Demography estimation

Shyam Gopalakrishnan
Aug 8th 2025

Outline

- Background – What + Why demography?
- Various ways to estimate demography
- Recap coalescent + HMM
- Estimate from one genome
  - PSMC
- Exercise

Demography

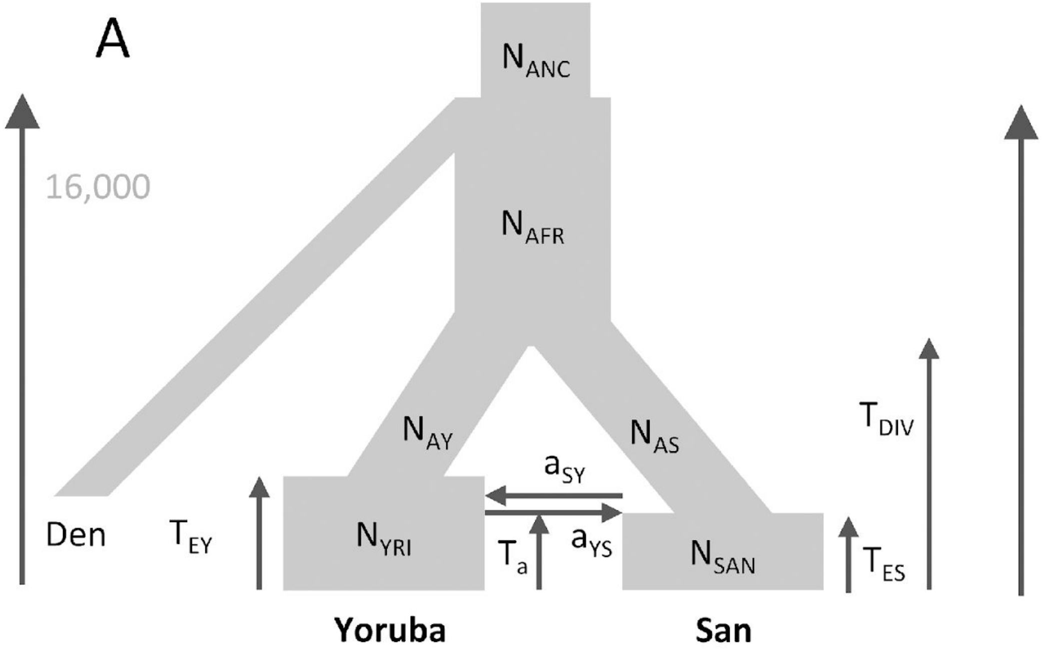
- What do we mean by demography?

Demography

- What do we mean by demography?
  - Population sizes
  - Migration rates
  - Population split times

Demography

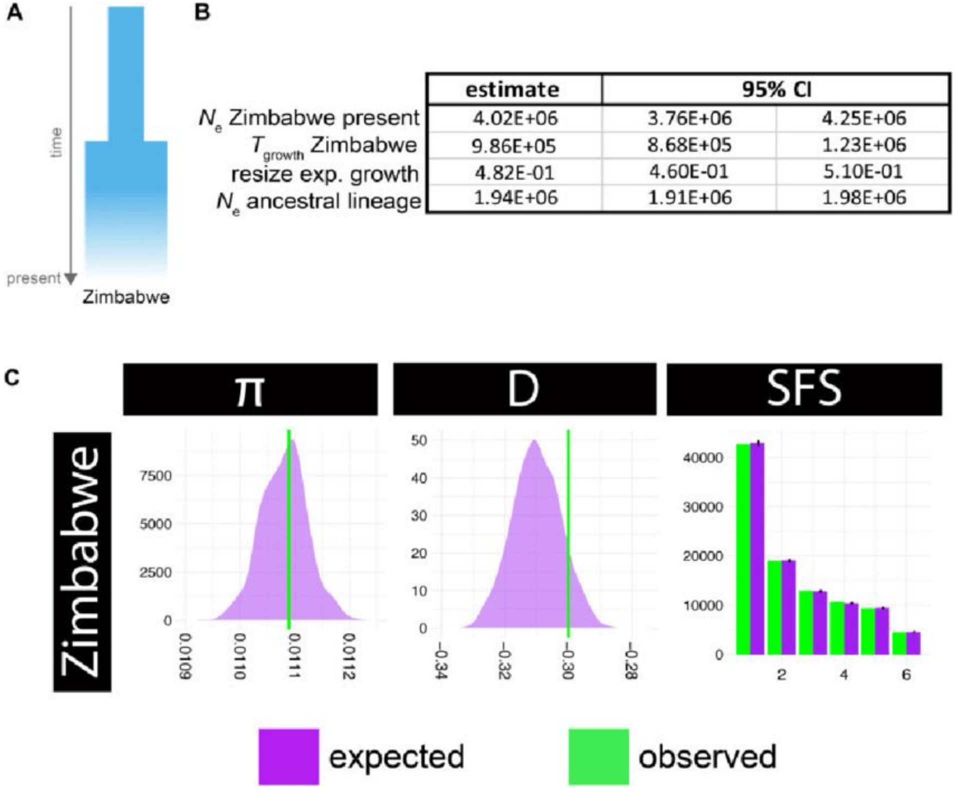
- What do we mean by demography?



Excoffier et al. 2013 PLoS Genet

Demography

- Demography for a single population?
  - Effective population size

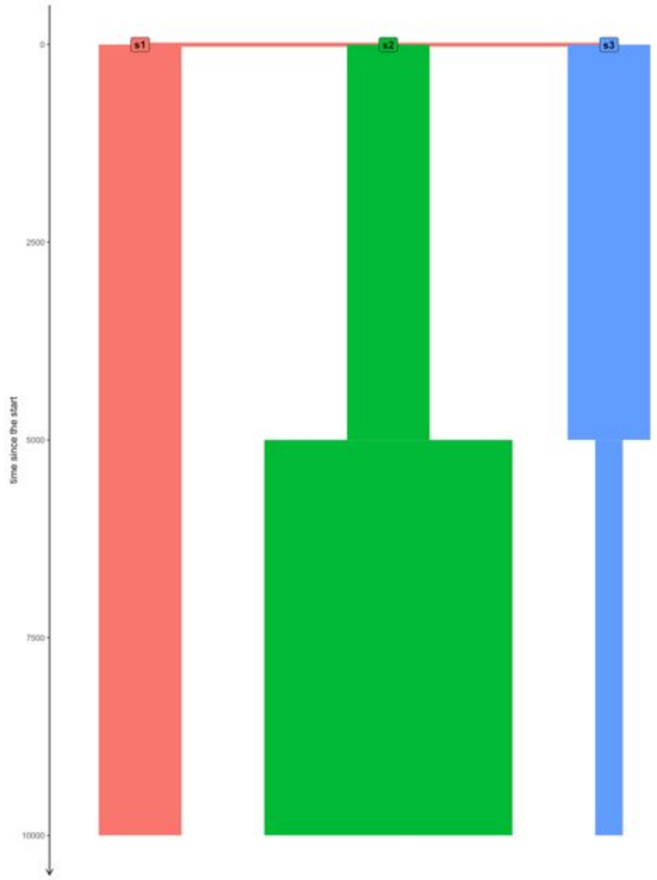


	estimate	95% CI	
N_e Zimbabwe present	4.02E+06	3.76E+06	4.25E+06
T_{growth} Zimbabwe	9.86E+05	8.68E+05	1.23E+06
resize exp. growth	4.82E-01	4.60E-01	5.10E-01
N_e ancestral lineage	1.94E+06	1.91E+06	1.98E+06

Arguello et al. 2019 GBE

Why care about demography?

- Demography allows us to characterize the neutral variation in the genome



How to estimate demography?

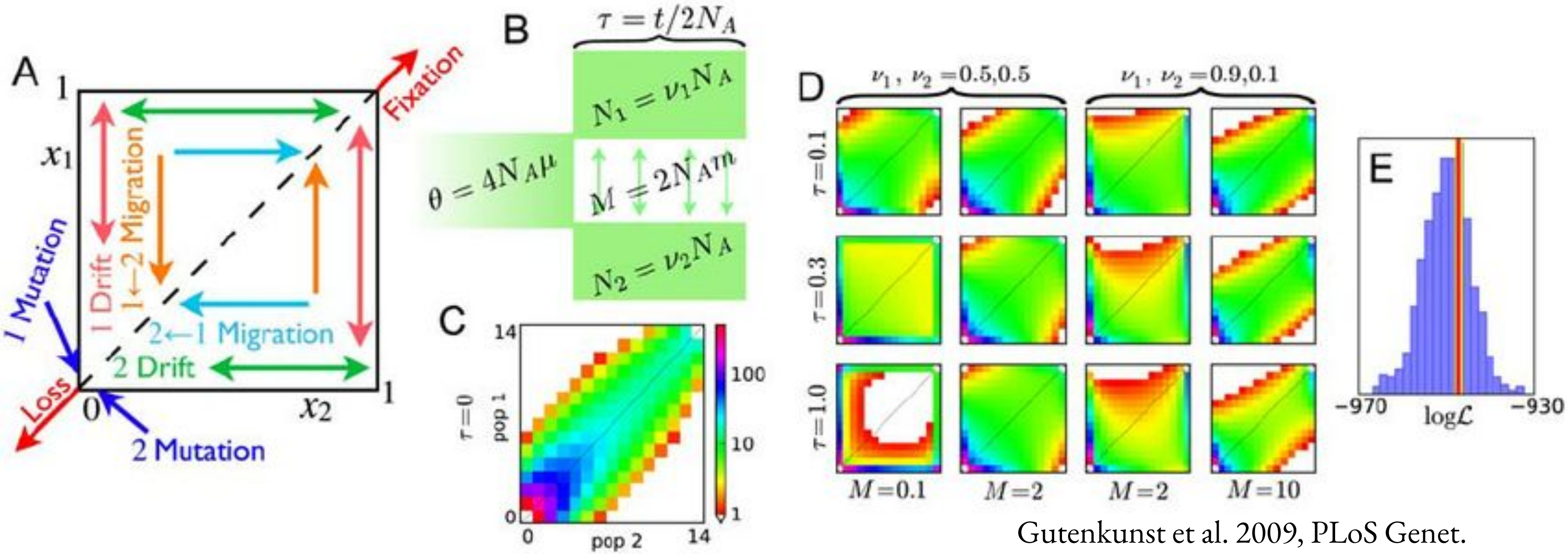
- Can you think of a way to estimate demography based on your lectures this week?

Estimating demography

- Usually using summary statistics
  - SFS
  - Linkage disequilibrium
  - A whole plethora of other summary statistics
- Define a model and find parameters that best fit the observed summary statistics
  - Various statistical techniques

Estimating demography

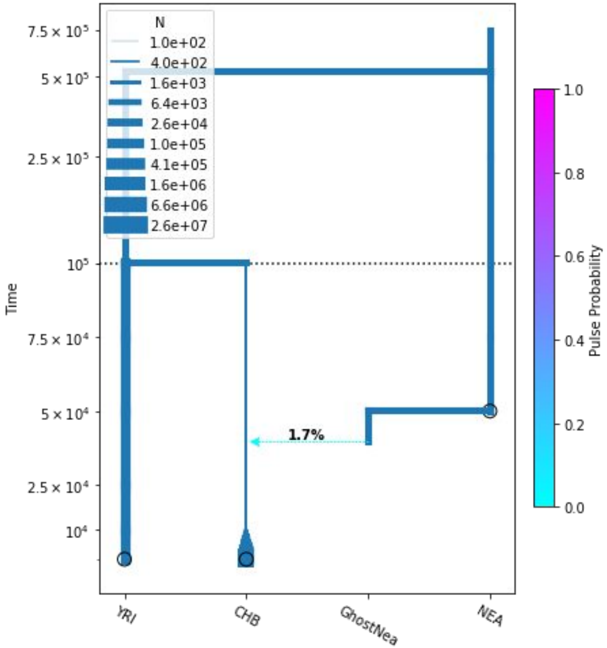
- SFS based demography estimation
  - dadi: diffusion approximation for demographic inference

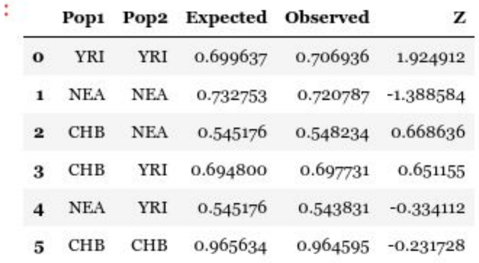


Gutenkunst et al. 2009, PLoS Genet.

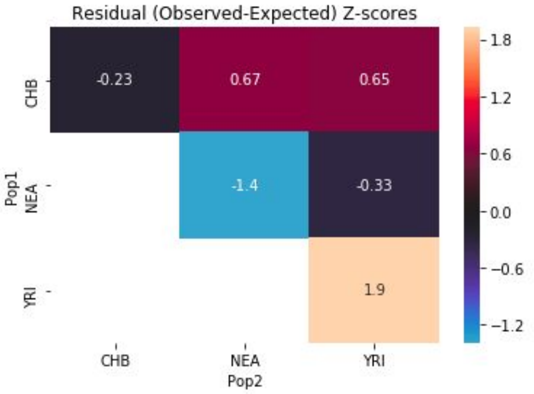
Estimating demography

- SFS based demography estimation
  - momi2: Moran models for inference





	Pop1	Pop2	Expected	Observed	Z
0	YRI	YRI	0.699637	0.706936	1.924912
1	NEA	NEA	0.732753	0.720787	-1.388584
2	CHB	NEA	0.545176	0.548234	0.668636
3	CHB	YRI	0.694800	0.697731	0.651155
4	NEA	YRI	0.545176	0.543831	-0.334112
5	CHB	CHB	0.965634	0.964595	-0.231728

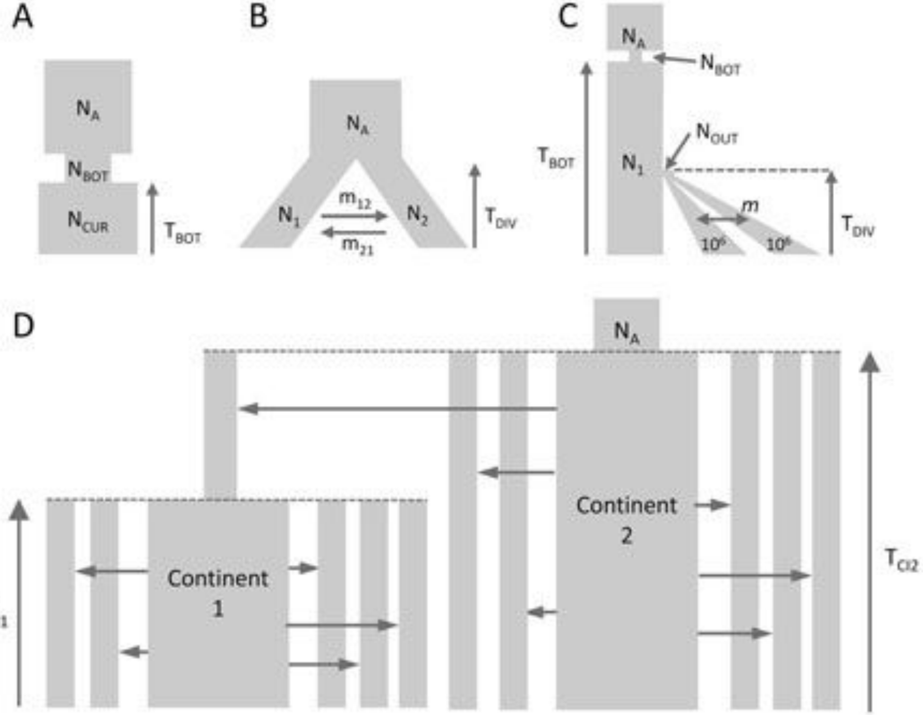


https://momi2.readthedocs.io/en/latest/tutorial.html

Kamm et al. 2020, JASA

Estimating demography

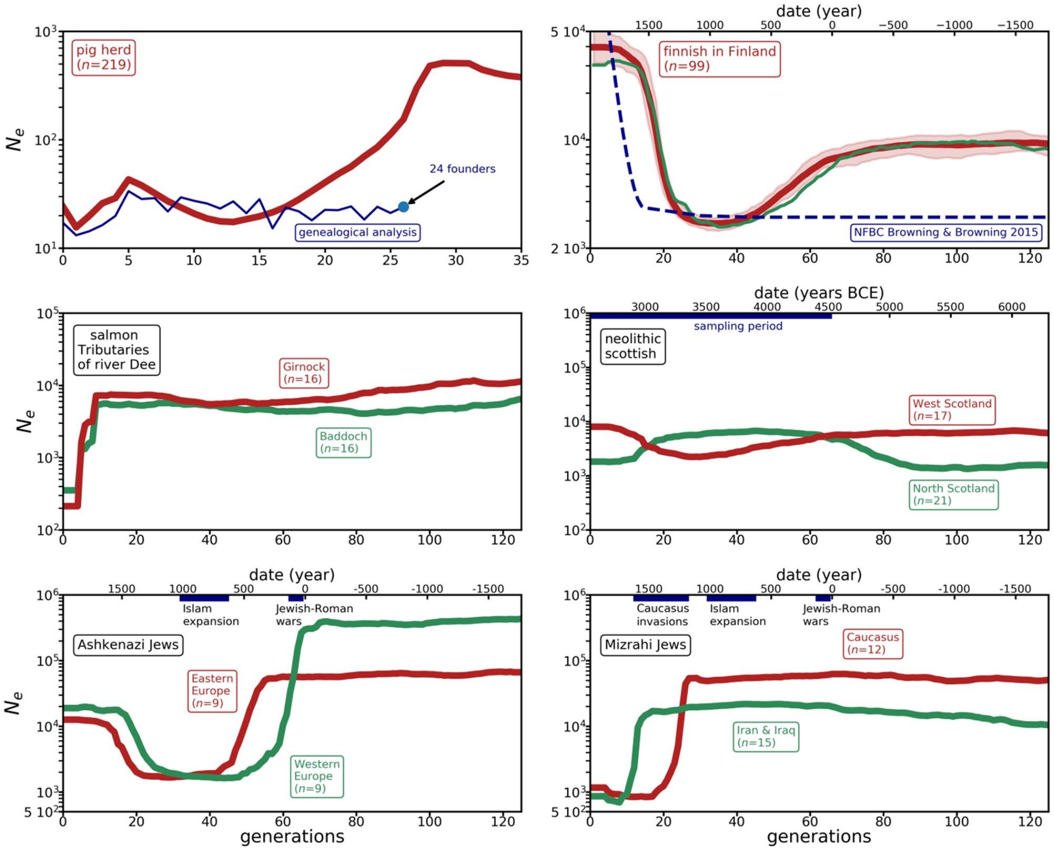
- SFS based demography estimation
  - fastsimcoal2



Excoffier et al. 2021, Bioninformatics

Estimating demography

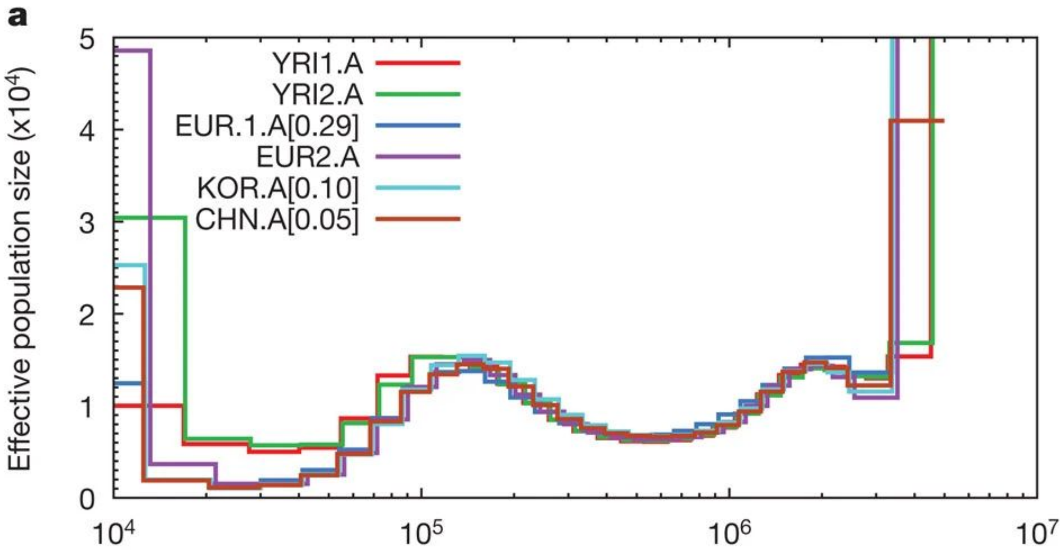
- Other metrics to estimate demography
  - Linkage disequilibrium (LD)

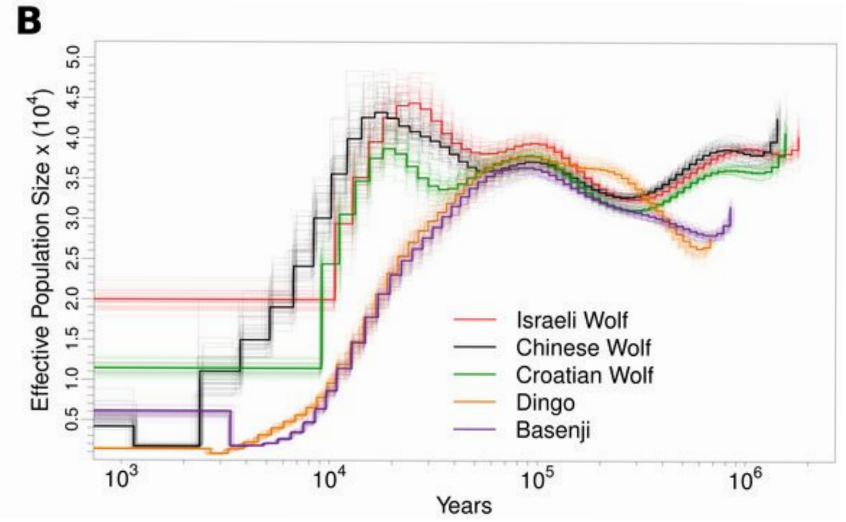


Santiago et al. 2020, MBE

Estimating demography

- Coalescent based (*)
  - PSMC

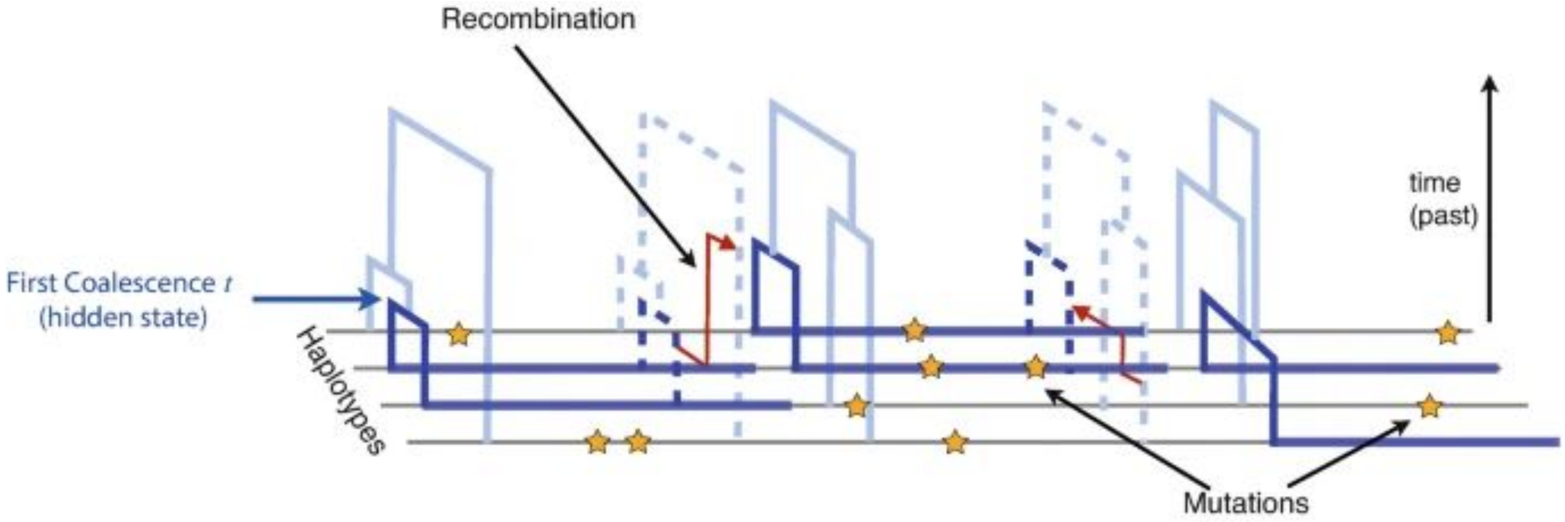




Li and Durbin 2011, Nature
Freedman et al. 2014, PLoS Genet.

Estimating demography

- Coalescent based (*)
  - MSMC



Schiffels and Durbin, 2014, Nature Genetics

Quick coalescent detour

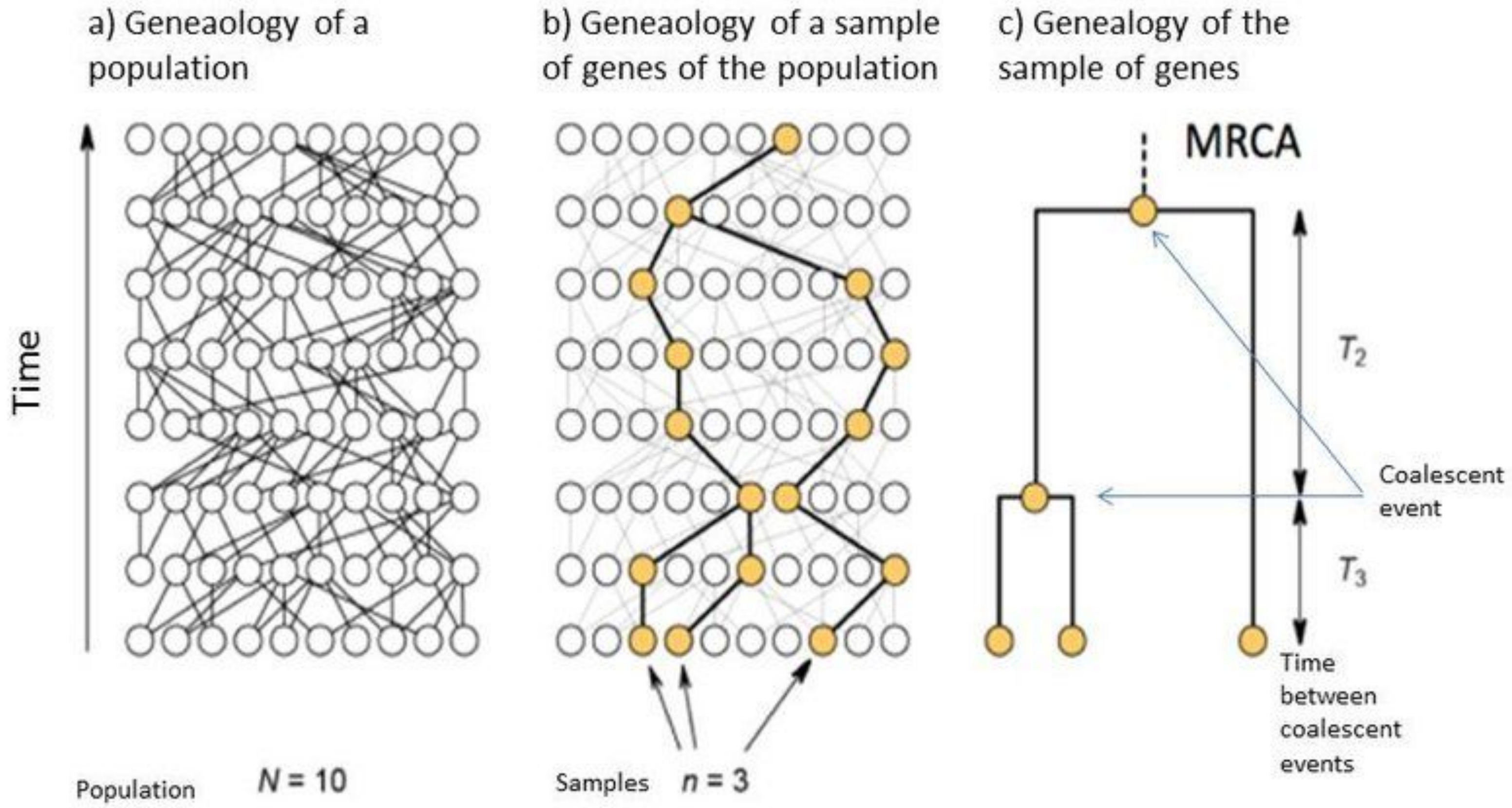

a) Geneaology of a population
b) Geneaology of a sample of genes of the population
c) Genealogy of the sample of genes
Time
MRCA
T_2
T_3
Coalescent event
Time between coalescent events
Population
$N = 10$
Samples
$n = 3$

Figure courtesy of Marie Louis

Coalescent to demography

Group discussion

Consider a pair of chromosomes

- Mutations
- Recombinations

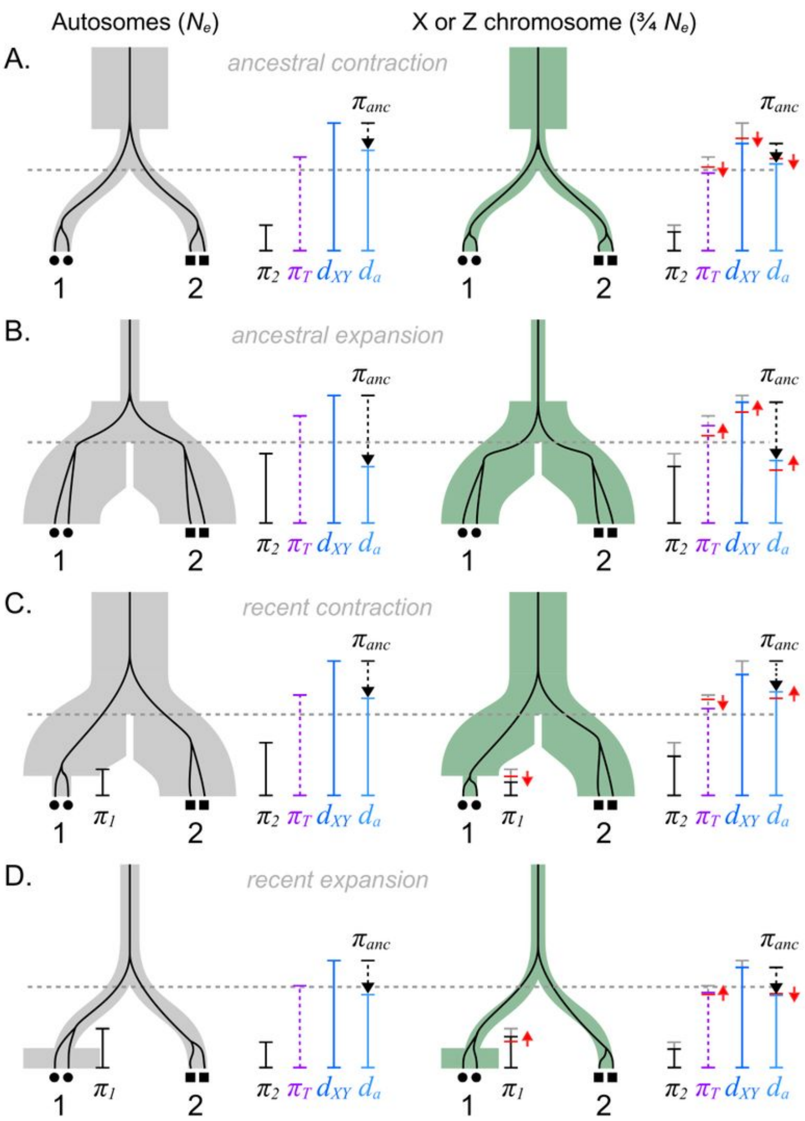
How would you use this information to estimate the demography, specifically Ne of one population?

Coalescent to demography

- Effect of changing effective population size, N_e

Coalescent to demography

- Effect of changing effective population size, N_e

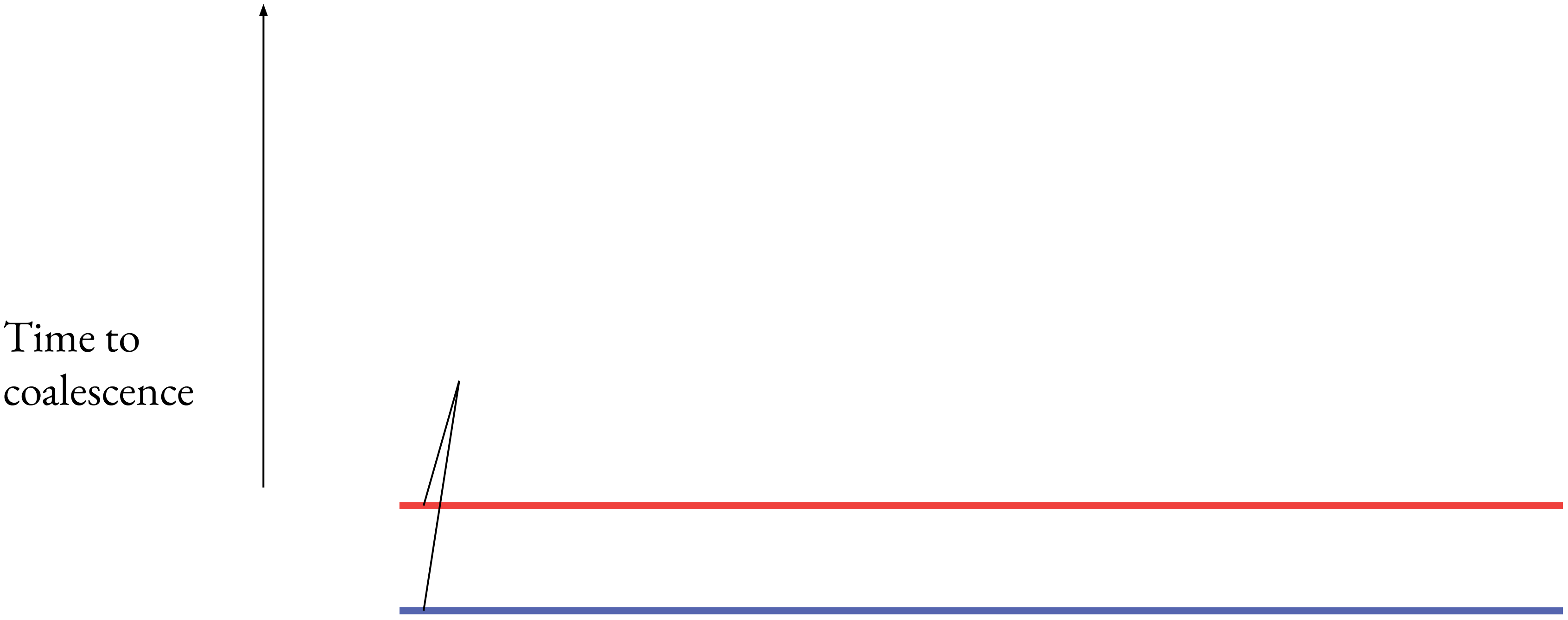


Information in one genome?

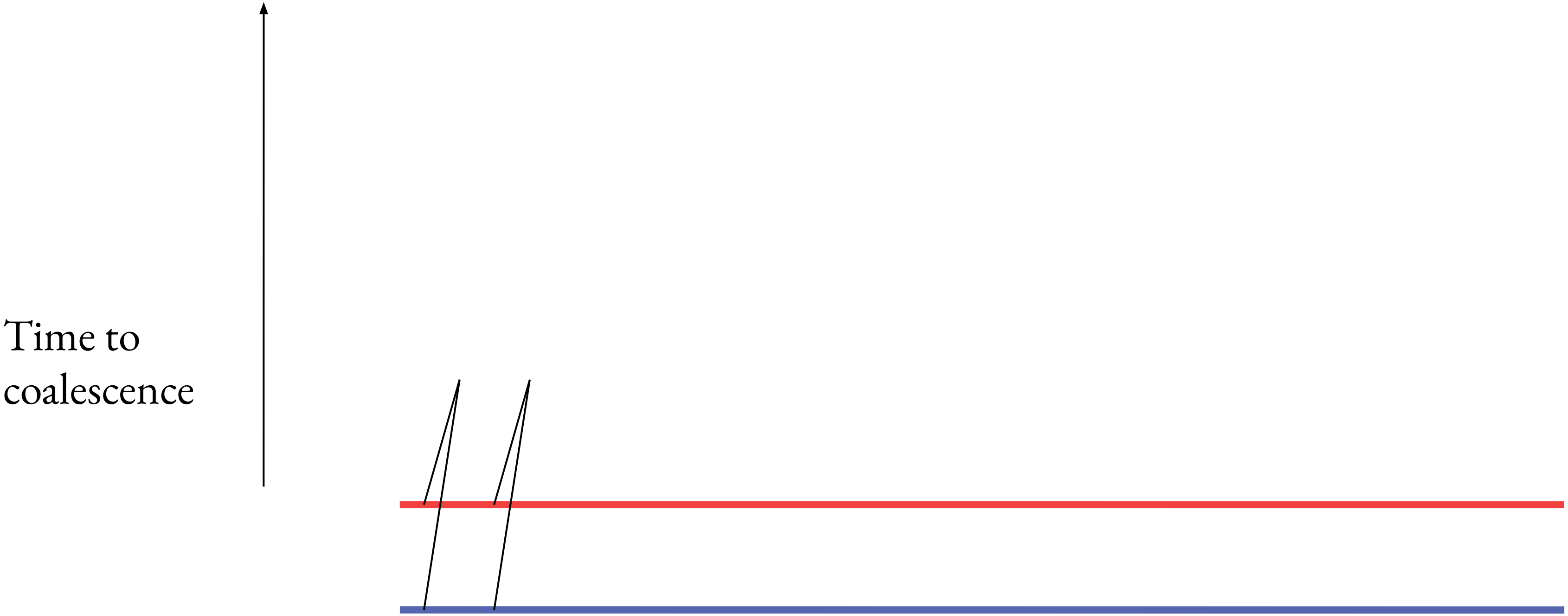
Information in one genome?

Time to
coalescence

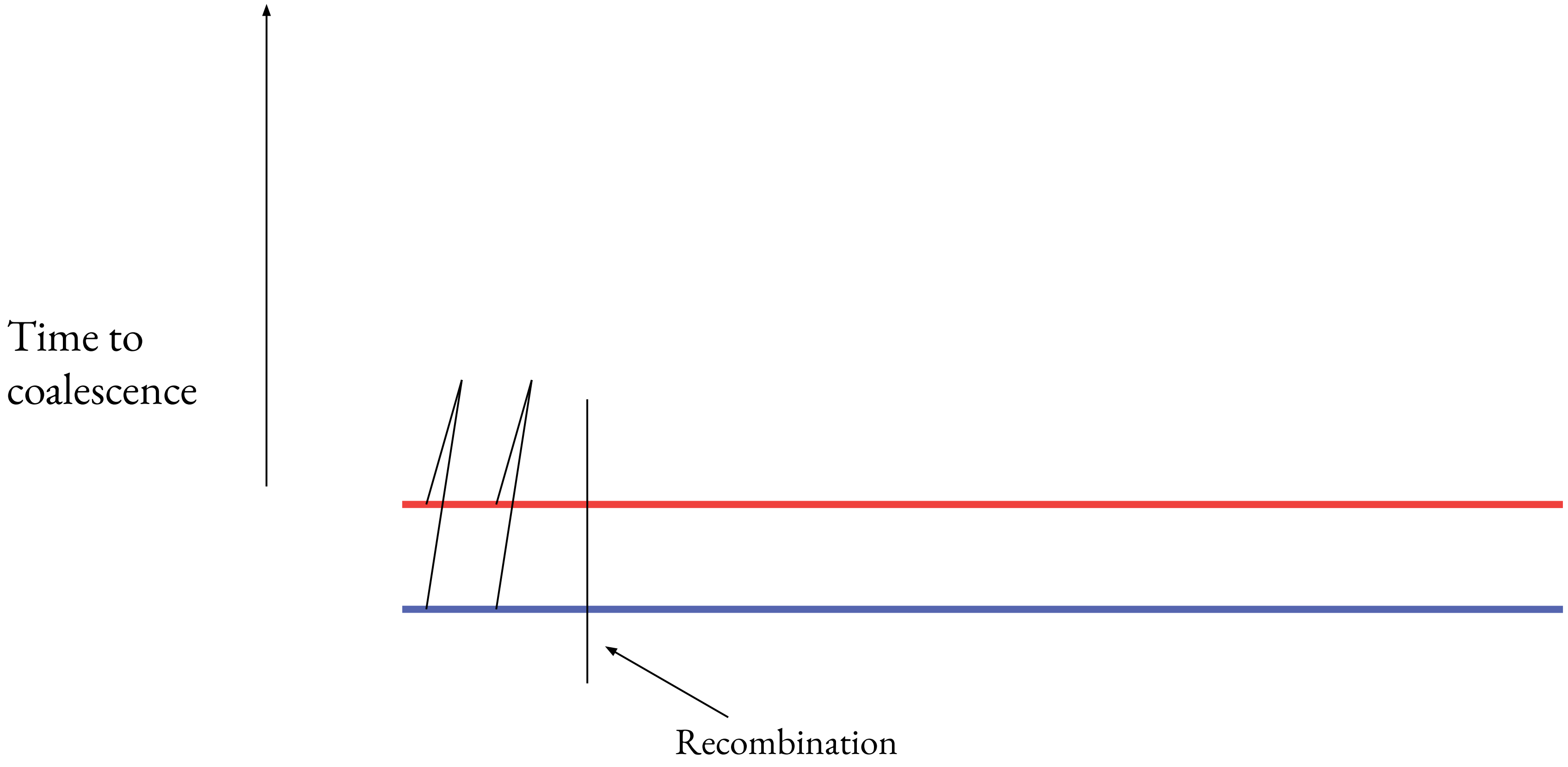
Information in one genome?



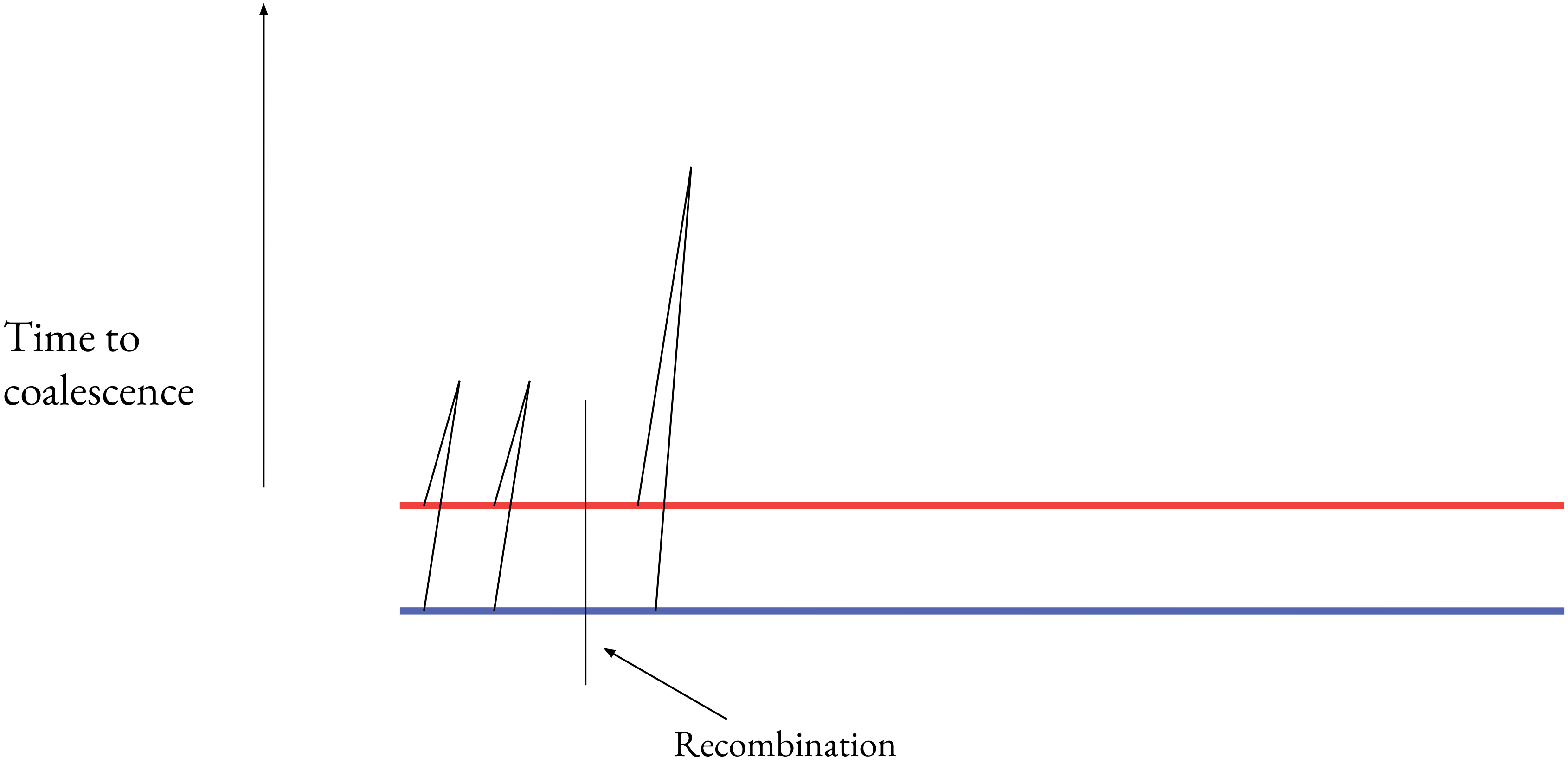
Information in one genome?



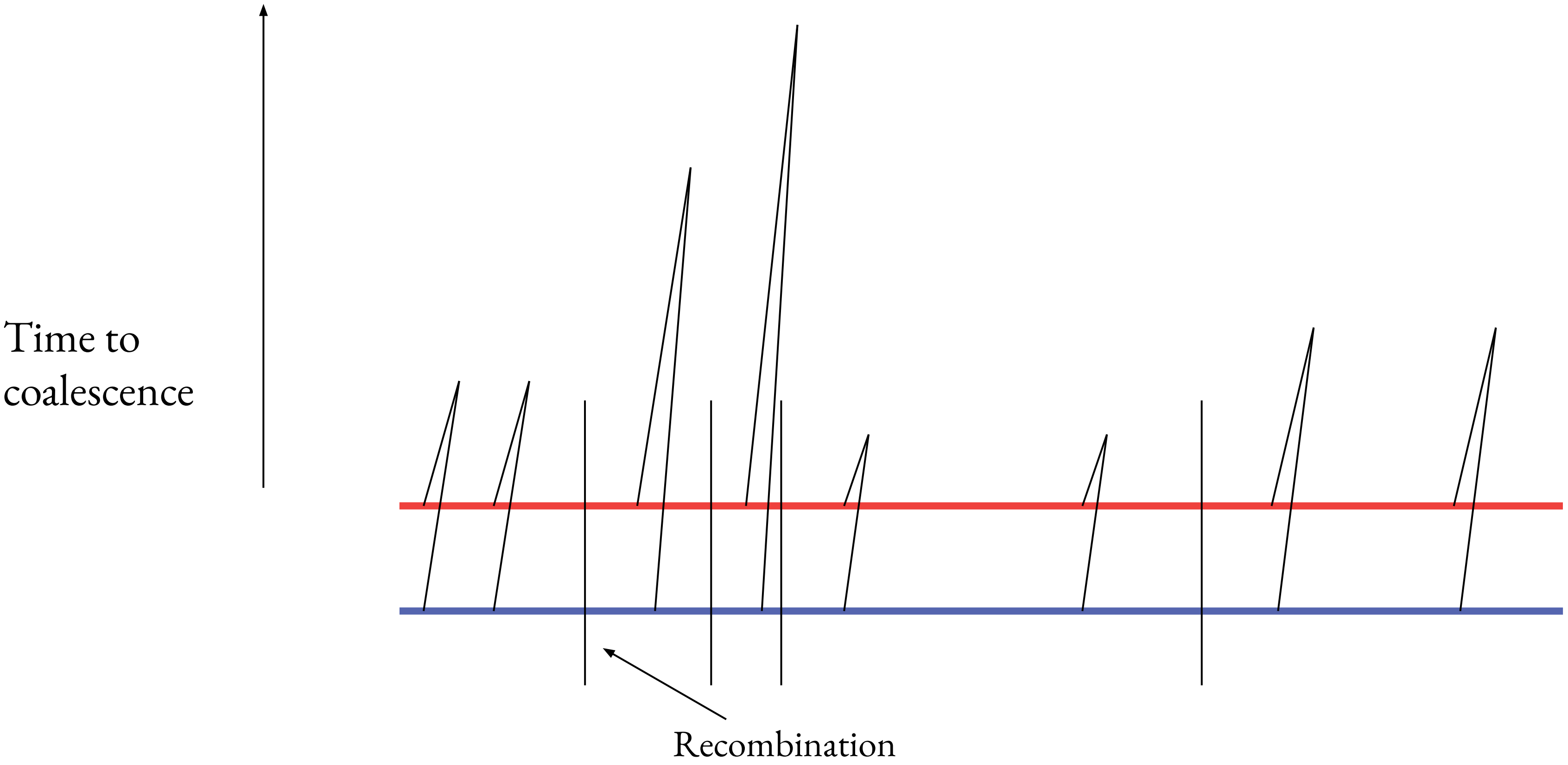
Information in one genome?



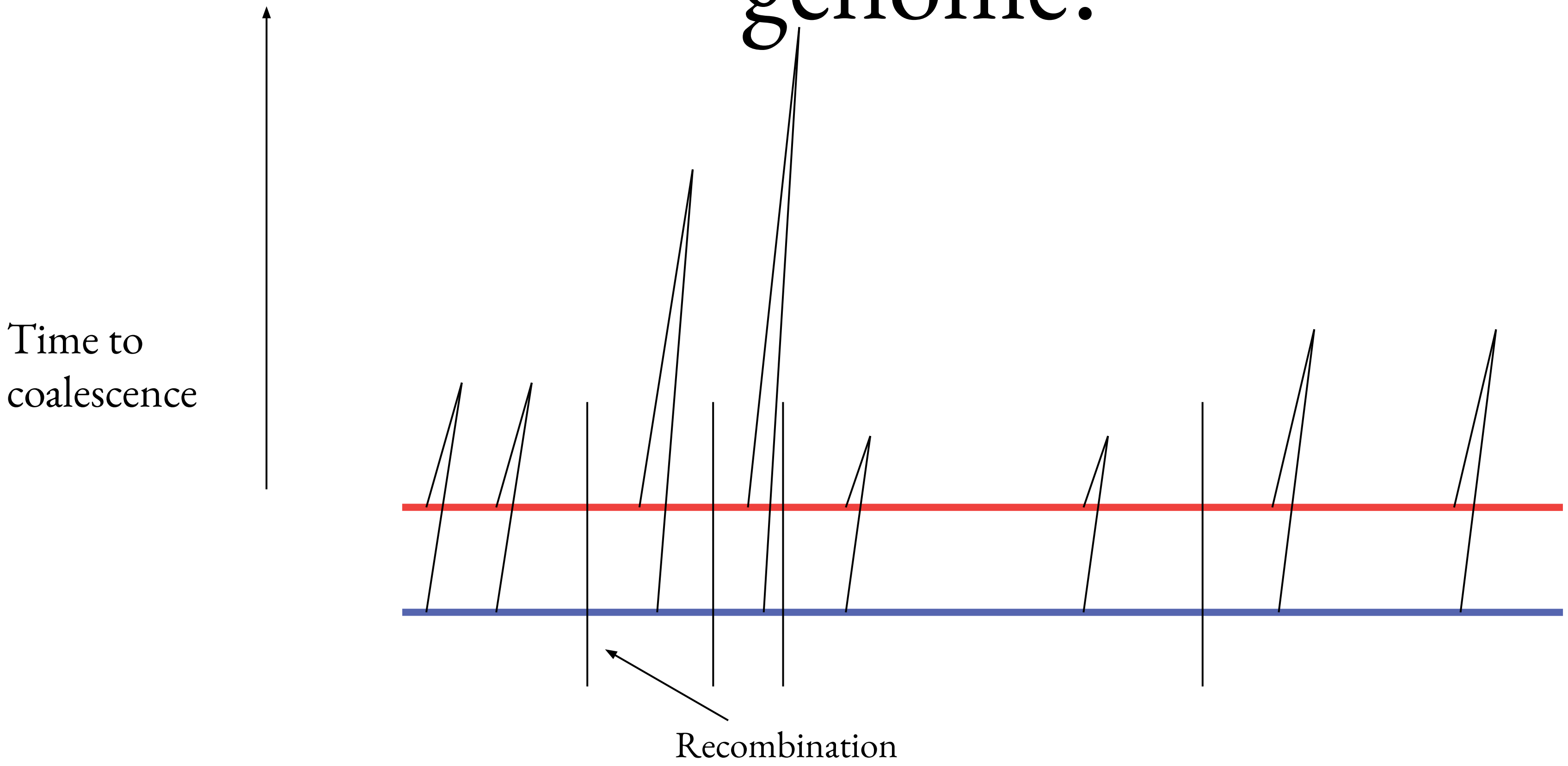
Information in one genome?



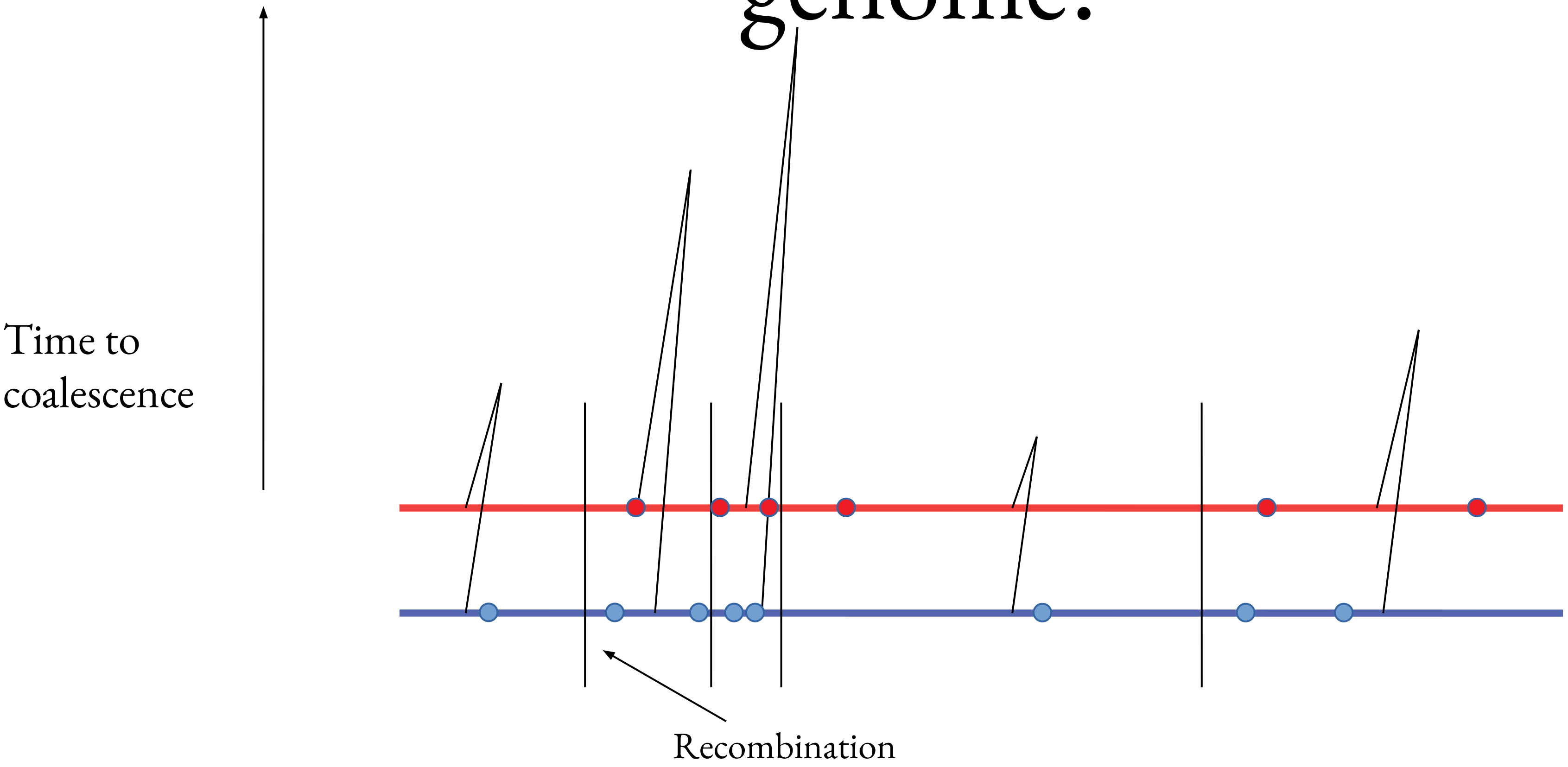
Information in one genome?



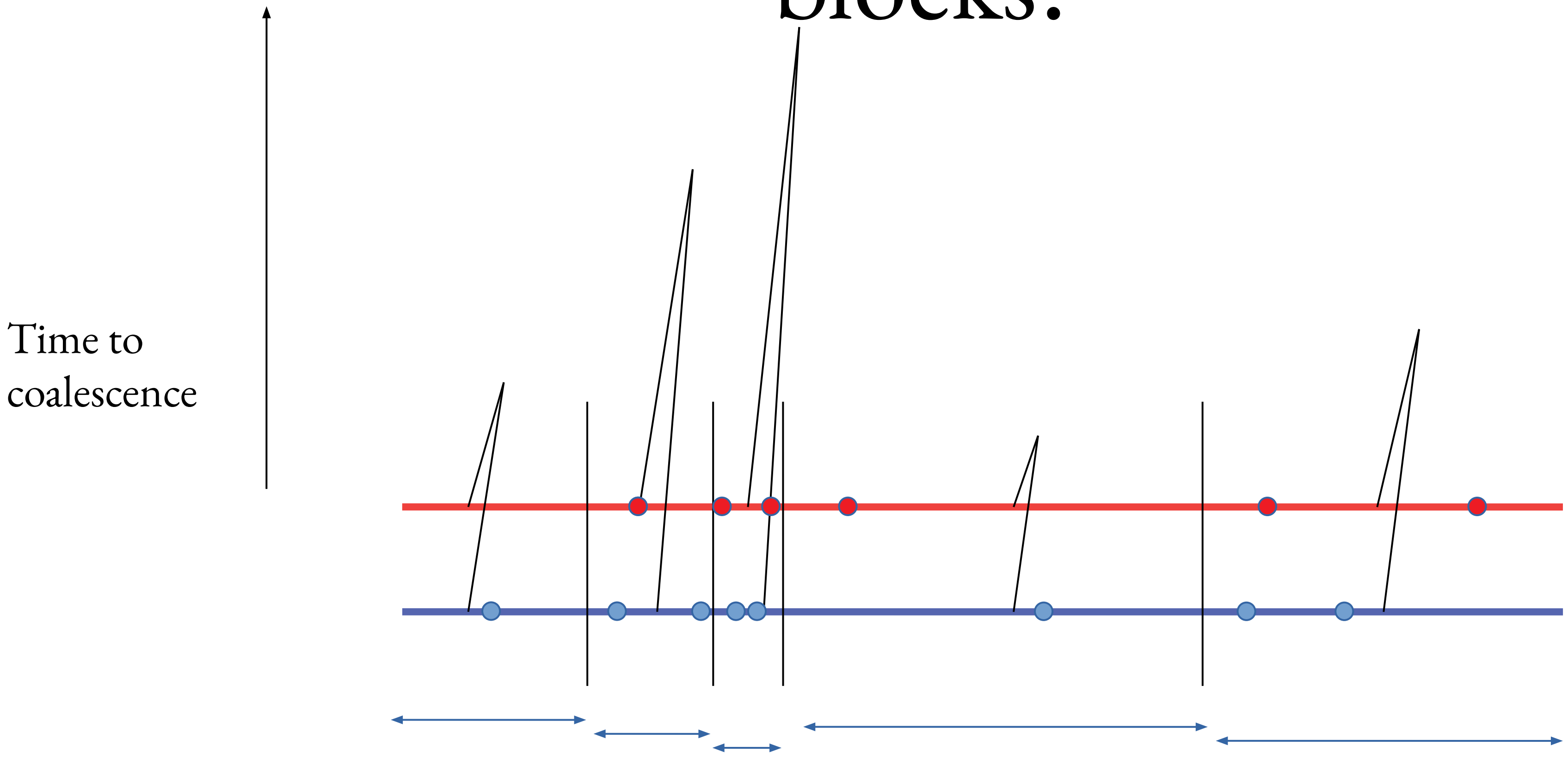
How will the mutations look on this genome?



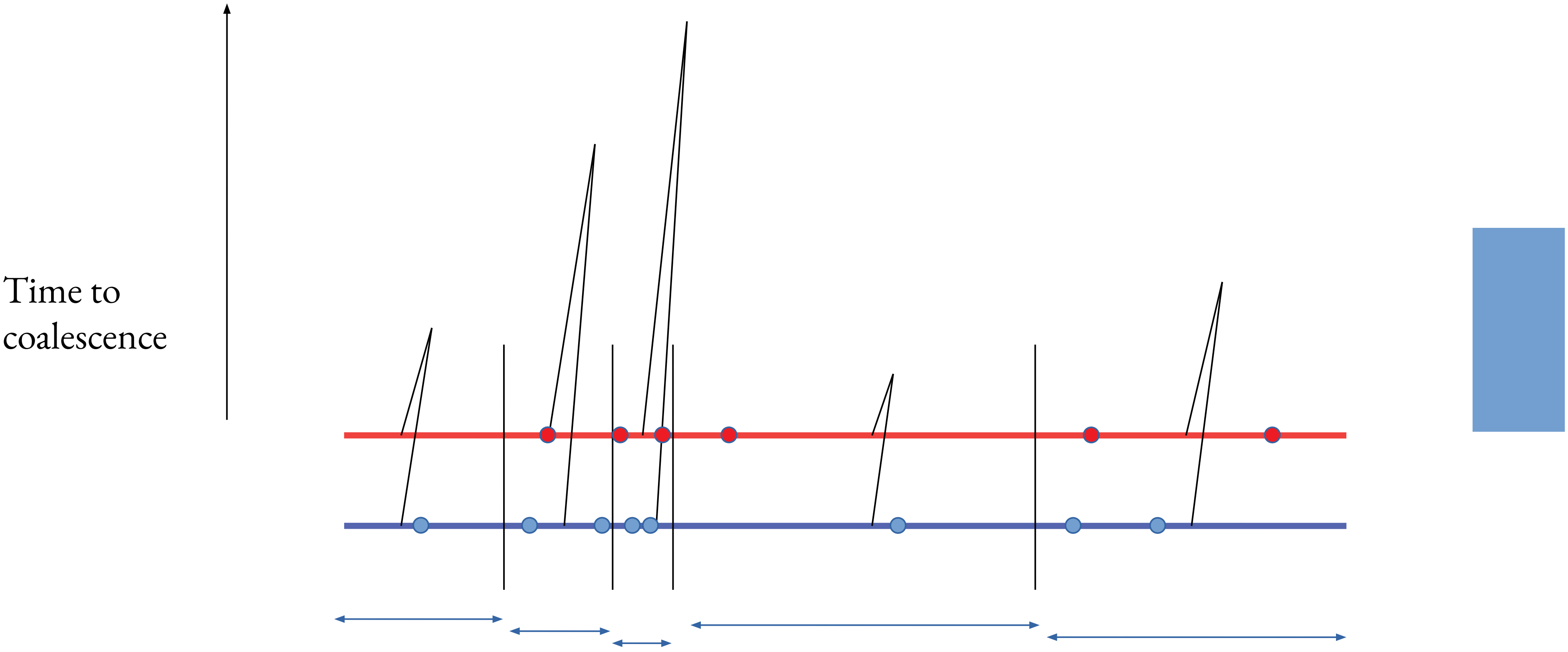
How will the mutations look on this genome?



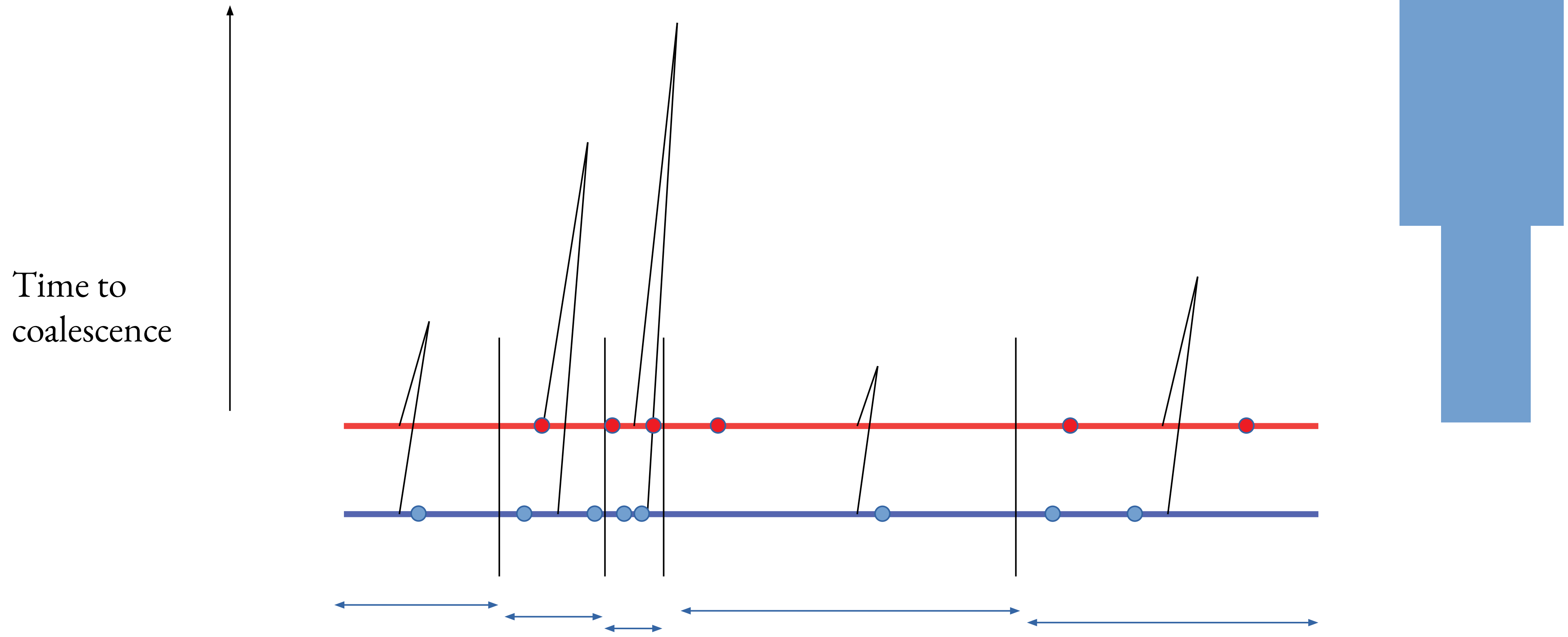
What about the size of the recombination blocks?



Reconstruct population history



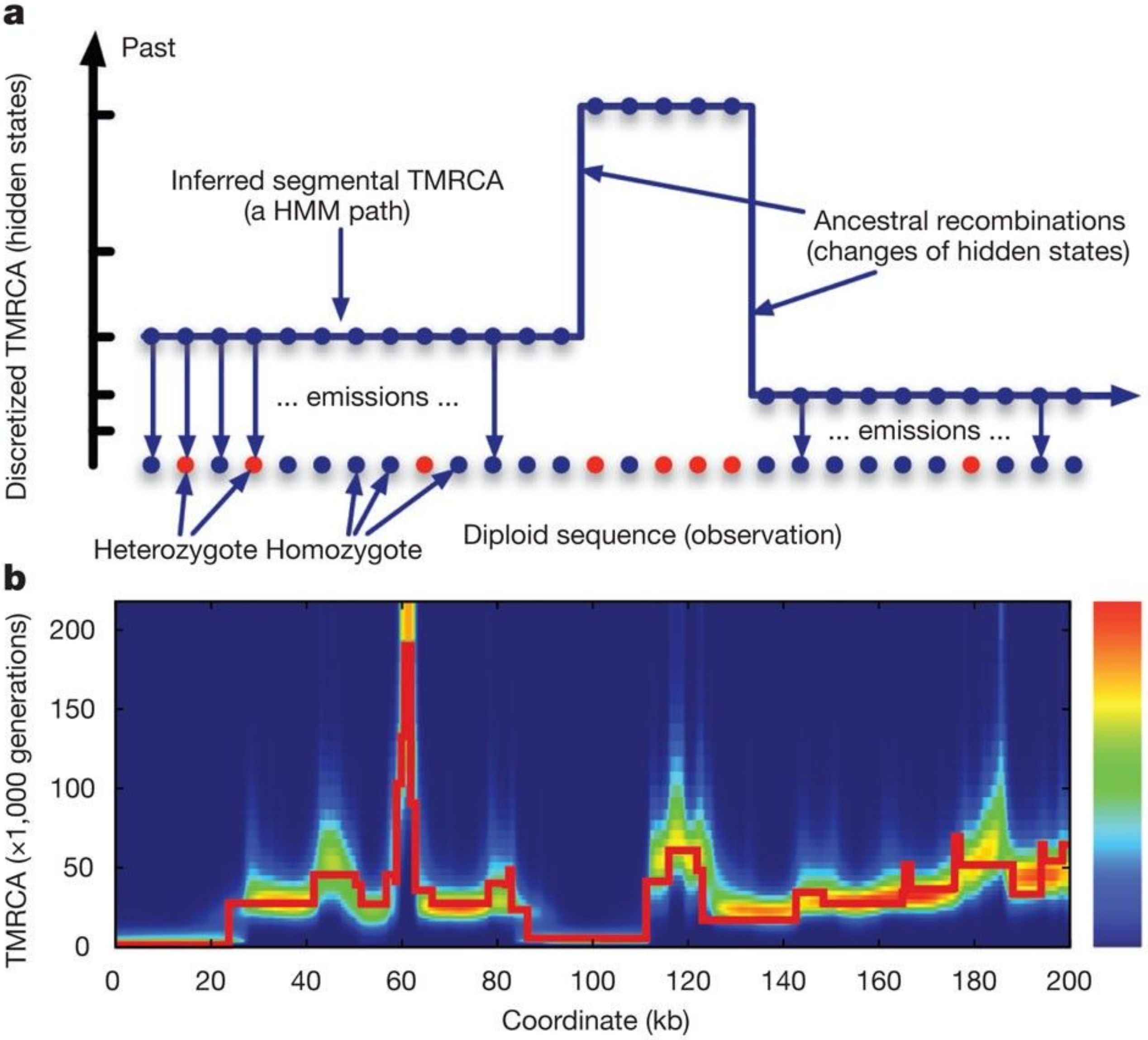
Reconstruct population history



PSMC: Pairwise Sequentially Markovian Coalescent

- Use recombination block sizes and density of mutations in genome to estimate population size, N_e
  - Hidden Markov Model (HMM)

PSMC


a
Past
Discretized TMRCA (hidden states)
Inferred segmental TMRCA
(a HMM path)
Ancestral recombinations
(changes of hidden states)
... emissions ...
... emissions ...
Heterozygote
Homozygote
Diploid sequence (observation)
b
TMRCA (×1,000 generations)
0
50
100
150
200
0
20
40
60
80
100
120
140
160
180
200
Coordinate (kb)

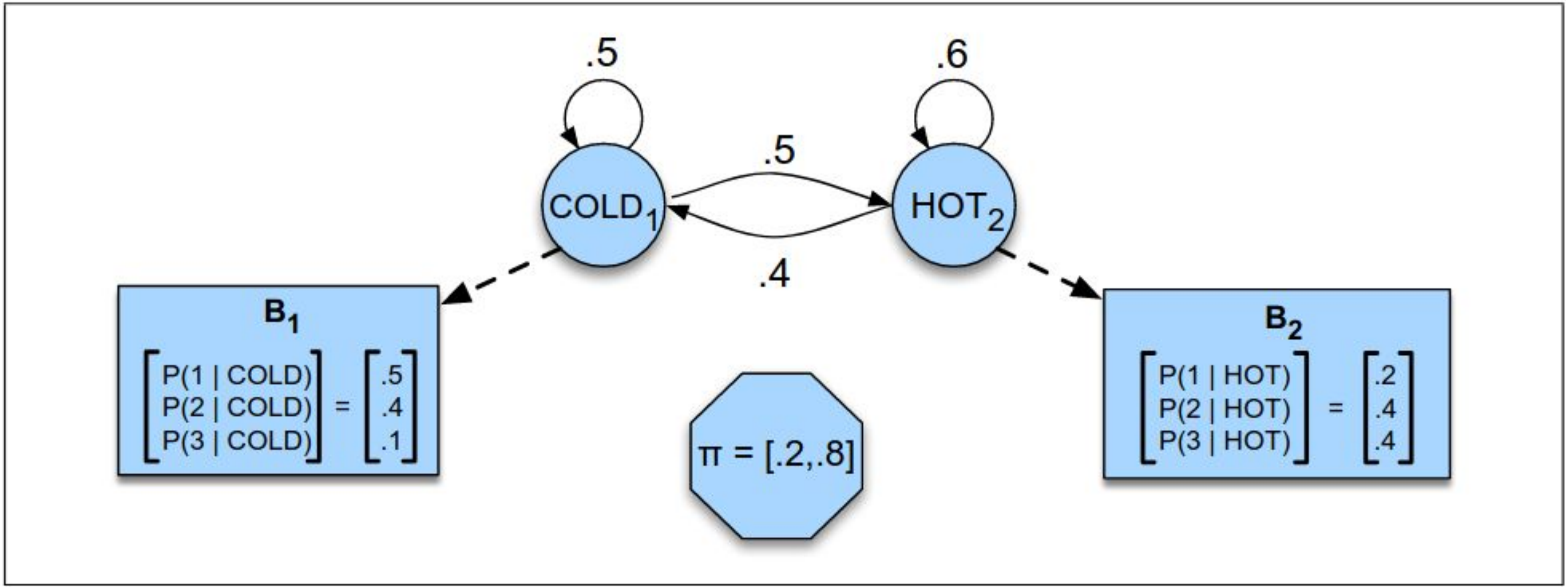
Quick detour #2: HMM

Eisner ice cream problem

Climatologist in 2800 AD who wants to understand the day to day temperature in 2000s, but all he has is the journal of Eisner who notes how many ice creams he eats every day.

Some simplifications + assumptions:
Only 2 day types - HOT or COLD
Eisner's ice cream consumption depends on day type
Day states form a markov chain – so today's day type affects tomorrow's day type.

Quick detour #2: HMM



B_1

$$\begin{bmatrix} P(1 \mid COLD) \\ P(2 \mid COLD) \\ P(3 \mid COLD) \end{bmatrix} = \begin{bmatrix} .5 \\ .4 \\ .1 \end{bmatrix}$$

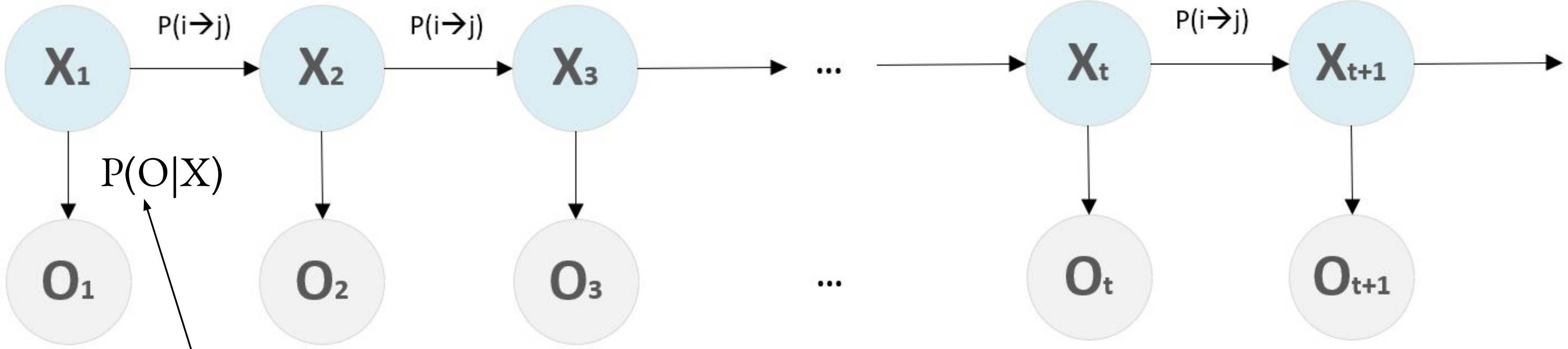
$\pi = [.2,.8]$

B_2

$$\begin{bmatrix} P(1 \mid HOT) \\ P(2 \mid HOT) \\ P(3 \mid HOT) \end{bmatrix} = \begin{bmatrix} .2 \\ .4 \\ .4 \end{bmatrix}$$

Quick detour #2: HMM

$P(X_{t+1}|X_t)$ → Transition probabilities



$P(O|X)$

Emission probabilities

Quick detour #2: HMM

Three problems we need to solve in a HMM:

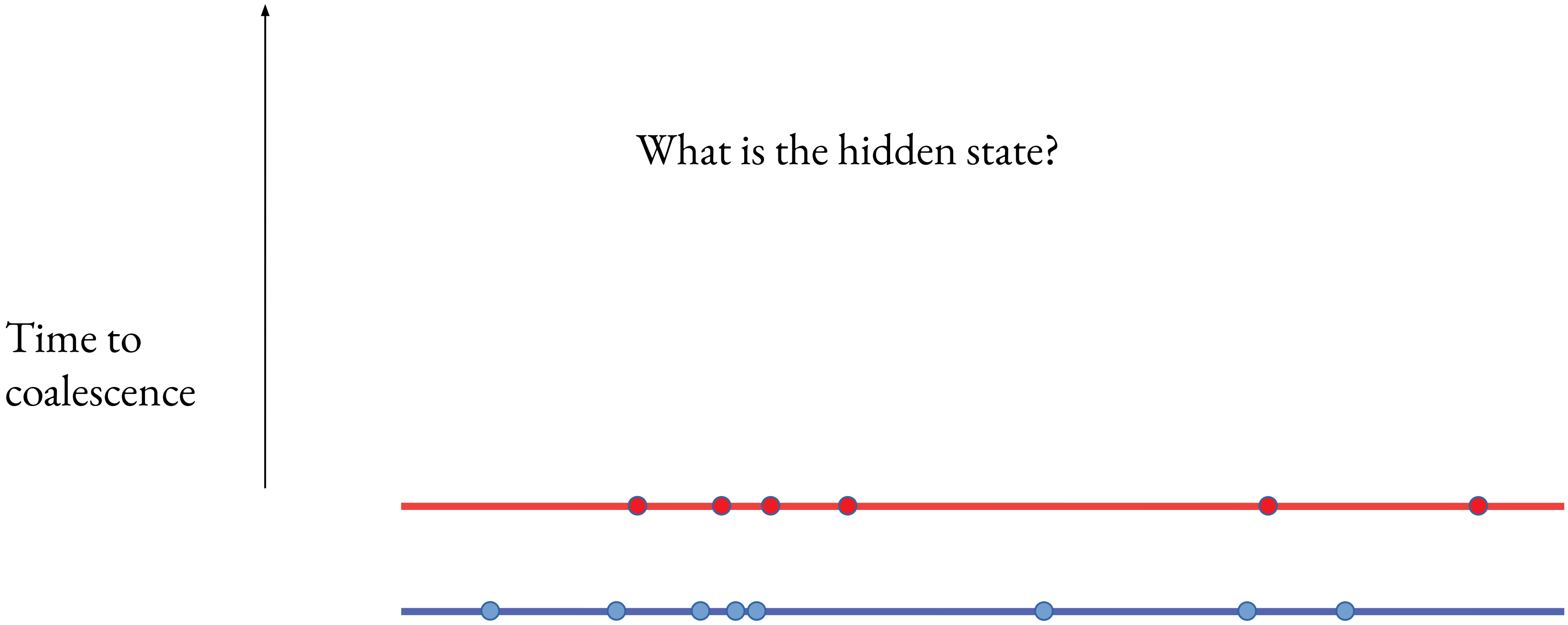
1. **Likelihood**: Given observations O, and parameters Θ of the HMM, we need to be able to compute the P(O | Θ)

2. **Decoding**: Given observations O, and parameters Θ, compute the best hidden state sequence X_1, X_2 ... X_n

3. **Learning**: Given observations O and the set of states in the HMM, learn the parameters Θ.

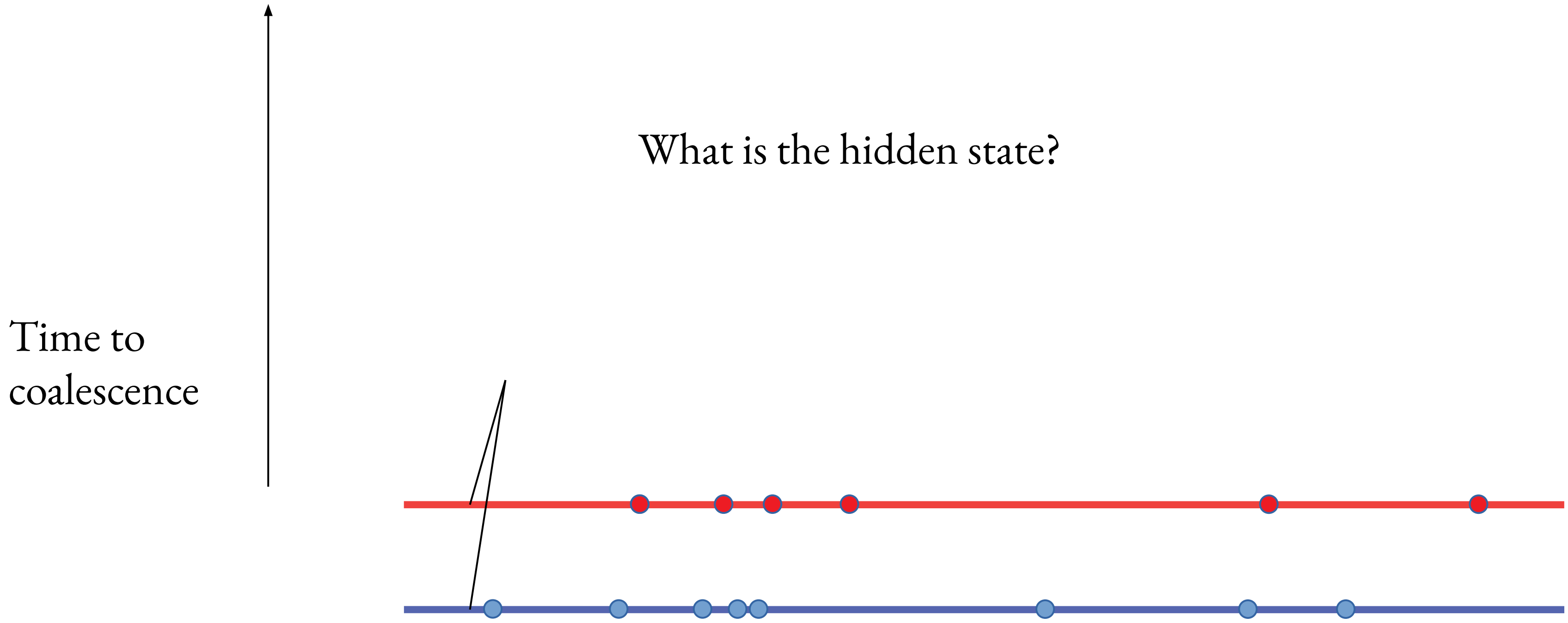
PSMC: HMM for genome

What is the hidden state?

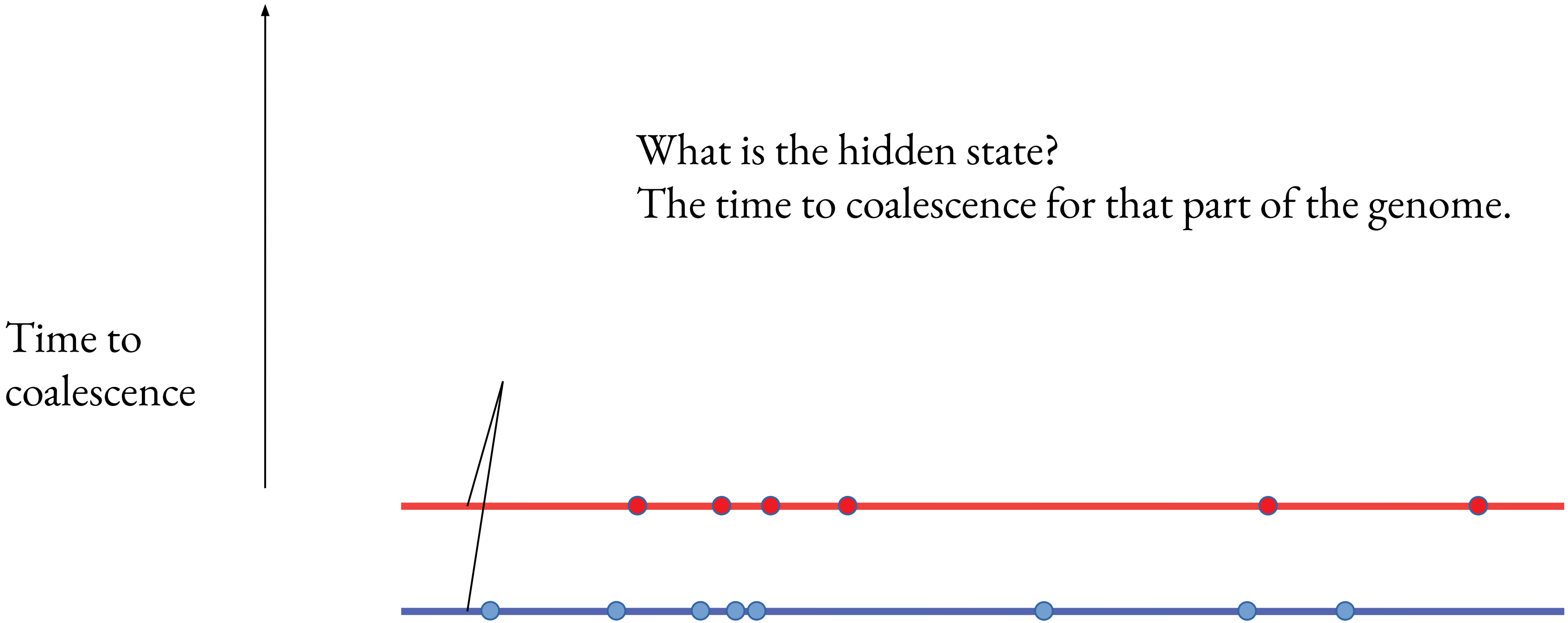


PSMC: HMM for genome

What is the hidden state?

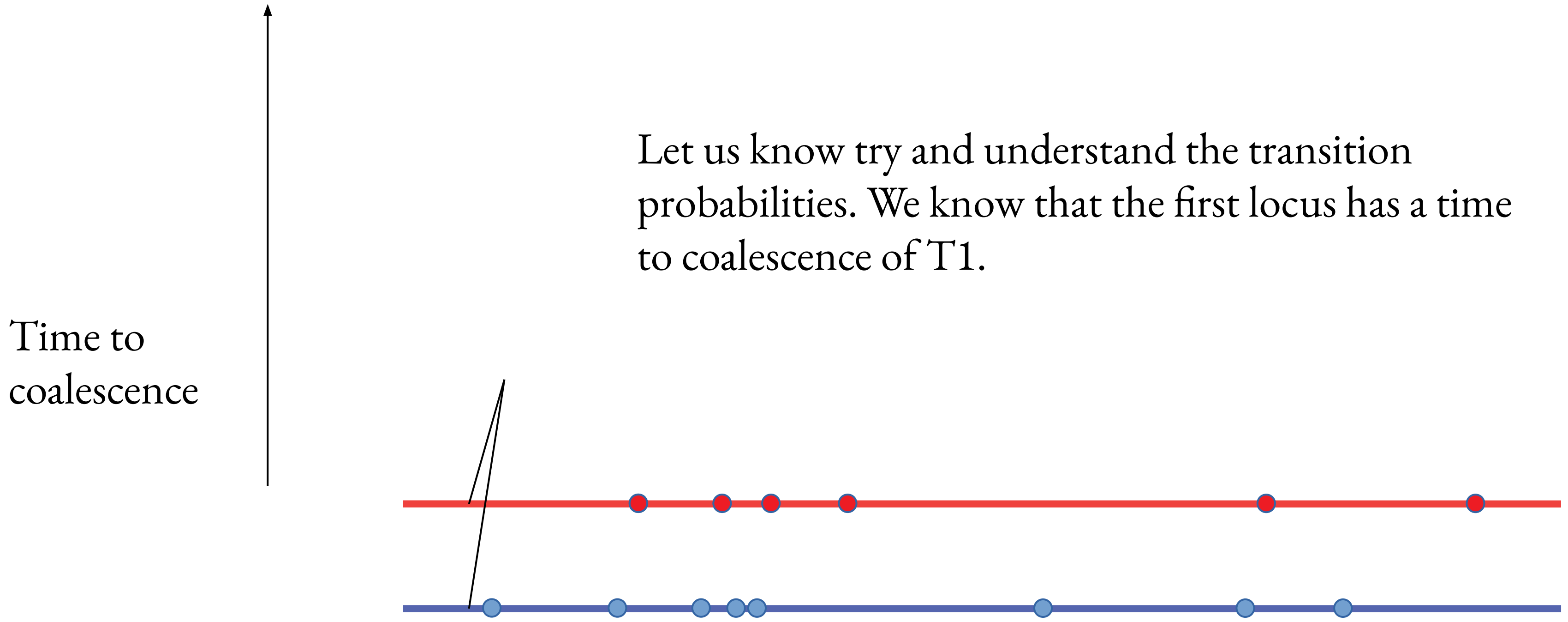


PSMC: HMM for genome



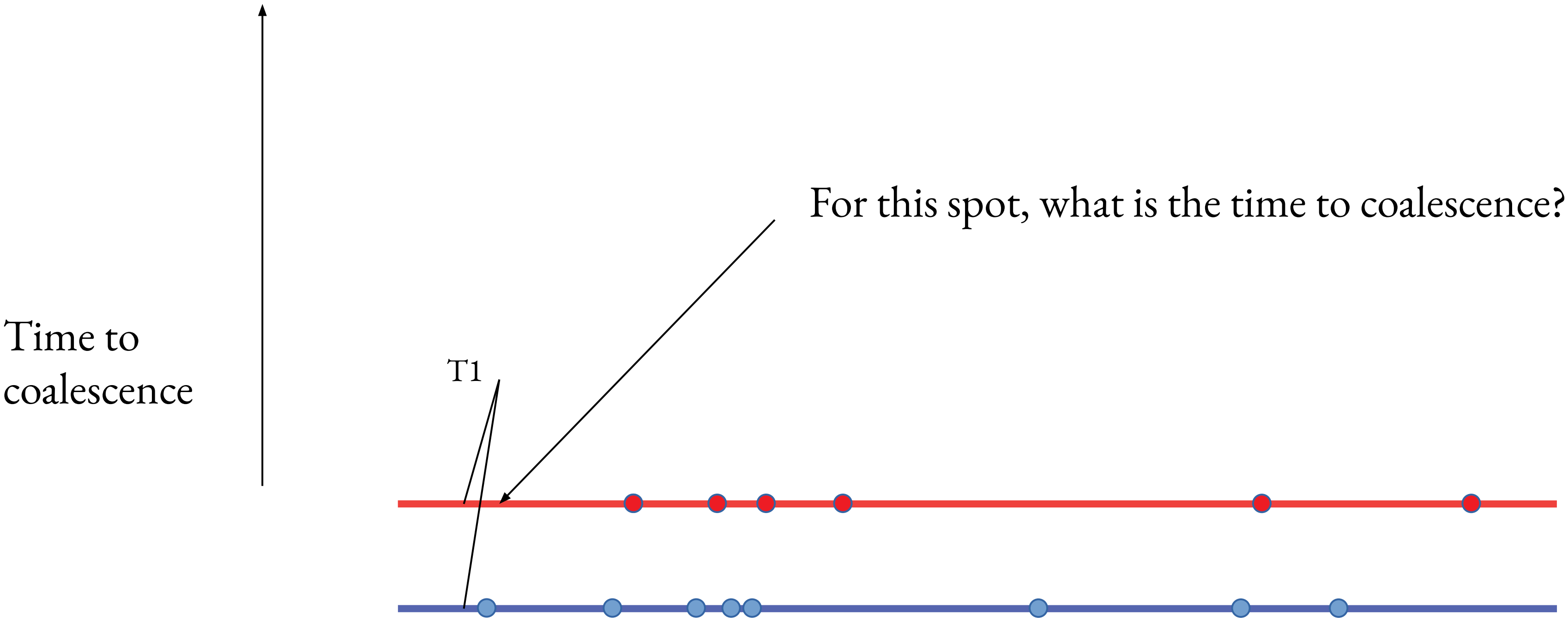
What is the hidden state?
The time to coalescence for that part of the genome.

PSMC: HMM for genome



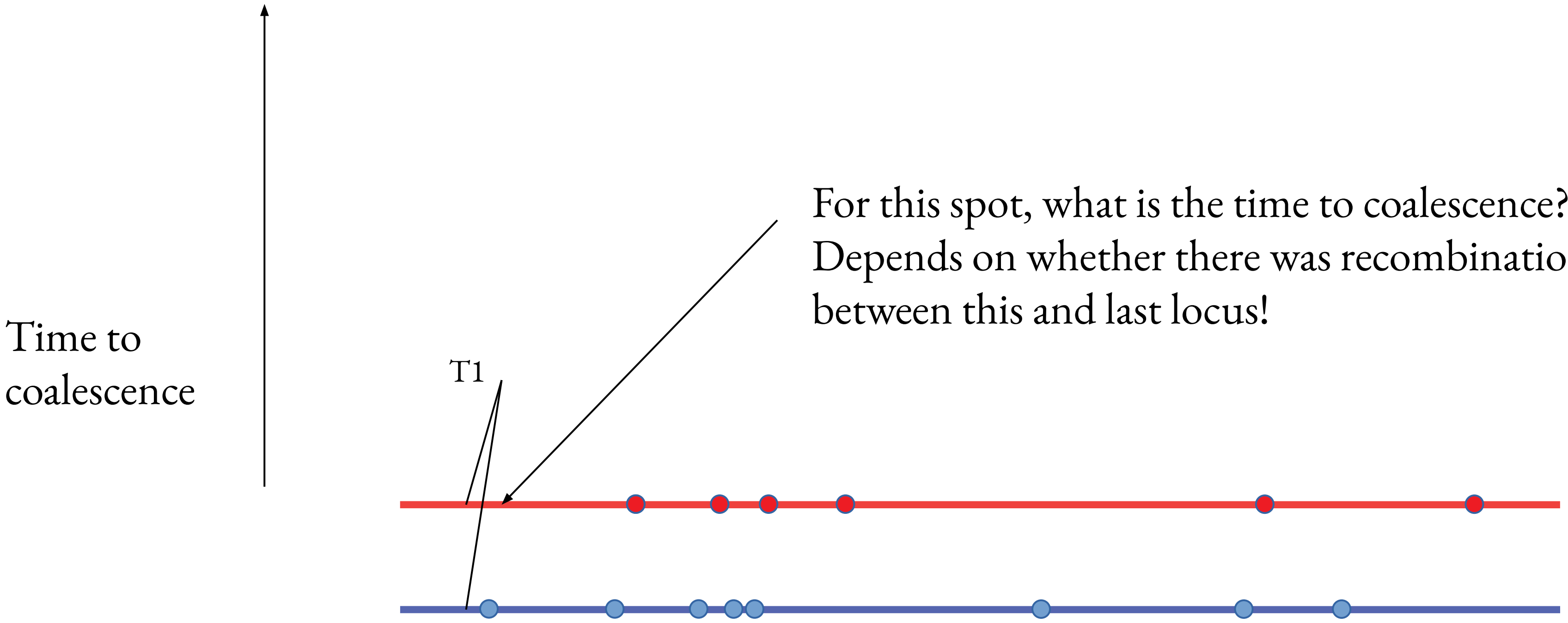
Let us know try and understand the transition probabilities. We know that the first locus has a time to coalescence of T1.

PSMC: HMM for genome



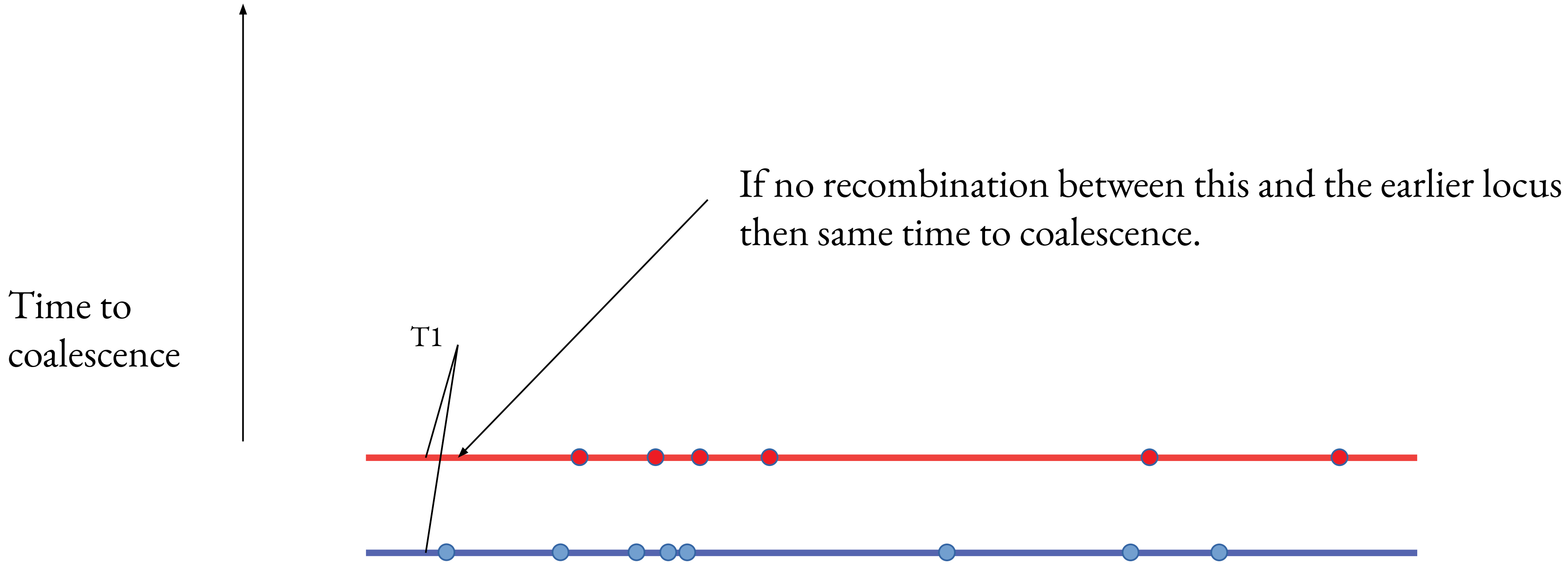
For this spot, what is the time to coalescence?

PSMC: HMM for genome



For this spot, what is the time to coalescence? Depends on whether there was recombination between this and last locus!

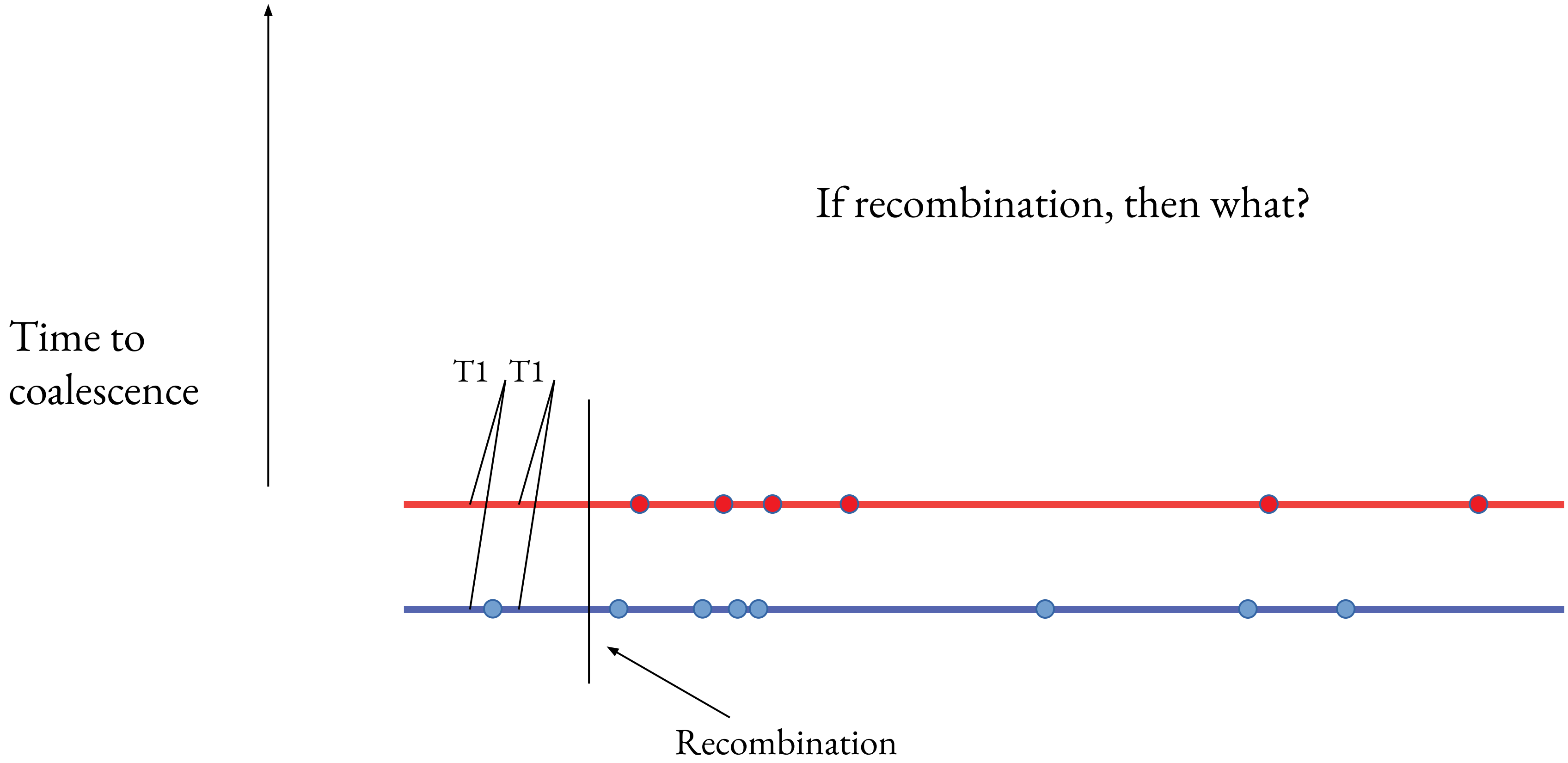
PSMC: HMM for genome



If no recombination between this and the earlier locus then same time to coalescence.

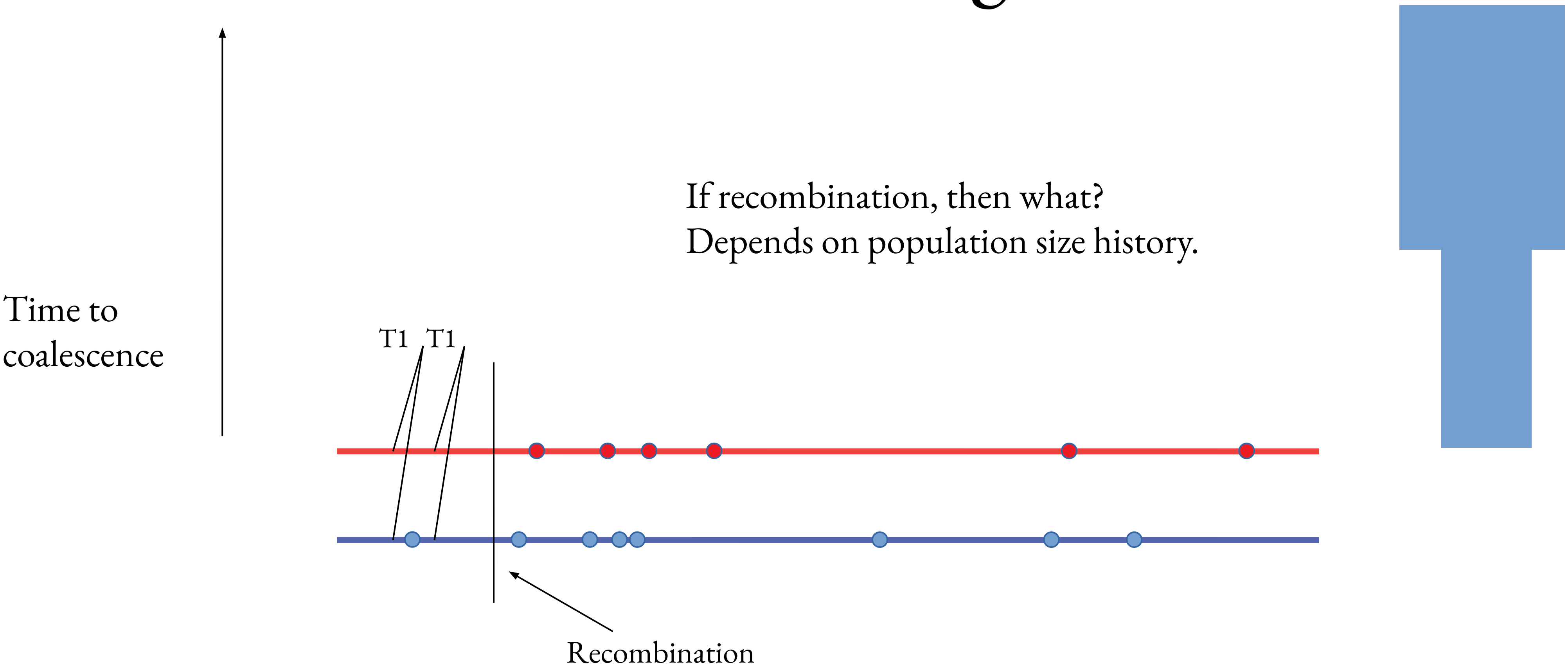
PSMC: HMM for genome

If recombination, then what?

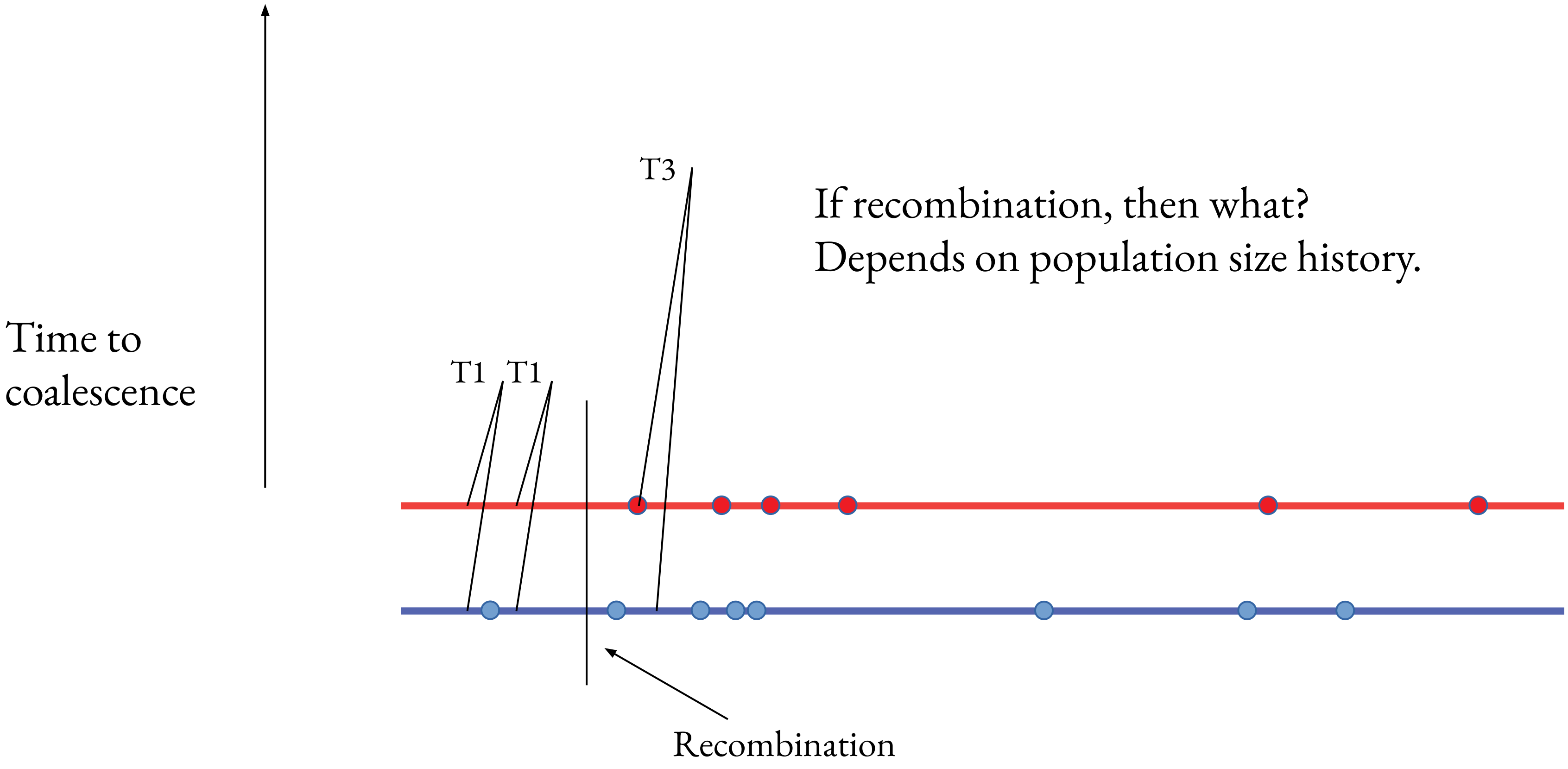


PSMC: HMM for genome

If recombination, then what?
Depends on population size history.



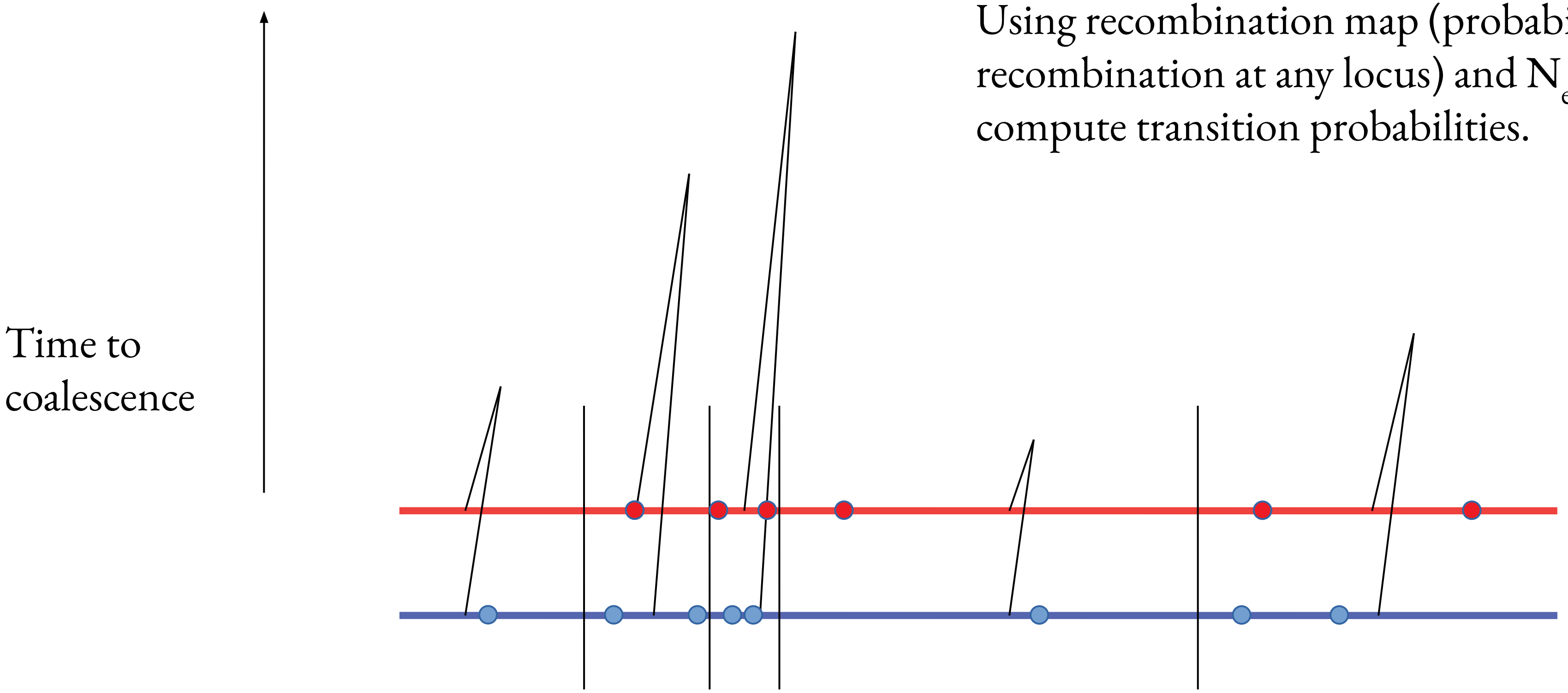
PSMC: HMM for genome



If recombination, then what?
Depends on population size history.

PSMC: HMM for genome

Using recombination map (probability of recombination at any locus) and N_e, we can compute transition probabilities.

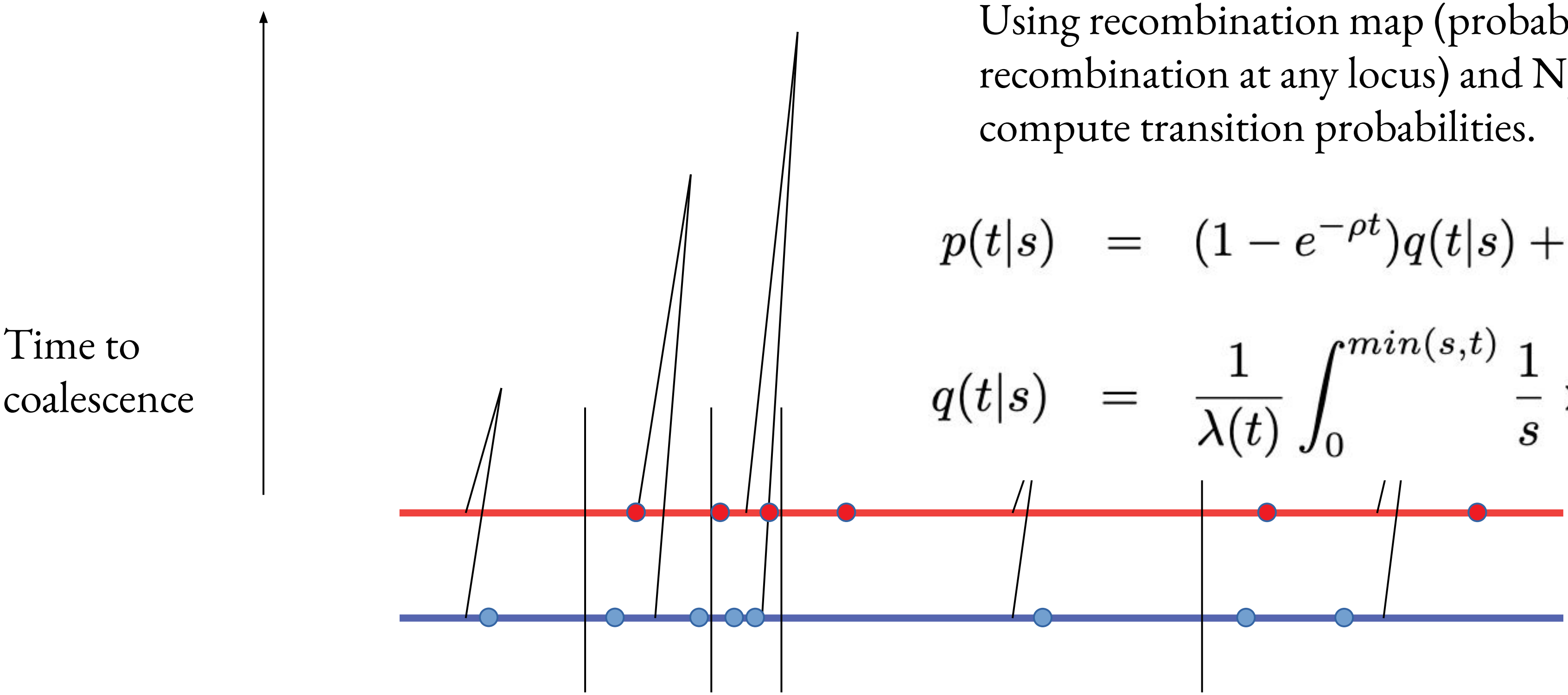


PSMC: HMM for genome

Using recombination map (probability of recombination at any locus) and N_e, we can compute transition probabilities.

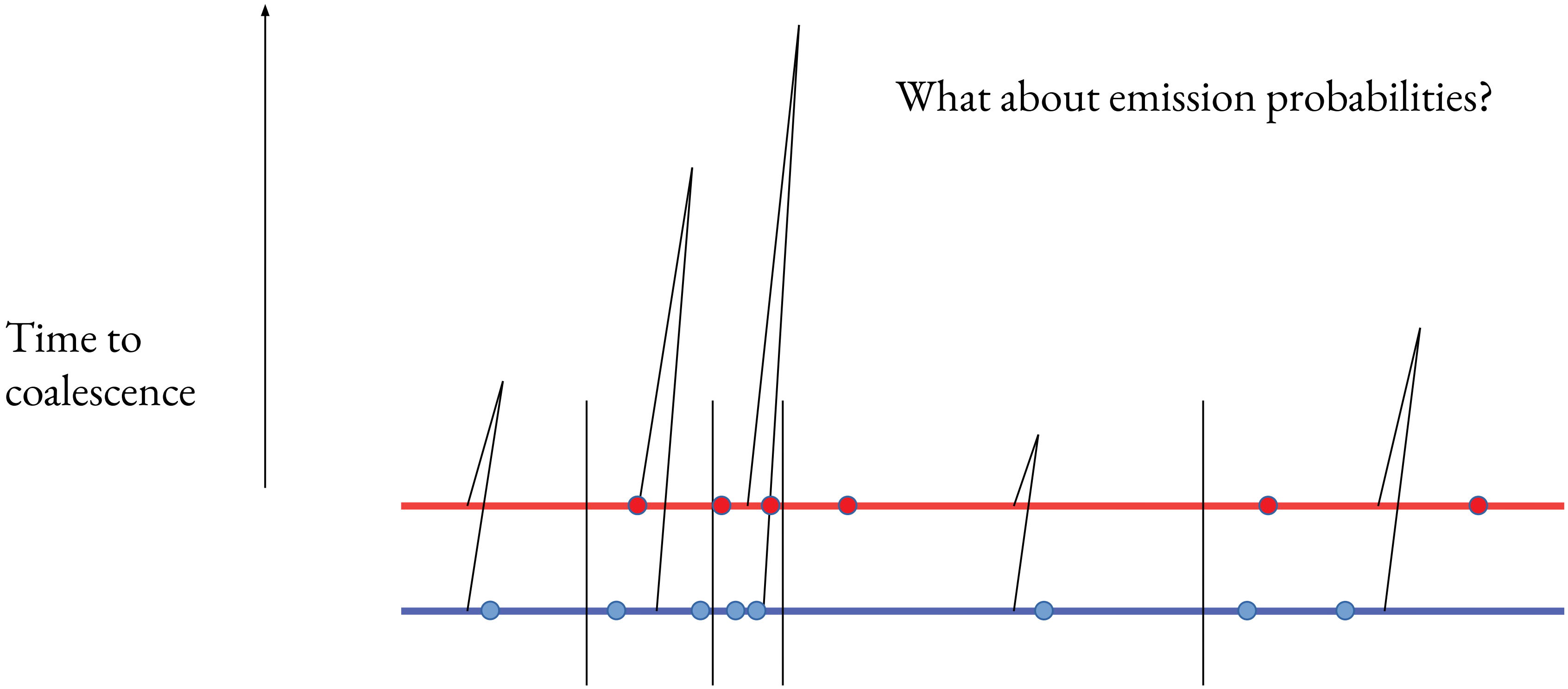
$$p(t|s) \quad = \quad (1-e^{-\rho t})q(t|s) + e^{-\rho s}\delta(t-s)$$

$$q(t|s) \quad = \quad \frac{1}{\lambda(t)}\int_0^{min(s,t)} \frac{1}{s} \times e^{-\int_u^t \frac{dv}{\lambda(v)}}$$



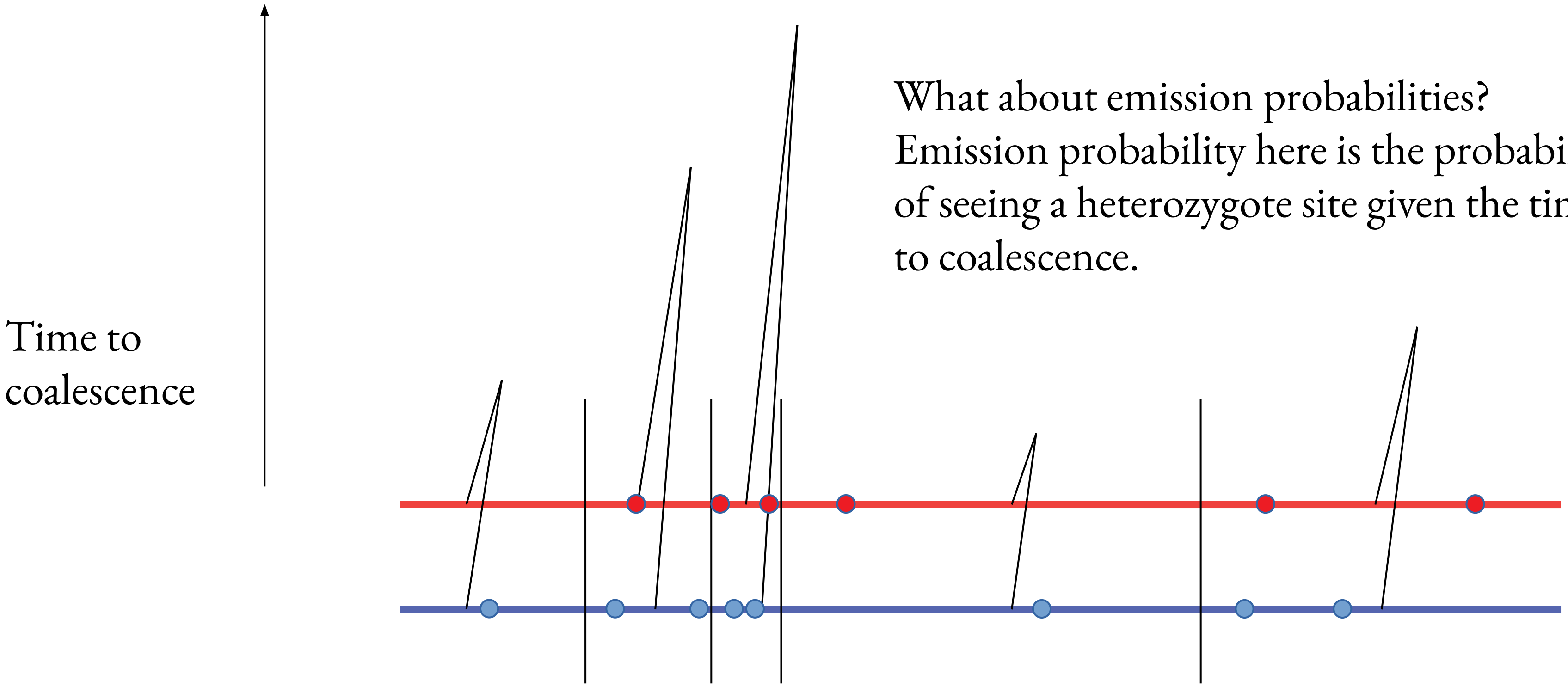
PSMC: HMM for genome

What about emission probabilities?



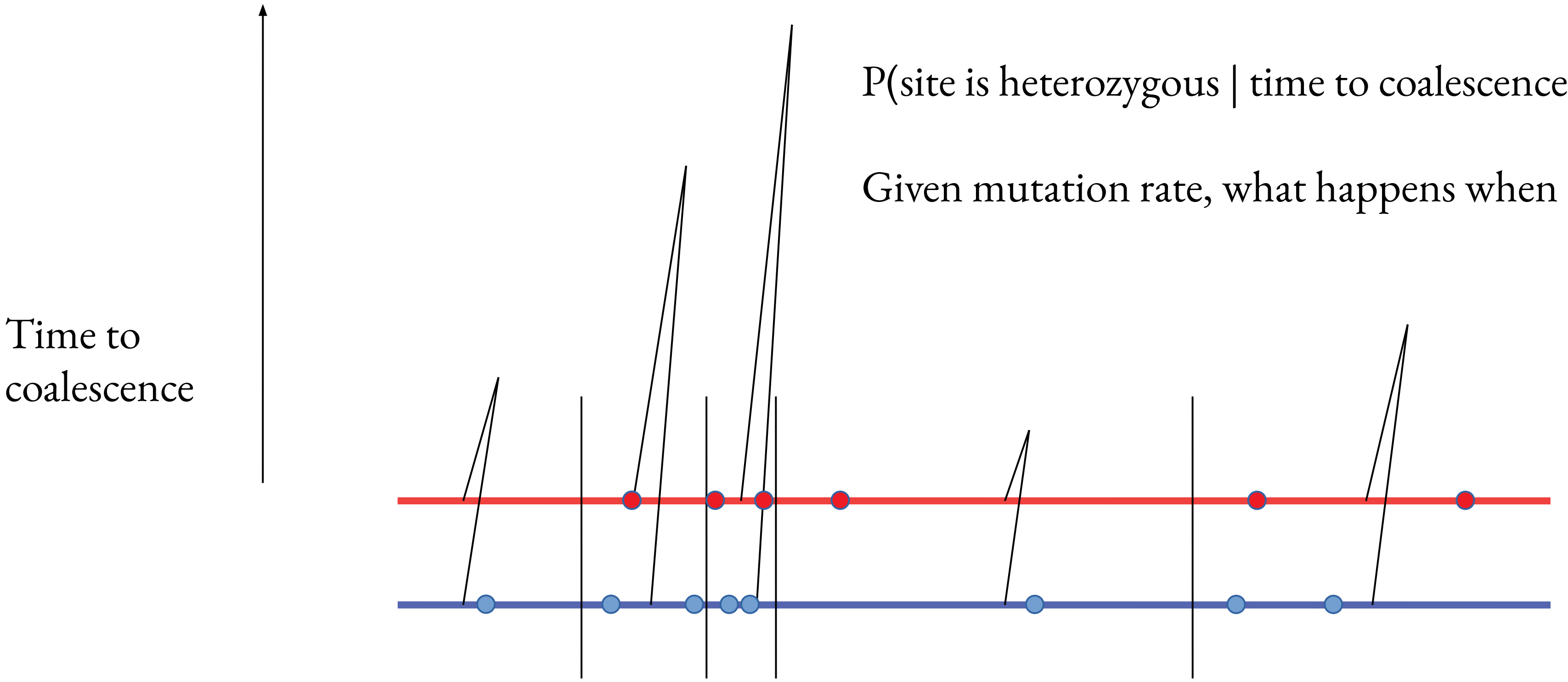
PSMC: HMM for genome

What about emission probabilities?
Emission probability here is the probability of seeing a heterozygote site given the time to coalescence.



P(site is heterozygous | time to coalescence = T)

PSMC: HMM for genome



P(site is heterozygous | time to coalescence = T)

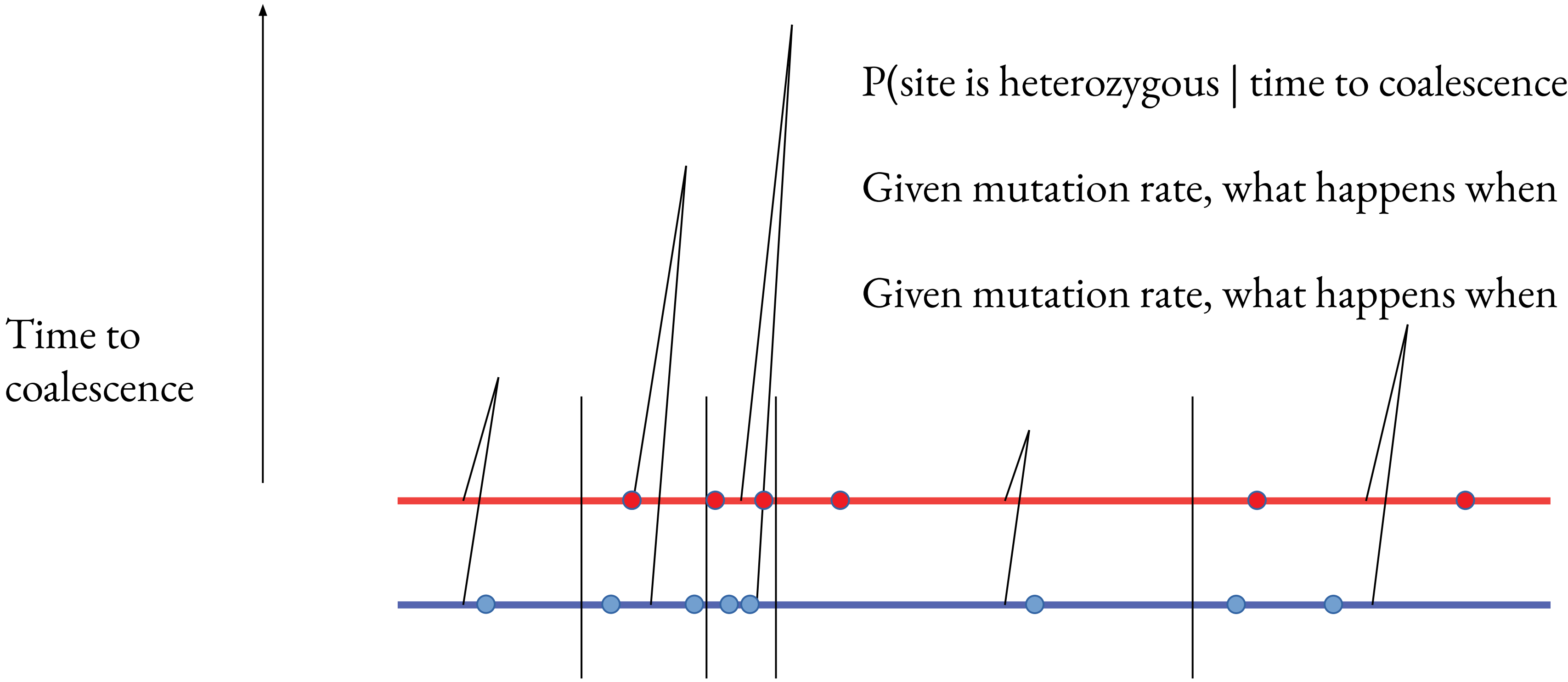
Given mutation rate, what happens when T ↑

PSMC: HMM for genome

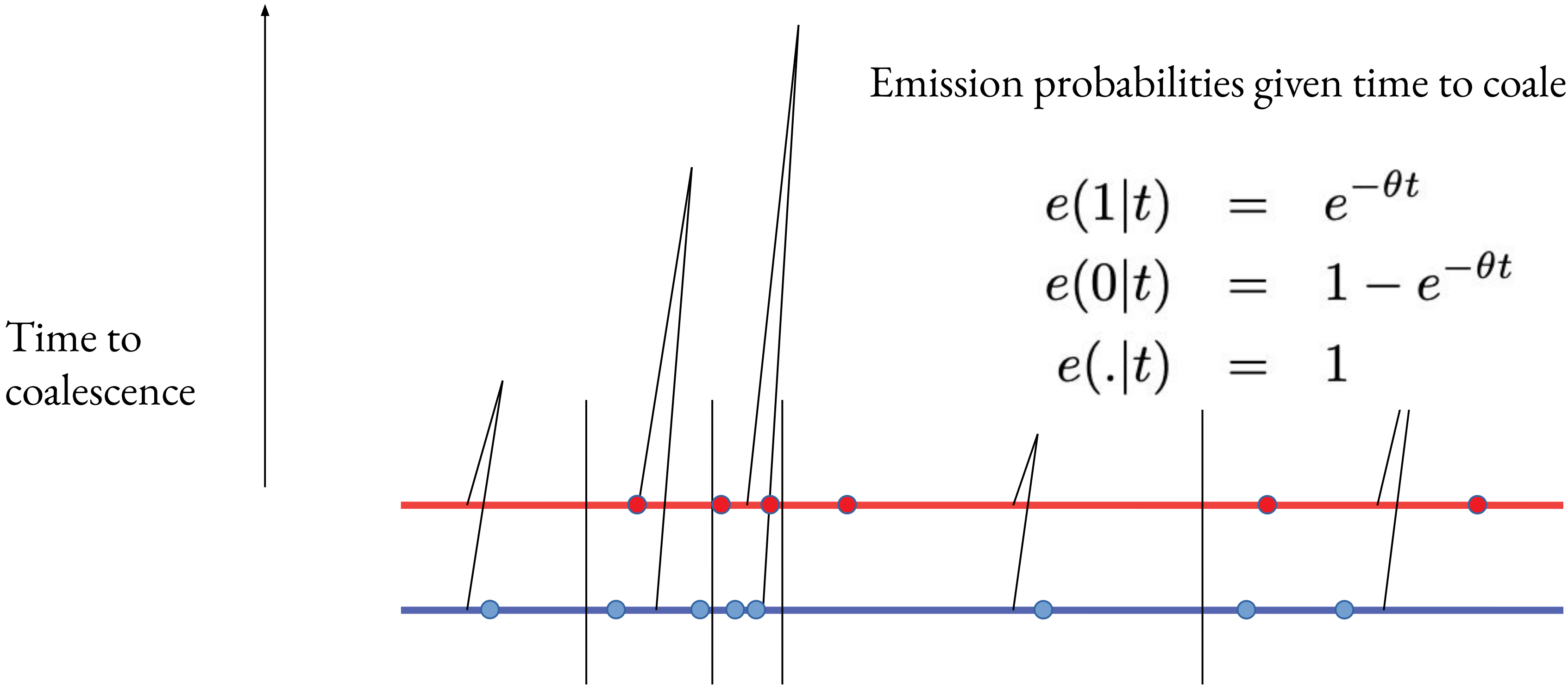
P(site is heterozygous | time to coalescence = T)

Given mutation rate, what happens when T ↑

Given mutation rate, what happens when T ↓



PSMC: HMM for genome

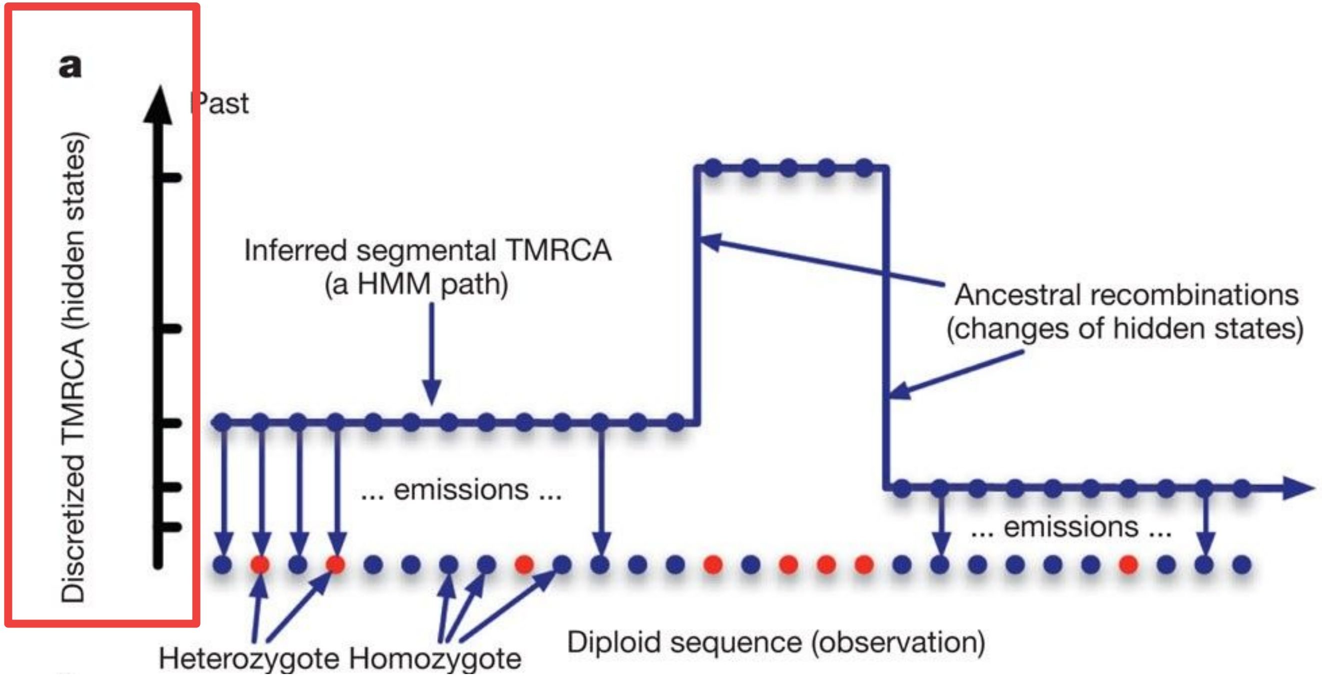


Emission probabilities given time to coalescence = T)

$$
\begin{aligned}
e(1|t) &= e^{-\theta t} \\
e(0|t) &= 1 - e^{-\theta t} \\
e(.|t) &= 1
\end{aligned}
$$

PSMC: Some missing bits

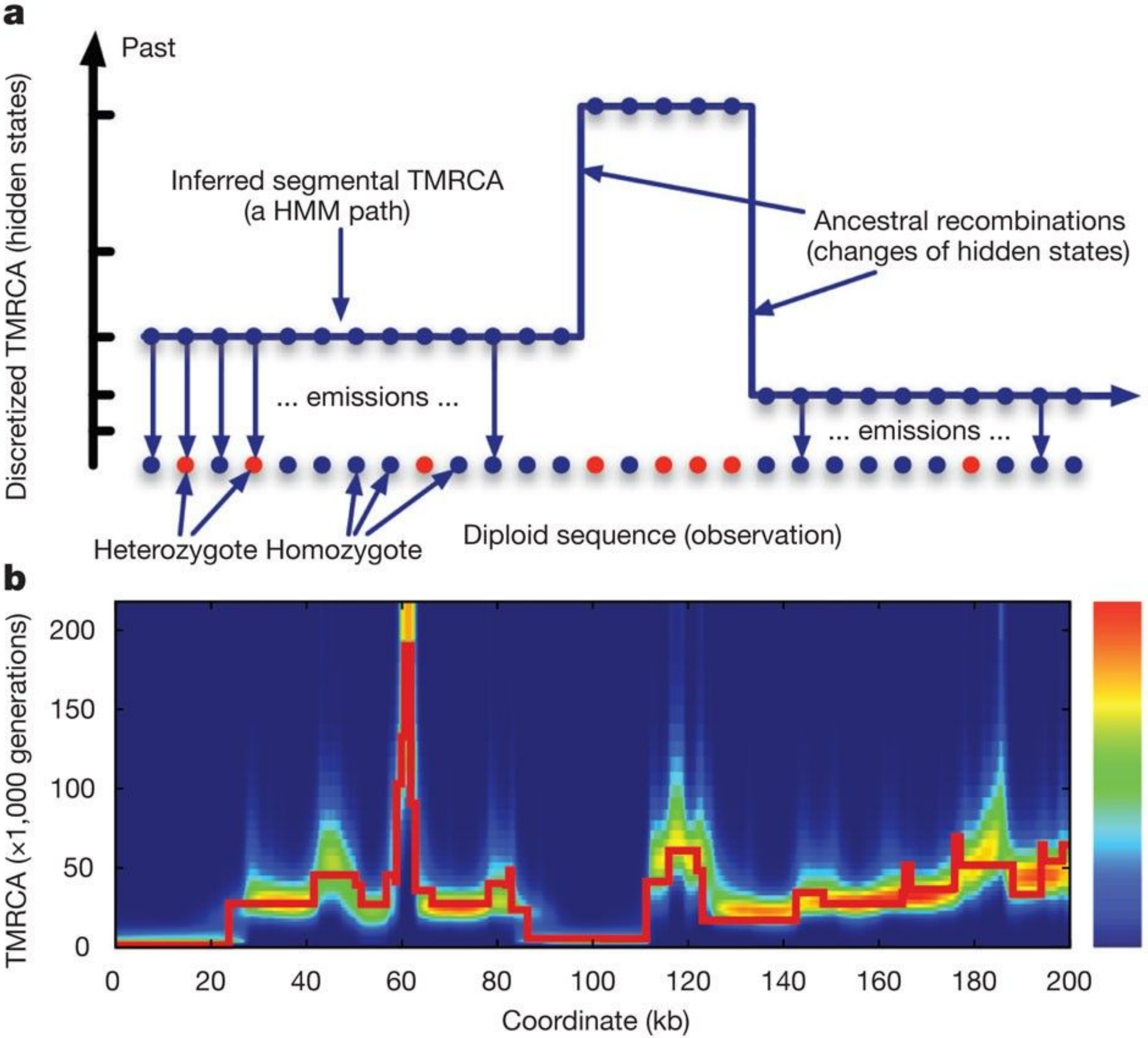
- Time is discretized
  - On log scale, so expected number of coalescent events in each bin ~ equal



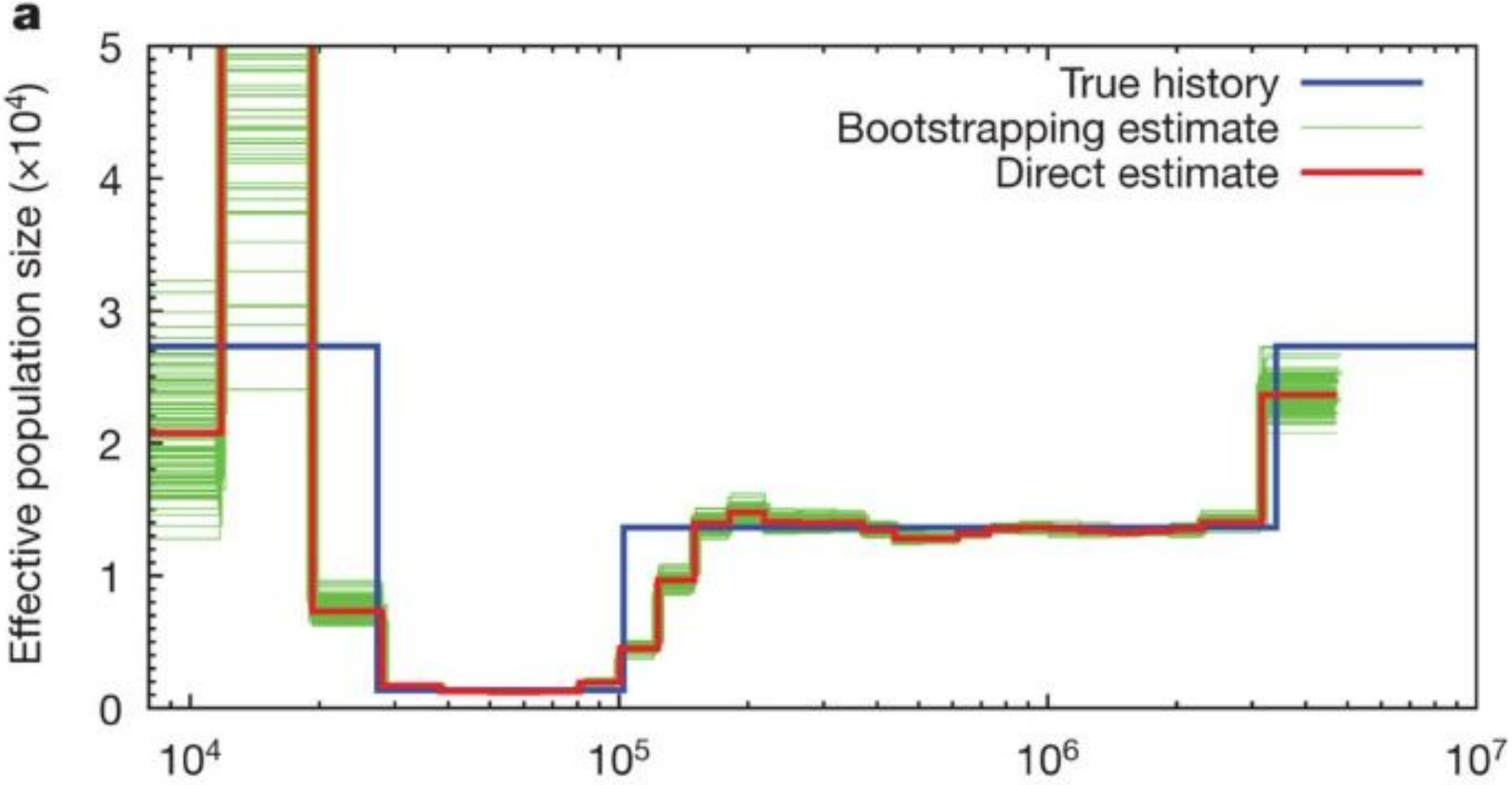
PSMC: Some missing bits

- Time is discretized
  - On log scale, so expected number of coalescent events in each bin ~ equal
- Mutation rate and mutation/recombination ratio are additional parameters
  - What happens if recombination rate is similar to mutation rate?

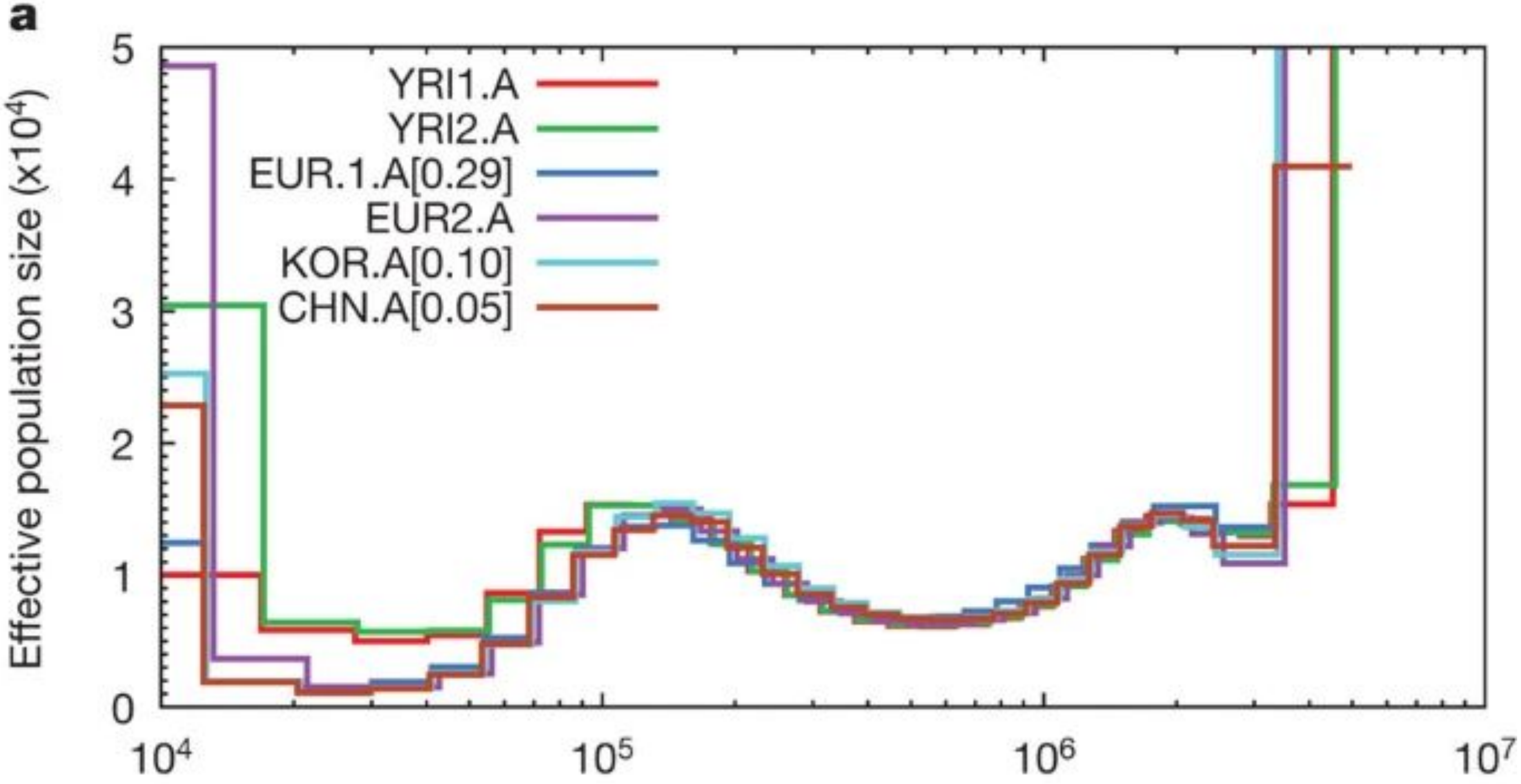
PSMC: Putting it all together



PSMC on simulated data



PSMC on human populations



Conclusions

- Many ways to skin a cat
  - SFS based, LD based, coalescent based, summary statistics based
- Pairwise Sequentially Markovian Coalescent
  - Lots of information in 1 genome
  - Remove problem with tree topology
  - Lots of methods in same framework – MSMC, SMC++, MiSTI

Exercises after the break