ell'} = 1 \middle| T_{MRCA}\right] p(T_{MRCA})\, dT_{MRCA}$$



$$= \int \underbrace{Pr\left[\pi_\ell = 1 \middle| T_{MRCA}\right] Pr\left[\pi_{\ell'} = 1 \middle| T_{MRCA}\right]}_{\text{mutations are neutral so can't influence each other}} p(T_{MRCA})\, dT_{MRCA}$$

$$= \int (2N_e T_{MRCA}) \cdot (2N_{e'} T_{MRCA}) \cdot p(T_{MRCA}) \cdot dT_{MRCA}$$

$$= (2N_e) \cdot (2N_{e'}) \cdot \langle T_{MRCA}^2 \rangle = (2N_e) \cdot (2N_{e'}) \cdot (2N^2)$$

$$= 2 \cdot (2N_e N) \cdot (2N_{e'} N)$$

$$= 2 \cdot Pr[\pi_e] \cdot Pr[\pi_{e'}] \geq Pr[\pi_e] Pr[\pi_{e'}]$$

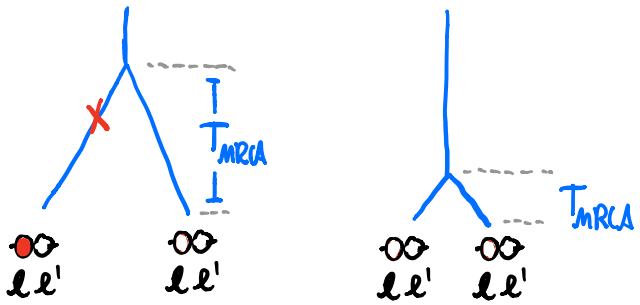
$\Rightarrow$ joint prob of mutations is <u>not</u> independent

$$Pr[\pi_{e'} = 1 \mid \pi_e = 1] = \frac{Pr[\pi_e = 1, \pi_{e'} = 1]}{Pr[\pi_e]} = 2\, Pr[\pi_{e'} = 1]$$

But previously said that neutral mutations can't influence each other directly...

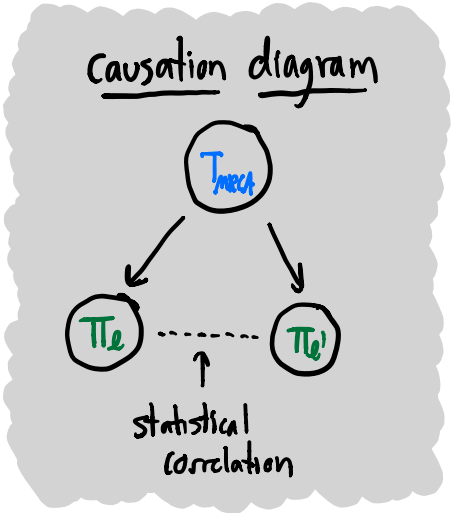
⇒ what's going on?

⇒ consider 2 trees:

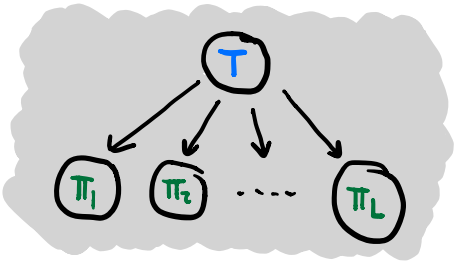


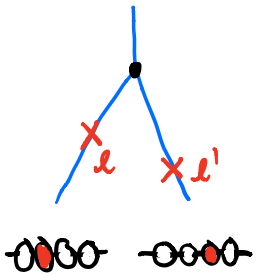
⇒ conditioned on $\pi_\ell = 1$, likely had bigger-than-avg T_{MRCA}

⇒ i.e. mutations don't interact, but are still coupled by shared genealogy



$\Rightarrow$ can keep adding more sites this way...





$\Rightarrow$ when $\mu_\ell T_{MRCA} \ll 1$, most mutations will occur @ unique site in genome

"infinite-sites approximation"

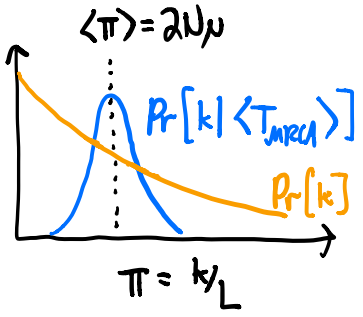
$\Rightarrow$ total # mut'ns (k) is Poisson Process w/ rate $U \equiv \sum_{\ell=1}^{L} \mu_\ell$

$$\Rightarrow \Pr[k \mid T_{MRCA}] = \frac{(2UT_{MRCA})^k}{k!} e^{-2UT_{MRCA}}$$

$$\Rightarrow \Pr[k] = \int \Pr[k \mid T_{MRCA}]\, p(T_{MRCA})\, dT_{MRCA}$$

$$= \int \frac{(2UT)^k}{k!} e^{-2UT} \frac{1}{N} e^{-T/N}\, dT$$

$$\Rightarrow \Pr[k] = \frac{(2N\nu)^k}{(2N\nu+1)^{k+1}}$$



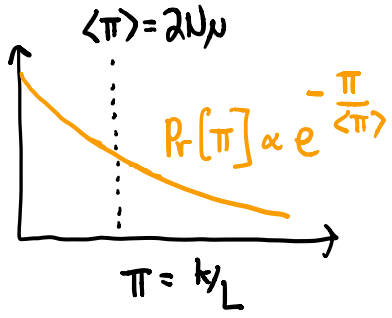
⇒ one advantage of coalescent approach:

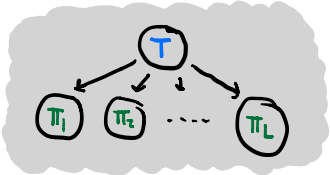
⇒ simple predictions for uncertainty in π (not just avg)

$$\text{e.g. } \mathrm{Var}(\pi) = \frac{\mathrm{Var}(k)}{L^2} = \frac{(1+2N\nu)2N\nu}{L^2}$$

$$\Rightarrow \text{or } C_V^2 \equiv \frac{\mathrm{Var}(\pi)}{\langle\pi\rangle^2} = \frac{1+2N\nu}{2N\nu} \geq 1$$

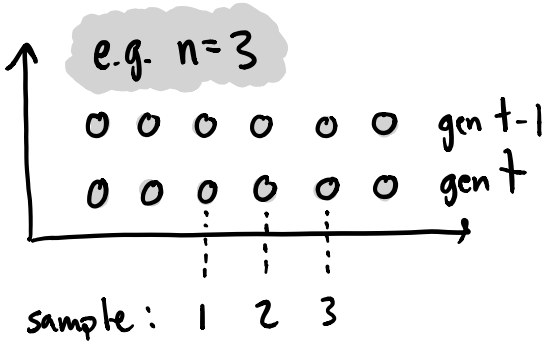
⇒ i.e. π does not self-average on a long asexual genome!





⇒ fluct'ns in T_{MRCA} affect many sites!

Larger sample sizes ($n > 2$)



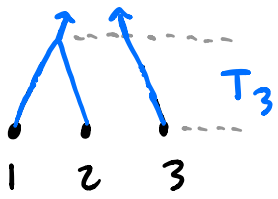
$\Rightarrow$ Prob that any 2 share ancestor is $\frac{1}{N}$ [$\times \binom{3}{2}$ pairs]

$\Rightarrow$ Prob that all 3 share ancestor $= N \cdot \left(\frac{1}{N}\right)\left(\frac{1}{N}\right)\left(\frac{1}{N}\right) = \frac{1}{N^2}$

$\Rightarrow$ when $N \gg 1$ → only need to worry about **pairwise coalescence**

(known as "Kingman's coalescent")

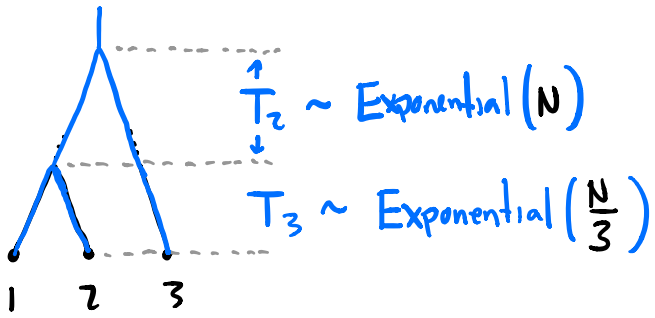
(all pairs are equally likely to coalesce)



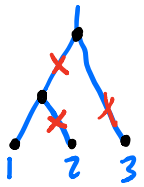
$\Rightarrow$ total prob of coalescence $= \frac{3}{N}$ per gen

$\Rightarrow T_3 \sim \text{Exponential}\left(\frac{N}{3}\right)$

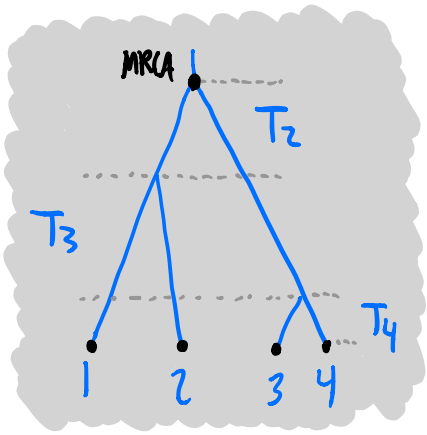
⇒ now we have sample of $n=2$... ⇒ repeat!



⇒ Done! can now paint on mutations...



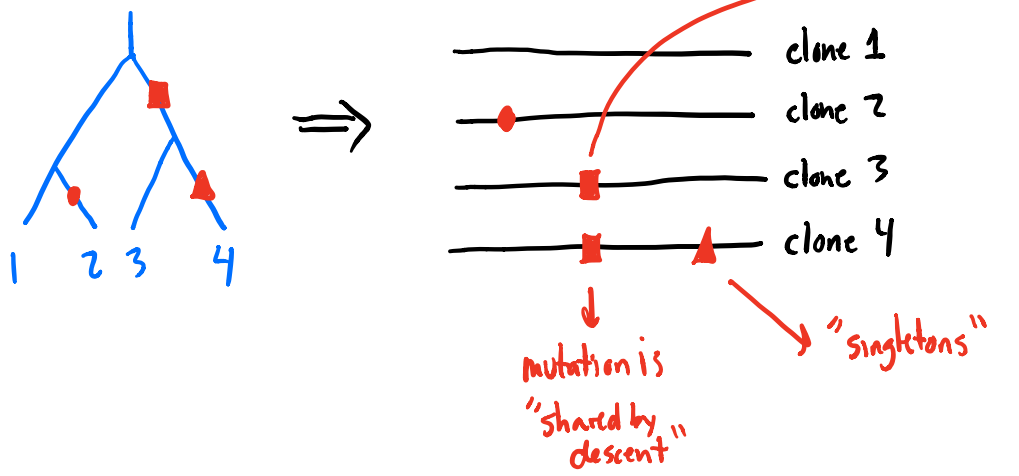
Easily generalizes to sample of size n:



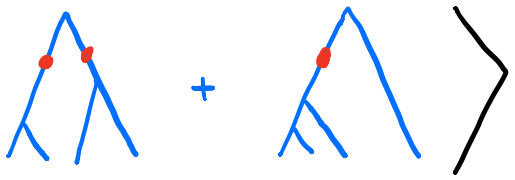
1. @ each step, only consider coalescence between pairs of lineages
2. Time until next coalescence event is $T_n \sim \text{Exponential}\left(N/\binom{n}{2}\right)$
3. choose random pair to coalesce

repeat!

4. then can paint mutations on @ end:



⇒ easy to simulate for $n>2$, but hard to calculate...

e.g. $\left\langle \begin{array}{c}\#\text{ doubletons in} \\ \text{sample } n=4\end{array} \right\rangle =$ 

$\Rightarrow$ must avg over: ① tree topologies
② branch lengths | topology
③ mutation painting | branch lengths

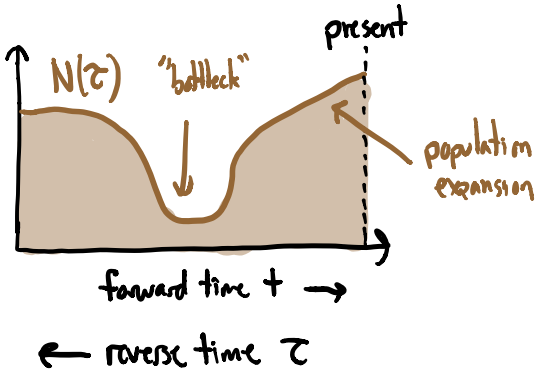
$\Rightarrow$ compare to single-locus prediction (easy!)

$$\left\langle \begin{array}{c}\#\text{ doubletons} \\ \text{in } n=4\end{array} \right\rangle = \int \binom{4}{2} f^2 (1-f)^{4-2} \cdot \left(\frac{2N\mu}{f}\right) \cdot df = N\mu$$

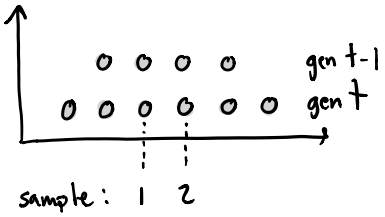
$\Rightarrow$ why use coalescent picture then??

Answer: coalescent picture makes it easy to model demography!

e.g. what if N was not constant, but varied historically in time:



⇒ coalescent picture still works, but coalescent prob ⟶ $1/N(\tau)$

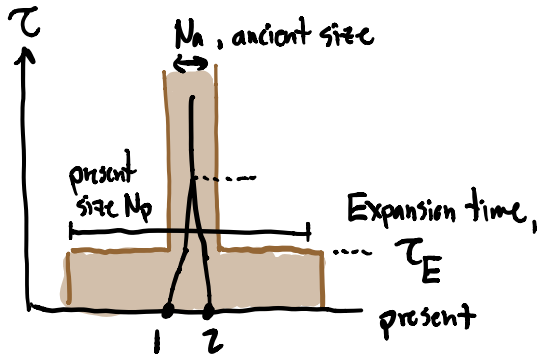


⇒ coalescence = "inhomogeneous" Poisson process:

$$\Rightarrow \Pr[T_2 > \tau] = \prod_{\tau'=1}^{\tau} \left[1 - \frac{1}{N(\tau')}\right] \approx e^{-\int_0^{\tau} \frac{d\tau'}{N(\tau')}}$$

$$\Rightarrow \Pr[T_2 = \tau] = \frac{1}{N(\tau)} e^{-\int_0^{\tau} \frac{d\tau}{N(\tau')}}$$

Simple example: rapid expansion in recent past



① if $N_p \to \infty$ $(\tau_E \ll N_p)$

$\Rightarrow$ no coalescence until τ_E

$\Rightarrow$ coalescence @ rate $\frac{1}{N_a}$ after

$\Rightarrow \langle T_2 \rangle = \tau_E + N_a$

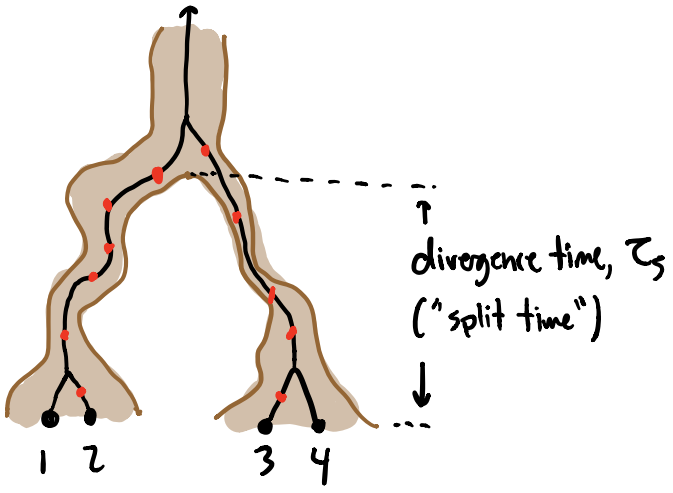
$$\Rightarrow \langle \pi \rangle = 2\mu\langle T_2 \rangle = 2\mu(\tau_E + N_a) \approx 2\mu N_a \quad (\text{if } \tau_E \ll N_a)$$

if $N_p \mu \sim 100 \Rightarrow$ why $\pi \sim 1e^{-03}$ in humans??

$\Rightarrow$ answer: $N(t)$ was smaller backward in time.

compare to $\frac{df}{dt} = \mu(1-f) - vf + \sqrt{\frac{f(1-f)}{N(t)}}\eta(t)$

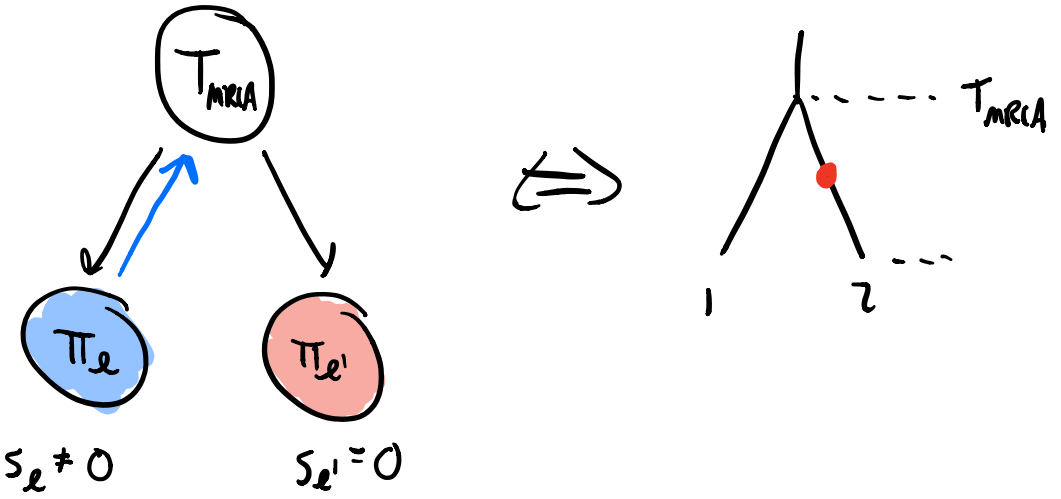
Can also easily add population structure



$\Rightarrow$ prob of coalescence between pop'ns = 0 until time $t = \tau_s$

$\Rightarrow$ much of pop gen is about inferring these demographic models

$\Rightarrow$ downside: hard to add selection back in to picture...



⇒ when is this going to be an issue?

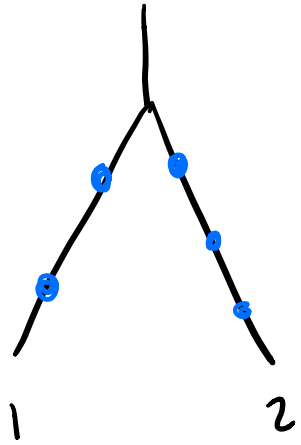
⇒ for $L=1$ case, needed $N|s| \ll 1$ for effectively neutral.

⇒ for $L \gg 1$, selection looks like $\left[X(\vec{y}) - \bar{X}(t)\right] f(\vec{y})$

vs

$sf(1-f)$ in $L=1$

⇒ suggests: $N\left|X(\vec{y}) - \bar{X}\right| \ll 1$ for neutrality



① assume effective neutrality:

⇒ total # mutations $\sim NU$

$$\left|X(\vec{y}) - X(\vec{y}_2)\right| = \sqrt{NU s^2}$$

⇒ self consistent:

$$(NU)(Ns)^2 \ll 1$$

e.g. $Ns \sim 0.1$ (neutral in single locus setting)

$$NU = \langle \pi \rangle L = \begin{cases} 10^4 & \text{for bacteria in a soil} \\ 10^6 & \text{for humans.} \end{cases}$$

$$\sqrt{10^4 \cdot (10^{-1})^2} = 10 \gg 1$$