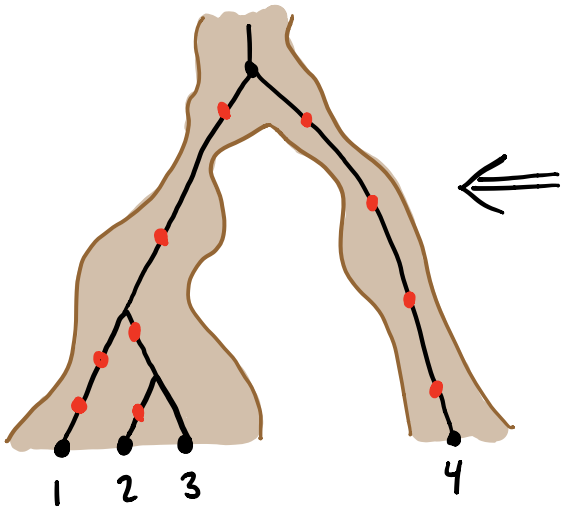
Announcements: PSET 4 DUE TUES; Office Hrs 12-1pm Monday

Last time: Coalescent theory for neutral + asexual genomes



$$\frac{\partial f(\vec{g})}{\partial t} = \cancel{\left[X(\vec{g}) - \bar{X}(t)\right] f(\vec{g})} + \sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}) - M(\vec{g} \to \vec{g}') f(\vec{g})$$

selection (nonlinear)

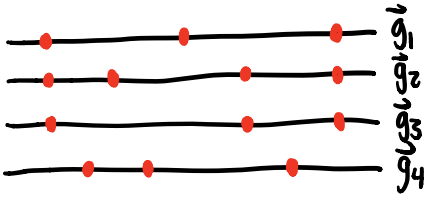
mutation (linear, "local")

$$+ \rho \sum \cancel{\left[g_{new} - \vec{g} \right] f(\vec{g})} - \rho f(\vec{g})$$

recombination (nonlinear, non-local)

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

genetic drift (stochastic)



2 simple rules:

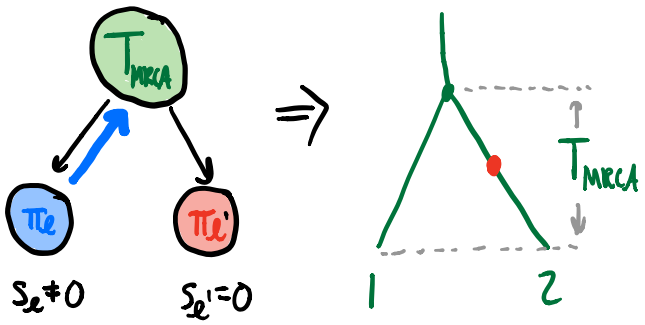
(i) genealogy: $p(\delta\delta) = 1/N(\tau)$

(ii) mutations: Poisson(μt)

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

⇒ Selection is hard

(alters causation diagram)



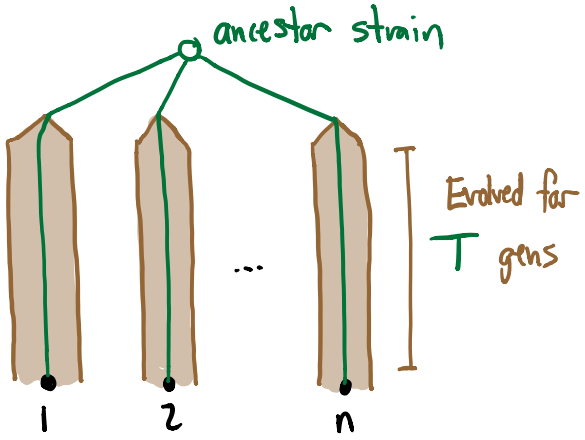
⇒ in some cases, coalescent picture can be salvaged if

① mainly care about predicting neutral sites (e.g. synonymous)

② can find some other way to predict genealogy

e.g. evolution experiment in HW 3, Problem 2:

⇒ picked 1 individual from each population



⇒ know exactly what genealogy looks like!

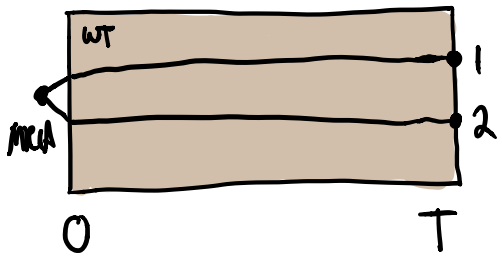
⇒ # synonymous mutations ~ Poisson($L_{syn} \mu T$)

⇒ why can't this work for larger samples?

⇒ why can't this work for larger samples?

⇒ consider 2 scenarios:

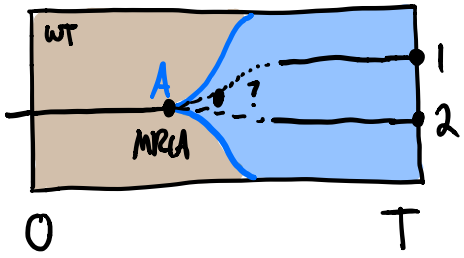
ⓐ Truly Neutral (N≫T)



→ $p_i = \frac{1}{N}$

prob ~ $\frac{T}{N} \ll 1$

ⓑ Selective Sweep

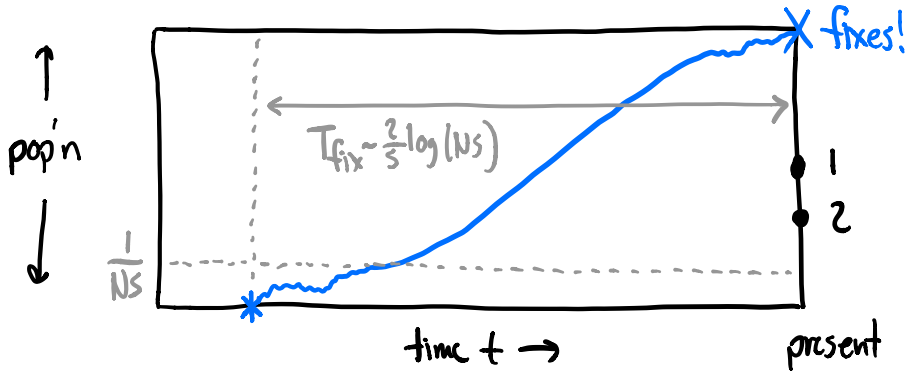


⇒ genealogies for $n \geq 2$ can be very different!

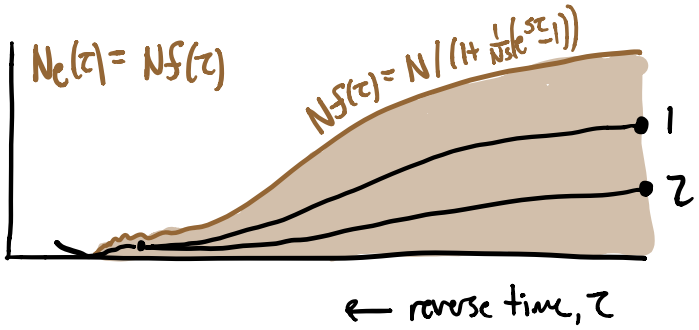
⇒ when selected mut'n is from successive mut'ns regime

⇒ can make some quantitative progress

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



↳ equivalent to demography problem!

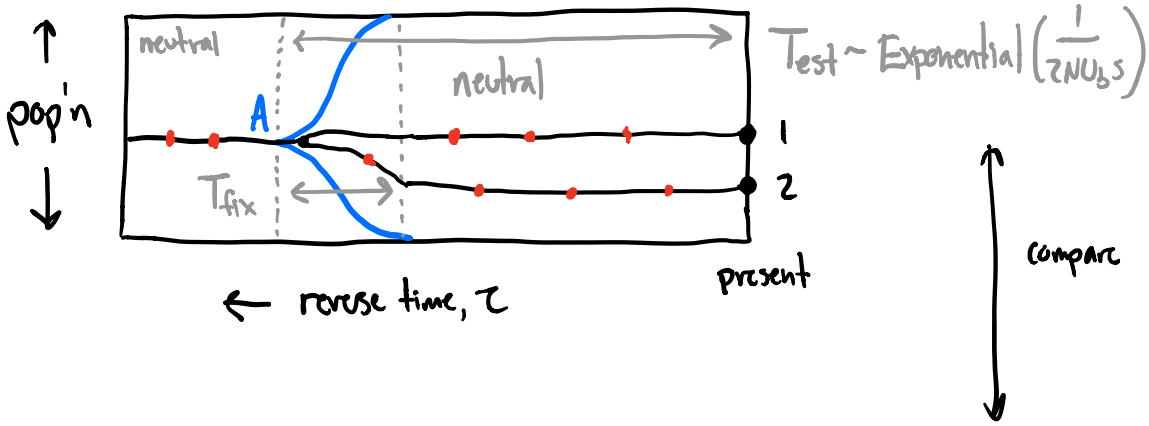


$$\Rightarrow \Pr[T_{MRCA} \geq \tau] = \exp\left[-\int_0^\tau \frac{d\tau'}{N_e(\tau')}\right] = e^{-\int_0^\tau \frac{(1-\frac{1}{Ns})+\frac{1}{Ns}e^{s\tau'}}{N} d\tau'}$$

$$= \exp\left[-\frac{(1-\frac{1}{Ns})\tau}{N} - \left(\frac{1}{Ns}\right)^2 (e^{s\tau}-1)\right] \approx \exp\left[-e^{-s(T_{fix}-\tau)}\right]$$

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

What if mutation had fixed before time of sampling?



Two characteristic regimes:

① if $N \ll T_{est}$ ⇒ coalescence before sweep ⇒ neutral!

② if $T_{est} \ll N$ ⇒ $T_{MRCA} \approx T_{est} = \text{Exponential}\left(\frac{1}{2NU_b s}\right)$

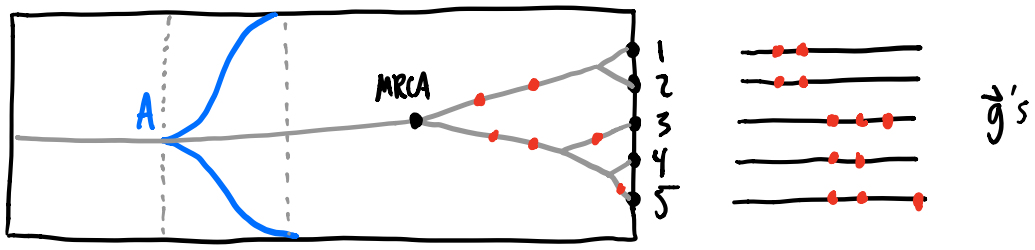
↳ $\pi_{syn} = 2\mu\langle T_{MRCA}\rangle = \left(\frac{N}{U_b}\right)\frac{1}{Ns}$

⇒ anticorrelated w/ N!

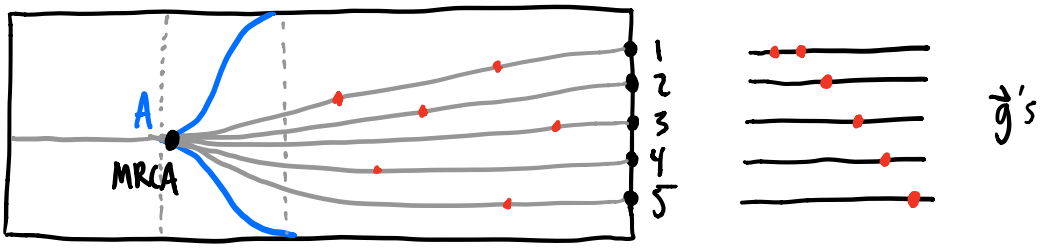
↳ "N_e" $\propto 1/N$

can extend to larger sample sizes:

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

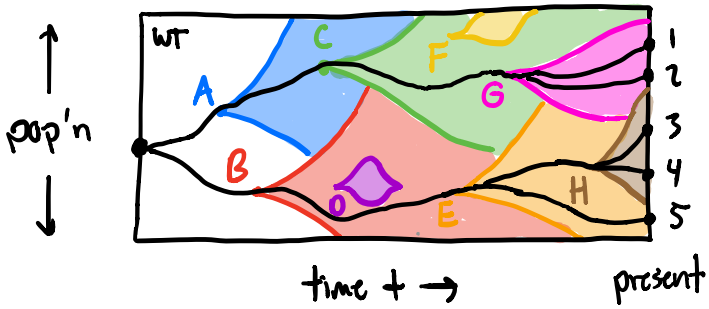


② $T_{fix} \ll T_{est} \ll N$ ⇒ "star-like genealogy"



⇒ not just a difference in scale ⇒ difference in shape!

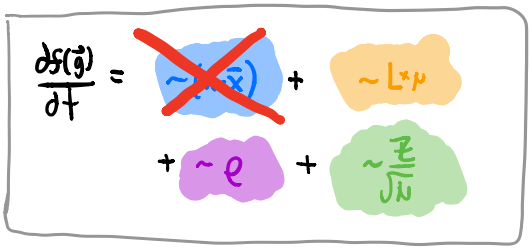
② $T_{est} \ll T_{fix} \ll N$ ⇒ "clonal interference"



(will revisit in final week)

Next: How can we account for recombination?

⇒ start w/ neutral case

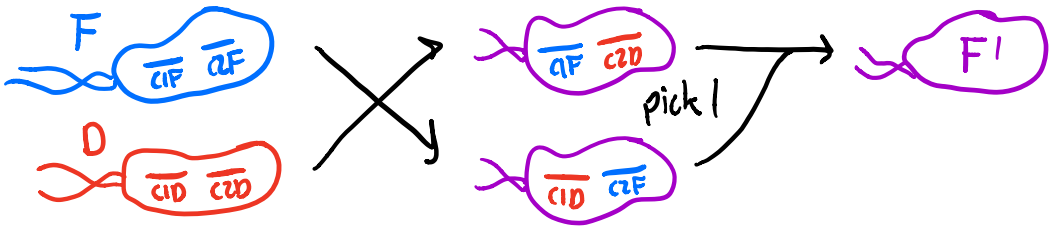


⇒ start w/ reassortment model of recombination

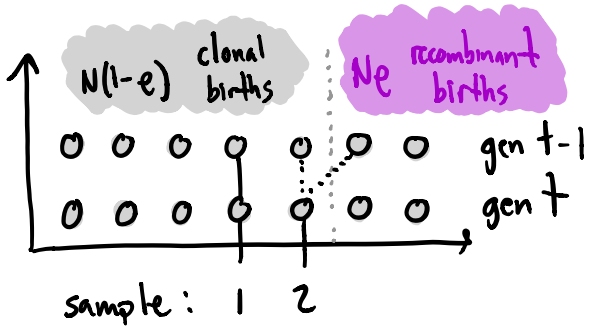
w/ 2 chromosomes of length L

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

⇒ Recall: @ per capita rate ρ:



Backwards in time:



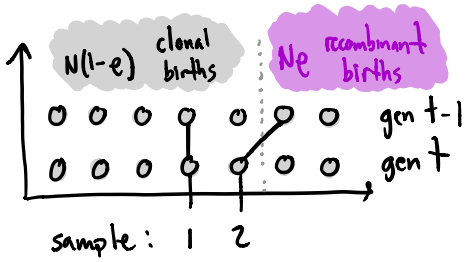
$\Rightarrow$ probability that individual was recombinant $= \frac{Ne}{N} = e$

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

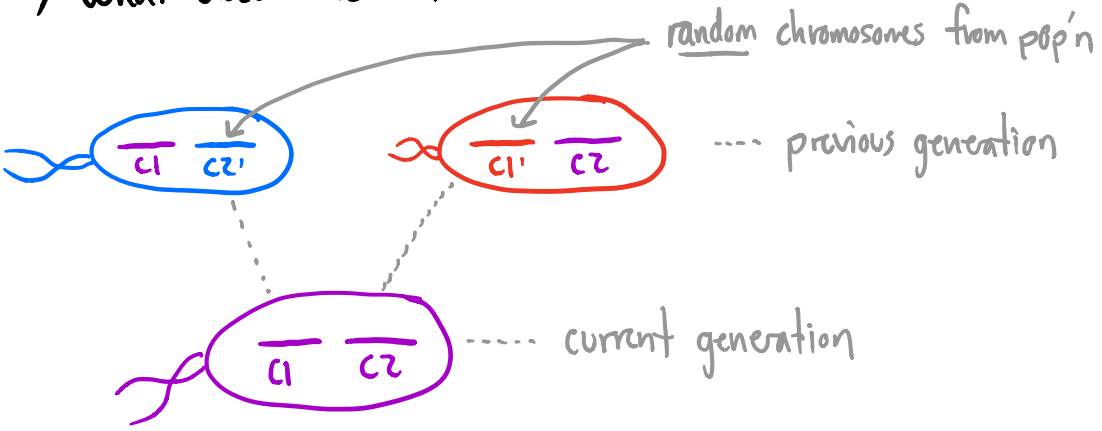
$\Rightarrow$ $Pr[\text{coalesce before recombine}] = \frac{\frac{1}{N}}{\frac{1}{N} + 2e} = \frac{1}{1+2Ne}$

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

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



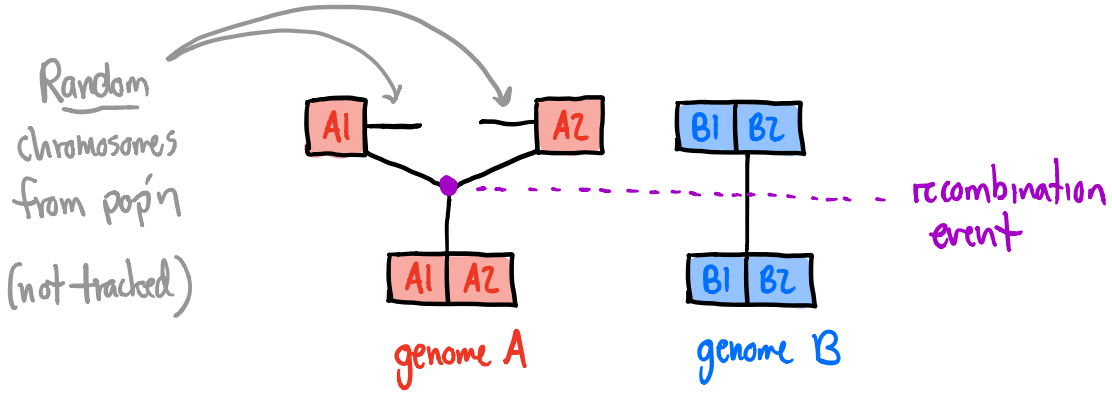
$\Rightarrow$ what does this look like?



⇒ ancestors of 2 chromosomes are different!

⇒ i.e. genealogies must separate!

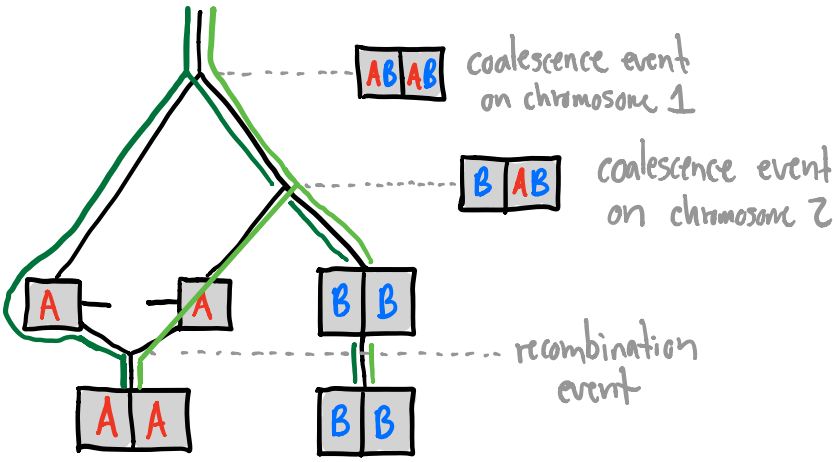
⇒ can represent this in coalescent picture as:



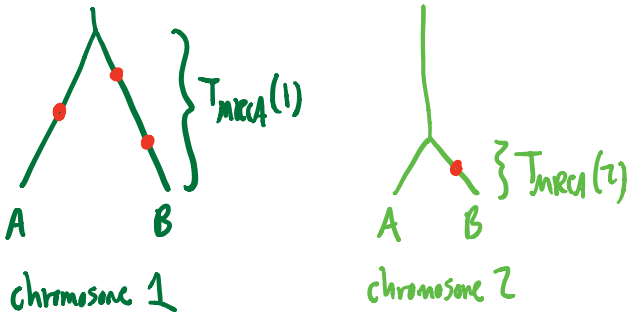
⇒ now coalescent continues w/ larger sample (n=3)

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

"ancestral
recomination
graph"
(ARG)



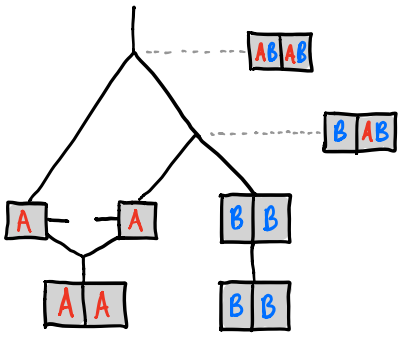
⇒ can extract genealogies for each chromosome:



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

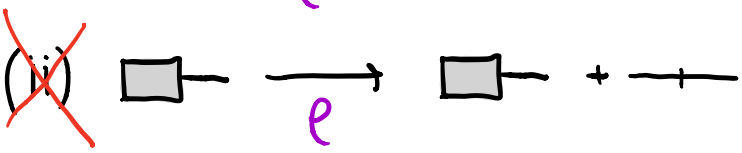
[compare to asexual case where $T_{MRCA}(1) = T_{MRCA}(2)$]

⇒ this was just one possible ARG...

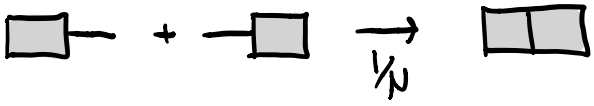


⇒ more generally, @ each step will have 4 types of events:

(i)  (recombination; rate ρ)

(ii)  (further recomb; don't care)

(iii)  (coalescence)

(iv)  (coalescence between 2 recombinants)

just as likely per pair!

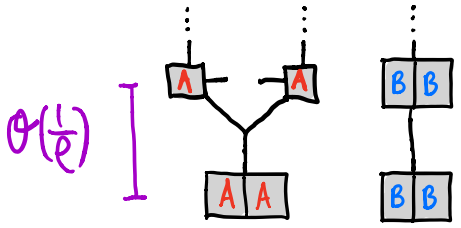
⇒ can we simulate this process in our heads when $N\rho \gg 1$?

Start with sample: 

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

② Total recombination rate = 2ρ

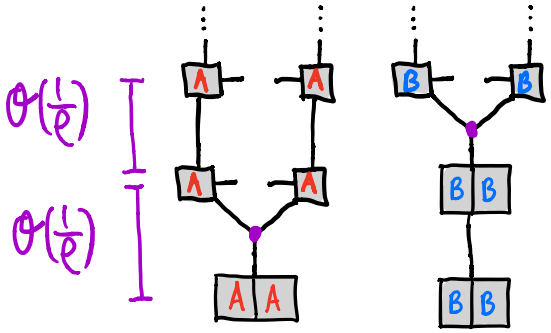
Step 2:



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

② recomb = ρ

Step 3:

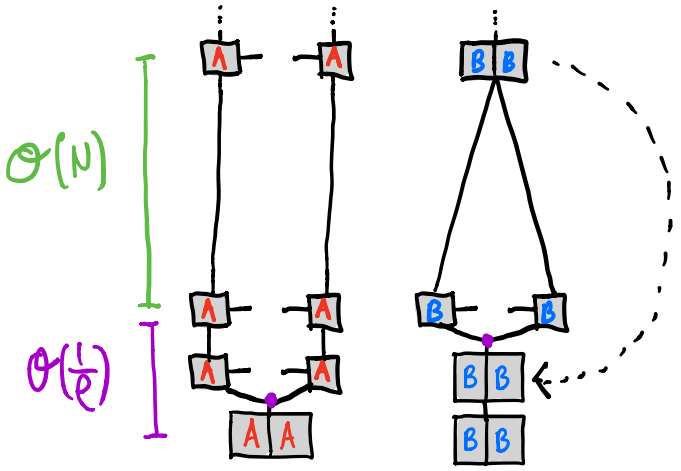


① coalescence = $\frac{1}{N}\binom{4}{2} = \frac{6}{N}$

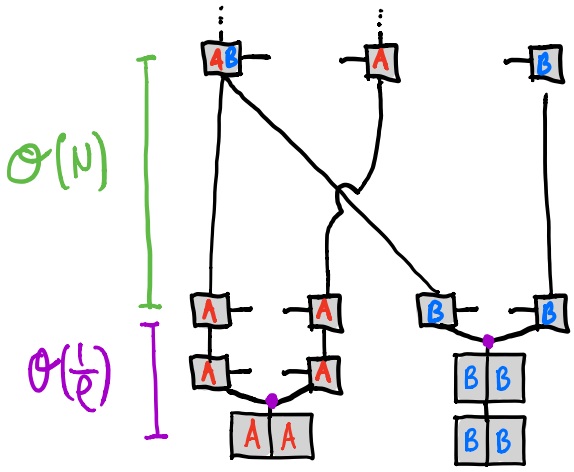
② recomb = 0

2 different types of coalescent events:

(i) coalescence of recombinant chromosomes (□+—)



(ii) coalescence involving sampled genetic material (□ + □)



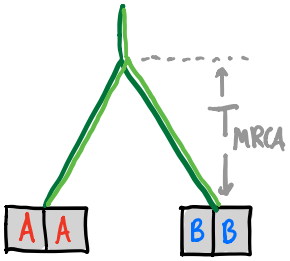
⇒ each coalescence happens ≈ independently w/ rate $1/N$.

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

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

Putting everything together:

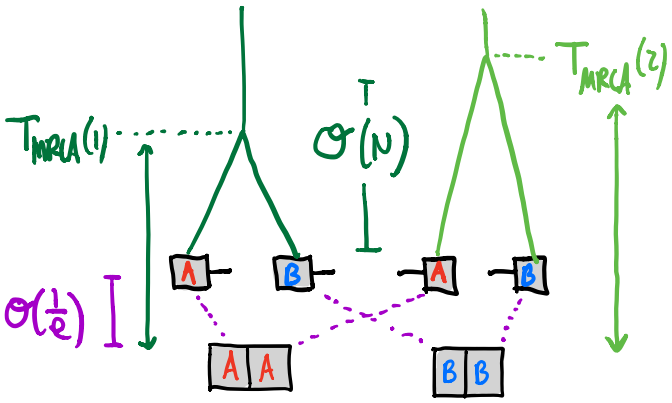
$Ne \ll 1$ (effectively asexual)



$T_{MRCA}(1) = T_{MRCA}(2)$

$\sim$ Exponential(N)

$Ne \gg 1$ (effectively independent)



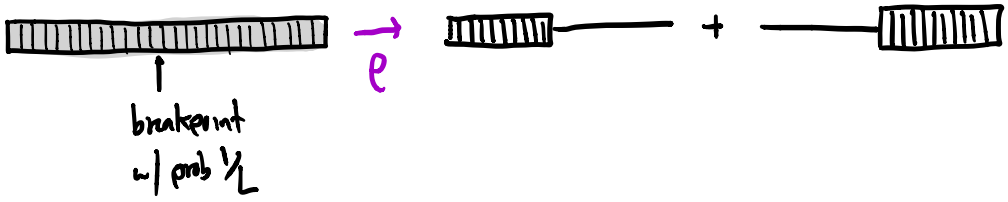
$T_{MRCA}(1), T_{MRCA}(2) \overset{iid}{\sim}$ Exponential(N)

$\Rightarrow$ same idea works for >2 chromosomes:

e.g. 

⇒ also works for other forms of recombination:

e.g. crossover:



⇒ 2nd event 

e.g. HGT / gene conversion:

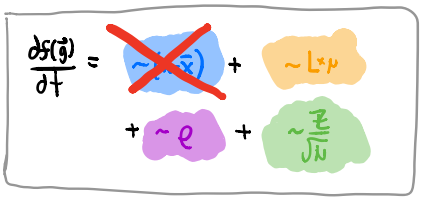
↳ 2nd 

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

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

⇒ again, hard to add

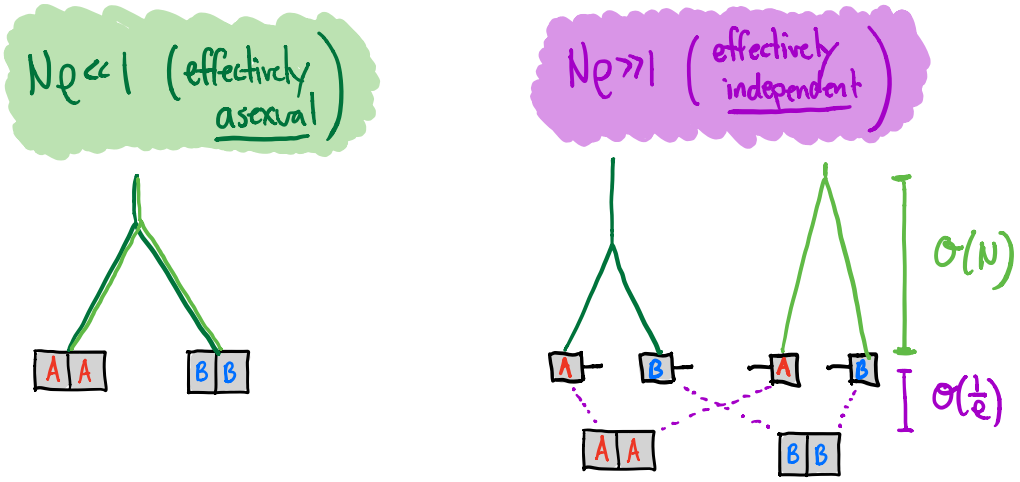
selection back to picture...



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

Quasi Linkage Equilibrium

Key result from coalescent picture:



$\Rightarrow$ can we apply similar idea in forward-time picture?

[working directly w/ $f(\vec{g},t)$]

$\Rightarrow$ to start, consider 2-locus model

w/o selection or mutation (i.e. genotypes already exist)

$\Rightarrow$ 4 genotypes: $\vec{g} = (0,0),(1,0),(0,1),(1,1)$

$\Rightarrow$ 4 genotype freqs: $f_{00}, f_{10}, f_{01}, f_{11}$

Multi-locus SDEs reduce to:

(i) $\frac{\partial f_{11}}{\partial t} = \underbrace{\rho\left[f_{10}f_{01} - f_{11}f_{00}\right]}_{\text{recombination}} + \sqrt{\frac{f_{11}}{N}}\eta_{11} - f_{11}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$ genetic drift

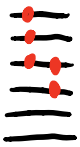
(ii) $\frac{\partial f_{10}}{\partial t} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{10}}{N}}\eta_{10} - f_{10}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

(ii) $\frac{\partial f_{01}}{\partial t} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{01}}{N}}\eta_{01} - f_{01}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

(iv) $\frac{\partial f_{00}}{\partial t} = \rho\left[f_{10}f_{01} - f_{11}f_{00}\right] + \sqrt{\frac{f_{00}}{N}}\eta_{00} - f_{00}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

$\Rightarrow$ Present day sample = Multinomial $(n, \vec{f})$

$(n_{11}, n_{10}, n_{01}, n_{00})$



$\Rightarrow$ Note: only 3 independent eqs (since $f_{11}+f_{10}+f_{01}+f_{00}=1$)

$\Rightarrow$ can eliminate $f_{00} = 1 - f_{11} - f_{10} - f_{01}$

& work w/ f_{11}, f_{10}, f_{01}

key idea: f_{11}, f_{10}, f_{01} is not only basis we can work with...

⇒ one alternative that is often used:

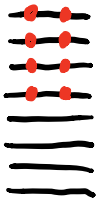
$$f_1 \equiv f_{11} + f_{10} \Rightarrow$$

$$f_2 \equiv f_{11} + f_{01}$$

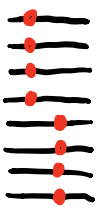
$$D \equiv f_{11} - f_1 f_2 \equiv f_{11} f_{00} - f_{10} f_{01}$$

⇒ Why? LD is measure of how double mutant deviates from model where mut'ns are independent

e.g. one high-LD scenario:



e.g. another high LD scenario:



⇒ sometimes write as correlation coefficient:

$$r \equiv \frac{D}{\sqrt{f_1(1-f_1)f_2(1-f_2)}}$$

⇒ why is f_1, f_2, D a good basis?

⇒ Rewrite SDEs:

$$\frac{df_1}{dt} =$$