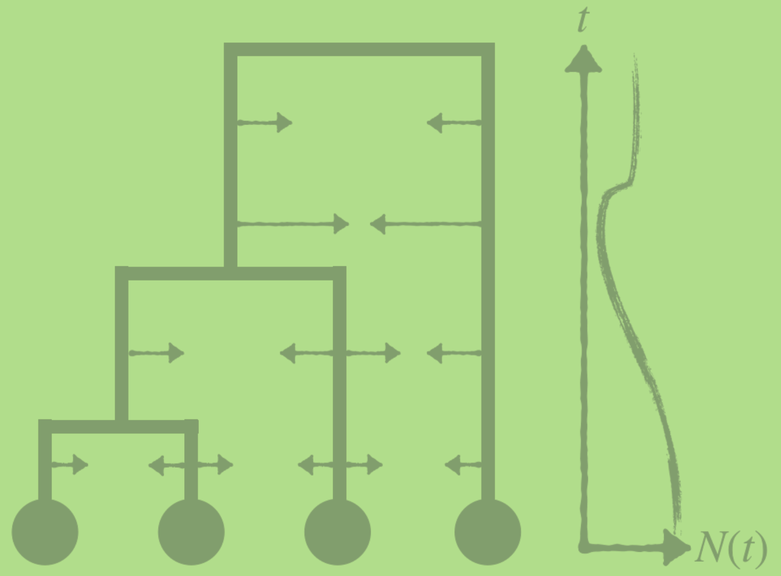
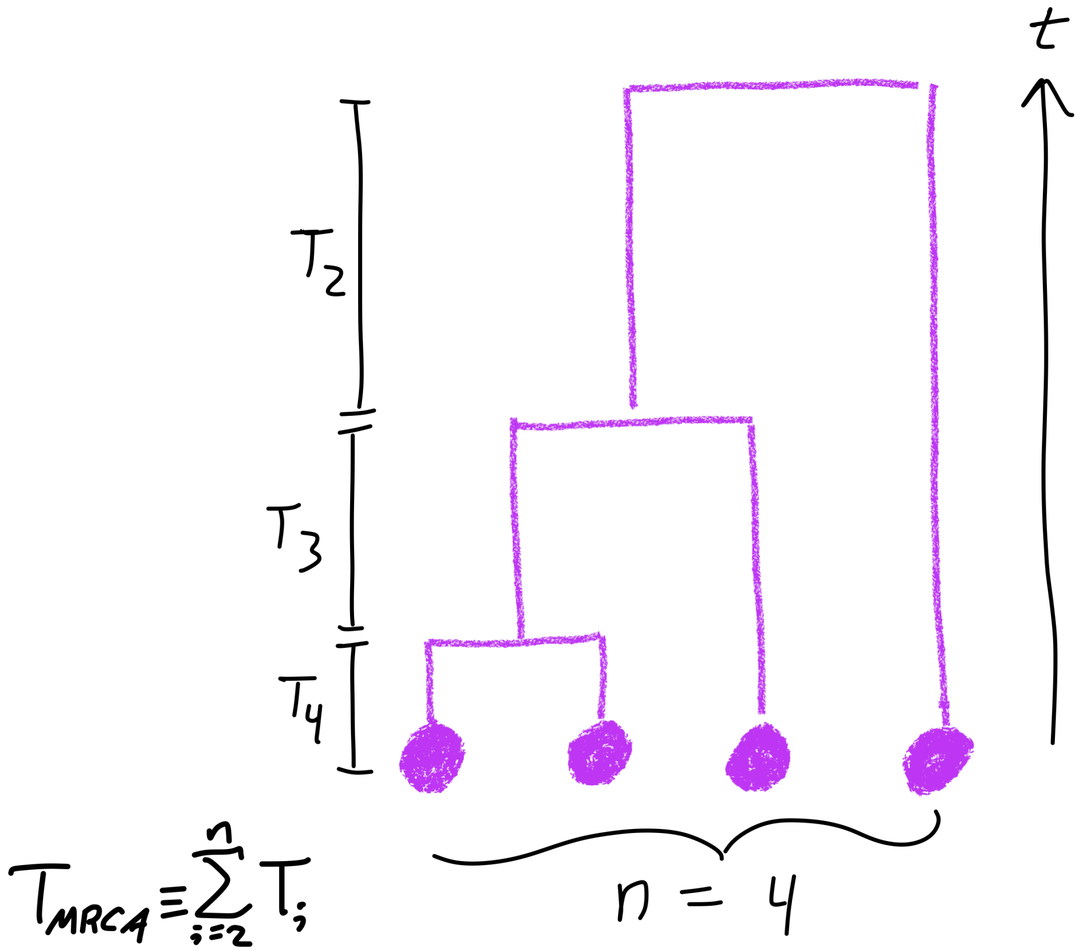
GENOME 541: population genetic inference

II. The coalescent with recombination



Previously on...



"intercoalescence times"

$$\mathbb{E}[T_i] = \frac{2N}{\binom{i}{2}}$$

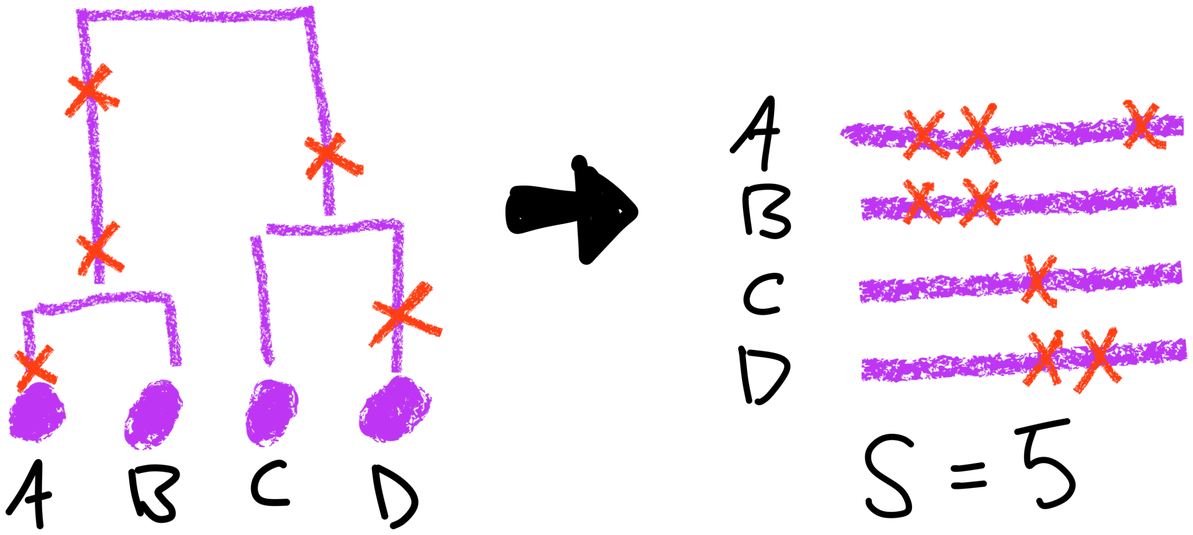
Exp. dist.

$$p(t_i) = \frac{\binom{i}{2}}{2N} e^{-\frac{\binom{i}{2}}{2N} t_i}$$

- Each pair is a process w/ rate $\frac{1}{2N}$
- The pairs race to coalesce

Previously on...

segregating sites, S, equals # mutations in the sample's history (infinite sites approximation)



Constant N case:

total branch length

$$\mathbb{E}[S] = \mu \mathbb{E}[T_{total}]$$

$$= \mu \sum_{i=2}^{n} i \mathbb{E}[T_i]$$

$$= \mu \sum_{i=2}^{n} i \frac{2N}{\binom{i}{2}}$$

$$= 4\mu N \sum_{i=1}^{n-1} \frac{1}{i}$$

Genetic diversity stats

 # segregating sites, S, equals # mutations in the sample's history (infinite sites approximation):

$$\mathbb{E}[S] = 4N\mu \sum_{i=2}^{n-2} \frac{1}{i}$$

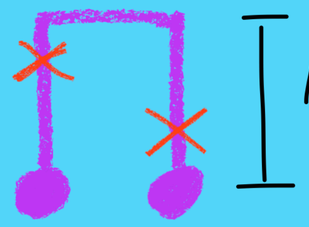
Pairwise divergence, π, # mutations in the history of two sampled haplotypes:

$$\mathbb{E}[\pi] = 4N\mu$$

Tajima's D: null hypothesis, standard neutral coalescent (constant N)

$$\frac{\mathbb{E}[S]}{\sum_{i=2}^{n-2} 1/i} - \mathbb{E}[\pi] = 0$$

Derivation:

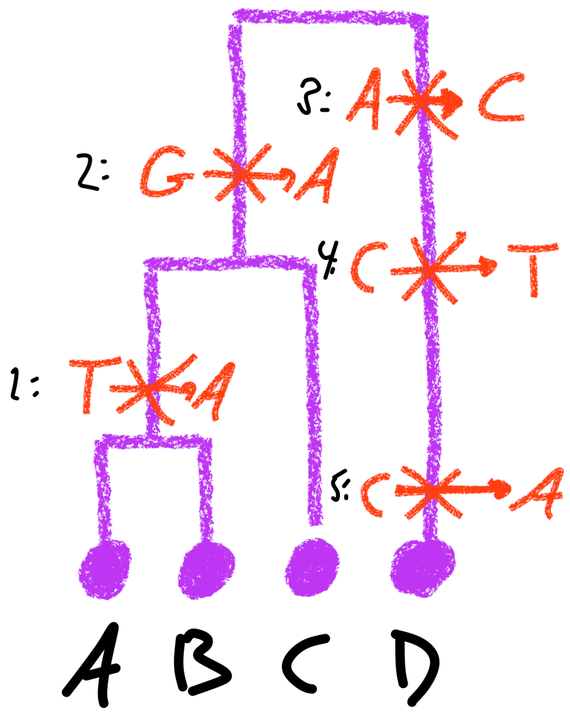


$$\mathbb{E}[T_2] = \frac{1}{1/2N} = 2N$$

$$\Rightarrow \mathbb{E}[\pi] = 4N\mu$$

Question: why 4?

Example



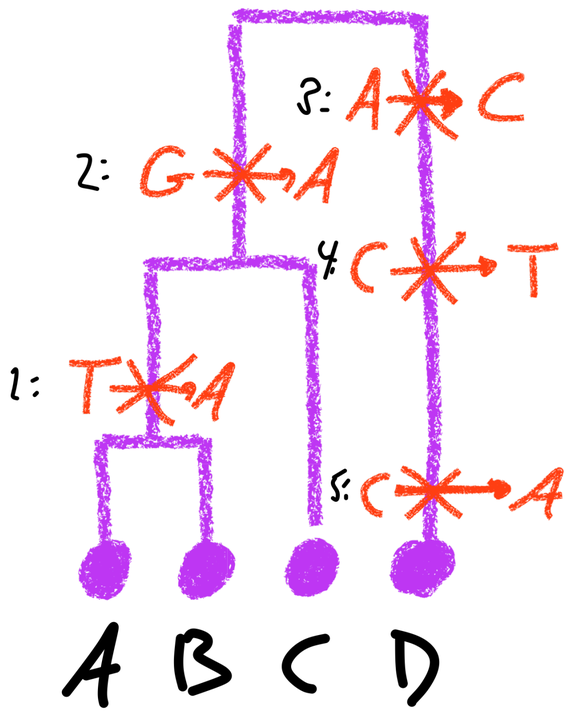
	1	2	3	4	5
A	A	A	A	C	C
B	A	A	A	C	C
C	T	A	A	C	C
D	T	G	C	T	A

$$S = 5$$

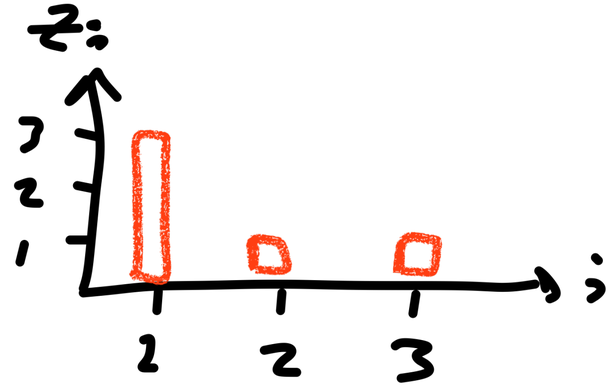
$$\pi = \frac{1}{\binom{4}{2}}\left(0+2+5+2+5+4\right) = \frac{8}{3}$$

$$D = \frac{5}{1+\frac{1}{2}+\frac{1}{3}} - \frac{8}{3} = \cdots \frac{2}{33}$$

Example: sample frequency spectrum (SFS)



	1	2	3	4	5
A	A	A	A	C	C
B	A	A	A	C	C
C	T	A	A	C	C
D	T	G	C	T	A



SFS: histogram of mutant allele frequencies

$\vec{z} = [z_1, z_2, \ldots, z_{n-2}]$, z_i = #mutations with frequency i in sample

Theory: sample frequency spectrum (SFS)

SFS: histogram of mutant allele frequencies

$\vec{z} = [z_1, z_2, \ldots, z_{n-2}]$, z_i = # mutations with frequency i in sample

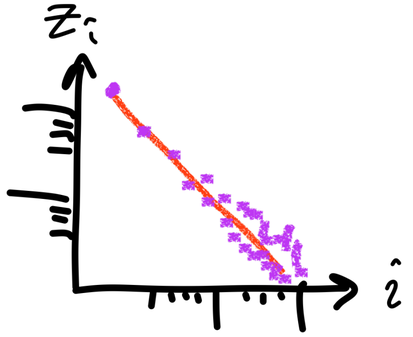
With a bit of work...

$$\mathbb{E}[z_i] = \frac{4N\mu}{i}$$

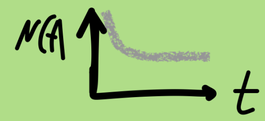
(for constant N) $i = 1, 2, \ldots, n-1$

$$\log z_i = \log\left(\frac{4N\mu}{i}\right)$$

$$= \log 4N\mu - \log i$$

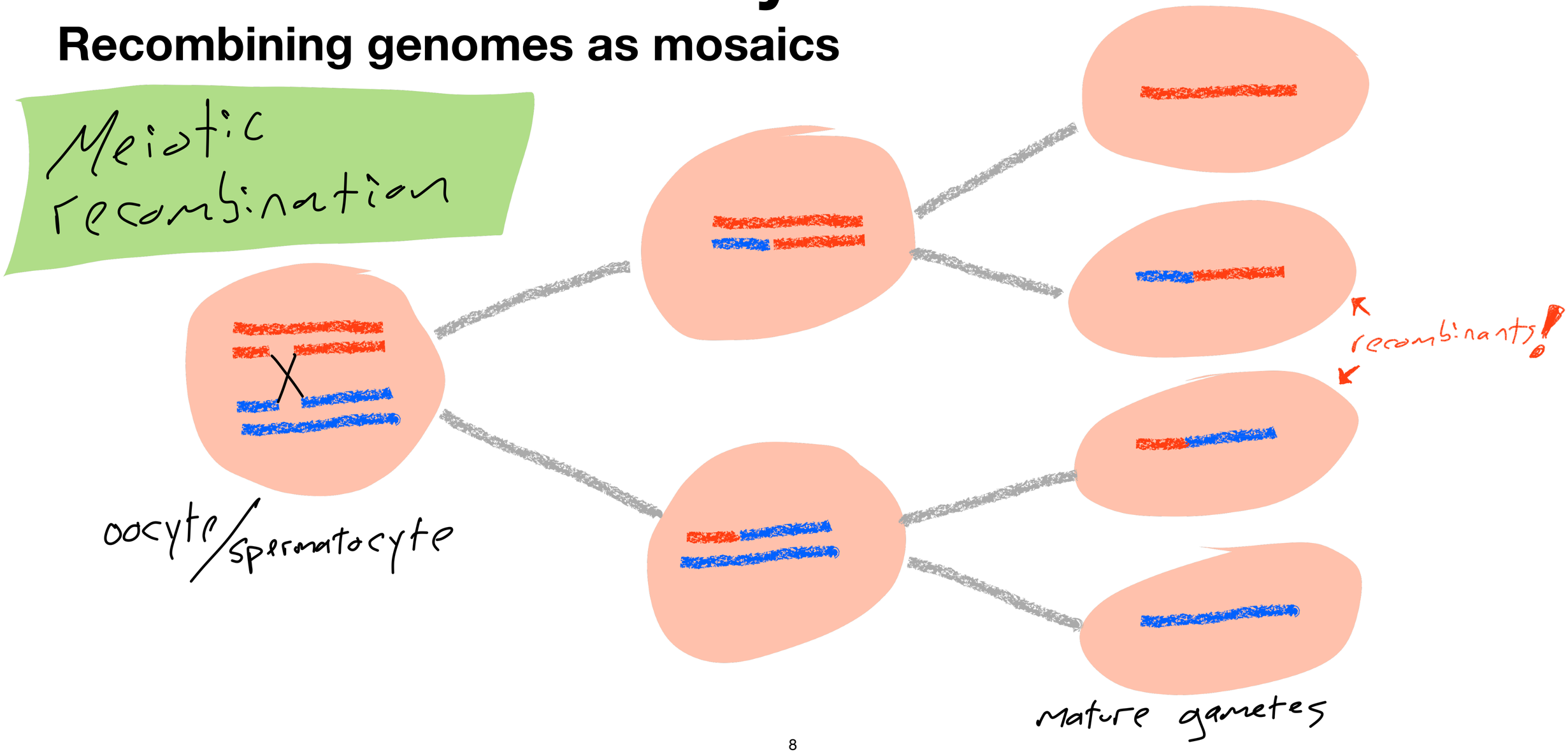


Question: what happens to SFS if $N(t)$ is changing?



Non tree-like ancestry

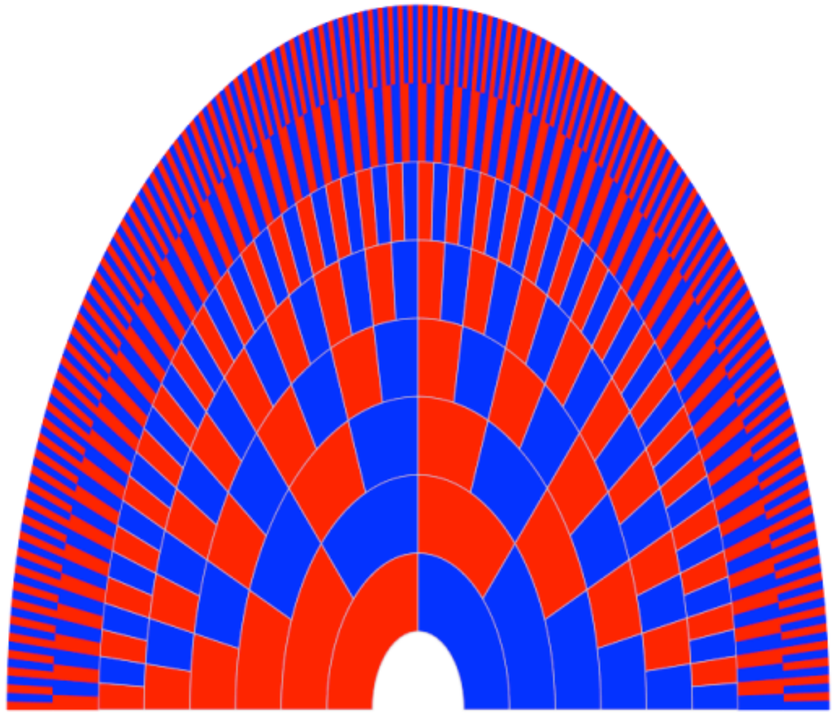
Recombining genomes as mosaics



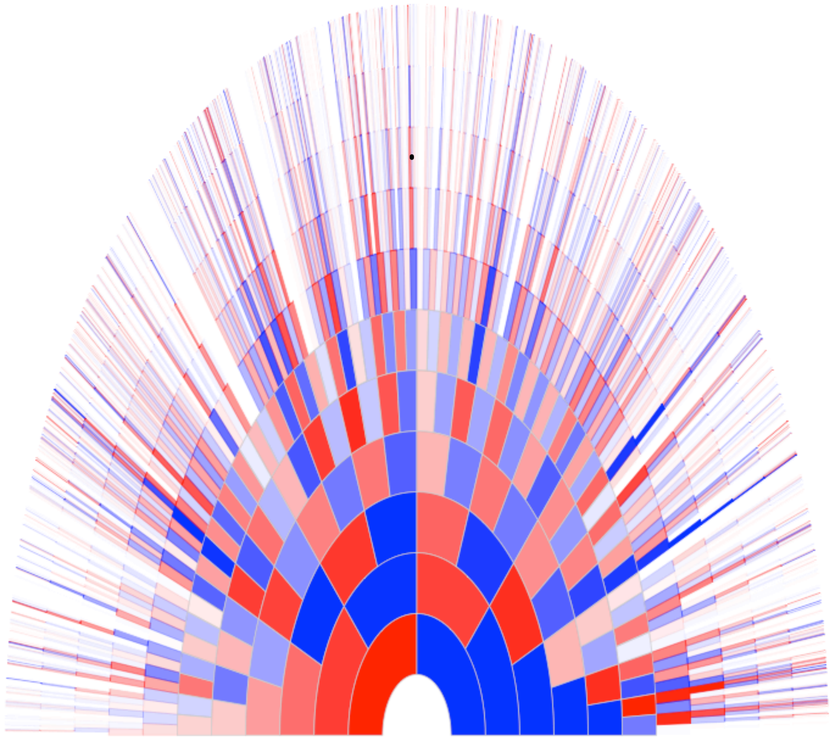
Non tree-like ancestry

Recombining genomes as mosaics

genealogical ancestry



genetic ancestry

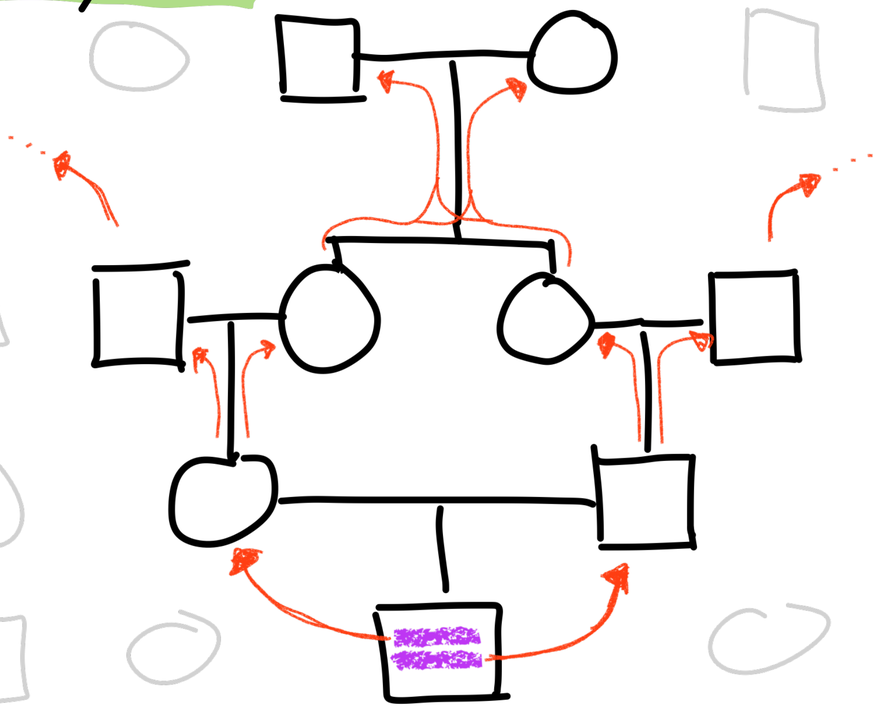


https://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/

Non tree-like ancestry

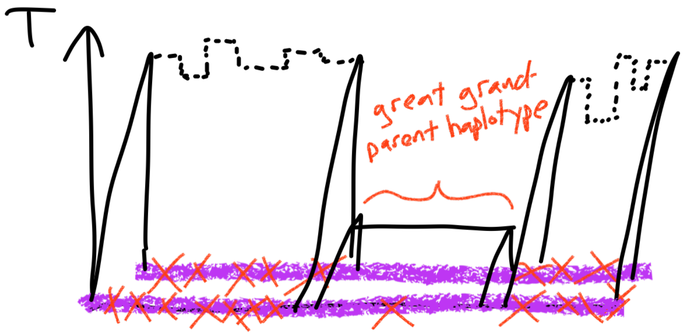
Recombining genomes as mosaics

1st-cousin mating



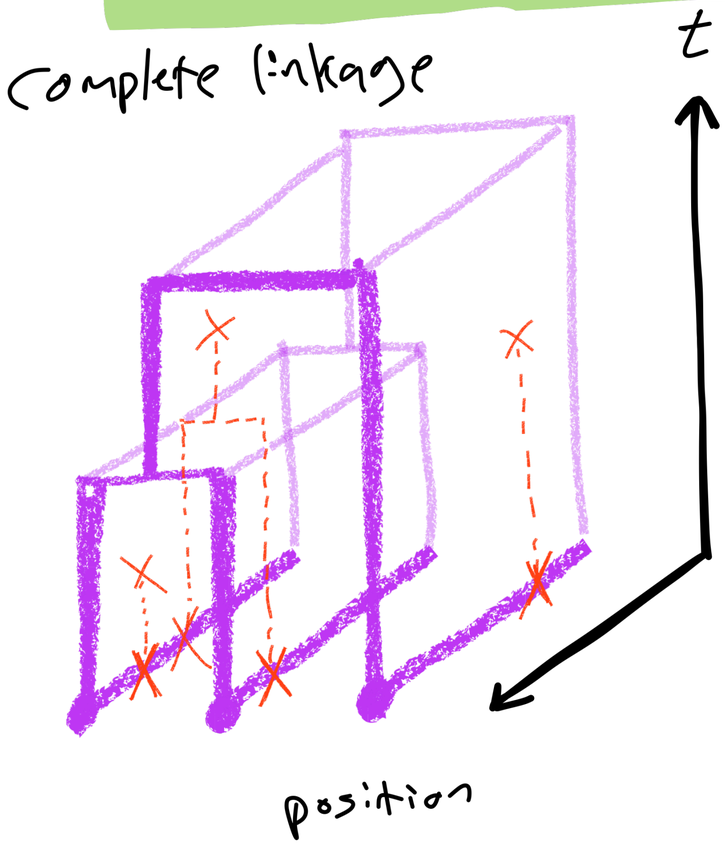
Question:

- probability [?] to coalesce in great grandparent?
- Otherwise much deeper coalescence

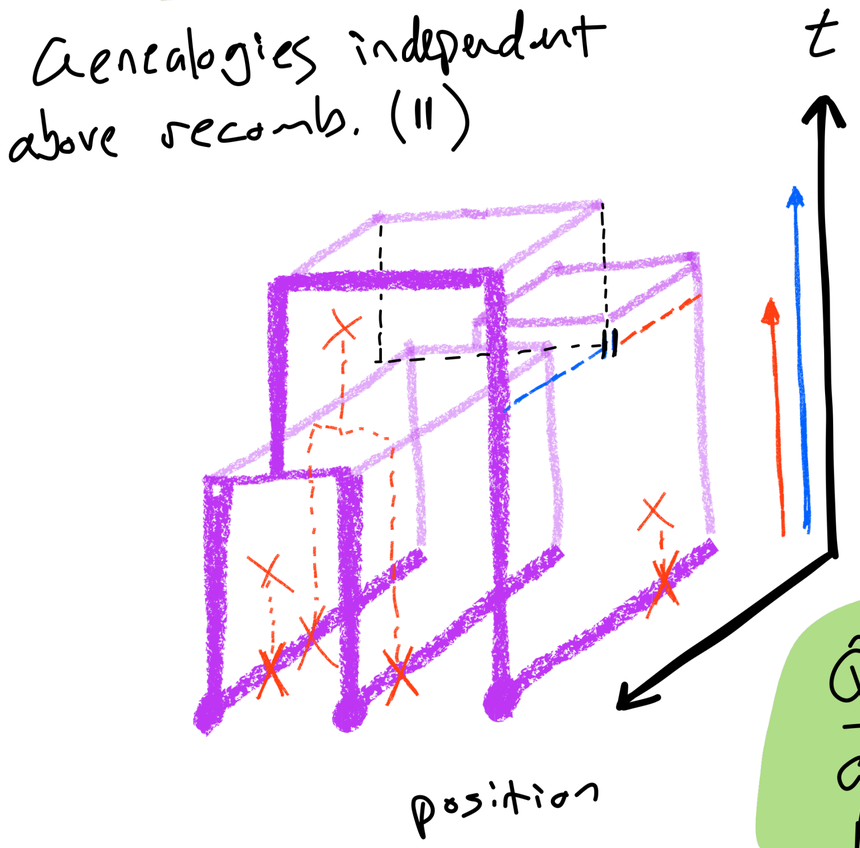


The coalescent with recombination

Without recombination:
coalescences and mutations



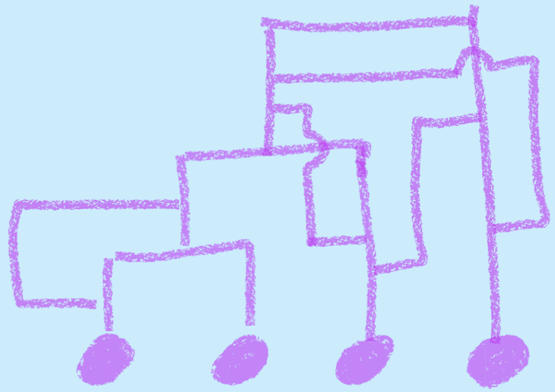
With recombination:
add Poisson recombination events!



Genealogies decorrelate (sample different TMRCA)

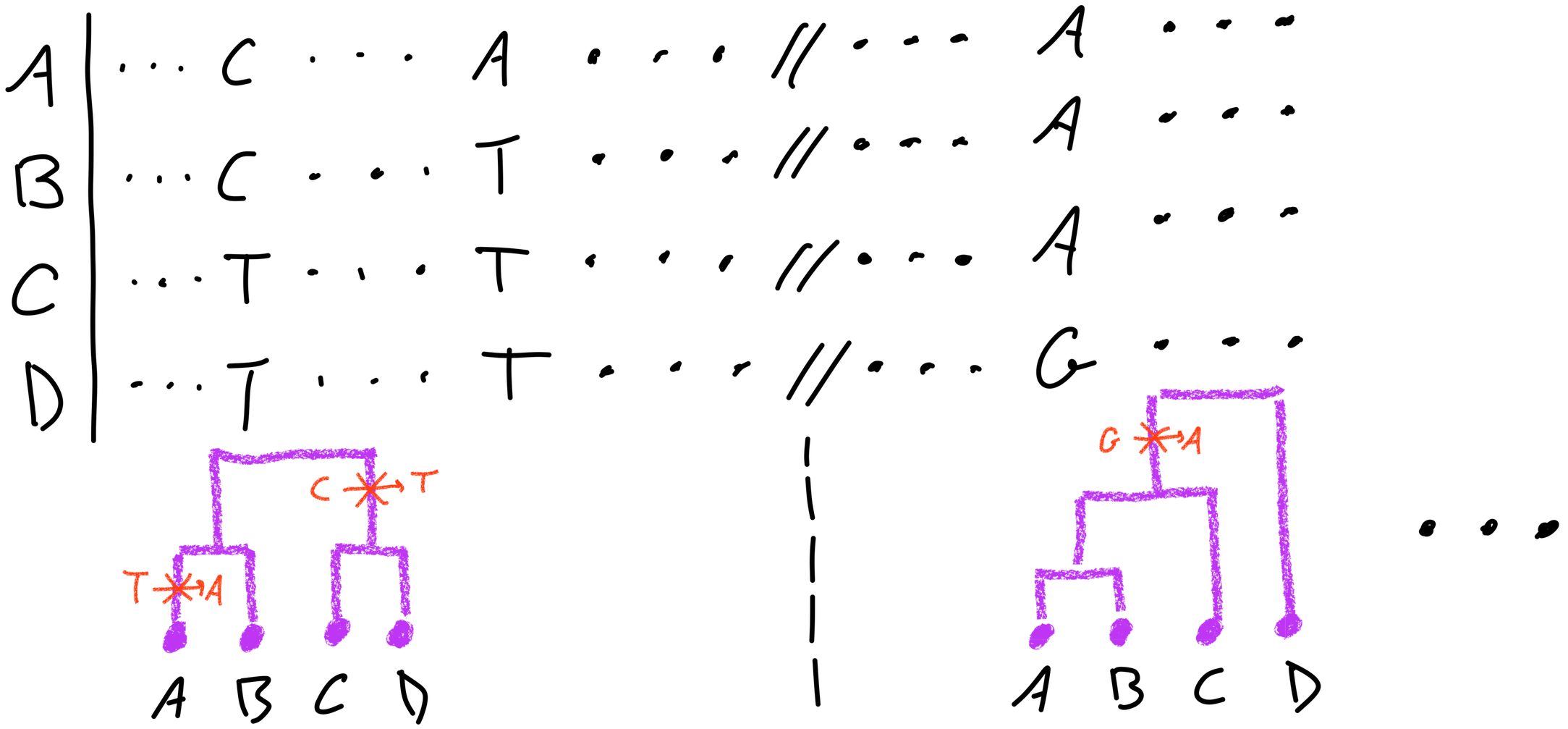
Question: does this change $\mathbb{E}[S]$, $\mathbb{E}[\pi]$, $\mathbb{E}[Z_i]$?

THE ANCESTRAL RECOMBINATION GRAPH



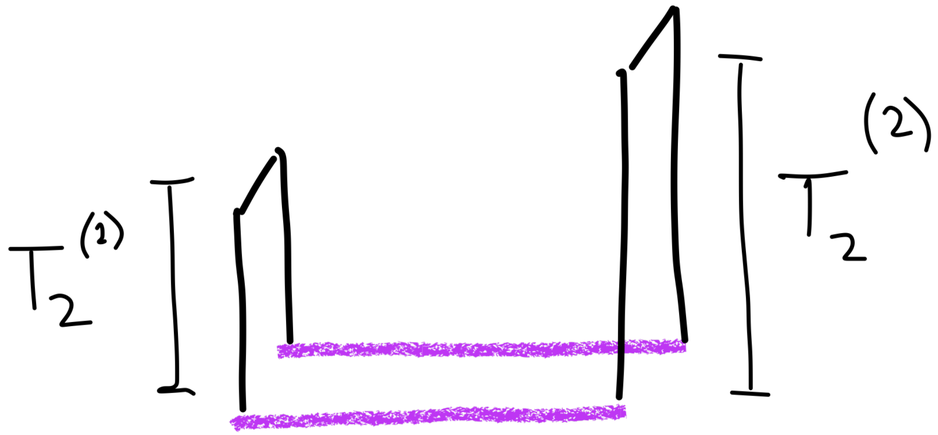
"ARG"

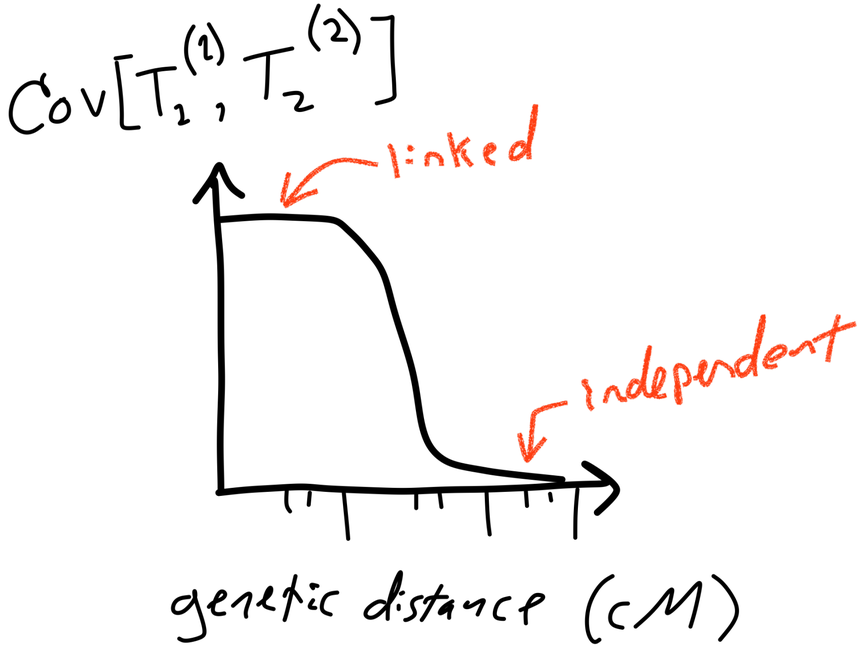
A sequence of trees



Decay of linkage

Larger genetic distance
$\Rightarrow$ recombination more likely





Deeper coalescence
$\Rightarrow$ shorter span

$$X \mid t \sim \exp(4Nrt)$$

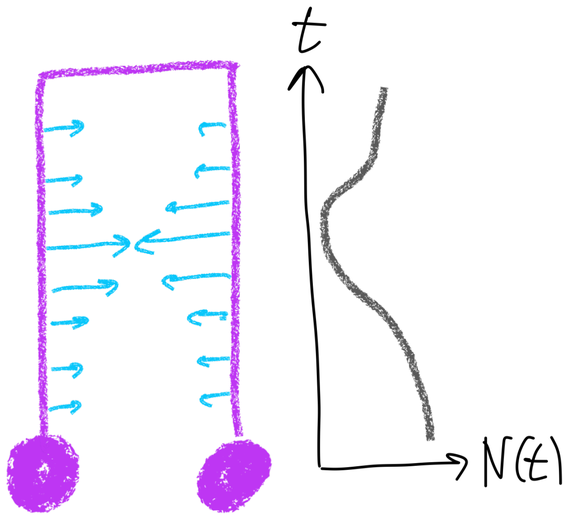
X: genetic dist. to recomb.

r: recombination rate

Previously on...

Population size determines coalescence rate

What if population size varies over time? $N(t)$



$N(t)$ distorts time scale from the standard coalescent

- time compressed when $N(t)$ is small
- time stretched when $N(t)$ is large

The details:

$$P(T_i = t_i) = \frac{\binom{i}{2}}{2N_{t_i}} \prod_{j=2}^{t_{i-1}} \left(1 - \frac{\binom{i}{2}}{2N_j}\right)$$

big N $\rightarrow$ $$p(t_i) = \frac{\binom{i}{2}}{2N(t)} e^{-\binom{i}{2}\int_0^t \frac{ds}{2N(s)}}$$

inhomogeneous Poisson process

Coalescent HMMs

Pairwise sequential Markov coalescent (PSMC)

Hidden state: T_{MRCA}

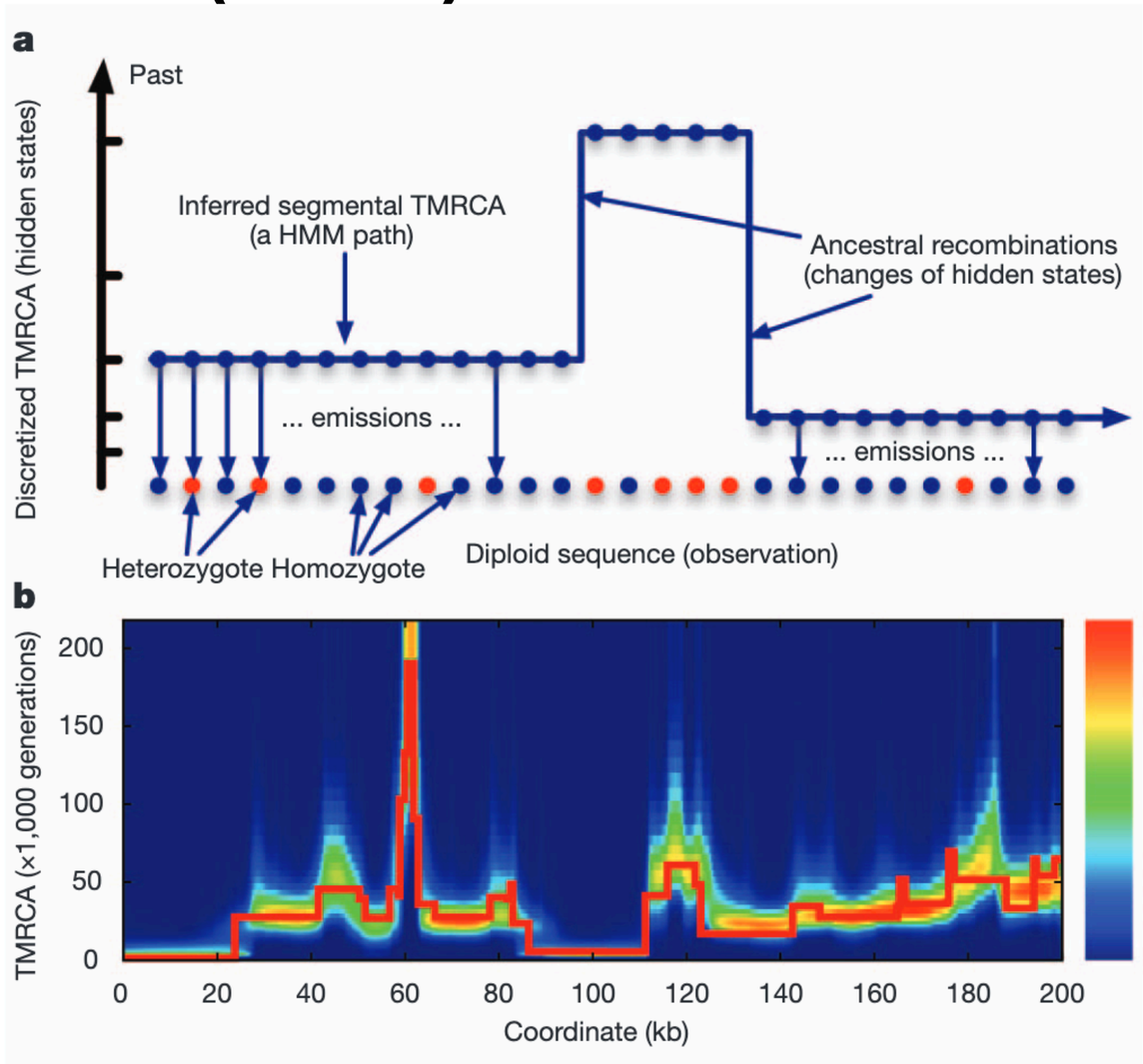
Emission: heterozygosity 00100110...

Transitions:

$$p(t|s) = \begin{cases} e^{-\rho s}, & s = t \\ (1-e^{-\rho s})\, q(t|s), & s \neq t \end{cases}$$

$\rho = 4Nr$

complicated integral of history $N(t)$



ARTICLE

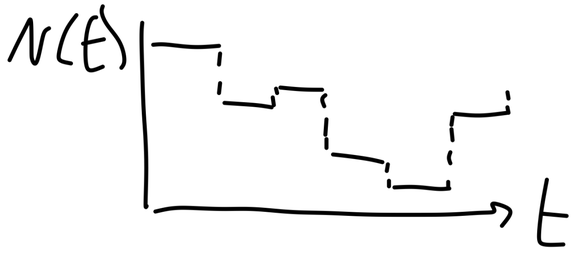
OPEN

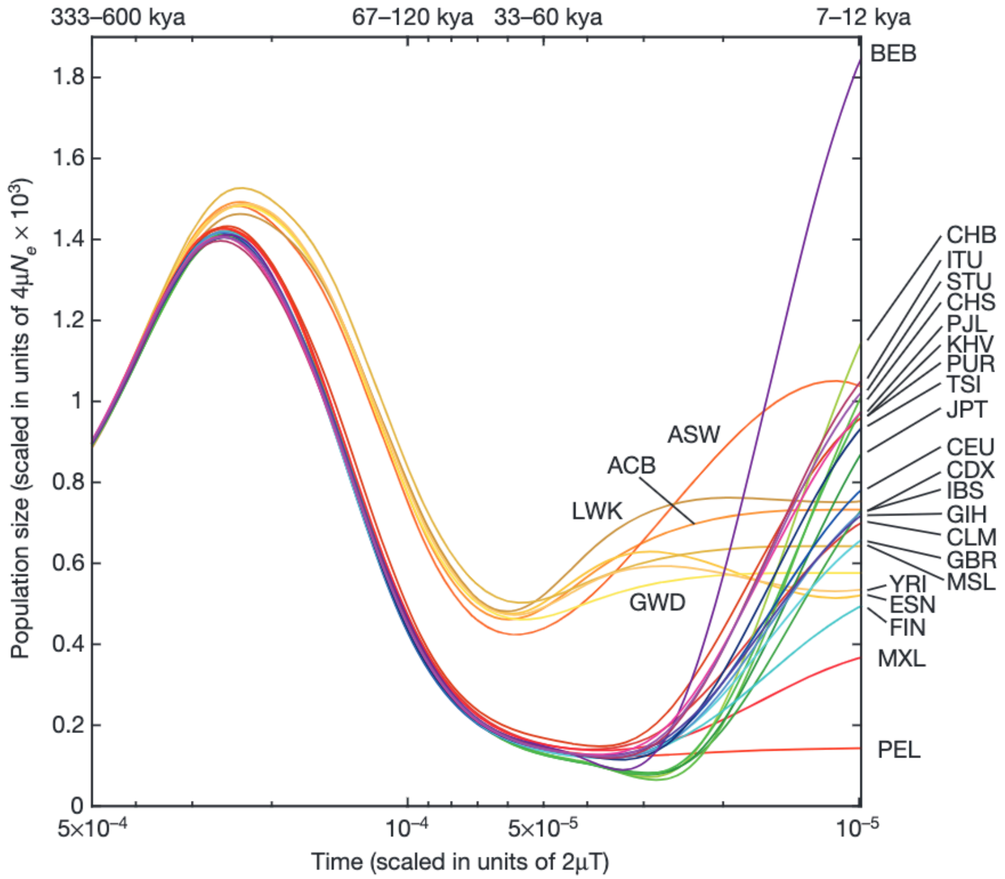
doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

Averaging indv. estimates to get pop estimate

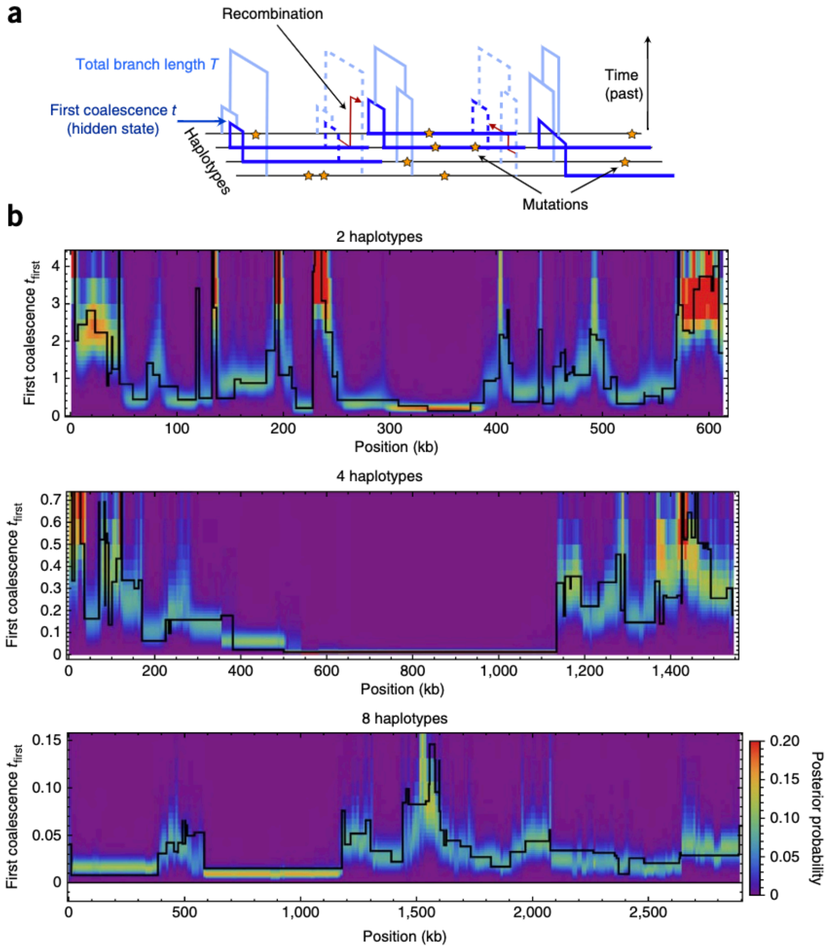




Generalizing beyond diploids

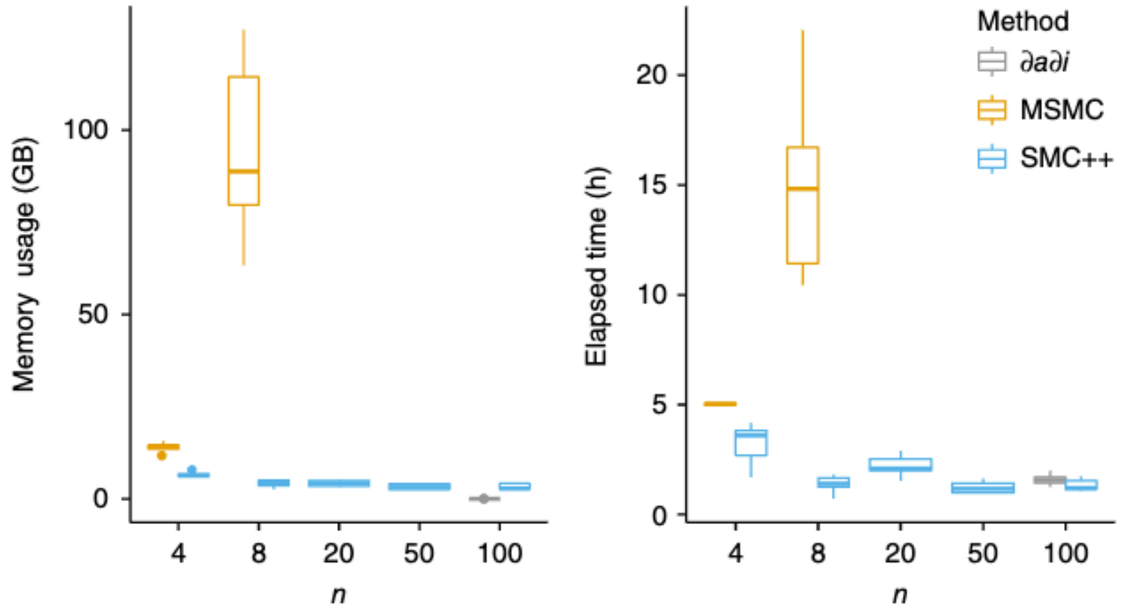
Inferring human population size and separation history from multiple genome sequences

Stephan Schiffels & Richard Durbin



Robust and scalable inference of population history from hundreds of unphased whole genomes

Jonathan Terhorst[1], John A Kamm[1,2] & Yun S Song[1–4]

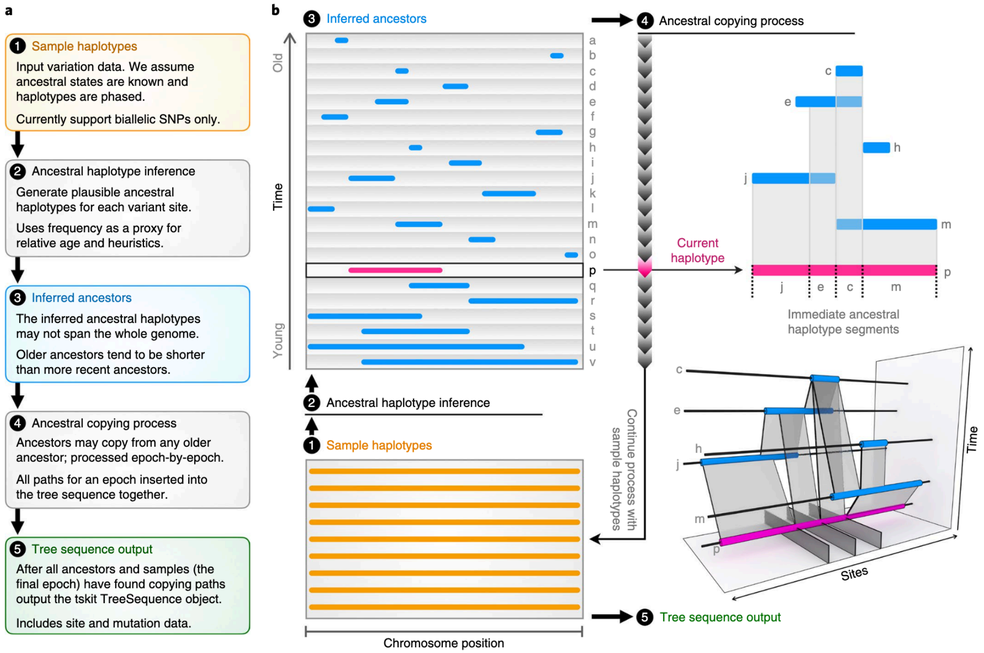


The future: tree sequence inference

ARTICLES
https://doi.org/10.1038/s41588-019-0483-y
nature genetics

Inferring whole-genome histories in large population datasets

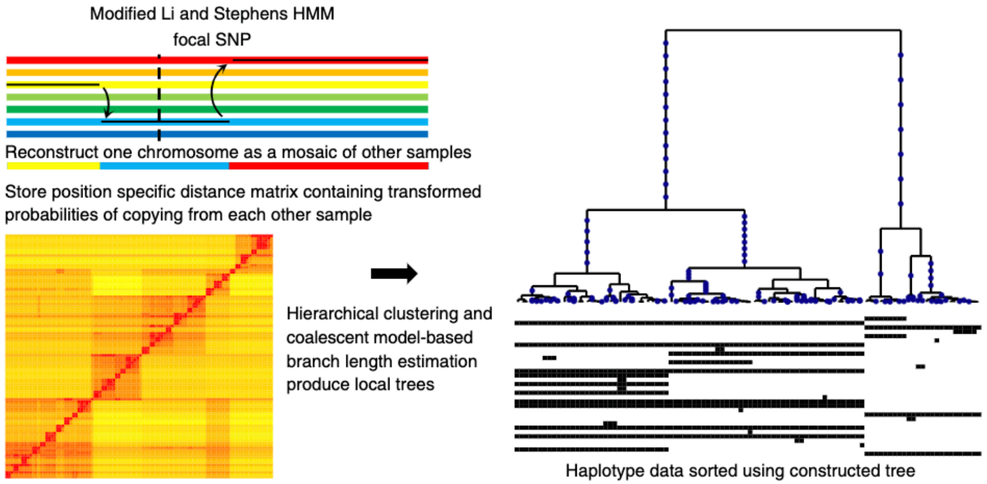
Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers and Gil McVean

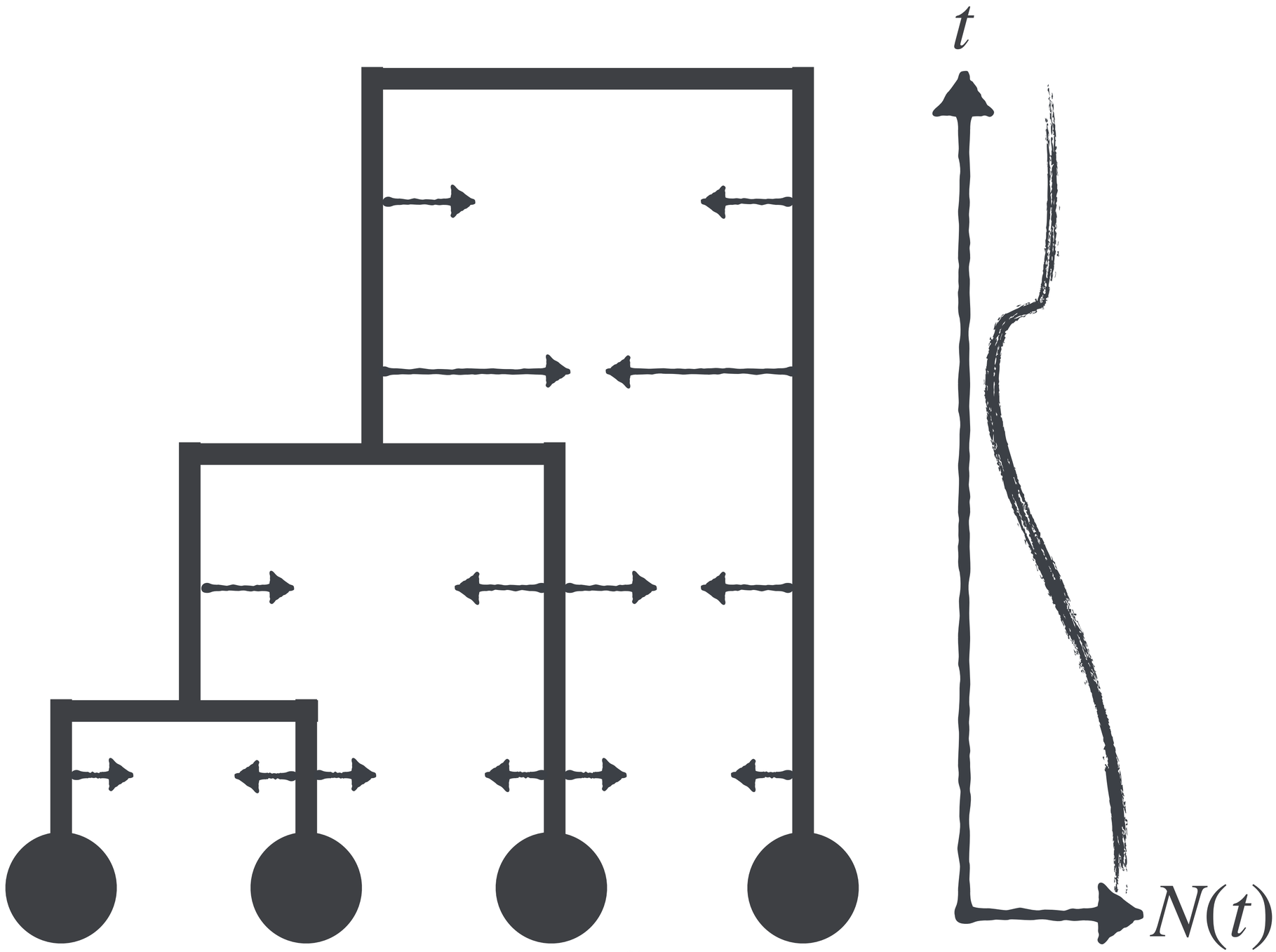


nature genetics
ARTICLES
https://doi.org/10.1038/s41588-019-0484-x

A method for genome-wide genealogy estimation for thousands of samples

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myers[1,3]*




t
N(t)