Recombination and the site frequency spectrum

Fernando Racimo

Adelaide, January 2018

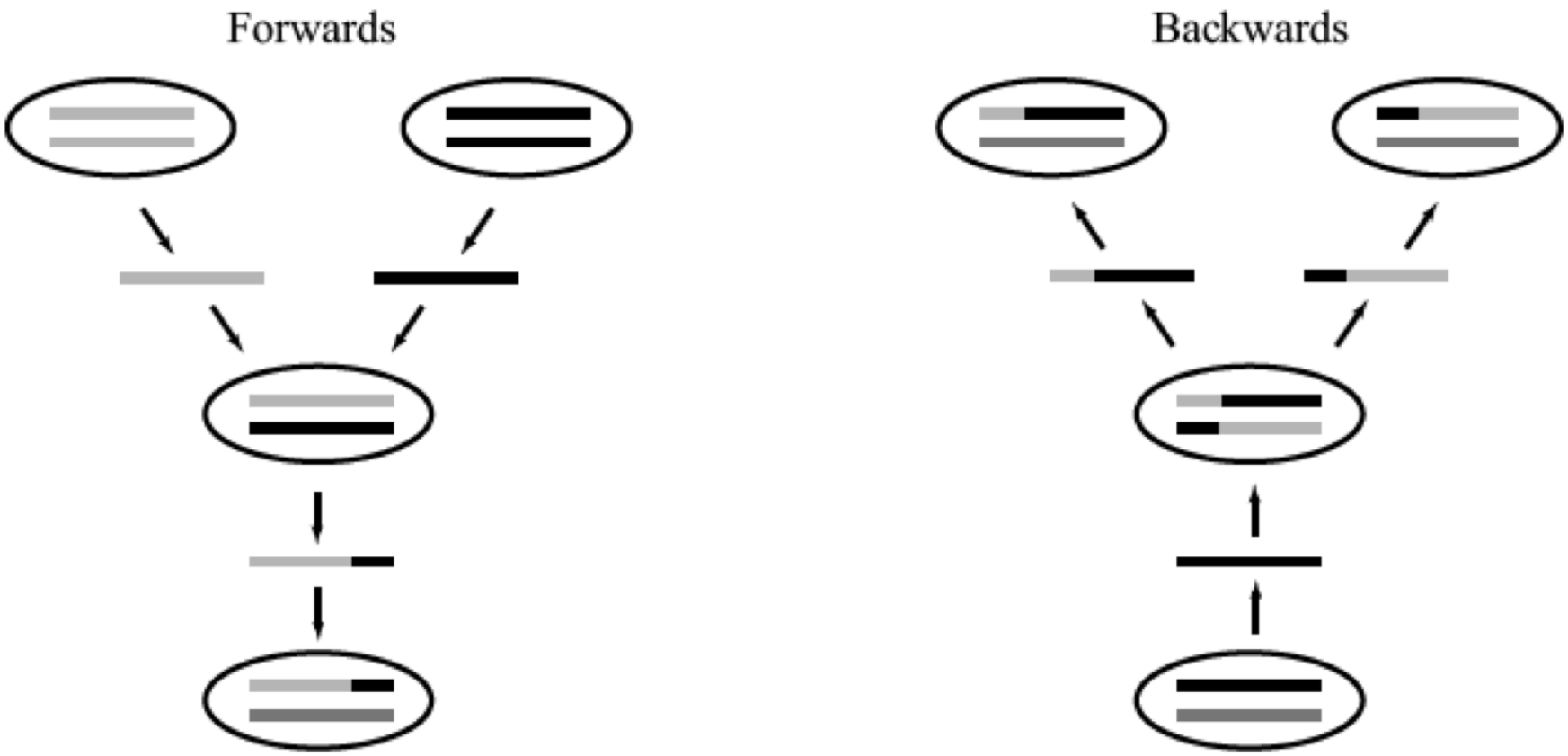
Today

- Recombination
- The site frequency spectrum

Today

- **Recombination**
- The site frequency spectrum

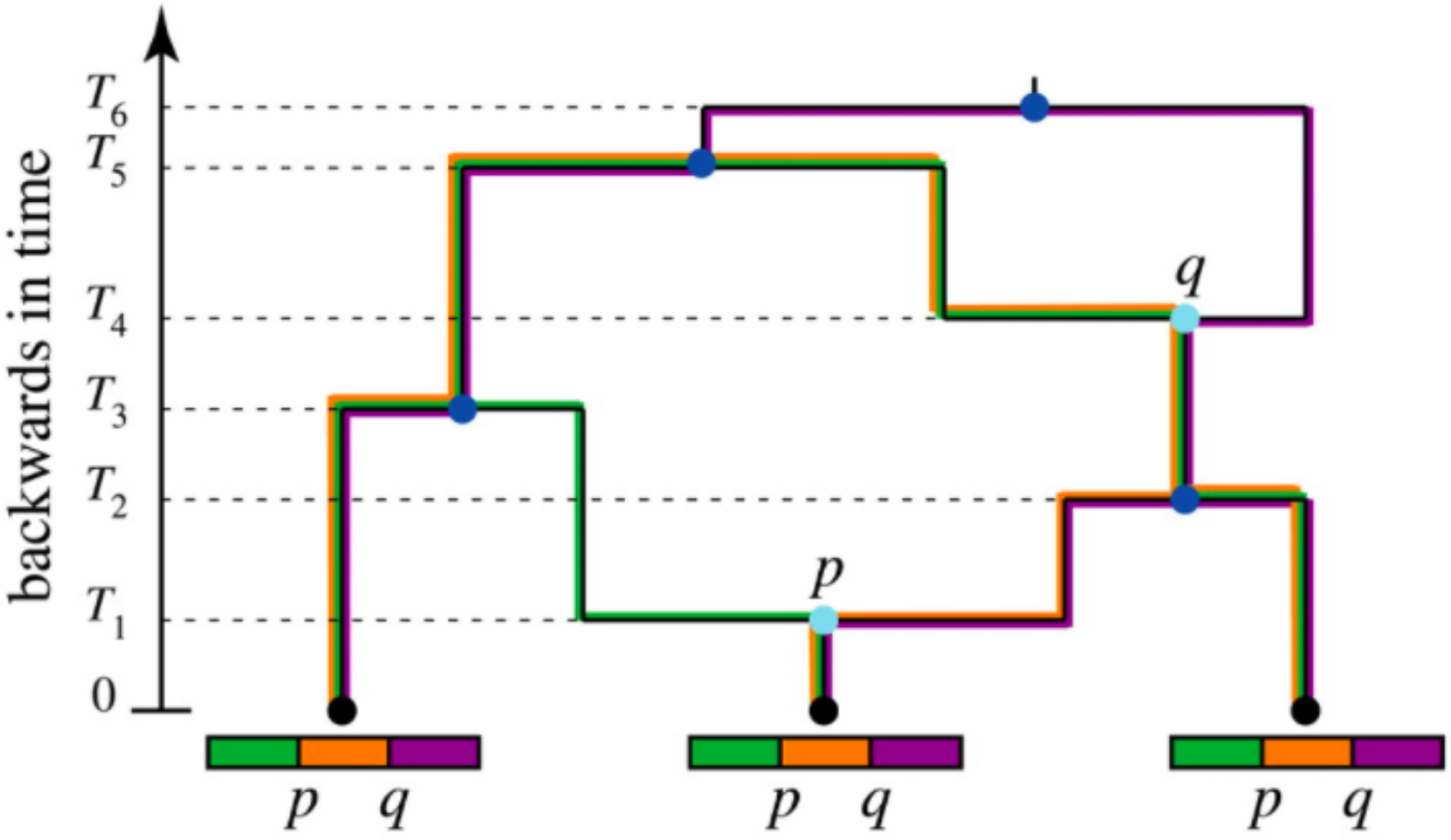
Recombination



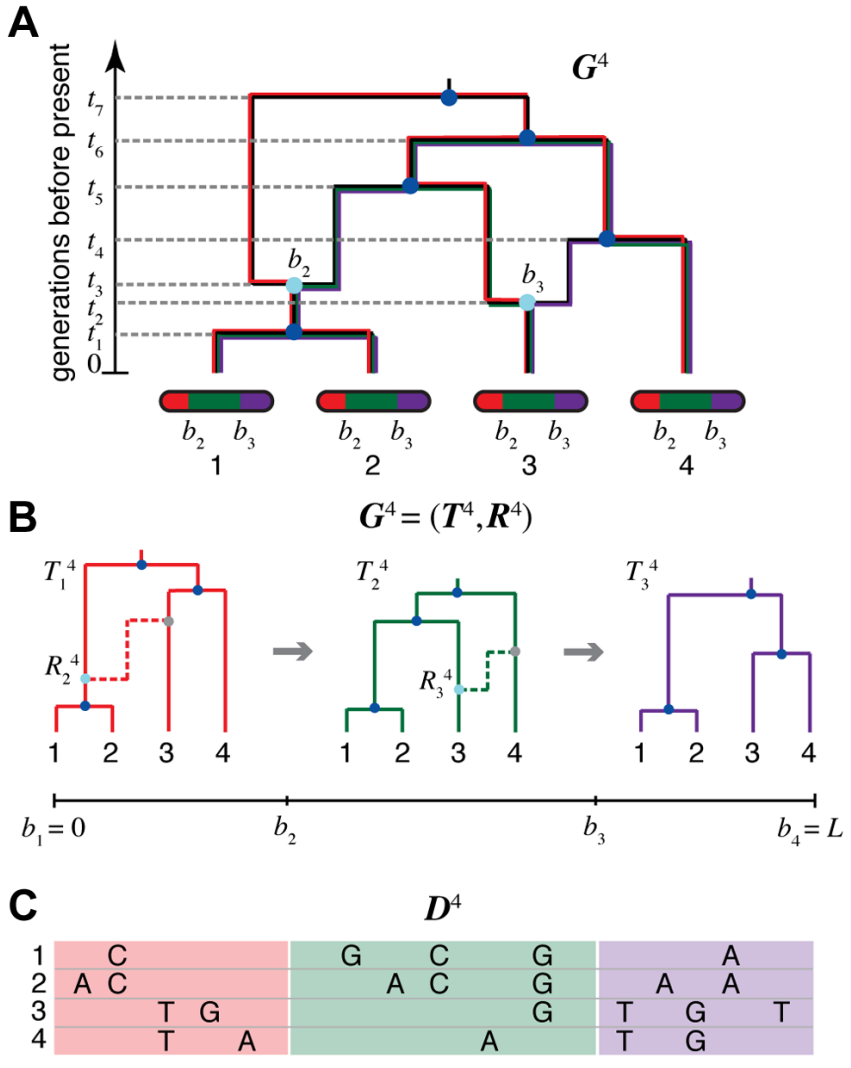
Recombination and coalescence

- Thinking backwards in time, the time to the next recombination event between two sites is exponentially distributed with rate $2c$, where c is the per-site recombination rate.
- Along a sequence, two processes happening in competition: coalescence and recombination
- Both are exponentially distributed (with different rates)
- For two sequences, $P[\textit{coalescence before recombination}] = \frac{1/(2N)}{1/(2N)+2c} = \frac{1}{1+4Nc}$
- In the coalescent timescale, the population-scaled recombination rate ($\rho = 4Nc$) competes against the population-scaled coalescent rate (1).

The ancestral recombination graph (ARG)

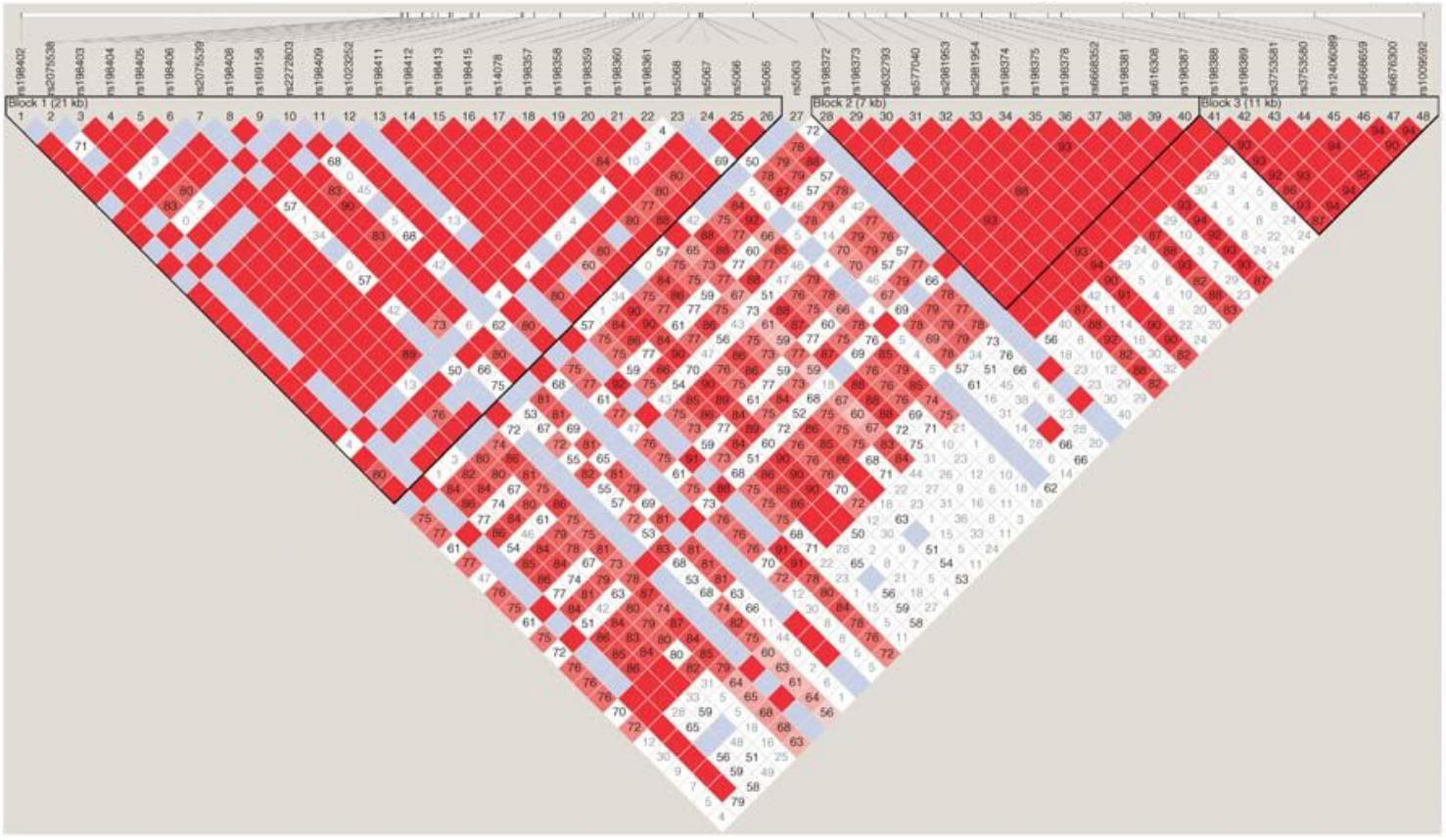


The ancestral recombination graph (ARG)



Linkage disequilibrium

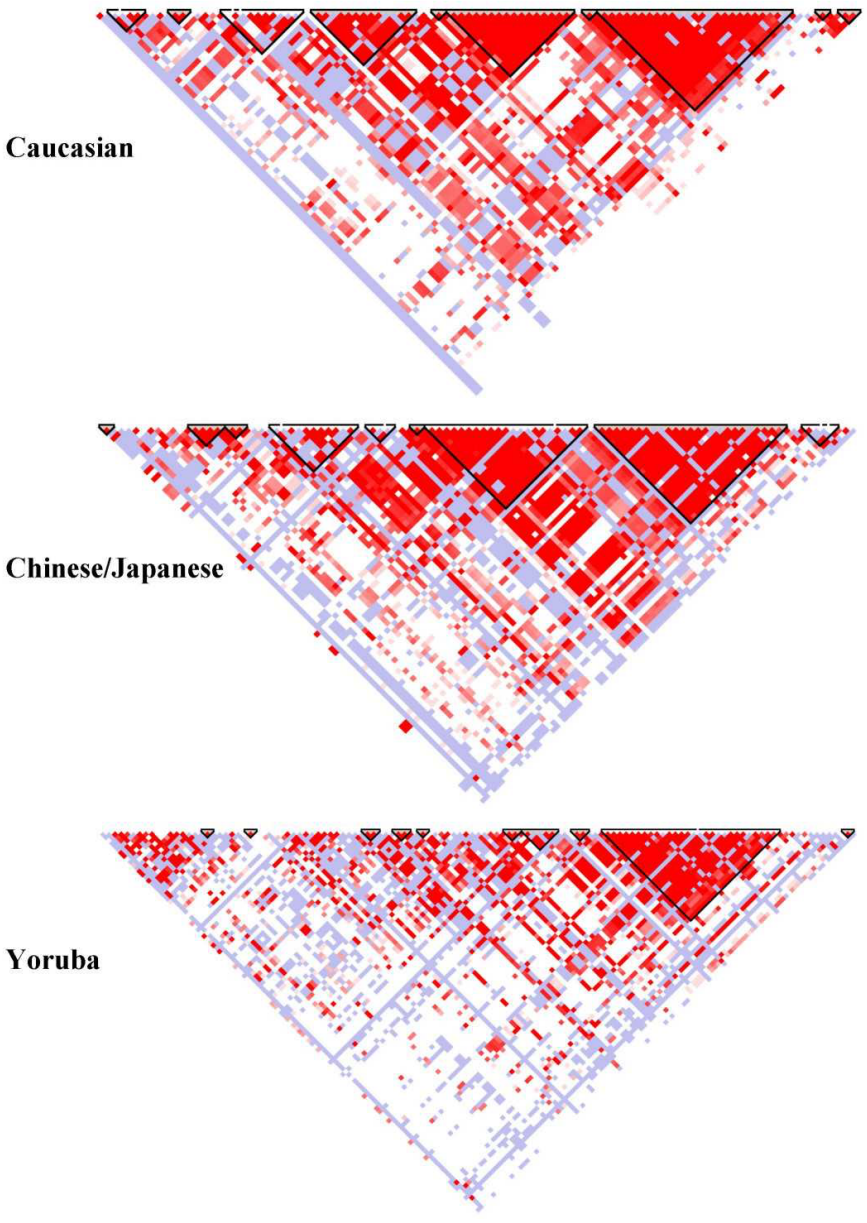
- Recombination (and lack thereof) along the genome causes some sites to be more correlated to each other than others
- Linkage disequilibrium (LD) is the non-random association of alleles at different loci along a genome



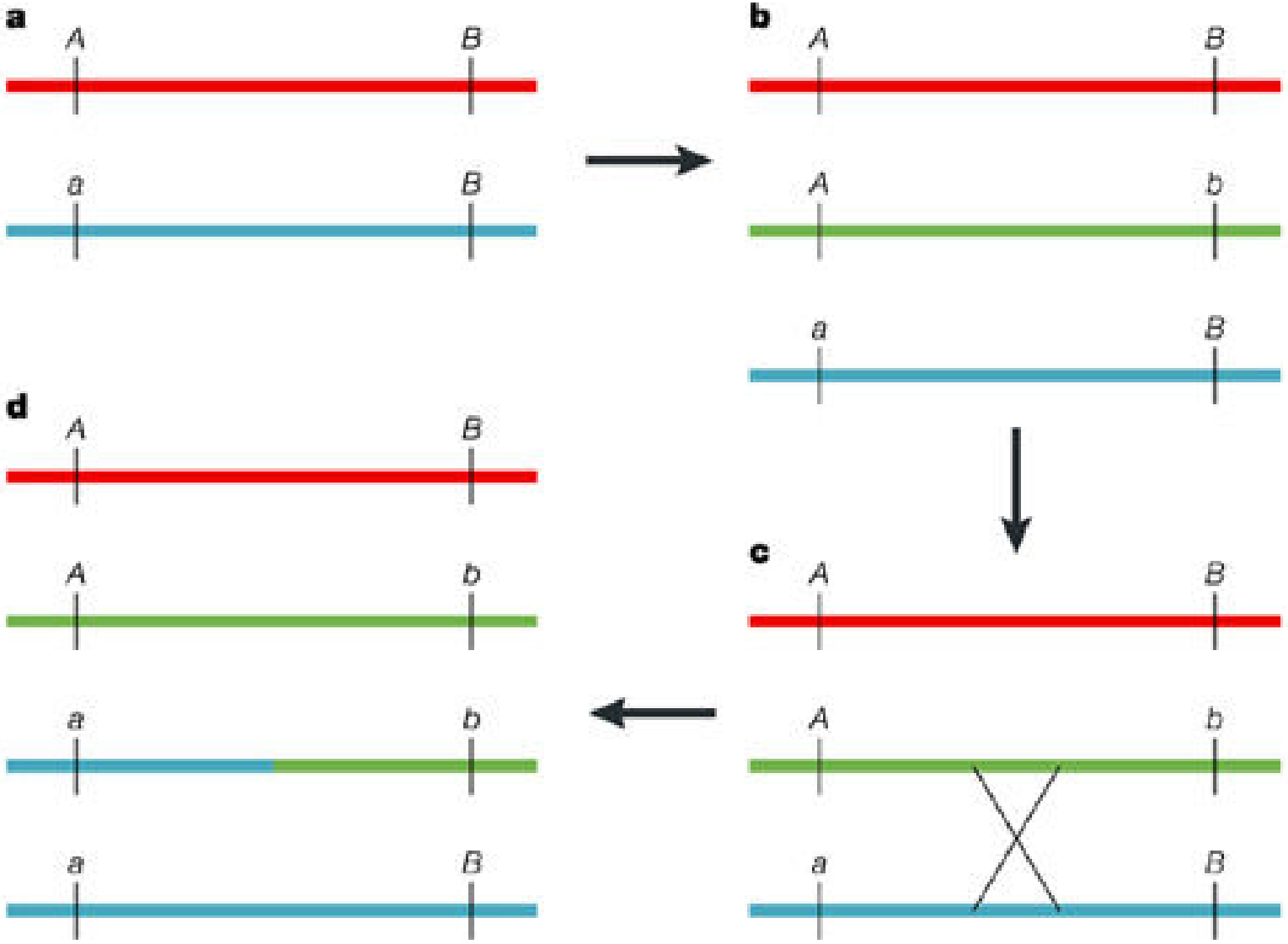
Linkage disequilibrium

- The LD structure in a given region will likely be different in different populations
- LD increases genome-wide as a consequence of bottlenecks
- LD also can increase locally as a consequence of a selective sweep

Linkage disequilibrium



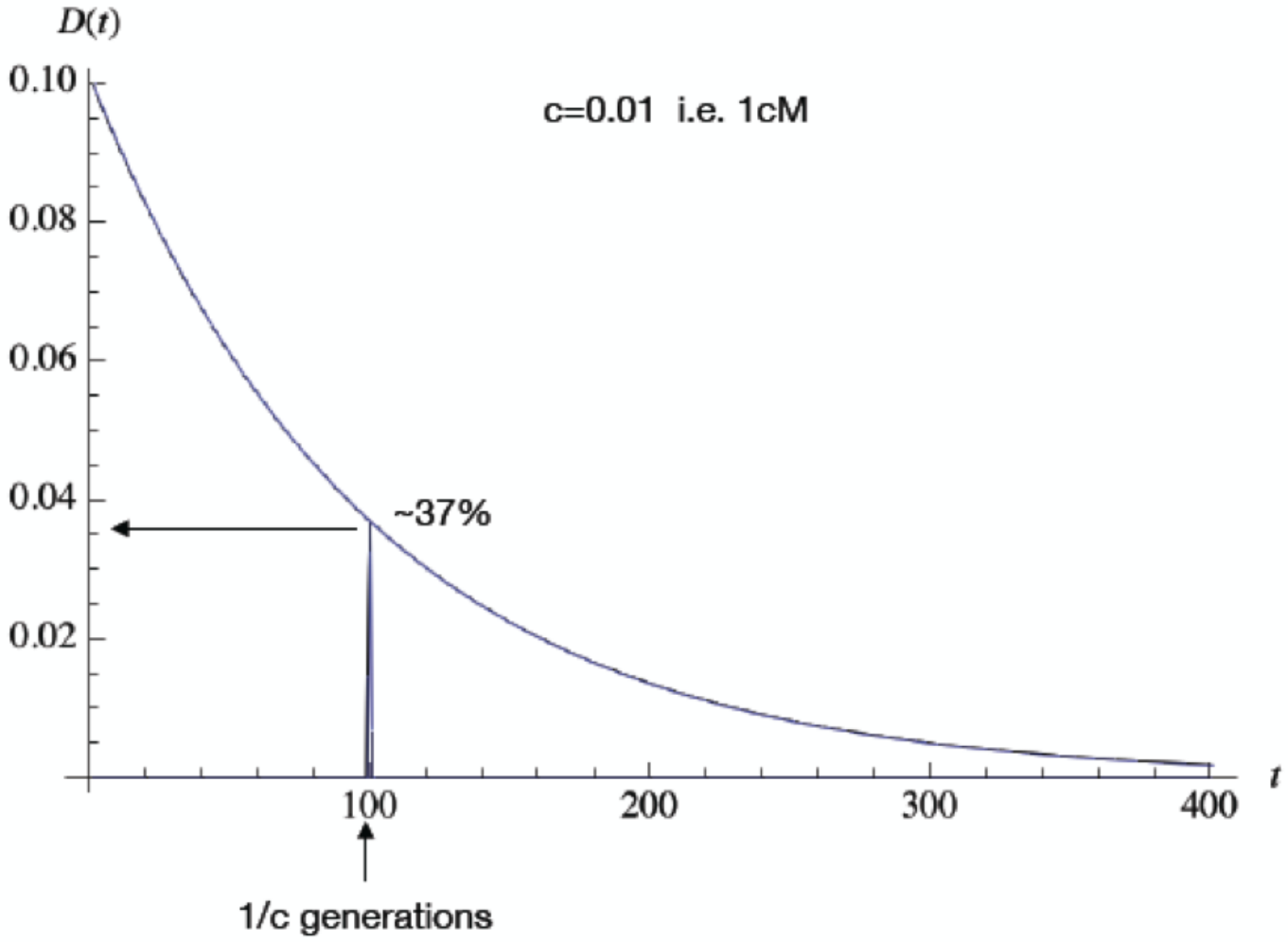
Measuring linkage disequilibrum



Measuring linkage disequilibrum

- D measures the difference between the observed value of f_{AB} and its expectation assuming independence between the two loci:
  $D = f_{AB} - f_A f_B$
- D can also be computed using all observed genotypes:
  $D = f_{AB} f_{aa} - f_{Ab} f_{aB}$
- D decays forwards in time as a function of the recombination rate c:
  $D(t) = (1 - c)^t D(0)$

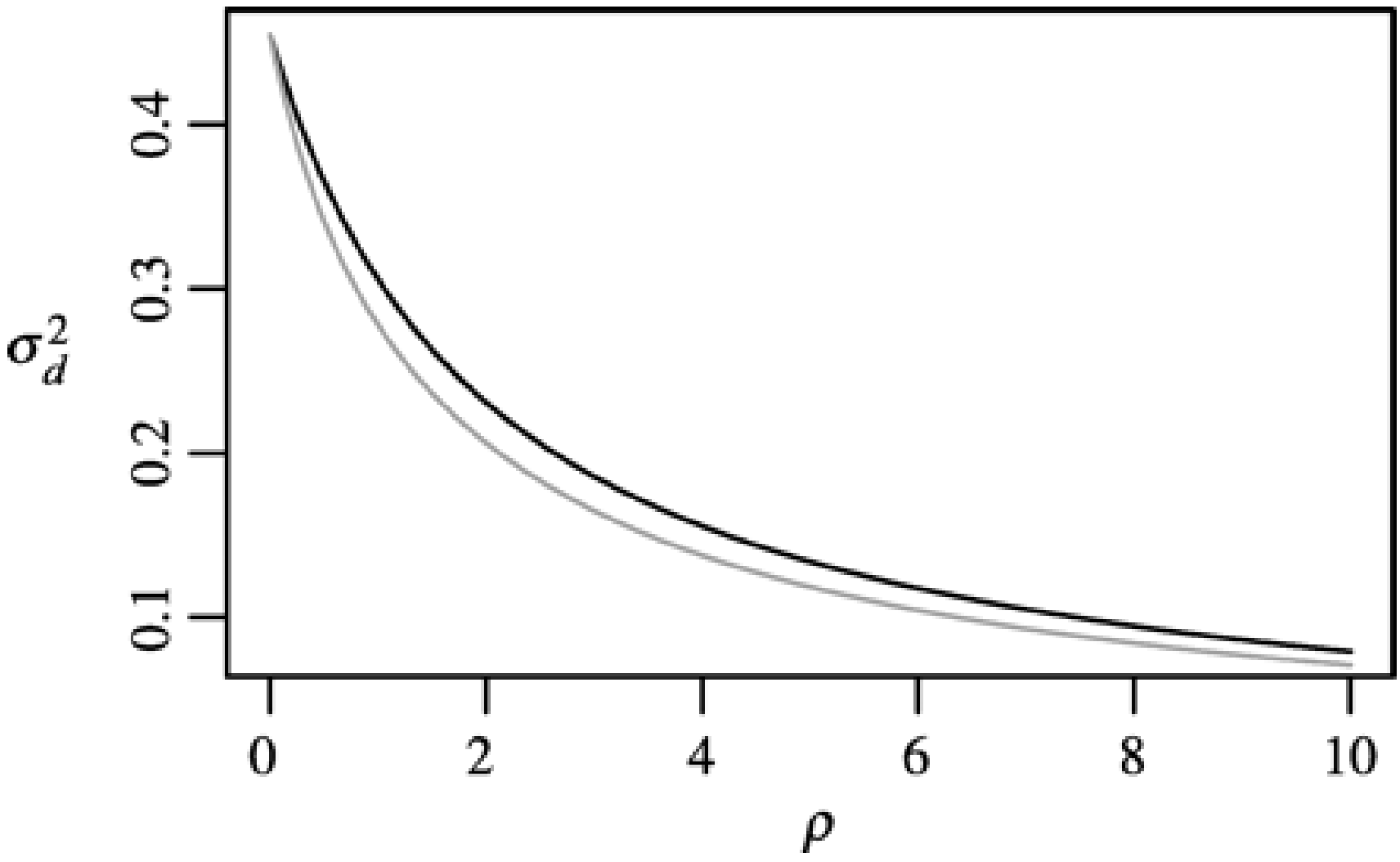
LD decay over time



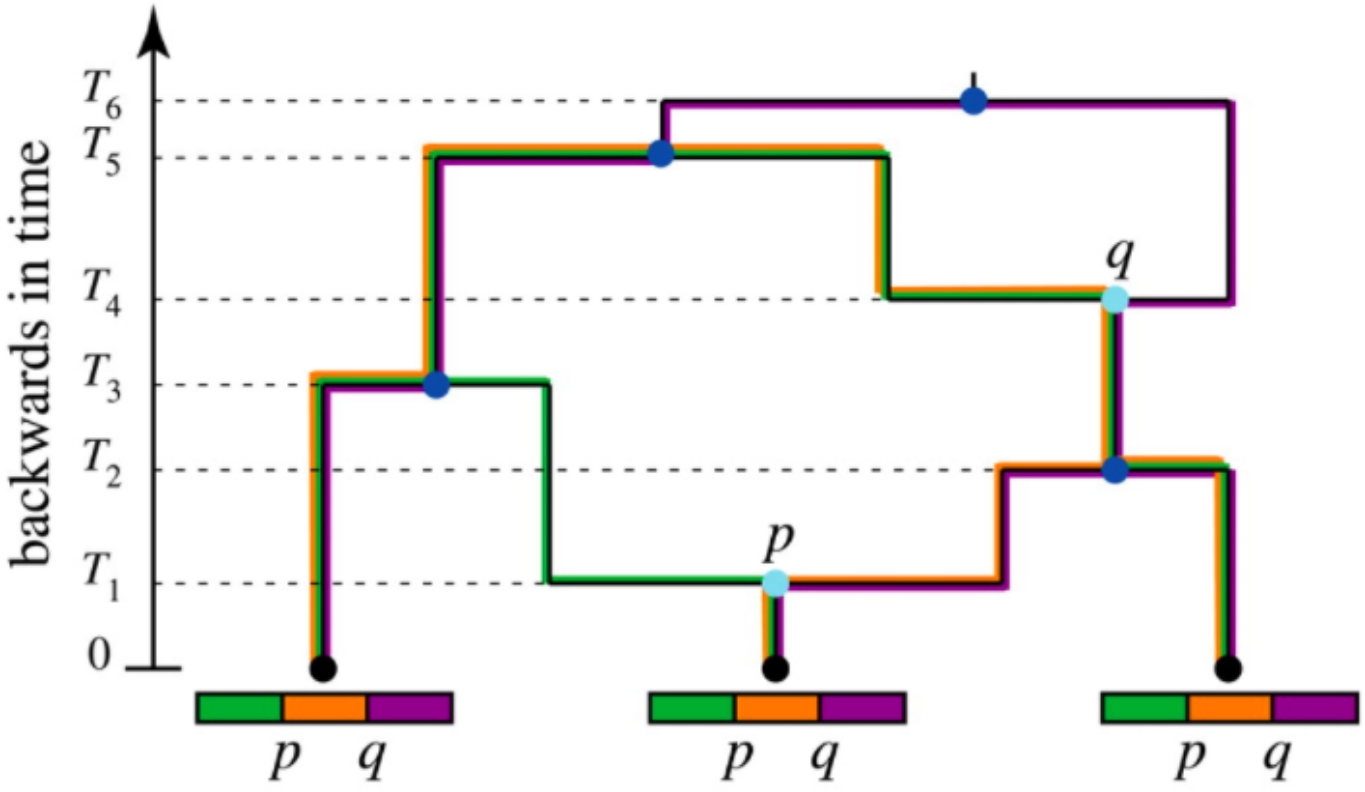
Measuring linkage disequilibrum

- An alternative measure of LD is: $r^2 = \frac{D^2}{f_A f_a f_B f_b}$
- The expectation of r^2 can be approximately expressed as a ratio of expectations:
  $E[r^2] \approx \frac{E[D^2]}{E[f_A f_a f_B f_b]}$
- Using this result, one can show (see Wakeley book) that $E[r^2]$ is approximately a function of the population-scaled recombination rate parameter ρ:
  $E[r^2] \approx \frac{\rho+10}{\rho^2+13\rho+22}$
- Thus, we can relate an **observation** (r^2) with a **parameter** in our model (ρ). This can allow us to infer the parameter or check if the neutral model is a good fit.

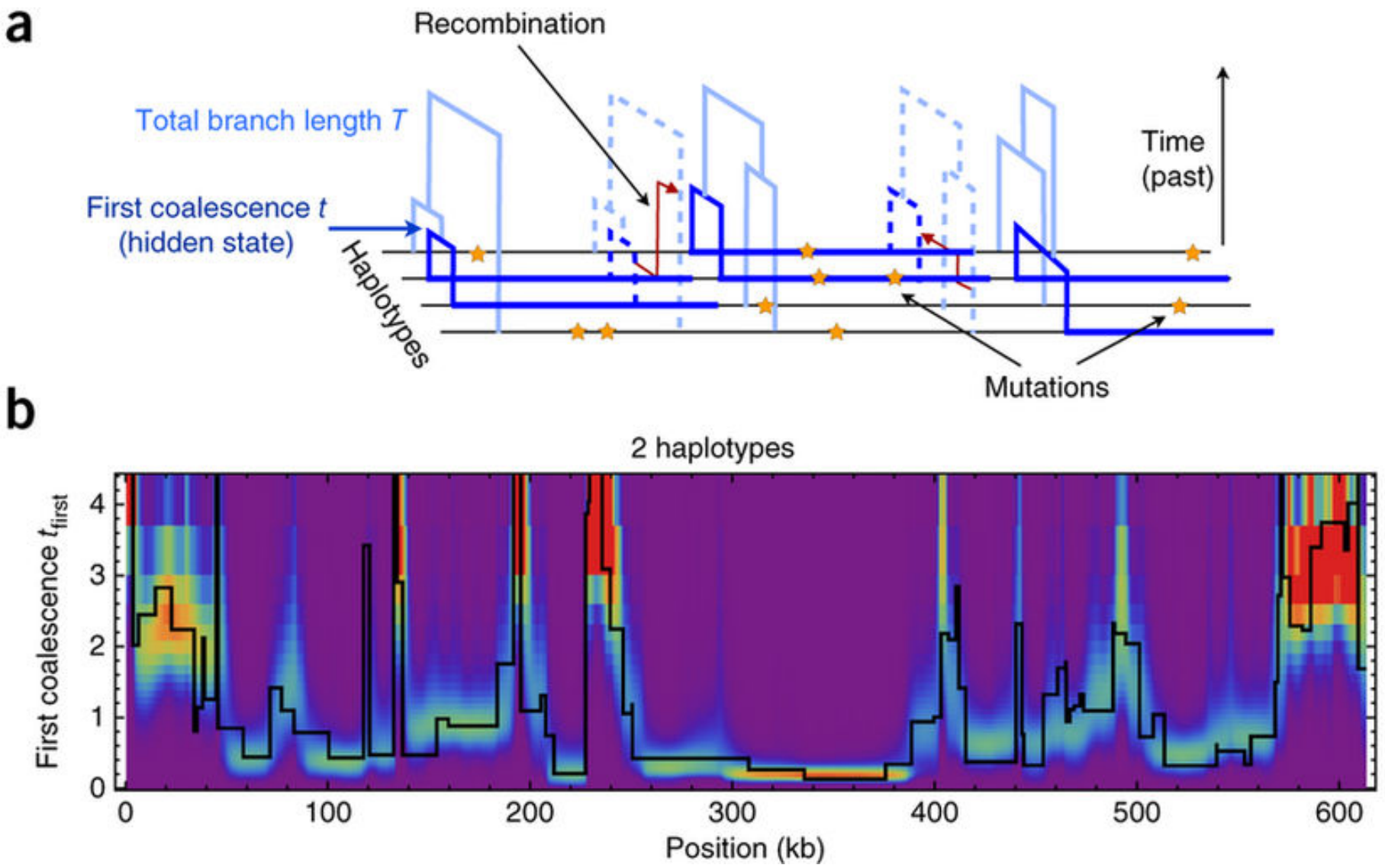
Measuring linkage disequilibrum



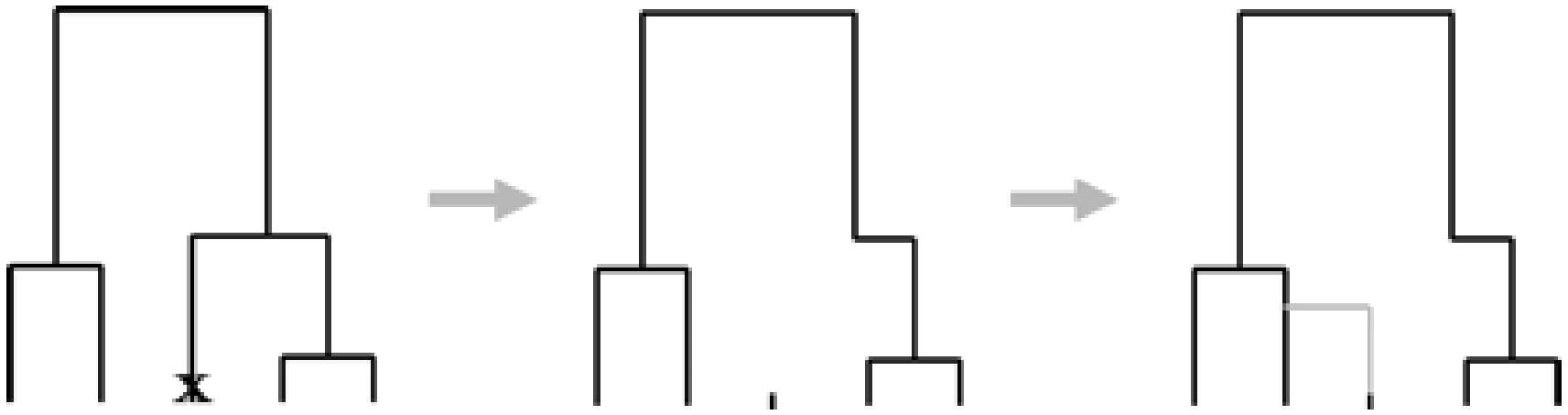
SMC: a Markovian approximation to the ARG



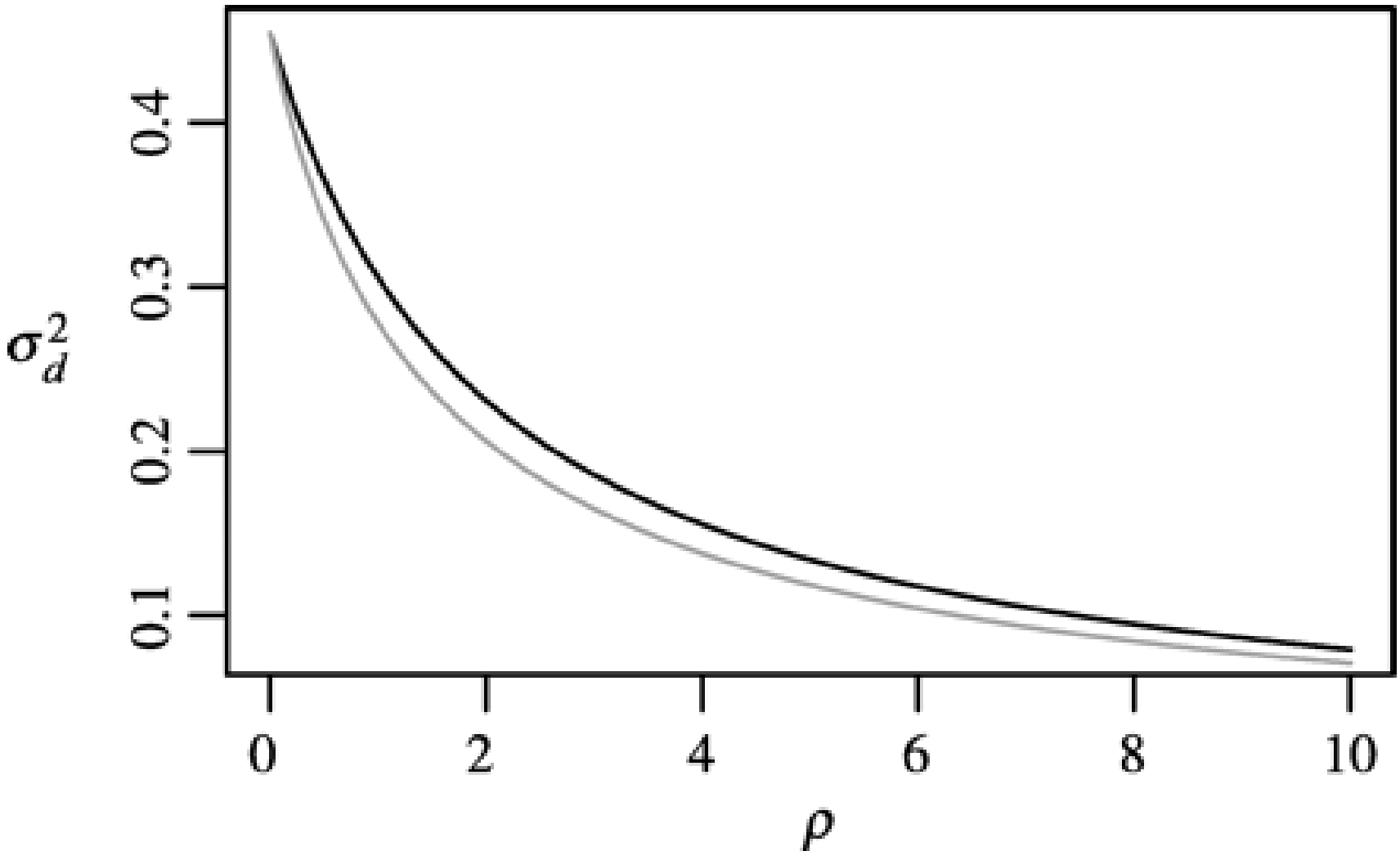
SMC: a Markovian approximation to the ARG



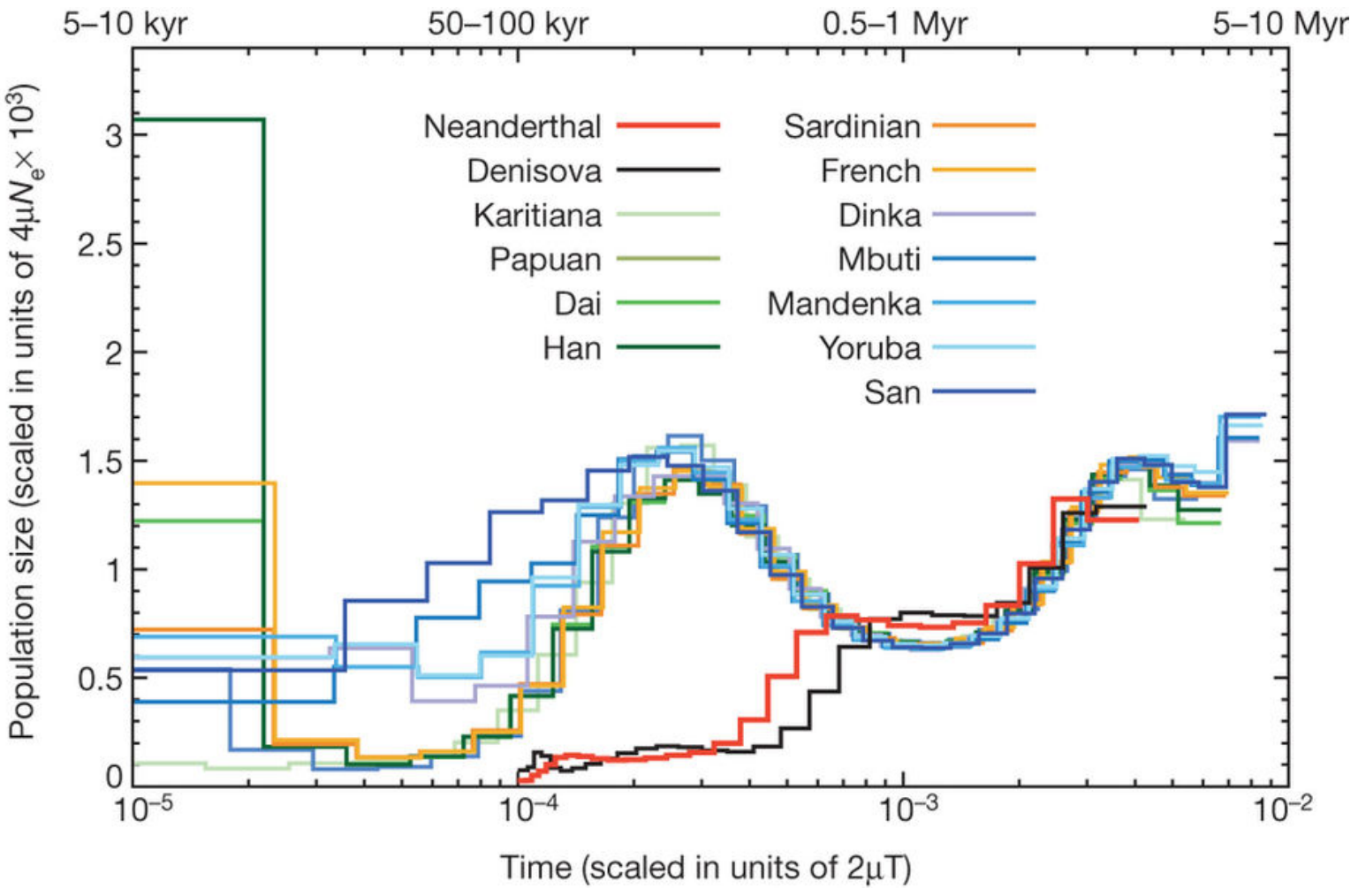
SMC: a Markovian approximation to the ARG



SMC: a Markovian approximation to the ARG



PSMC


5–10 kyr
50–100 kyr
0.5–1 Myr
5–10 Myr
Neanderthal
Denisova
Karitiana
Papuan
Dai
Han
Sardinian
French
Dinka
Mbuti
Mandenka
Yoruba
San
Population size (scaled in units of 4μN_e × 10^3)
0
0.5
1
1.5
2
2.5
3
10^{-5}
10^{-4}
10^{-3}
10^{-2}
Time (scaled in units of 2μT)

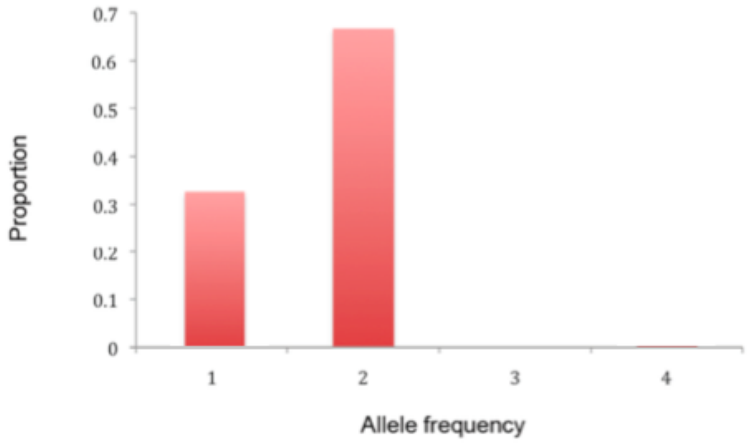
Today

- Recombination
- **The site frequency spectrum**

Site-frequency spectrum (SFS)

- Like π and θ_W, the SFS is a summary of population genetic data.
- Each bin represents the proportion of sites with a particular minor allele frequency.

```
Sequence 1 aggtatgcta gaaccctaga aagacacaga gatagacaag
Sequence 2 aggtatgcta gaaacctaga tagacacaga gatagacaag
Sequence 3 aggtatgcta gaaacctaga tagacacaga gatagacaag
Sequence 4 aggtatgctg gaaccctaga tagacacaga gatagacaag
Sequence 5 aggtatgctg gaaccctaga tagacacaga gatagacaag
```



Folded vs. Unfolded SFS

- To build an unfolded SFS, we need knowledge of the ancestral state of each SNP.
- Each bin represents the proportionf of sites with a particular **derived** allele frequency.

Ancestral sequence a c a

```
Sequence 1 aggtatgcta gaaccctaga aagacacaga gatagacaag
Sequence 2 aggtatgcta gaaacctaga tagacacaga gatagacaag
Sequence 3 aggtatgcta gaaacctaga tagacacaga gatagacaag
Sequence 4 aggtatgctg gaaccctaga tagacacaga gatagacaag
Sequence 5 aggtatgctg gaaccctaga tagacacaga gatagacaag
```

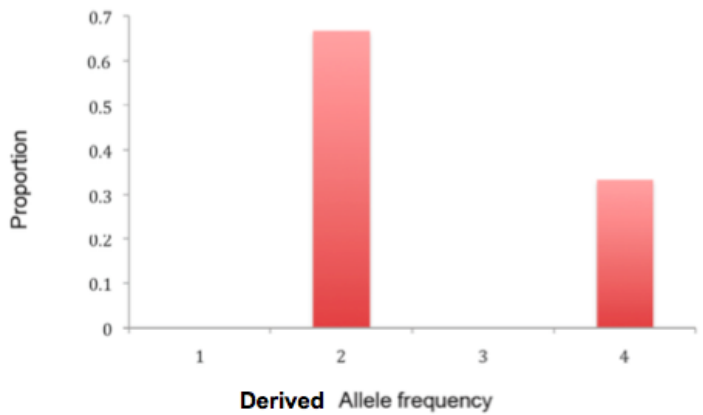
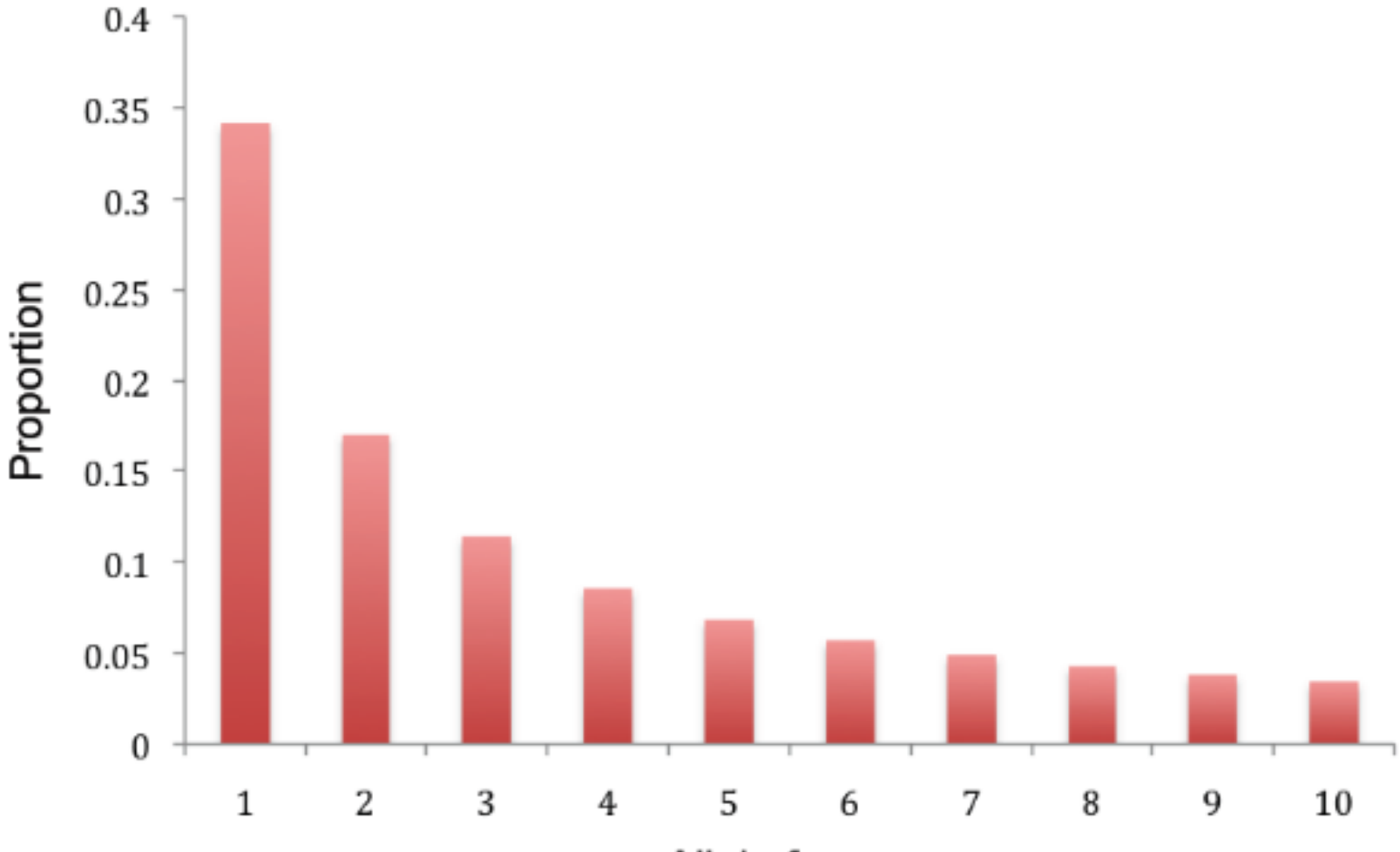


Figure 3.5. The Site Frequency Spectrum (SFS) for the DNA sequence data example from the Infinite sites model section.

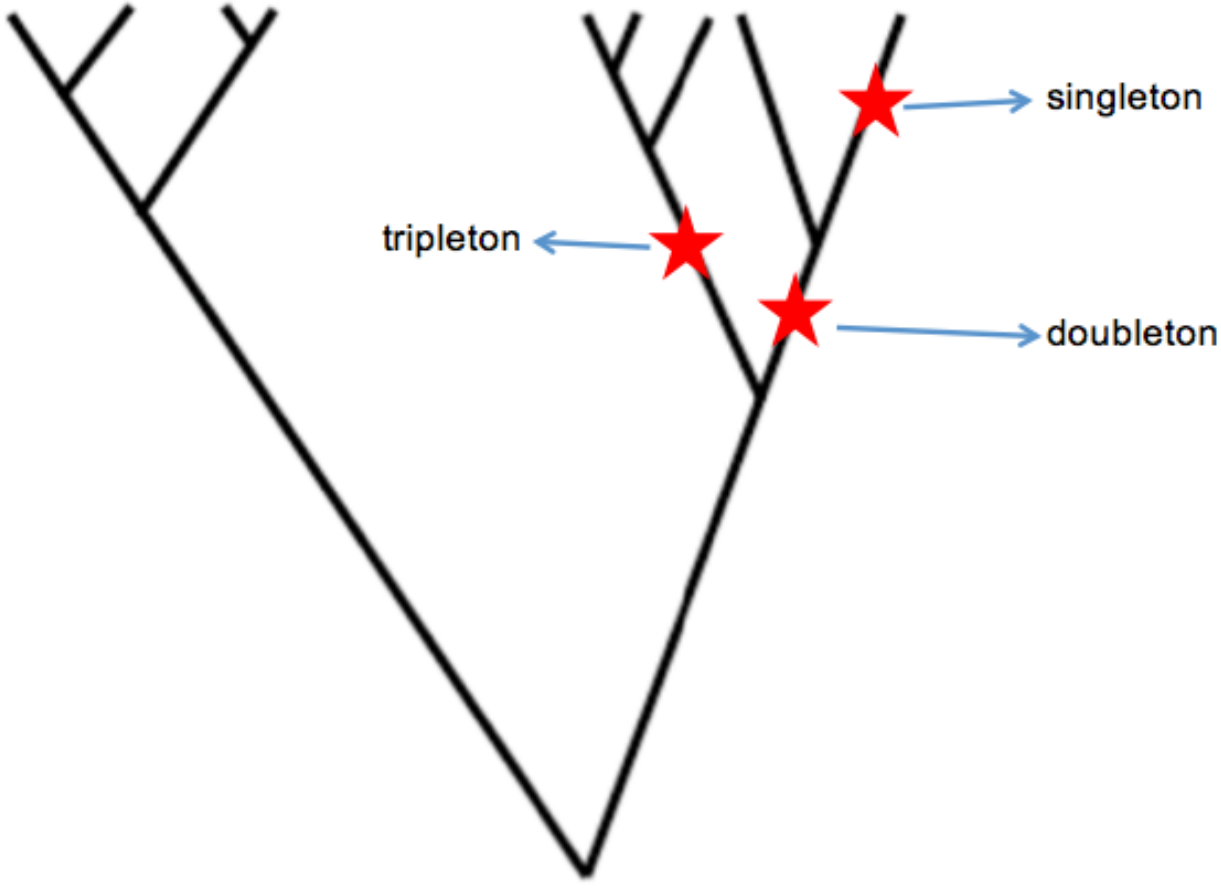
Site-frequency spectrum (SFS)

- SFS under the neutral coalescent model for a sample of n=11 haploid individuals.



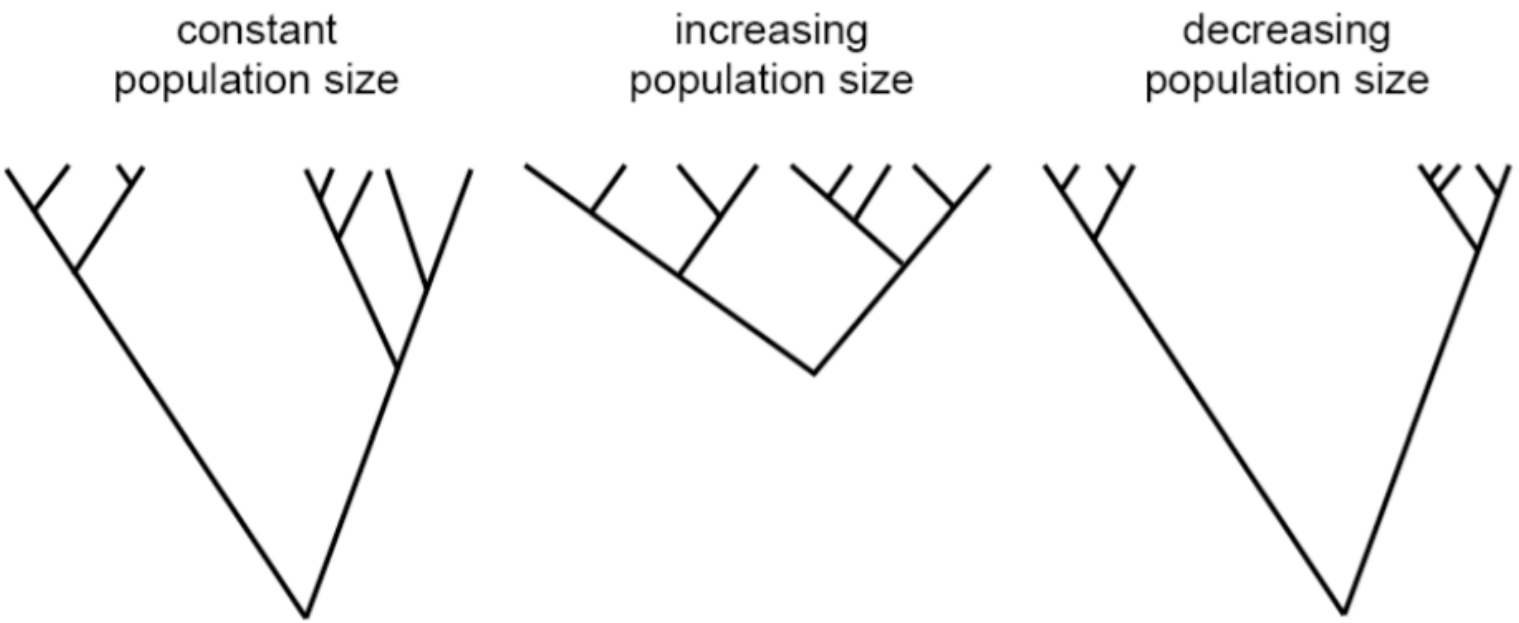
SFS and coalescent trees

- The shape of the SFS depends on the underlying coalescent trees.

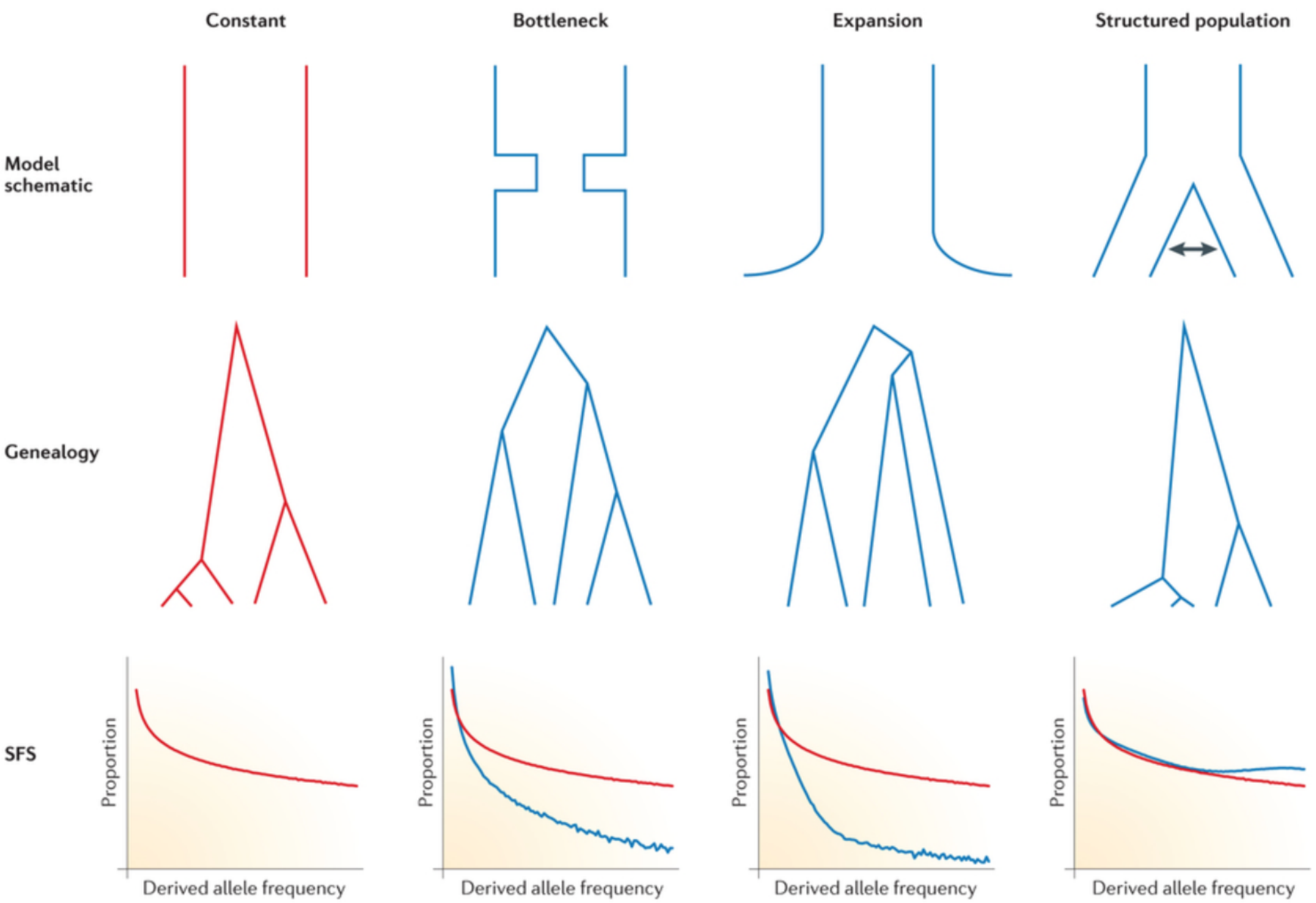


SFS and coalescent trees

- In turn, the coalescent trees along the genome depend on the population demography.
- Thus, we can use the SFS to learn about past demography.



SFS for different demographic models

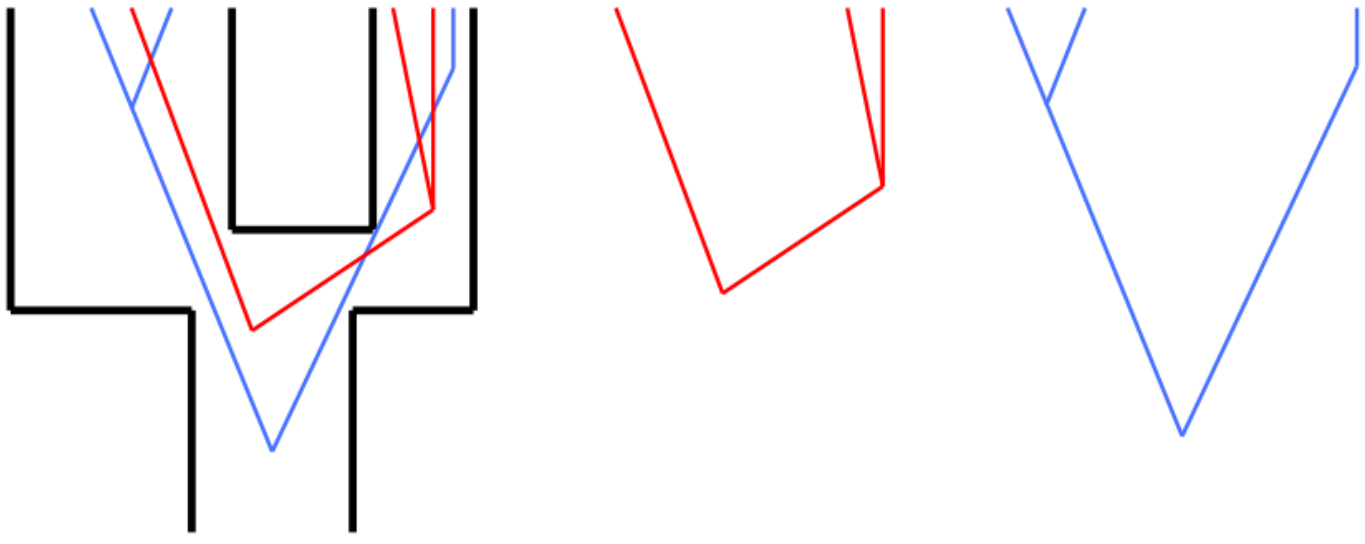


Some complications

- Coalescent trees are different in different parts of the genome (due to recombination).

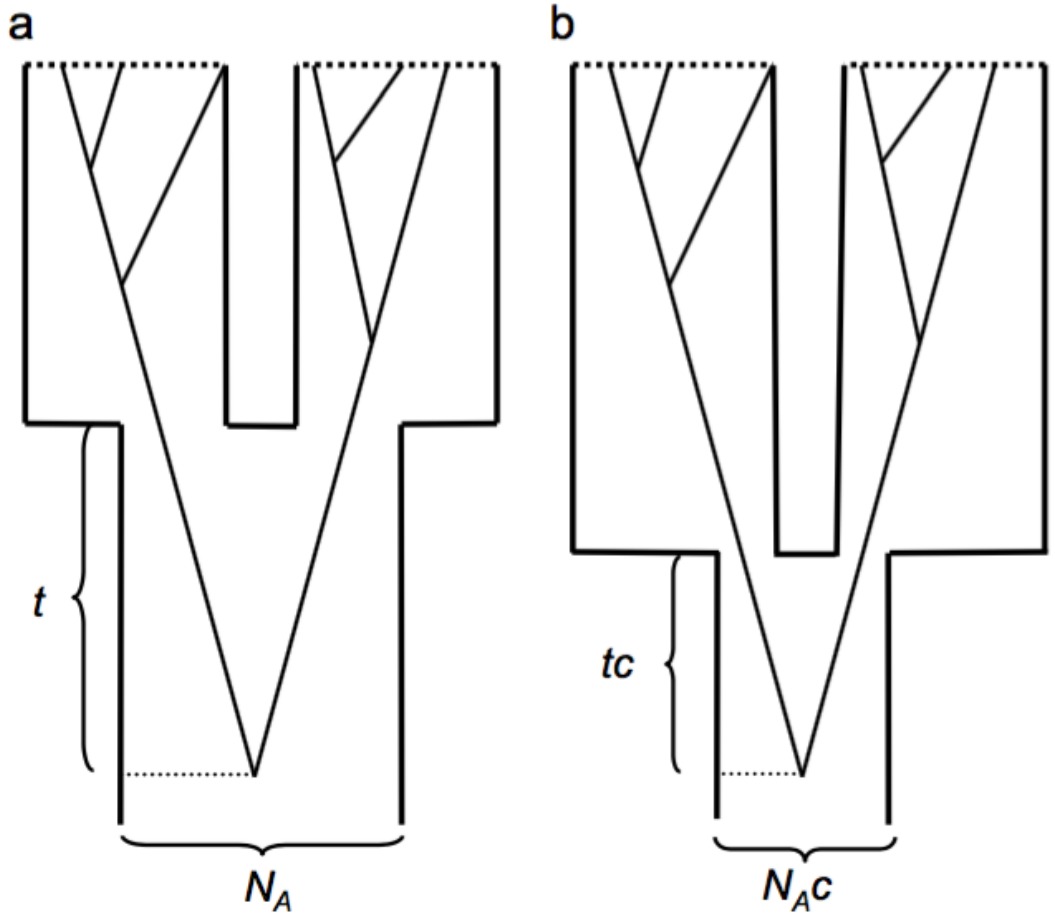
Some complications

- The same demography can randomly produce different trees.



Some complications

- Different demographies can produce the same tree.

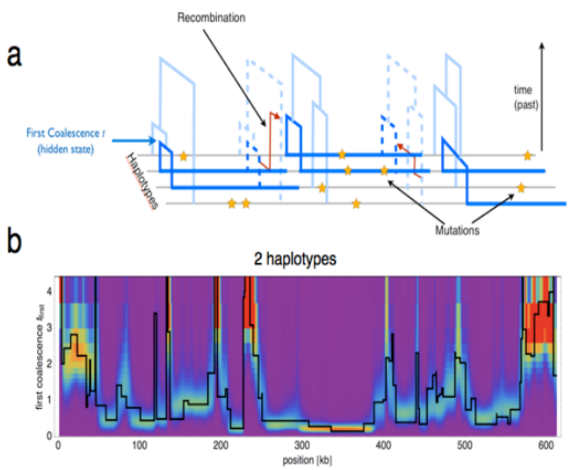


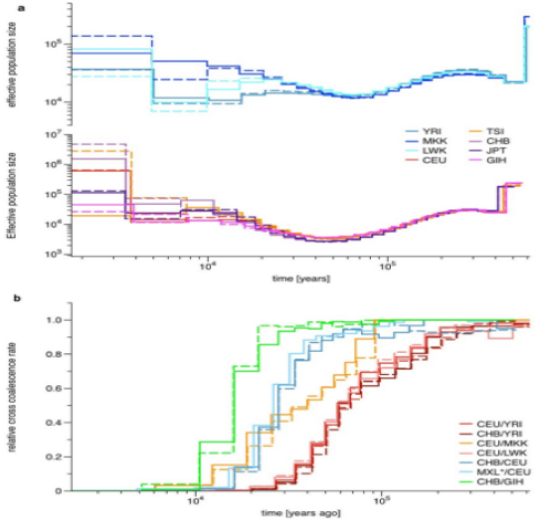
Demographic inference

- A demographic history does not determinisitcally produce a single tree, and coalescent trees do not deterministically produce a single SFS.
- However, some types of data are more likely under certain trees, and some trees are more likely under certain historical models.
- Important probability distributions: $P[data|tree]$ and $P[tree|history]$

Demographic inference

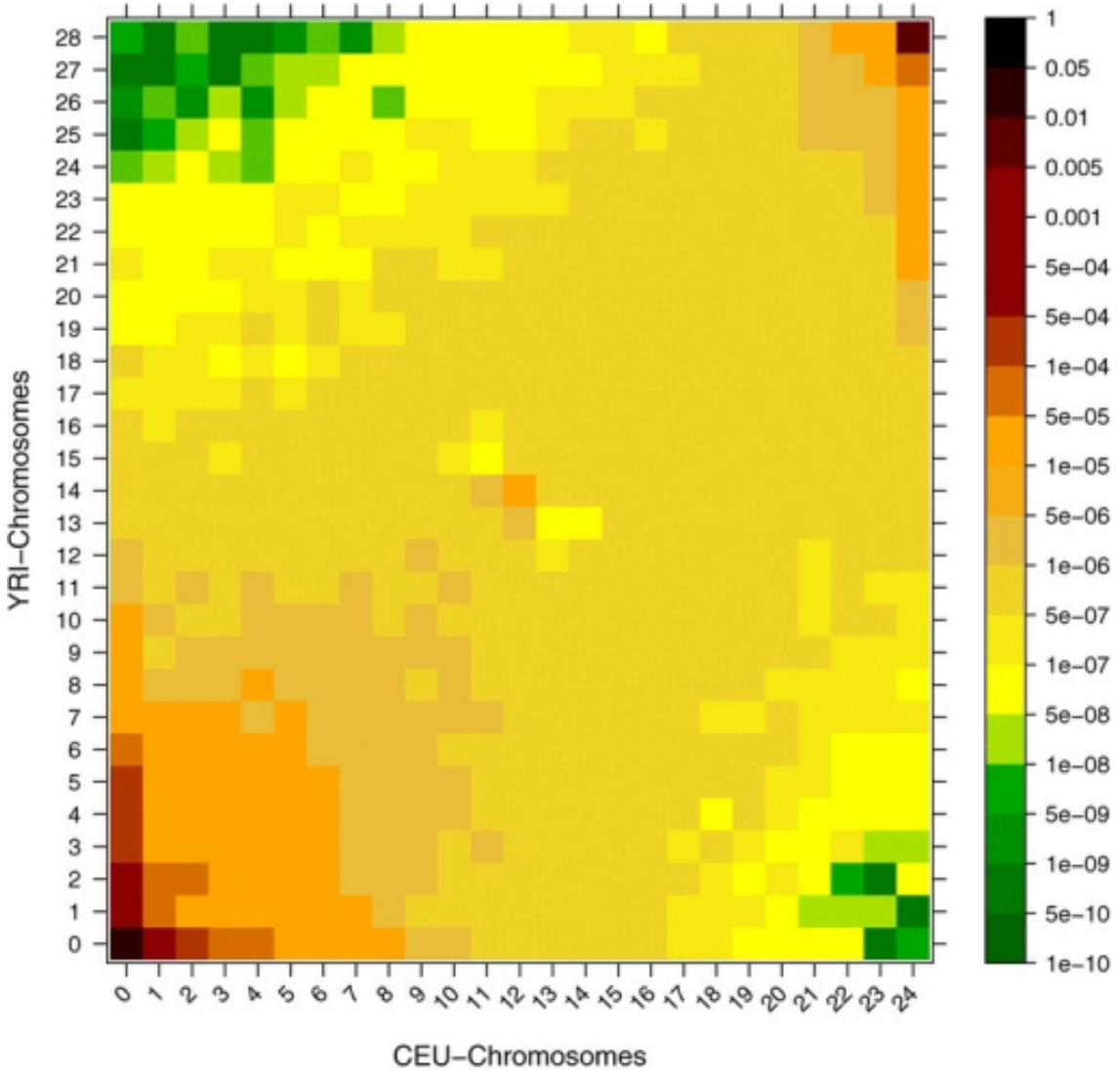
- To get at the likelihood of the data from the demographic history, we use Felsenstein's equation:
- $P[data|history] = \sum_{tree\ space} P[data|tree]P[tree|history]$
- In general, very hard to compute over the entire genome (use of approximations).
- This is the basis of many demographic inference programs: fastsimcoal2, MSMC, PSMC, etc.



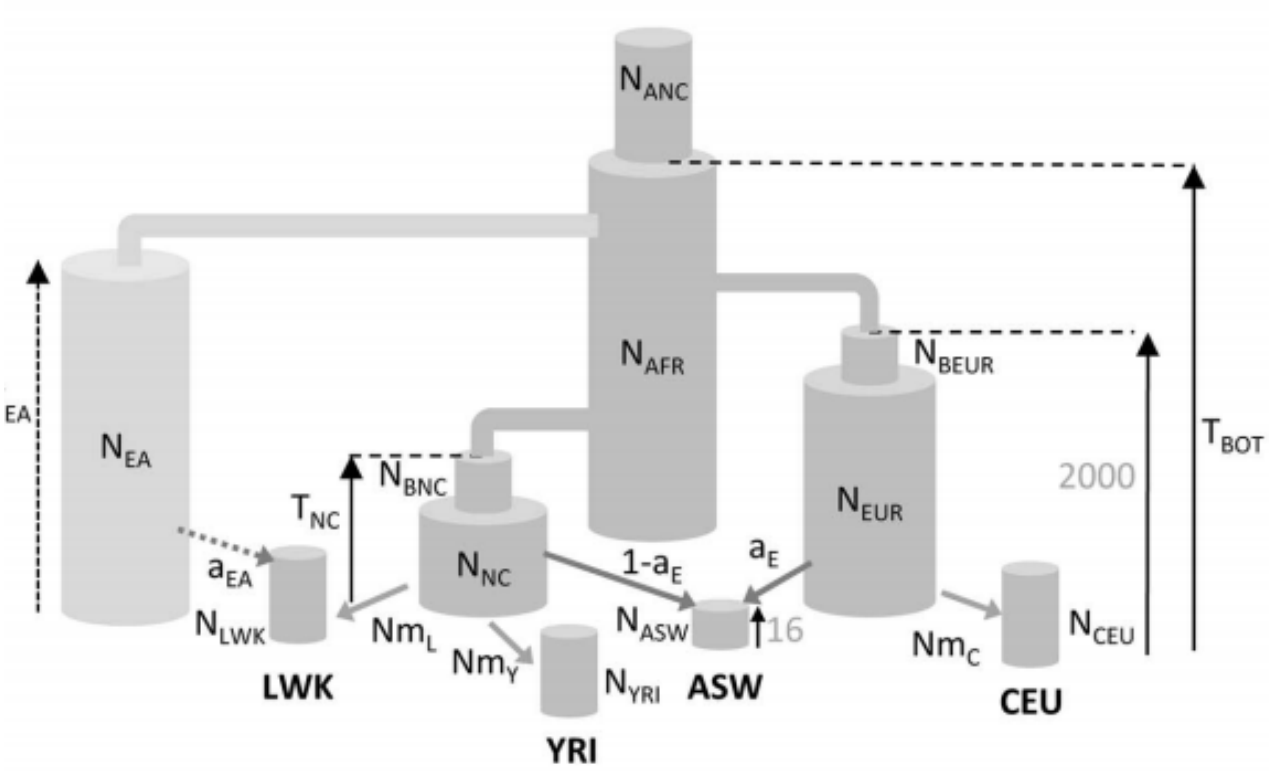


Two-dimensional spectrum

- We can also compute the joint SFS for two (or more) populations.



Demographic inference



Exercises today

- https://github.com/FerRacimo/DemographicCourseAdelaide2018/blob/master/PopSizeTutorial.md