Sequentially Markovian Coalescence

André E. R. Soares

Populations Genomics in Practice 2023

The idea

To reconstruct the demographic history of a species from genome sequences.

The most common methods to do so

PSMC

doi:10.1038/nature10231

Inference of human population history from individual whole-genome sequences

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

MSMC

Published: 22 June 2014

Inferring human population size and separation history from multiple genome sequences

Stephan Schiffels & Richard Durbin

Nature Genetics **46**, 919–925 (2014) | Cite this article

The most common methods to do so

PSMC

Unphased genome

One diploid individual

MSMC

Phased genomes

Many individuals

Coalescence methods

So it has some **

Some assumptions and particularities:

- Neutral evolution

- Panmitic population: equal coalescence probabilities for all pairs of lineages

- **Coalescence rate is related to population size.** (as seen earlier)

If we have a genealogy that shows when the coalescent events happened, we then can infer the population size of a population.

Sequentially Markovian Coalescent

Think different

Traditional coalescence methods

- They work backwards

- They build full genealogies from modern/extant samples.

Sequentially Markovian Coalescence

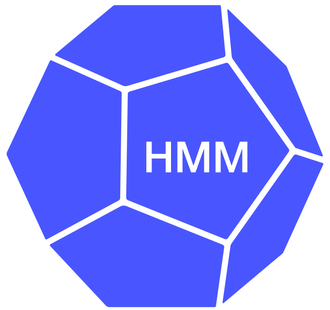
- They go along the genome

- It builds a “local" genealogy for each locus on the genome.

How does it work?

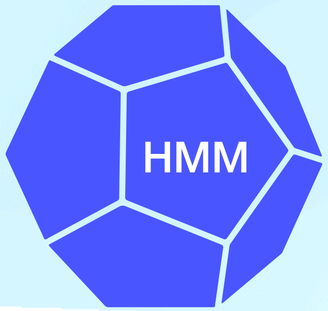
PSMC uses a HMM

HMM = Hidden Markov Model

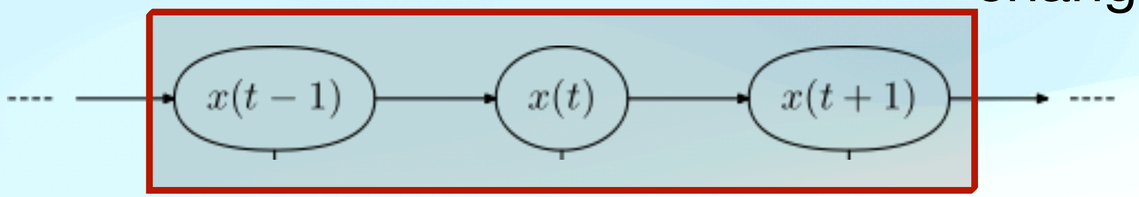


How does it work?

Some SMC methods will use a HMM

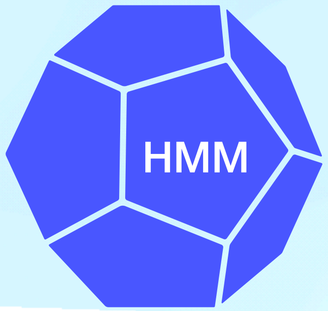


Some states that change with time

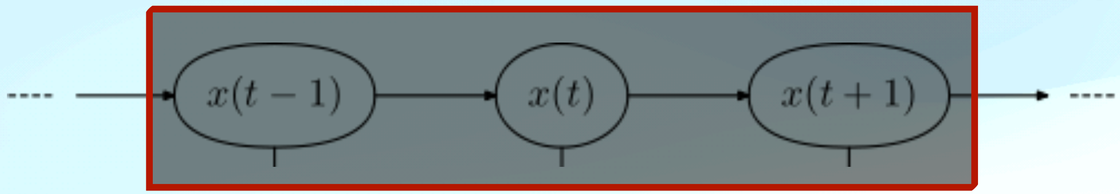


How does it work?

Some SMC methods will use a HMM

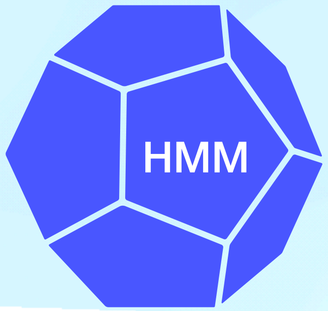


But they’re hidden (you can’t observe them)



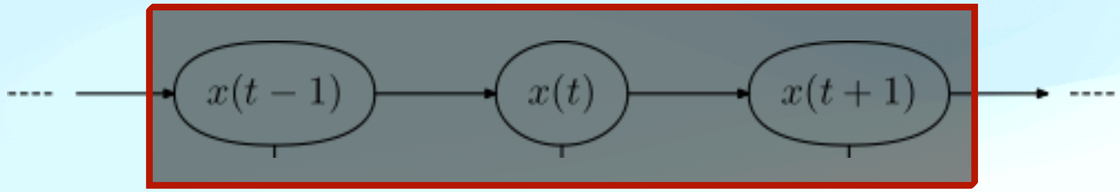
How does it work?

Some SMC methods will use a HMM



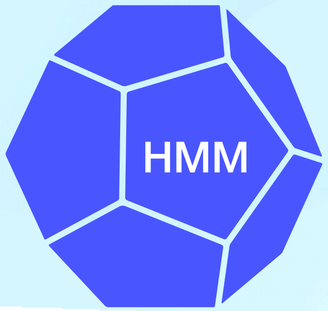
Markov process

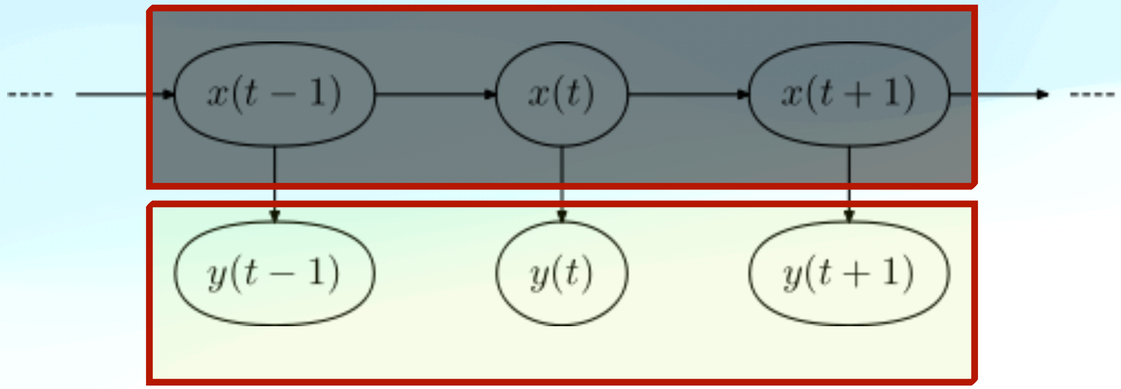
But they’re hidden (you can’t observe them)



How does it work?

Some SMC methods will use a HMM

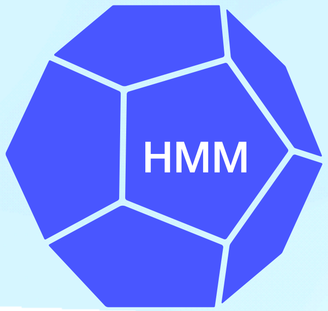


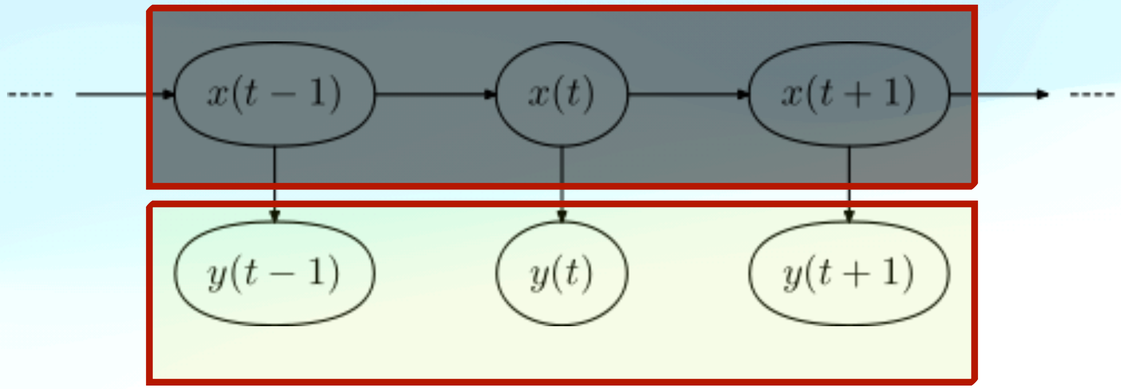


But you can observe
something connected to them

How does it work?

Some SMC methods will use a HMM

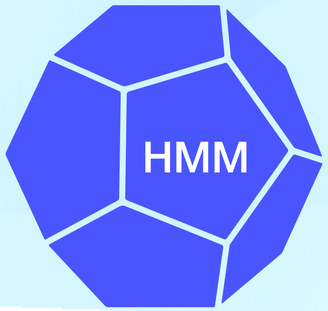


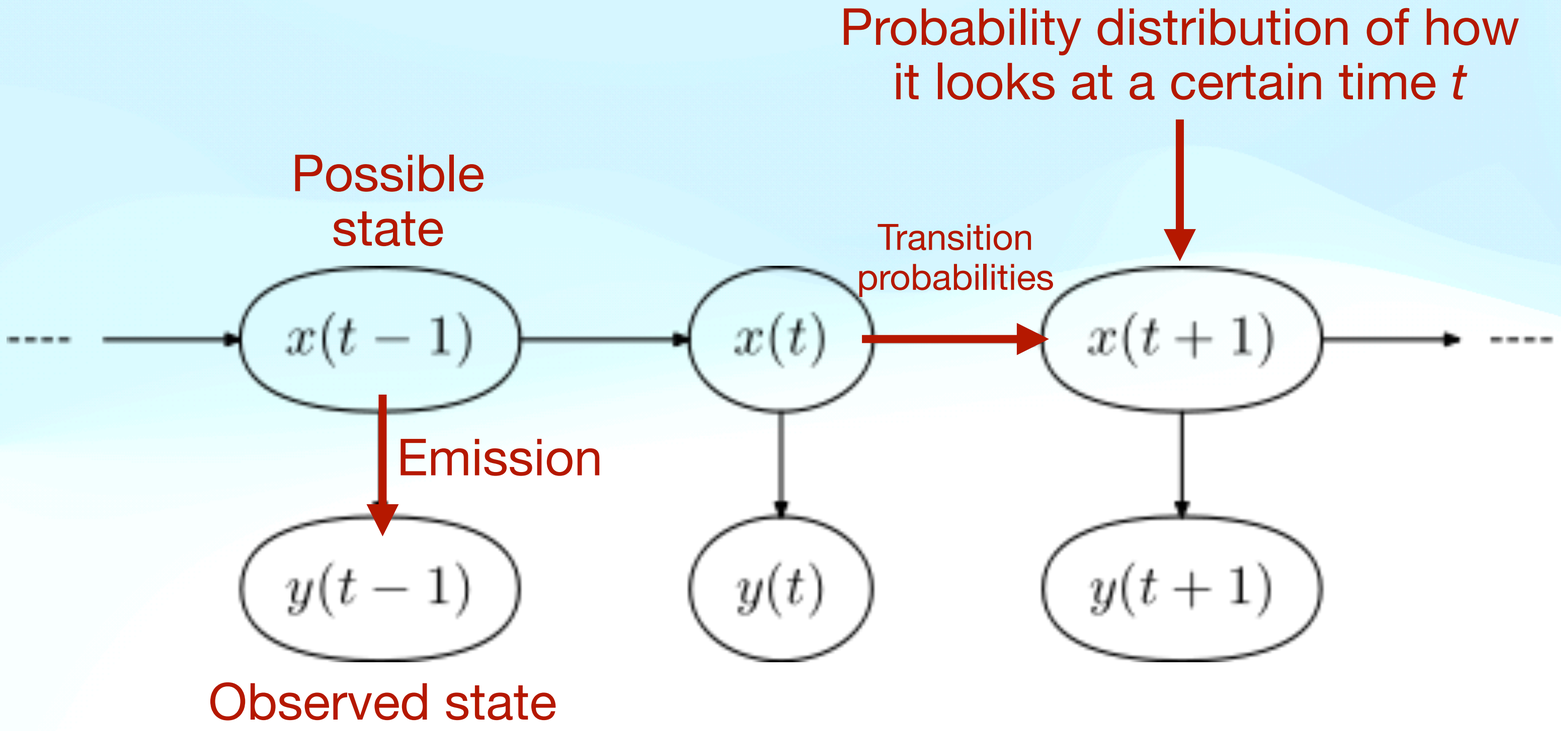


It’s really powerful for dynamic systems and is used a lot in computational biology (annotation, etc)

How does it work?

Some SMC methods will use a HMM

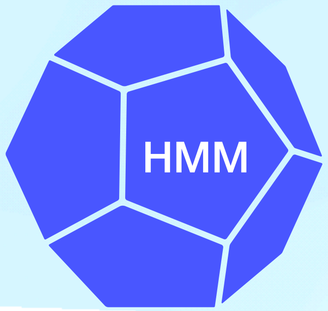


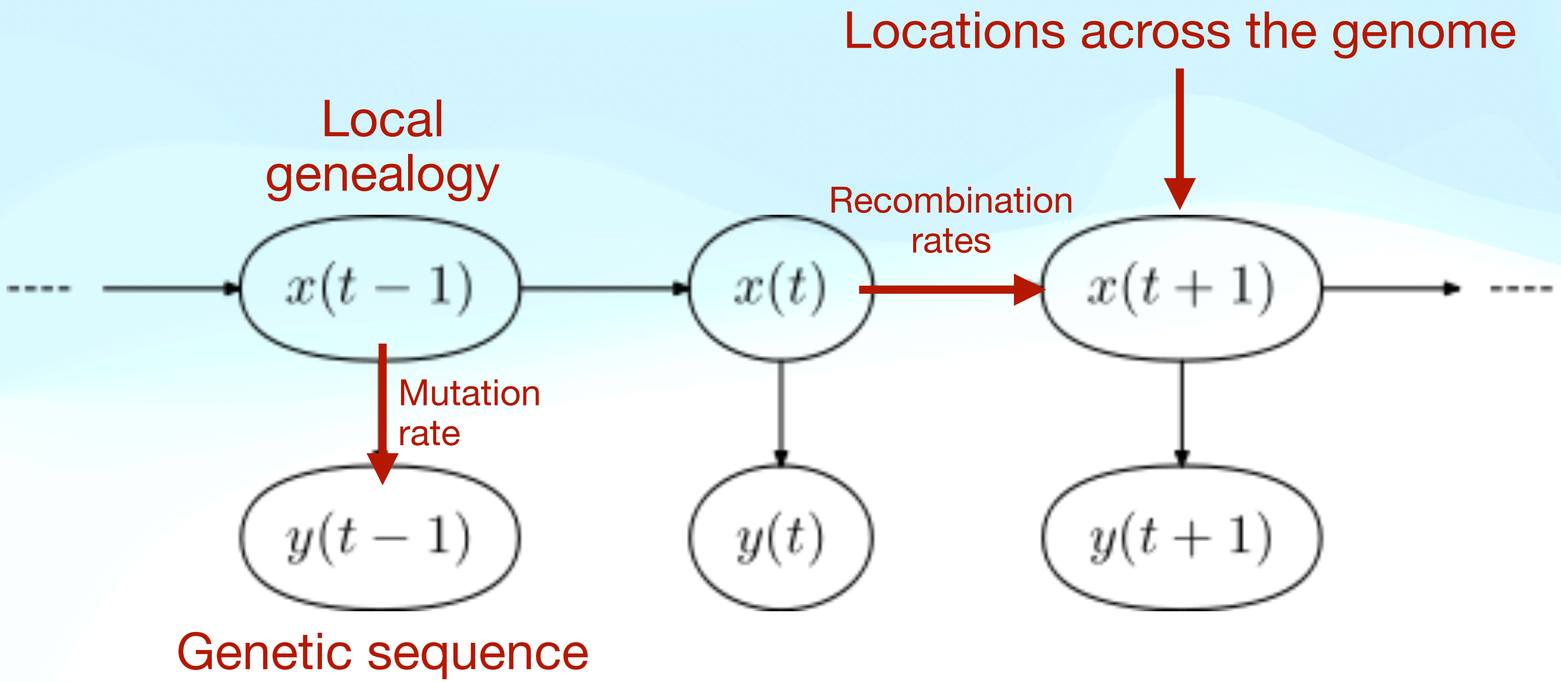


Naming each part of it…

How does it work?

Some SMC methods will use a HMM

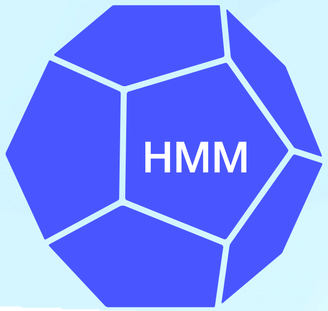


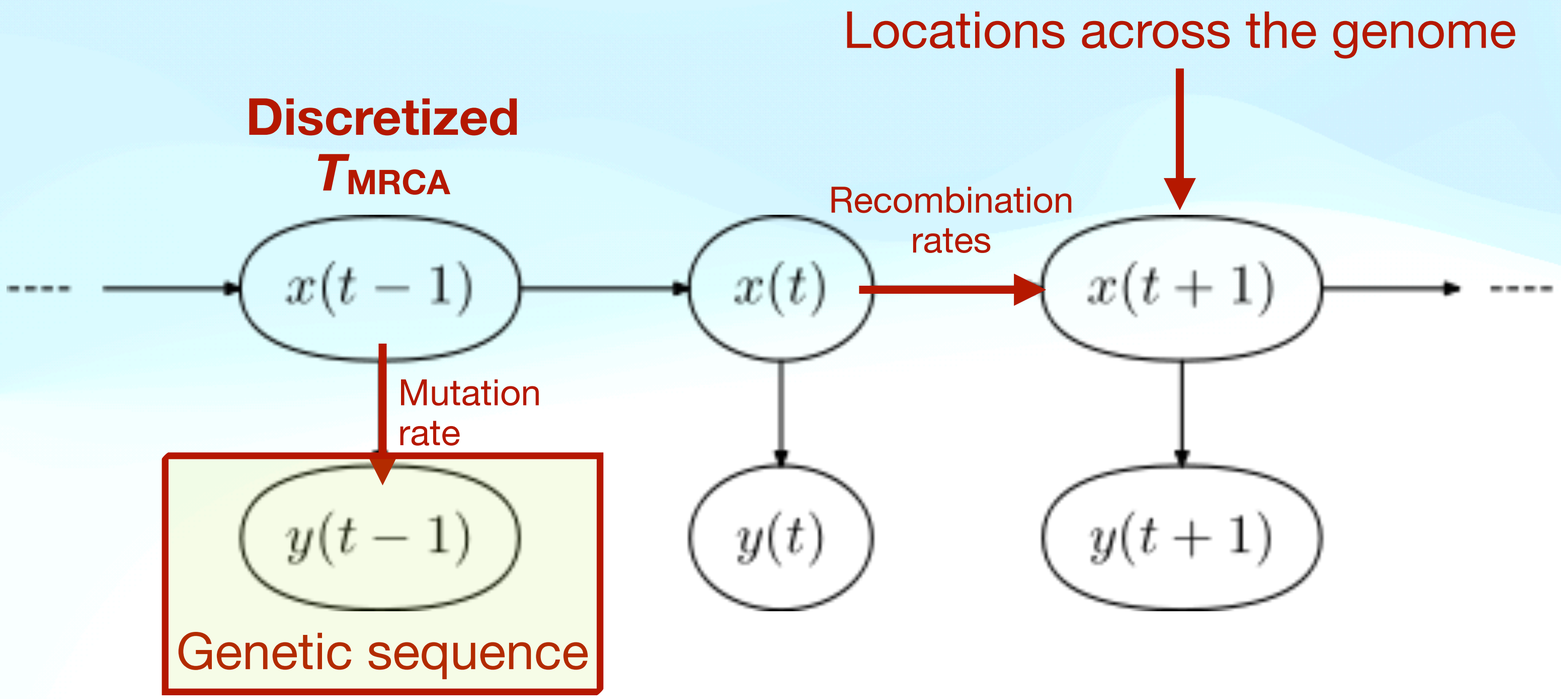


How SMC methods work

How does PSMC work?

Some SMC methods will use a HMM

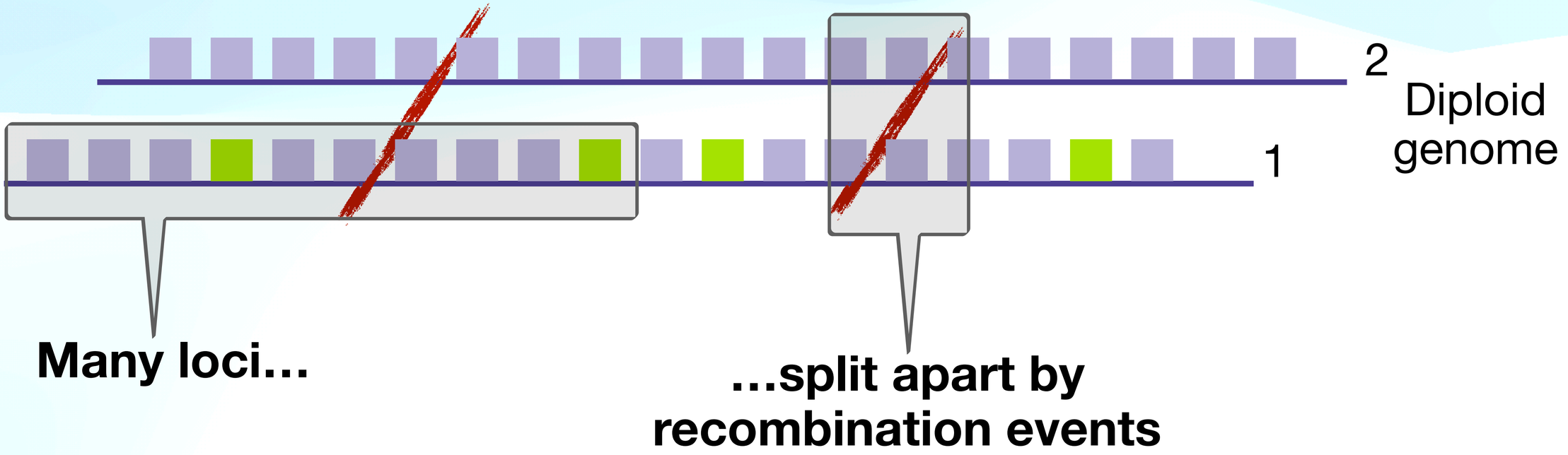


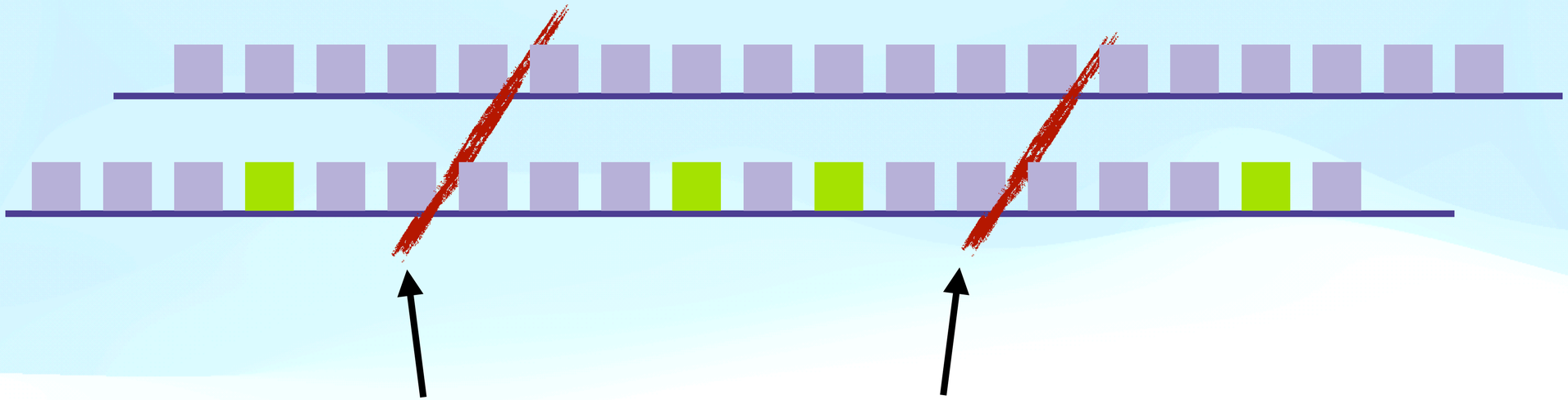


- It creates 100 bp windows
- 1 for hets, 0 for homs

How does SMC work?

In a whole genome we have…



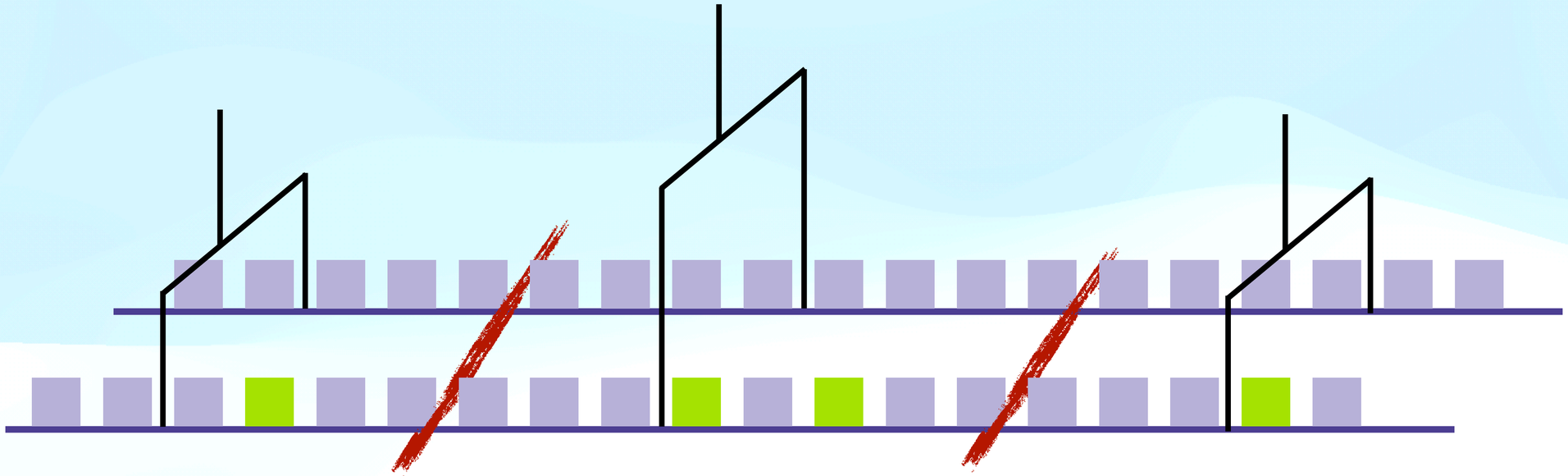


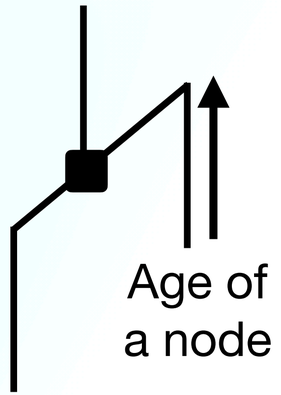
The SMC is capable of calculating the probability of recombination breakpoints.

Nucleotides states

Recombination breakpoints

Since a “tree" between two sequences is known, the only variable that change is T_{MRCA}.

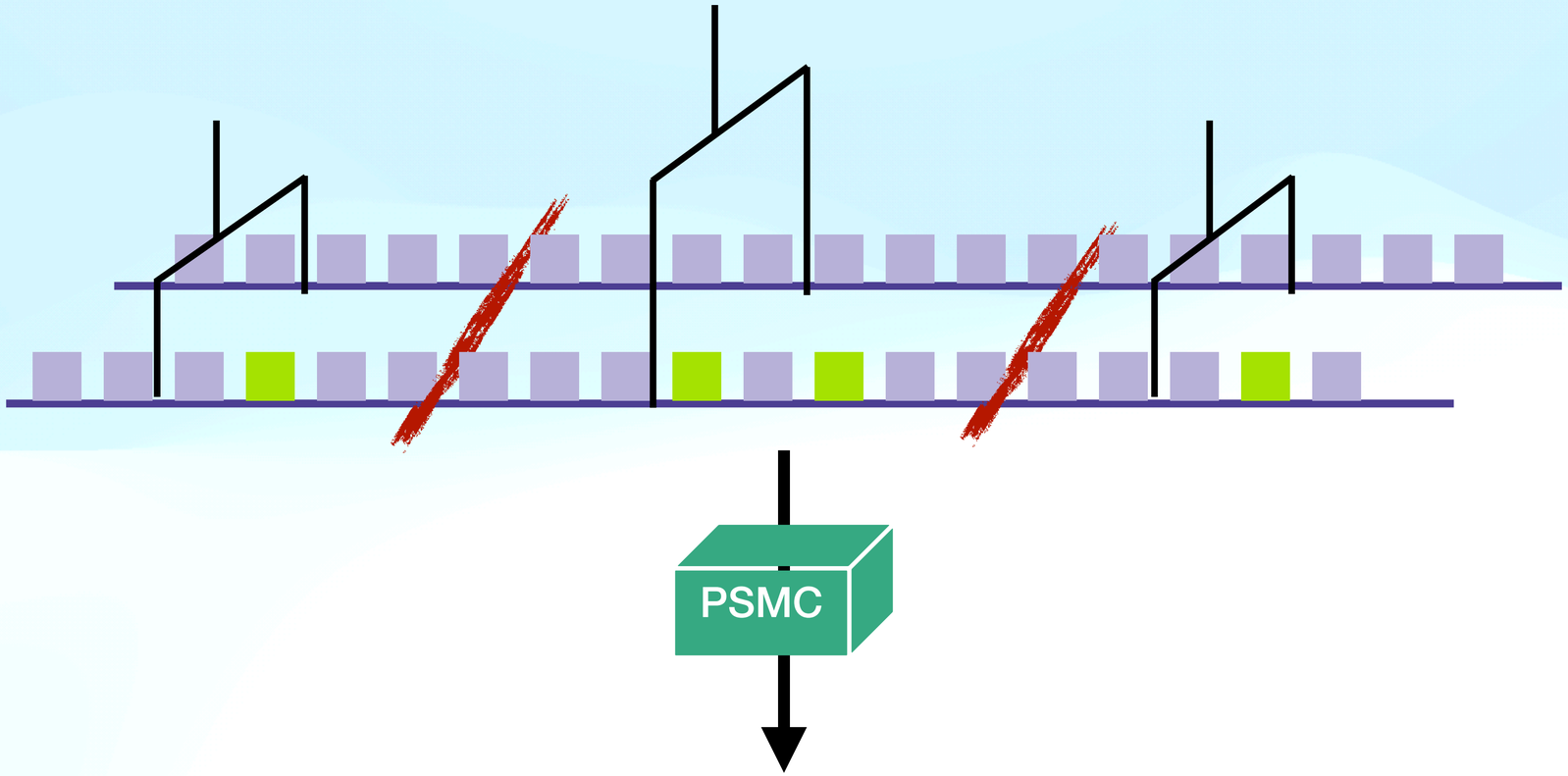




Nucleotides states

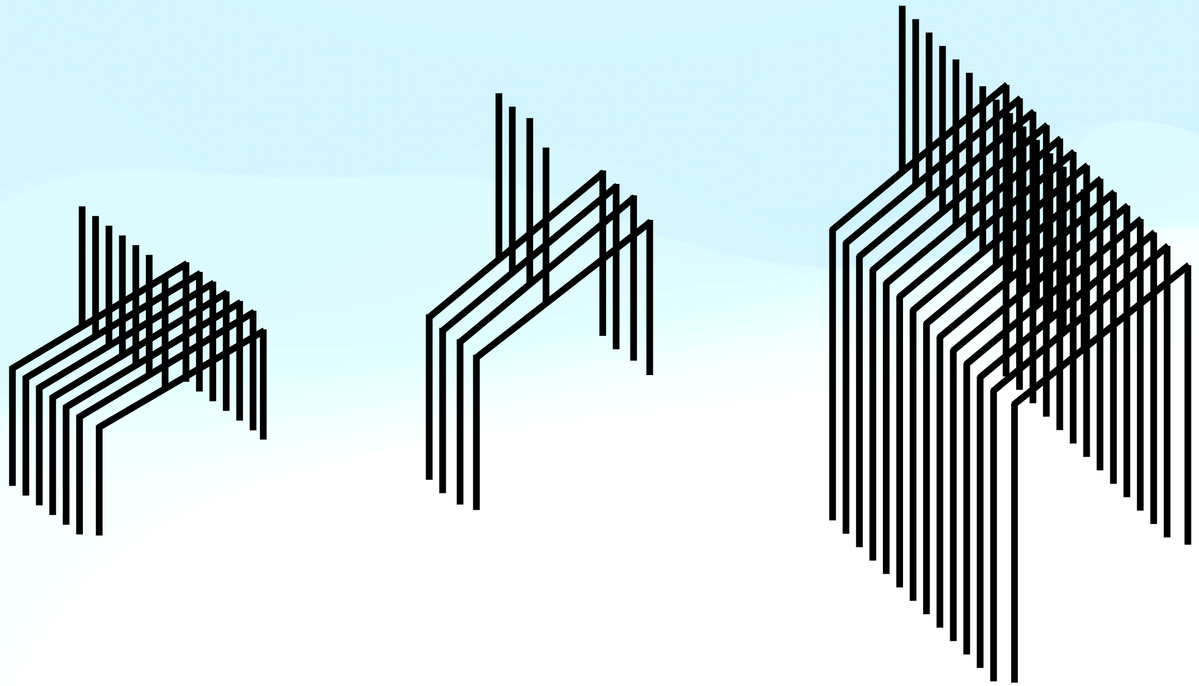
Recombination breakpoints

So when PSMC runs…



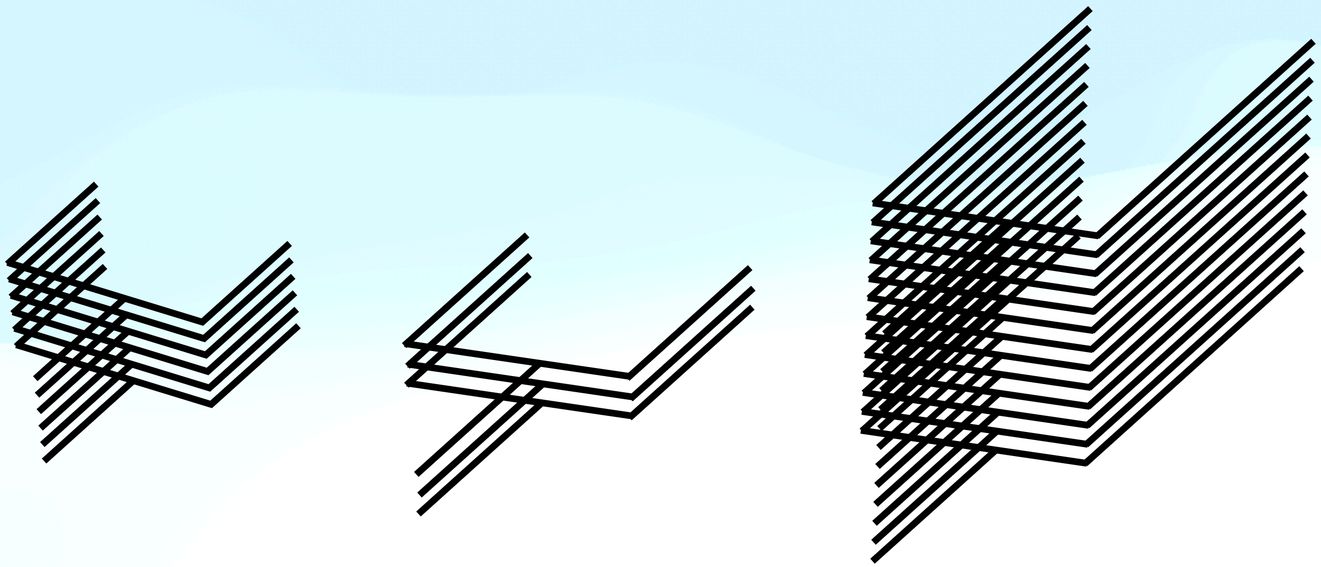
It writes down the likelihood of getting the **genomic data**, the **recombination break points**, and the T_{MRCA} along the sequence, per window (default is 100 bp).

You can count how many events happened at a certain time period



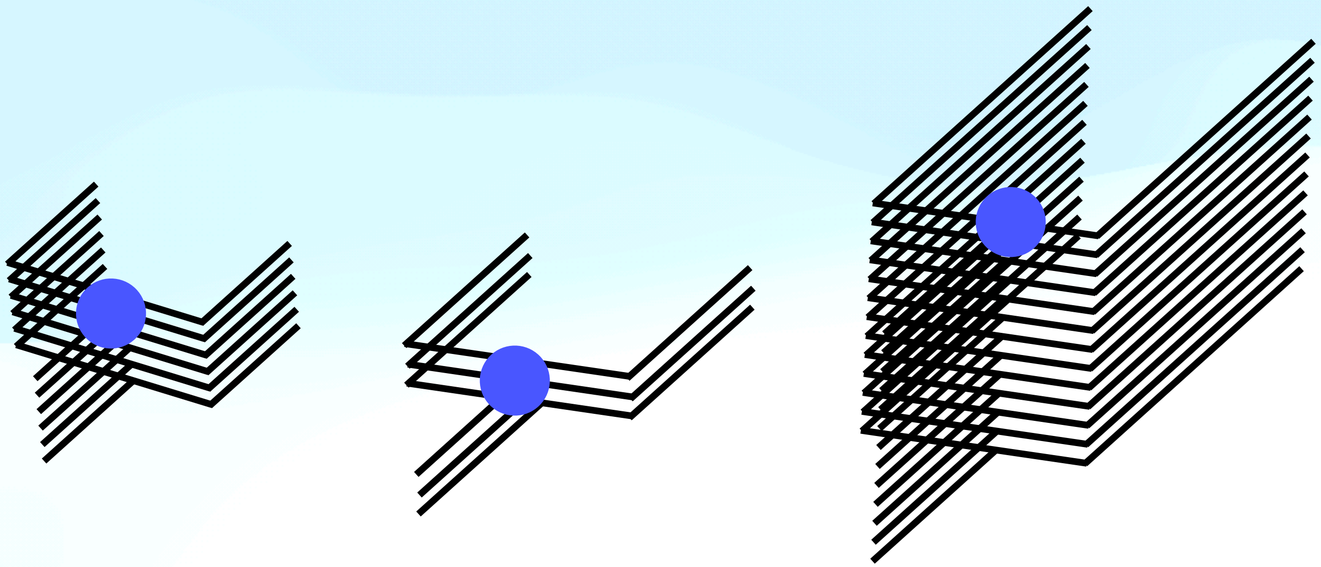
PSMC and MSMC use this information to reconstruct the effective population size through time

You can count how many events happened at a certain time period



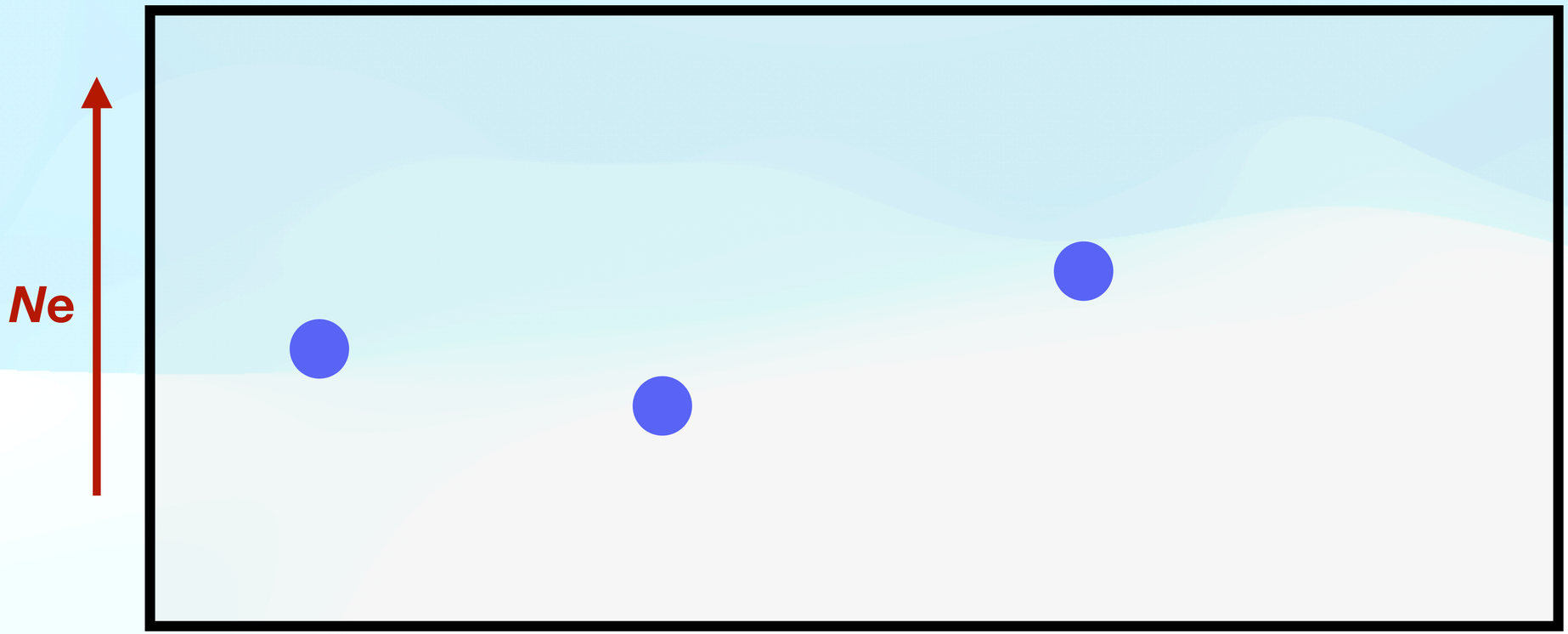
PSMC and MSMC use this information to reconstruct the effective population size through time

You can count how many events happened at a certain time period



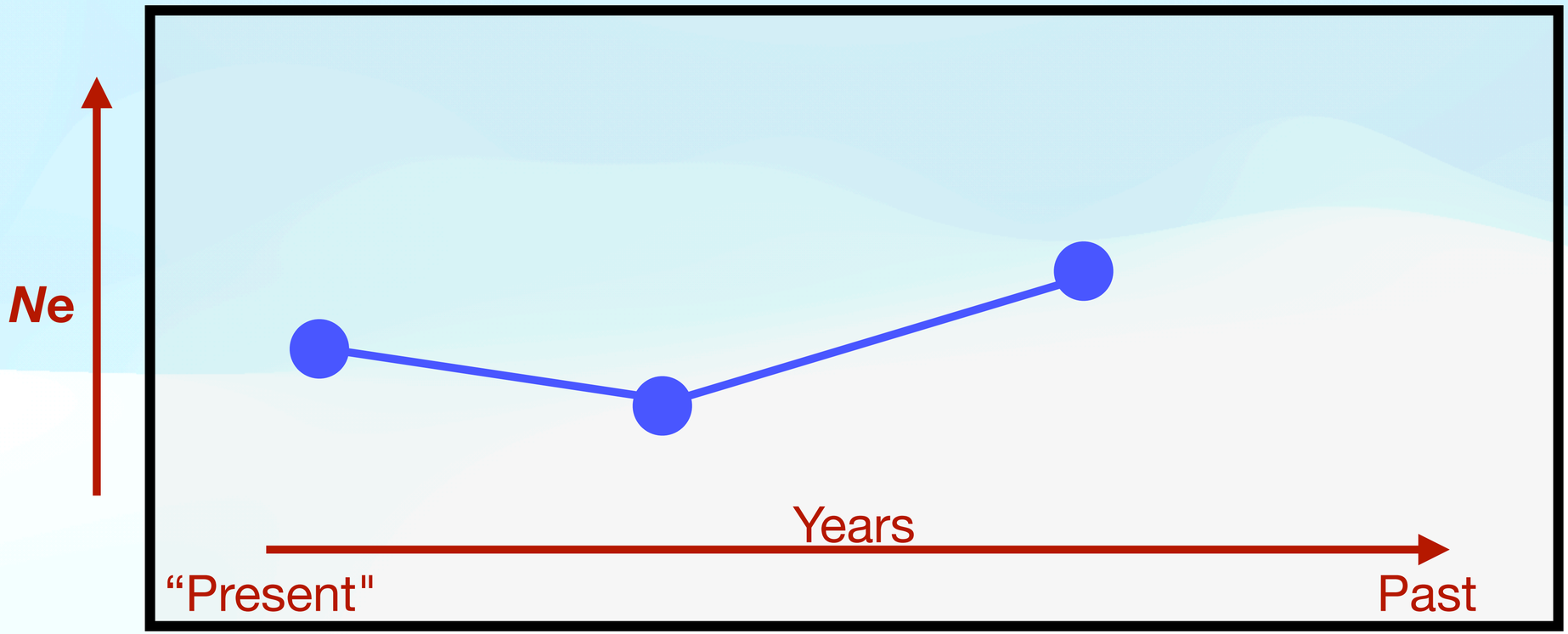
PSMC and MSMC use this information to reconstruct the effective population size through time

You can count how many events happened at a certain time period



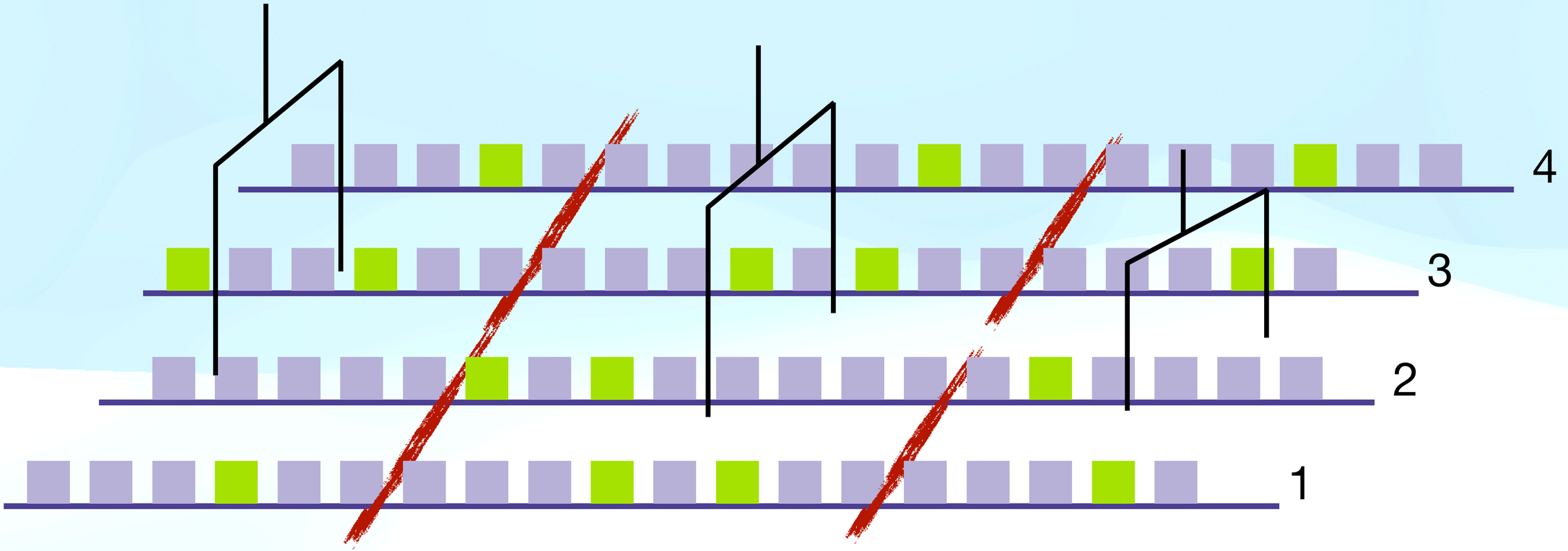
PSMC and MSMC use this information to reconstruct the effective population size through time

You can count how many events happened at a certain time period



PSMC and MSMC use this information to reconstruct the effective population size through time

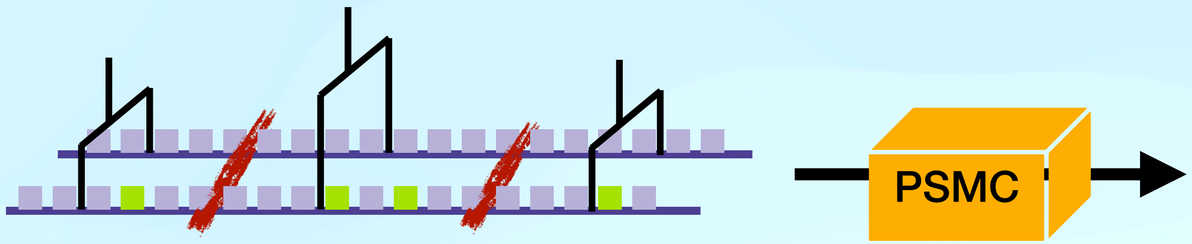
What about MSMC?

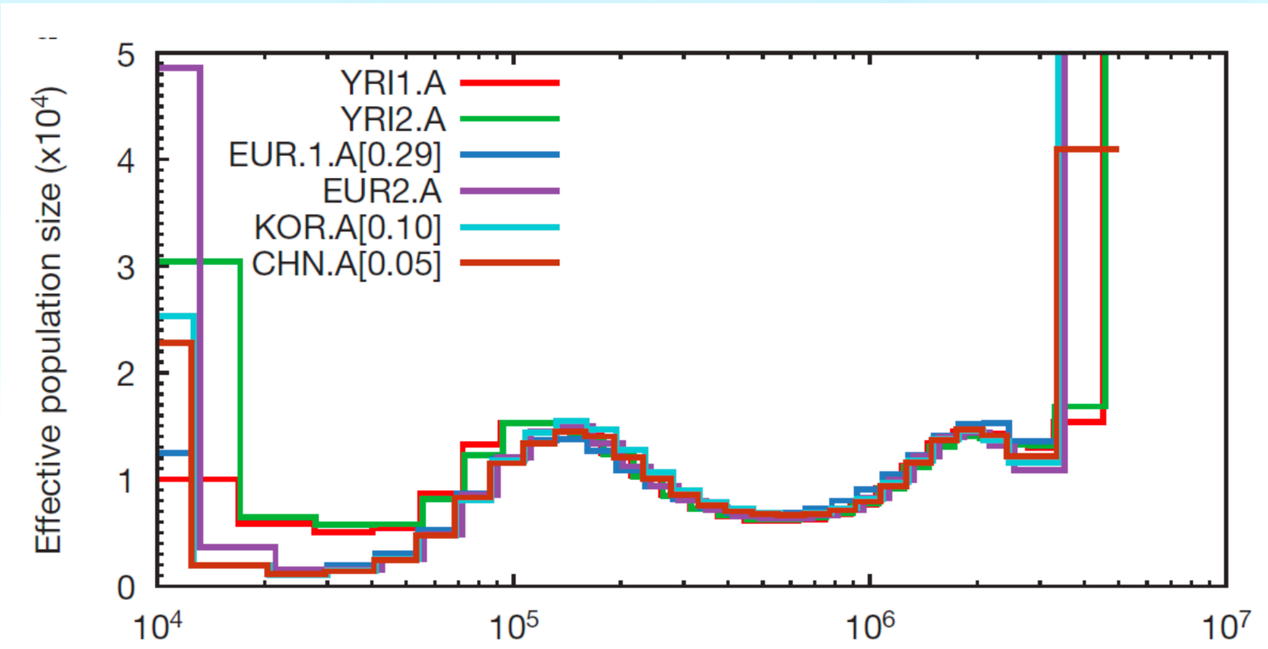


PSMC: Local genealogy is the time to the T_{MRCA}, because only one tree

MSMC: Uses only some of the local trees, the ones that describe the T_{MRCA} of two alleles at a locus

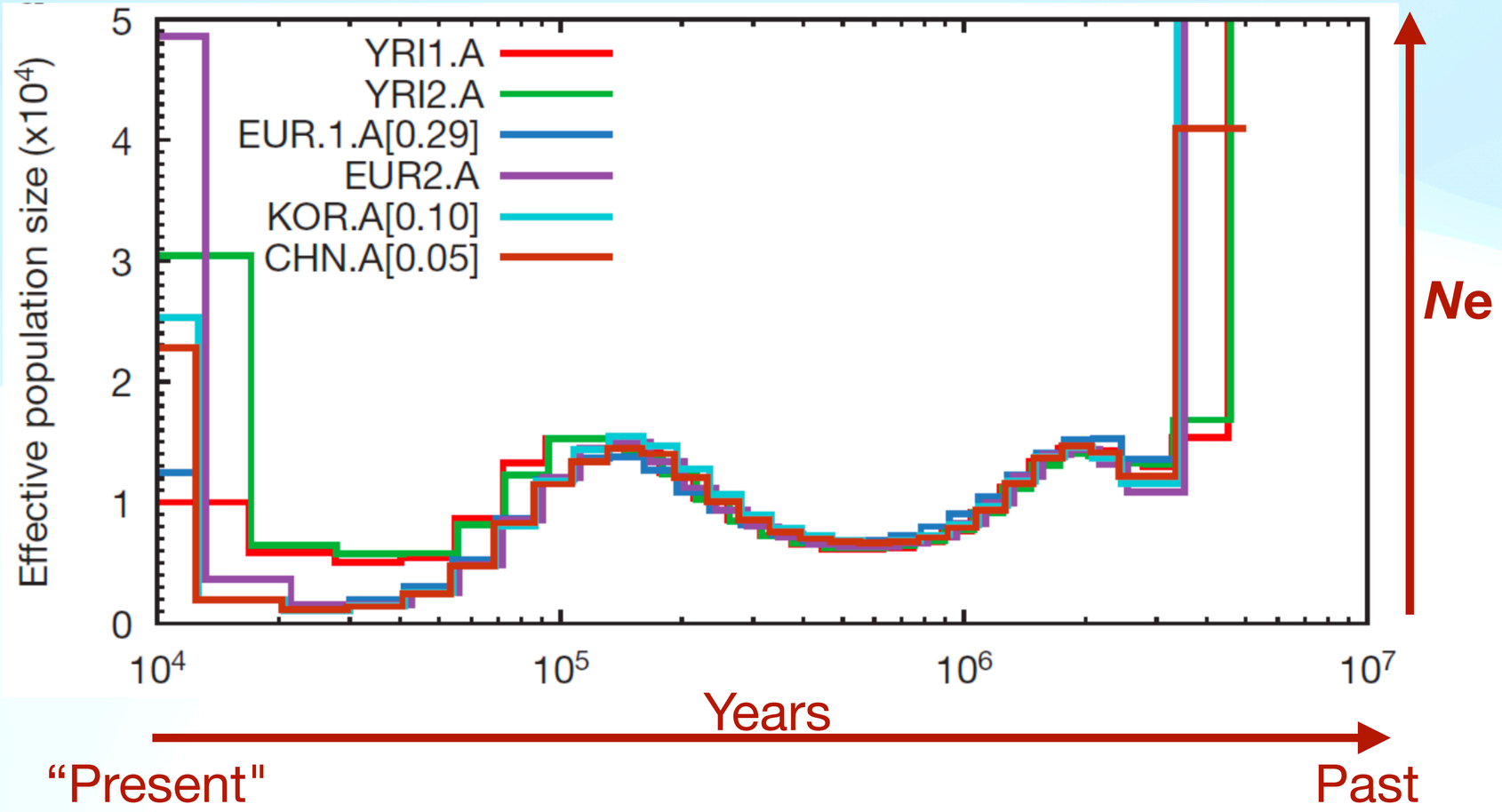
What does it look like?



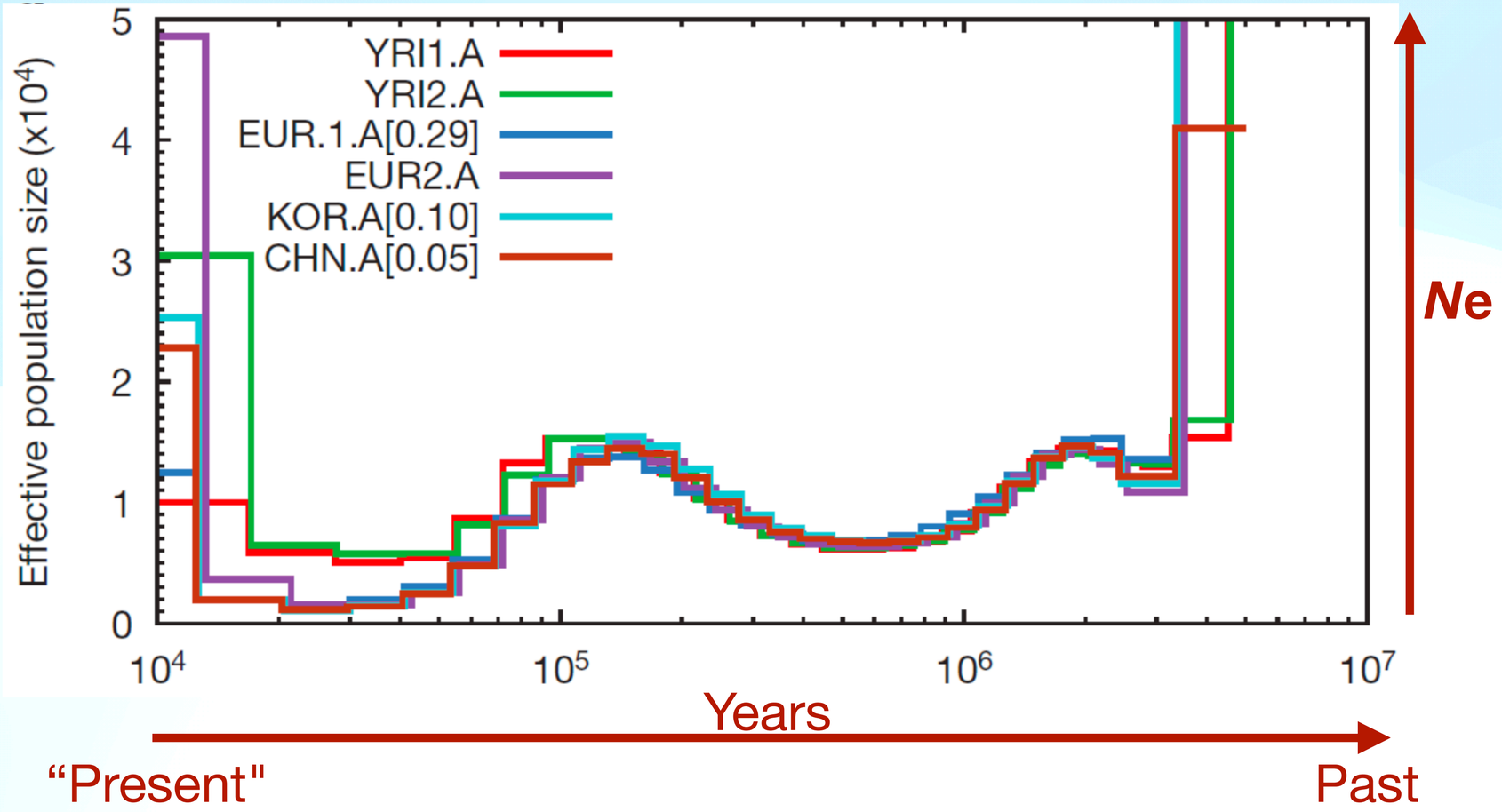


* You need to know the **generation time** and **mutation rate** to scale the graph.

PSMC plot

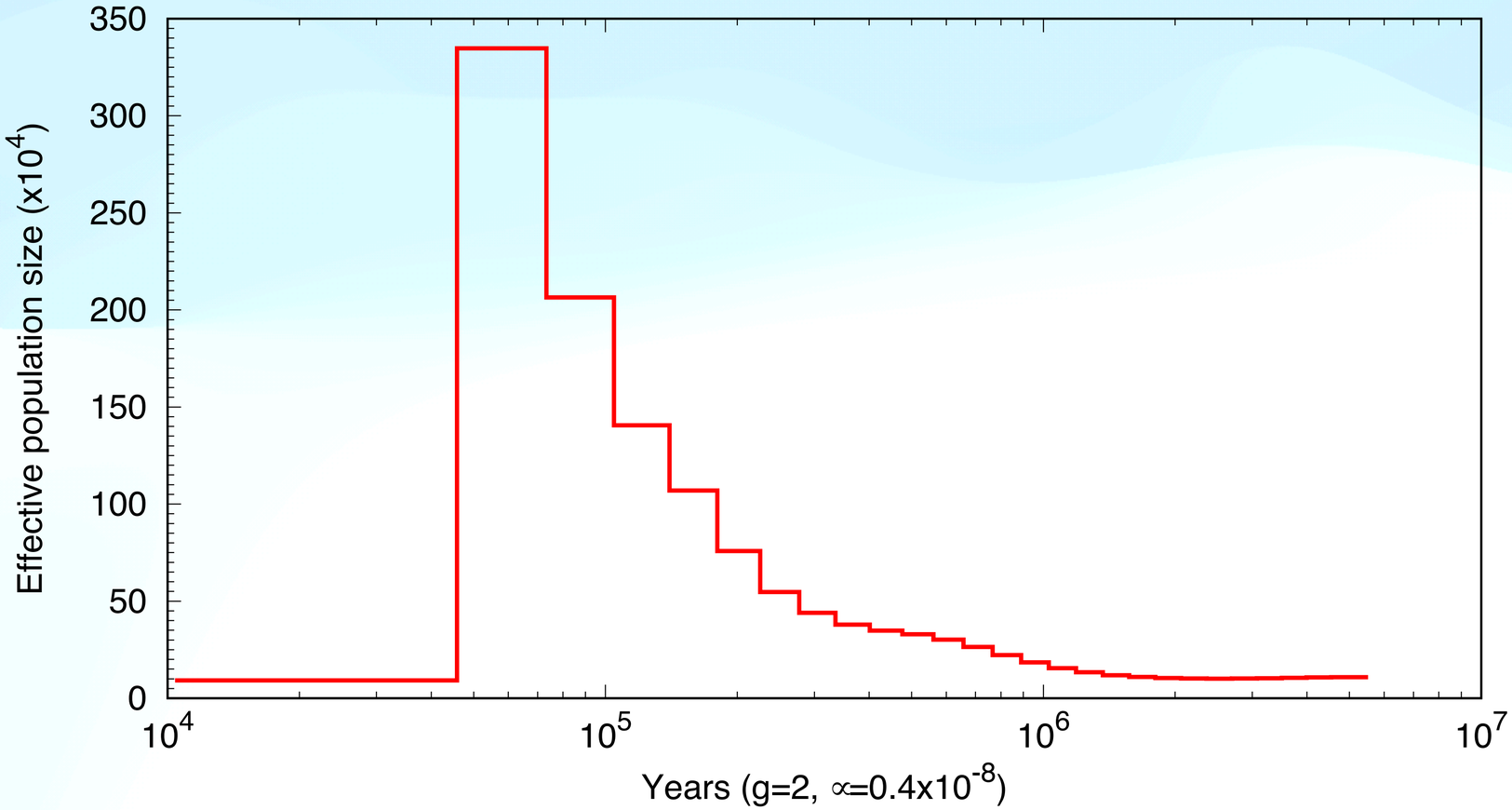

Effective population size (x10⁴)
5
4
3
2
1
0
YRI1.A
YRI2.A
EUR.1.A[0.29]
EUR2.A
KOR.A[0.10]
CHN.A[0.05]
10⁴
10⁵
10⁶
10⁷
Years
"Present"
Past
Ne

PSMC plot


Effective population size (x10⁴)
5
4
3
2
1
0
YRI1.A
YRI2.A
EUR.1.A[0.29]
EUR2.A
KOR.A[0.10]
CHN.A[0.05]
10⁴
10⁵
10⁶
10⁷
Years
"Present"
Past
Ne

PSMC plot - example 01

What do you see?



PSMC: the weak points

- It’s really dependable on proper generation times and mutation rates;

- Sudden drops in population will show as a steady decline over time instead of a sudden bottleneck;

- Population structure will strongly affect the results (as it affects coalescence times);

- Not great with zoo animals;

- Doesn’t have a framework for testing hypotheses, must use bootstrap replicates instead.

PSMC: the weak points

Factors that affect the coalescent rates will affect the PSMC too, like natural selection and nonrandom mating.

PSMC *vs* MSMC

- It tends to out-perform PSMC even in single genomes;

- Multiple individuals help increase the chance for alleles with coalescence events in the recent past;

- Deep coalescences are relatively rare in all cases.

PSMC and beyond

MSMC2: Expands on MSMC, that’s limited to eight samples in most cases.

SMC++ (Terhorst et al. 2017): higher accuracy

ASMC (Palamara et al. 2018): 2-3x faster

XSMC (Ki & Terhorst, 2020)

Gamma-SMC (Schweiger & Durbin, published in Advance August 10, 2023): Up to 20x faster, can detect loci under positive selection.

Reading suggestions

REVIEW ARTICLE

Ecology and Evolution Open Access WILEY

A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data

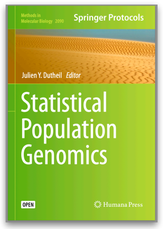
Niklas Mather | Samuel M. Traves | Simon Y. W. Ho

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Limits and convergence properties of the sequentially Markovian coalescent

Thibaut Paul Patrick Sellinger | Diala Abu-Awad | Aurélien Tellier

MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent

Stephan Schiffels and Ke Wang