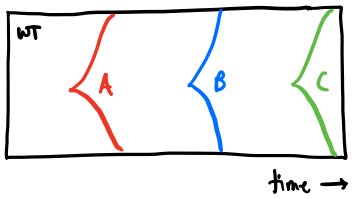
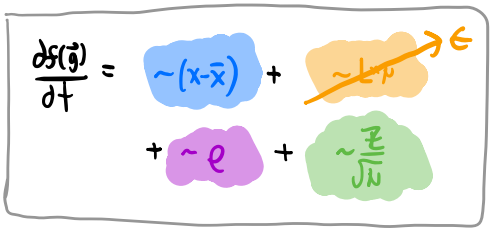
Chapter 11

Neutral theory and the coalescent

Neutral theory & the Coalescent

Successive mutations:

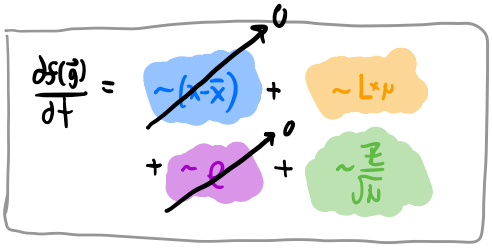




⇒ ~1 variant present @ high freqs ⇒ solved by reducing to L=1 model

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

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



⇒ one other limit that's well understood:

neutral evolution in nonrecombining genome

when $X(\vec{g})=0$ + $e=0$, left with: $(\mu_e = \nu_e)$

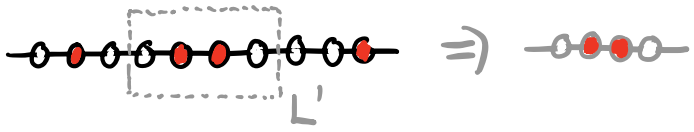
$$\frac{df(\vec{g})}{dt} = \sum_{|\vec{g}'-\vec{g}|=1} \sum_e \mu_e f(\vec{g}')\left[g_e(1-g_e') + (1-g_e)g_e'\right] - \sum_e \mu_e f(\vec{g})$$

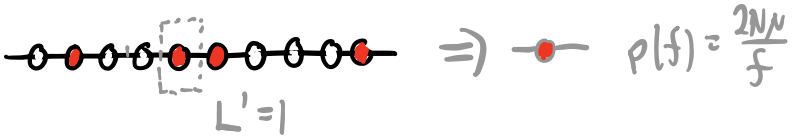
incoming mutations — outgoing mutations

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

genetic drift

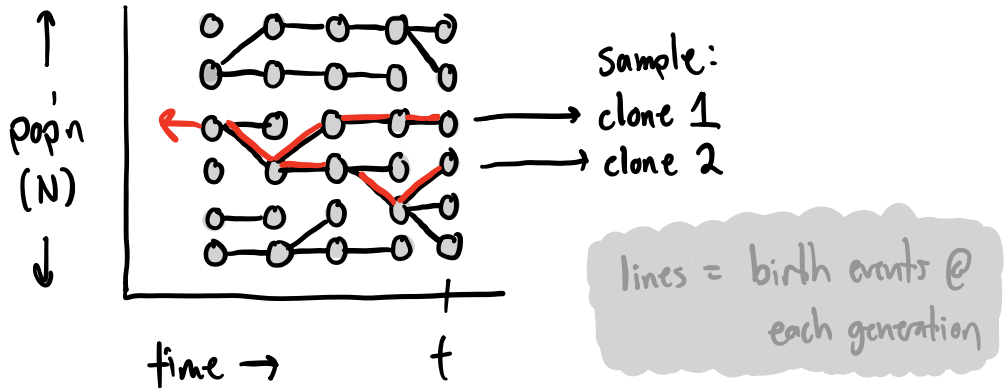
Key insight: sites don't actually influence each other (because neutral)

e.g. 



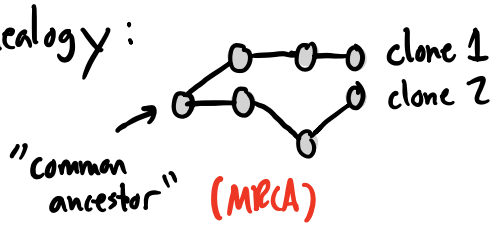
⇒ 2nd key insight: can take $L'=0$ —

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

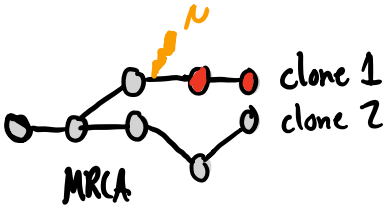


⇒ key insight: lines also = genealogical relationships backward in time!

Genealogy:

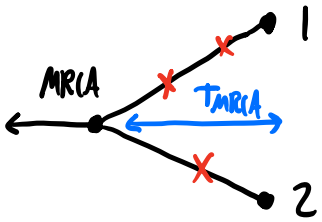


differences between sampled individuals = mutations on genealogy



$\Rightarrow$ Mut'ns @ site $\ell \approx$ Poisson Process

w/ rate μ_ℓ on each branch



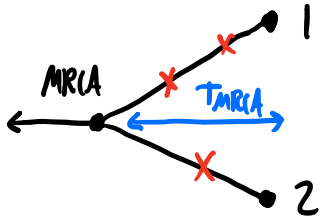
↳ total # mut'ns $\sim$ Poisson$(2 T_{MRCA} \mu_\ell)$

$\Rightarrow$ 2 extreme limits:

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

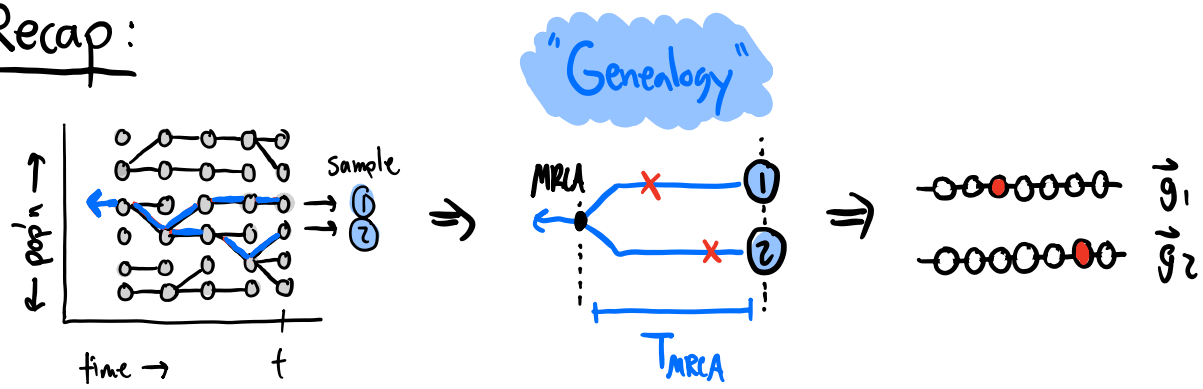
$\Rightarrow \Pr\left[\begin{matrix}\text{genetic diff} \\ \text{@ site } \ell\end{matrix} \middle| T_{MRCA}\right] \approx$

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



$\Rightarrow \Pr\left[\begin{matrix}\text{genetic diff} \\ \text{@ site } L\end{matrix} \middle| T_{MRCA}\right] =$

Recap:



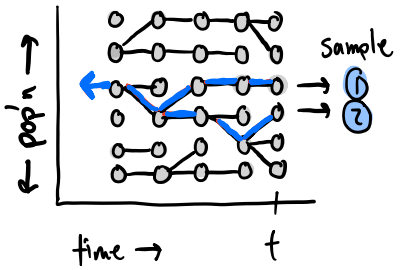
⇒ Given genealogy (T_{MRCA}), mutations occur as Poisson Process along each branch ("mutation painting")

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

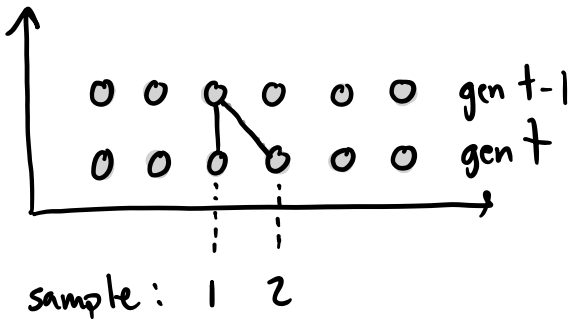
Question: what determines genealogy (T_{MRCA})?

⇒ **Note**: T_{MRCA} is random quantity

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



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



"coalesced"

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

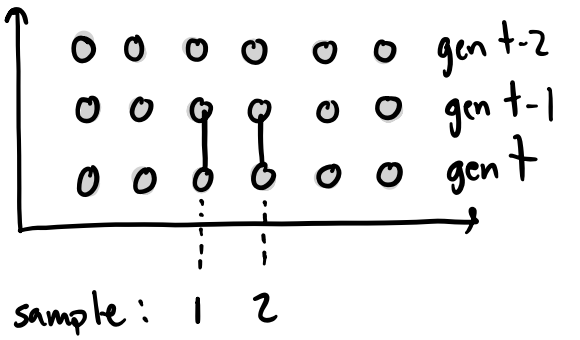
prob that both draw same

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

↳ # possible ancestors

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

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



Prob = $(1 - \frac{1}{N})$

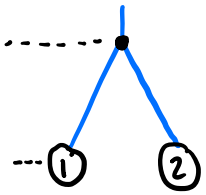
Process repeats itself w/ next gen:

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

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

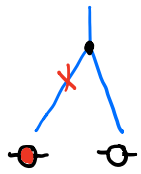
$\Rightarrow$ coalescence is also a Poisson Process w/ rate $\frac{1}{N}$!

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



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

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



$$\Pr[\text{difference @ site } \ell] = \int \underbrace{\Pr(\text{diff @ } \ell \mid T_{MRCA})}_{\text{mutation painting}} \underbrace{p(T_{MRCA})}_{\text{coalescent}} \, dT_{MRCA}$$

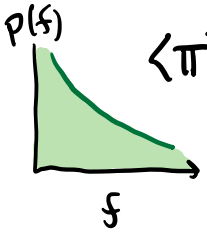
$$\underset{(NT \ll 1)}{\approx} \int 2\mu_\ell T_{MRCA} \cdot p(T_{MRCA}) \, dT_{MRCA}$$

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

$$= 2N\mu_\ell$$

$\Rightarrow$ matches our previous result

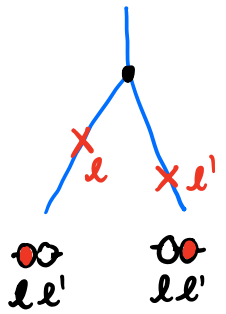
Since $\langle \pi \rangle \equiv Pr(\text{difference @ site } \ell)$



$\langle \pi \rangle \approx 2N_e\mu$

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

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



$$= \int Pr\left[\pi_\ell = 1 \,\middle|\, T_{MRCA}\right] Pr\left[\pi_{\ell'} = 1 \,\middle|\, T_{MRCA}\right] p(T_{MRCA})\, dT_{MRCA}$$

mut'ns are neutral, so can't affect each other!

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

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

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

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

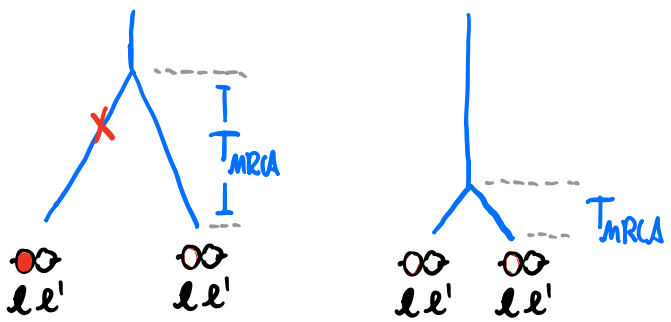
Upshot: joint prob of mut'ns is **not** independent:

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

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

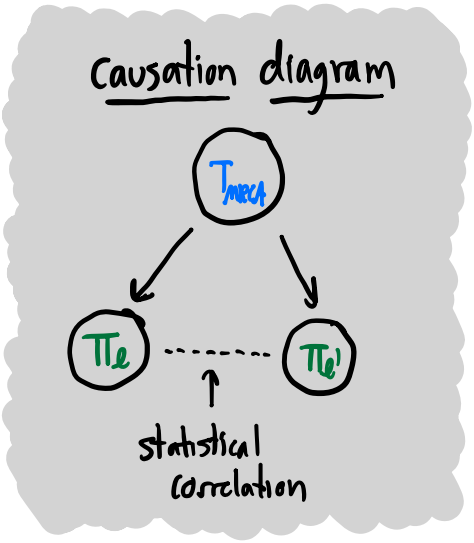
$\Rightarrow$ what's going on?

$\Rightarrow$ consider 2 trees:

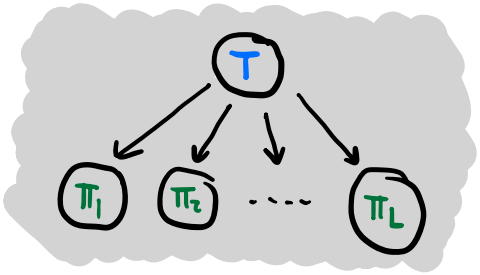


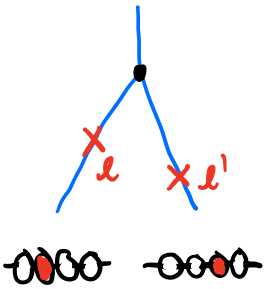
$\Rightarrow$ conditioned on $\pi_\ell = 1$, likely had bigger-than-avg T_{MRCA}

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



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





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

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

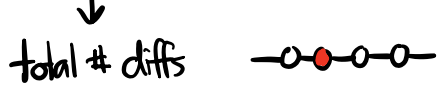
$$\Rightarrow \quad \Pr[k \mid T_{MRCA}] = \frac{(2UT_{MRCA})^k}{k!} e^{-2UT_{MRCA}}$$

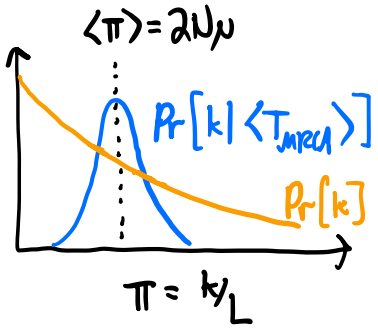
$$\Rightarrow \quad \Pr[k] = \int \Pr[k \mid T_{MRCA}]\, p(T_{MRCA})\, dT_{MRCA}$$

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

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

k = total # diffs btw 2 genomes





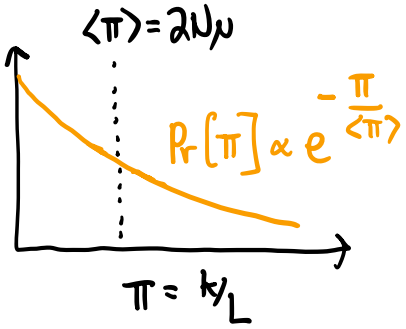
⇒ one advantage of coalescent approach:

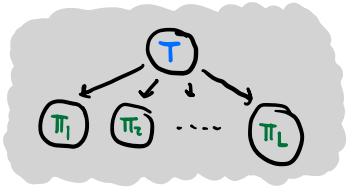
⇒ simple predictions for uncertainty in π (not just avg)

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

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

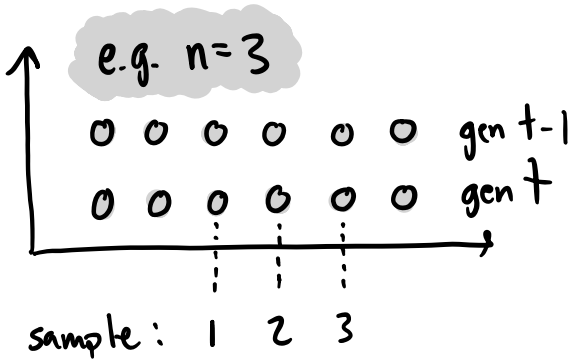
$\Rightarrow$ i.e. π does not <u>self-average</u> on a long <u>asexual</u> genome!





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

Larger sample sizes ($n > 2$)



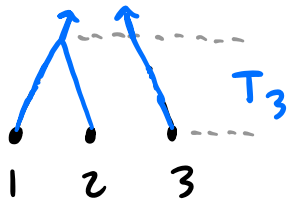
$\Rightarrow$ Prob that any 2 share ancestor is $\frac{1}{N}$ [$\times \binom{3}{2}$ pairs]

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

(known as "Kingman's coalescent")

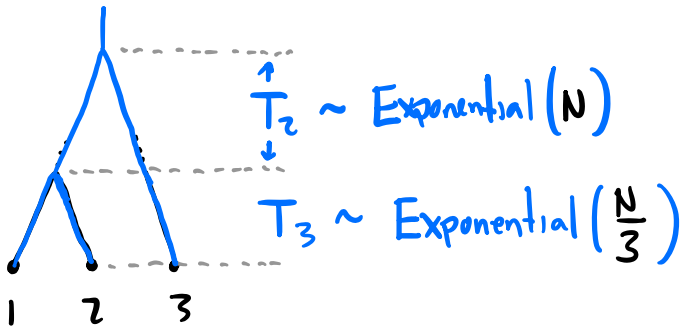
(all pairs are equally likely to coalesce)



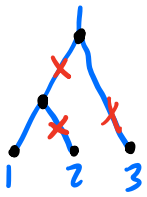
$\Rightarrow$ total prob of coalescence $= \frac{3}{N}$ per gen

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

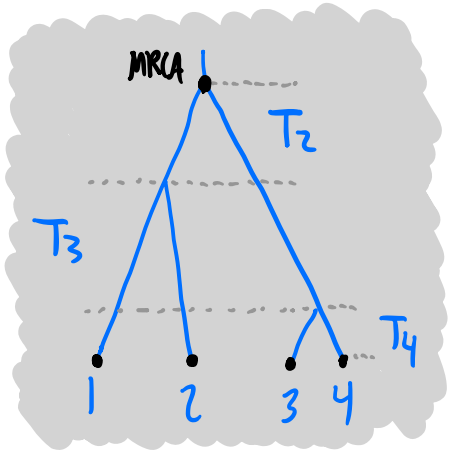
⇒ now we have sample of $n=2$... ⇒ repeat!



⇒ Done! can now paint on mutations...



Easily generalizes to sample of size n:



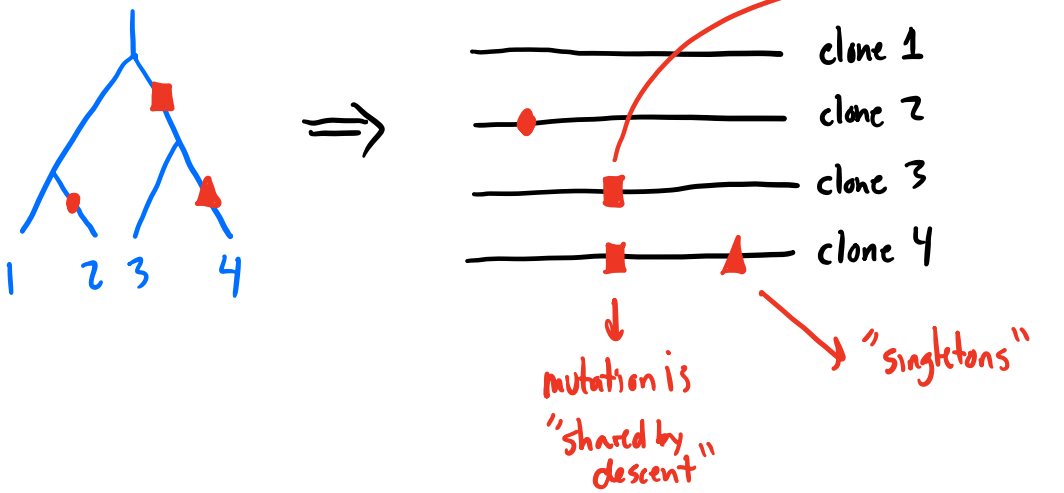
① @ each step, only consider coalescence between pairs of lineages

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

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

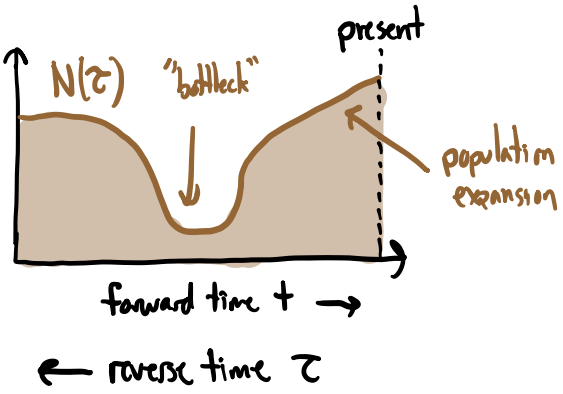
$\Rightarrow$ compare to single-locus prediction (easy!)

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

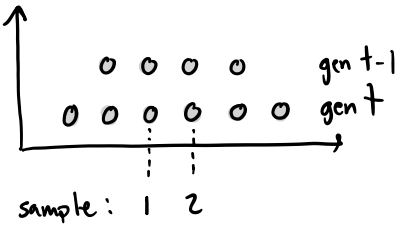
$\Rightarrow$ why use coalescent picture then??

Answer: coalescent picture makes it easy to model demography!

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



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

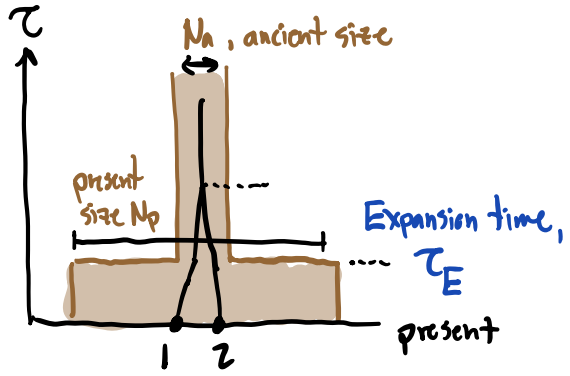


⇒ coalescence = "inhomogeneous" Poisson process:

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

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

Simple example: rapid expansion in recent past



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

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 forward time calc... $\frac{df}{dt} = \mu(1-f) - \nu f + \sqrt{\frac{f(1-f)}{N(t)}}\eta(t) \Rightarrow p(f,t)$)

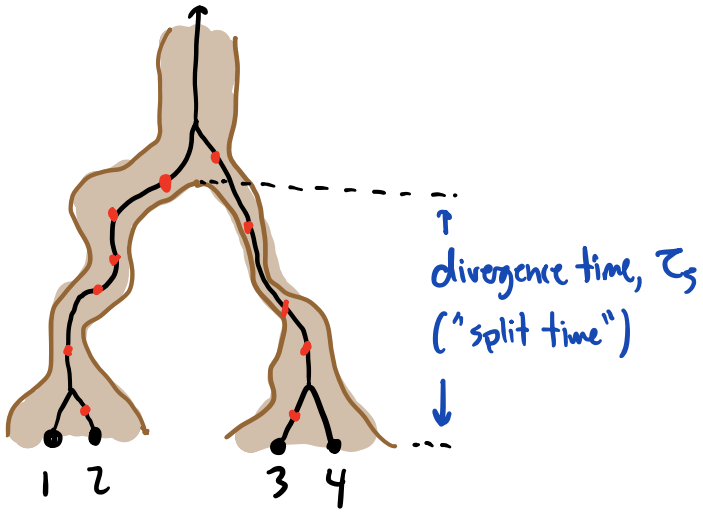
can revisit our earlier puzzle: if $N_p \cdot \mu \sim 100$ in humans

why $\langle \pi \rangle \sim 10^{-3}$?

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

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

Can also easily add population structure



divergence time, τ_s
("split time")

$\Rightarrow$ Pr(coalescence) btween pop'ns = 0 until time $t = \tau_s$

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

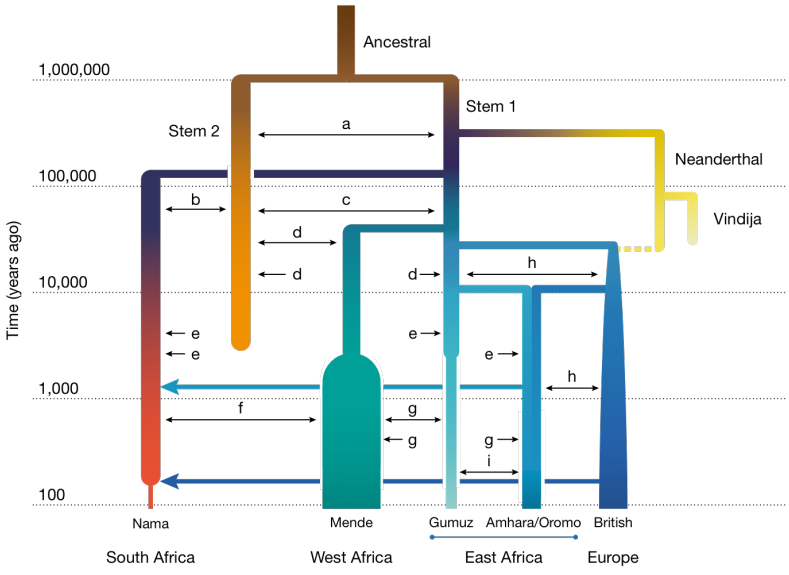
e.g.:

Article

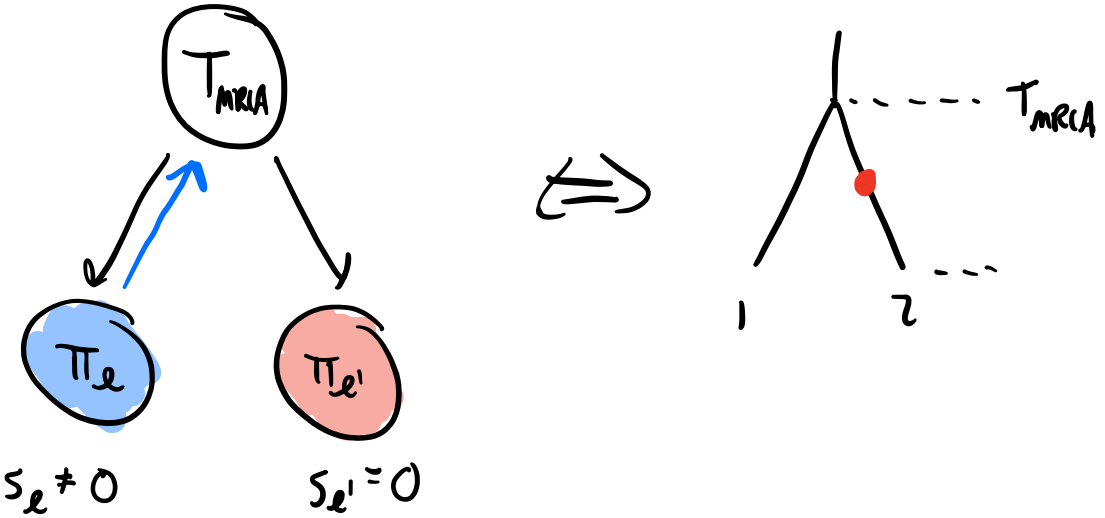
A weakly structured stem for human origins in Africa

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



⇒ downside: hard to add selection back in to picture...

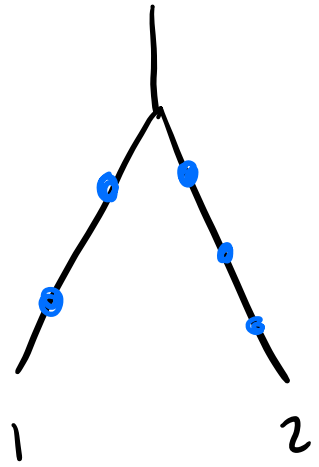


⇒ when is this going to be an issue?

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

⇒ for $L \gg 1$, selection looks like $[X(\vec{s}) - \bar{X}(t)] f(\vec{s})$ vs $sf(1-f)$ in $L=1$

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



① assume effective neutrality:

$\Rightarrow$ total # mutations $\sim NU$

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

$\Rightarrow$ self consistent:

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

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

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

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