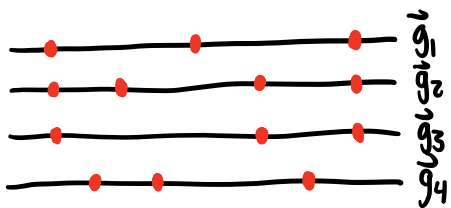
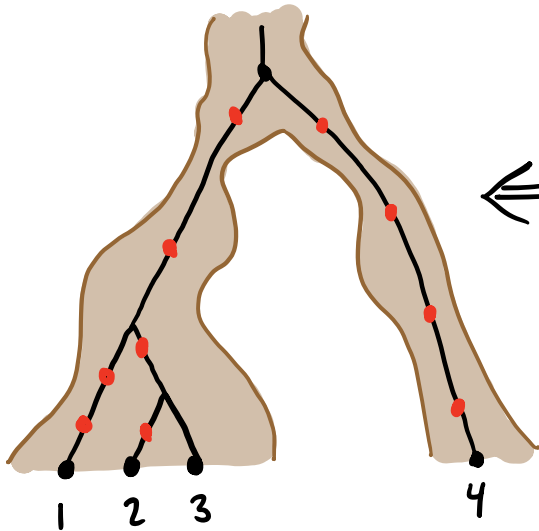


Chapter 12

Genealogies with selection and recombination

Last time: Coalescent theory for **neutral** + **asexual** genomes



$$\frac{d\bar{f}(\bar{s})}{dt} = \underbrace{[\bar{X}(\bar{s}) - \bar{X}(t)] f(\bar{s})}_{\text{selection (non-linear)}} + \underbrace{\sum_{\bar{j}} M(\bar{s} \rightarrow \bar{j}) f(\bar{j}) - M(\bar{j} \rightarrow \bar{s}) f(\bar{j})}_{\text{mutation (linear, "local")}}$$

$$+ e \sum_{\bar{j}, \bar{k}} \underbrace{r(\bar{s} \rightarrow \bar{j}, \bar{k}) f(\bar{j}) - r(\bar{s} \rightarrow \bar{k}, \bar{j}) f(\bar{k})}_{\text{recombination (non-linear, non-local)}}$$

$$+ \underbrace{\sqrt{\frac{f(\bar{s})}{N}} \eta(t) - f(\bar{s}) \sum_{\bar{j}} \sqrt{\frac{f(\bar{j})}{N}} \eta(t)}_{\text{genetic drift (stochastic)}}$$

2 simple rules:

(i) genealogy: $p(\delta\delta) = 1/N(\tau)$

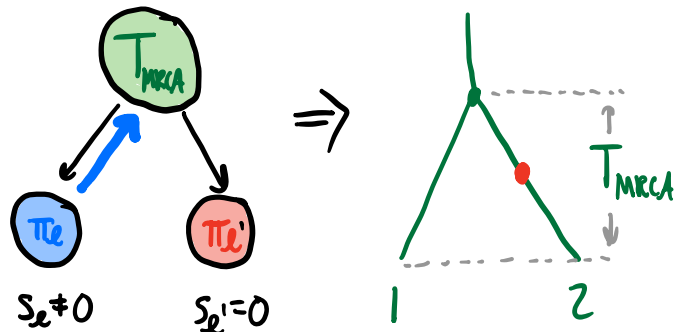
(ii) mutations: Poisson(μt)

⇒ E.g. pairwise diversity:

$$\langle \pi \rangle = 2\mu \langle T_{MRCA} \rangle = 2\mu \int_0^{\infty} e^{-\int_0^{\tau} \frac{dc'}{N(c')}} d\tau$$

Today: How can we get **selection** + **recombination** back in picture?

⇒ Selection is hard
(alters causation)
diagram



⇒ in some cases, coalescent picture can be salvaged if

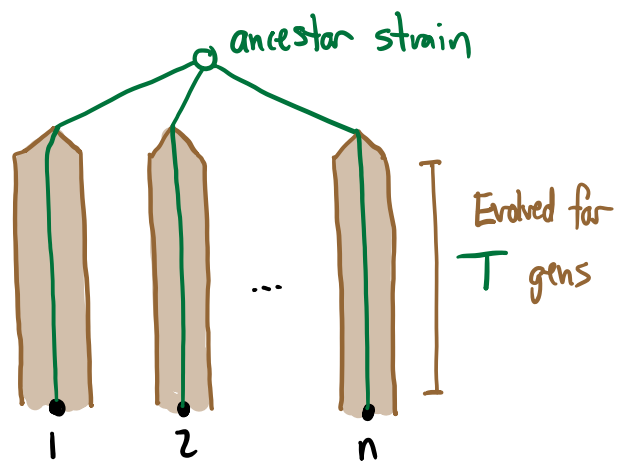
① mainly care about predicting **neutral sites** (e.g. synonymous mut's)

② can find some other way to predict **genealogy**

Simple example:

evolution experiment
in HW 3, Problem 2:

⇒ picked 1 individual
from each population



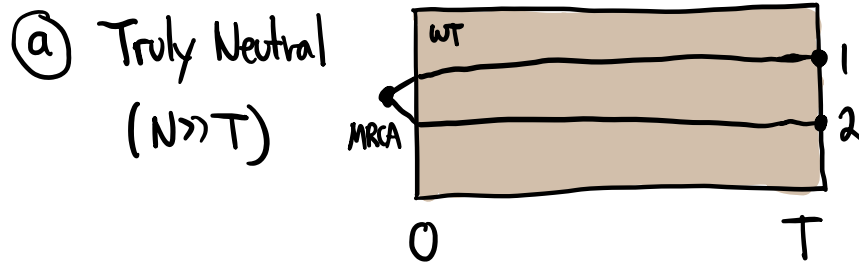
⇒ know exactly what genealogy looks like!

⇒ # synonymous mut's / clone \sim Poisson($L_{syn} \mu T$)

⇒ why can't this work for larger samples?

⇒ why can't this work for larger samples?

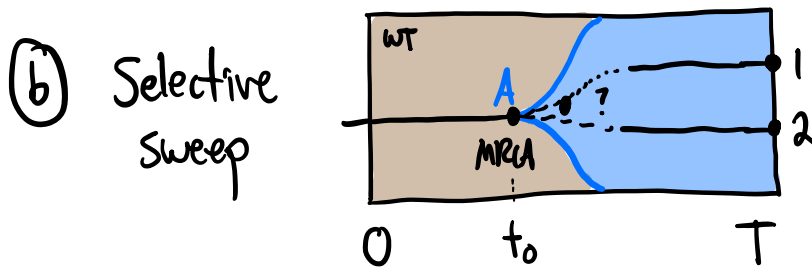
⇒ consider 2 scenarios:



$$\Rightarrow p(\text{obs}) = \frac{1}{N} (\text{per-gen}) \Rightarrow \Pr(T_{\text{MRCA}} < T \ll N) \approx T/N \ll 1$$

i.e., \approx no coalescence during experiment!

"diff is weak"



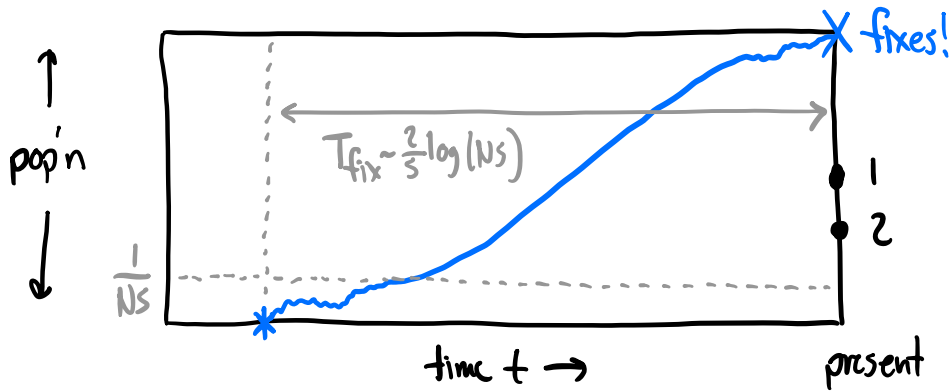
lineages must coalesce by t_0 !

⇒ genealogies for $n \geq 2$ can be very different!

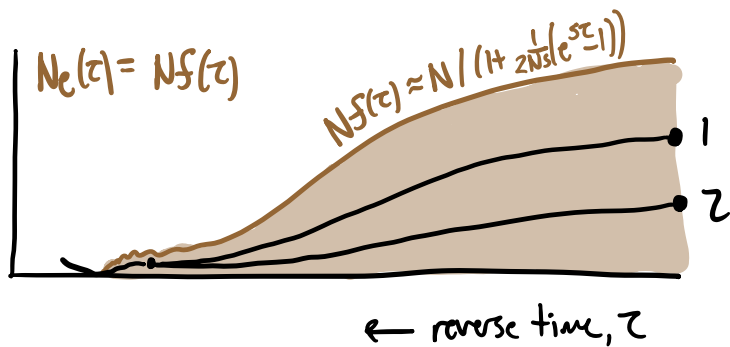
⇒ when **selected mut'n** is from **successive mut'n's regime**

⇒ can make some quantitative progress

in this case, know entire trajectory of **selected mut'n**:



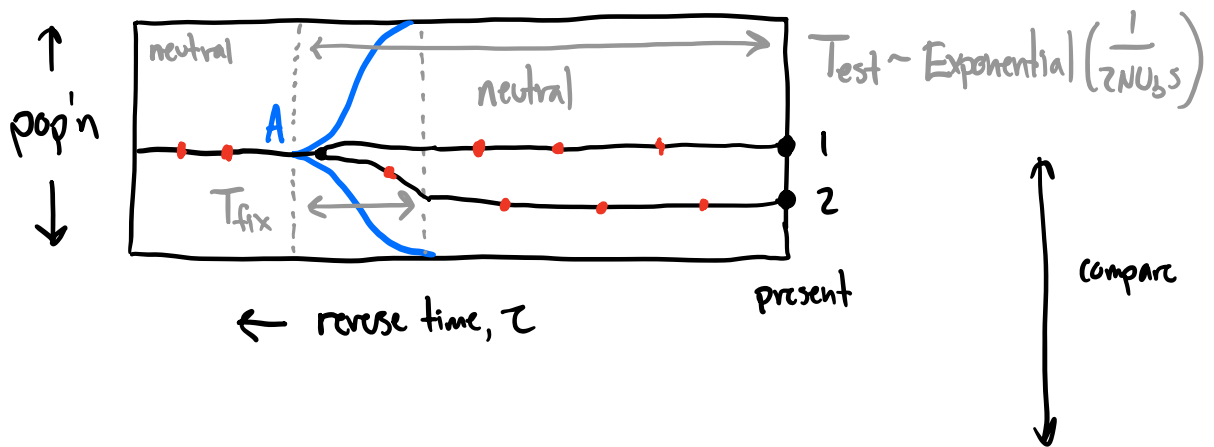
equivalent to demography problem!



$$\begin{aligned} \Rightarrow \Pr[T_{MRCA} \geq \tau] &= \exp\left[-\int_0^\tau \frac{dz'}{N_e(z')}\right] = e^{-\int_0^\tau \frac{(1 - \frac{1}{2Ns}) + \frac{1}{2Ns} e^{sz'}}{N} dz'} \\ &= \exp\left[-\frac{(1 - \frac{1}{2Ns})\tau}{N} - \frac{1}{2(Ns)^2} (e^{s\tau} - 1)\right] \approx \exp\left[-2e^{-s(T_{fix} - \tau)}\right] \end{aligned}$$

\Rightarrow no coalescence until $\tau \sim T_{fix} \pm \mathcal{O}(\frac{1}{s})!$ [when $f(z) \approx \frac{1}{Ns}$]

what if mutation had fixed before time of sampling?



Two characteristic regimes:

$$T_{MRCA} \sim \text{Exp}(N)$$

① if $N \ll T_{test} \Rightarrow$ coalescence before sweep \Rightarrow neutral!

② if $T_{test} \ll N \Rightarrow T_{MRCA} \approx T_{test} = \text{Exponential}(\frac{1}{2N\mu_b s})$

$$\hookrightarrow \pi_{syn} = 2\mu \langle T_{MRCA} \rangle = \left(\frac{N}{U_b}\right) \frac{1}{Ns}$$

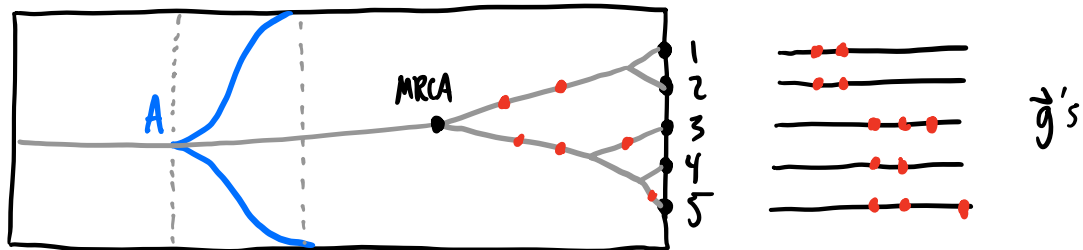
\Rightarrow anticorrelated w/ $N!$

$$\hookrightarrow "N_e" \propto 1/N$$

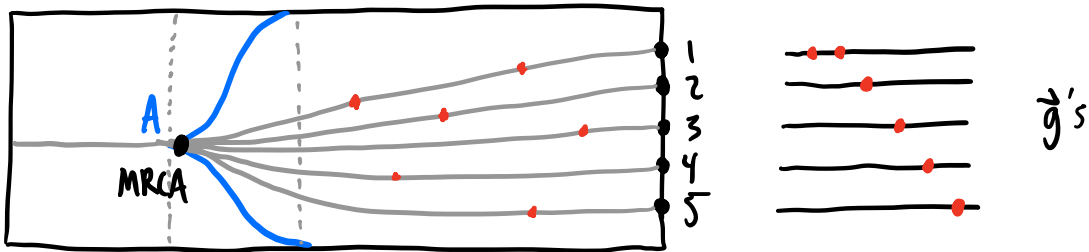
(doesn't make much sense to think about it as an "Ne")

can extend to larger sample sizes:

① $T_{fix} \ll N \ll T_{est} \Rightarrow$ effectively neutral

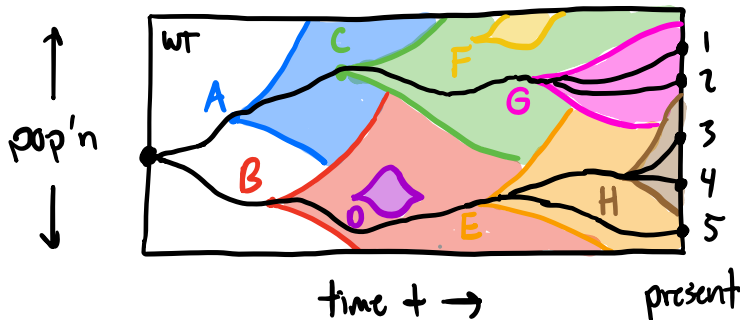


② $T_{fix} \ll T_{est} \ll N \Rightarrow$ "star-like genealogy"



\Rightarrow not just a difference in scale \Rightarrow difference in shape!

③ $T_{est} \ll T_{fix} \ll N \Rightarrow$ "clonal interference"



(will revisit in final week)

Next: How can we account for **recombination**?

⇒ start w/ neutral case

$$\frac{ds(\bar{g})}{dt} = \cancel{\sim \frac{1}{N}} + \sim L\mu + \sim e + \sim \frac{e}{N}$$

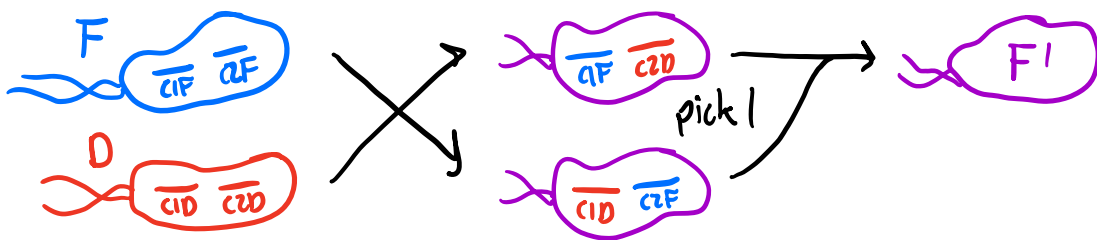
⇒ start w/ **reassortment** model of recombination

w/ 2 chromosomes of length L

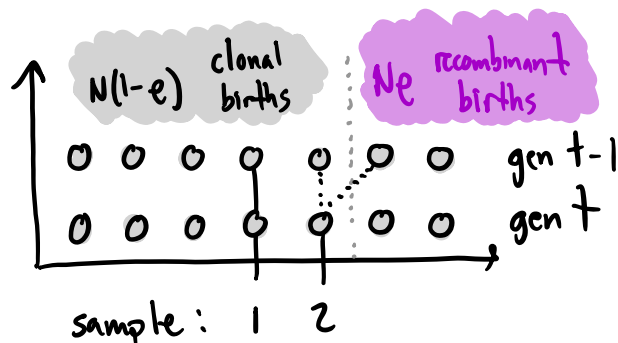


(e.g. HA + NA genes in influenza)

⇒ Recall: @ per capita rate e :



Backwards in time:



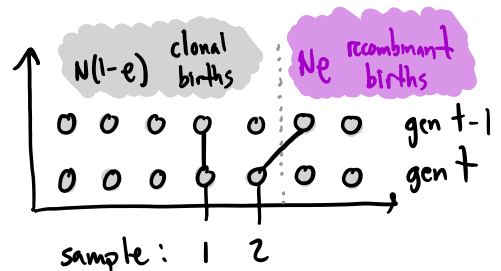
⇒ probability that individual was **recombinant** = $\frac{Ne}{N} = e$

⇒ coalescence probability = $\frac{1}{N}$ (same as before)

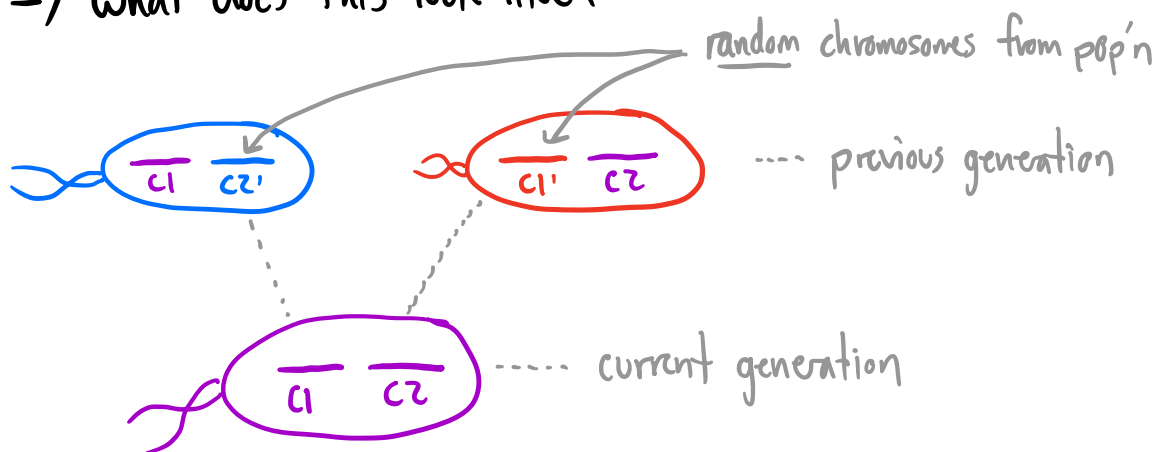
$$\Rightarrow \Pr[\text{coalesce before recombine}] = \frac{\frac{1}{N}}{\frac{1}{N} + 2e} = \frac{1}{1 + 2Ne}$$

⇒ if $Ne \ll 1 \rightarrow$ effectively asexual!

⇒ if $Ne \gg 1$, good chance that some ancestral individuals were result of **recombination event...**



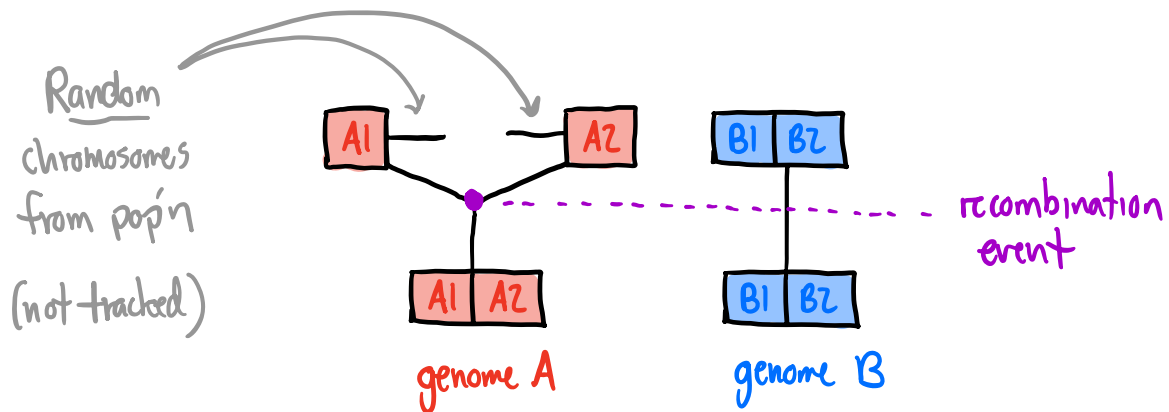
⇒ what does this look like?



⇒ ancestors of 2 chromosomes are different!

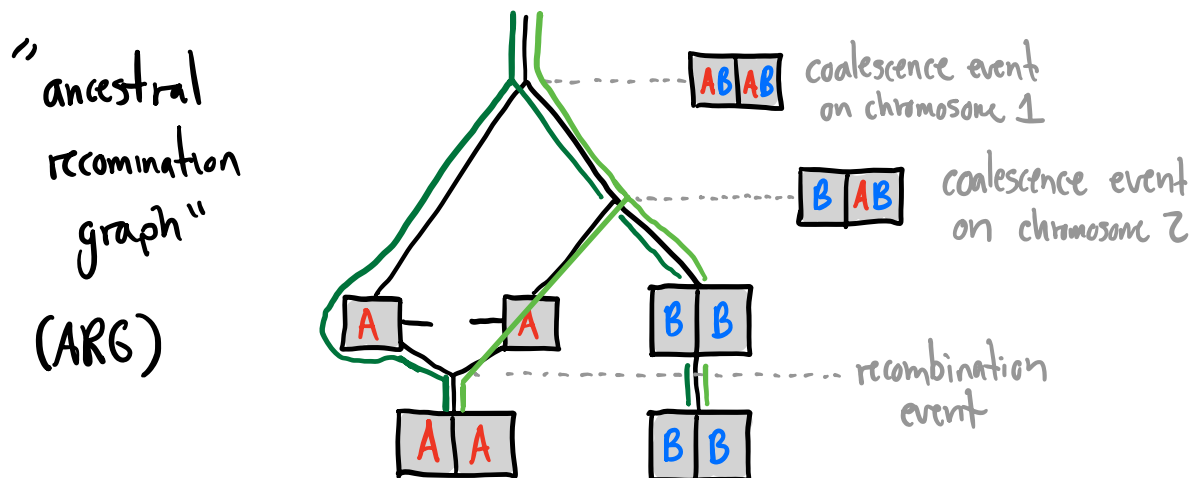
⇒ i.e. genealogies must separate!

⇒ can represent this in coalescent picture as:

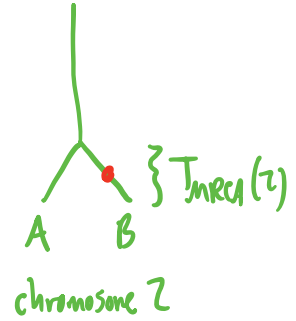
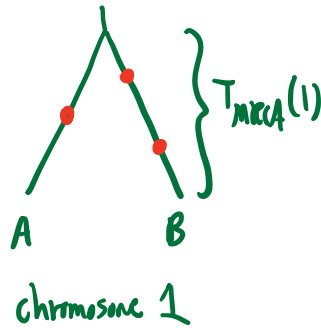


⇒ now coalescent continues w/ larger sample ($n=3$)

⇒ e.g. if no more recomb events, could have:



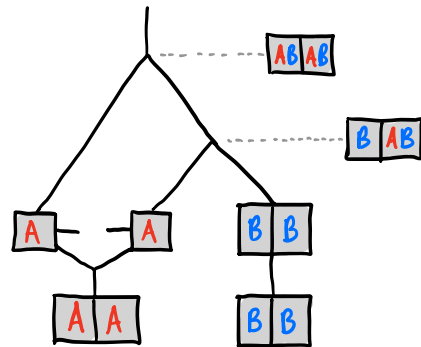
⇒ can extract genealogies for each chromosome:



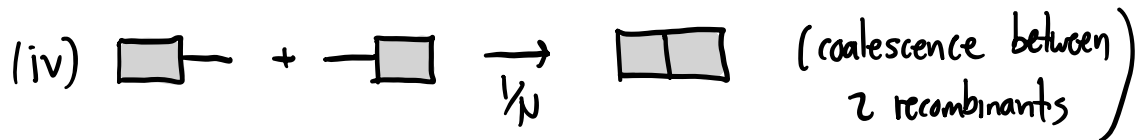
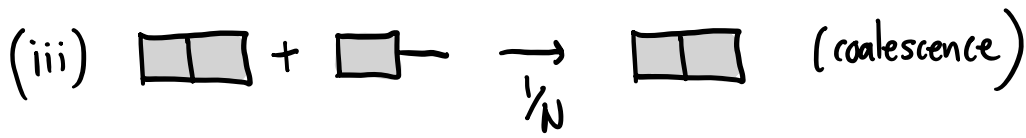
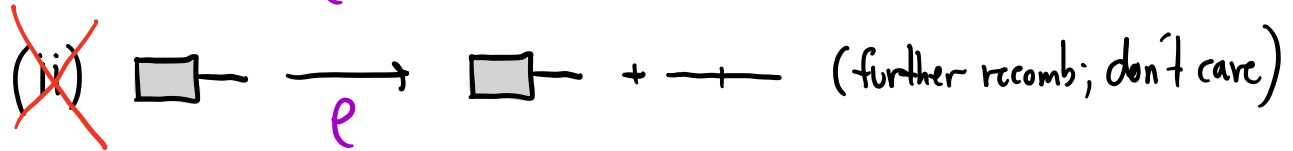
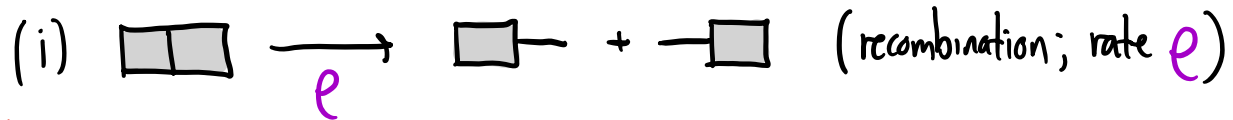
⇒ i.e. recombination allows genealogies to differ @ different locations along genome

[compare to asexual case where $T_{MRCA(1)} = T_{MRCA(2)}$]

⇒ this was just one possible ARG...



⇒ more generally, @ each step will have 4 types of events:



just as likely per pair!

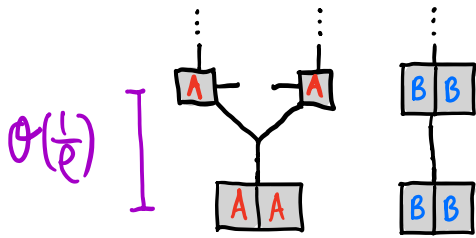
⇒ can we simulate this process in our heads when $N_e \gg 1$?

Start with sample:  

① Total coalescence rate = $1/N$ (1 pair)

② Total recombination rate = $2e$

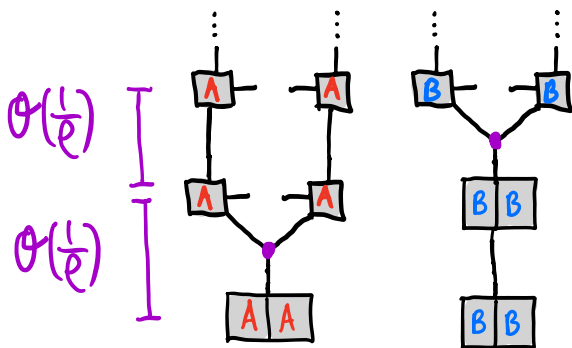
Step 2:



① coalescence = $\frac{1}{N} \cdot \binom{3}{2} = \frac{3}{N}$

② recomb = ρ

Step 3:

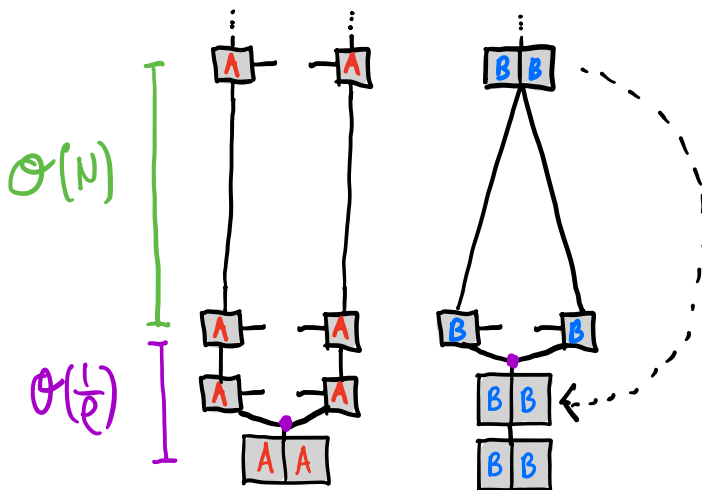


① coalescence = $\frac{1}{N} \binom{4}{2} = \frac{6}{N}$

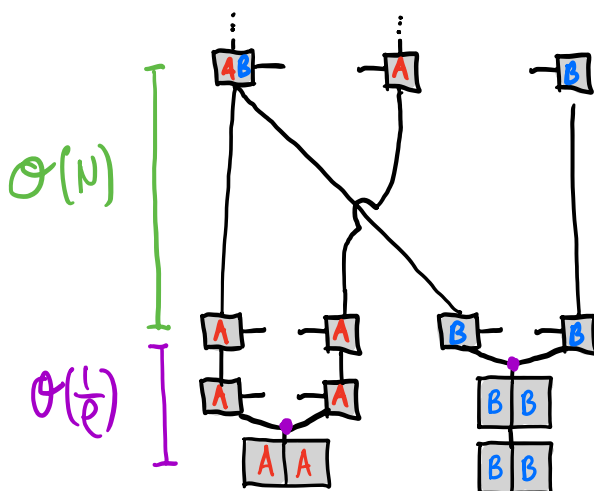
② recomb = 0

2 different types of coalescent events:

(i) coalescence of recombinant chromosomes ($\square + -$)



(ii) coalescence involving sampled genetic material ($\square + \square$)



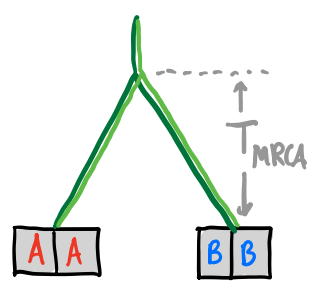
\Rightarrow each coalescence happens \approx independently w/ rate $1/N$.

$\Rightarrow T_{MRCU}(1) \sim \text{Exponential}(N)$

$T_{MRCU}(2) \sim \text{Exponential}(N)$

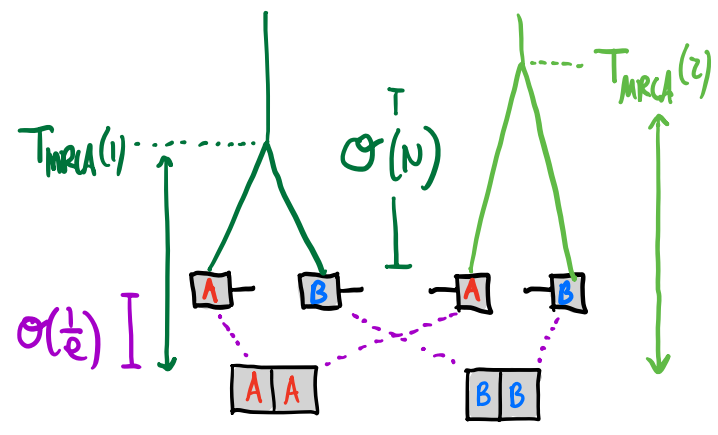
Putting everything together:

$Nq \ll 1$ (effectively asexual)



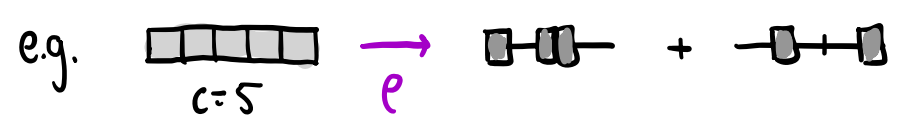
$T_{MRCA(1)} = T_{MRCA(2)}$
 $\sim \text{Exponential}(N)$

$Nq \gg 1$ (effectively independent)



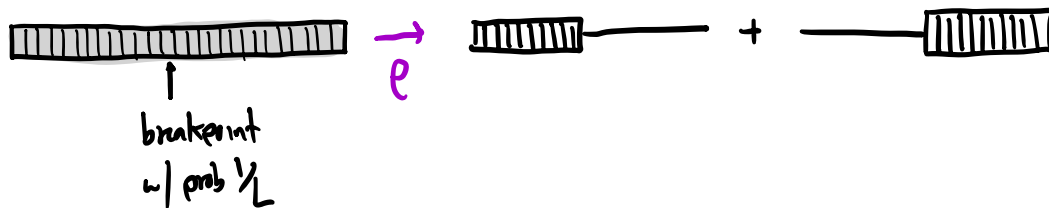
$T_{MRCA(1)}, T_{MRCA(2)} \overset{iid}{\sim} \text{Exponential}(N)$

\Rightarrow same idea works for > 2 chromosomes:

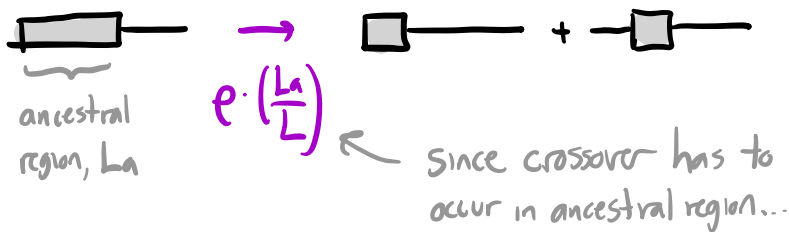


⇒ also works for other forms of recombination:

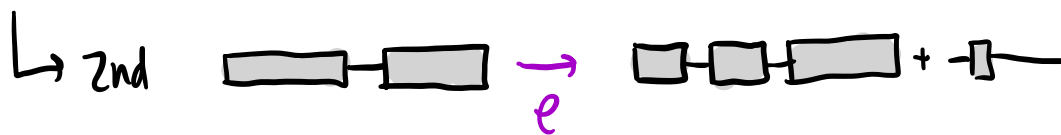
e.g. crossover:



⇒ 2nd event




e.g. HGT / gene conversion:



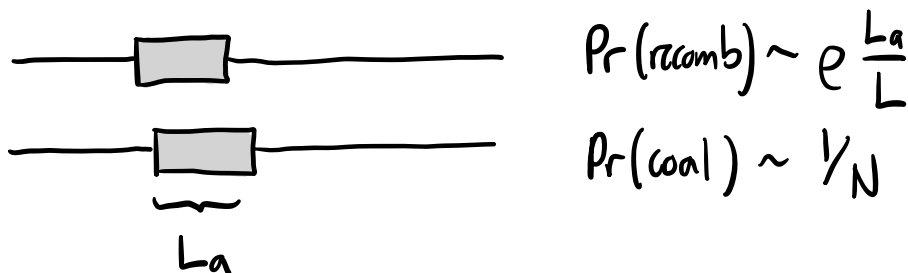
⇒ easy to simulate, but hard to calculate (even for $n=2$!)

⇒ effective sample size $\sim 2^{\text{effective \# chromosomes}}$

New feature for longer genomes:

even if $N_e \gg 1$, ancestral chunks (——) will eventually get small enough that $\Pr(\text{recomb}) \sim \Pr(\text{coal})$...

⇒ can we estimate when?



$$\Pr(\text{recomb}) \sim \Pr(\text{coal}) \Rightarrow e^{-\frac{L_a^*}{L}} \sim \frac{1}{N} \Rightarrow L_a^* \sim \frac{L}{N_e}$$

Upshot: on length scales $\lesssim L_a^*$, sites are likely to share ancestry

e.g. humans: $L_a^* \sim \frac{10^8}{10^{4.5} \cdot 10^0} \approx 1-10\text{kb}$

$$\Rightarrow \frac{L}{L_a} \sim \frac{3 \times 10^9 \text{ bp}}{10^{3-4} \text{ bp}} \sim 3 \times 10^{5-6} \text{ chunks/genome}$$

⇒ again, hard to add

selection back to picture...

$$\frac{ds(\vec{q})}{dt} = \cancel{\sim \vec{q}} + \sim L \cdot \mu + \sim e + \sim \frac{Z}{\sqrt{u}}$$

⇒ Next: back to forward-time
approach to see if
we can make some progress...