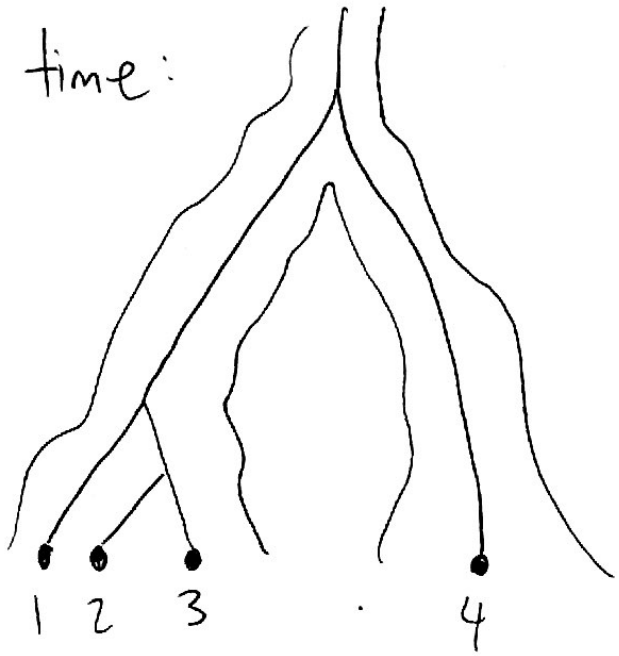
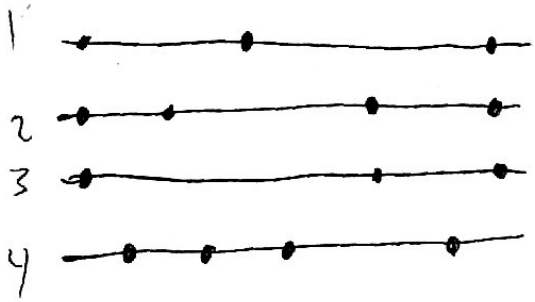
Coalescent theory II : selection & recombination

Last time:





mutations: poisson process w/ rate U.

we introduced a backward-in-time picture (coalescent theory) for describing sequences from a long neutral & non-recombing genome.

⇒ powerful because it separated genealogy of sample from mutations that occur conditioned on genealogy.

genealogy: kingman coalescent

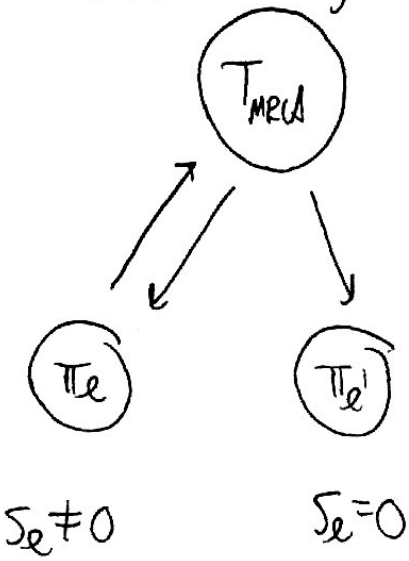
$p_c \sim 1/N(t)$

⇒ lots of use in pop. gen of higher organisms.

downside is that very hard to get selection back into picture.

⇒ basic problem is that causation diagram gets reversed:

causation diagram



genealogy

* there still is a genealogy between individuals in sample, so can paint on neutral mutations if you know TMRCA.

(i.e., neutral mutations ≈ time)

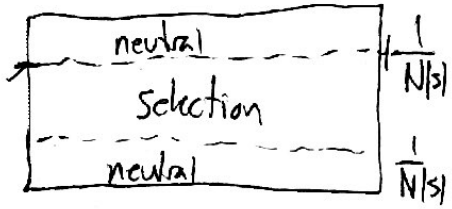
$\Rightarrow$ but can't paint on selected mutations in this way

$\Rightarrow$ even worse, no longer know how to calculate T_{MRCA} because it will depend on which mutations are there.

(catch-22)

$\Rightarrow$ when is this likely to be a problem? let's try to estimate.

$\Rightarrow$ remember for single site,



$\Rightarrow$ needed $N|s| \ll 1$ for neutral.

for longer genome, selection term is now $\left[x(\vec{g}) - \bar{x}(t)\right] f(\vec{g})$

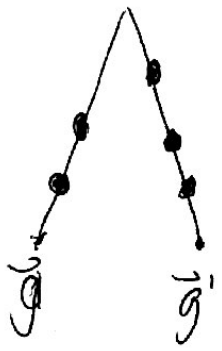
$\hookrightarrow \bar{x}(t) = \sum_{\vec{g}} x(\vec{g}) f(\vec{g})$

$\Rightarrow$ suggests that need $|X(\vec{g}) - X(\vec{g}')| \ll \frac{1}{N}$ for most pairs of genomes

can we estimate this difference w/ a self consistency argument?

$\Rightarrow$ if neutral limit is good approx, then # mutations between individuals is

$\sim 2U\langle T_2 \rangle$



but mutations will occur on branches equally, & tend to cancel out. (Central limit theorem).

$\Rightarrow$ if each mutation has effect $\pm s$, then

$$|X(\vec{g}) - X(\vec{g}')| \sim \sqrt{2U\langle T_2 \rangle s^2} \sim \sqrt{NUs^2}$$

$\Rightarrow$ neutral model is good approx if $\sqrt{NUs^2} \ll \frac{1}{N}$

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

$\Rightarrow$ if $NU \gg 1$, can be violated even if $Ns \ll 1$

e.g. if $Ns \approx 0.1$ (nearly neutral in single-locus setting)

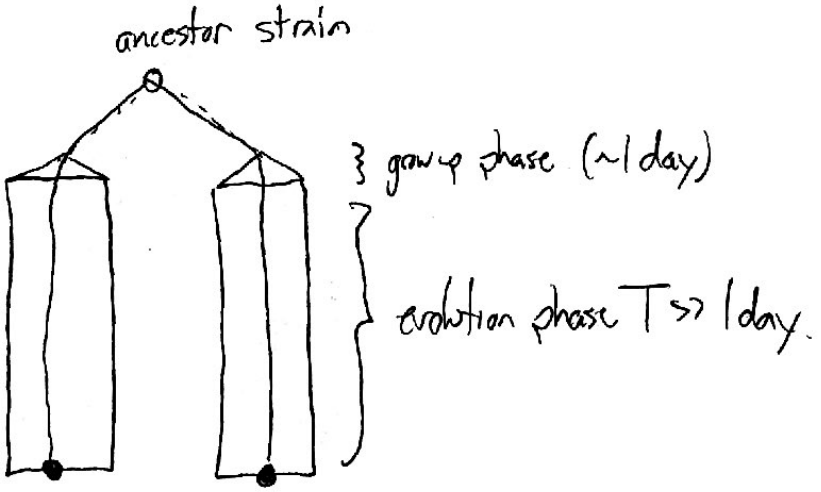
and $NU = \langle \pi \rangle L \approx \begin{cases} 10^4 & \text{for bacteria in gut} \\ 10^6 & \text{for humans} \end{cases}$ (ignoring recombination, for now)

$\Rightarrow$ then $\left| X(\vec{s}) - X(\vec{s}') \right| N \sim \sqrt{NU(Ns)^2} = \sqrt{10^4 \times 10^{-2}} = 10 \gg 1$

$\Rightarrow$ in this case, neutral model would not be good approx.
& we'll have to turn to different methods...

In some cases, neutral picture can still be salvaged if we only care about predicting neutral mutations (e.g. synonymous mutations) and we can find some other way to predict genealogy.

e.g. in evolution experiment:



$\Rightarrow$ if pick 1 individual from each pop'n, we know what their genealogy looks like (up to initial grow up period.)

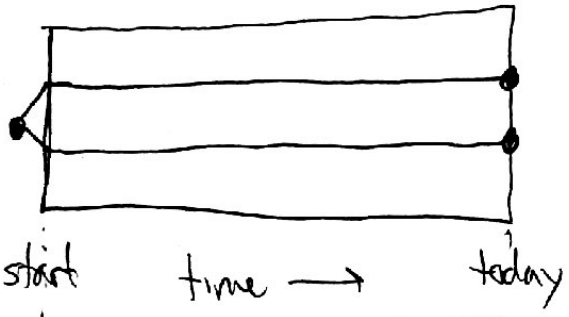
mutations ~ Poisson

$\underline{\text{neutral}}$ mutations btw $\simeq$ Poisson(2UT) regardless of any kind of selection (or anything else) w/in pop'n.
1 & 2

⇒ why doesn't this work for larger sample sizes?

⇒ let's consider 2 scenarios:

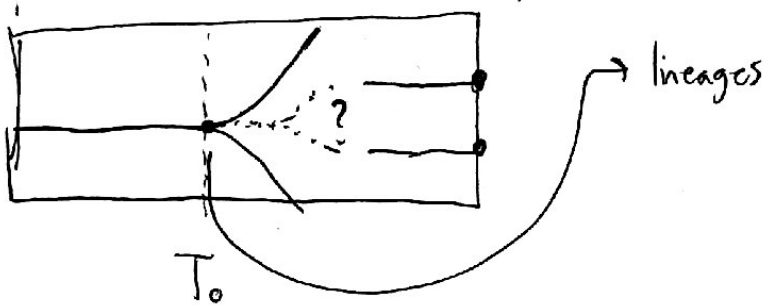
(a) Neutral



if N >> T

⇒ no coalescence until start of experiment.

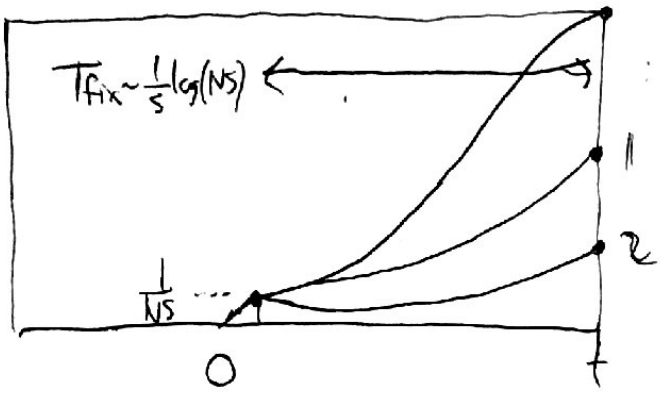
(b) Sweep



⇒ mutations w/in population very different in 2 scenarios.

⇒ in case where selected mutation is from SSWM regime can still make progress (common trend ⇒ reduce to single locus selection & things will be easier)

in this case:



$\Rightarrow$ [illegible] • sweep acts like effective population w/ size $Nf(t)$

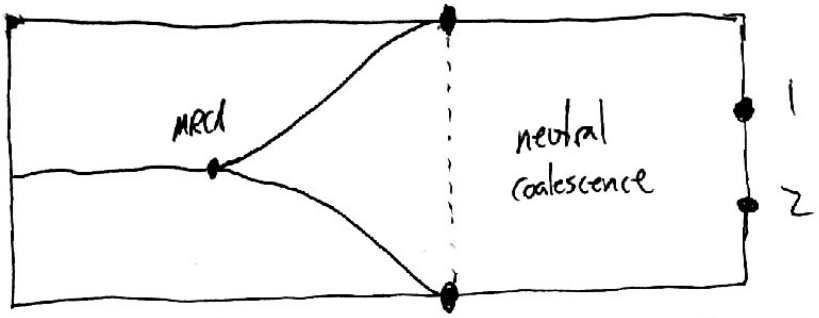
$\Rightarrow$ since $Ns \gg 1$ $\Rightarrow$ $T_{fix} \ll N$ $\Rightarrow$ no coalescence until $f(t)$ gets small!

How small? Need $\frac{1}{N(t)} \approx 1$ (decent chance of coalescing in that amount of time.)

$\Rightarrow$ no coalescence until $f(t) \sim \frac{1}{Ns}$ $\left(t \sim \frac{1}{s}\right)$

$\Rightarrow$ $T_2 \approx \frac{1}{s}\log(Ns) \pm O\left(\frac{1}{s}\right)$ ← difference is small

[illegible] [illegible] mutation had fixed before time of sampling?



$T_{est} \sim 1/N\mu s$ (exponential distn)

$T_{fix} \sim \frac{1}{s}\log(Ns) \ll T_{est}$ (SSWM)

$\Rightarrow$ 2 regimes:

(1) if $N\theta \ll T_{est}$ $\Rightarrow$ coalescence before sweep. (looks neutral)
$(N\mu/Ns) \ll 1$

(2) if $N\theta \gg T_{est}$ $\Rightarrow$ no coalescence until $\approx \Theta(\frac{1}{s})$ of establishment time.
$(N\mu)(Ns) \gg 1$

$T_2 \approx T_{est} \sim$ Exponential $\left(\frac{1}{N\mu s}\right)$

$\Rightarrow$ in this case $\pi \approx \langle T_2 \rangle V_0 \approx \frac{V_0}{V_b}\frac{1}{Ns}$ $\rightarrow$ anti-correlated w/ pop size.

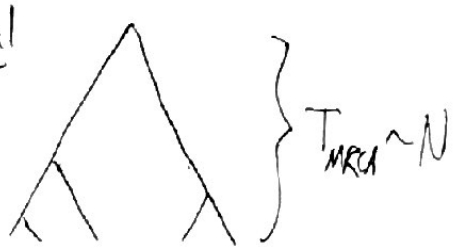
$\Rightarrow$ works for larger sample sizes:

simple selective sweep:

"star like genealogy"
(mostly singletons)

↳ similar to population expansion

neutral

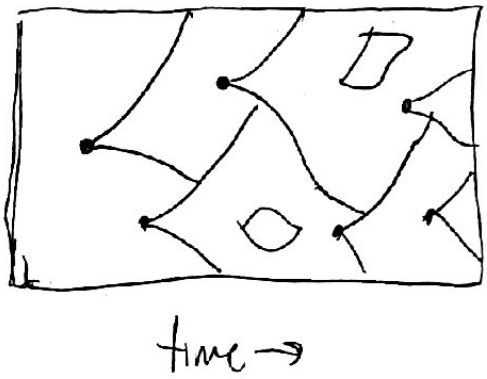


regular kingman genealogy
(last coalescence is longest)

---

w/ more than 1 selected mutation, things get complicated. (clonal interference)

⇒ will revisit in a few lectures



---

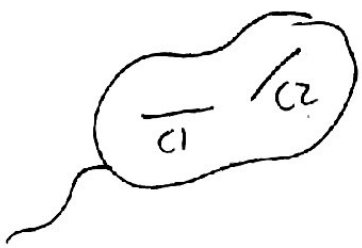
so far, have been talking about asexual populations

⇒ what about recombination? does coalescent picture work there too?

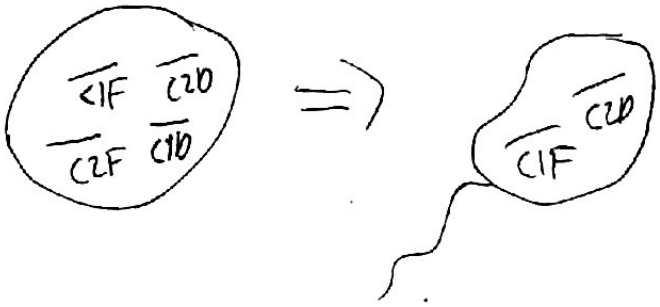
let's go back to neutral case and consider recombination.

⇒ to keep things simple, let's consider a reassortment model of recombination w/ 2 chromosomes of length L

eg. HA + NA genes in influenza.



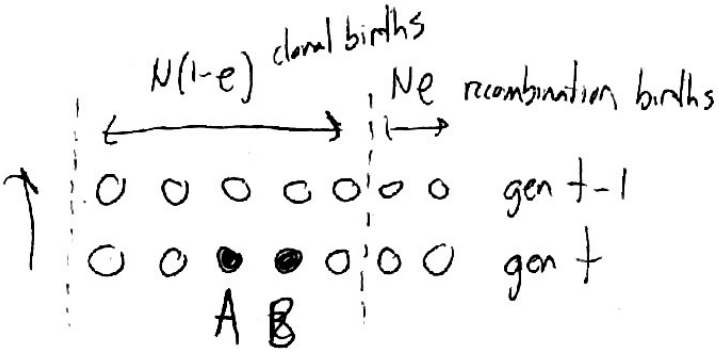
then @ rate ρ, individual mates w/ another



♂ replaces chromosome w/ one from mate.

(i.e., ρ = successful recombination)

$\Rightarrow$ now let's see if we can look back in time:



$\Rightarrow$ in last generation, $N(1-\rho)$ individuals reproduced asexually

$N\rho$ individuals were result of recombination event.

$\Rightarrow$ probability of coalescence $\approx 1/N$

$\Rightarrow$ probability that individual is recombinant $= \dfrac{N\rho}{N(1-\rho)} \approx \rho$

$\Rightarrow$ so w/ prob $e^{-T_{MRCA}(2\rho)}$, there were no recombination events in genealogy of sample.

in this case, T_{MRCA} is same as ancestral population,

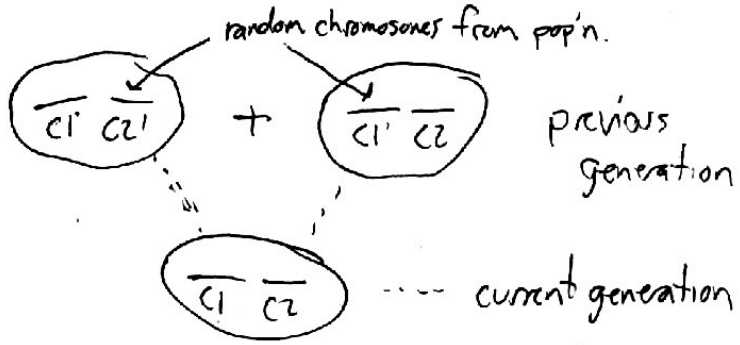
$T_{MRCA} \sim$ Exponential (N), so $P(\text{no recomb}) = \int e^{-T_{MRCA}(2\rho) - T_{MRCA}/N} \frac{1}{N} e \, dT$

$$\approx \frac{1}{1+2N\rho}$$

$\Rightarrow$ hence if $N\rho \ll 1$, genomes behave as if effectively asexual.

$\Rightarrow$ however, if $N\rho \gg 1$, very good chance that in history of sample, one of ancestors will be result of recombination event.

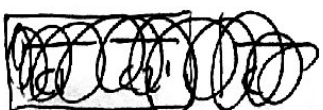
$\Rightarrow$ what does this look like?

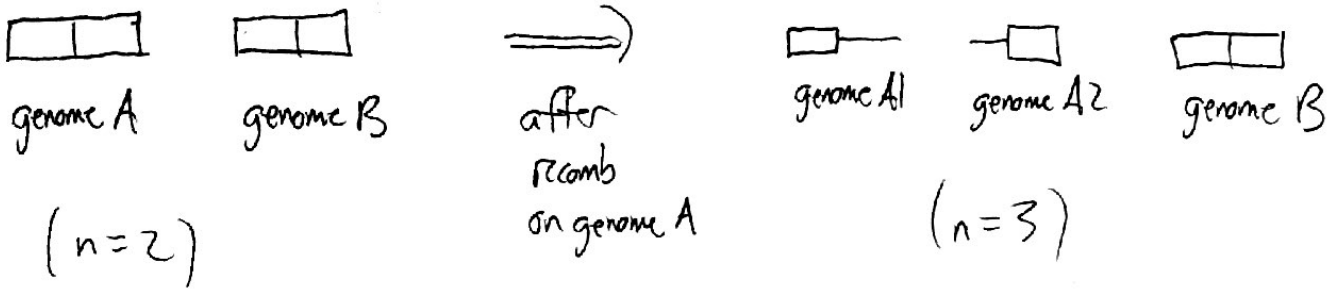


$\Rightarrow$ in this case, ancestors of 2 chromosomes are <u>different</u>

i.e., genealogies of 2 chromosomes separate.

$\Rightarrow$ in coalescent picture, effectively increases our sample size:

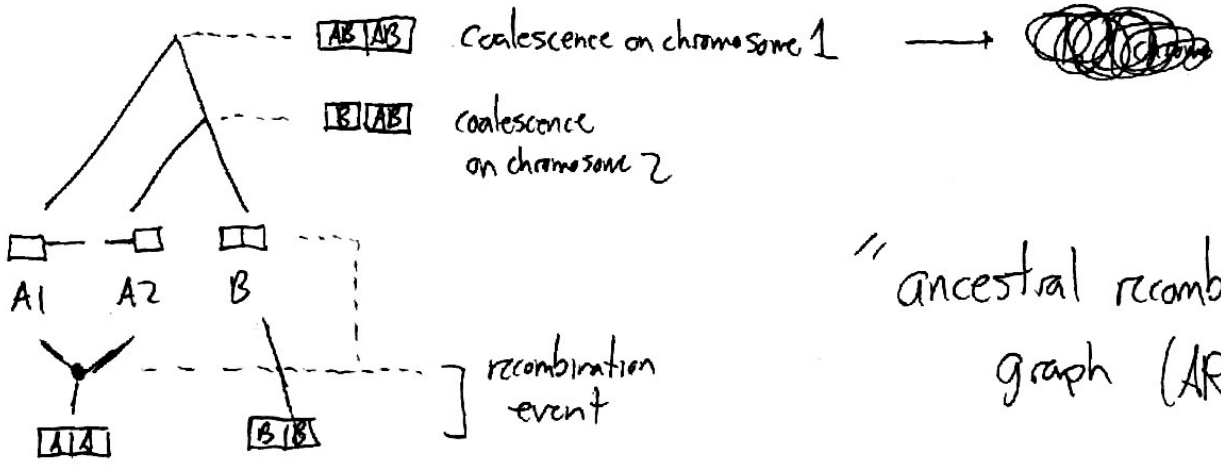




$(n=2)$ $(n=3)$

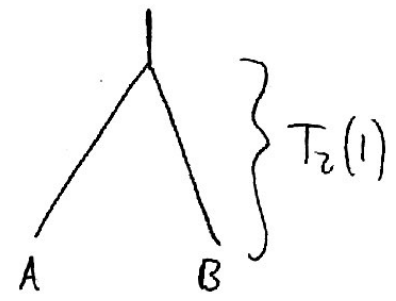
Now coalescent process continues w/ larger sample size:

eg if no other recombination events, could get:

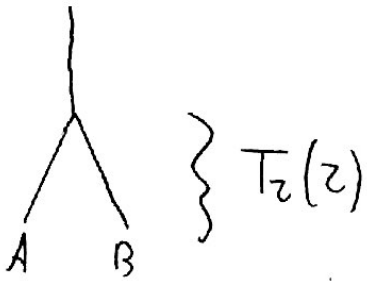


"ancestral recombination graph (ARG)"

$\Rightarrow$



genealogy of chromosome 1

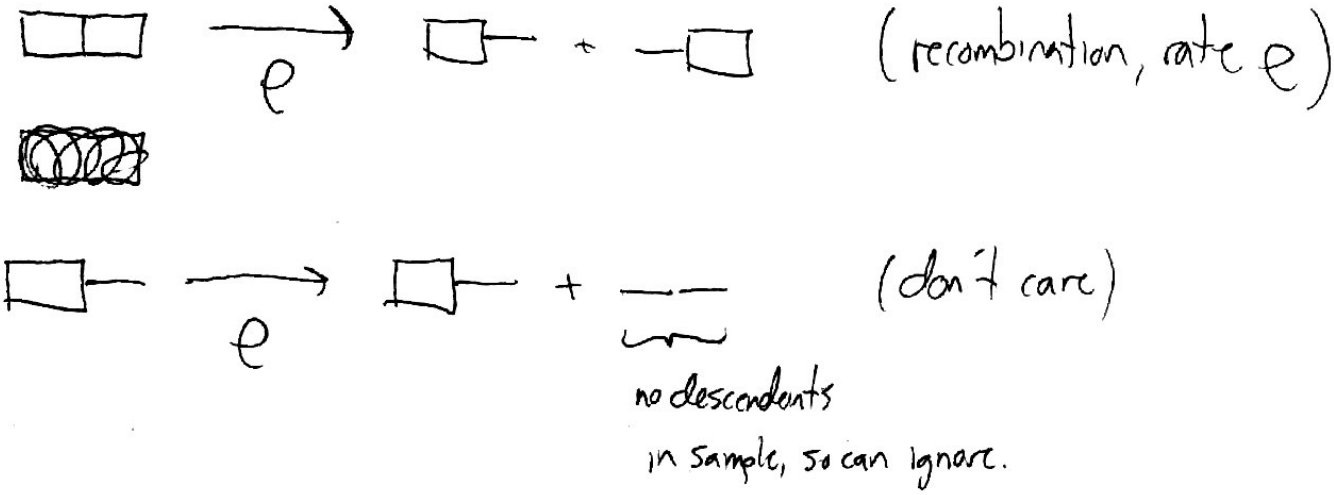


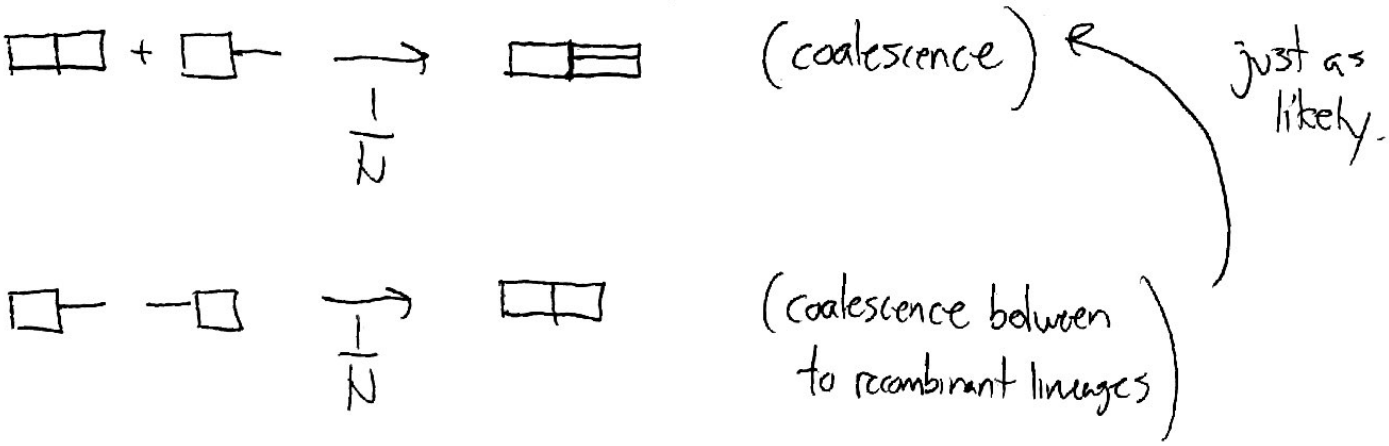
genealogy of chromosome 2.

$\Rightarrow$ i.e., recombination allows genealogies to be different @ different sites in genome. (in asexual case, $T_2(1) = T_2(2)$)

of course, this is just one possible arg a.

@ each stage, can have events:





let's see if we can simulate this in our heads in limit that $N\rho \gg 1$

(1) start w/ [A A] [B B] coalescence $\frac{1}{N}$, recomb 2ρ

(2) recomb happens first w/ high prob. ($\frac{1}{N} \ll 2\rho$),

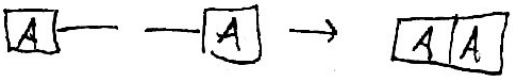
[A]— —[A] [B B] time $T_{rec} \sim \frac{1}{2\rho}$.

Now coalescence $\frac{\binom{3}{2}}{N}$, recomb ρ

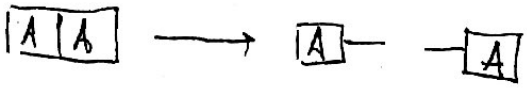
$\Rightarrow$ recomb happens first w/ high prob. $\left(\frac{\binom{3}{2}}{N} \ll \rho\right)$

$\Rightarrow$ must now happen in other lineage: (Time ~ $1/\rho$)

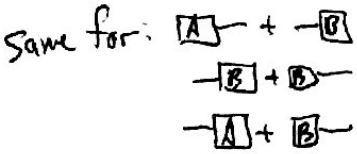
Now coalescence = $\frac{\binom{4}{2}}{N}$, recomb = 0. $\Rightarrow$ next event must be coalescence. ($T \sim N$)

$\Rightarrow$ 4 possible pairs. if 

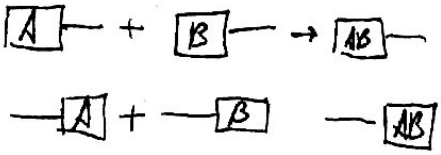
then next event will be recomb.



(only time ~ $1/\rho \ll N$)

Same for: 

$\Leftarrow$ so it's almost as if this coalescence event never happened

$\Rightarrow$ can instead keep track of the 2 coalescence events that won't immediately get undone by recomb: 

each of the two events happens independently w/ rate $\frac{1}{N}$

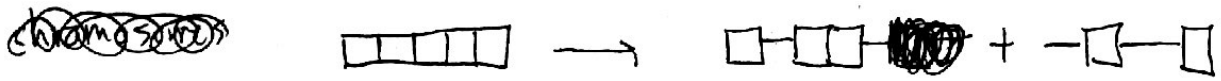
$$\Rightarrow T_2(1), T_2(2) \overset{\text{independent}}{\sim} \text{Exponential}(N)$$

Putting everything together, we have:

$$p\left(T_2(1), T_2(2)\right) \approx \begin{cases} \frac{1}{N} e^{-\frac{T_2(1)}{N}} \delta\left(T_2(1) - T_2(2)\right) & \text{if } Ne \ll 1 \\ \left[\frac{1}{N} e^{-\frac{T_2(1)}{N}}\right] \cdot \left[\frac{1}{N} e^{-\frac{T_2(2)}{N}}\right] & \text{if } Ne \gg 1 \end{cases}$$

i.e., effectively asexual vs effectively independent.

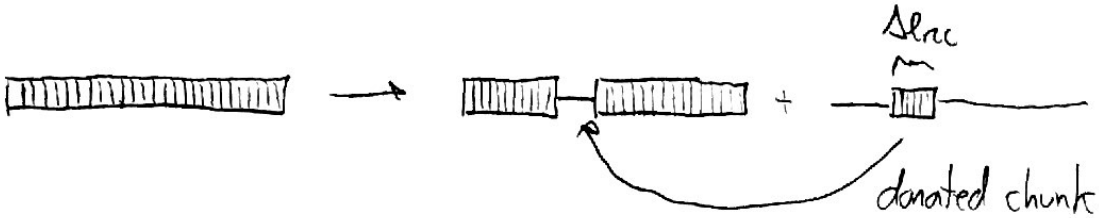
$\Rightarrow$ works for more than 2 chromosomes:



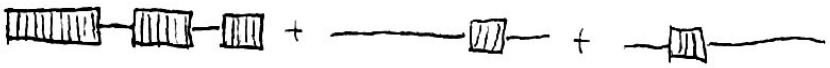
$\Rightarrow$ also works for other modes of recombination:

① crossover recomb:

(2) HGT



$\Rightarrow$ next recomb:



$\Rightarrow$ need $\sim \frac{L}{\Delta\ell_{rec}}$ events to mix up whole genome.

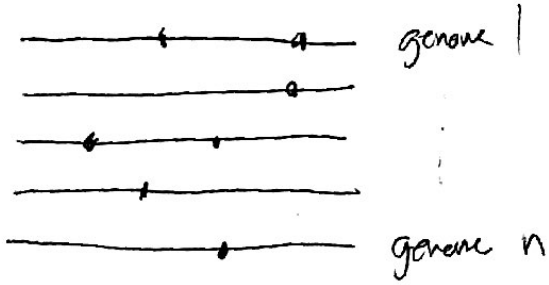
$\Rightarrow$ but actually calculating statistics of ARG is very hard

(because like coalescent tree w/ 2^c samples, where c = # of chromosomes.)

$\Rightarrow$ essentially intractable (lots of approx models out there...)

$\Rightarrow$ then you have to paint mutations on top of this.

$\Rightarrow$ hard to directly know what data looks like:



$\Rightarrow$ next time: back to forward time approach to calculate some summary statistics.