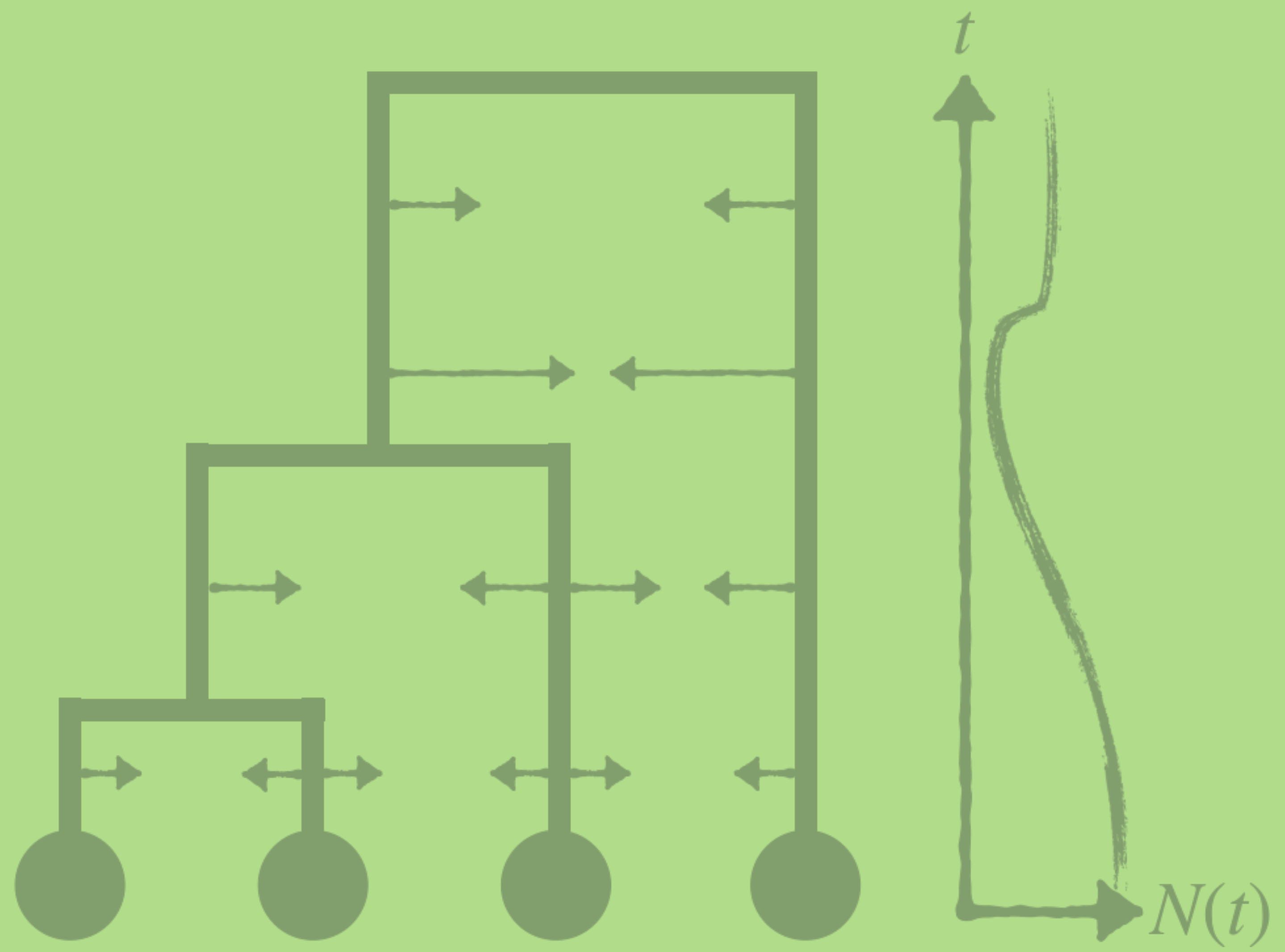
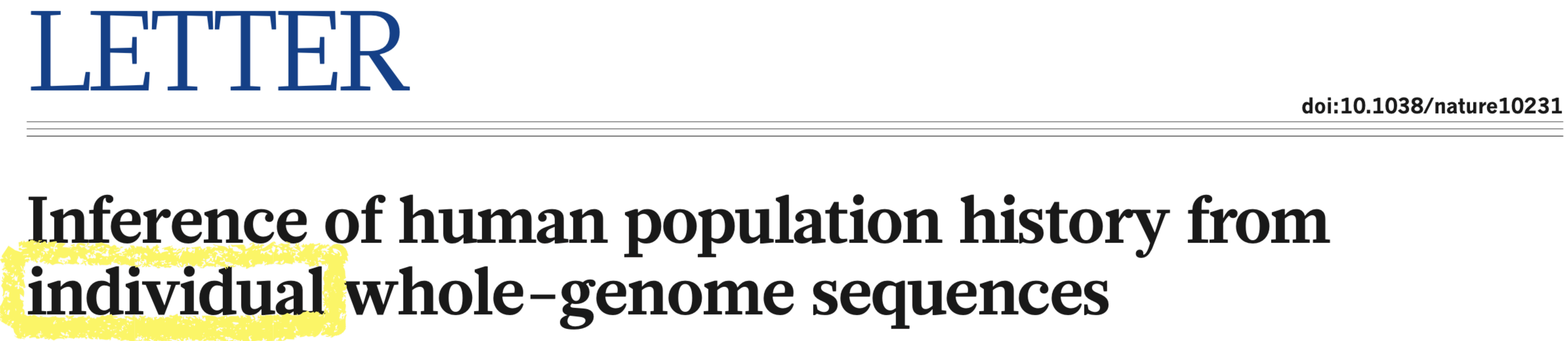
GENOME 541: population genetic inference

I. The coalescent (again)



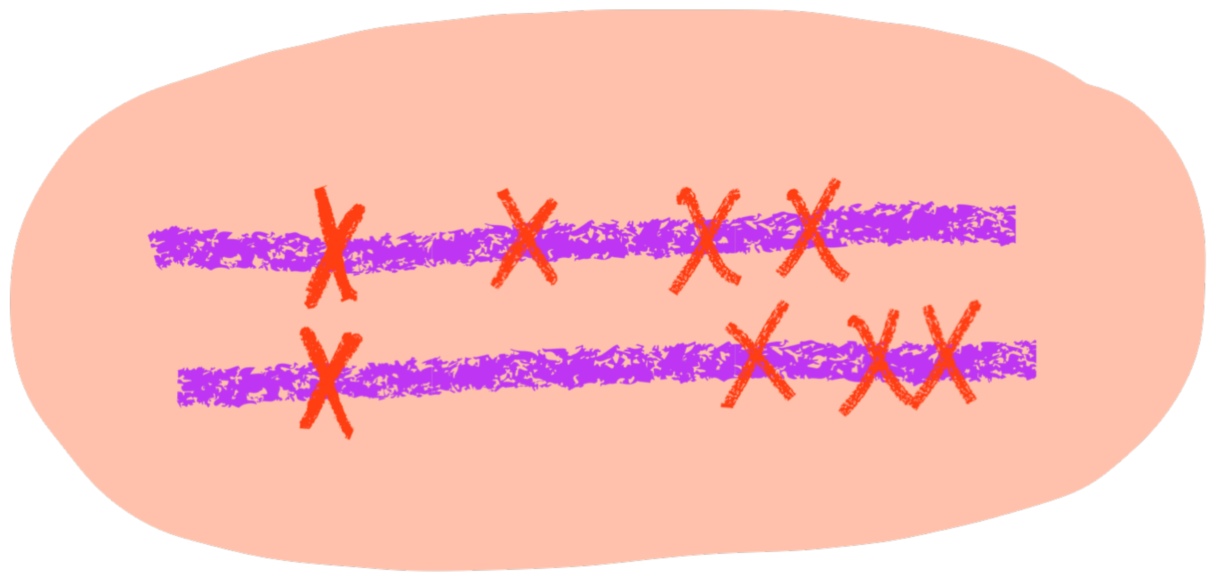


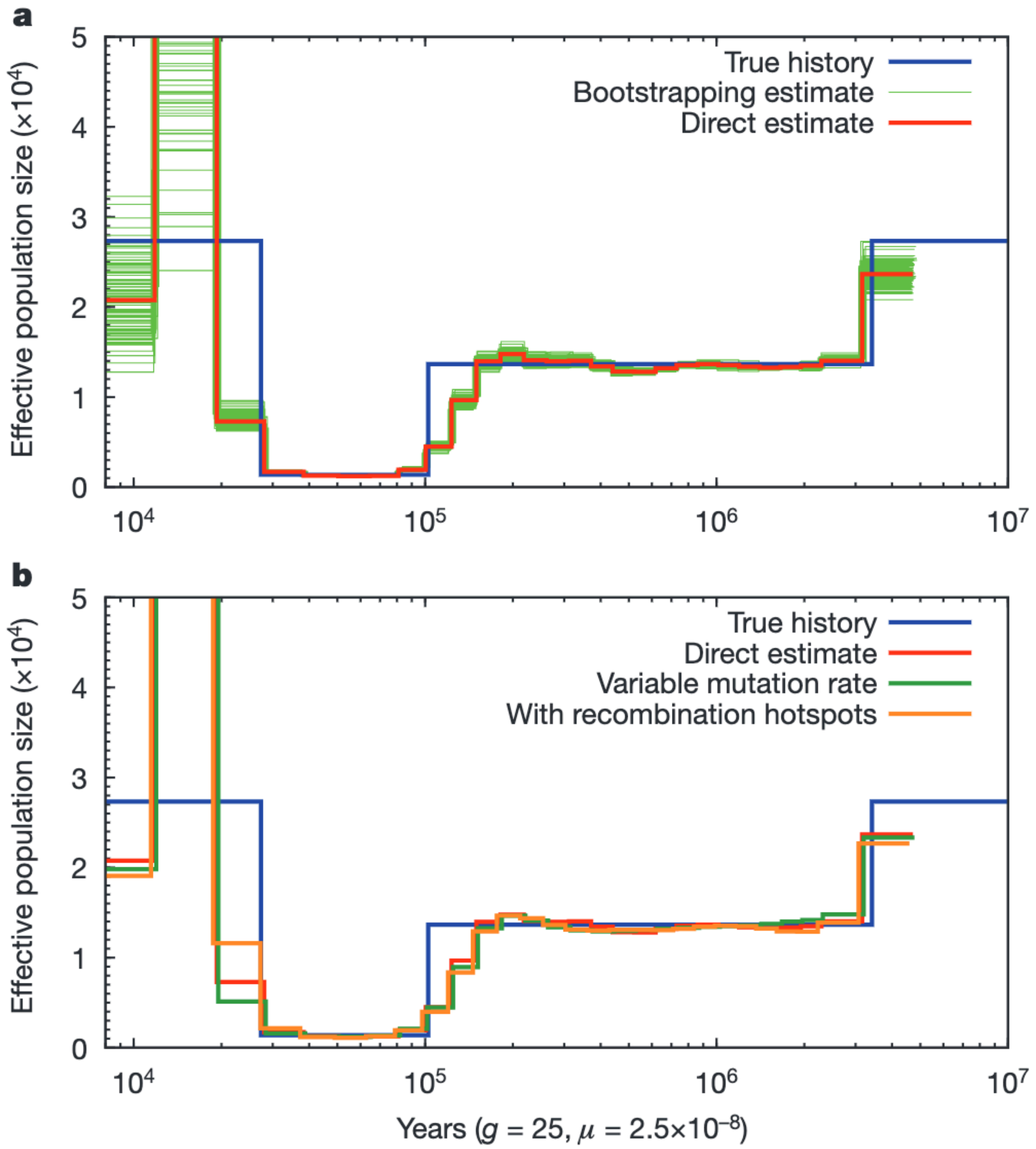
LETTER

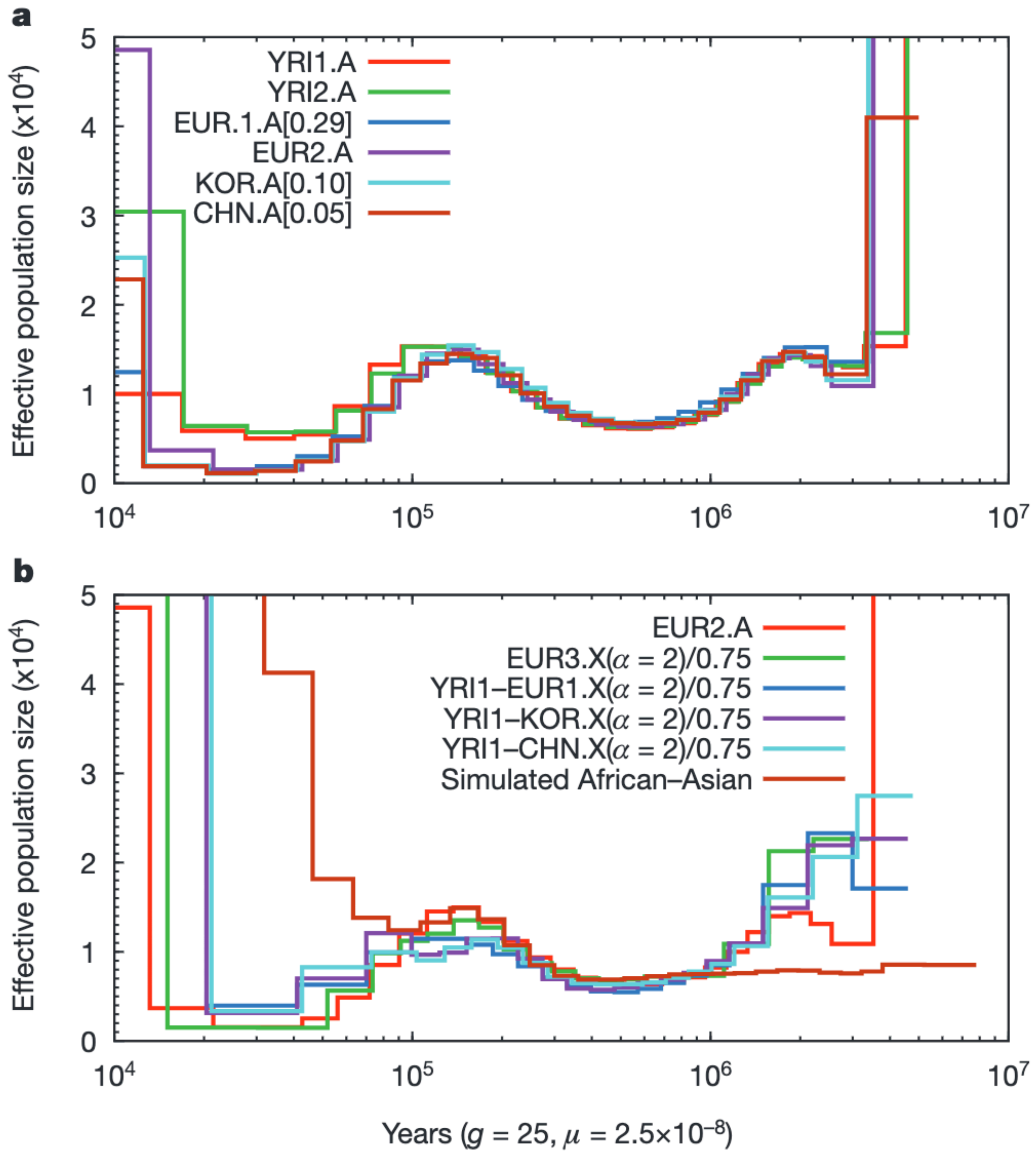
doi:10.1038/nature10231

Inference of human population history from individual whole-genome sequences

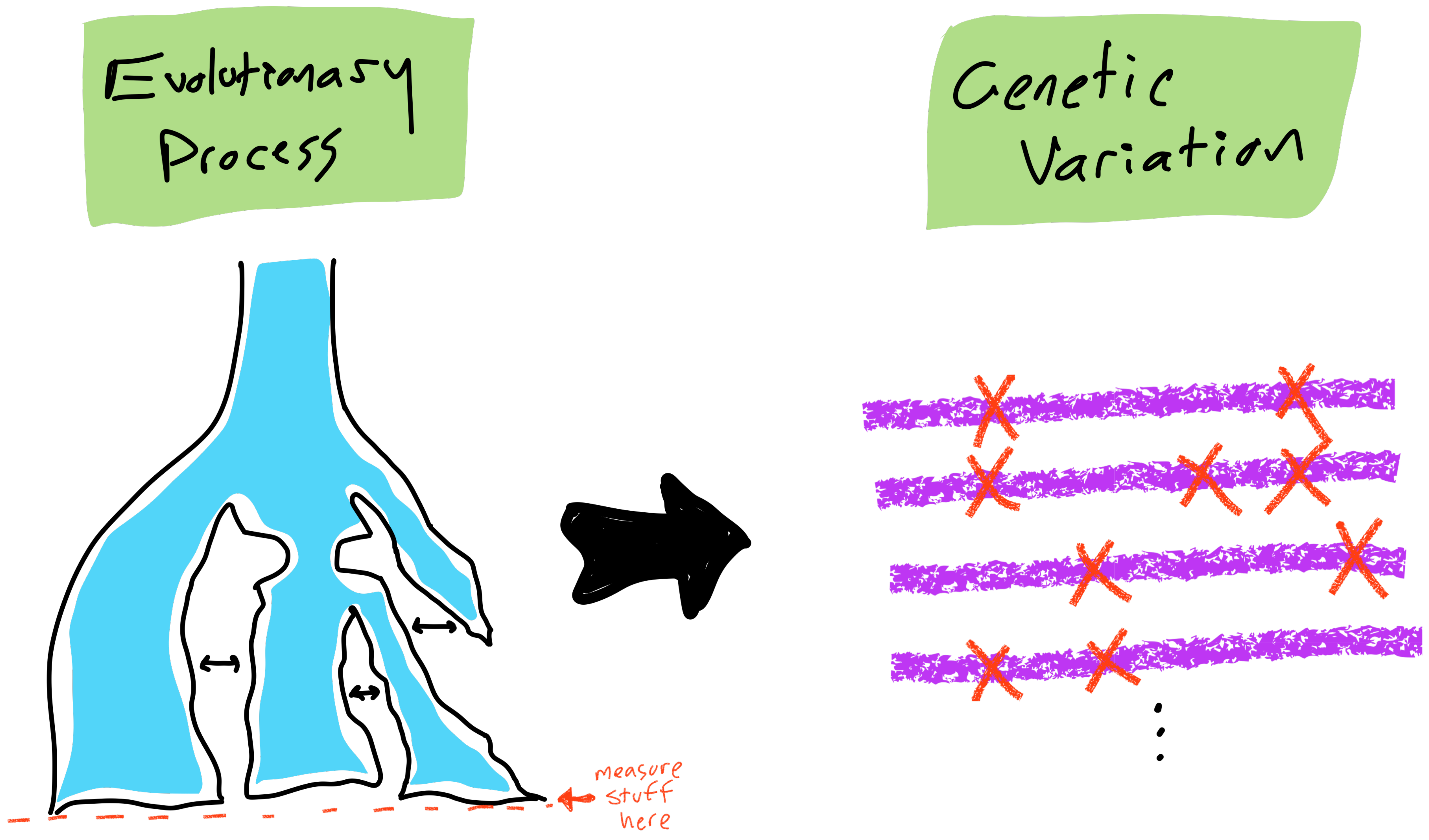
Heng Li[1,2] & Richard Durbin[1]



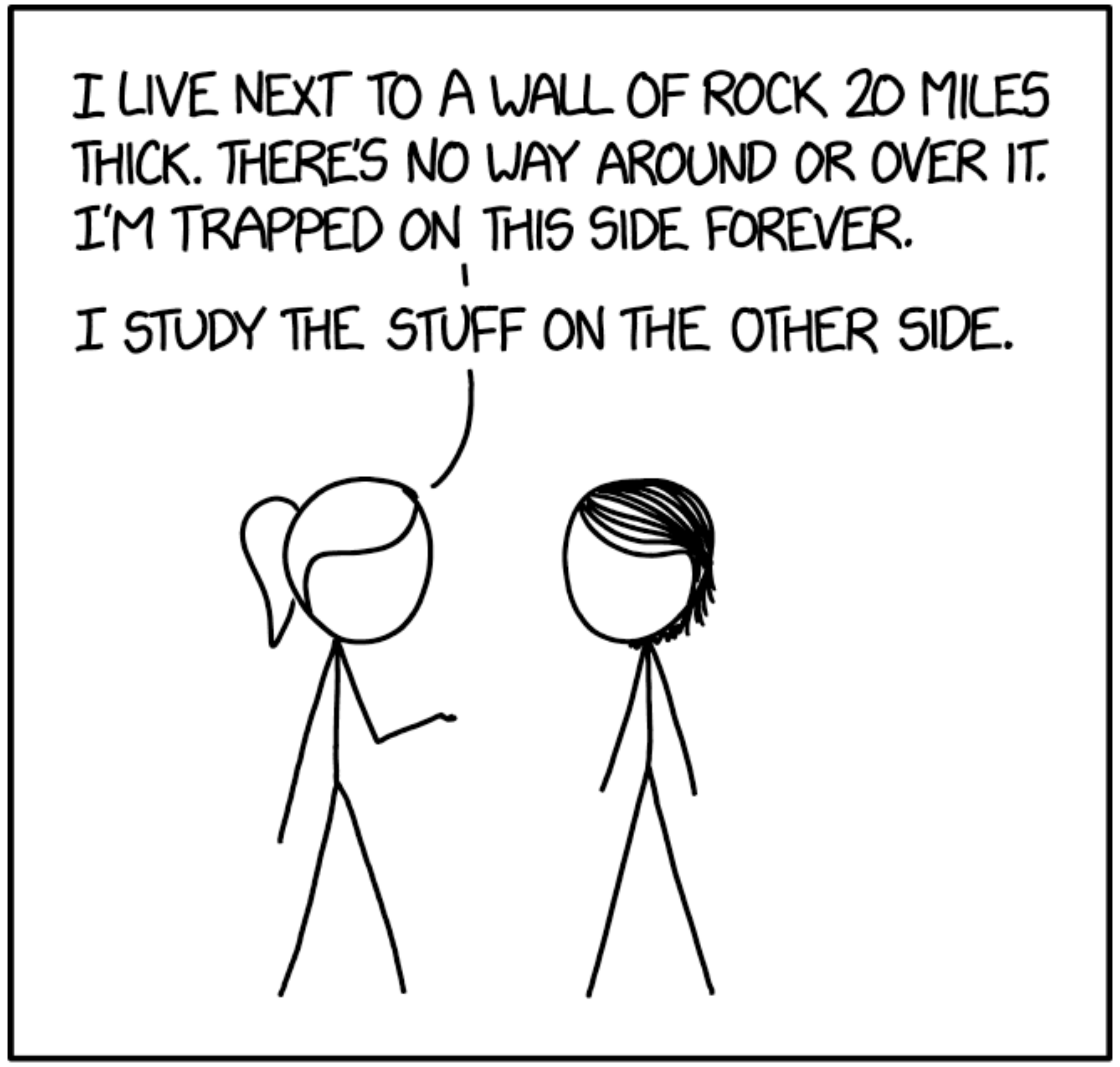




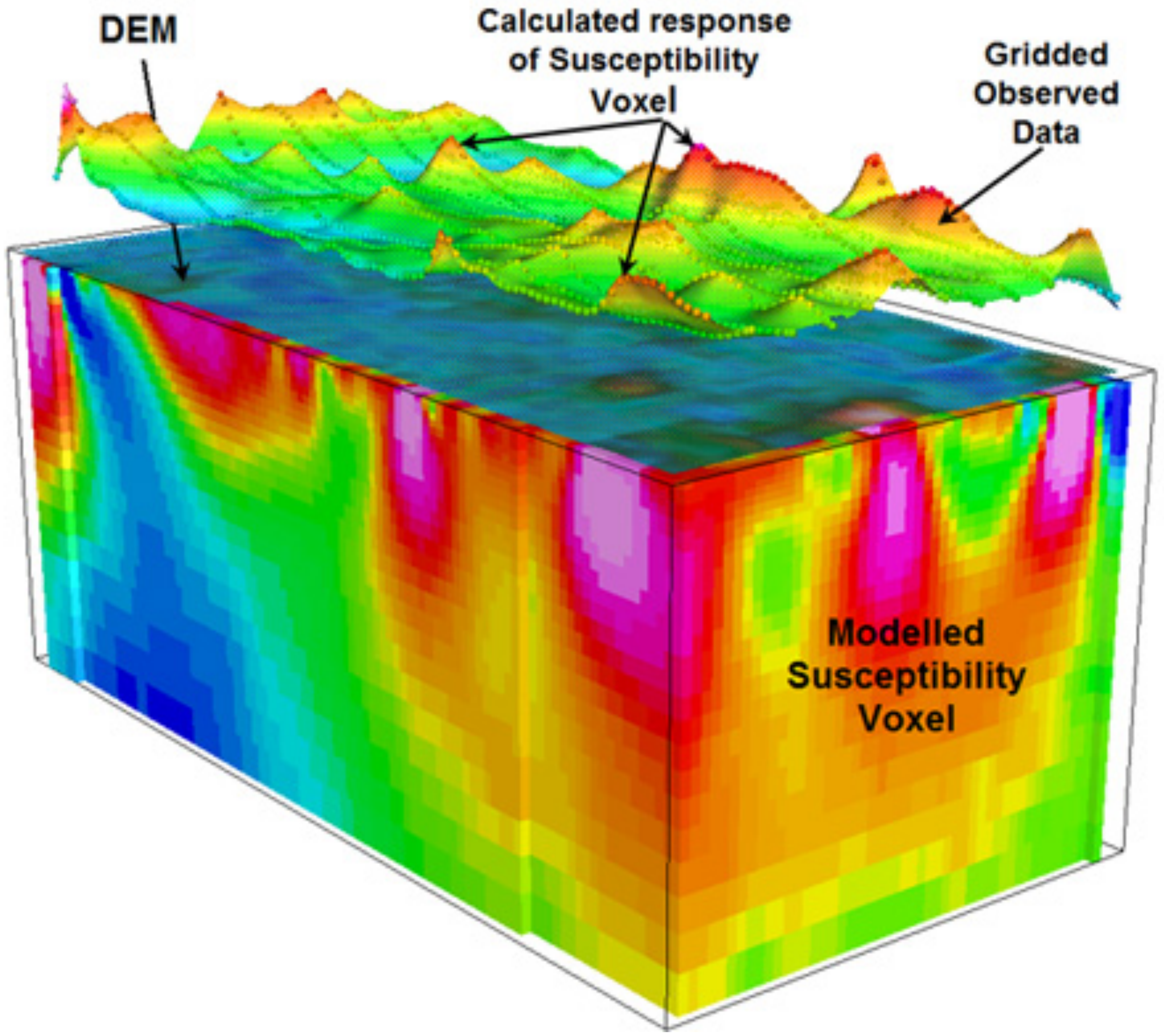
The big picture



Inverse problems


I LIVE NEXT TO A WALL OF ROCK 20 MILES THICK. THERE'S NO WAY AROUND OR OVER IT. I'M TRAPPED ON THIS SIDE FOREVER.
I STUDY THE STUFF ON THE OTHER SIDE.

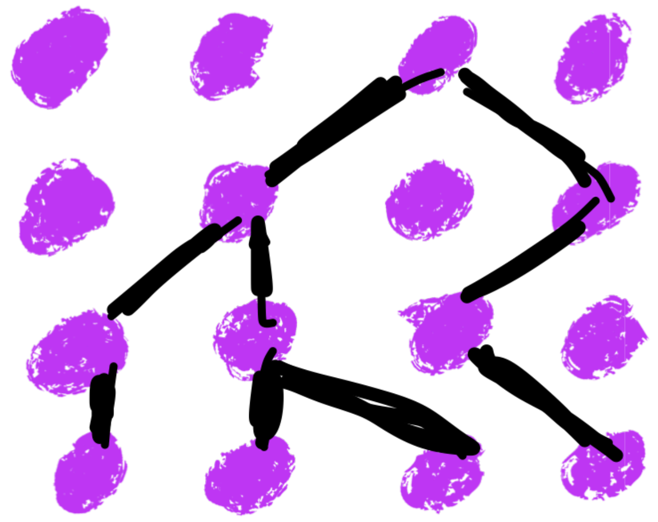
MANTLE GEOLOGY SEEMS LIKE THE MOST FRUSTRATING FIELD.


DEM
Calculated response of Susceptibility Voxel
Gridded Observed Data
Modelled Susceptibility Voxel

$$\underset{\substack{\uparrow \\ \text{data}}}{g(x)} = \int \underset{\substack{\uparrow \\ \text{model}}}{K(x,y)}\, \underset{\substack{\uparrow \\ \text{stuff we want} \\ \text{to know about}}}{f(y)}\,\mathrm{d}y$$

Inference frameworks

Wright-Fisher process at the core



Summary statistics and simple calculations (small data, simple population)

π : average pairwise divergence

S: # segregating sites

T_D : deviation from neutrality

Probabalistic models (big data, more complex populations)

Forward time:

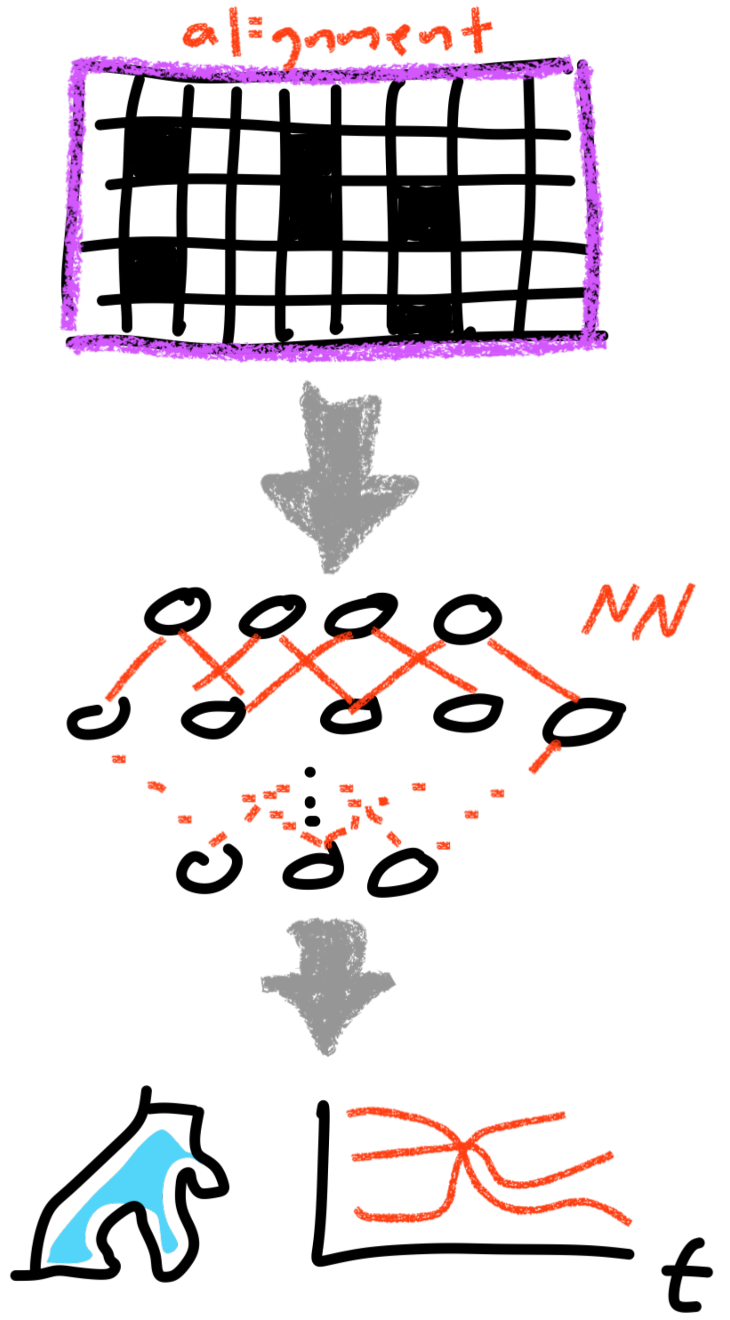
- PDEs / diffusion
- selective sweeps

Reverse time:

- coalescent geneology
- coalescent HMM

Machine learning (big data, arbitrarily complex population)

Supervised learning

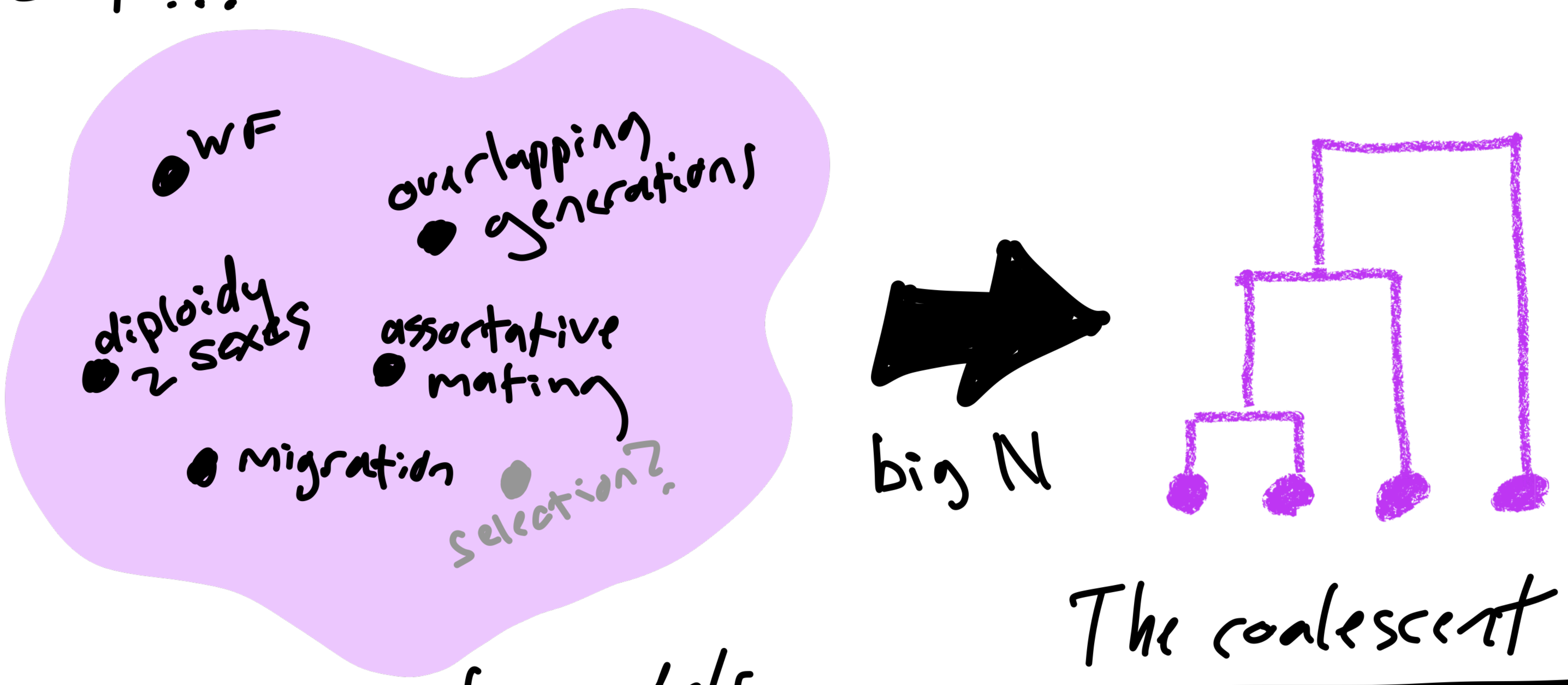


Problem

The Wright-Fisher process is wrong!

Question: what's missing w/ WF model?

It turns out...



- WF
- overlapping generations
- diploidy
- 2 sexes
- assortative mating
- migration
- selection?

Universe of models

(WF just one)

The coalescent limit

Universality

$PV = nRT$

Oh nos, where are the molecules?

Cole-escent theory

Inspired by GENOME 551 with Cole Trapnell

- 20 students in a class, numbered 1, 2, …, 20
- Each day of class, professor rolls 20-sided die to choose a student to call on

Question: How many classes do I expect it to take me to get called on?

Each class I get called w/ Prob. $\frac{1}{20}$

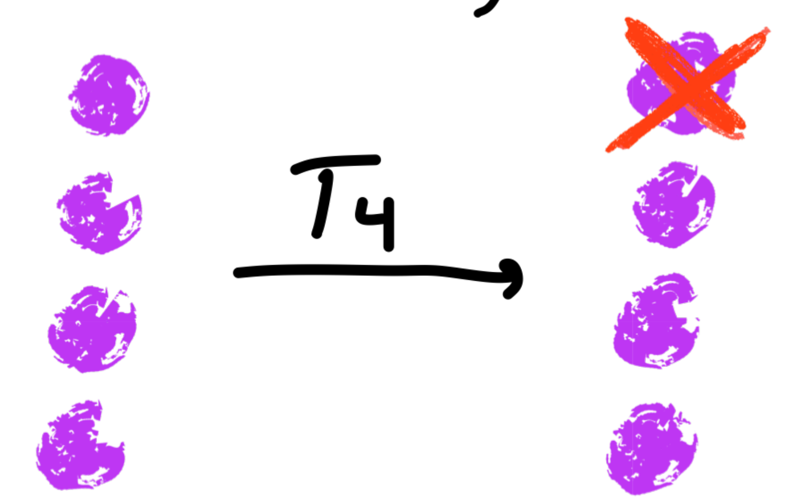
$$\Longrightarrow \mathbb{E}[T] = \frac{1}{1/20} = 20 \text{ classes}$$

Cole-escent theory

Question: How many classes do you expect until all 20 students have been called on?

When there are i students left to call on, the prob of rolling one of these remaining i is $\frac{i}{20}$

$$\mathbb{E}[T_i] = \frac{1}{i/20} = \frac{20}{i}$$



$i = 20, 19, \ldots, 1$

So

$$\mathbb{E}[T_{20\to 0}] = \mathbb{E}[T_{20}] + \mathbb{E}[T_{19}] + \ldots \mathbb{E}[T_1]$$

$$= 1 + \frac{20}{19} + \ldots + \frac{20}{1} = 1 + 20\sum_{i=1}^{19}\frac{1}{i} \approx 72$$

Cole-escent theory

Generalize: Class size N (and N-sided die)
sample of $n \leq N$ students

$$\mathbb{E}[T_{n\to 0}] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{1}{i/N} = N \sum_{i=1}^{n} \frac{1}{i}$$

$$p = \frac{i}{N}$$

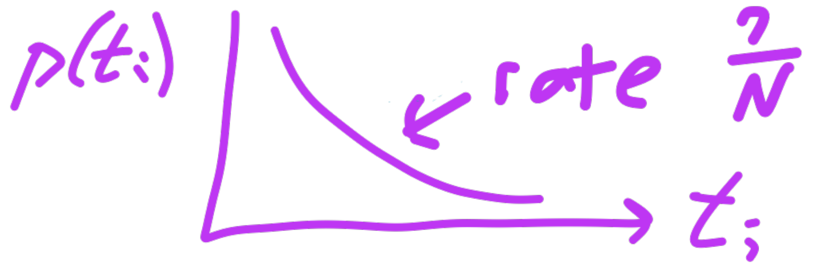
Each interval is geometrically distributed

$$\mathbb{P}(T_i = t_i) = \frac{i}{N}\left(1 - \frac{i}{N}\right)^{t_i - 1}$$

For large N:

$$p(t_i) \simeq \frac{i}{N} e^{-\frac{i}{N} \underset{\substack{\uparrow \\ \text{continuous}}}{t_i}}$$

→ approx. exponential dist.



Memoryless:

$$p(t_i | t_{i+1}) = p(t_i)$$

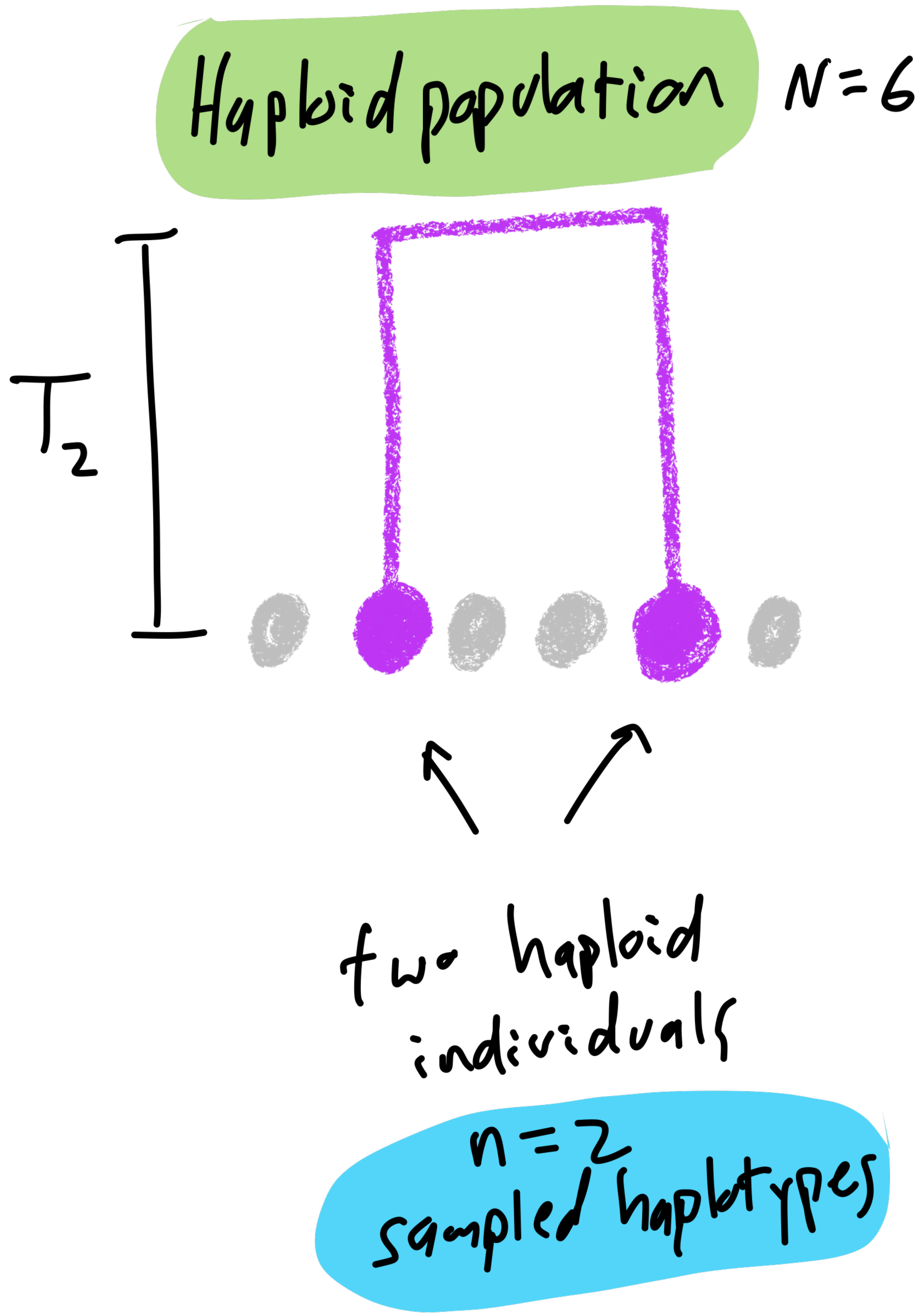
Coalescent theory

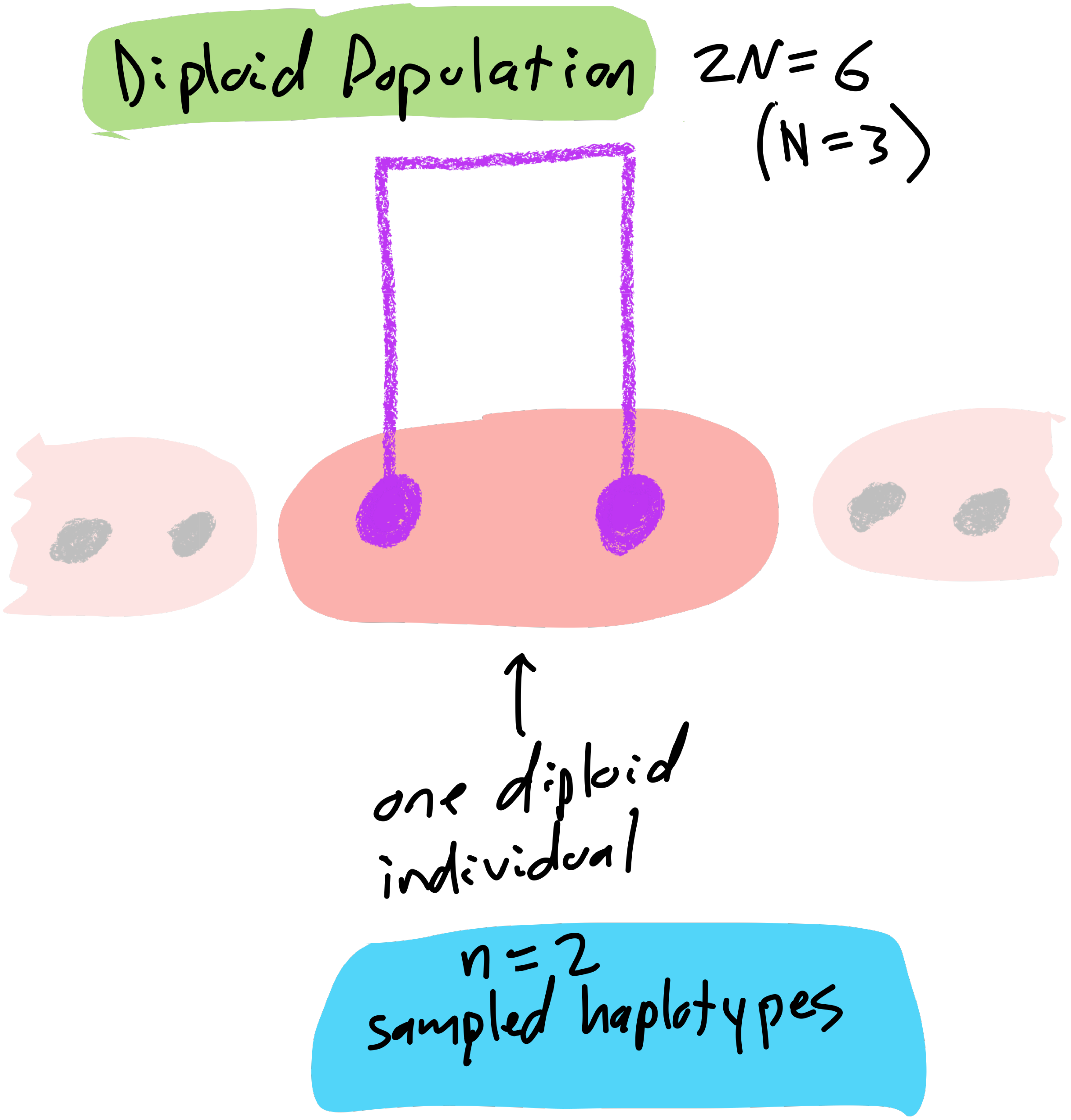
Only slightly fancier

	Cole-escent	coalescent
time	classes into the future	generations into the past
events	dice rolls, one student	coalescences, pairs of individuals
rate of events w/ i individuals	$\frac{i}{N}$ i ← students uncalled N ← class size	$\frac{\binom{i}{2}}{2N}$ $\binom{i}{2}$ ← $=\frac{1}{2}i(i-1)$, # pairs N ← population size ↑ if diploid

Coalescent theory

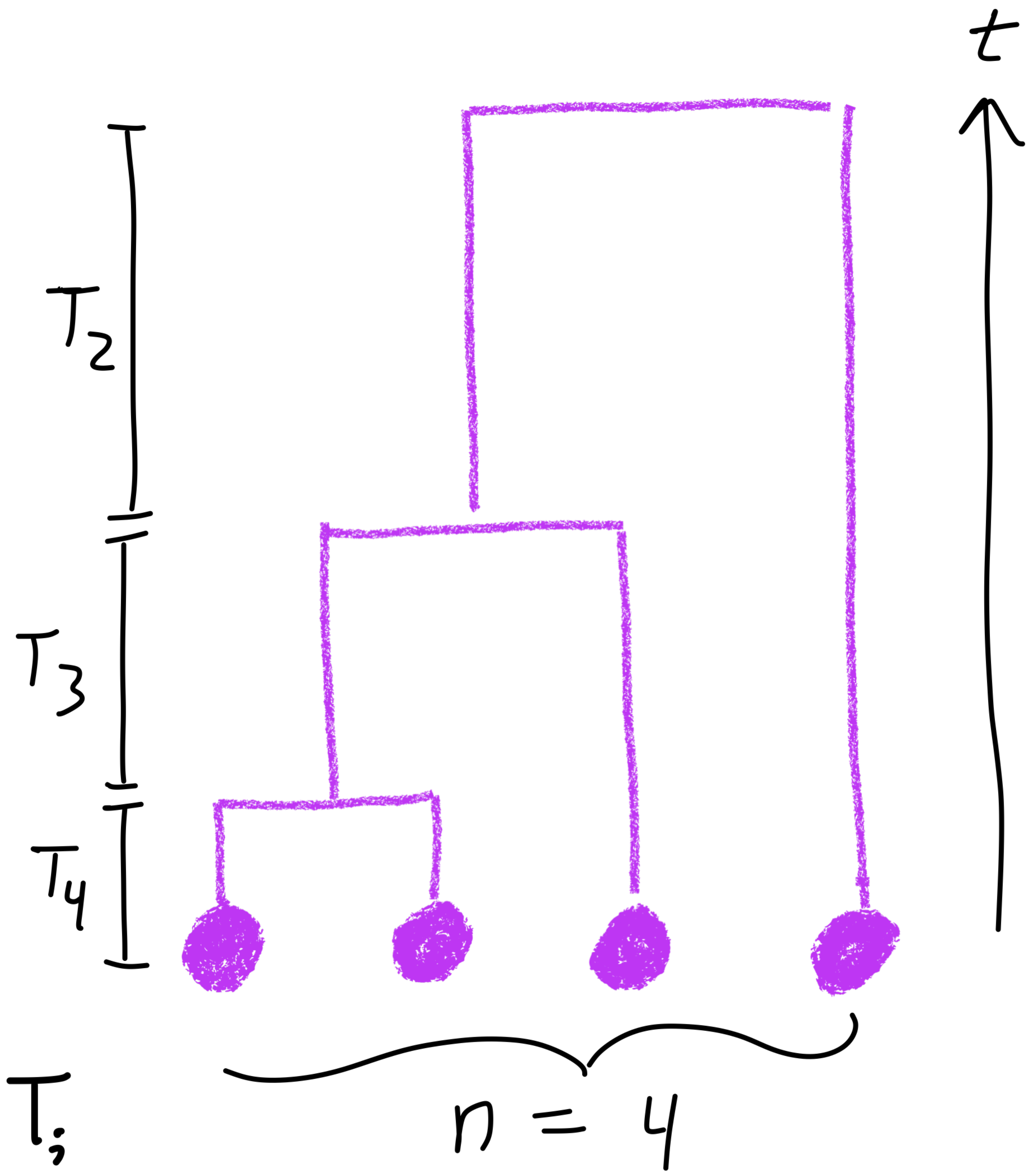
Watch those factors of 2!





Coalescent theory

Updating the previous results



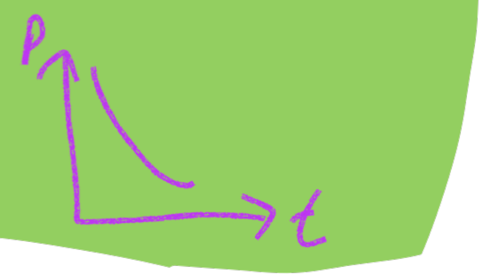
$T_{MRCA} = \sum_{i=2}^{n} T_i$

"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

Coalescent theory

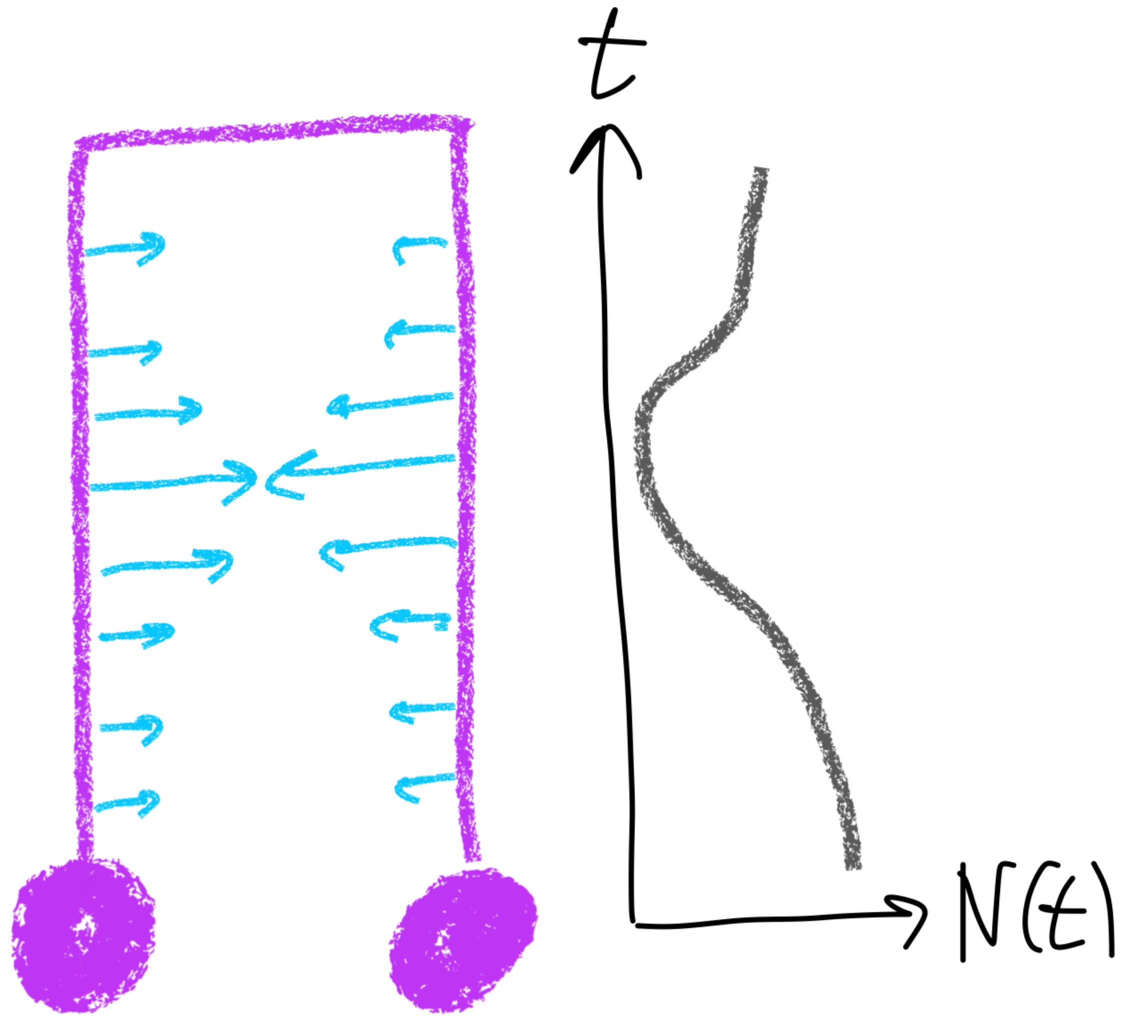
Population size determines coalescence rate

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

[Like students adding/dropping mid-quarter in Cole-escent theory]

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:

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

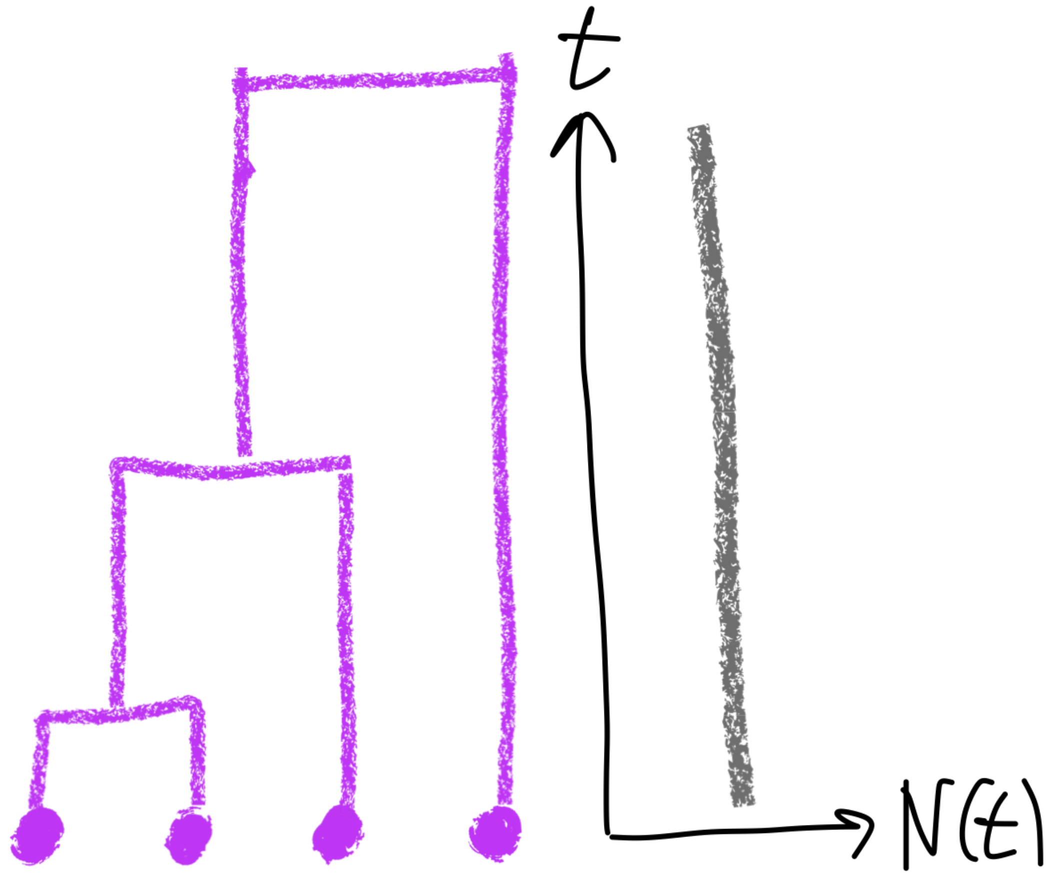
$$\xrightarrow{\text{big } N} p(t_i) = \frac{\binom{i}{2}}{2N(t)} e^{-\binom{i}{2}\int_0^t \frac{ds}{2N(s)}}$$

inhomogeneous Poisson process

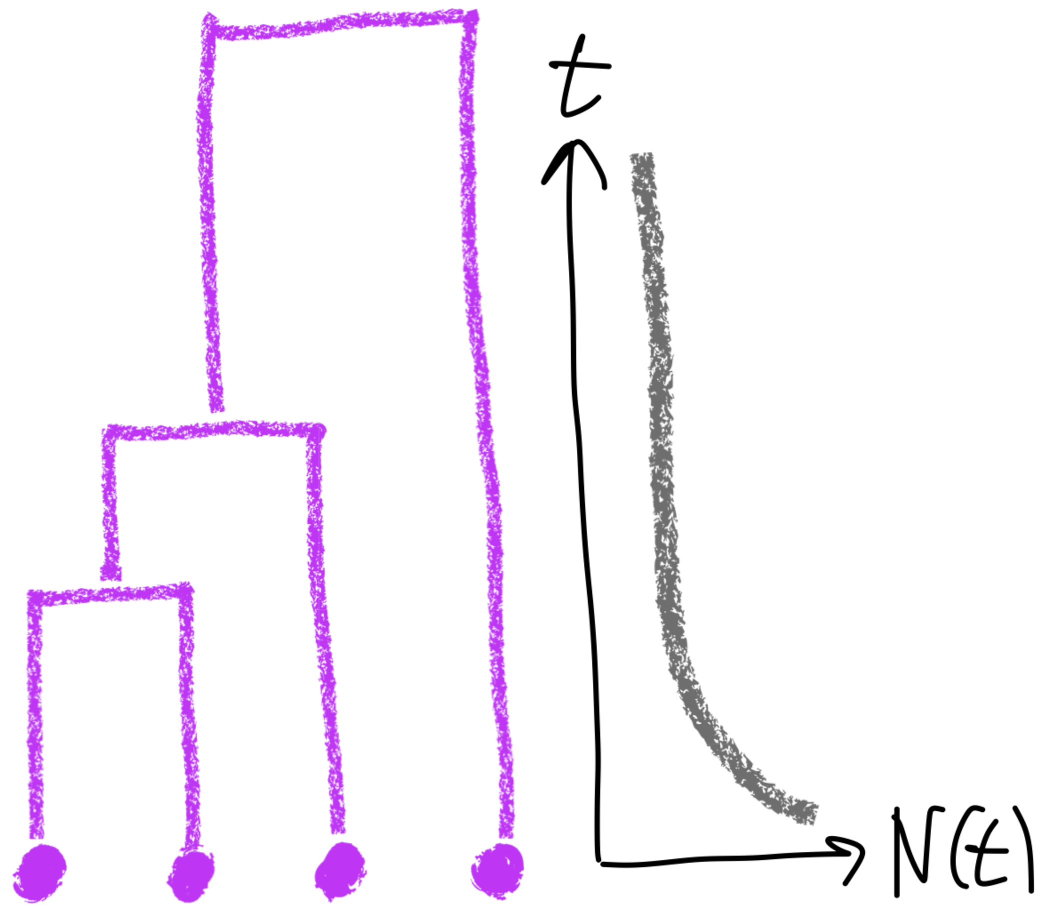
Coalescent theory

Population size determines coalescence rate

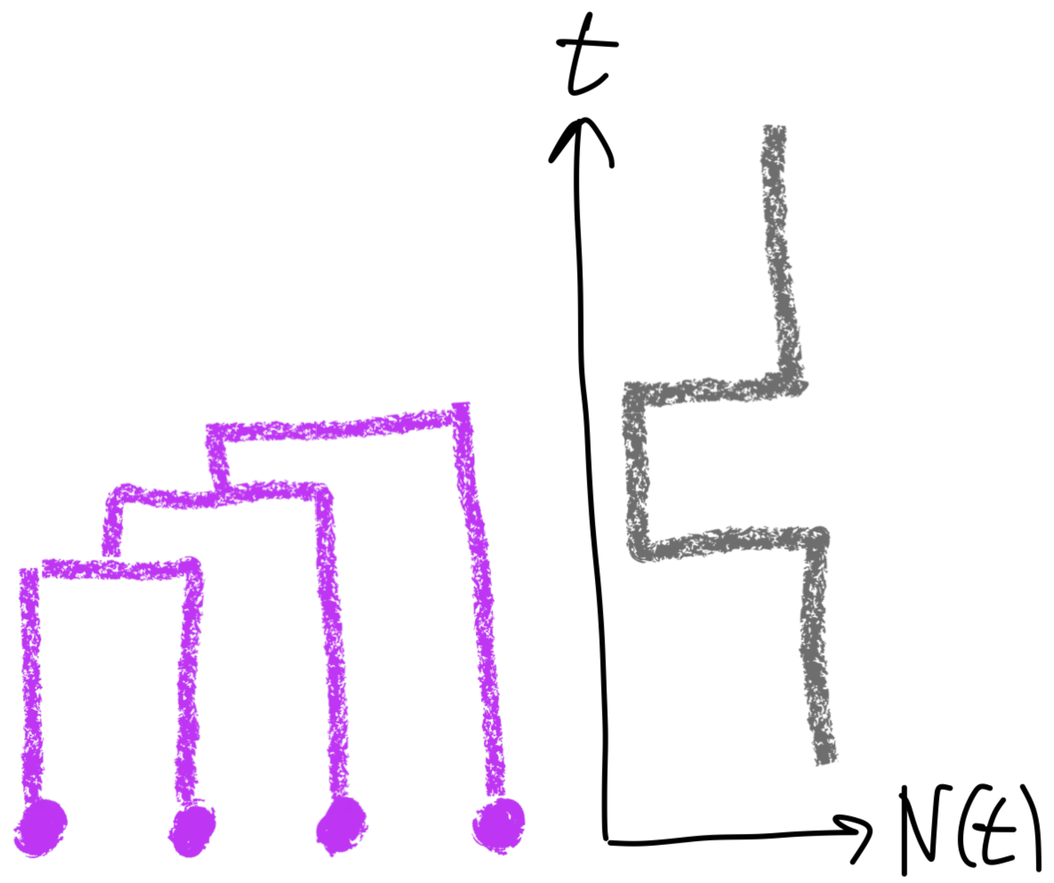
Constant



Exponential growth

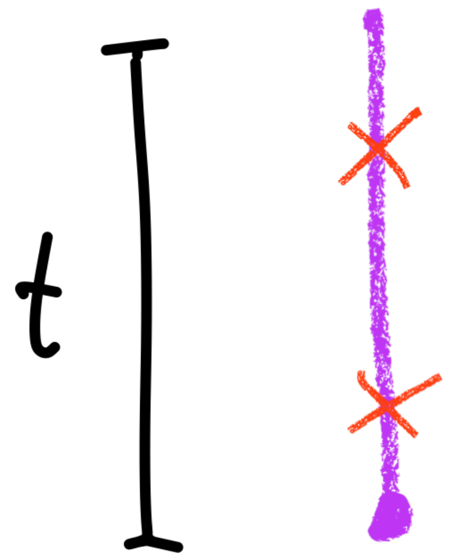


Bottleneck



Coalescent theory

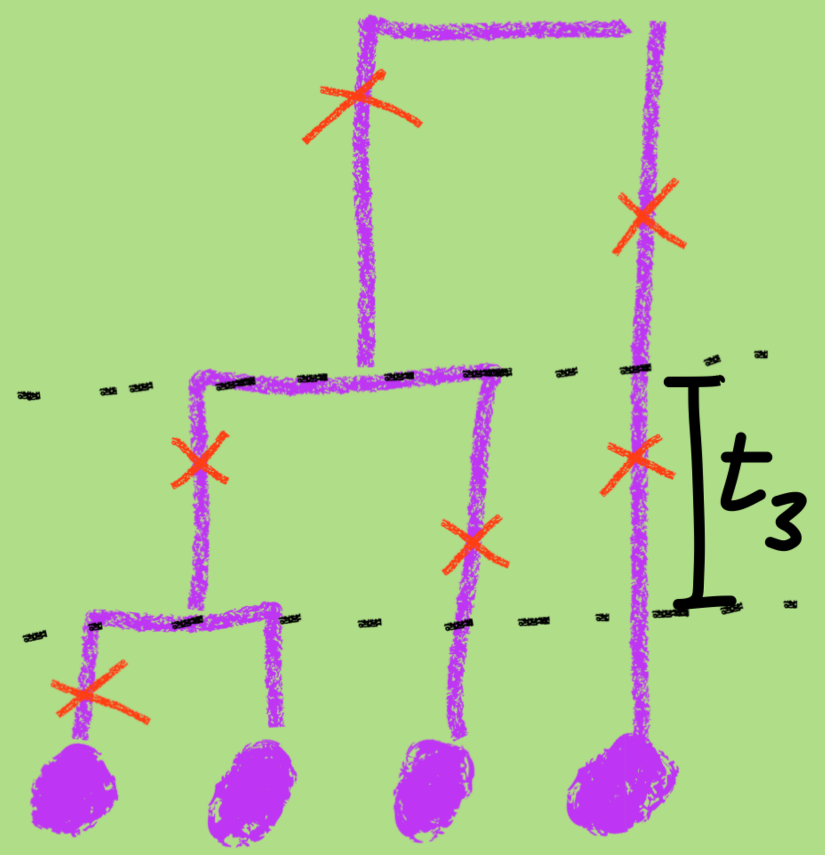
Mutations



mutations on branch of length t is Poisson rv w/ mean μt

$$\mathbb{P}(k \mid t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

mutations per genome per generation



mutations in intercoalescent interval i, of length t_i, is Poisson rv w/mean

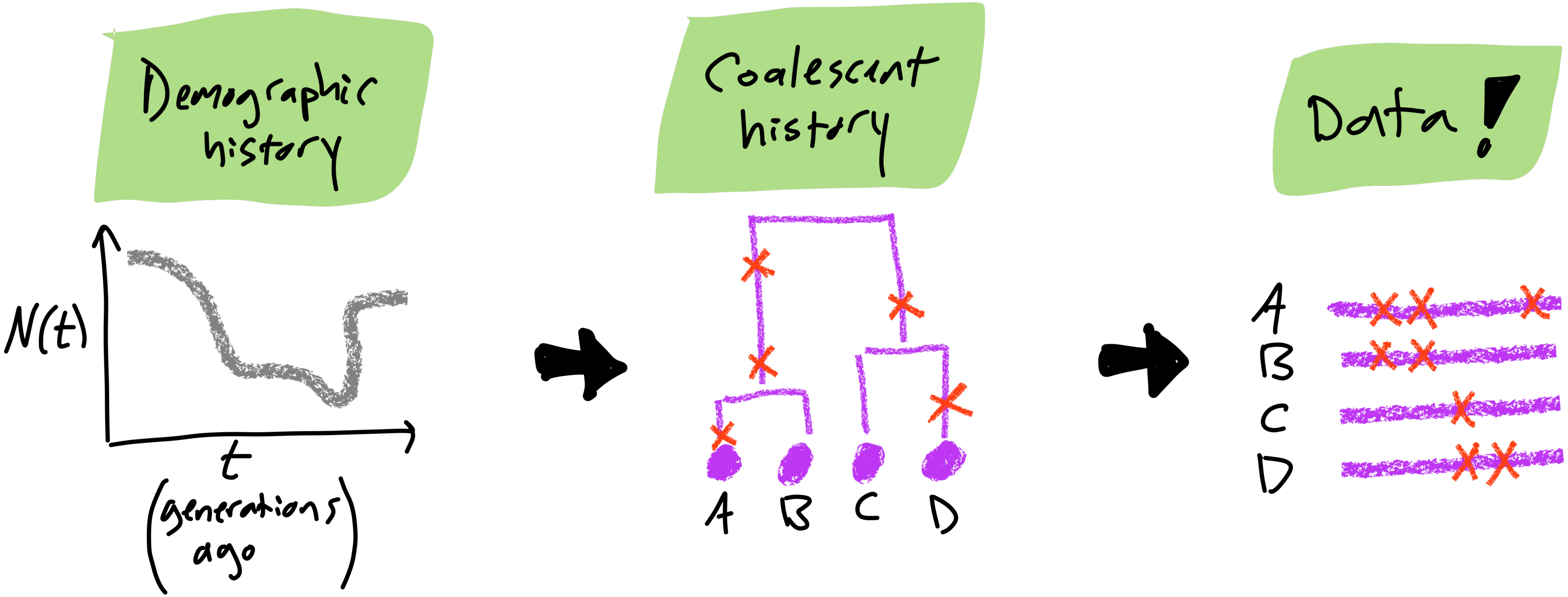
$$i \mu t_i$$

lines in interval i

duration of interval i

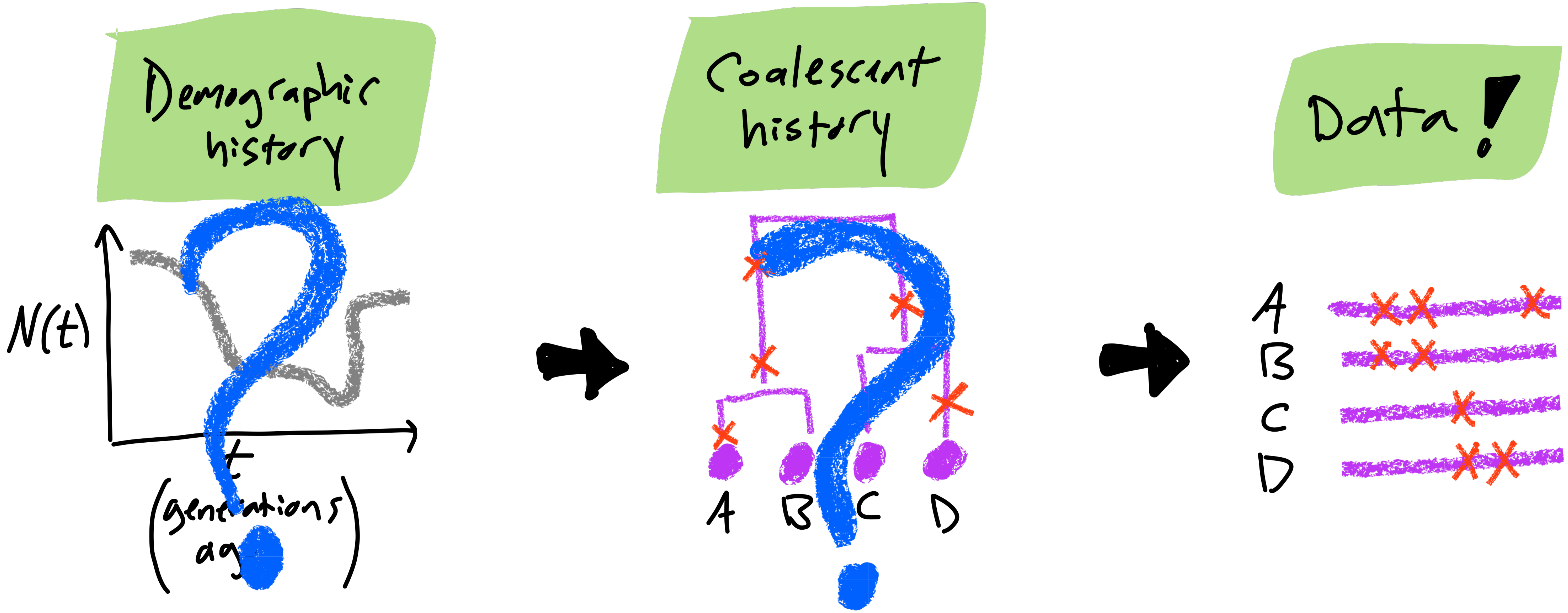
Coalescent theory

Genetic diversity



Coalescent theory

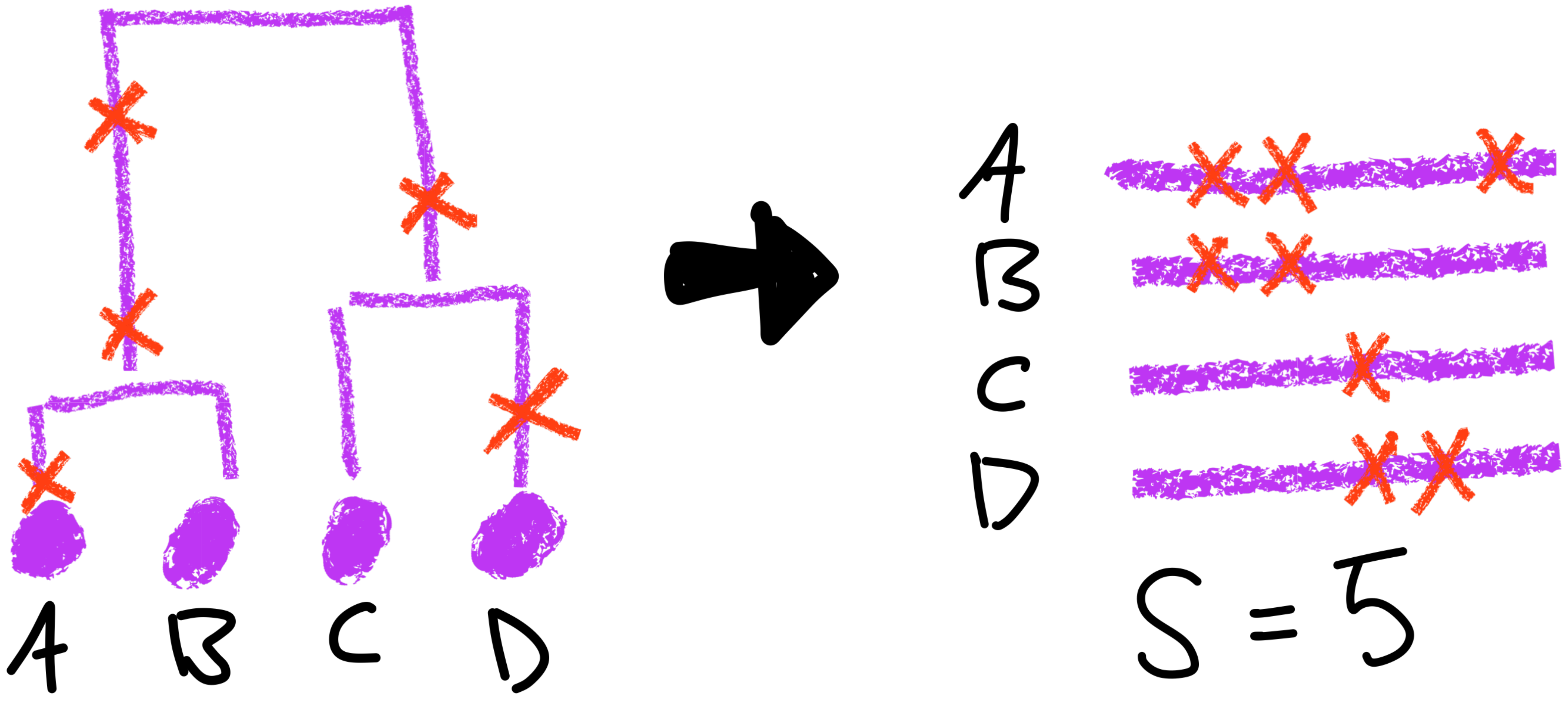
Genetic diversity



Coalescent theory

Genetic diversity

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



$S = 5$

Constant N case:

total branch length

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

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

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

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

Recap

Good to know for homework

Intercoalescent times are indep. exponential rvs

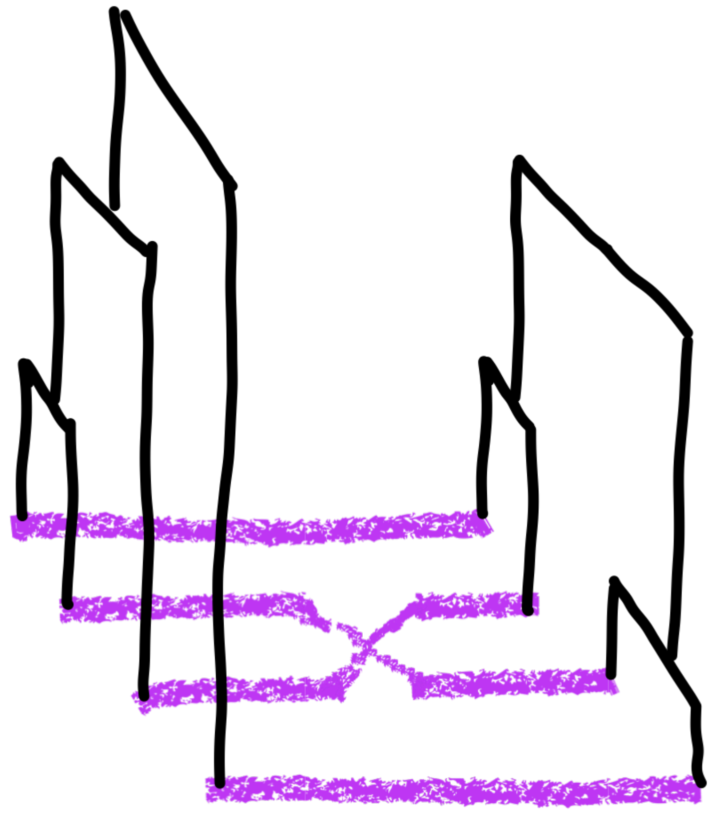
$$\overset{\text{intercoalescent interval } i}{T_i} \sim \exp\left(\frac{2\overset{\text{diploid pop size}}{N}}{\binom{i}{2}}\right), \quad \text{for} \quad i = \overset{\text{\# sampled haplotypes}}{n}, n-1, \ldots, 2$$

mutations on branches are indep. Poisson rvs

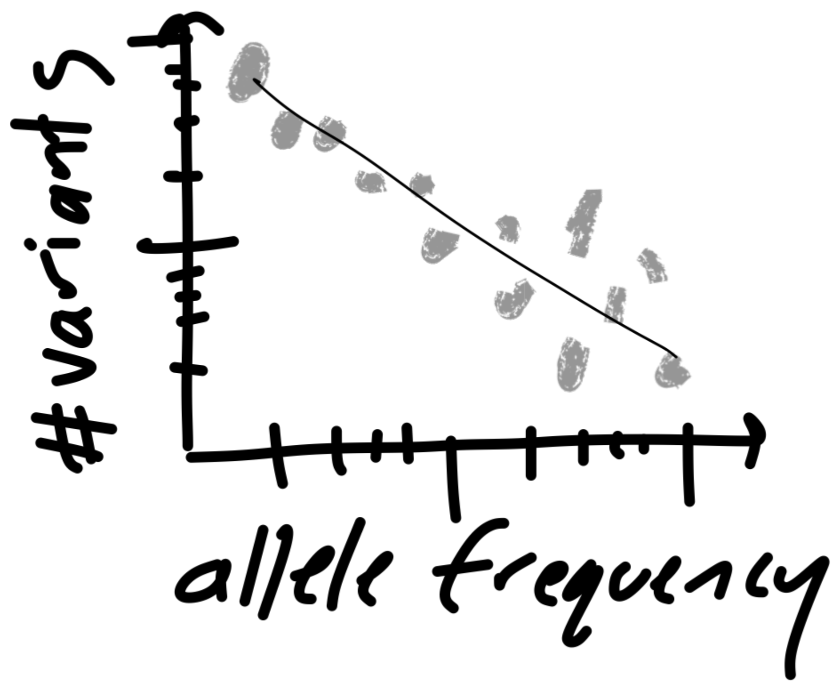
$$\underset{\text{\# mutations}}{k} \sim \text{Pois}\left(\underset{\text{mutation rate}}{\mu}\,\underset{\text{branch length}}{t}\right)$$

Next time

- The coalescent with recombination



- Sample Frequency spectrum (SFS)



- Coalescent hidden Markov model