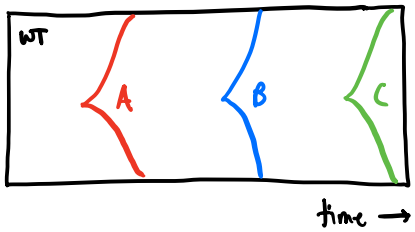
Chapter 11

Neutral theory and the coalescent

Neutral theory & the Coalescent

Successive mutations:



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

⇒ ~1 variant present @ high freqs ⇒ solved by reducing to L=1 model

* But genomes in data separated by <u>multiple</u> mut'ns
(e.g. humans, 2 individuals differ by ~1 mut / 1000 bp)

⇒ need to understand what's going on in these cases...

$$\frac{\partial f(\vec{g})}{\partial t} = \cancel{\sim(x-\bar{x})}^{\,0} + \sim L\times\mu + \cancel{\sim\rho}^{\,0} + \sim\frac{z}{\sqrt{N}}$$

⇒ one other limit that's well understood:

<u>neutral evolution</u> in <u>nonrecombining genome</u>

when $X(\vec{g})=0$ + $e=0$, left with: $(\mu_\ell = \nu_\ell)$

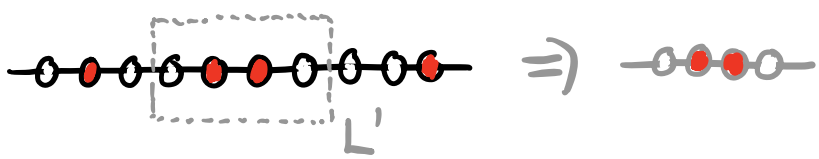
$$\frac{\partial f(\vec{g})}{\partial t} = \underbrace{\sum_{|\vec{g}'-\vec{g}|=1} \sum_\ell \mu_\ell f(\vec{g}')\left[g_\ell(1-g_\ell') + (1-g_\ell)g_\ell'\right]}_{\text{incoming mutations}} - \underbrace{\sum_\ell \mu_\ell f(\vec{g})}_{\text{outgoing mutations}}$$

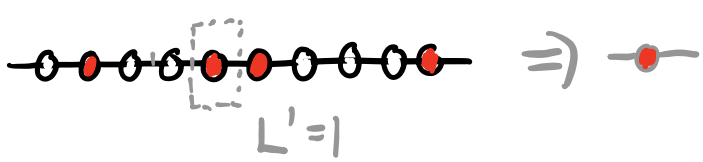
$$+ \sqrt{\frac{f(\vec{g})}{N}}\,\eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}}\,\eta(\vec{g}')$$

genetic drift

<u>Key insight</u>: sites don't actually <u>influence</u> each other (because neutral)

e.g.

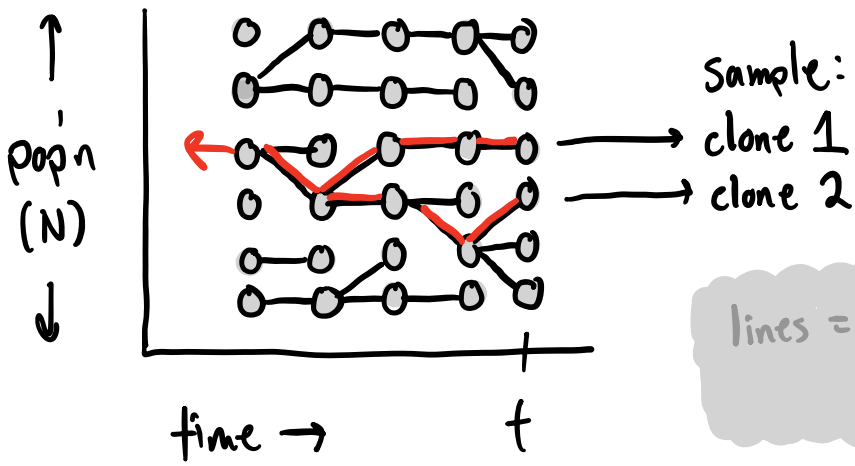




$p(f) = \frac{2N\mu}{f}$

$\Rightarrow$ <u>2nd key insight</u>: can take $L'=0$ —

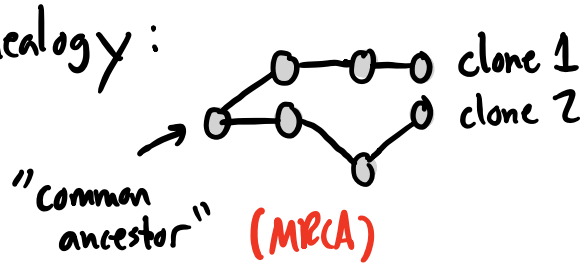
E.g. simulation of neutral pop'n in Wright-Fisher model:



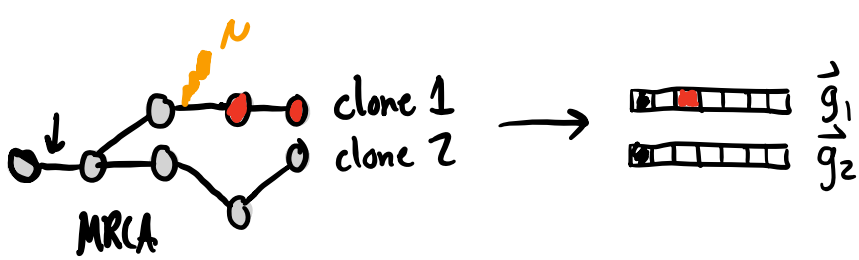
lines = birth events @ each generation

⇒ <u>key insight</u>: lines also = genealogical relationships backward in time!

Genealogy:

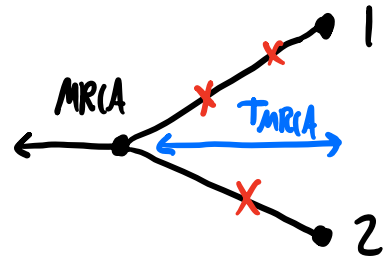


⇓ <u>differences</u> between sampled individuals = <u>mutations on genealogy</u>



$\Rightarrow$ Mut'ns @ site ℓ $\approx$ <u>Poisson Process</u>

w/ rate μ_ℓ on each branch



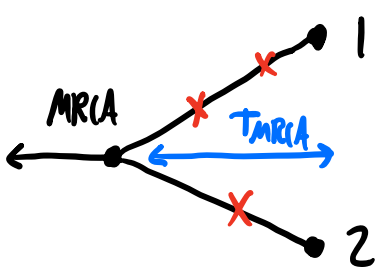
$\hookrightarrow$ total # mut'ns $\sim \text{Poisson}(2 T_{MRCA} \mu_\ell)$

$\Rightarrow$ 2 extreme limits:

(1) $\mu_\ell T_{MRCA} \ll 1$ $\Rightarrow$ 0 or 1 mutations on whole tree

$$\Rightarrow \Pr\left[\begin{matrix}\text{genetic diff}\\ \text{@ site } \ell\end{matrix} \,\middle|\, T_{MRCA}\right] \approx 2 T_{MRCA} \mu_\ell$$

(2) $\mu_\ell T_{MRCA} \gg 1$ $\Rightarrow$ lots of forward & backward mutations along each branch.



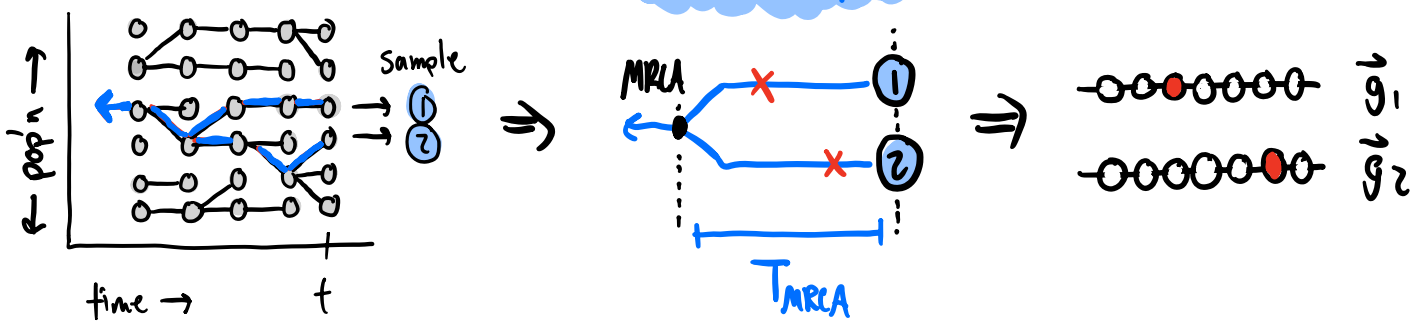
$$\Rightarrow \Pr\left[\begin{matrix}\text{genetic diff}\\ \text{@ site } L\end{matrix} \,\middle|\, T_{MRCA}\right] = \underbrace{2}_{\#\text{ states}} \cdot \underbrace{\frac{1}{2} \cdot \frac{1}{2}}_{\Pr(\text{same state})}$$

$$= \frac{1}{2}$$

(more generally $\sim 1/4$ if A, C, T, G's ...)

Recap:

"Genealogy"



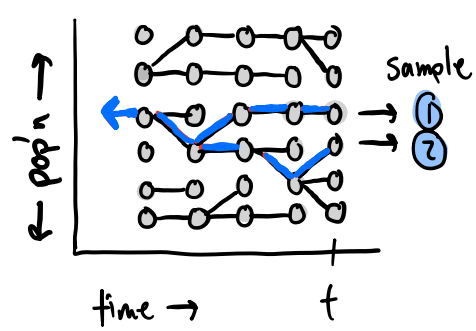
⇒ Given genealogy (T_{MRCA}), mutations occur as Poisson Process along each branch ("mutation painting")

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

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

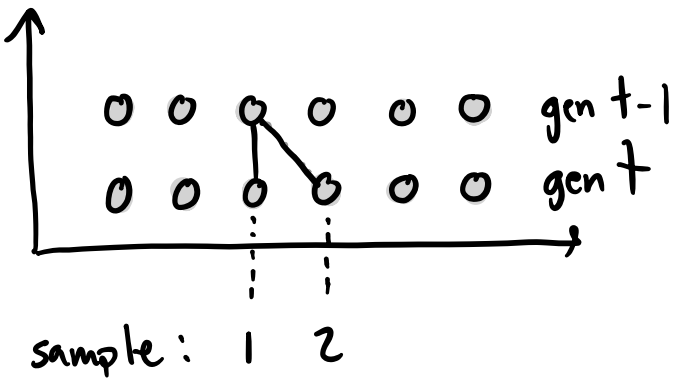
⇒ Note: T_{MRCA} is random quantity

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



⇒ can predict using tool called "coalescent theory"

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



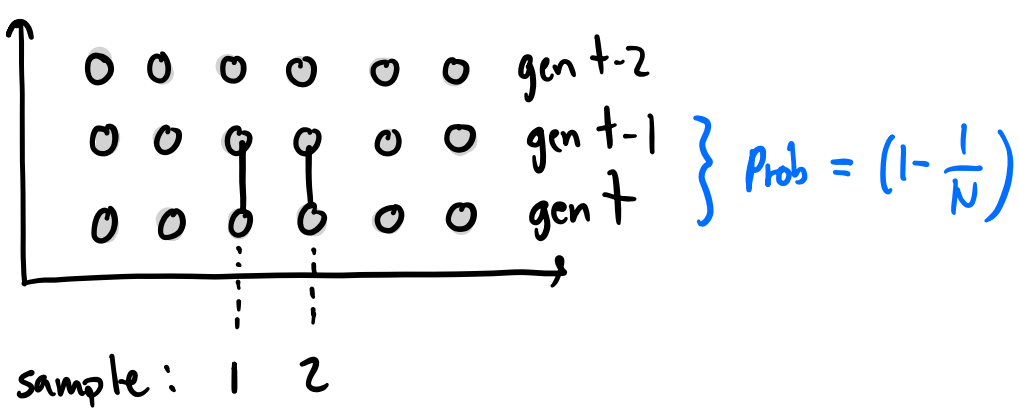
⇒ Two individuals share ancestor in previous gen w/ probability: "coalesced"

$$N \times \underbrace{\left(\frac{1}{N}\right) \times \left(\frac{1}{N}\right)}_{\text{prob that both draw same}} = \frac{1}{N}$$

N → # possible ancestors

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

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



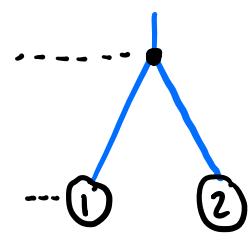
Process repeats itself w/ next gen:

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

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

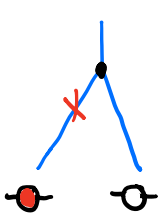
$\Rightarrow$ coalescence is also a <u>Poisson Process</u> w/ rate $\frac{1}{N}$!

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



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

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



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

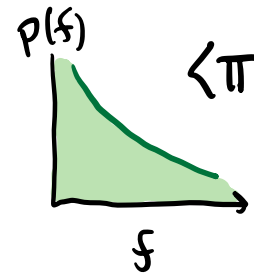
$$\underset{(\mu T \ll 1)}{\approx} \int 2\mu_\ell T_{MRCA} \cdot p(T_{MRCA})\, dT_{MRCA}$$

$$= 2\mu_\ell \langle T_{MRCA} \rangle = 2\mu_\ell N$$

$$= 2N\mu_\ell$$

$\Rightarrow$ matches our previous result

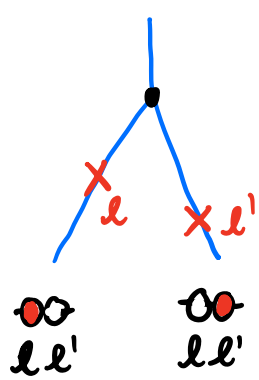
since $\langle \pi \rangle \equiv \Pr\left(\begin{matrix}\text{difference} \\ \text{@ site } \ell\end{matrix}\right)$



$\langle \pi \rangle \approx 2N_e N$ ✓

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

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



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

$$= \int \left(2\mu_\ell T_{MRCA}\right) \cdot \left(2\mu_{\ell'} T_{MRCA}\right) \cdot p(T_{MRCA}) \cdot dT_{MRCA}$$

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

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

$$= 2\cdot Pr[\pi_\ell]\cdot Pr[\pi_{\ell'}] \geq Pr[\pi_\ell]\,Pr[\pi_{\ell'}]$$

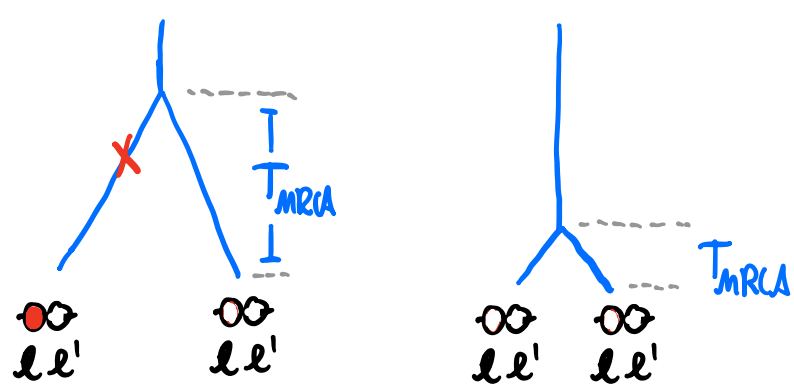
<u>Upshot</u>: joint prob of mut'ns is <u>not</u> independent:

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

<u>But</u> previously said that <u>neutral</u> <u>mutations</u> can't influence each other <u>directly</u>...

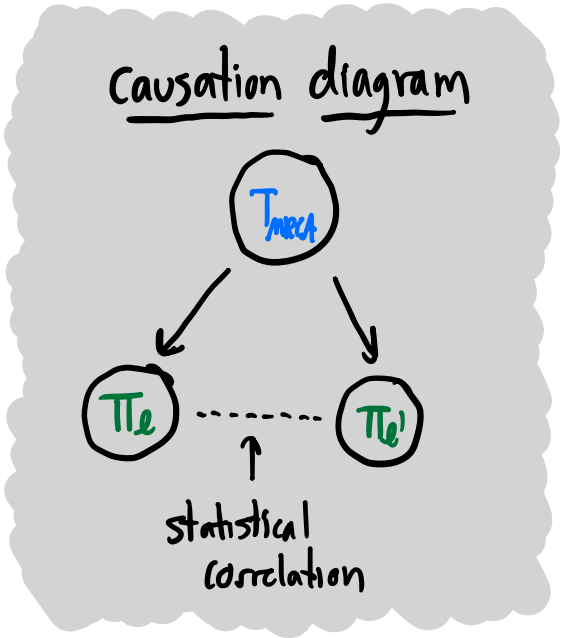
$\Rightarrow$ what's going on?

$\Rightarrow$ consider 2 trees:

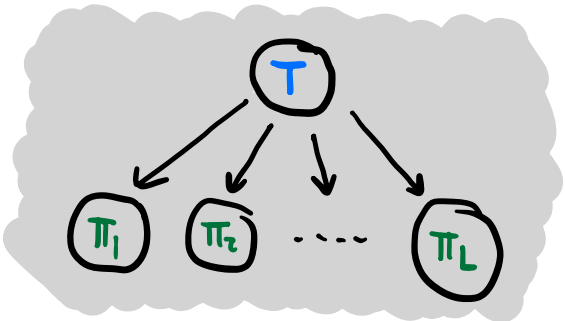


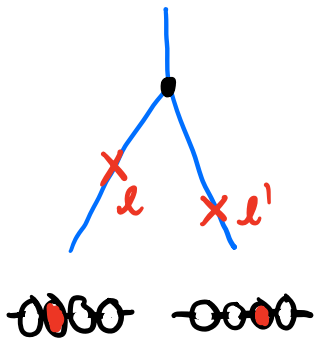
$\Rightarrow$ conditioned on $\pi_\ell = 1$, likely had bigger-than-avg T_{MRCA}

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



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





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

"infinite-sites approximation"

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

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

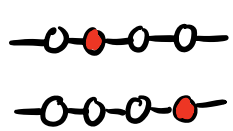
$$\Rightarrow \quad \Pr[k] = \int \Pr[k \mid T_{MRCA}]\, p(T_{MRCA})\, dT_{MRCA}$$

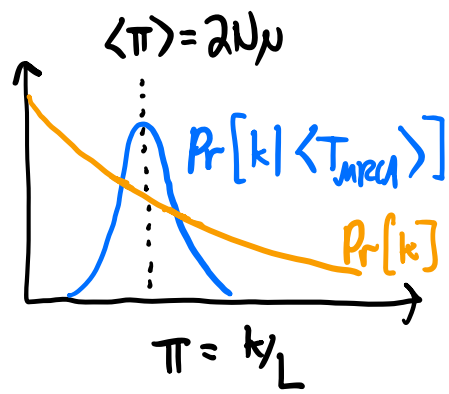
$$= \int \frac{(2UT)^k}{k!} e^{-2UT} \frac{1}{N} e^{-T/N}\, dT$$

geometric dist'n.

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

total # diffs btw 2 genomes





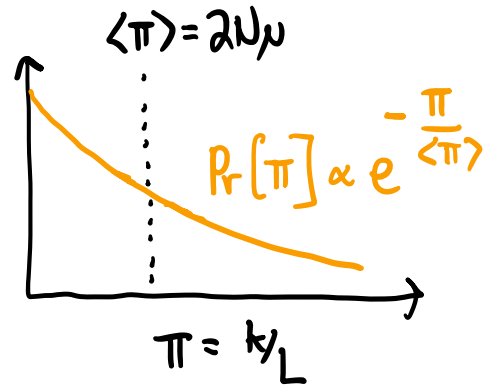
$\Rightarrow$ one advantage of coalescent approach:

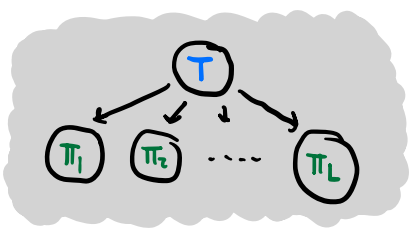
$\Rightarrow$ simple predictions for <u>uncertainty</u> in π (not just avg)

e.g. $\mathrm{Var}(\pi) = \dfrac{\mathrm{Var}(k)}{L^2} = \dfrac{(1+2N\mu)\,2N\mu}{L^2}$

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

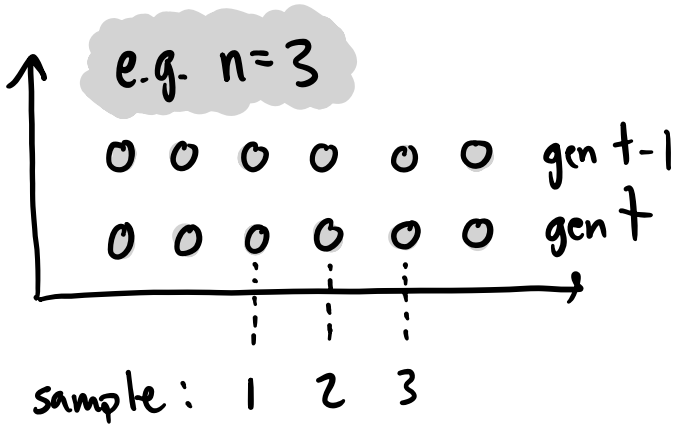
$\Rightarrow$ i.e. π does not <u>self-average</u> on a long <u>asexual</u> genome!





$\Rightarrow$ fluct'ns in T_{MRCA} affect <u>many</u> sites!

Larger sample sizes $(n>2)$



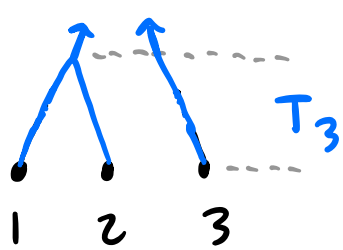
$\Rightarrow$ Prob that <u>any 2</u> share ancestor is $\left(\frac{1}{N}\right)\left[\times \overset{\sim 3}{\binom{3}{2}} \text{ pairs}\right]$

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

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

(known as "Kingman's coalescent")

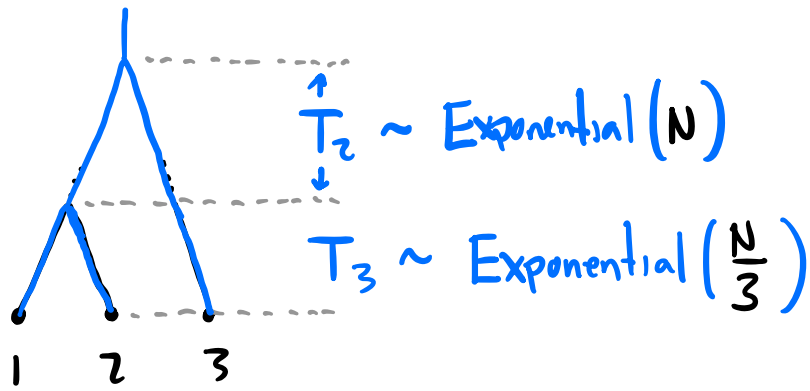
(all pairs are equally likely to coalesce)



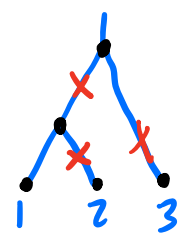
$\Rightarrow$ total prob of coalescence $= \frac{3}{N}$ per gen

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

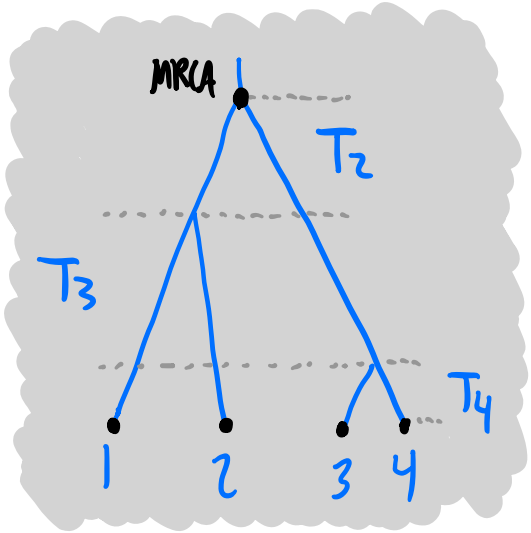
$\Rightarrow$ now we have sample of $n=2$... $\Rightarrow$ repeat!



$\Rightarrow$ Done! can now paint on mutations...



Easily generalizes to sample of size n:



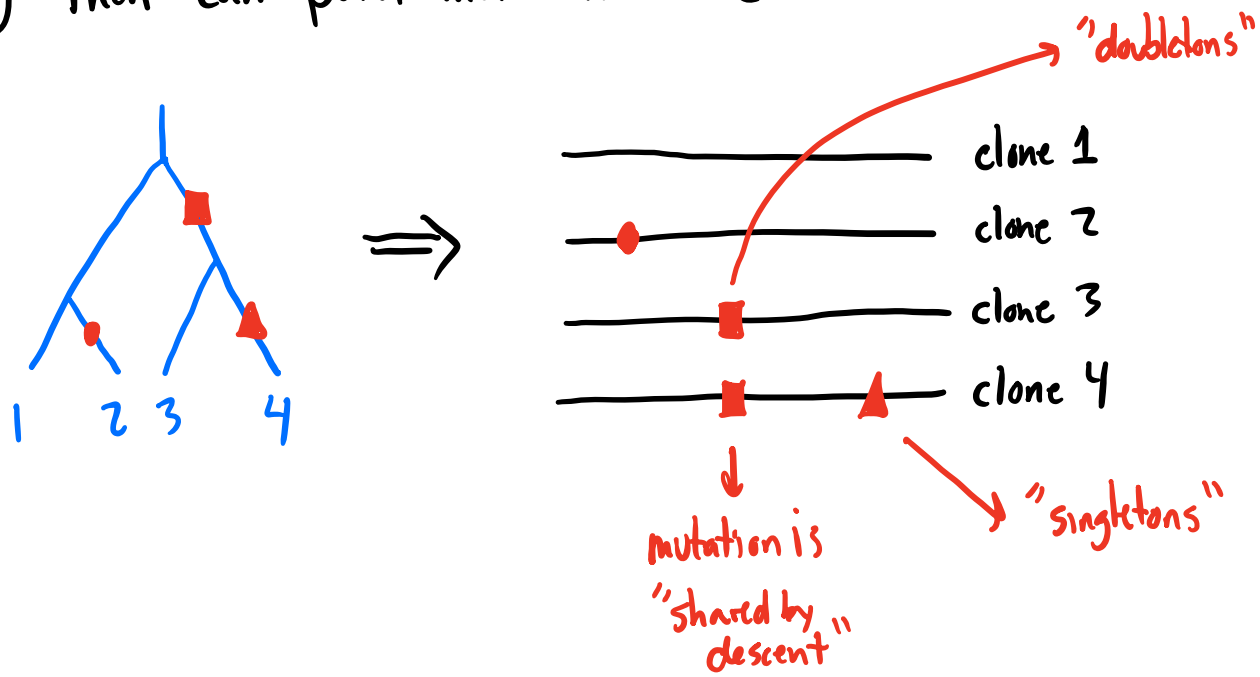
① @ each step, only consider coalescence between pairs of lineages

② Time until next coalescence event is $T_n \sim \text{Exponential}\left(N/\binom{n}{2}\right)$

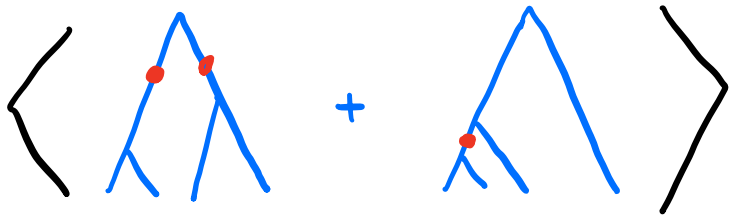
③ choose random pair to coalesce

repeat!

④ then can paint mutations on @ end:



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

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

$\Rightarrow$ must avg over: ① tree topologies

② branch lengths | topology

③ mutation painting | branch lengths

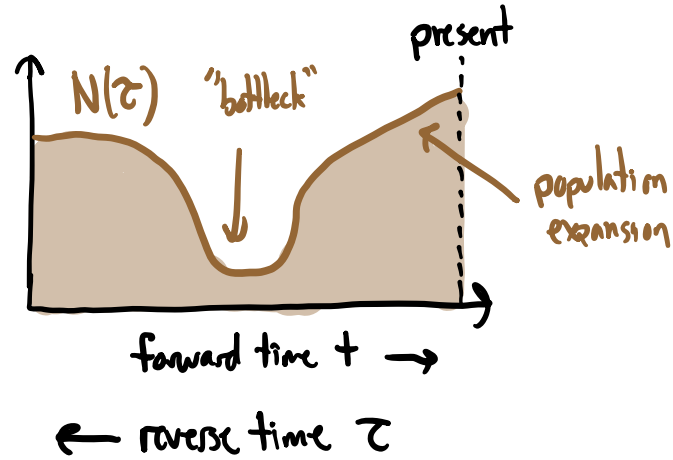
$\Rightarrow$ compare to single-locus prediction (easy!)

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

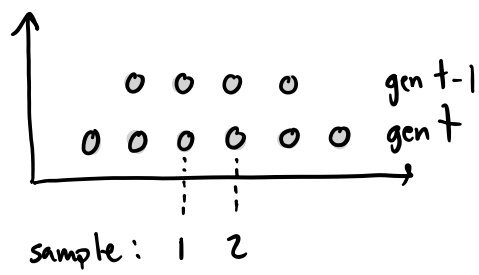
$\Rightarrow$ why use coalescent picture then??

Answer: coalescent picture makes it easy to model demography!

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



$\Rightarrow$ coalescent picture still works, but coalescent prob $\longrightarrow 1/N(\tau)$

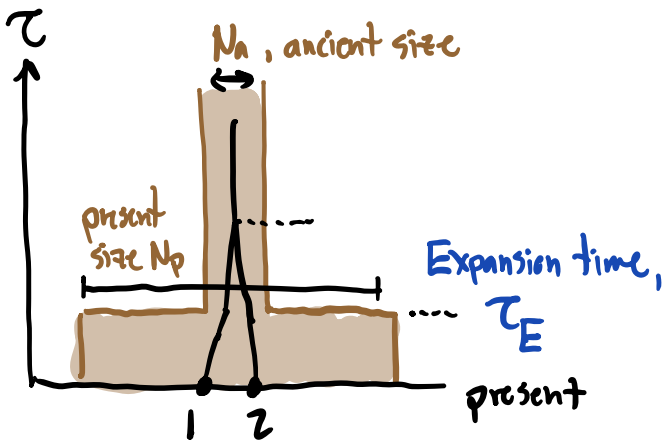


$\Rightarrow$ coalescence = "inhomogeneous" Poisson process:

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

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

Simple example: rapid expansion in recent past



If $N_p \approx \infty$ & $\tau_E \ll N_p$:

① no coalescence until τ_E

② coalescence @ rate $\frac{1}{N_a}$ after

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

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

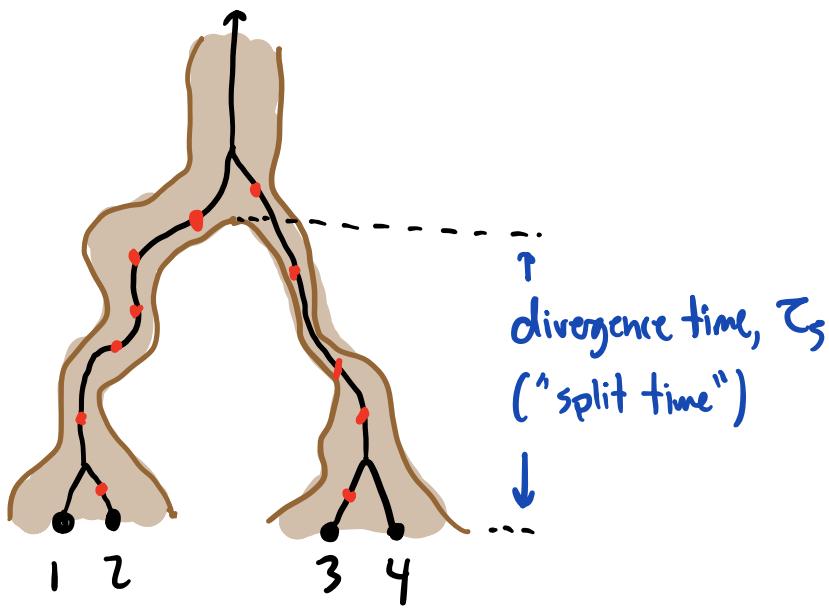
(compare to foward time calc... $\frac{\partial f}{\partial t} = \mu(1-f) - \nu f + \sqrt{\frac{f(1-f)}{N(t)}}\,\eta(t) \Rightarrow p(f,t)$)

can revisit our earlier puzzle: if $N_p \cdot \mu \sim 100$ in humans
why $\langle \pi \rangle \sim 10^{-3}$?

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

$$\Rightarrow N_a \approx 10^5 \quad \left(\tau_E \ll 10^5 \text{ gens}\right)$$

Can also easily add <u>population structure</u>



$\Rightarrow$ Pr(coalescence) <u>btween</u> pop'ns = 0 until time $\tau = \tau_s$

$\Rightarrow$ much of pop gen is about <u>inferring</u> these <u>demographic models</u>

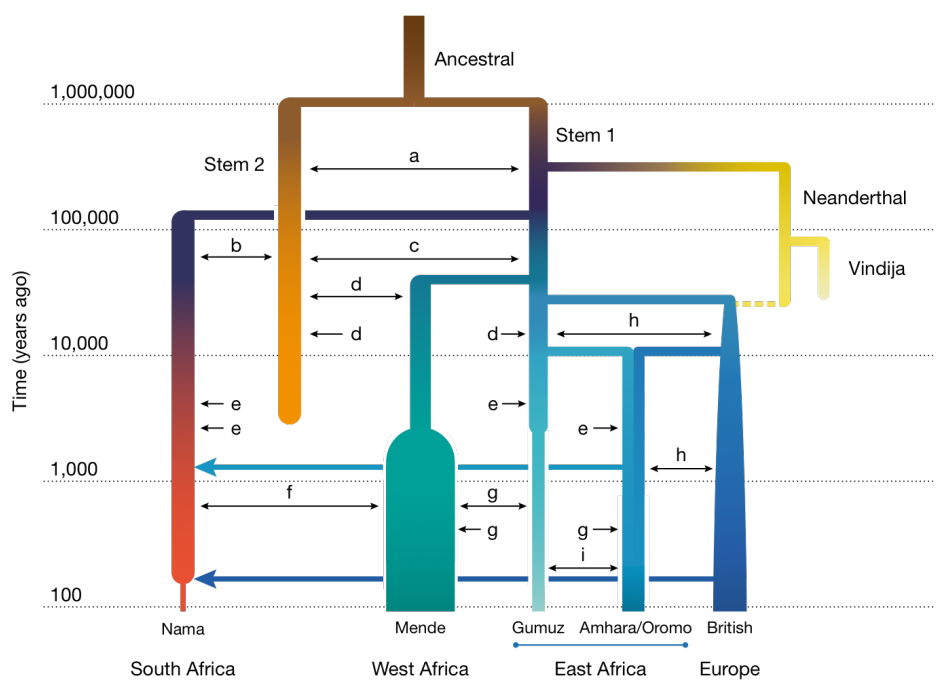
e.g.:

Article

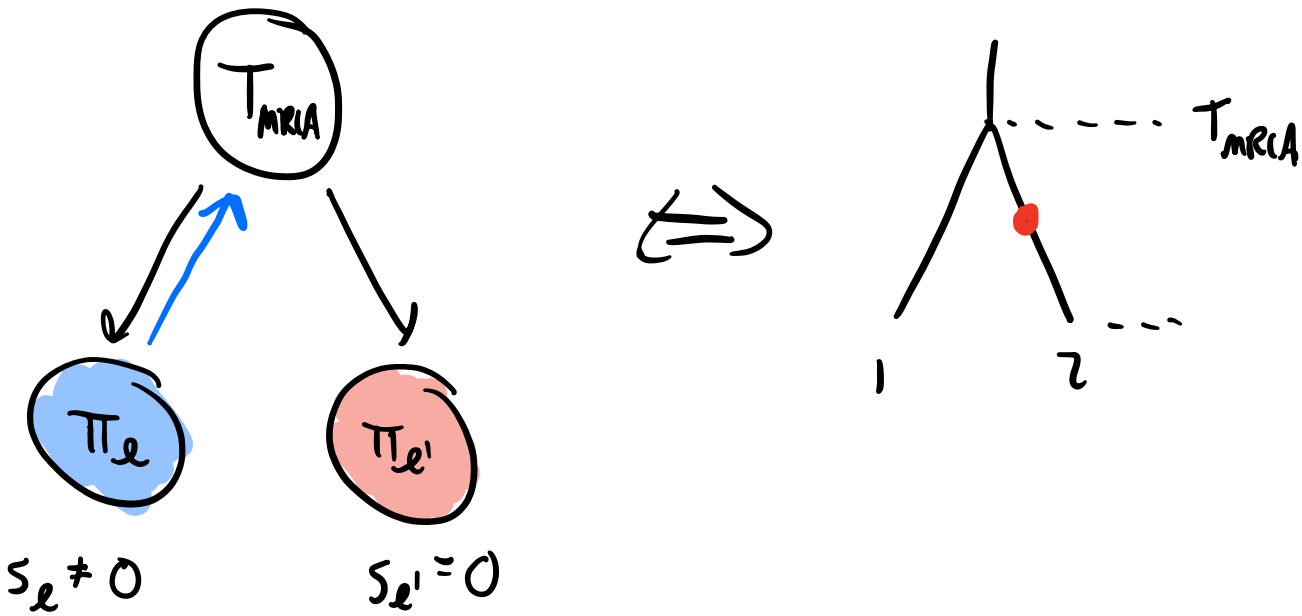
A weakly structured stem for human origins in Africa

Aaron P. Ragsdale[1], Timothy D. Weaver[2], Elizabeth G. Atkinson[3], Eileen G. Hoal[4,5,6], Marlo Möller[4,5,6], Brenna M. Henn[2,7] ✉ & Simon Gravel[8] ✉

Published online: 17 May 2023



$\Rightarrow$ <u>downside</u>: hard to add selection back in to picture...



$\Rightarrow$ when is this going to be an issue?

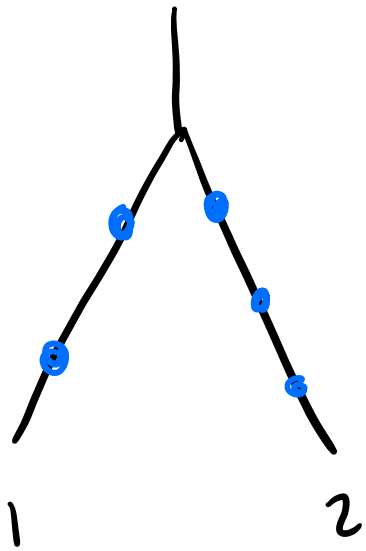
$\Rightarrow$ for $L=1$ case, needed $N|s| \ll 1$ for effectively neutral.

$\Rightarrow$ for $L \gg 1$, selection looks like $\left(X(\vec{g}) - \bar{X}(t)\right) f(\vec{g})$

vs

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

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



① assume effective neutrality:

$\Rightarrow$ total # mutations $\approx NU$

$$|X(\vec{g}_1) - X(\vec{g}_2)| = \sqrt{NU s^2}$$

$\Rightarrow$ self consistent:

$$\boxed{(NU)(Ns)^2 \ll 1}$$

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

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

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