

Demography estimation

using PSMC

Shyam Gopalakrishnan

Aug 23rd 2024

Outline

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

Demography

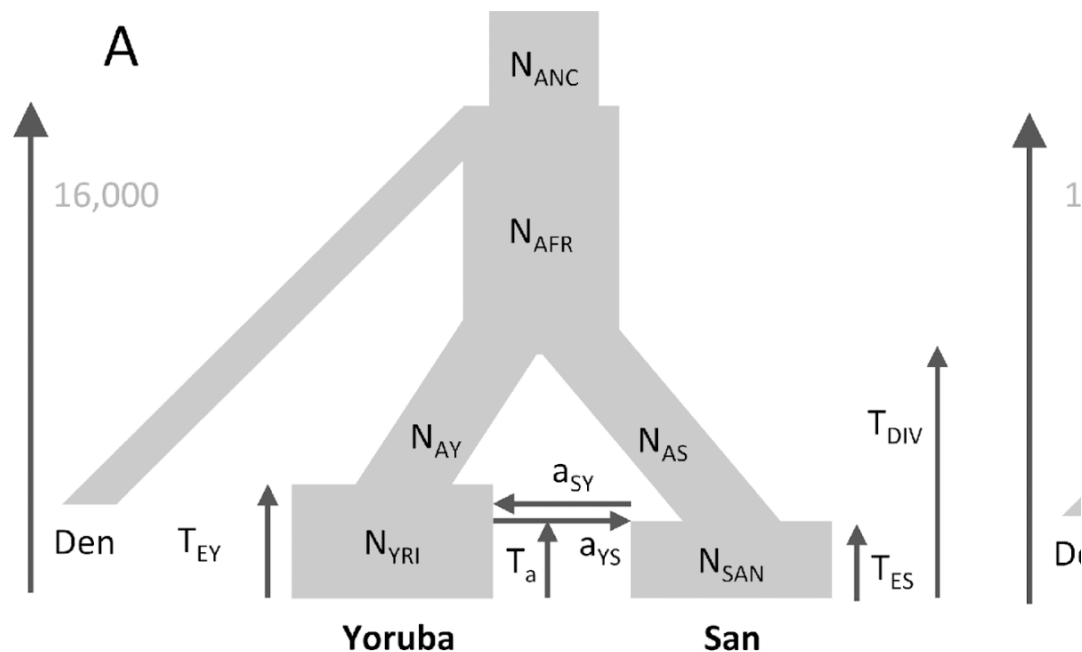
- What do we mean by demography?

Demography

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

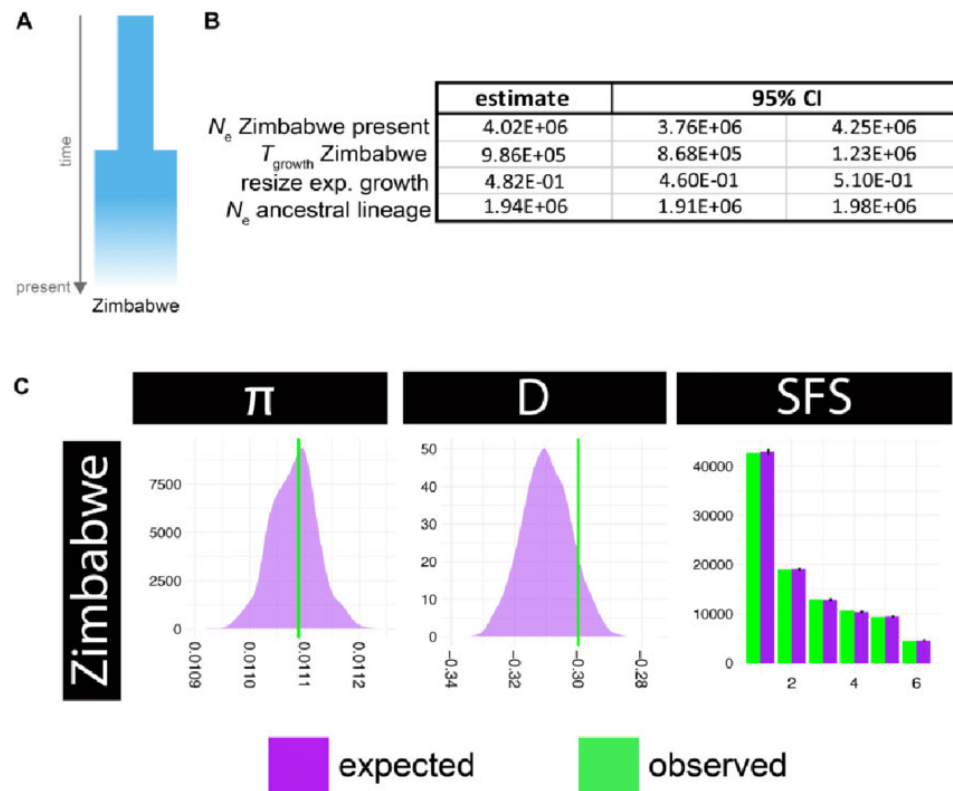
Demography

- What do we mean by demography?



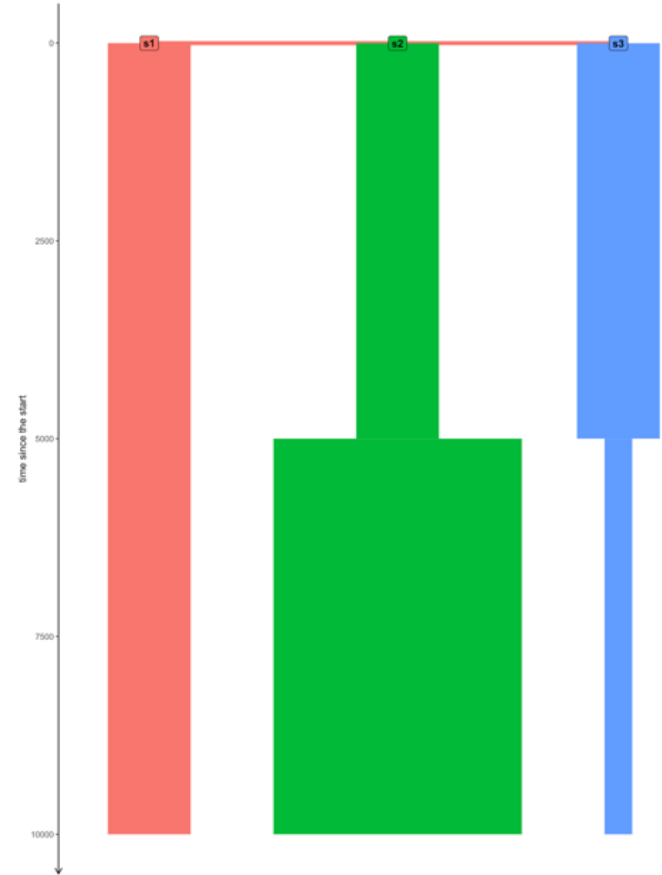
Demography

- Demography for a single population?
 - Effective population size



Why care about demography?

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



How to estimate demography?

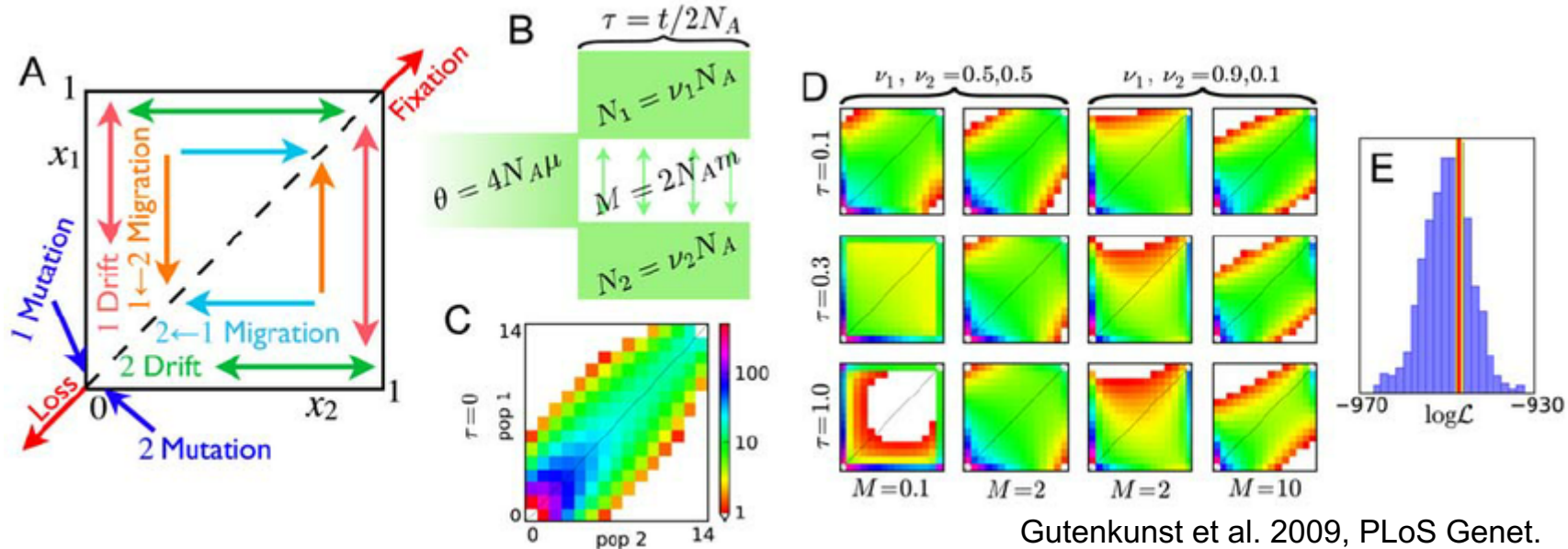
- Can you think of a way to estimate demography based on your morning exercises?

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

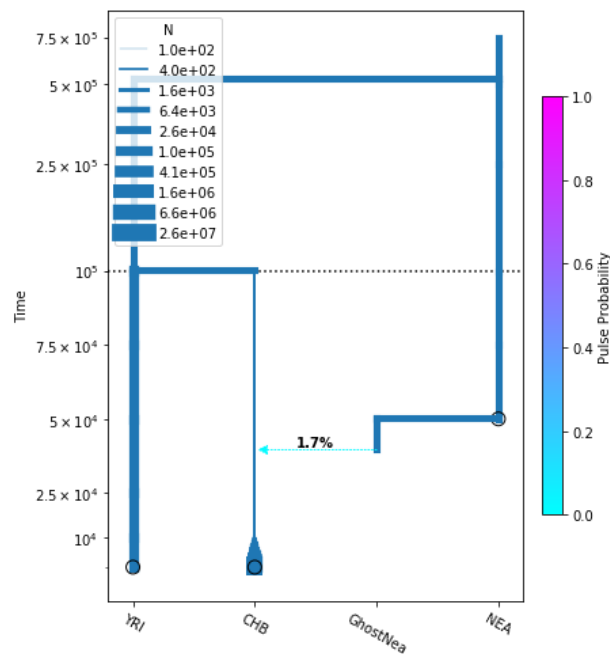
Estimating demography

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

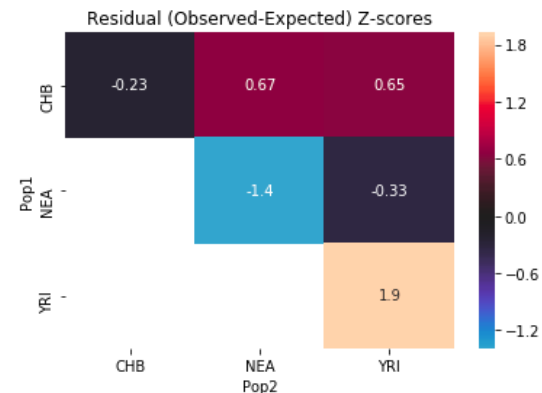


Estimating demography

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

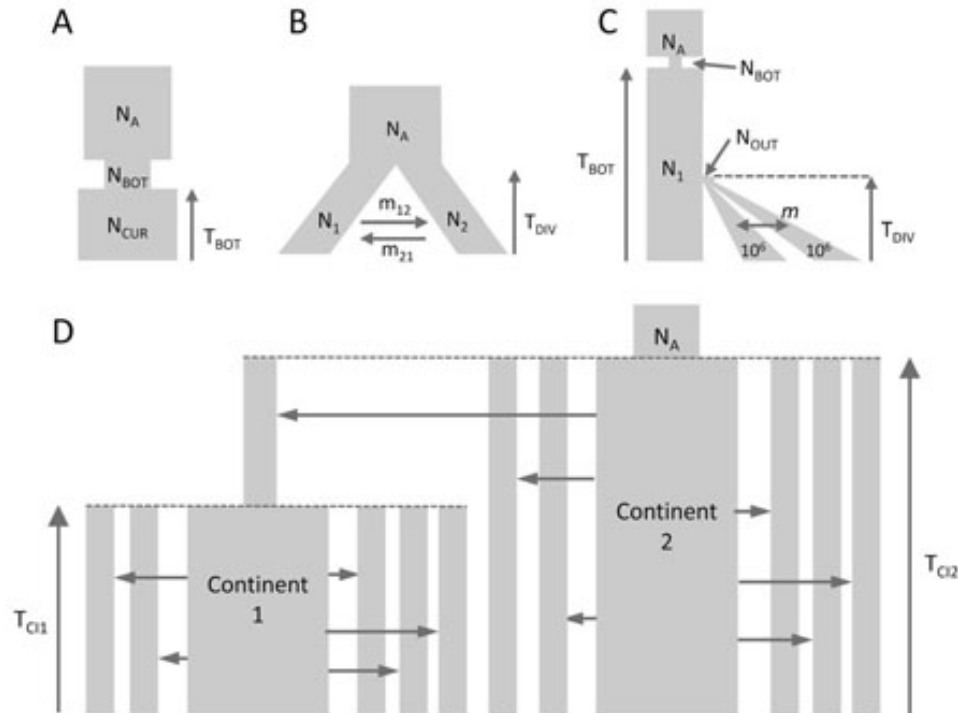


	Pop1	Pop2	Expected	Observed	Z
0	YRI	YRI	0.699637	0.706936	1.924912
1	NEA	NEA	0.732753	0.720787	-1.388584
2	CHB	NEA	0.545176	0.548234	0.668636
3	CHB	YRI	0.694800	0.697731	0.651155
4	NEA	YRI	0.545176	0.543831	-0.334112
5	CHB	CHB	0.965634	0.964595	-0.231728



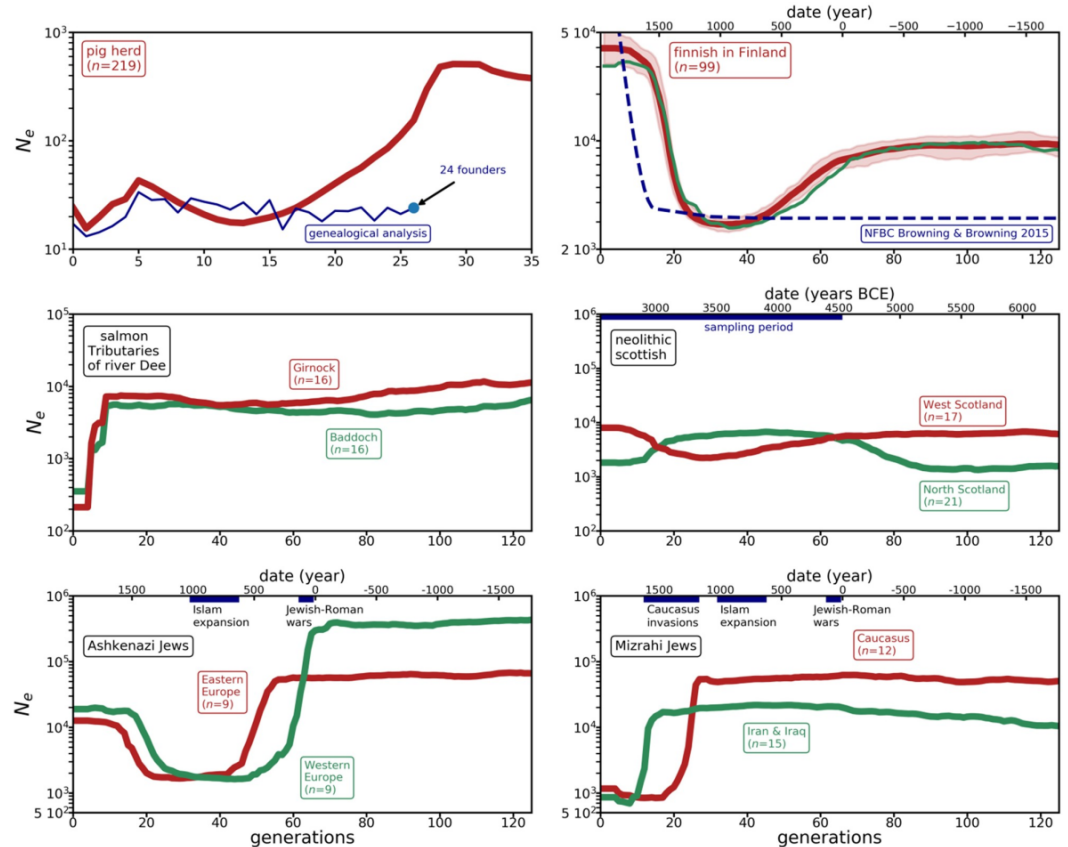
Estimating demography

- SFS based demography estimation
 - fastsimcoal2



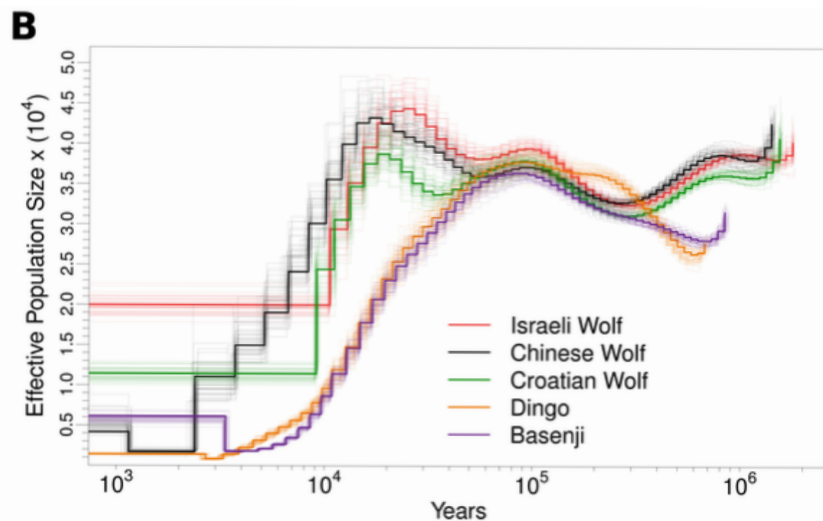
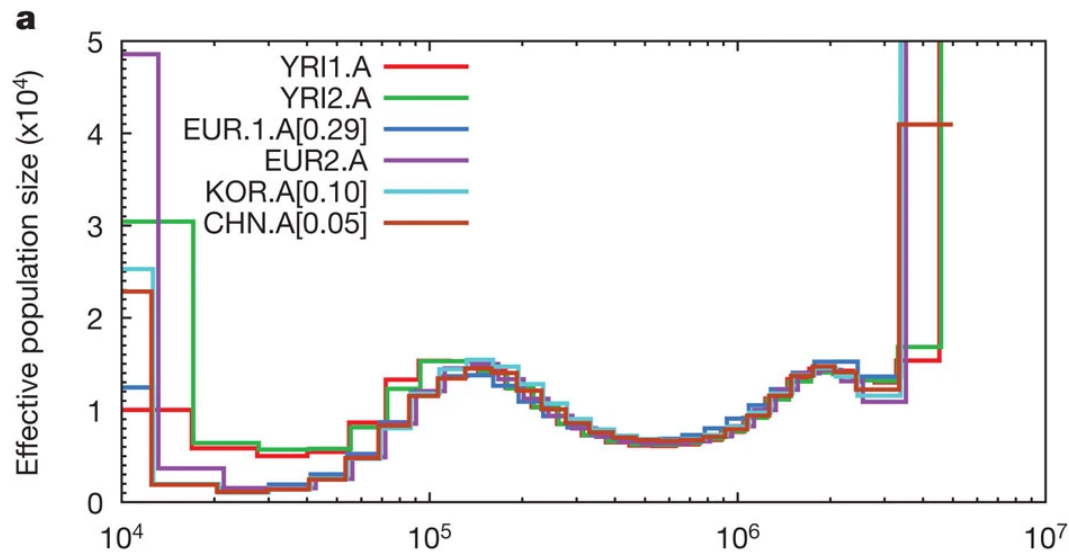
Estimating demography

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



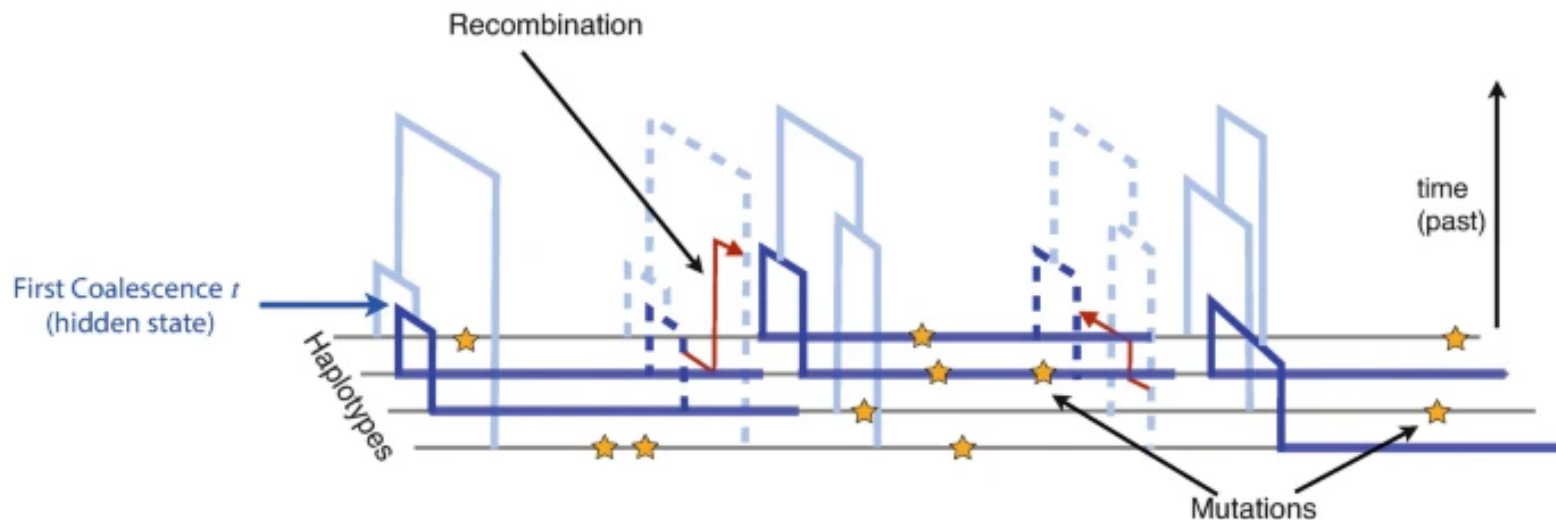
Estimating demography

- Coalescent based (*)
 - PSMC

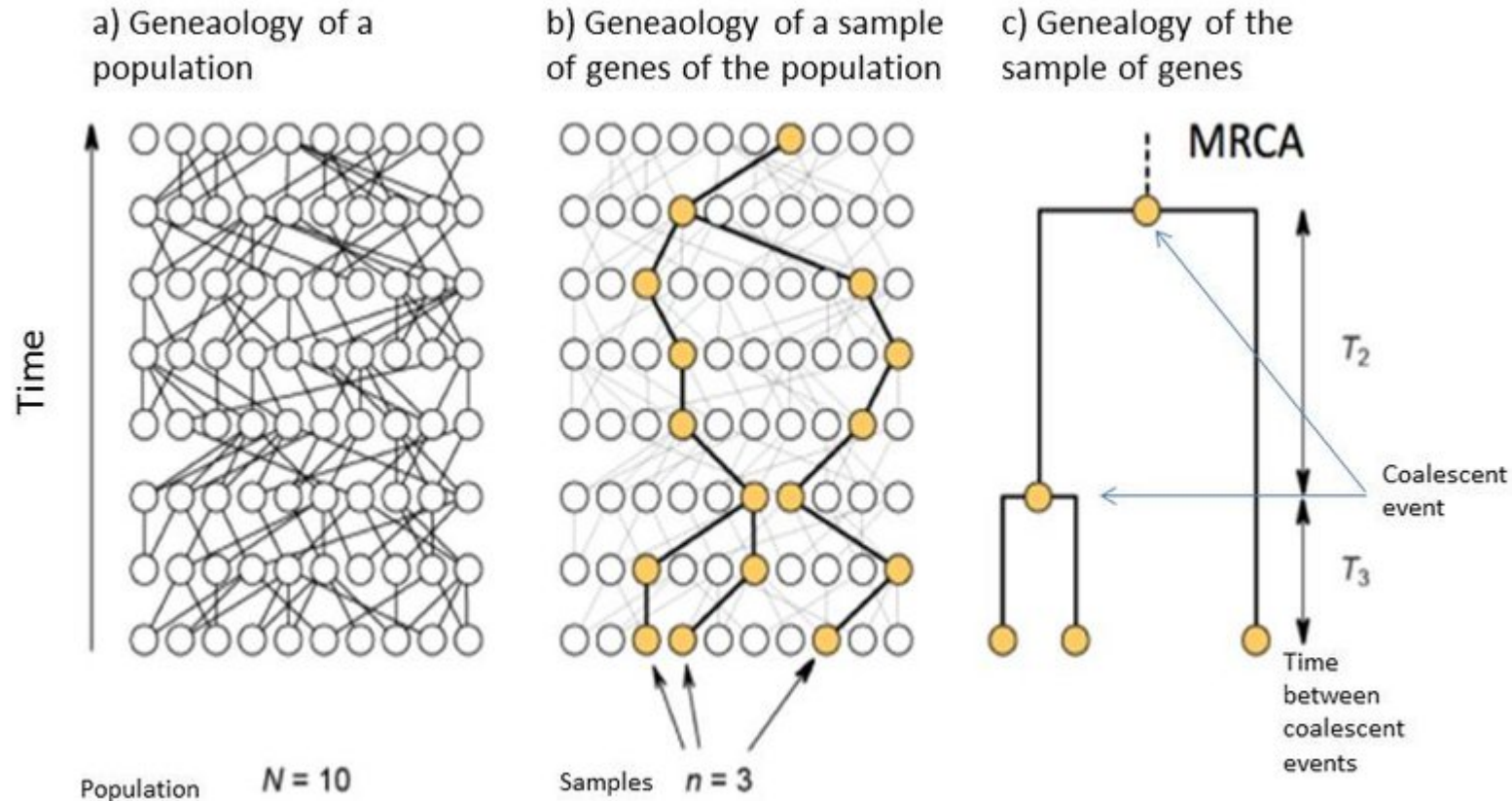


Estimating demography

- Coalescent based (*)
 - MSMC



Quick coalescent detour



Coalescent to demography

Group discussion

- Mutations in a pair of chromosomes
- Recombinations in a pair of chromosomes

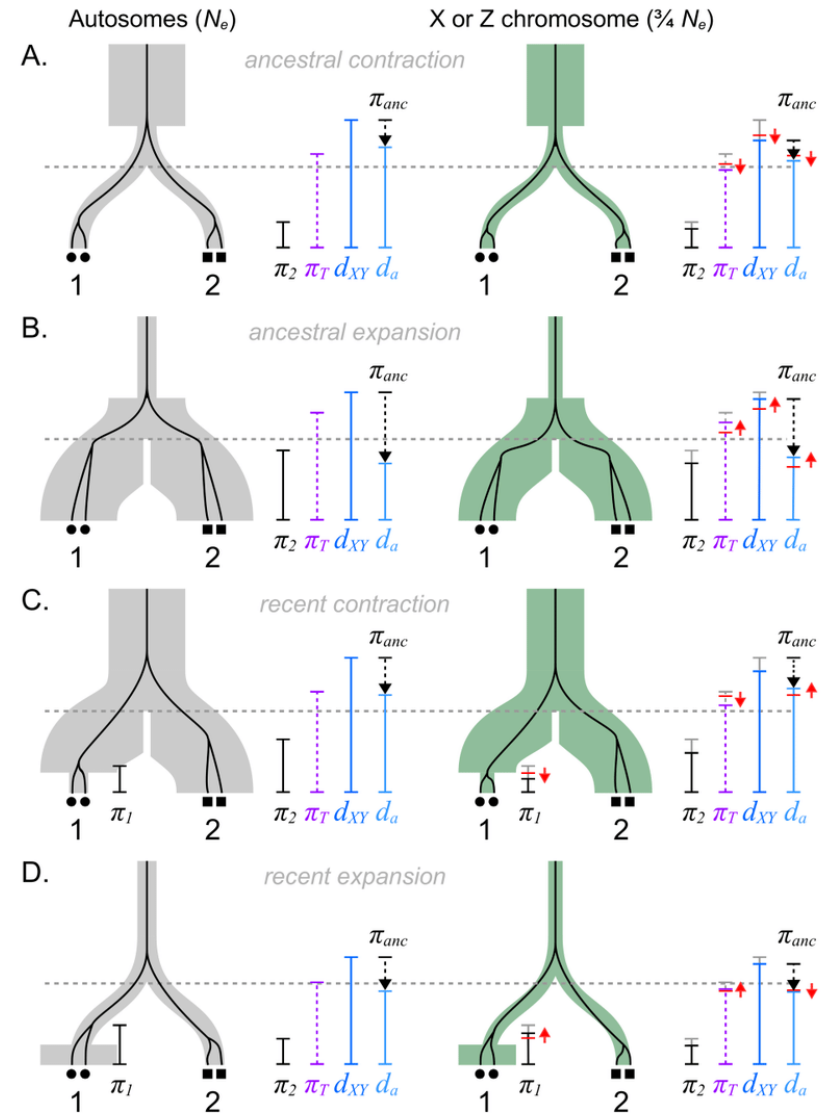
How would you use this information to estimate demography of one population?

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?



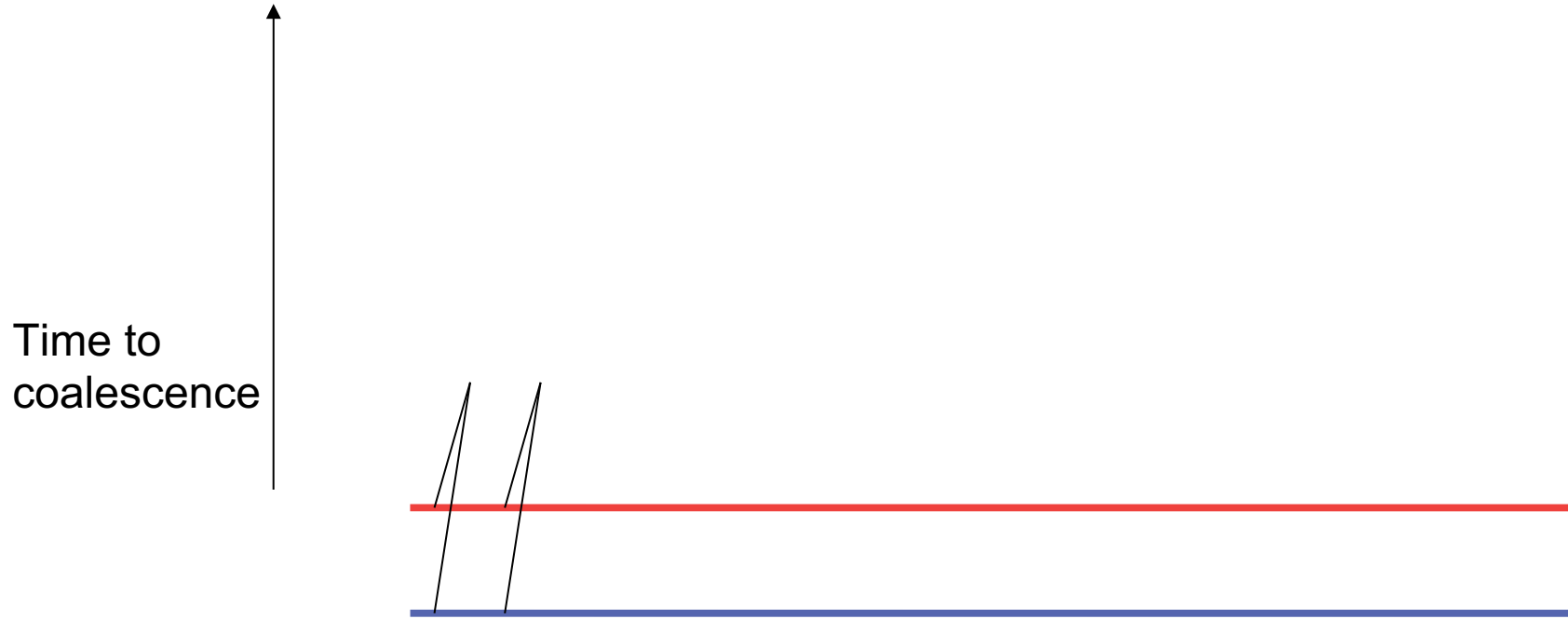
Information in one genome?



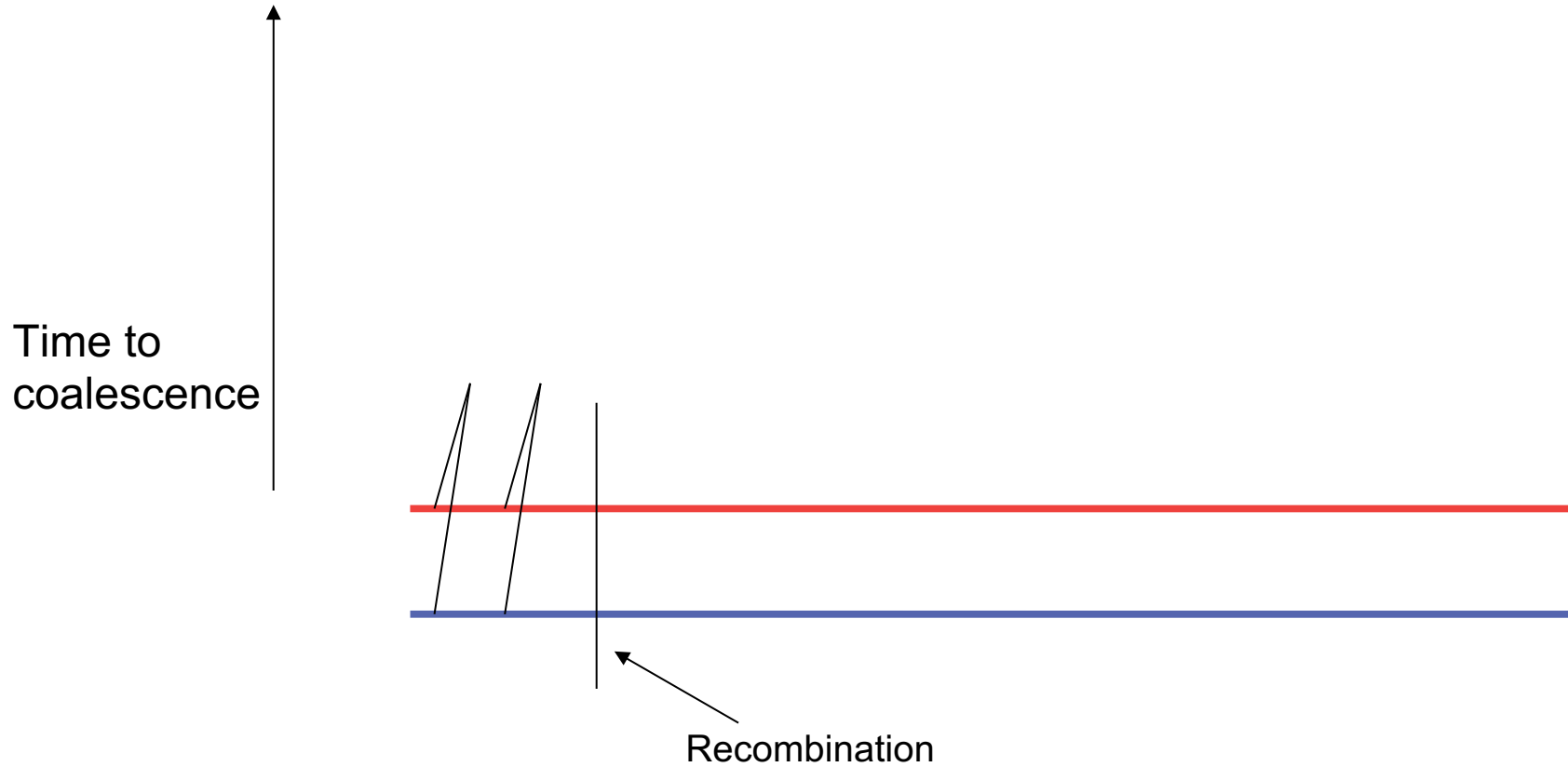
Information in one genome?



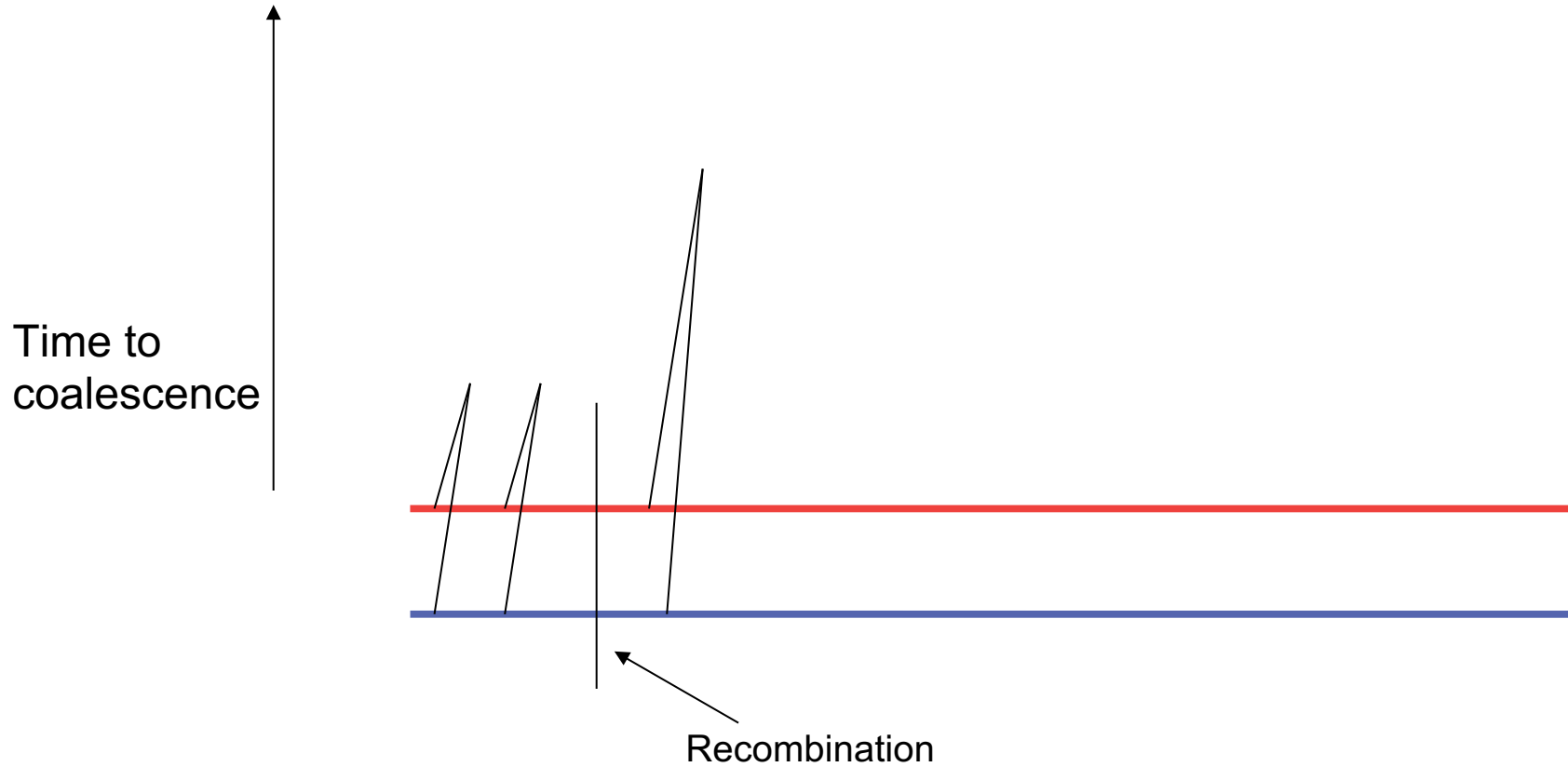
Information in one genome?



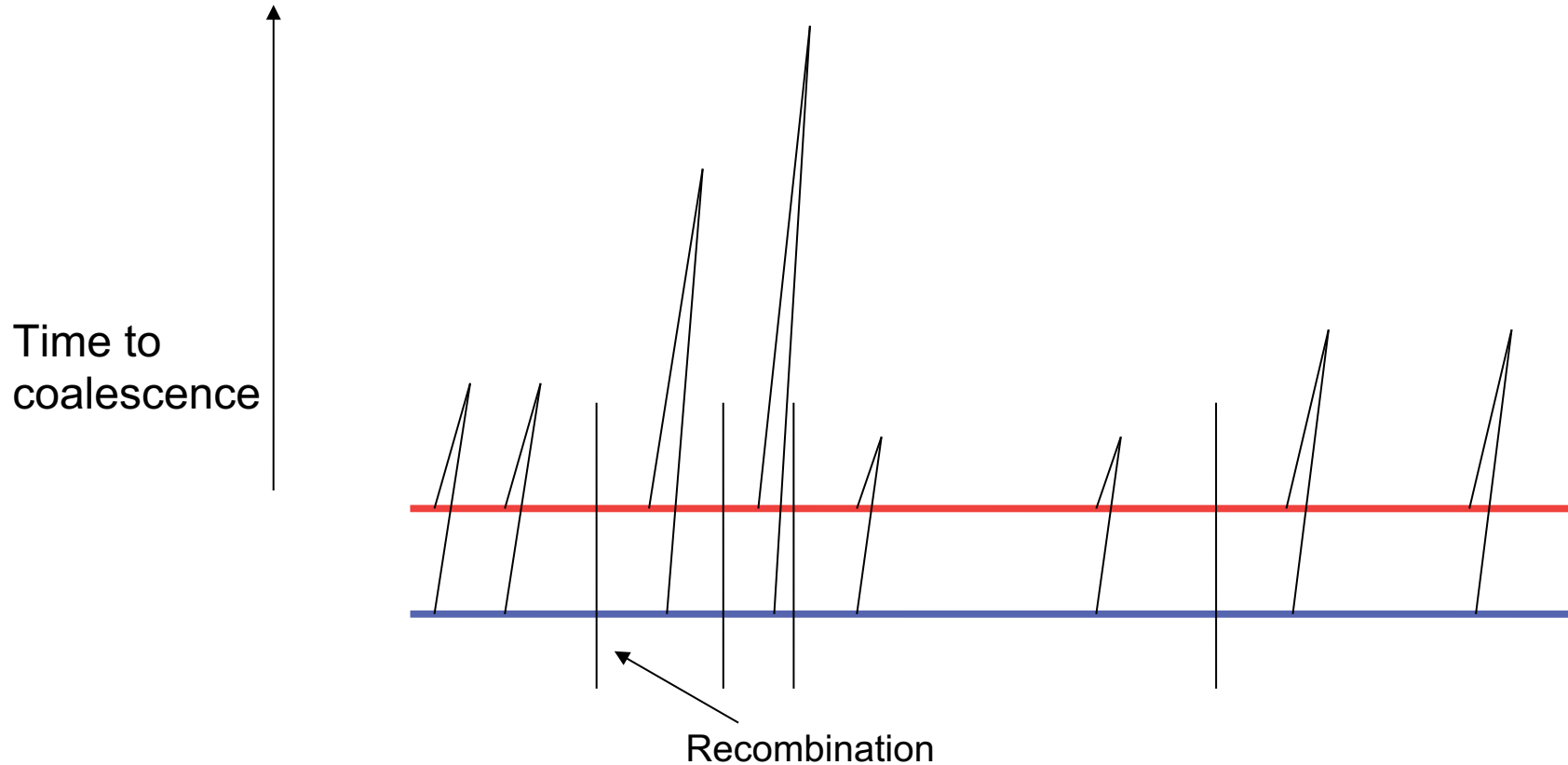
Information in one genome?



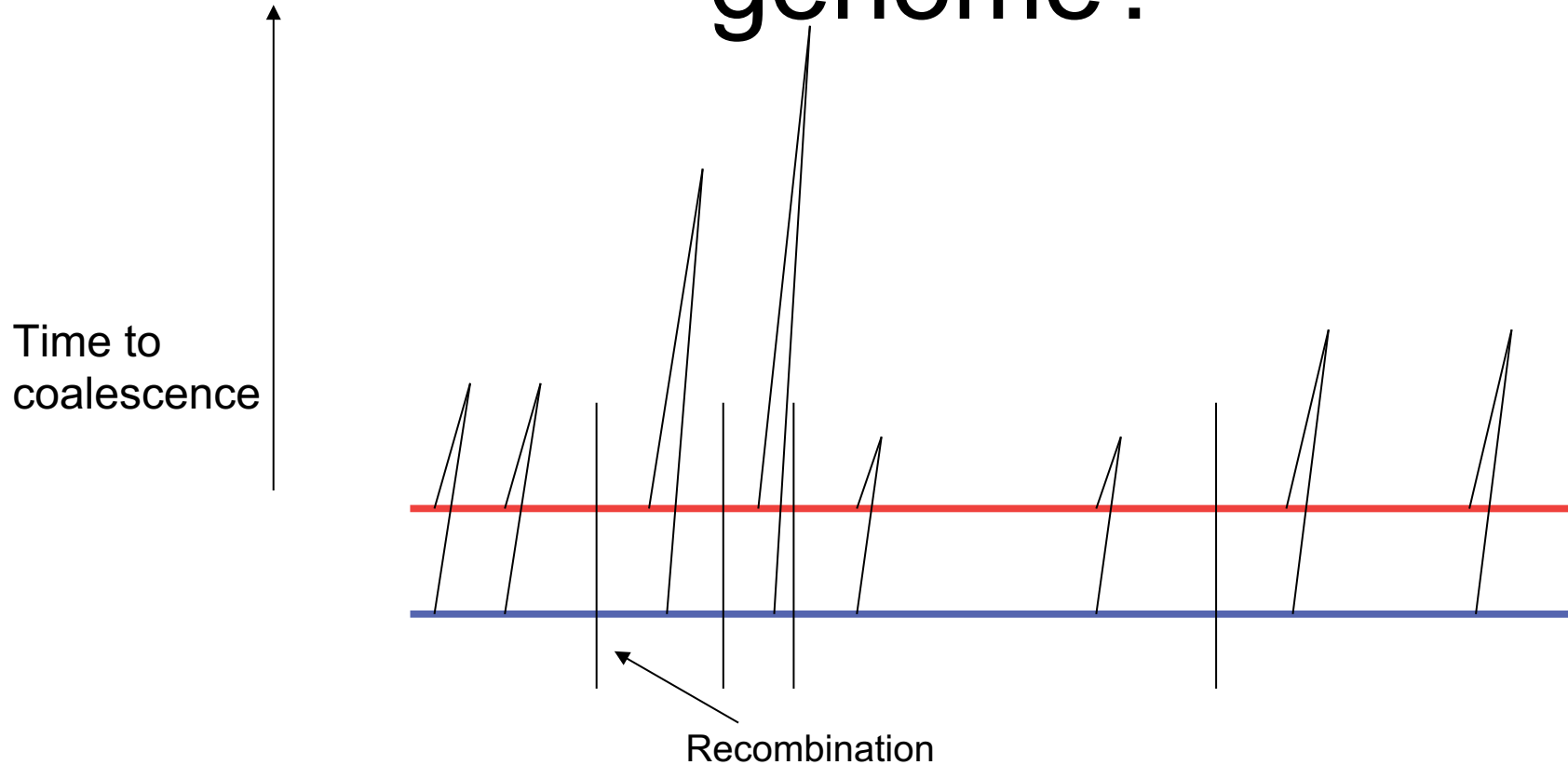
Information in one genome?



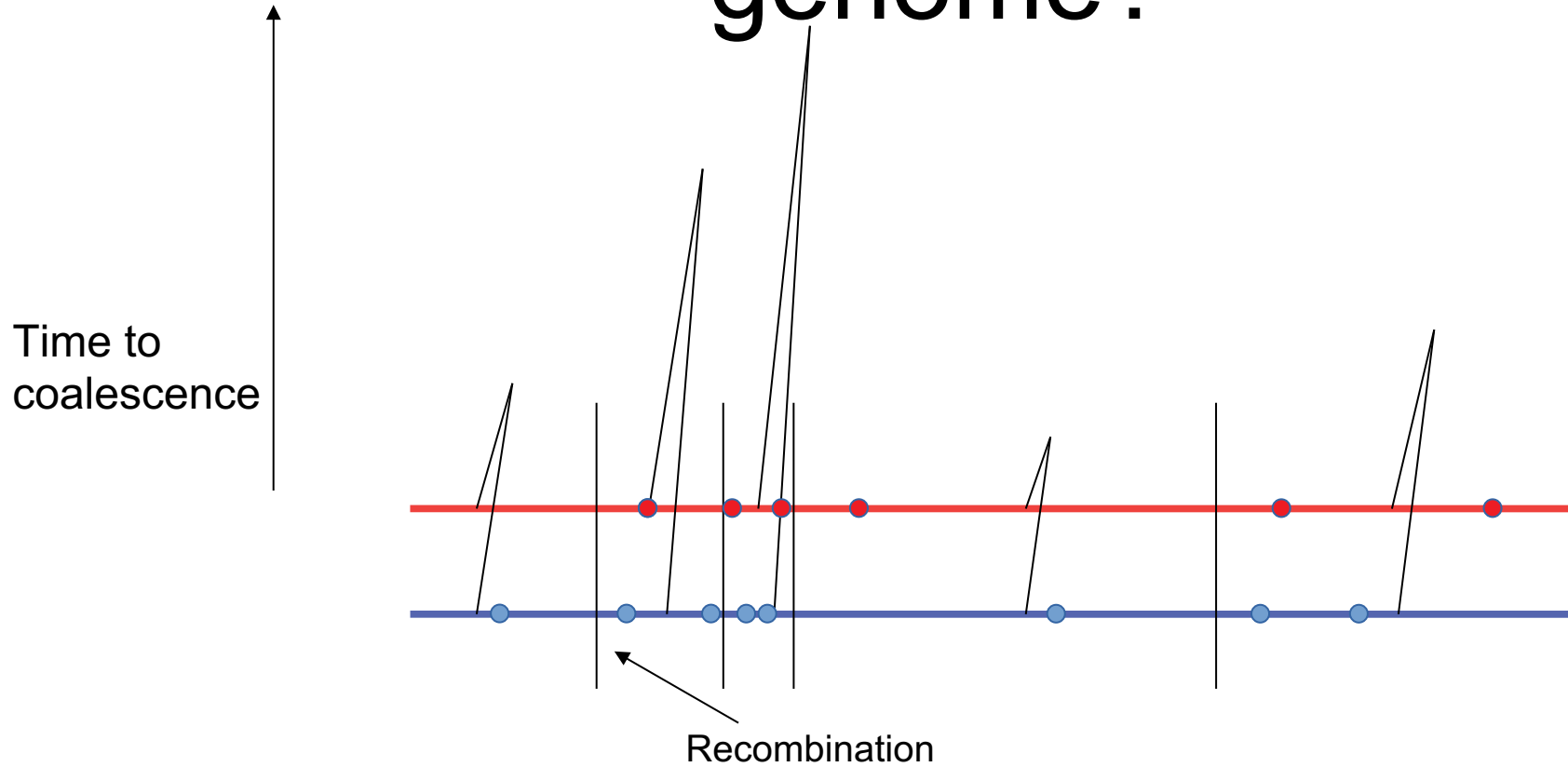
Information in one genome?



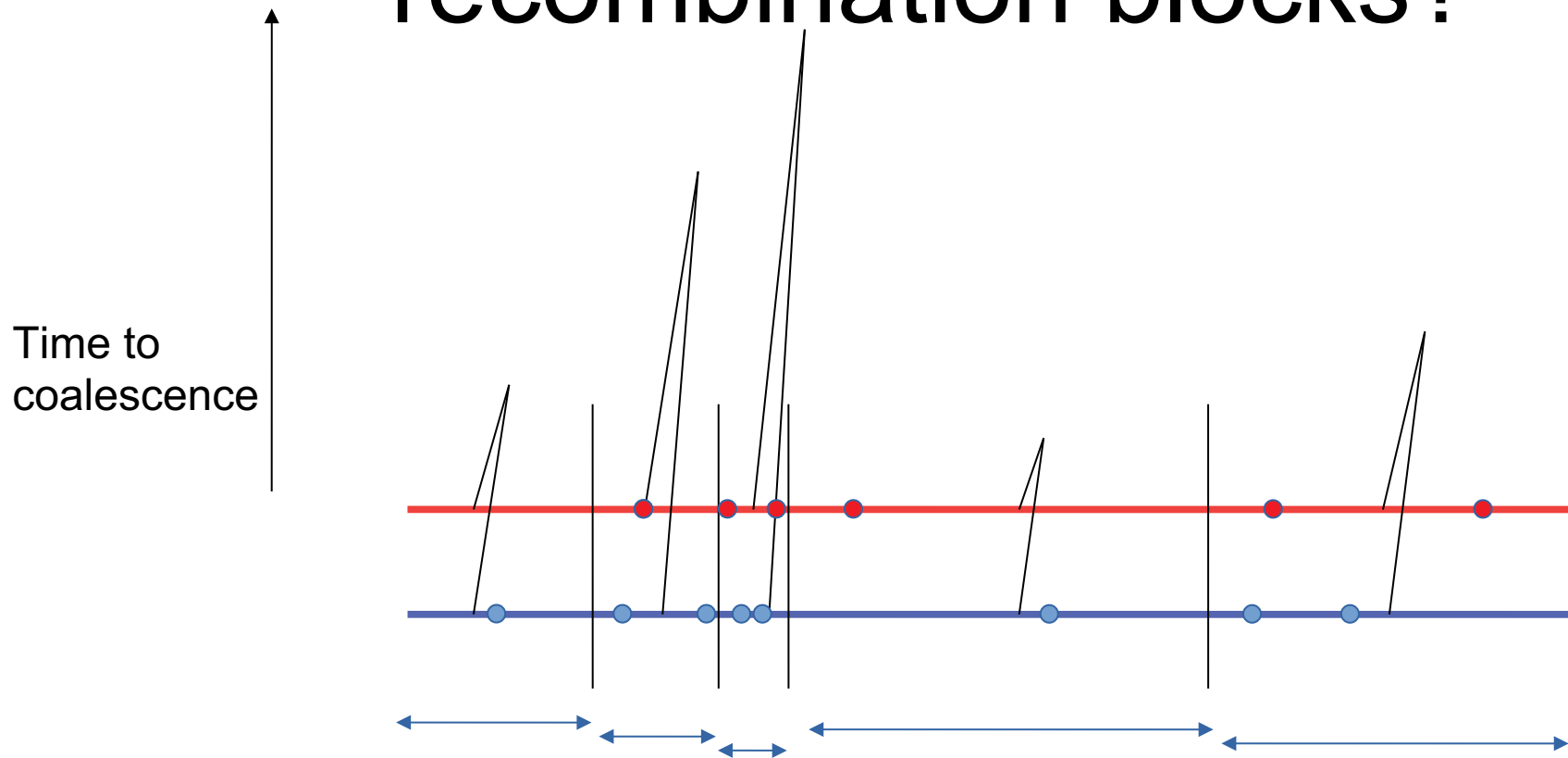
How will the mutations look on this genome?



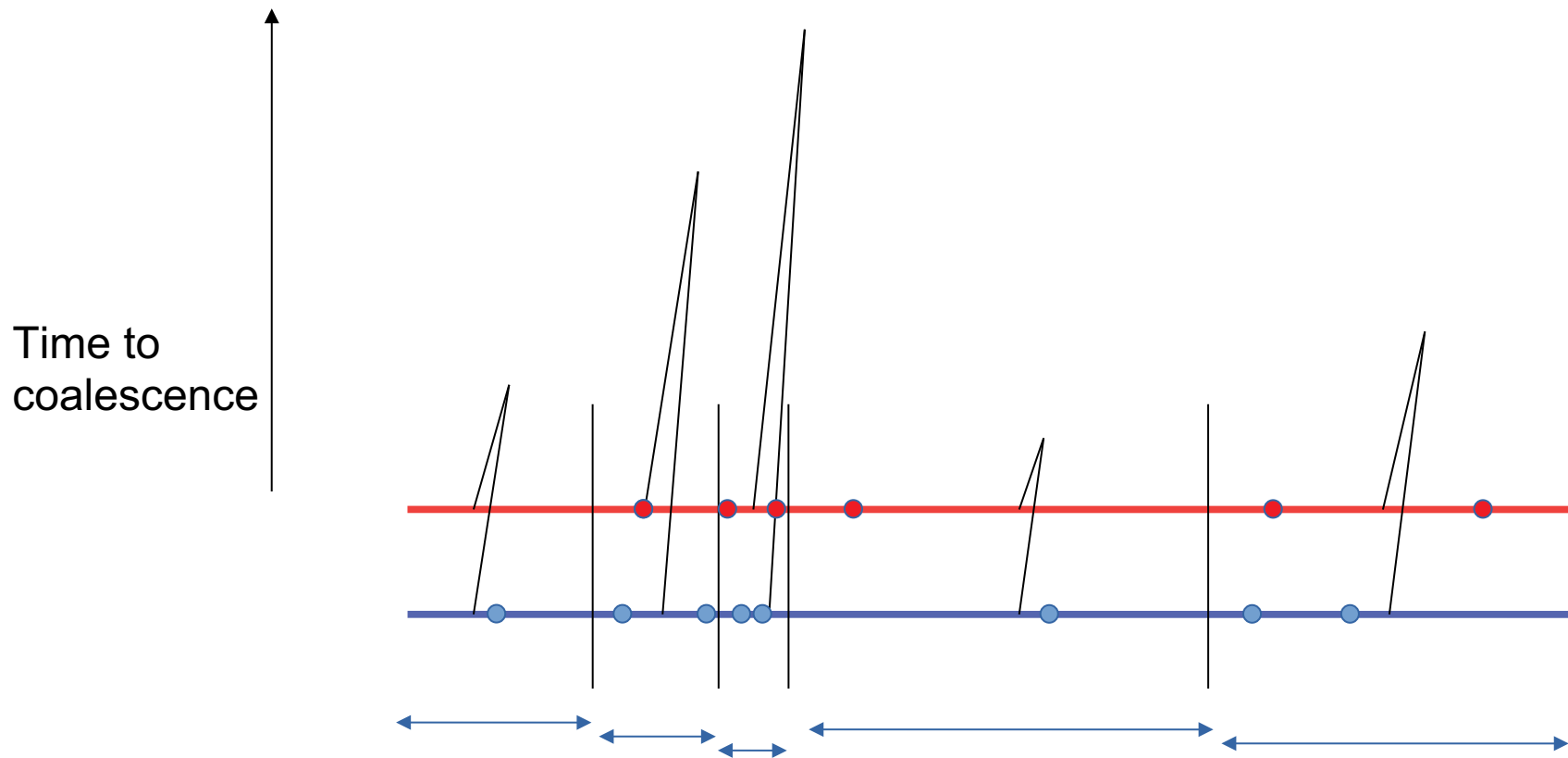
How will the mutations look on this genome?



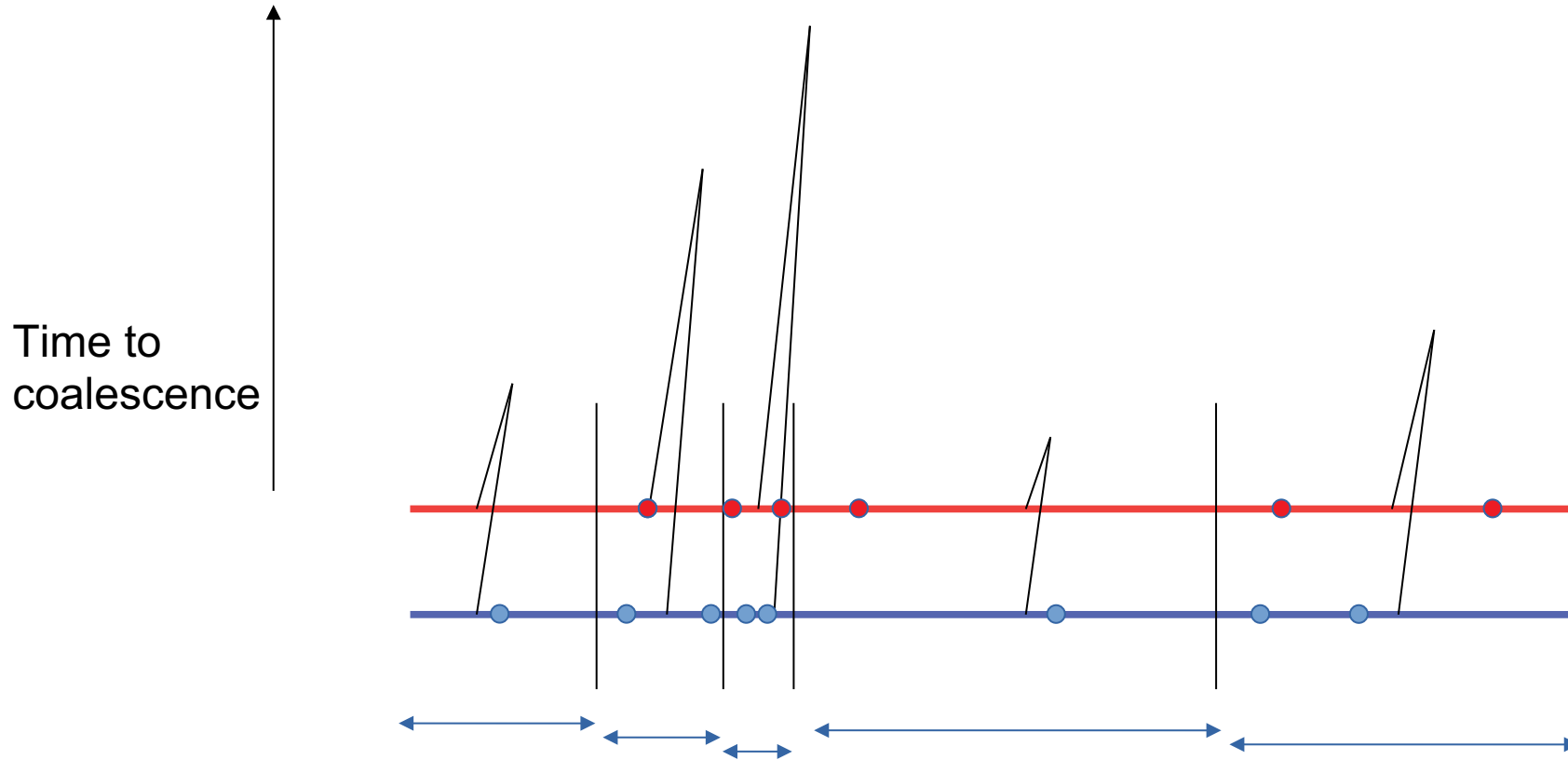
What about the size of the recombination blocks?



Reconstruct population history



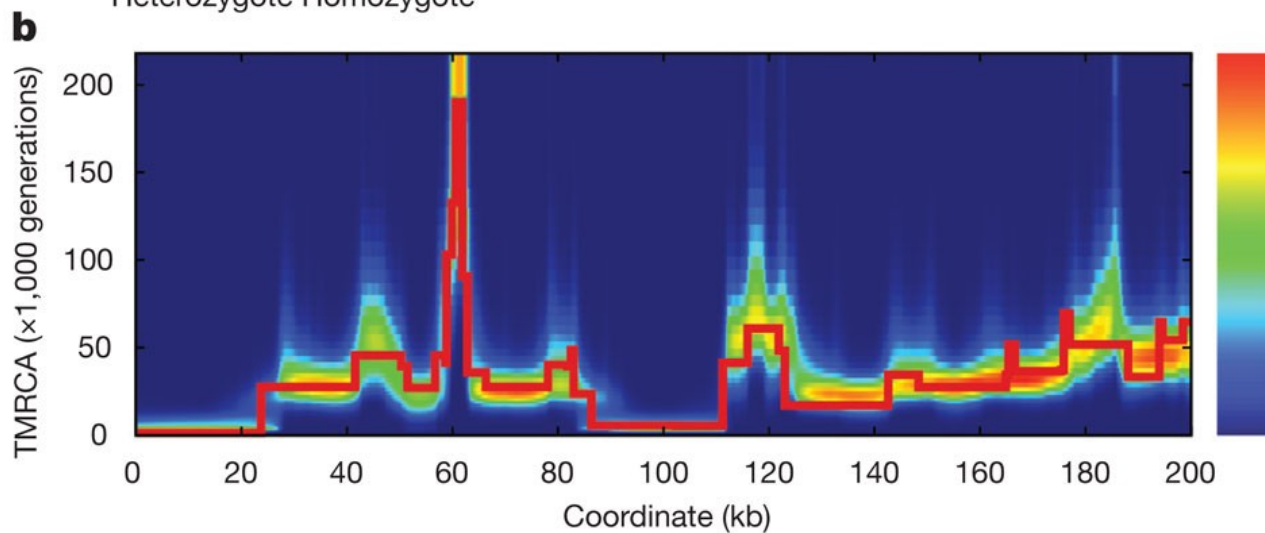
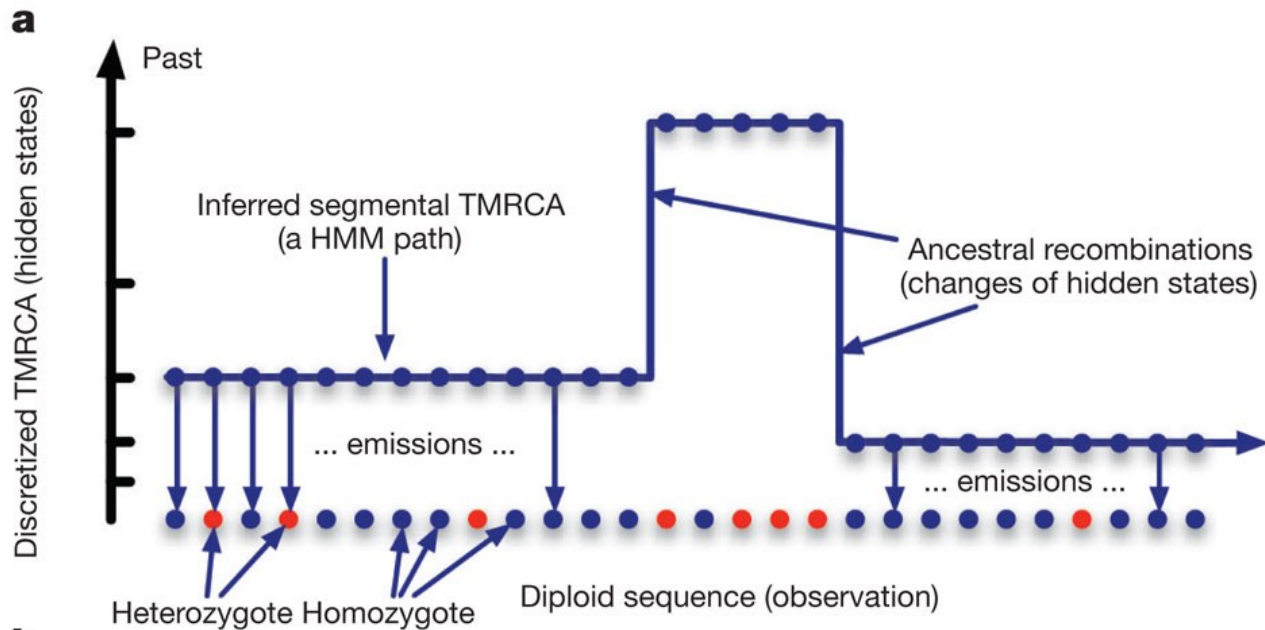
Reconstruct population history



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



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.

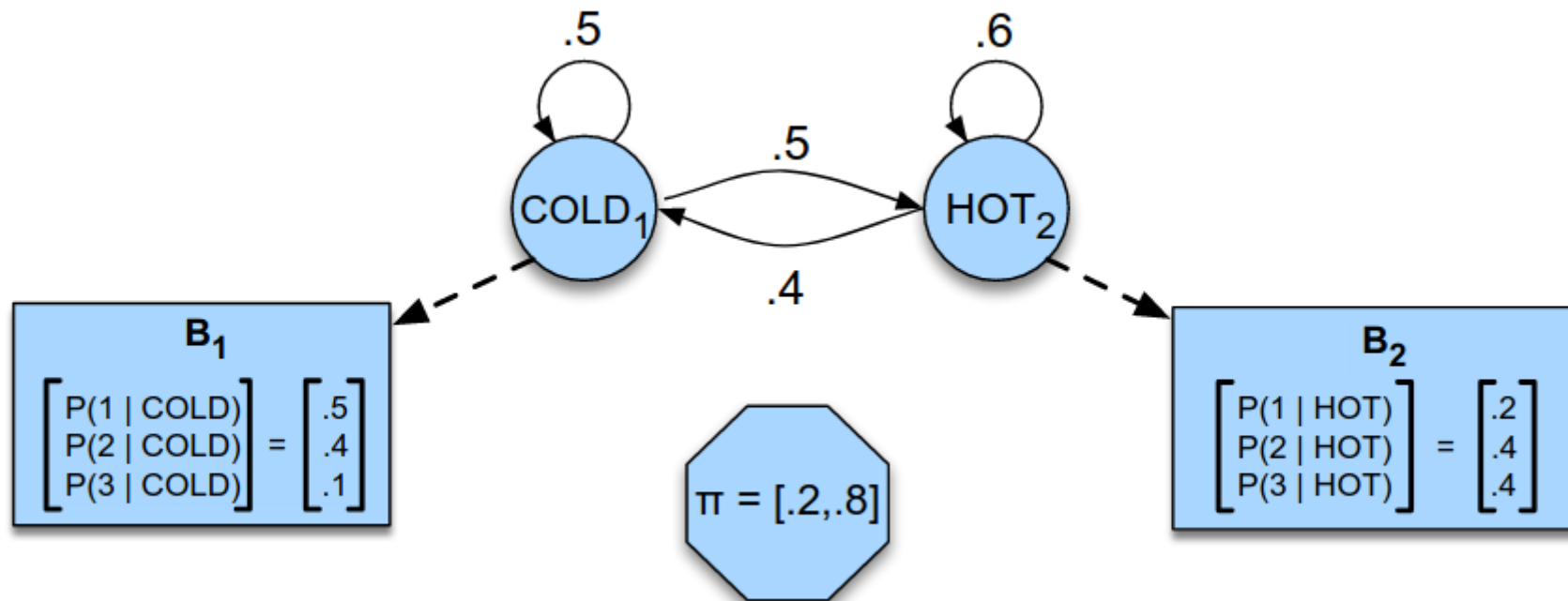
Some simplifications + assumptions:

Only 2 day types - HOT or COLD

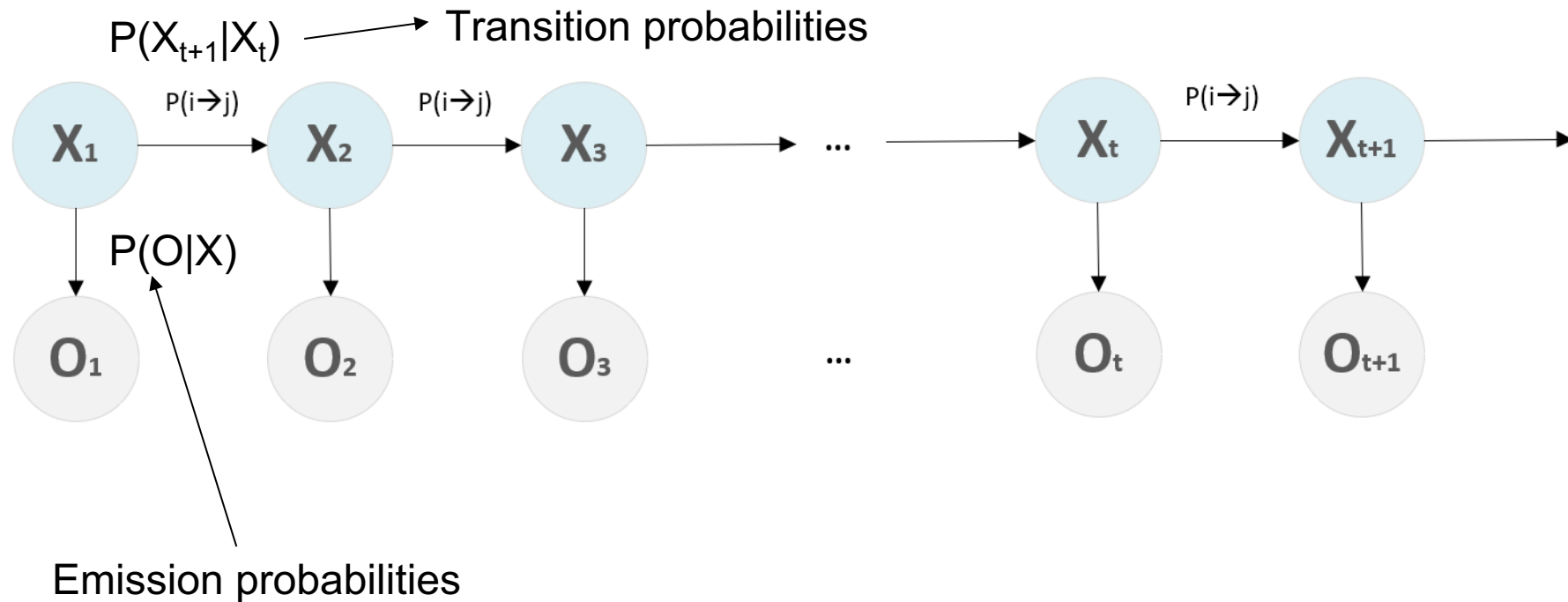
Eisner's ice cream consumption depends on day type

Day states form a markov chain – so today's day type affects tomorrow's day type.

Quick detour #2: HMM



Quick detour #2: HMM

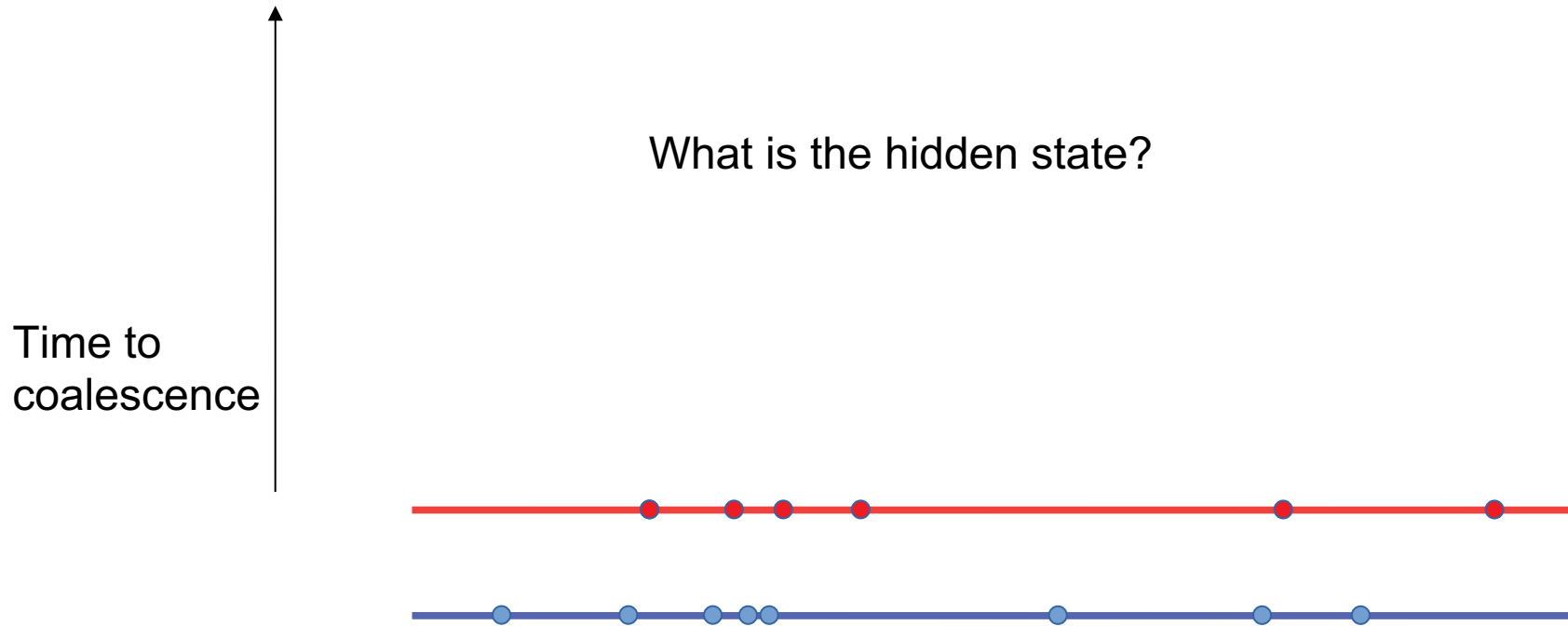


Quick detour #2: HMM

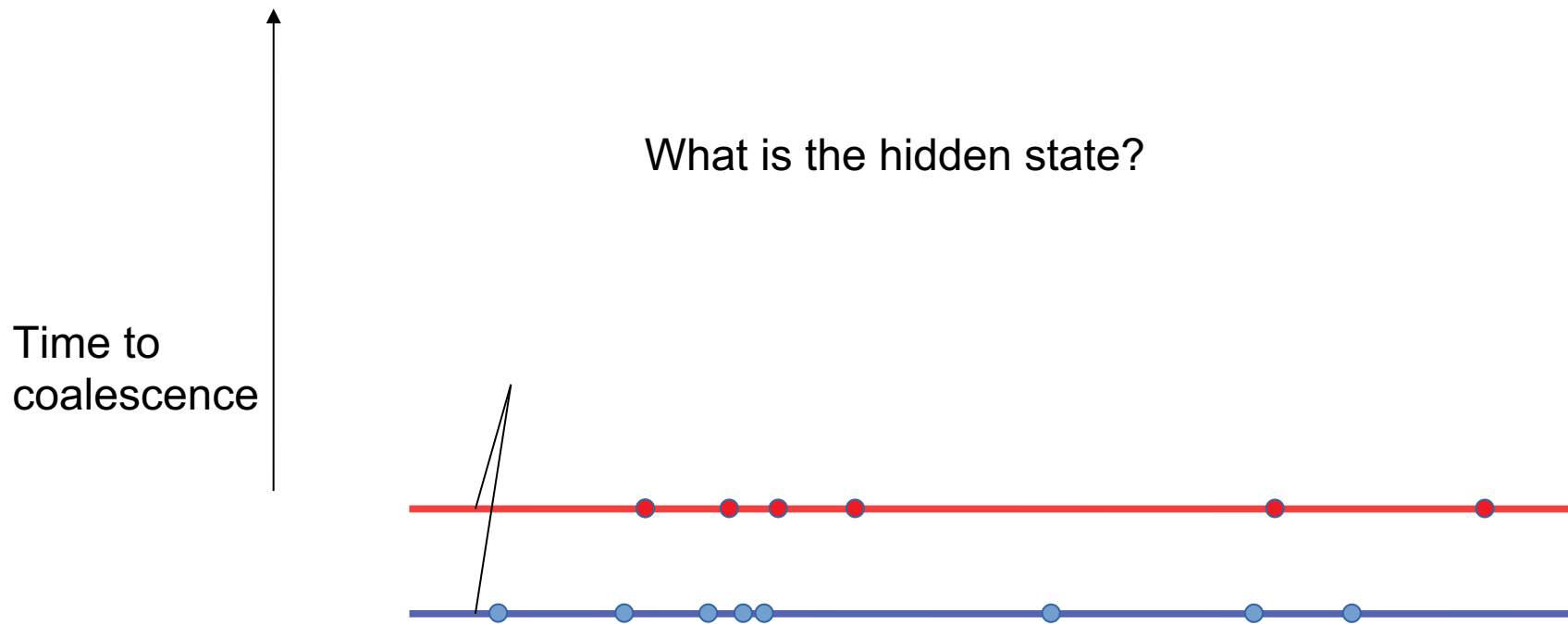
Three problems we need to solve in a HMM:

1. **Likelihood**: Given observations O , and parameters Θ of the HMM, we need to be able to compute the $P(O \mid \Theta)$
2. **Decoding**: Given observations O , and parameters Θ , compute the best hidden state sequence $X_1, X_2 \dots X_n$
3. **Learning**: Given observations O and the set of states in the HMM, learn the parameters Θ .

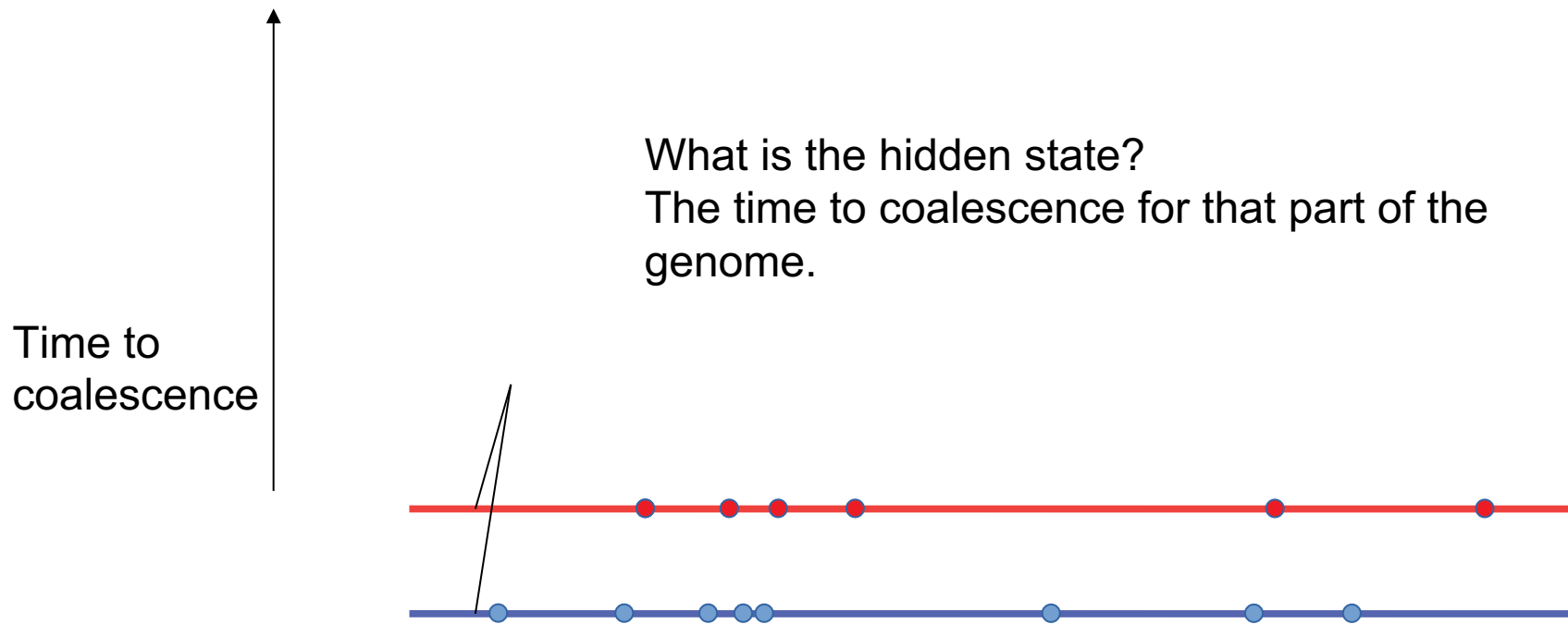
PSMC: HMM for genome



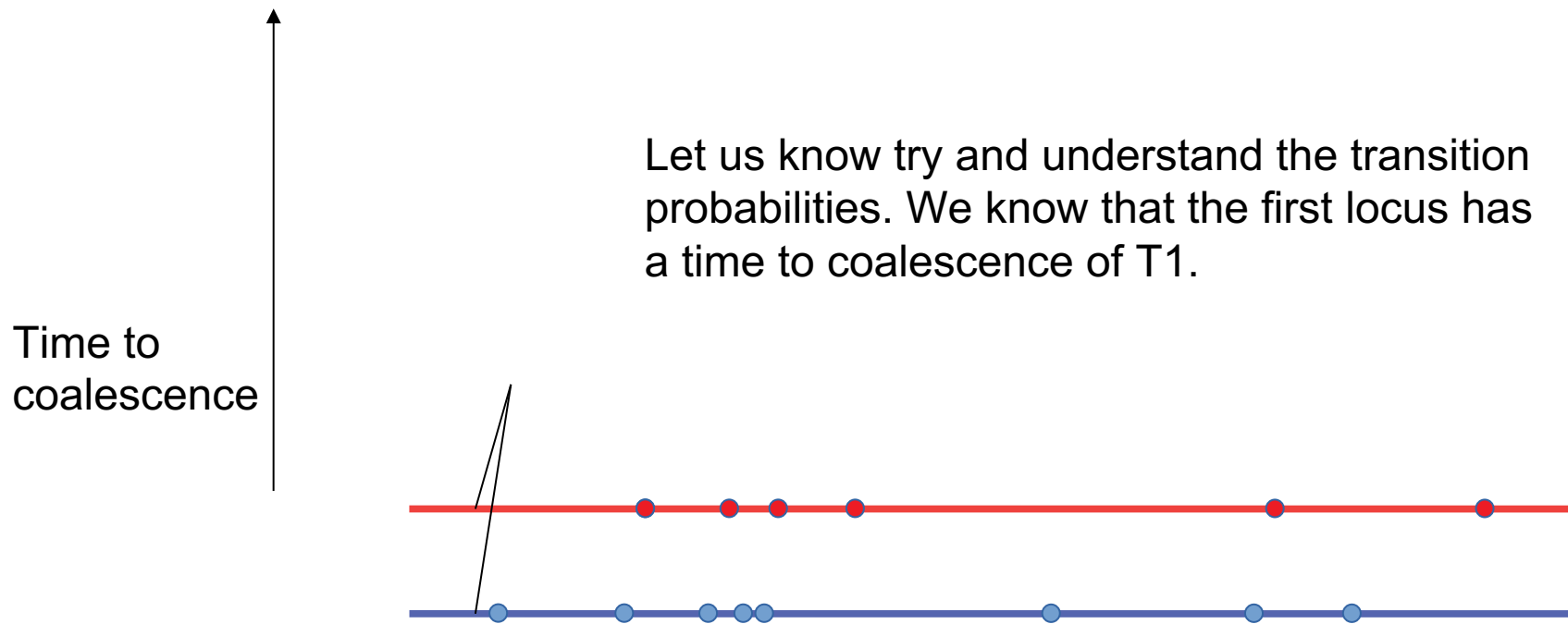
PSMC: HMM for genome



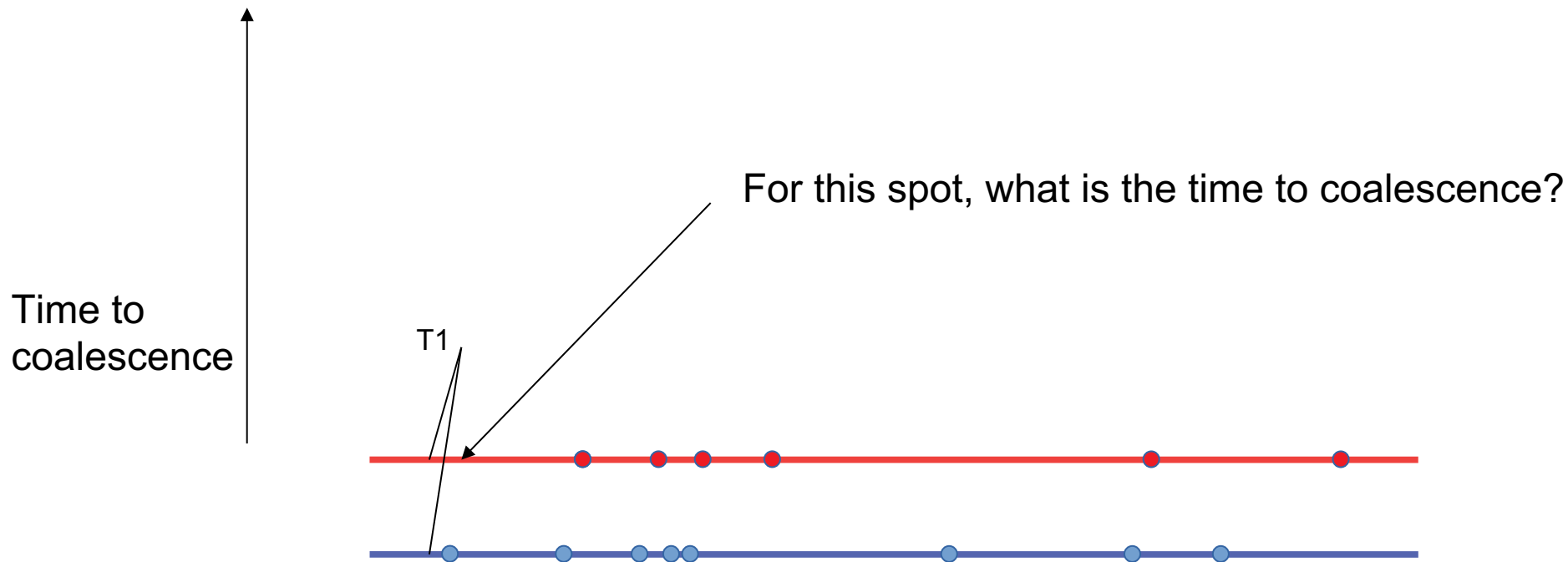
PSMC: HMM for genome



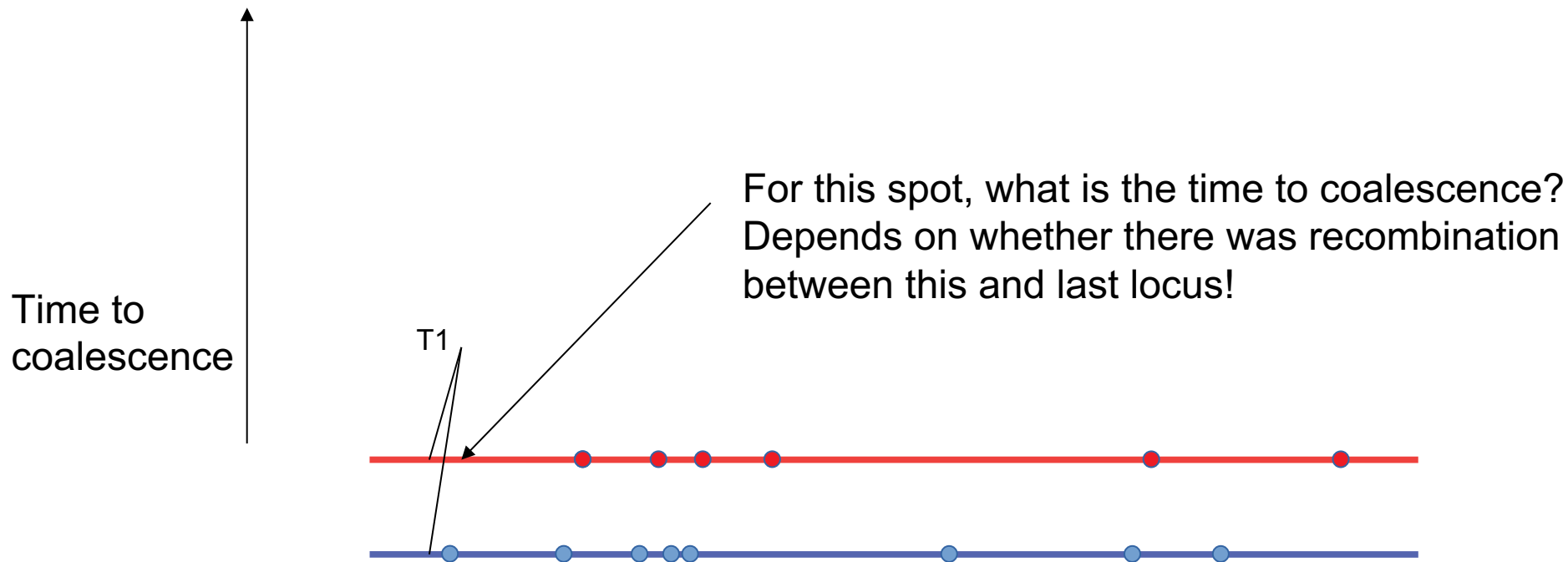
PSMC: HMM for genome



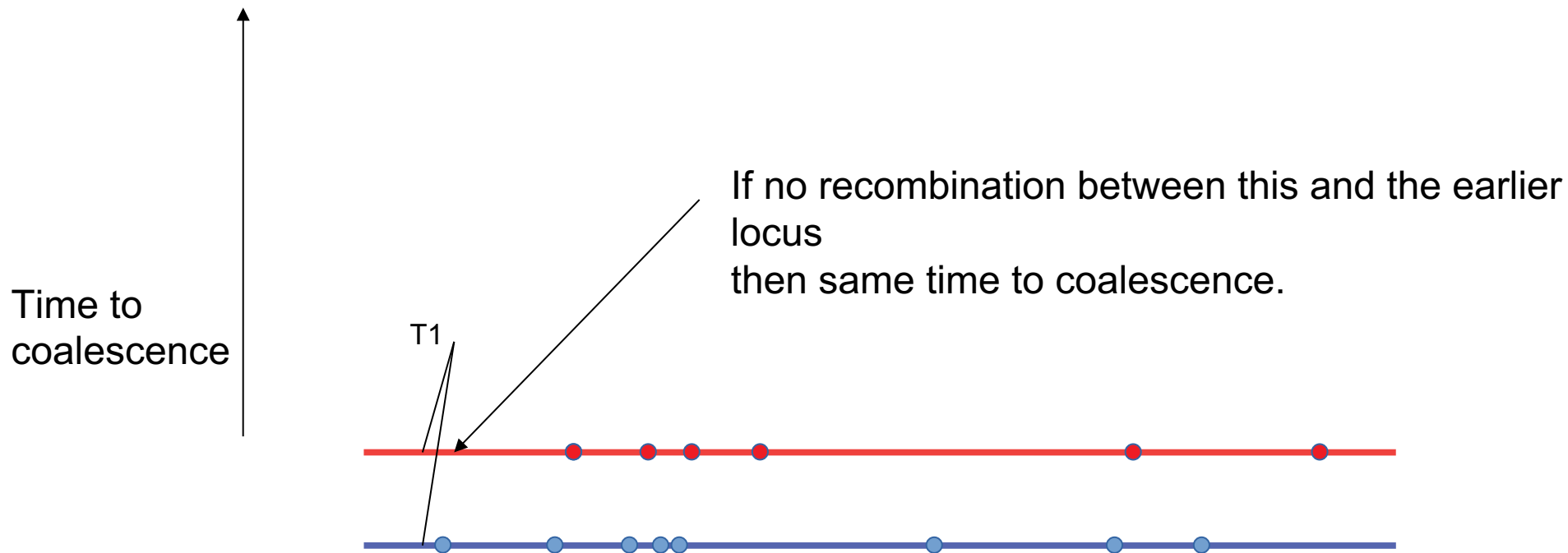
PSMC: HMM for genome



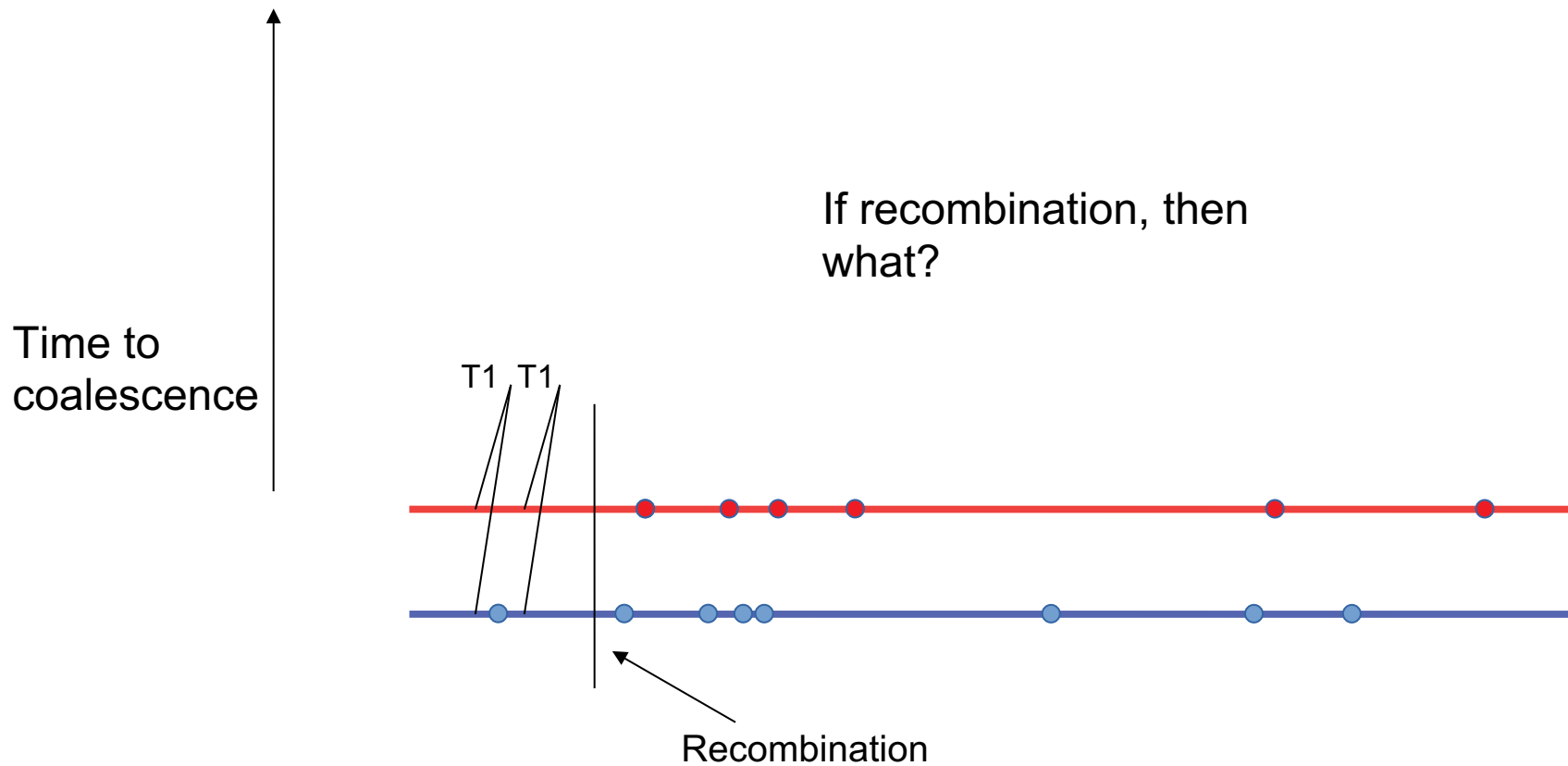
PSMC: HMM for genome



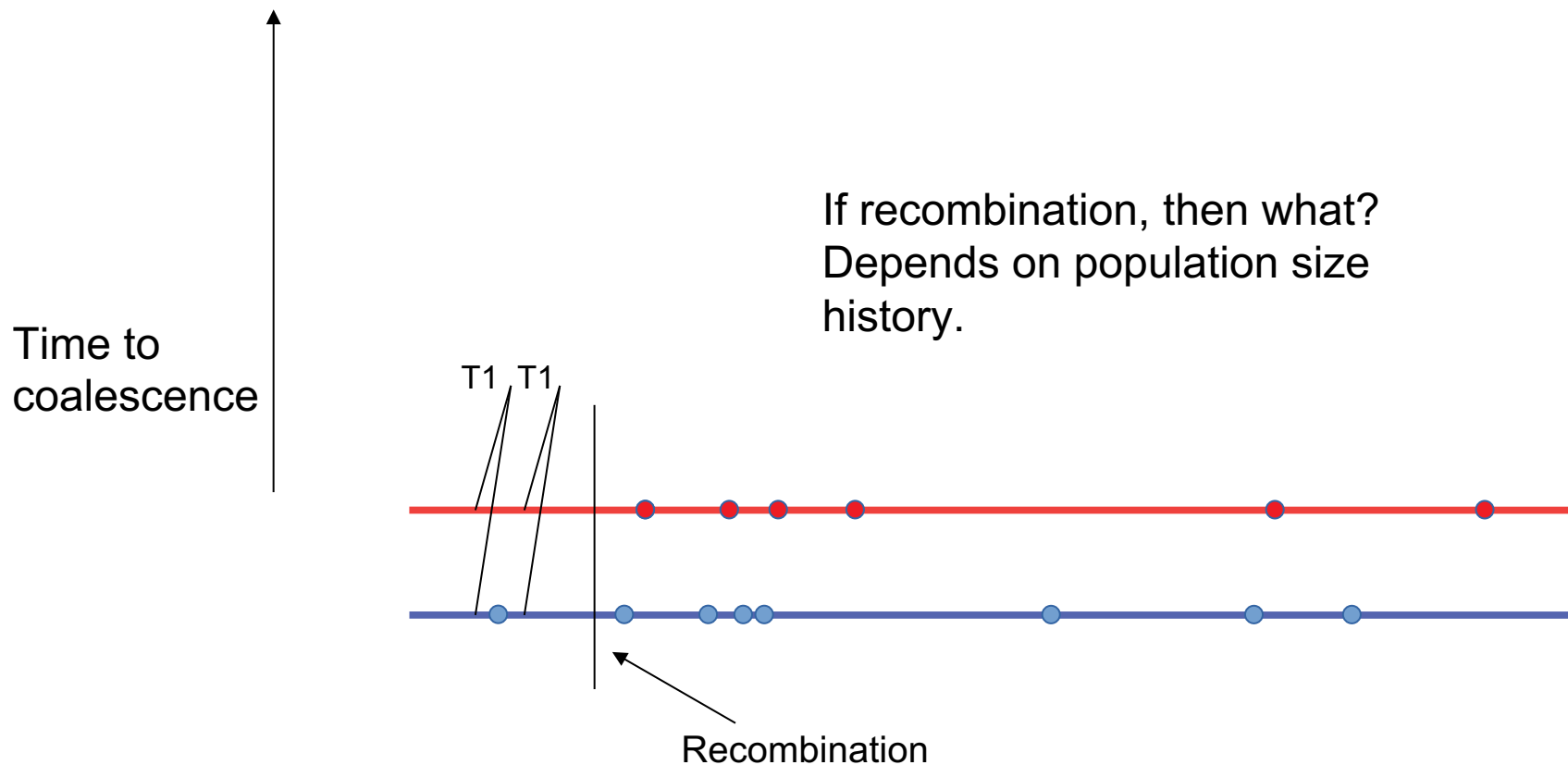
PSMC: HMM for genome



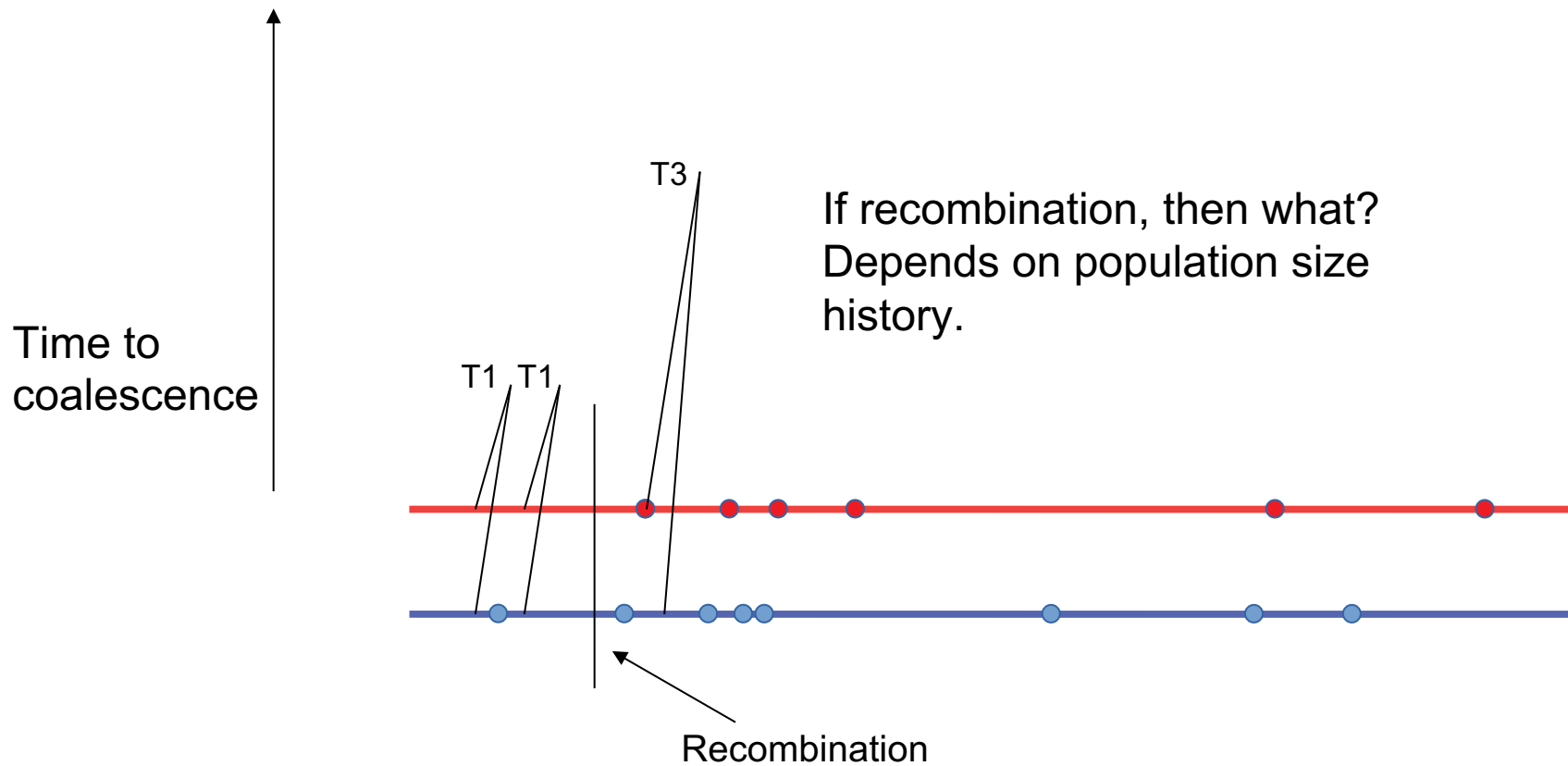
PSMC: HMM for genome



PSMC: HMM for genome

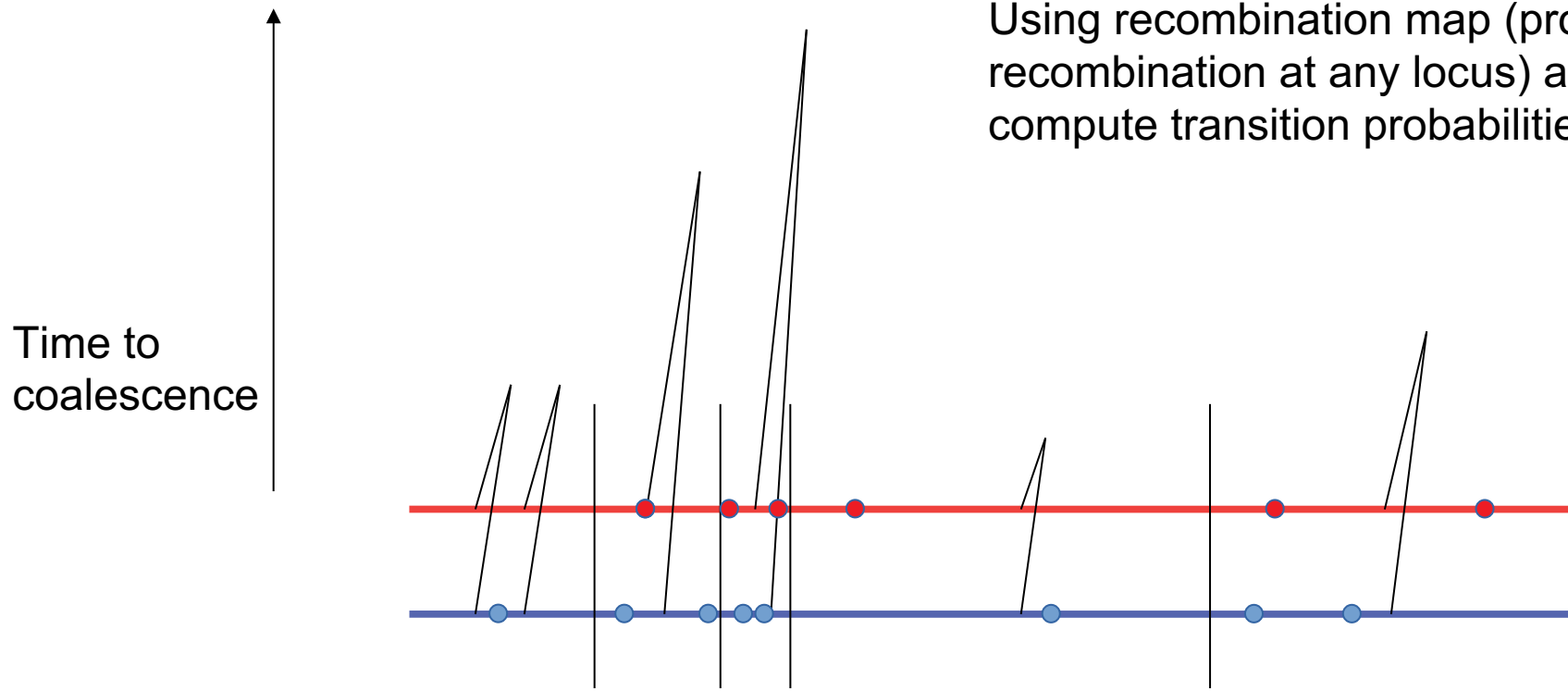


PSMC: HMM for genome

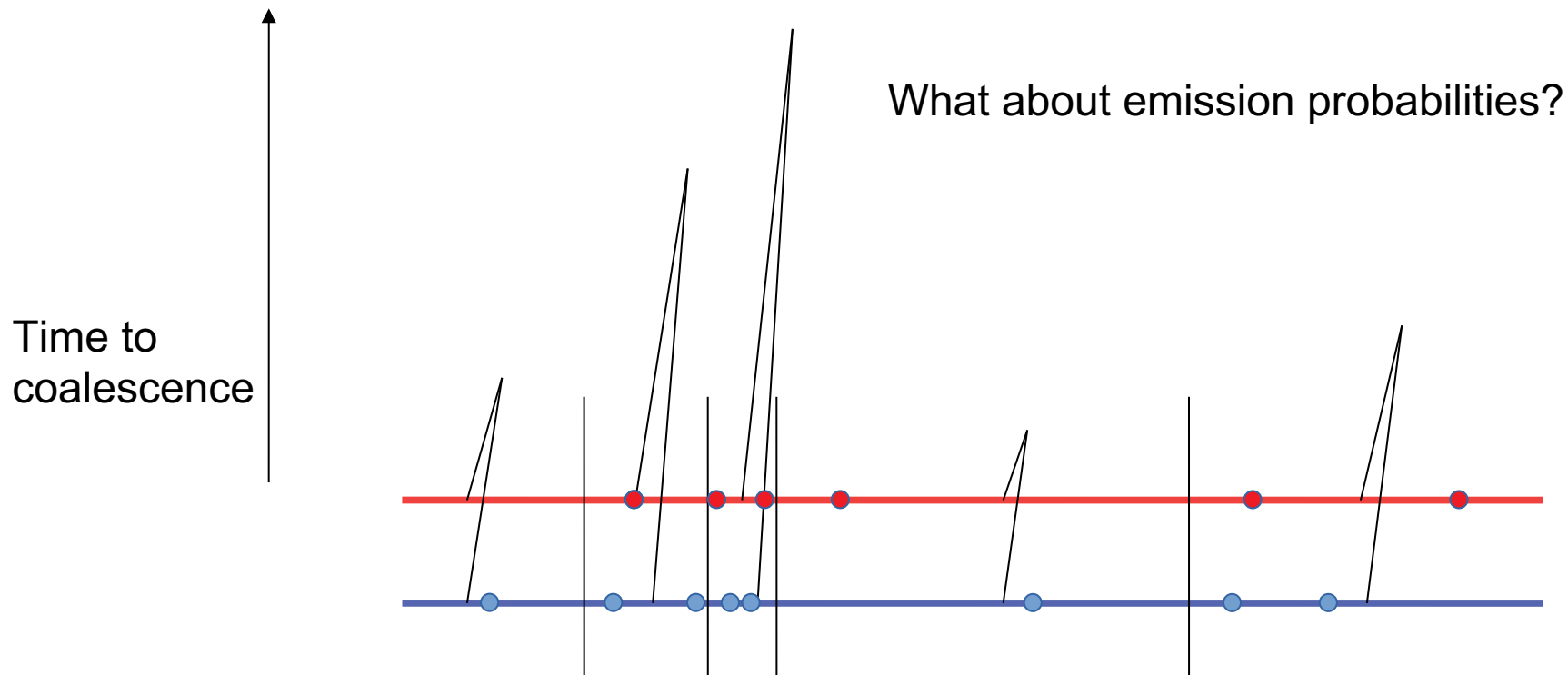


PSMC: HMM for genome

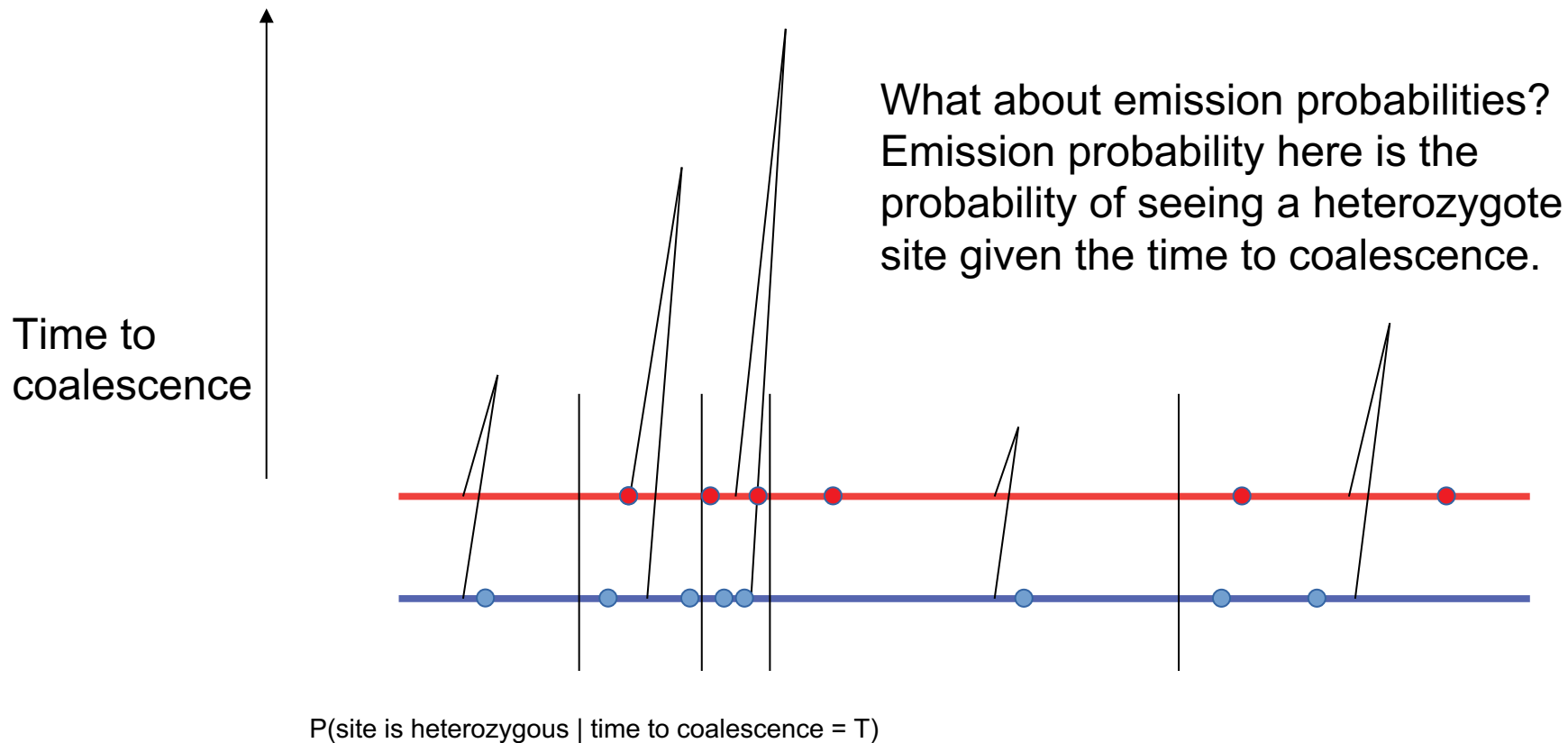
Using recombination map (probability of recombination at any locus) and N_e , we can compute transition probabilities.



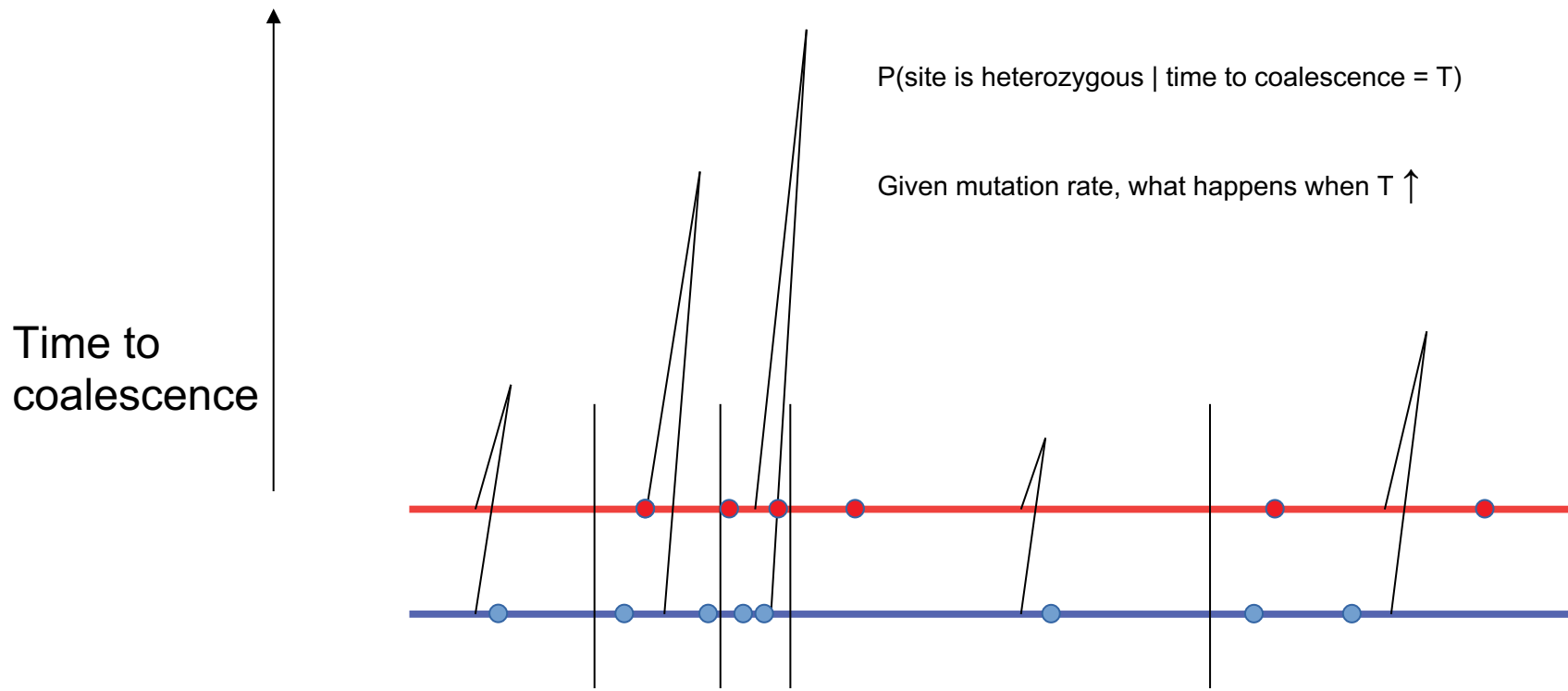
PSMC: HMM for genome



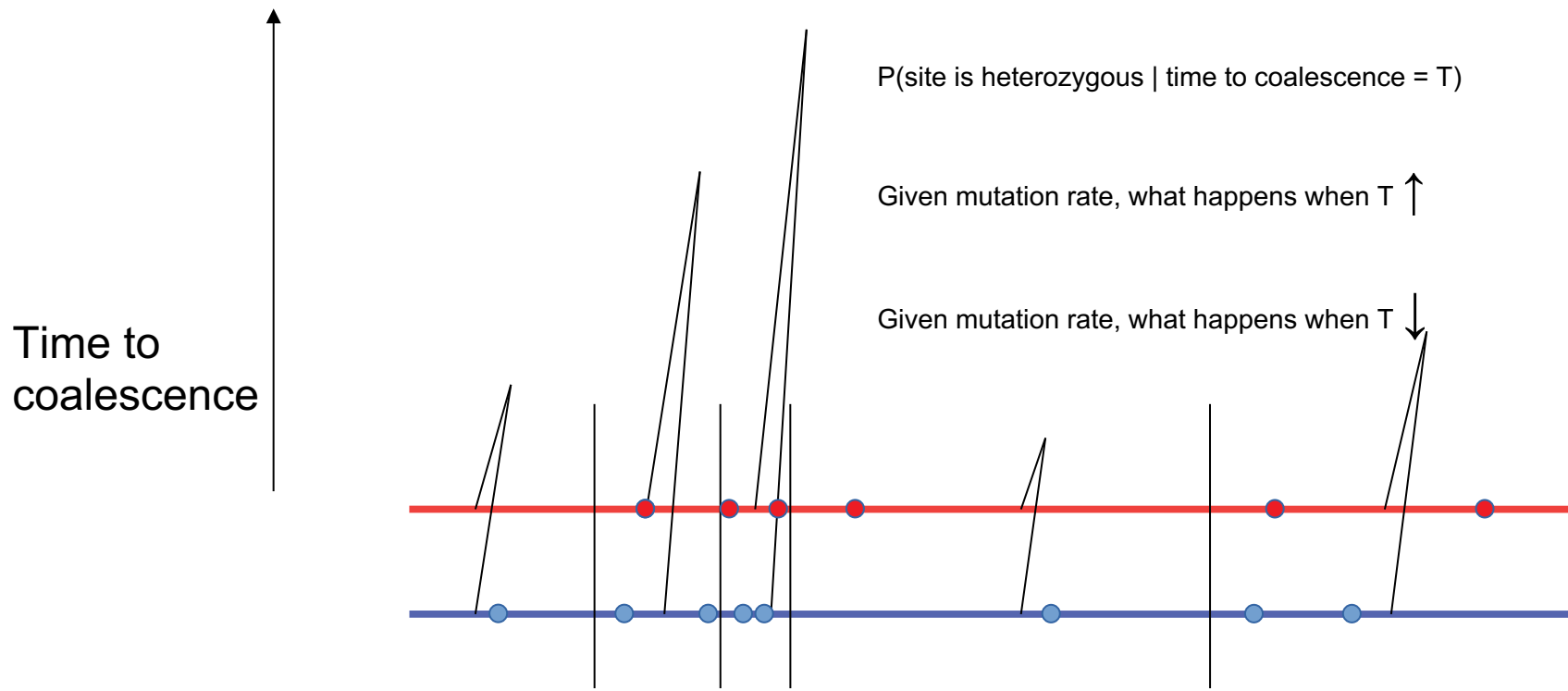
PSMC: HMM for genome



PSMC: HMM for genome



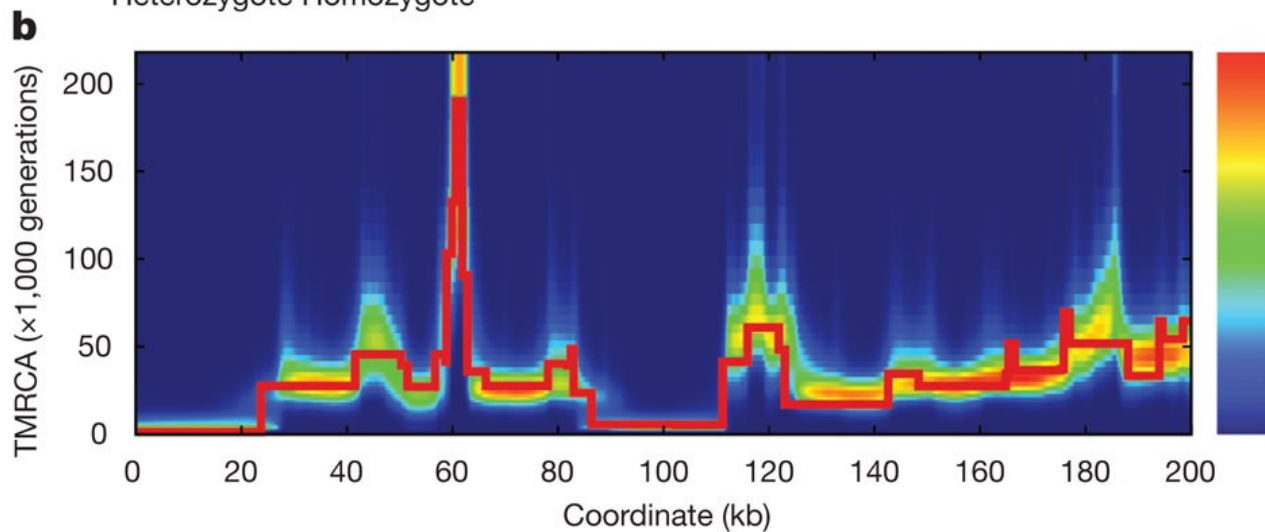
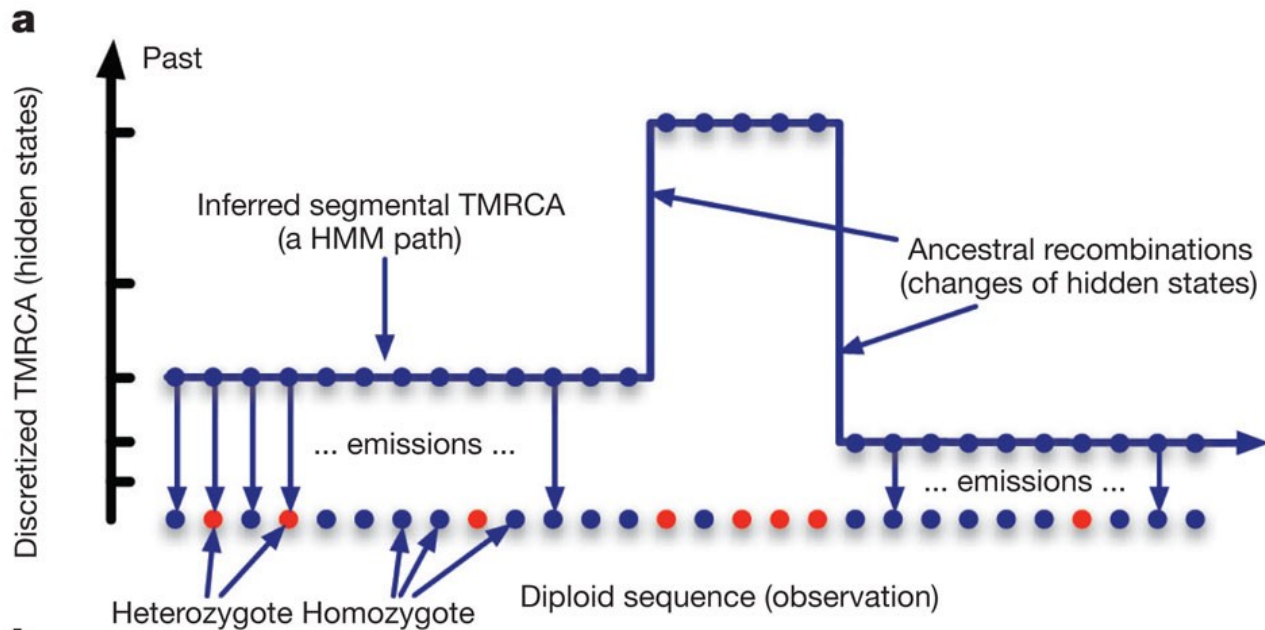
PSMC: HMM for genome



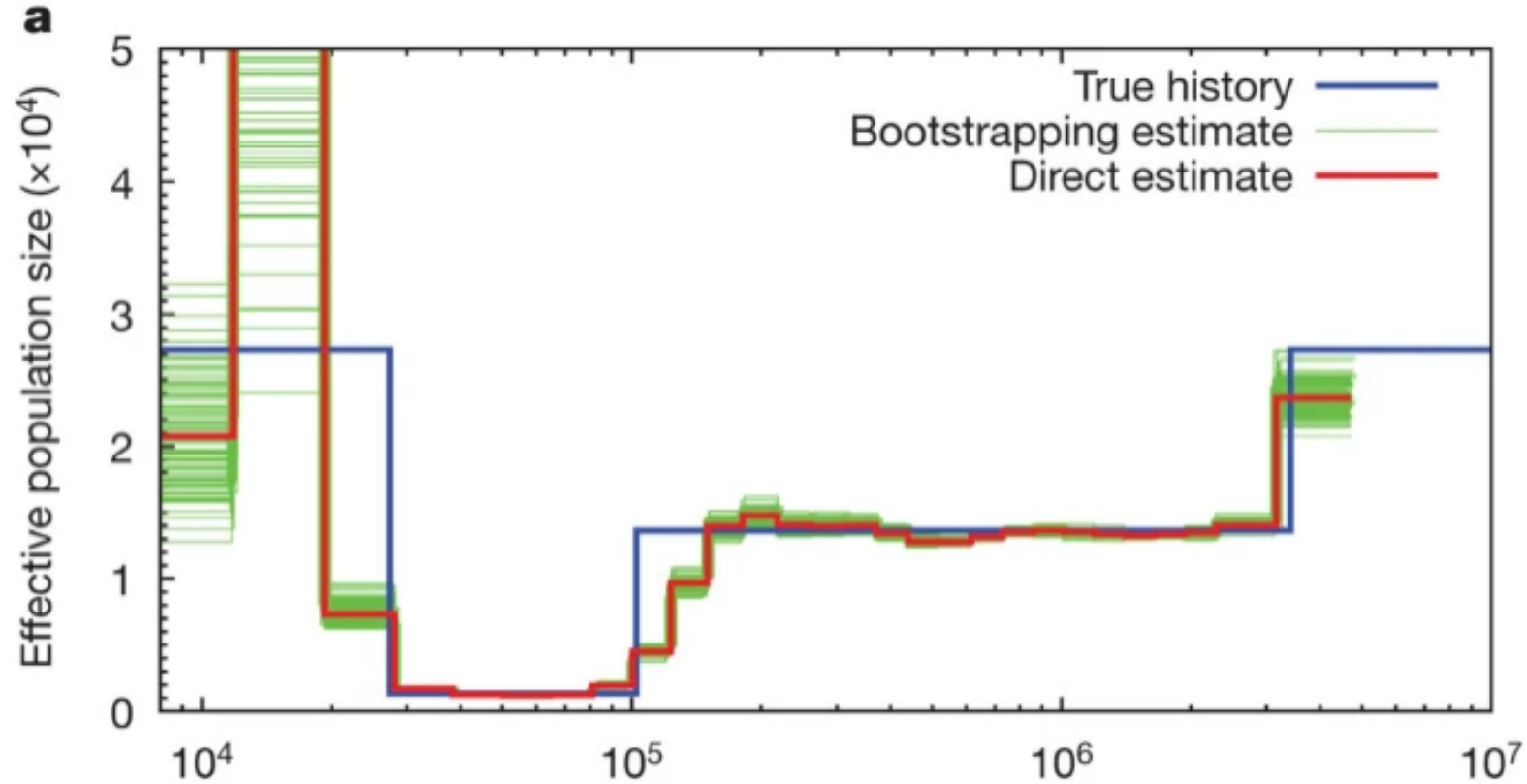
PSMC: Some missing bits

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

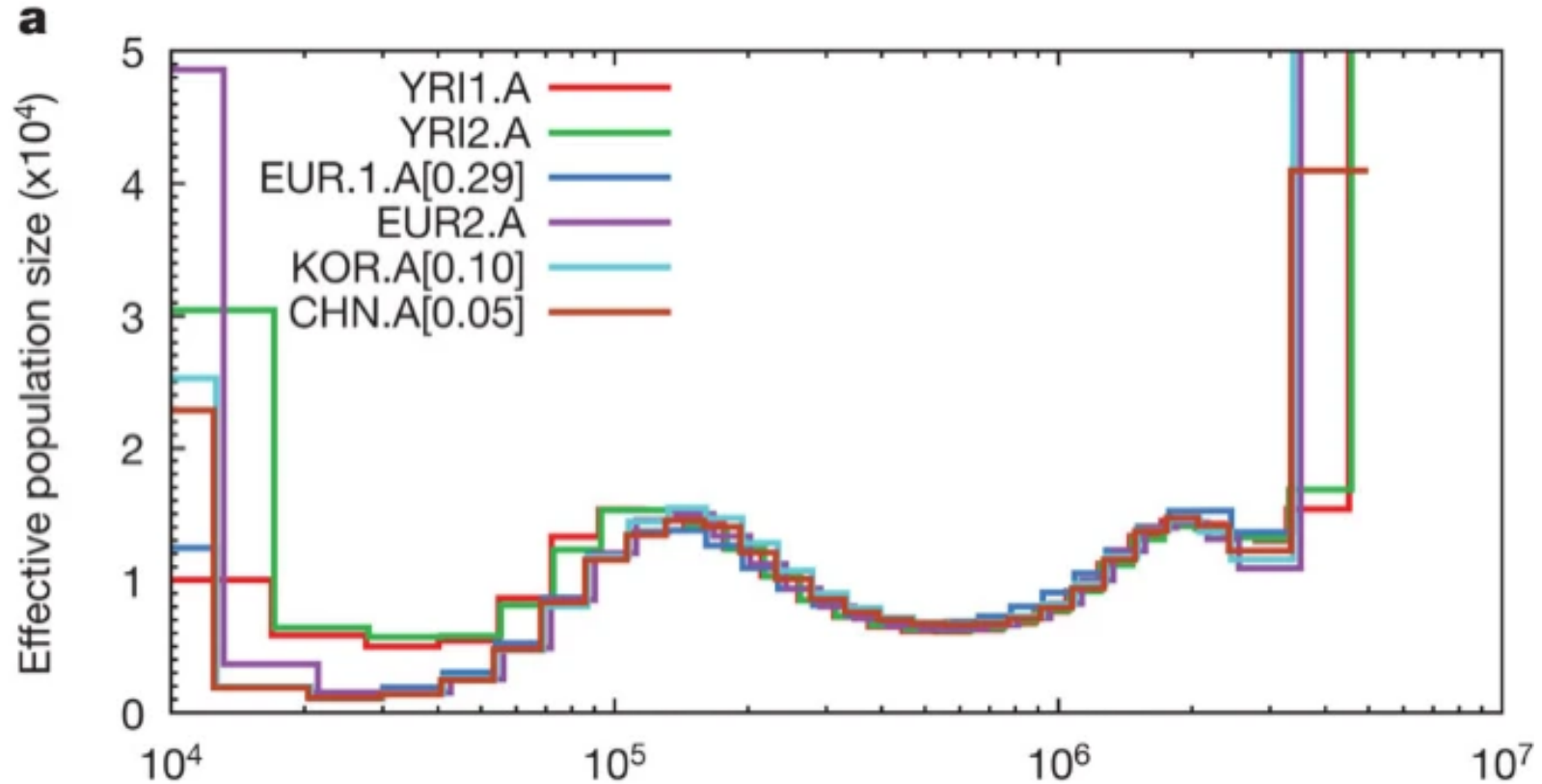
PSMC: Putting it all together



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
 - Many methods in same framework – MSMC, SMC++, MiSTI

- Exercises after the break