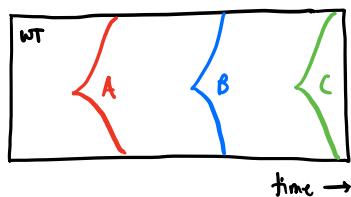


Chapter II

Neutral theory and the coalescent

Neutral theory & the Coalescent

Successive mutations:



$$\frac{dS(\vec{q})}{dt} = \sim(x - \bar{x}) + \sim L\mu \rightarrow e + \sim Q + \sim \frac{Z}{JN}$$

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

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

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

$$\frac{dS(\vec{q})}{dt} = \sim(x - \bar{x}) + \sim L\mu + \sim Q + \sim \frac{Z}{JN}$$

⇒ one other limit that's well understood:

neutral evolution in nonrecombining genome

when $X(\vec{g})=0$ & $e=0$, left with: $(\nu_e = \bar{\nu}_e)$

$$\frac{\partial f(\vec{g})}{\partial t} = \sum_{|\vec{g}' - \vec{g}|=1} \sum_e \mu_e f(\vec{g}') \left[g_e(1-g_e') + (1-g_e)g_e' \right] - \underbrace{\sum_e N_e f(\vec{g})}_{\text{outgoing mutations}}$$

incoming mutations

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

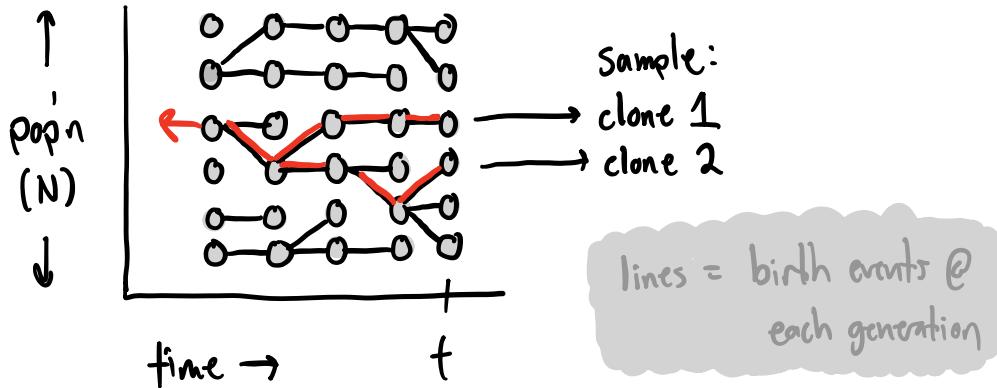
Key insight: sites don't actually influence each other (because neutral)

e.g. 

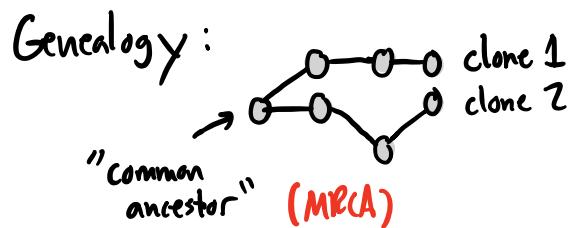
$$\Rightarrow \rho(f) = \frac{2N}{f}$$

\Rightarrow 2nd key insight: can take $L' = 0$ —

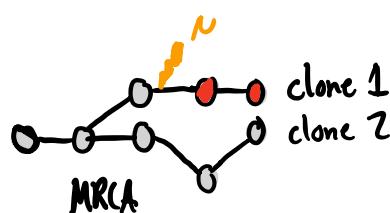
E.g. simulation of neutral pop'n in Wright-Fisher model:



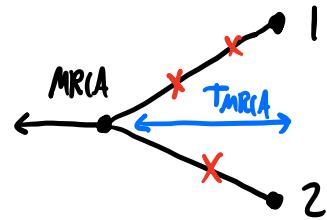
⇒ key insight: lines also = **genealogical relationships**
backward in time!



↓ differences between sampled individuals
= mutations on genealogy



\Rightarrow Mut's @ site $\ell \approx$ Poisson Process
w/ rate μ_e on each branch



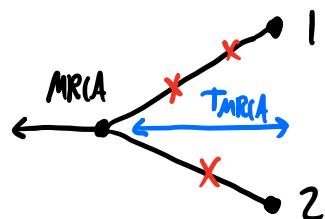
\hookrightarrow total # mut's
 $\sim \text{Poisson}(2T_{\text{MRCAs}}\mu_e)$

\Rightarrow 2 extreme limits:

(1) $\mu_e T_{\text{MRCAs}} \ll 1 \Rightarrow$ 0 or 1 mutations on whole tree

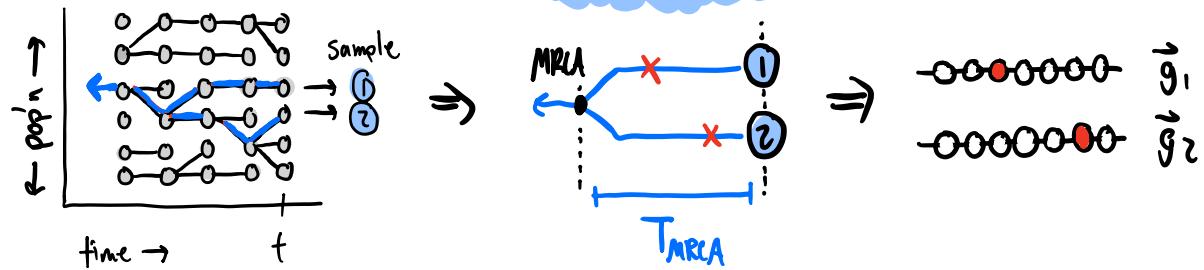
$$\Rightarrow \Pr[\text{genetic diff} @ \text{site } \ell \mid T_{\text{MRCAs}}] \approx$$

(2) $\mu_e T_{\text{MRCAs}} \gg 1 \Rightarrow$ lots of forward & backward mutations along each branch.



$$\Rightarrow \Pr[\text{genetic diff} @ \text{site } \ell \mid T_{\text{MRCAs}}] =$$

Recap:



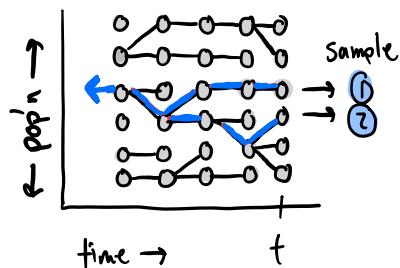
\Rightarrow Given genealogy (T_{MRCA}), mutations occur as Poisson Process along each branch ("mutation painting")

$$\Pr[\text{difference @ site } l \mid T_{\text{MRCA}}] \approx \begin{cases} 2\mu_e T_{\text{MRCA}} & \text{if } \mu T_{\text{MRCA}} \ll 1, \\ 1/2 & \text{else.} \end{cases}$$

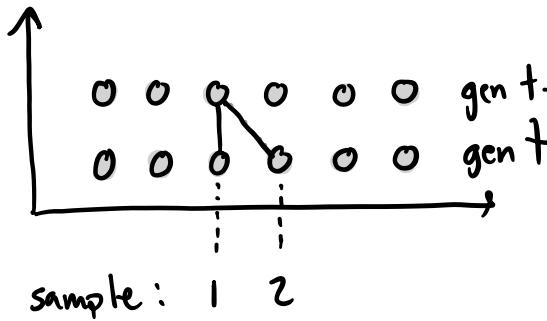
Question: what determines genealogy (T_{MRCA})?

\Rightarrow Note: T_{MRCA} is random quantity

(genealogy will vary from
Sample-to-Sample &
Simulation-to-Simulation...)



\Rightarrow key insight: start from present & work backward in time:



"coalesced"

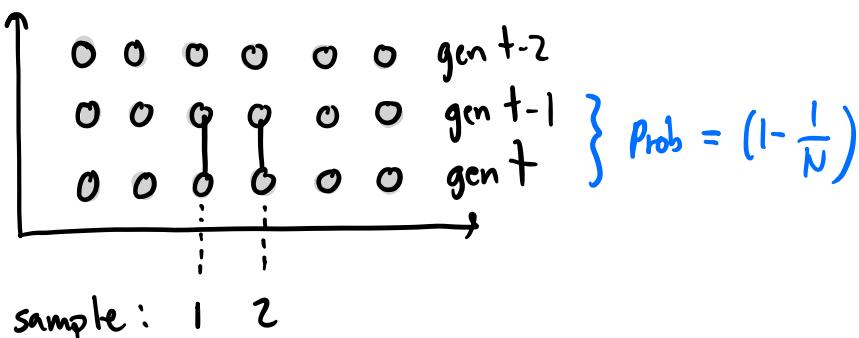
\Rightarrow Two individuals share ancestor in previous gen w/ probability:

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

$\underbrace{}$ prob that both draw same
 \downarrow # possible ancestors

\Rightarrow w/ probability $\frac{1}{N}$ $\Rightarrow T_{\text{MRCA}} = 1$

\Rightarrow otherwise, diff ancestors in gen $t-1 \Rightarrow$ repeat!



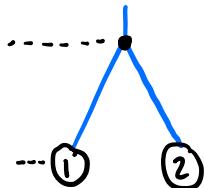
Process repeats itself w/ next gen:

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

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

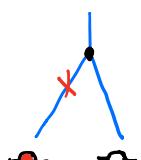
\Rightarrow coalescence is also a Poisson Process w/ rate $\frac{1}{N}$!

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



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

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

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

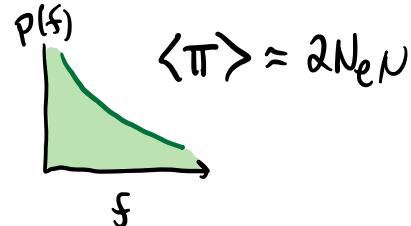
$$\underset{(NT \ll 1)}{\approx} \int 2N_e T_{\text{MRCA}} \cdot p(T_{\text{MRCA}}) dT_{\text{MRCA}}$$

$$= 2N_e \langle T_{\text{MRCA}} \rangle$$

$$= 2N\mu_e$$

\Rightarrow matches our previous result

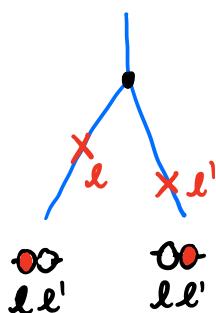
since $\langle \pi \rangle \equiv \Pr(\text{difference} @ \text{site } e)$



\Rightarrow Distribution of T_{MRCA} becomes more important

when considering mutations @ multiple sites, e.g.

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



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

$$= \int (2N_e T_{\text{MRCA}}) \cdot (2\mu_e T_{\text{MRCA}}) \cdot p(T_{\text{MRCA}}) \cdot dT_{\text{MRCA}}$$

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

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

$$= 2 \cdot \Pr(\pi_e) \cdot \Pr(\pi_{e'}) \geq \Pr(\pi_e) \Pr(\pi_{e'})$$

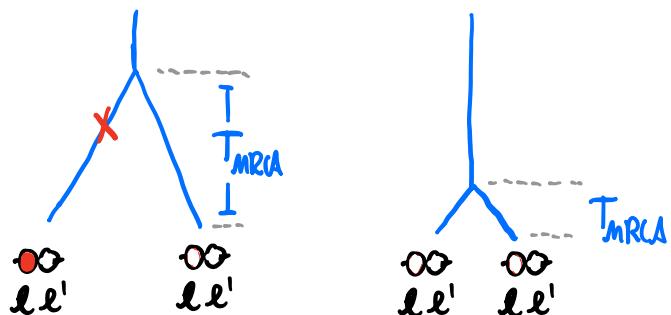
Upshot: joint prob of mut's is not independent:

$$\Pr(\pi_{e'} = 1 \mid \pi_e = 1) = \frac{\Pr(\pi_e = 1, \pi_{e'} = 1)}{\Pr(\pi_e)} = 2 \Pr(\pi_{e'} = 1)$$

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

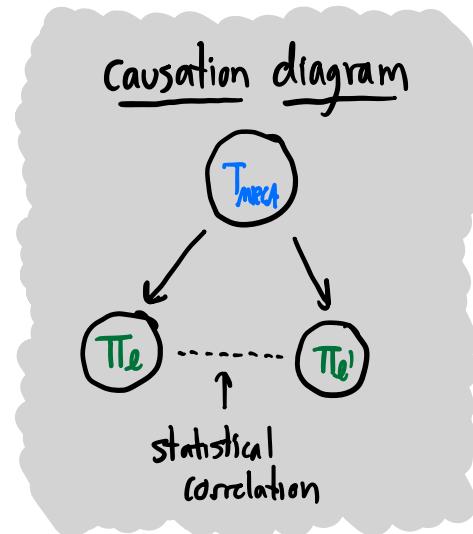
\Rightarrow what's going on?

\Rightarrow consider 2 trees:

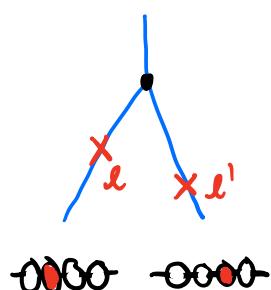
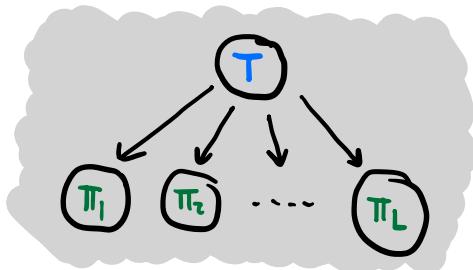


\Rightarrow conditioned on $\pi_e = 1$, likely had bigger-than-avg T_{MRCA}

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



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



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

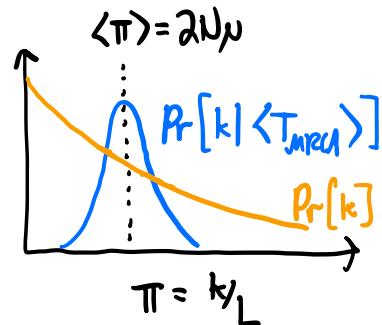
\Rightarrow total # mut's (k) is Poisson Process w/ rate $U \equiv \sum_{e=1}^L \mu_e$

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

$$\begin{aligned} \Rightarrow \Pr[k] &= \int \Pr[k | T_{\text{MRCA}}] \rho(T_{\text{MRCA}}) dT_{\text{MRCA}} \\ &= \int \frac{(2^U T)^k}{k!} e^{-2^U T} \frac{1}{N} e^{-T/N} dT \end{aligned}$$

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

total # diffs ---o---o---
 btw 2 genomes ---o---o---o---



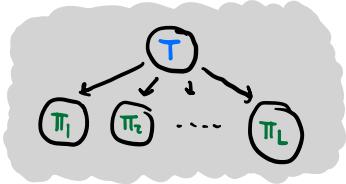
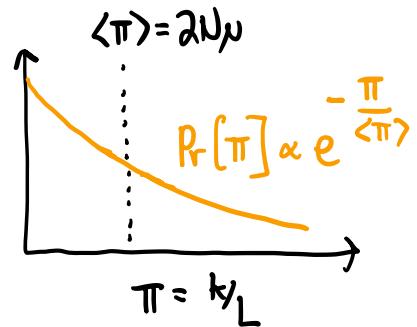
\Rightarrow one advantage of coalescent approach :

\Rightarrow simple predictions for uncertainty in π (not just avg)

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

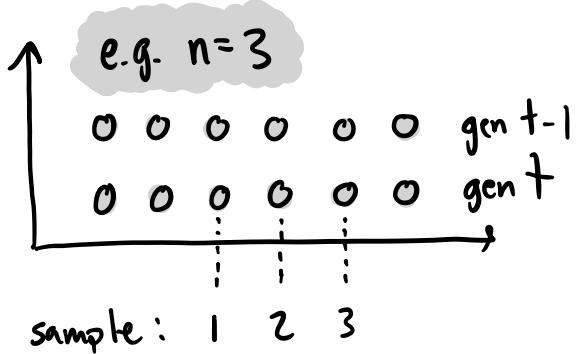
$$\Rightarrow \text{ or } C_V^2 \equiv \frac{\text{Var}(\pi)}{\langle \pi \rangle^2} = \frac{1+2NU}{2NU} \geq 1$$

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



\Rightarrow fluctuations in T_{MRCA} affect many sites!

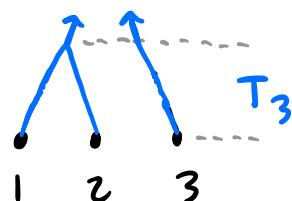
Larger sample sizes ($n > 2$)



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

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

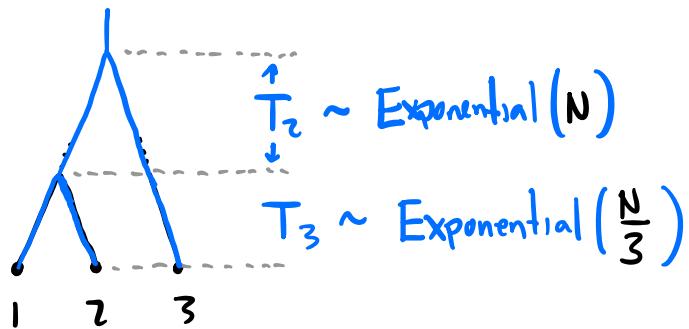
\Rightarrow when $N \gg 1 \rightarrow$ only need to worry about **pairwise coalescence**
 (known as "Kingman's coalescent")
 (all pairs are equally likely to coalesce)



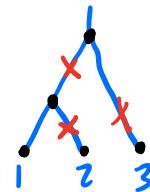
\Rightarrow total prob of coalescence = $\frac{3}{N}$ per gen

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

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

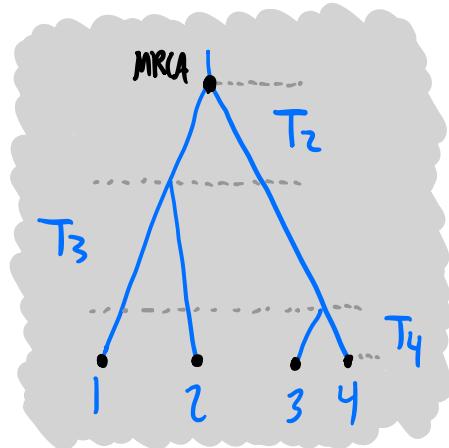


\Rightarrow Done! can now paint on mutations...

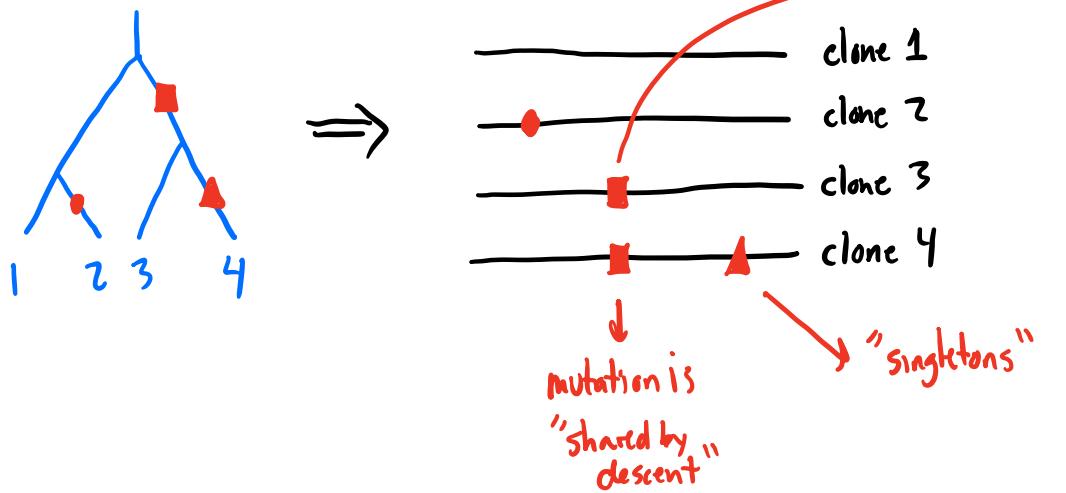


Easily generalizes to sample of size n:

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



- ④ then can paint mutations on @ end:



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

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

\Rightarrow must avg over:

- ① tree topologies
- ② branch lengths | topology
- ③ mutation painting | branch lengths

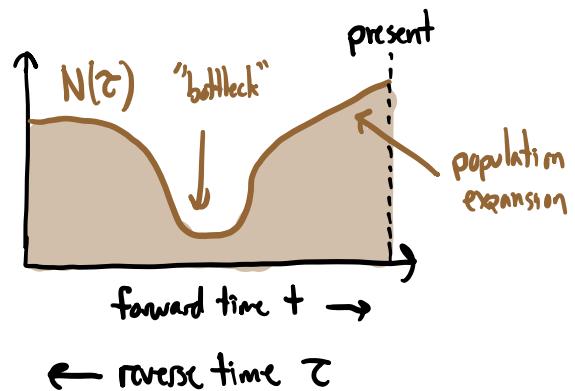
\Rightarrow compare to single-locus prediction (easy!)

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

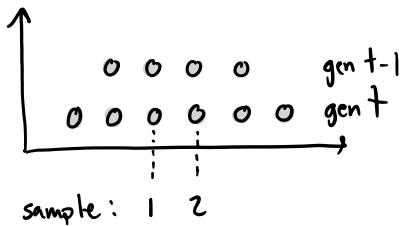
\Rightarrow why use coalescent picture then??

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 $\rightarrow Y_{N(\tau)}$

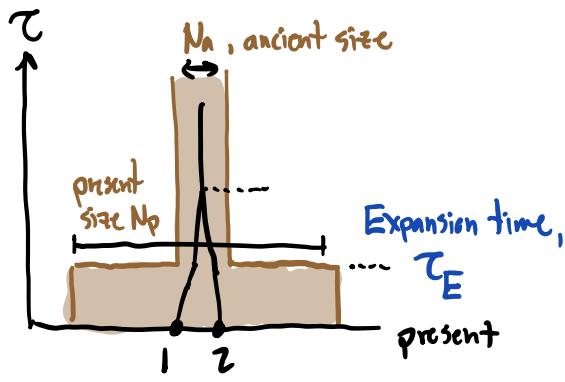


⇒ Coalescence = "inhomogeneous" Poisson process:

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

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

Simple example: rapid expansion in recent past



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

- ① no coalescence until τ_E
 - ② coalescence @ rate $\frac{1}{N_a}$ after
- $$\Rightarrow \langle \tau_2 \rangle = \tau_E + N_a$$

$$\Rightarrow \langle \pi \rangle = 2N \langle \tau_2 \rangle = 2N(\tau_E + N_a) \approx 2NN_a \quad (\text{if } \tau_E \ll N_a)$$

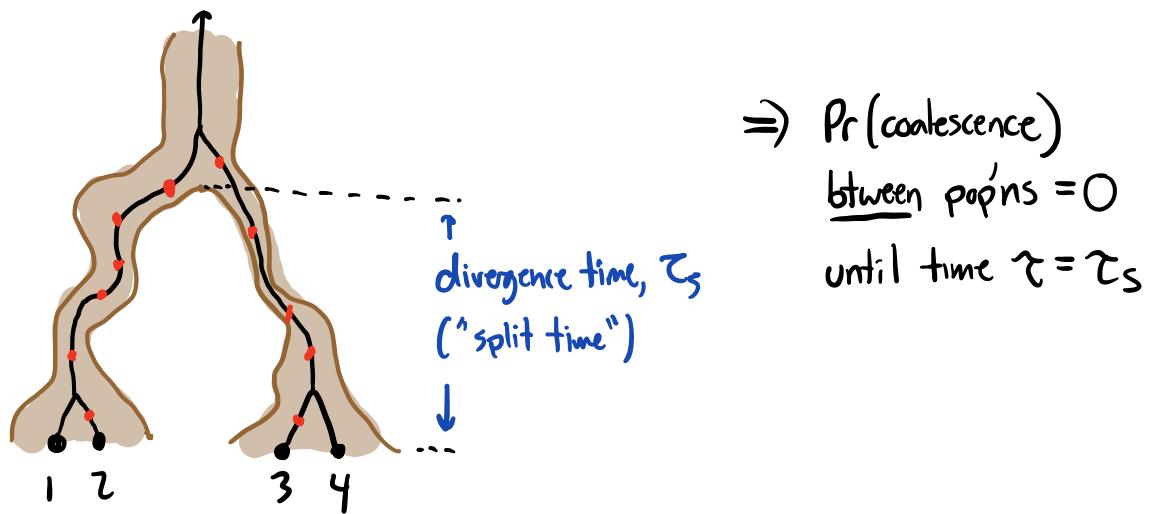
(Compare to forward time calc... $\frac{df}{dt} = N(1-f) - vf + \sqrt{\frac{f(1-f)}{N(t)}} \eta(t) \Rightarrow p(f,t)$)

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



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

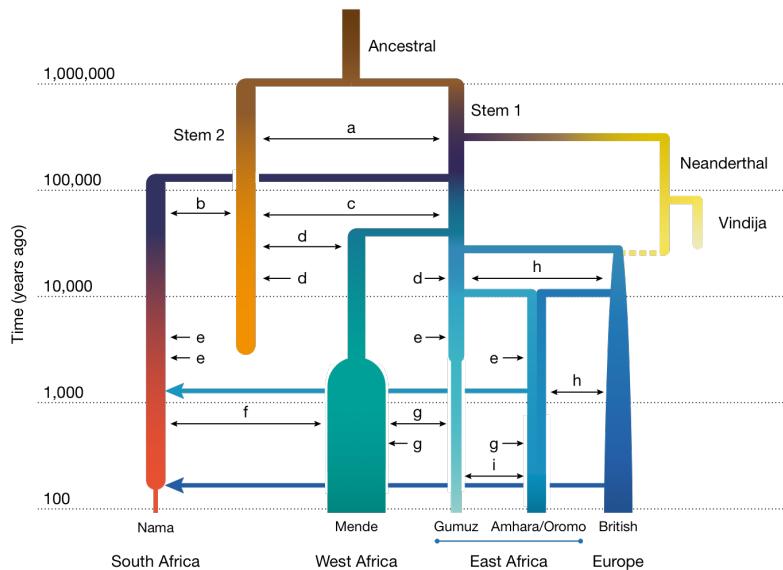
e.g. :

Article

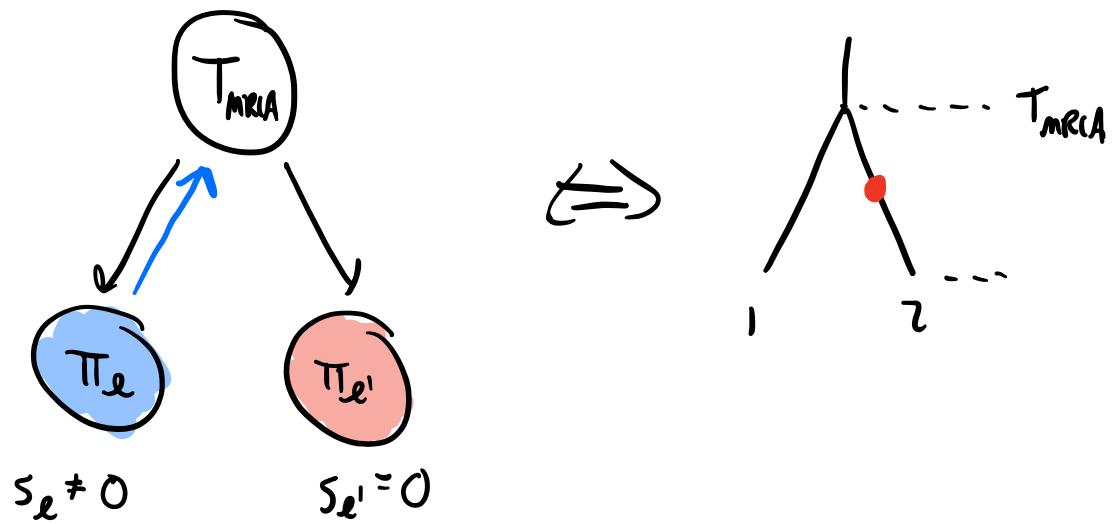
A weakly structured stem for human origins in Africa

Aaron P. Ragsdale¹, Timothy D. Weaver², Elizabeth G. Atkinson³, Eileen G. Hoal^{4,5,6},
 Marlo Möller^{4,5,6}, Brenna M. Henn^{2,7,8} & Simon Gravel⁸

Published online: 17 May 2023



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

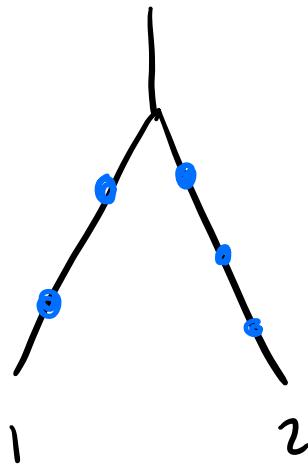


\Rightarrow when is this going to be an issue?

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

\Rightarrow for $L \gg 1$, selection looks like $\left(\bar{x}(\vec{s}) - \bar{x}(t) \right) f(\vec{s})$
 vs
 $sf(1-f)$ in $L=1$

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



① assume effective neutrality:

\Rightarrow total # mutations $\approx N_U$

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

\Downarrow self consistent:

$$\boxed{(N_U)(N_S)^2 \ll 1}$$

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

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

$$\downarrow \quad \sqrt{10^4 \cdot (10^{-1})^2} = 10 \rightarrow 1$$