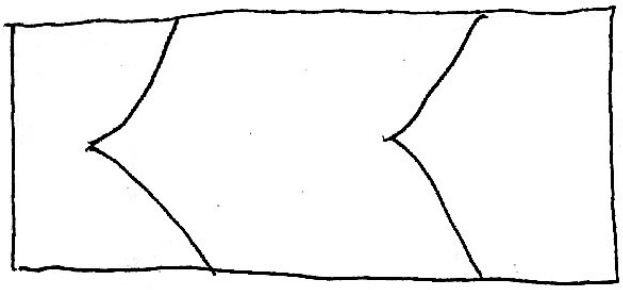
Neutral Theory and the Coalescent

In the last lecture, we discussed the behavior of multi-locus models of evolution in the successive mutations regime



where only ~1 genetic variant is present @ high frequencies @ any given time.

In this case, behavior was exactly solvable because it reduced to an effective single locus model.

In general, genomes in data are separated by multiple mutations (e.g. humans, ~ 1 mutation every 1000bp between 2 individuals)

$\Rightarrow$ we need to understand what's going on in these cases.

today, we'll focus on one limit that is very well understood: neutral evolution on a nonrecombining genome.

it might seem unrealistic (and it might be unrealistic, @ least for bacteria), but the neutral model is extremely influential in how people have come to think about & talk about genetic data, so it's worth being aware of it.

one motivation for considering this limit is that selection & recombination contribute nonlinear terms to the full multi-locus SDE we wrote down last class. Setting $X(\vec{g}) = \text{const}$ & $e=0$ leads to a linear system, and we'd expect that linear systems should be possible to analyze (@ least in principle). At first glance, linear system is still pretty complicated:

$$\frac{df(\vec{g})}{dt} = \left[\sum_{\substack{\vec{g}' \text{ s.t.} \\ |\vec{g}'-\vec{g}|=1}} \sum_{e=1}^{L} \mu_e f(\vec{g}') \left[g_e(1-g_e') + (1-g_e) g_e' \right] - \right] - f(\vec{g}) \sum_{e=1}^{L} \mu_e$$

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

insight comes from realizing that the sites don't actually influence each other (because they are neutral).

⇒ i.e., we could focus on a subset of the genome L', and write down a corresponding neutral model for the genotypes @ just these loci, and it would have the same form.

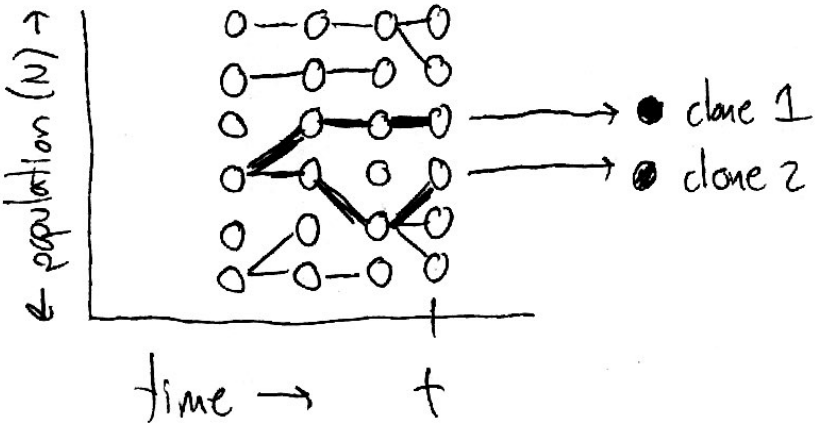
⇒ we saw an extreme form of this a few lectures ago where we took $L'=1$, and saw that the behavior of a single site in a long neutral genome is indistinguishable from a single-locus model.

⇒ thus, in a similar way as in the SSWM limit, we might be able to understand what's going on due to some effective reduction to a single locus behavior that we already understand

⇒ in this case, the true insight comes from going one step further and considering a zero locus model, (that is, taking out mutations entirely.)

to explain what I mean, suppose we simulate a neutral population in a Wright fisher model, and we sample 2 random individuals from the population @ the end.

If we examined the guts of the simulation, it would look like:

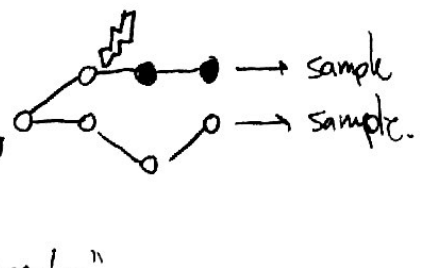


where lines illustrate birth events that form next generation.

going backward in time, lines also illustrate genealogical relationships between the sampled (or unsampled individuals).

In this case, ● ● the 2 clones would have a genetic difference @ a particular site if one of the divisions along the genealogy resulted in a mutation, before the common ancestor!

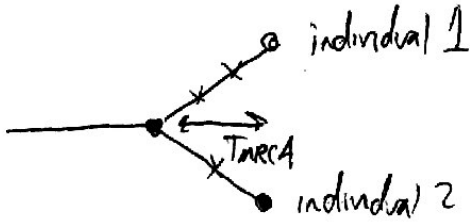
e.g.



* since mutations are neutral, they can have no influence on this hypothetical genealogy itself.

$\Rightarrow$ we are free to "paint" them on after the fact.

E.g. if 2 individuals shared a common ancestor T_{MRCA} generations ago, then mutations occur as a Poisson process @ rate μ_e on each branch (length T_{MRCA})



$\Rightarrow$ 2 extreme limits:

(1) $\mu_e T_{MRCA} \ll 1 \Rightarrow$ 0 or 1 mutations along whole tree.

$$\Pr\left[\begin{matrix}\text{genetic difference} \\ @ \text{ site } e\end{matrix}\right] = 1-(1-\mu_e)^{2T_{MRCA}} \approx 2\mu_e T_{MRCA}$$

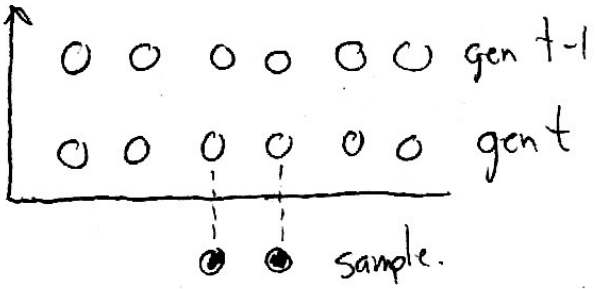
(2) $\mu_e T_{MRCA} \Rightarrow$ lots of forward & reverse mutations along each branch of tree.

$$\Pr\left[\begin{matrix}\text{genetic difference} \\ @ \text{ site } L\end{matrix}\right] = \frac{1}{2}$$

$\Rightarrow$ seems pretty straight forward. question is what sets T_{MRCA}?

In principle, this is a random quantity, since genealogies in W.F. simulation will vary if you re-run the simulation.

$\Rightarrow$ Suppose we start from present day population & work back in time:



all individuals in gen t must have some ancestor in previous generation. $\Rightarrow$ in WF model, these ancestors are chosen uniformly @ random from previous generation (w/ replacement.)

$\Rightarrow$ probability that two individuals share a common ancestor is $1/N$ $\Rightarrow$ in this case, we say they have "coalesced"

$\Rightarrow$ w/ probability $\frac{1}{N}$, $T_{MRCA} = 1$ ← also known as "coalescence time"

$\Rightarrow$ w/ probability $(1-\frac{1}{N})$, 2 individuals descend from different parents in previous generation.

then process repeats itself:

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

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

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

i.e., $\langle T_{MRCA} \rangle = N$, $\sqrt{\text{Var}(T_{MRCA})} = N$

$\Rightarrow$ the total probability of observing a mutation @ site ℓ can then be obtained by integrating over T_{MRCA}:

$$\Pr\left[\begin{smallmatrix}\text{difference}\\ @\ \text{site}\ \ell\end{smallmatrix}\right] = \int \Pr(\text{diff} \mid T_{MRCA})\, p(T_{MRCA})\, dT_{MRCA}$$

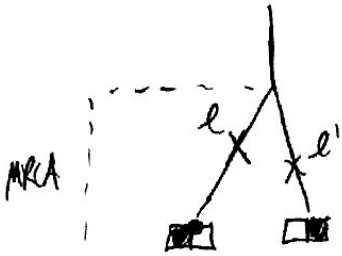
$$(\text{when } N_e \text{ const}) \simeq \int 2\mu_\ell T_{MRCA}\, p(T_{MRCA})\, dT_{MRCA} = 2\mu_\ell \langle T_{MRCA} \rangle = 2\mu_\ell N$$

↳ this matches w/ what we derived for $\langle \pi \rangle$ before, since

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

distribution of T_{MRCA} becomes important when we consider mutations @ multiple sites. e.g.

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



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

mutations are neutral, so can't affect each other.

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

$$= 2 \cdot \left(2\mu_\ell N\right) \cdot \left(2\mu_{\ell'} N\right) \quad \left[\text{since } \langle T^2 \rangle = 2N\right]$$

$$= 2 \cdot \Pr\left[\pi_\ell = 1\right] \Pr\left[\pi_{\ell'} = 1\right] \geq \Pr\left[\pi_\ell = 1\right] \Pr\left(\pi_{\ell'} = 1\right)$$

$\Rightarrow$ joint probability of having a mutation at both sites is <u>not</u> independent

$$\Rightarrow \text{Cov}\left(\pi_\ell, \pi_{\ell'}\right) = \frac{\langle \pi_\ell \pi_{\ell'} \rangle - \langle \pi_\ell \rangle \langle \pi_{\ell'} \rangle}{\langle \pi_\ell \rangle \langle \pi_{\ell'} \rangle} = 1$$

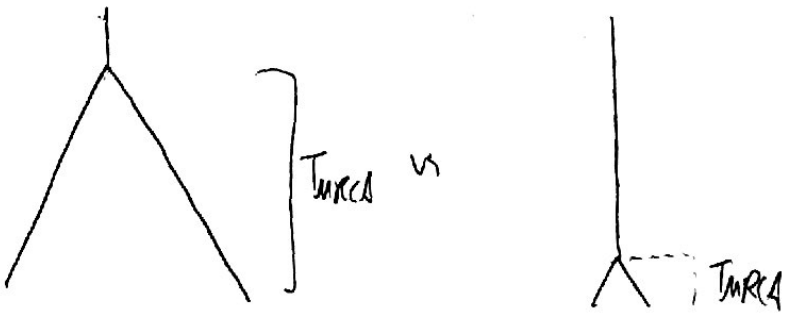
twice as likely!

$\Rightarrow$ knowing that you have a mutation @ site ℓ makes it more likely to observe a mutation @ site ℓ'!

But previously said that mutations don't influence other directly...

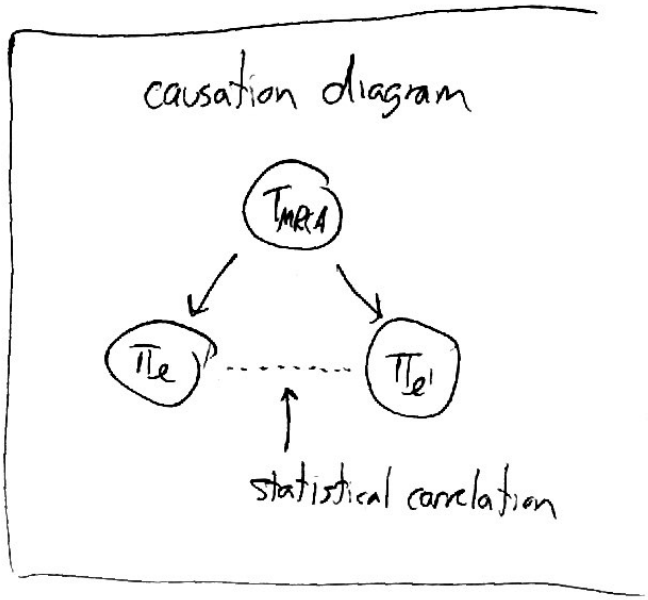
⇒ what's going on?

⇒ mutations more likely to occur when T_{MRCA} is bigger



⇒ conditioned on $\pi_\ell = 1$, probably had a bigger than avg T_{MRCA} so more likely to have mutation @ site ℓ' too.

i.e., mutation processes do not interact, but are still coupled together through random genealogy that happens to be present in a given population



⇒ can keep adding more sites in this way. In limit that $\mu_e T_{MRCA} \ll 1$, then most mutations will occur @ a unique site in the genome. ("infinite sites" assumption). then conditioned on T_{MRCA}, [illegible] the total # of mutations occurs as a Poisson process w/ rate $U \equiv \sum_{\ell=1}^{L} N_e$ per gen.

i.e., if k = total # of mutational differences between the two individuals, then $\Pr[k|T_{MRCA}] = \frac{(2UT_{MRCA})^k}{k!} e^{-2UT_{MRCA}}$

$$\Pr[k] = \int \Pr[k|T_{MRCA}] P(T_{MRCA})\, dT_{MRCA} = \int \frac{(2UT_{MRCA})^k}{k!} \frac{1}{N} e^{-(2U+\frac{1}{N})T_{MRCA}} dT_{MRCA}$$

$$= \frac{(2NU)^k}{(2NU+1)^{k+1}} \Rightarrow \text{geometric distribution w/ prob } \frac{2NU}{1+2NU}.$$

(again, broader than $\Pr[k|T_{MRCA}]$)

$\Rightarrow$ So one advantage of coalescent approach is that it provides simple predictions for uncertainty in π rather than just $\langle \pi \rangle$.

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

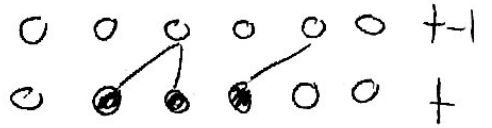
or $CV \equiv \frac{\mathrm{Var}(\pi)}{\langle \pi \rangle^2} = \frac{(1+2NU)}{2NU} \geq 1$

i.e., π does not self average, on a long a sexual genome.

$\Rightarrow$ fluctuations in T_{MRCA} influence large #s of sites simultaneously.

can also consider sample sizes larger than 2.

e.g. for n=3:



prob that any 2 share a common ancestor in previous generation is $\frac{1}{N}$

$\Rightarrow \binom{3}{2} = 3$ total pairs.

$\Rightarrow$ probability that <u>all 3</u> share an ancestor in previous generation is $N \cdot \left(\frac{1}{N}\right)\left(\frac{1}{N}\right)\left(\frac{1}{N}\right) = \frac{1}{N^2} \ll \frac{1}{N}$ (when $N \gg 1$)

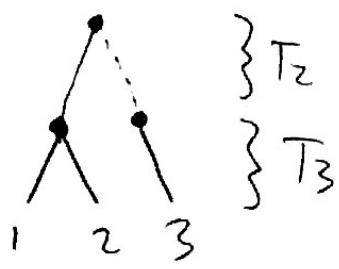
- N: # of ancestors possible
- $\left(\frac{1}{N}\right)$: prob that each one chooses that [redacted] ancestor.

$\Rightarrow$ only have to worry about <u>pairwise coalescence</u> (known as "kingman coalescent") (all pairs equally likely to coalesce)

[redacted] total # of pairs is 3, so total probability of getting a coalescence blw a pair is $3/N$.

$\Rightarrow$ time until this happens is $T_3 \equiv$ Exponential$\left([redacted]\frac{N}{3}\right)$

$\Rightarrow$ once this happens, choose random pair to coalesce. then have effective sample of n=2.



$\Rightarrow$ time till coalescence of these is

[redacted] $T_2 \equiv$ Exponential(N) as before

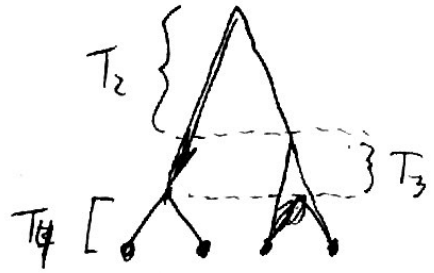
at this point, all share common ancestor, so done!

$\Rightarrow$ can then paint mutations on as before

$\Rightarrow$ easily generalizes to sample of size n:

@ each step, need only consider coalescence btw pairs of individuals $\Rightarrow$ time until next coalescence event = $T_n \equiv \text{Exponential}\left(\frac{N}{\binom{n}{2}}\right)$

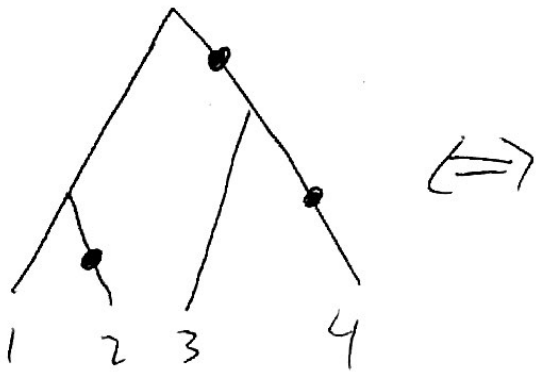
$\Rightarrow$ once you draw T_n, choose a random pair to coalesce. then repeat w/ sample of size n-1. $\longrightarrow$ n=2.



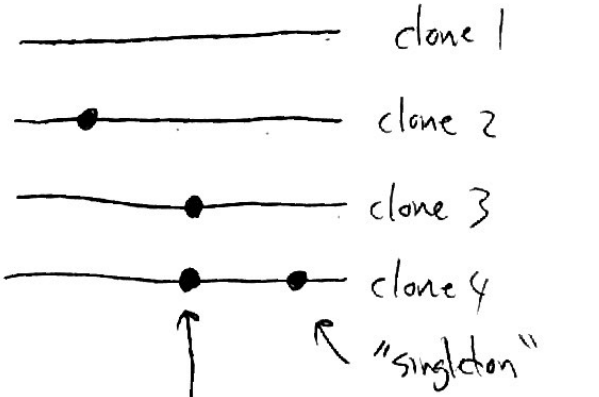
$\Rightarrow$ then can paint mutations on @ end.

$\Rightarrow$ mutations that occur higher up in tree are present in all individuals that inherit from that branch.

e.g.



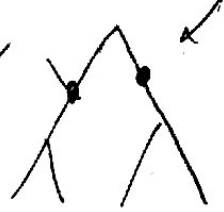
$\Longleftrightarrow$



mutation shared by descent

(only one mutation event)

easy to simulate this process, for any n ("coalescent simulations") but much harder to calculate anything analytically for $n \geq 4$.

e.g. $\langle$ # of doubletons in sample of size 4 $\rangle$ = ( + )

all trees that look like this

i.e. avg over mutations occurring
+ branch lengths
+ topologies

$\Rightarrow$ gets very hard for $n \geq 4$, even when $n \to \infty$.

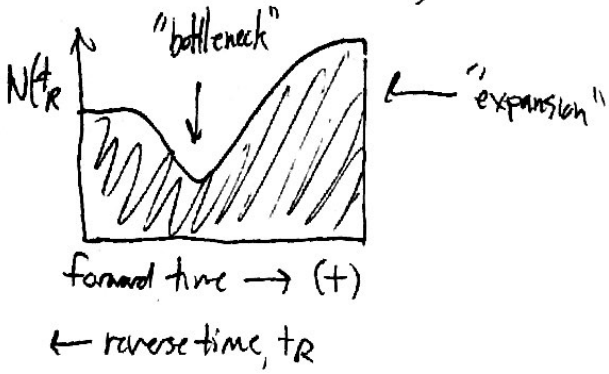
$\Rightarrow$ compare to single locus prediction:

$$\left\langle \begin{matrix} \# \\ \text{doubletons} \end{matrix} \text{ in sample of size 4} \right\rangle = \int [\text{illegible}] \cdot \binom{4}{2} f^2 (1-f)^{4-2} \cdot \frac{2N\mu}{f} \, df$$

$$= \frac{N\mu}{5} \quad \text{(easy!)}$$

why would we ever use coalescent picture then?

$\Rightarrow$ answer is that coalescent picture makes it very easy to model demography. e.g. if N was not constant, but varied historically back in time, $N(t_R)$



then coalescence picture still works, except that now coalescence probability changes in each generation, $\frac{1}{N(t_R)}$

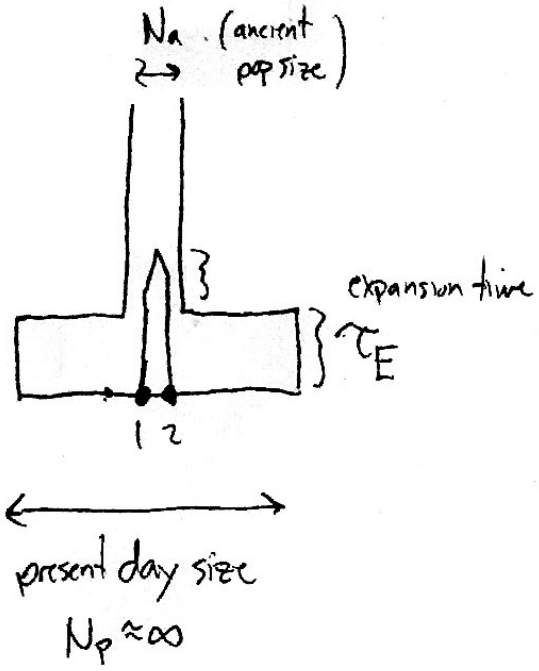
$\Rightarrow$ time to coalescence is inhomogeneous poisson process:

e.g.  $p(T_{\bullet n} = t) = \frac{\binom{n}{2}}{N(t)} e^{-\int_0^t \frac{\binom{n}{2}}{N(t')} dt'}$

$\Rightarrow$ if N is larger $\Rightarrow$ less likely to coalesce.
if N smaller $\Rightarrow$ more likely to coalesce.

$\Rightarrow$ otherwise, everything is the same (same topologies, same mutation painting)

e.g. simple case: rapid expansion from $N = N_a$ to  T_e generations ago. $N = N_p \approx \infty$



$\Rightarrow$ no coalescence for first τ_E generations (assuming $\tau_E \ll N_p$)

$\Rightarrow$ then regular coalescence @ rate $\frac{1}{N_a}$ afterwards.

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

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

- Hence, in this case $\langle \pi \rangle$ is dominated by historical (ancestral) population size N_a, and has nothing to do w/ population size now.

  ↖ sometimes people will call this effective pop size (N_e). → don't!

$\Rightarrow$ addresses key paradox we had w/ human data @ beginning of class. if $N\mu \gg 1$ why $\pi \sim 10^{-3}$?

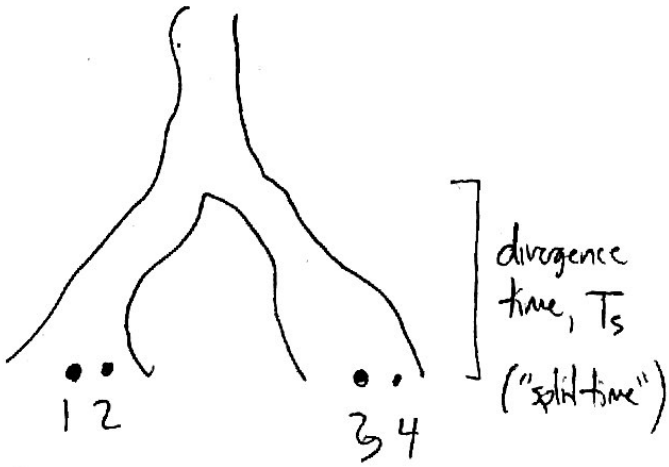
$\Rightarrow$ commonly accepted answer is that $N(t)$ was smaller in back in time (e.g. out of africa).

$\Rightarrow$ genetic data let's us measure $N(t)$.

coalescent picture makes this much easier than corresponding forward time calculation:

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

$\Rightarrow$ can also easily add in population structure, e.g. isolated subpopulations:



$\Rightarrow$ prob coalescence between pop'ns $\propto 0$ until T_s.

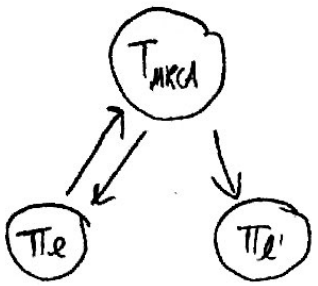
$\Rightarrow$ much of human pop gen is about infering these population demographic models (much fancier of course)

[illegible]

downside is that it is very hard to add selection back in to this picture.

$\Rightarrow$ basic problem is causation diagram gets reversed:

[illegible]



$\Rightarrow$ there still is a genealogy... so can still paint on neutral mutations (if you know T_{MRCA})

$\Rightarrow$ but can't paint on selected mutations & worse... don't know what T_{MRCA} dist'n is any more!

$\Rightarrow$ this scenario is called "linked selection". will revisit in a later lecture.

when is this a problem?

i.e. when is neutral limit a good approx?

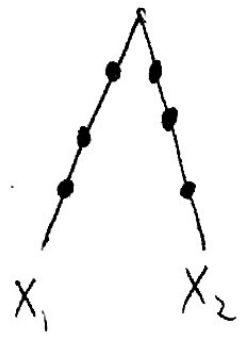
$\Rightarrow$ @ least need $Ns_e \ll 1$ (as in single locus case)

$\Rightarrow$ but since looking @ multi-locus problem, need

$$|X(\vec{s}) - X(\vec{s}')| \cdot N \ll 1$$

$\Rightarrow$ can we estimate these typical fitness differences using a self consistency argument?

$\Rightarrow$ if neutral model is good approx, then
mutations btw individuals is $\sim 2N\langle T_2 \rangle L$



$\Rightarrow$ mutations occur on different branches equally so tend to cancel each other out.

$\Rightarrow$ if each one has fitness effect $\pm s$.

then $X_1 - X_2 \sim \pm\sqrt{2\underbrace{\mu\langle T_2\rangle}_{\pi} \cdot L \cdot s^2}$ (from C.L.T.)

so neutral model is good approx if $(NU)(Ns)^2 \ll 1$

$\Rightarrow$ if $NU \gg 1$, can be bad even if Ns is very small!

e.g. if $Ns = 0.1$ (hard to select on individually)

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

then genome wide ΔX's are big!

$(10^4 \times 10^{-2} = 10^2 \gg 1!)$

$\Rightarrow$ in this case, we'll have to turn to alternative methods to understand evolution in a long genome...