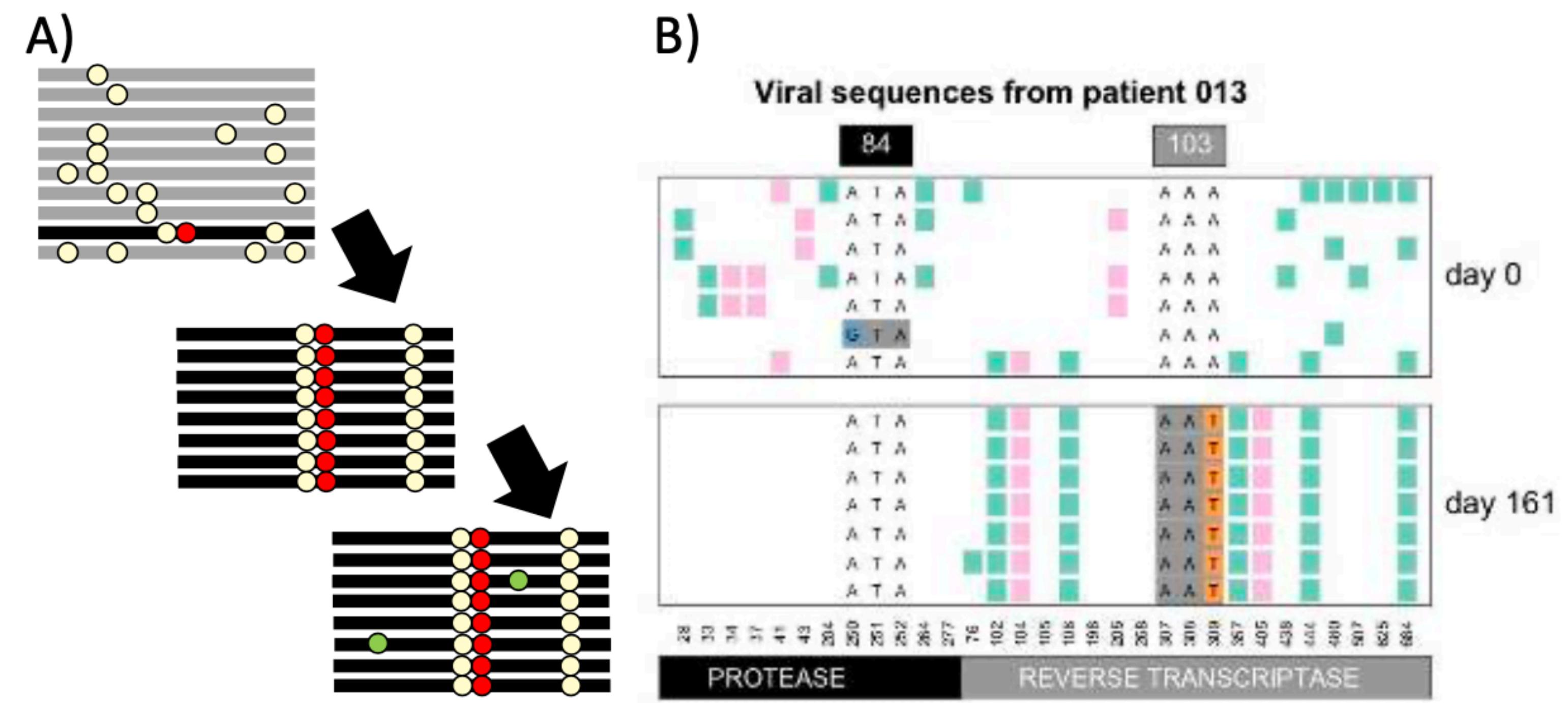


Lecture 14: The Effects of Linked Selection (Ch13)

Population genetic PCB4553/6685 3/19/2023

Hitchhiking of neutral alleles due to selective sweeps



Hitchhiking under complete linkage

- Expected time to fixation for a selected mutation:
 - $\tau = 4 \log(2N)/s$ (Chapter 10)
- Alleles at loci fully linked to the selected locus must coalesce τ generations back
- Usually τ much smaller than $2N$
- Reduced genetic diversity

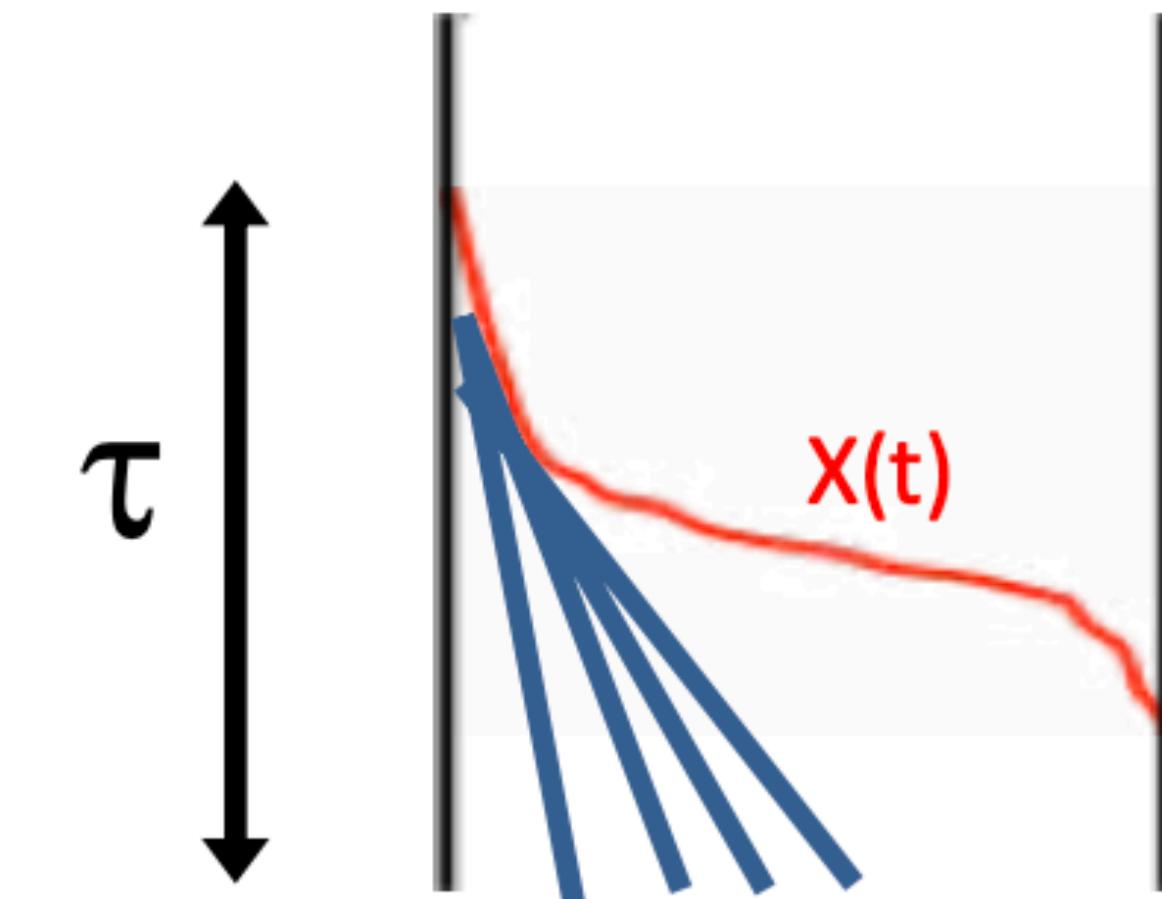


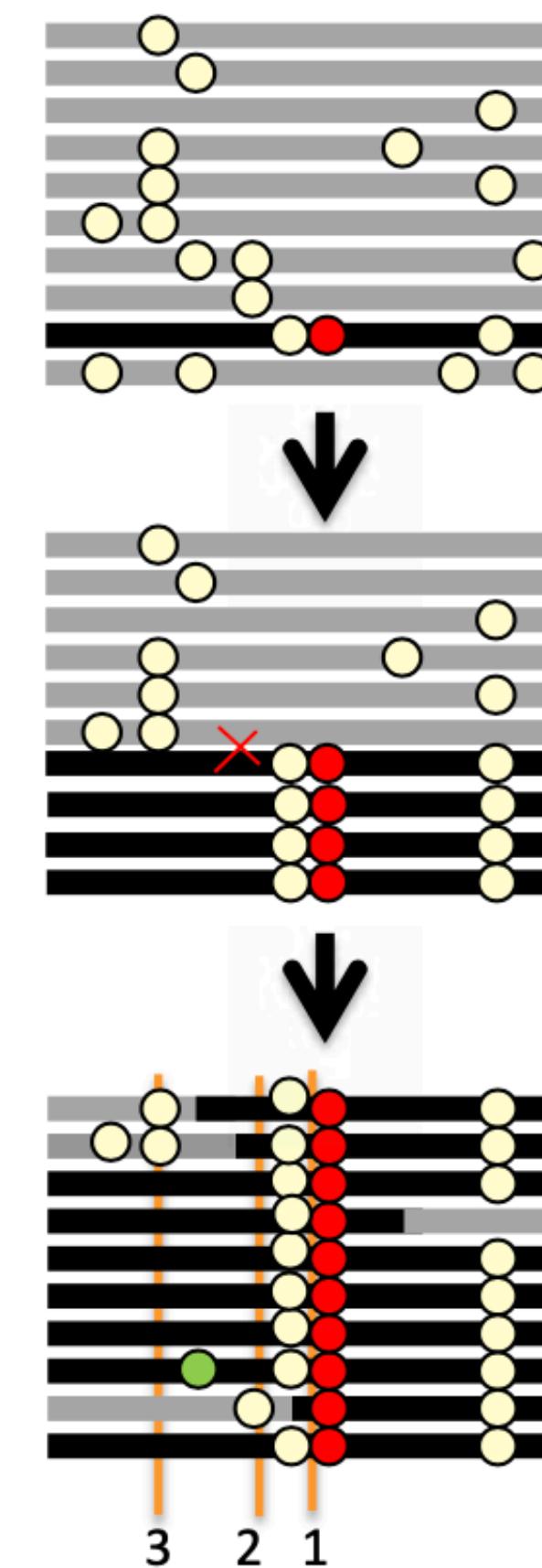
Figure 13.2: The coalescent of 4 lineages, marked in blue, at a locus completed linked to our selected allele. The frequency trajectory of the selected allele $X(t)$ is shown in red.

Hitchhiking for completely unlinked loci?

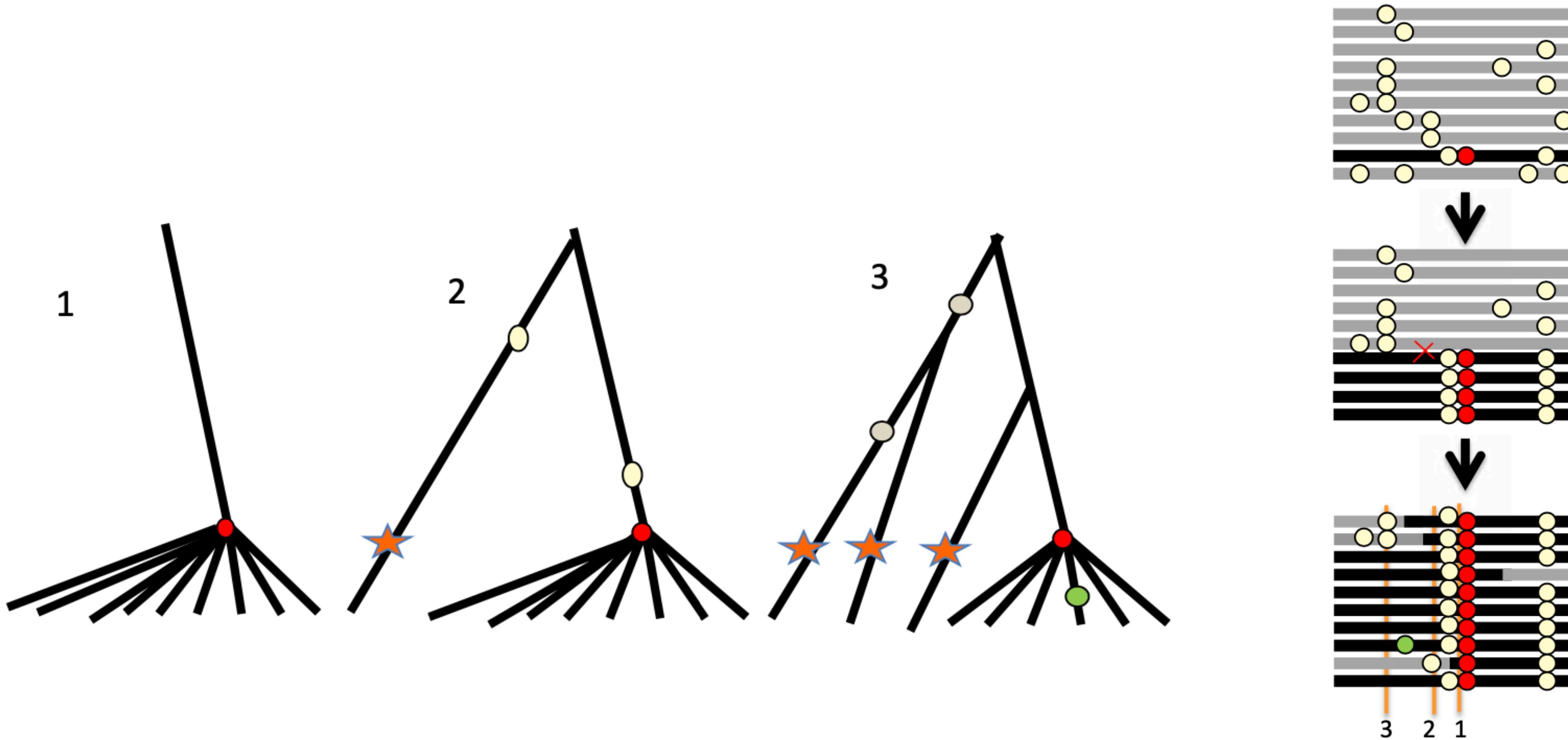
- All alleles have equal probability of associating with the selected allele
- Allele frequency does not change
- No hitchhiking

Sweep in recombining regions (partial linkage)

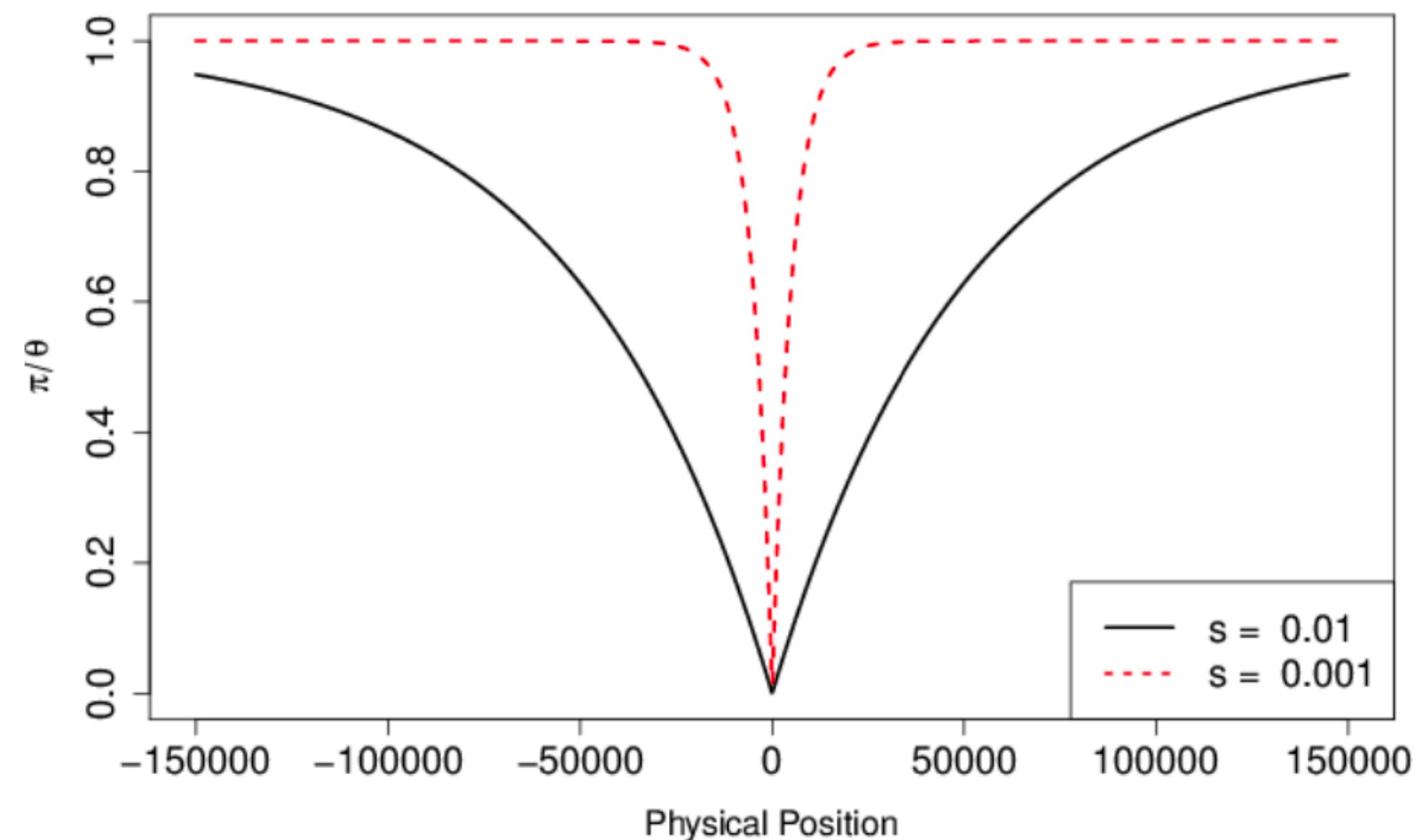
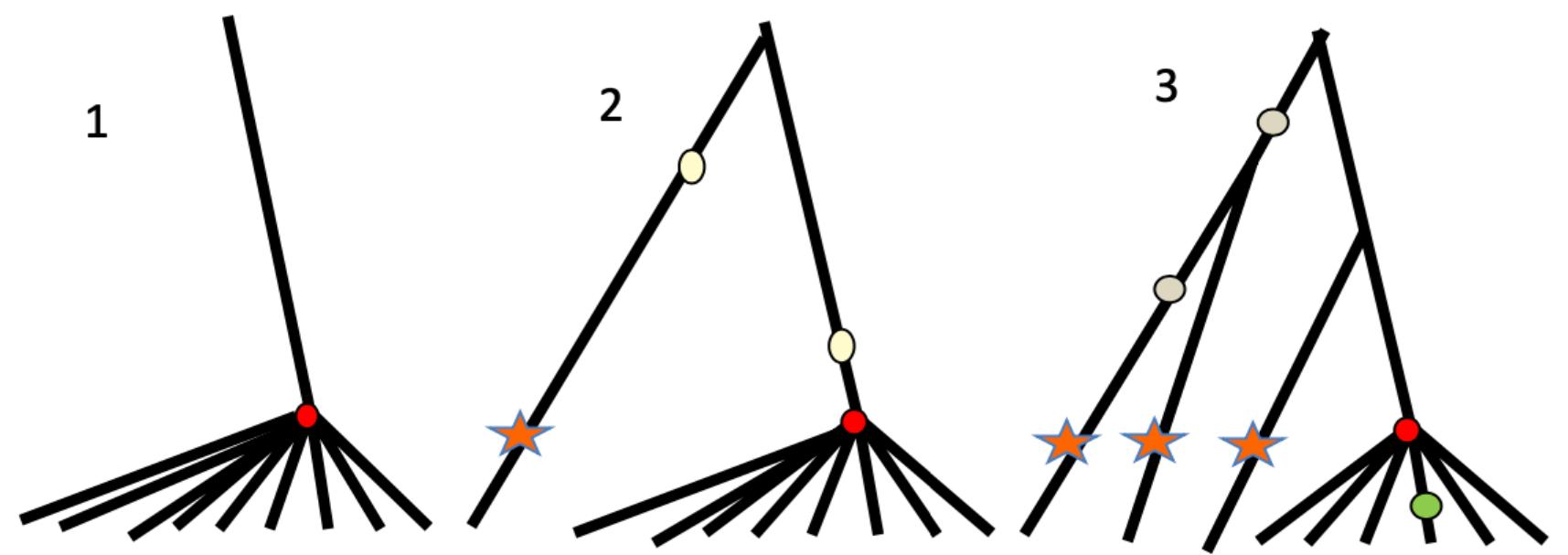
- Selected mutation arises on a particular haplotype, which increases in frequency in the pop
- Recombination events can occur between haplotypes carrying and not carrying the selected allele, in individuals who are heterozygote for the selected allele.
 - alleles that were not present on the original selected haplotype avoid being swept out
 - alleles originally linked to selected locus fails hitchhiking all the way to fixation.



Coalescent genealogies at different distance to the selected locus

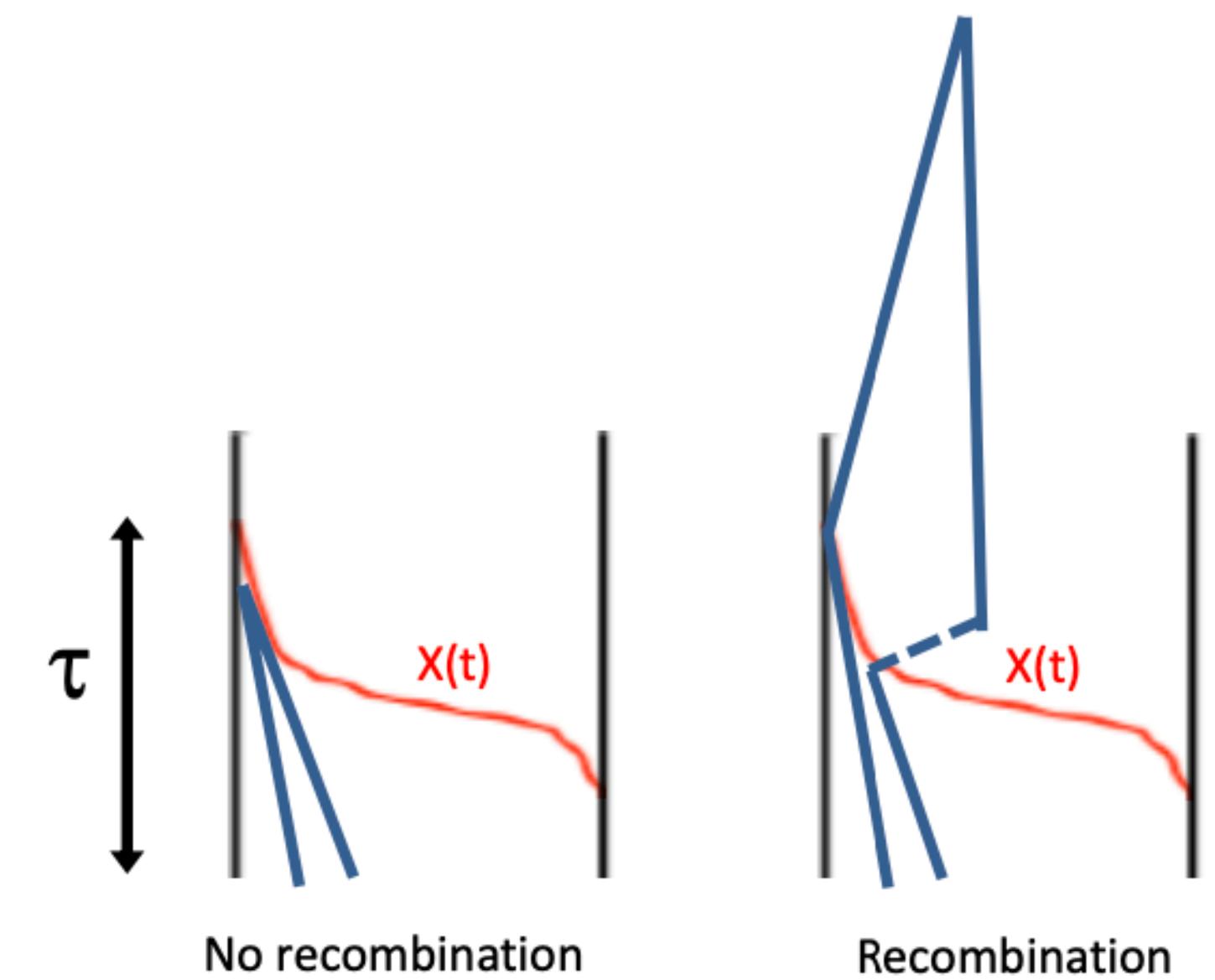


- As one move even further out from the selected site, more lineages descended from recombinant haplotypes that coalesce neutrally much deeper in time than τ
- allowing diversity to recover to background levels as we move away from the selected site

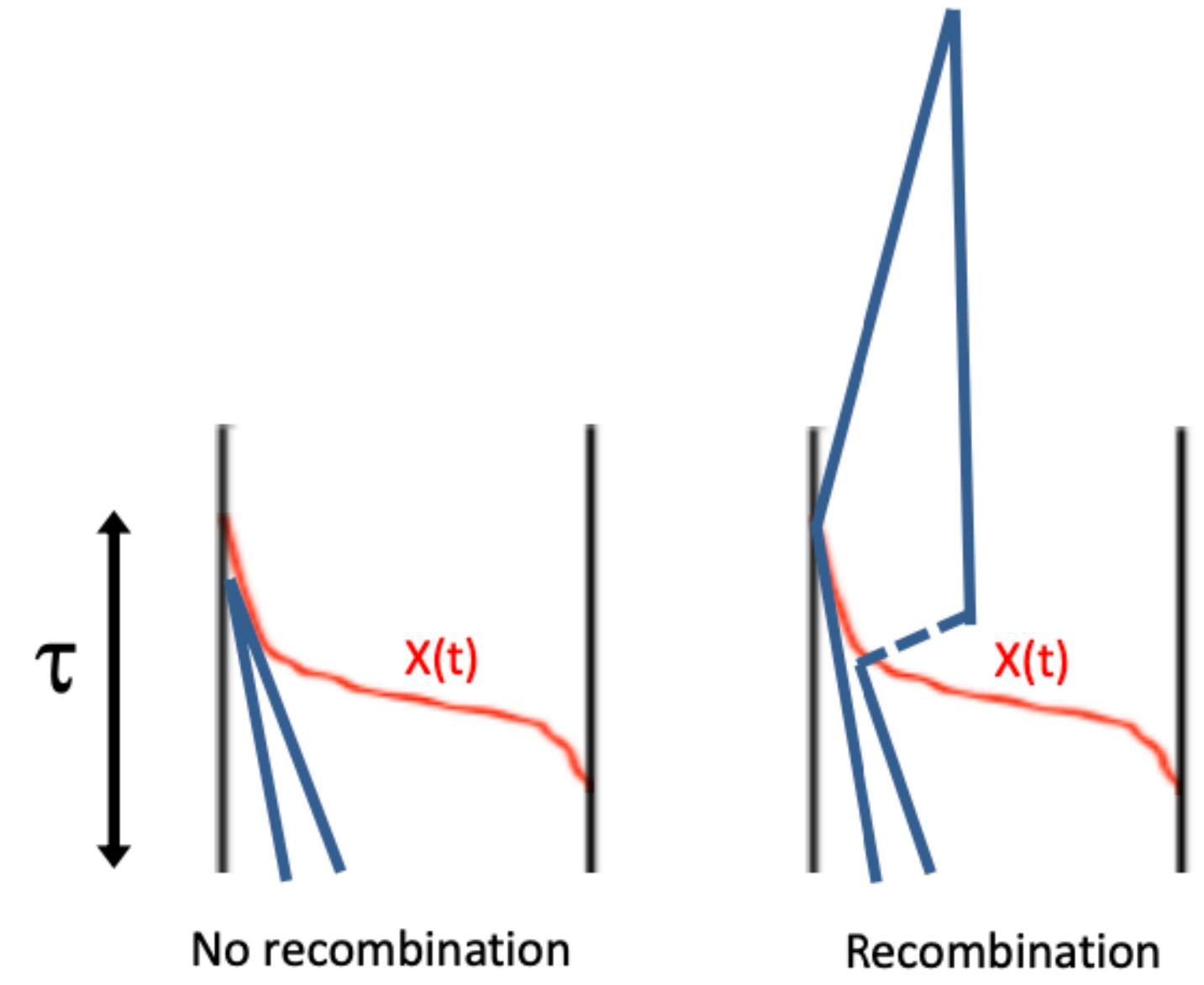


Quantitative model

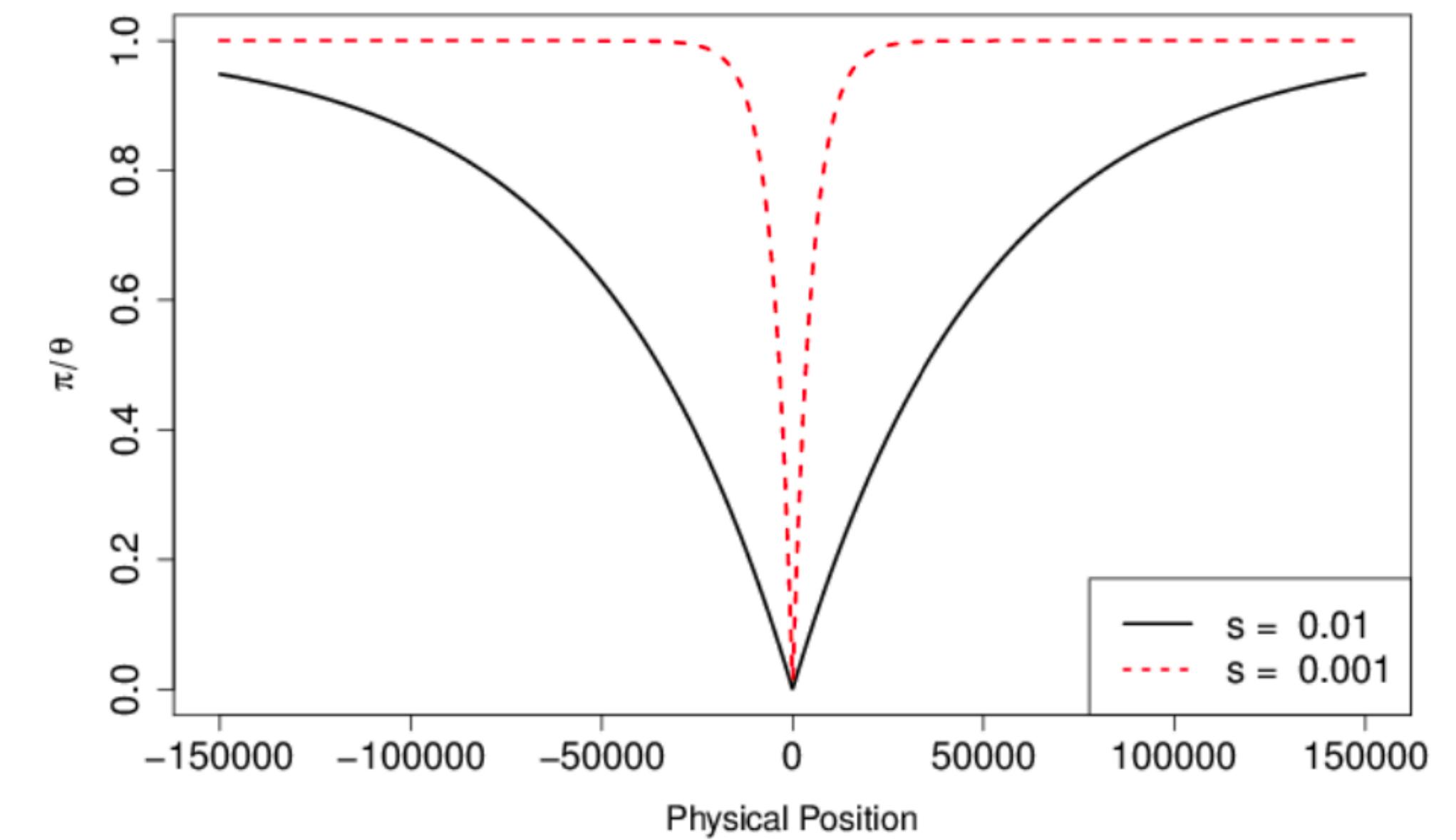
- What is the coalescent time for a pair of alleles at locus partially linked to the selected locus?
- probability no recombination occurs
 - $(1 - c)^\tau \approx e^{-c\tau}$
- Prob an allele does not descend from a recombinant haplotype
- $p_{NR} = e^{-c\tau/2}$
- Why 1/2?
- The neutral allele spent 1/2 its time in heterozygotes



- The probability that neither of our lineages is descended from a recombinant haplotype
 - p_{NR}^2
 - Coalescent time is τ
- Prob one or both descend from recombinant haplotype
 - $1 - p_{NR}^2$
 - Coalescent time is $\tau + 2N$
- Putting together
- $E(T_2) = \tau \times p_{NR}^2 + (1 - p_{NR}^2)(\tau + 2N) \approx (1 - p_{NR}^2)2N$

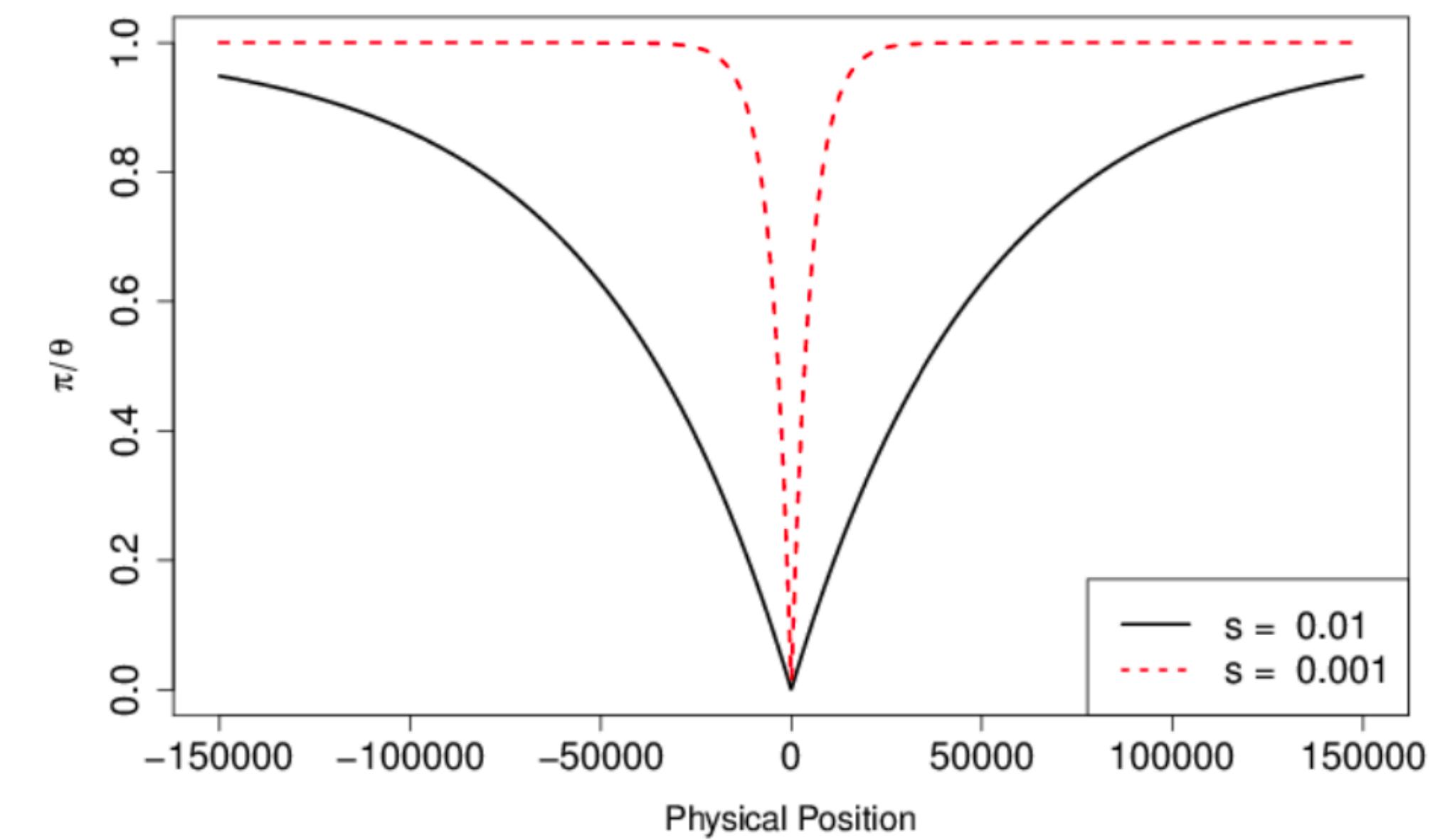


- Expected pairwise diversity for neutral loci due to hitchhiking
- $E(\pi_c) = 2\mu E(T_2) \approx \pi_0(1 - e^{-c\tau})$



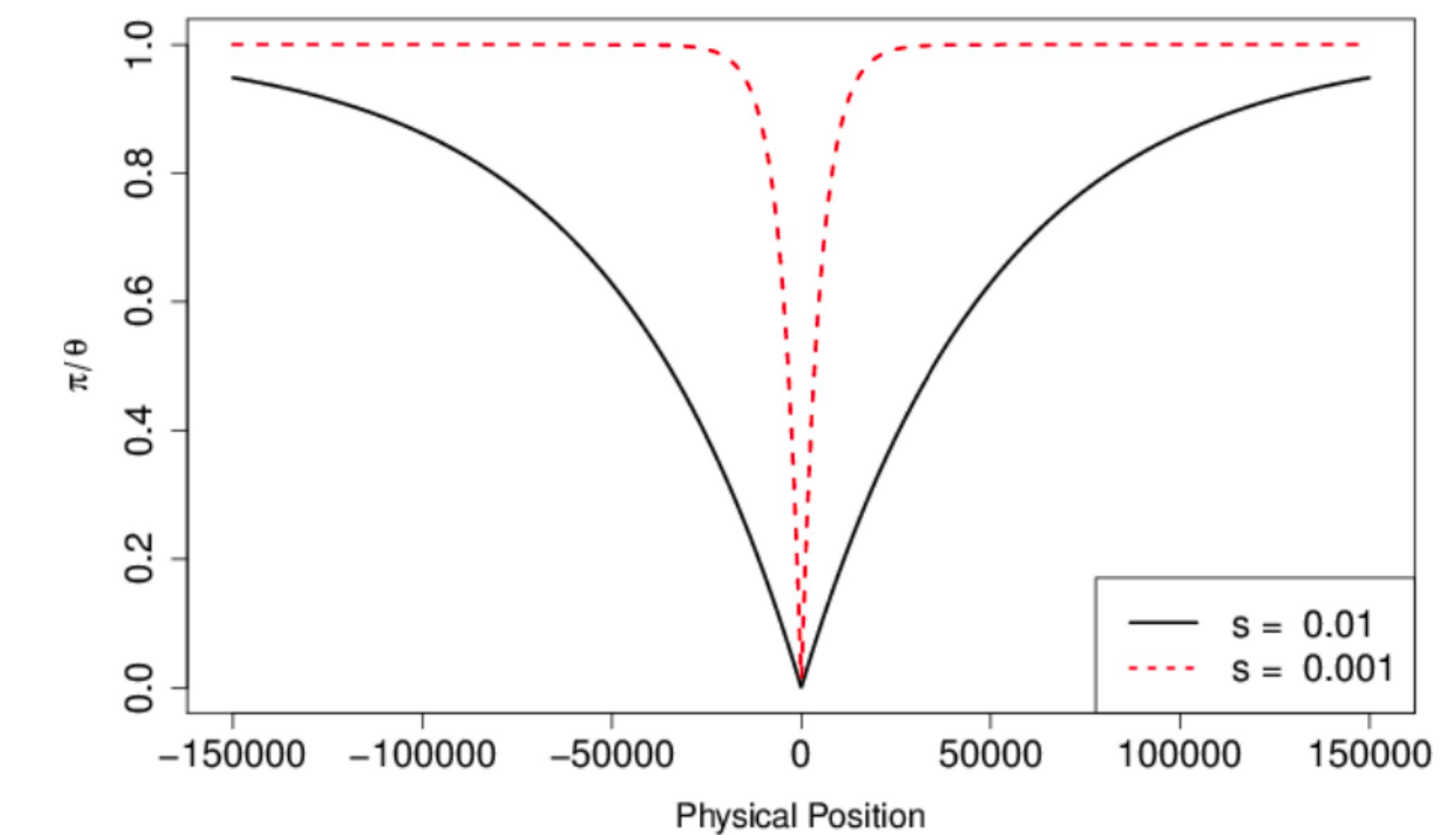
Physical scale

- physical scale over which diversity is reduced
 - How wide is the trough?
 - Physical distance (base pairs) to selected site where 50 % diversity is recovered
- c_{BP} per base pair per generation
 - Assume $c_{BP} = 10^{-8}$
- Recombination rate for a locus ℓ base pairs away from the selected site
 - $c = c_{BP}\ell$
- $l^* = -\frac{\log(0.5)}{c_{BP}\tau}$



Physical scale

- $l^* = -\frac{\log(0.5)}{c_{BP}\tau}$
- $\tau = 4 \log(2N)/s$
- $s = 0.1\%$ reduce diversity ~ 10 kb
- $s = 1\%$ would affect ~ 100 kb

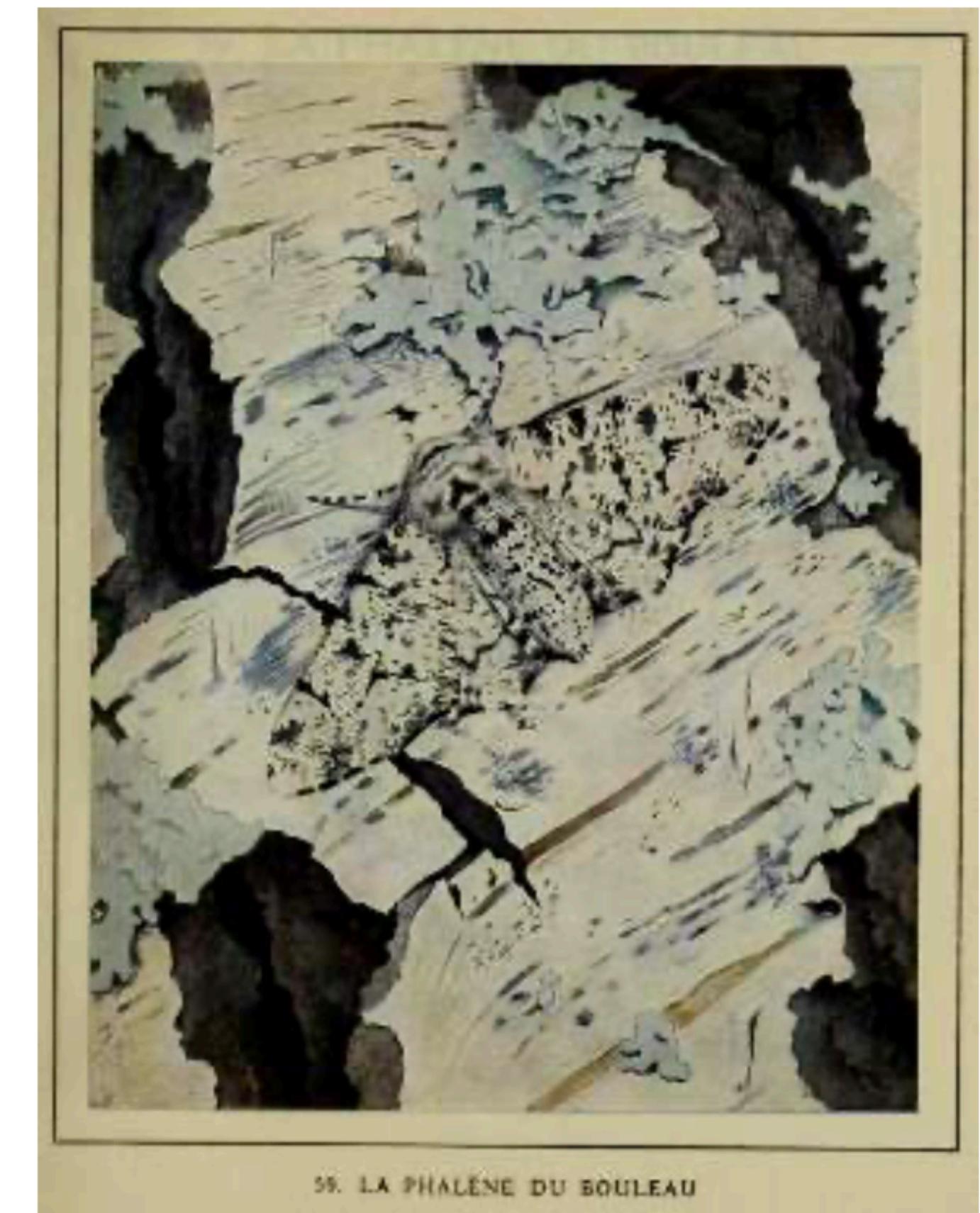


Question 1.

VAN'T HOF *et al.* (2011) identified the genetic basis of melanism in the peppered moth (*Biston betularia*). This allele swept to fixation in northern parts of the UK; a classic case of adaptation to industrial pollution (made famous by the work of KETTLEWELL, see MAJERUS (2009) and COOK *et al.* (2012)). The genetic basis of melanism is a transposable element (TE) inserted into a pigmentation gene. VAN'T HOF *et al.* found that diversity is suppressed in a broad region around the TE. Specifically, on the background of the TE, it takes roughly 200 kb in either direction for diversity levels to recover to 50% of genome-wide levels.

Random facts: In all moths and butterflies only males recombine; chromosomes are transmitted without recombination in females. The recombination rate in males is 2.9 cM/Mb. Peppered moths have an effective population size of roughly a hundred thousand individuals. Kettlewell used to eat moths when out collecting them in the field (personal communication, Art. Shapiro).

- A) Briefly explain how this pattern offers further evidence that the melanic allele was favoured by selection.
- B) Using this information, and assuming the allele's effects on fitness are additive, what is your estimate of the age of the allele?
- C) What is your estimate of the selection coefficient favouring this melanic allele?



Other signals of selective sweeps

- As neutral diversity levels slowly recover through an influx of new mutations after a sweep, there is a strong skew towards low frequency derived alleles, a pattern that persists for many generations
- One can use Tajima's D to scan the data in windows along the genome for detecting selective sweeps

- A selective sweep will decrease the within-population diversity (H_S) surrounding the selected site, without affecting the diversity between different populations.
- This will create F_{ST} peaks between weakly differentiated pops

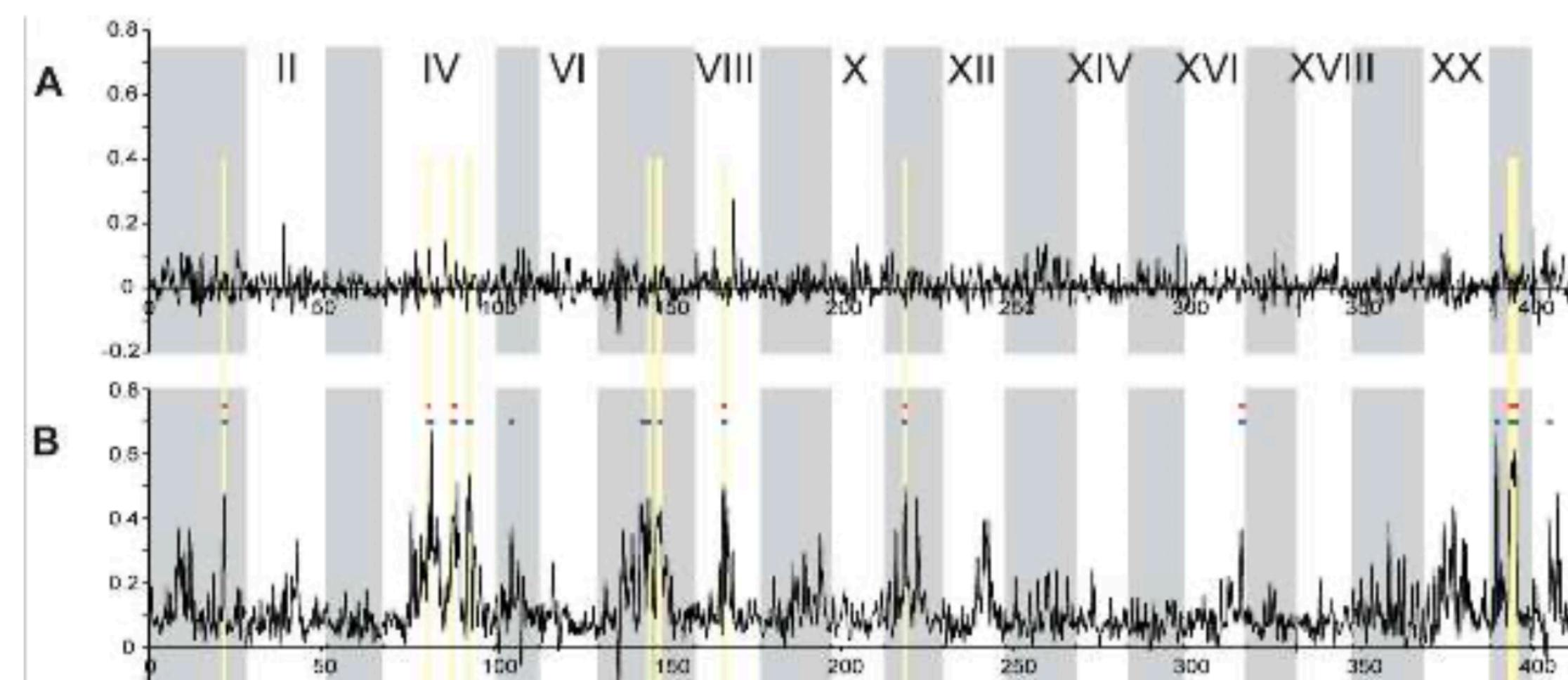
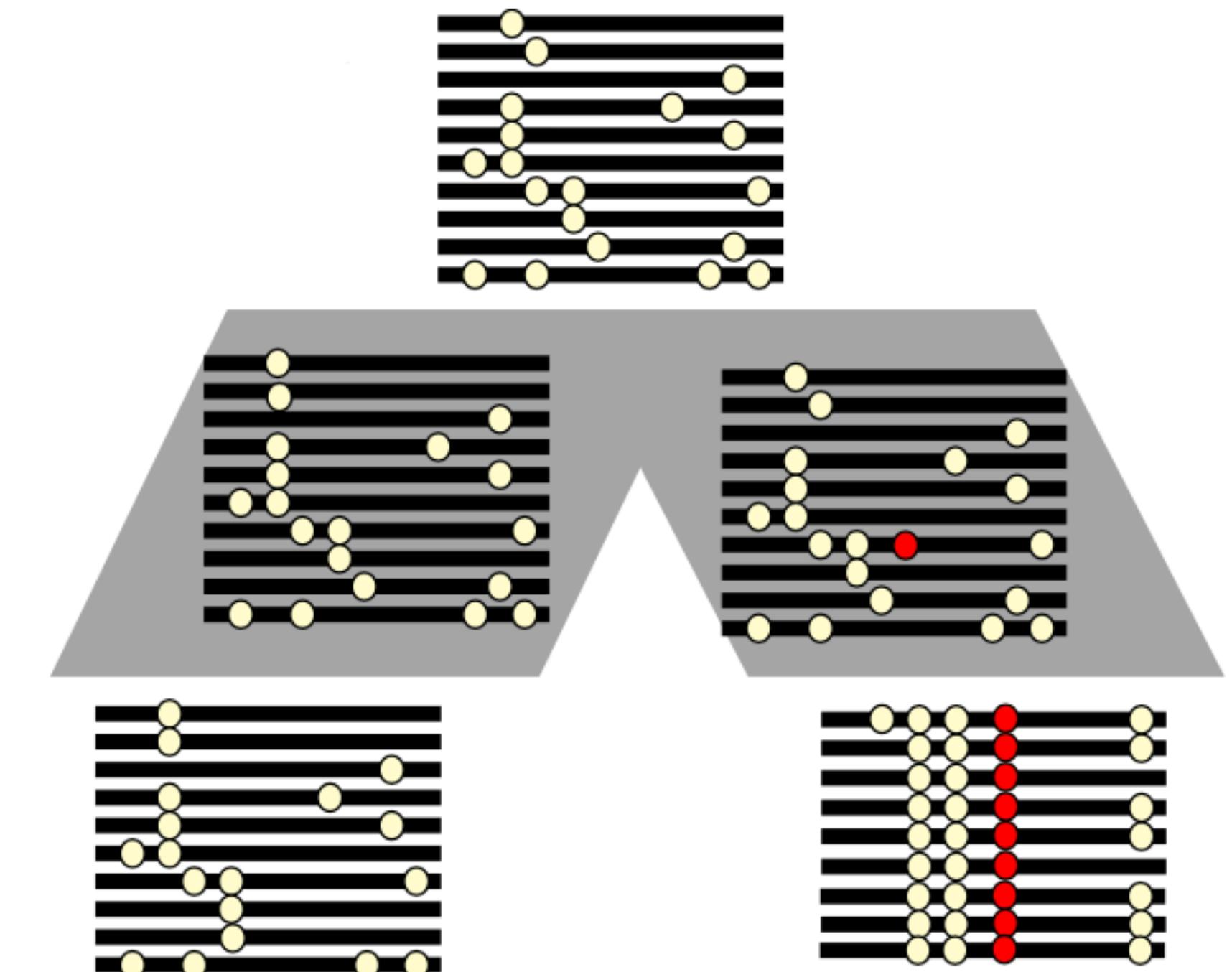
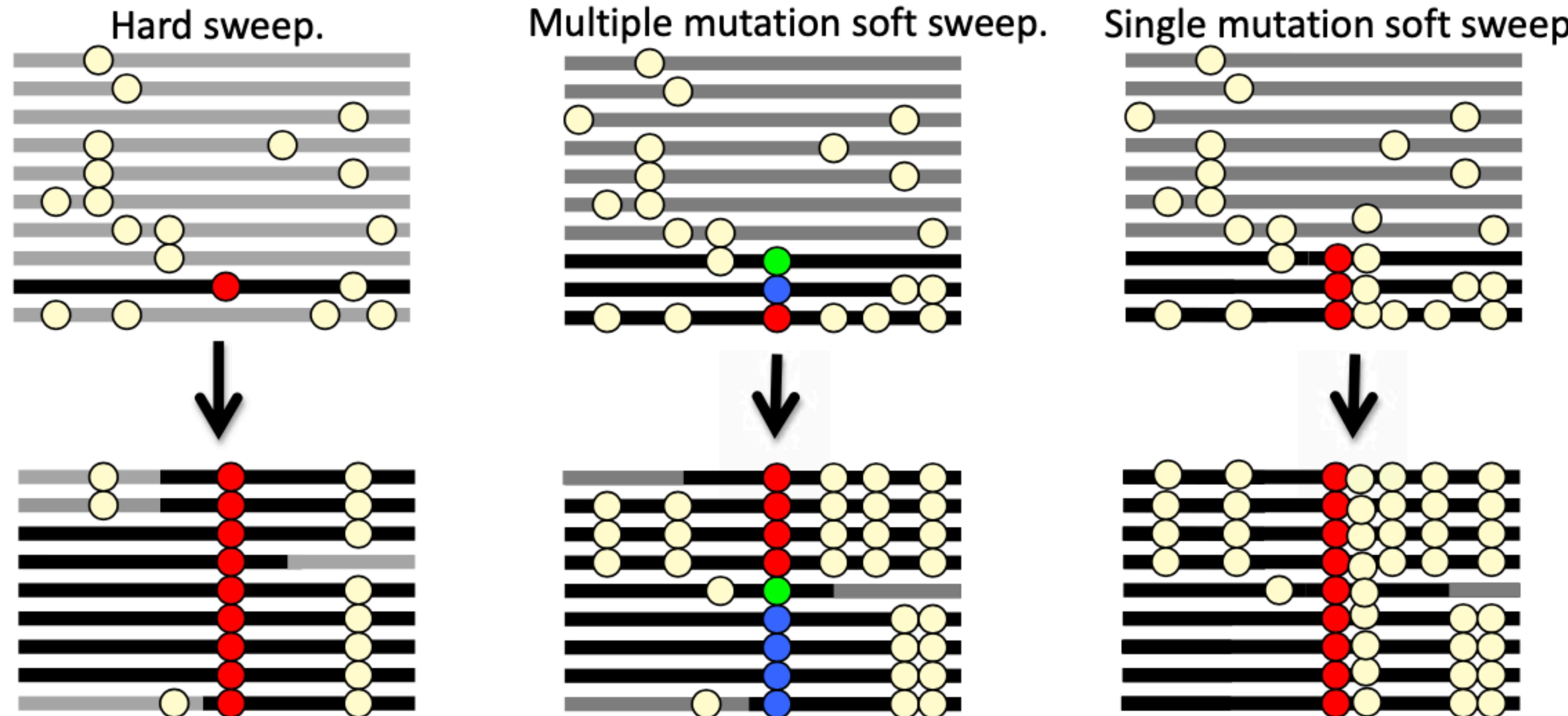


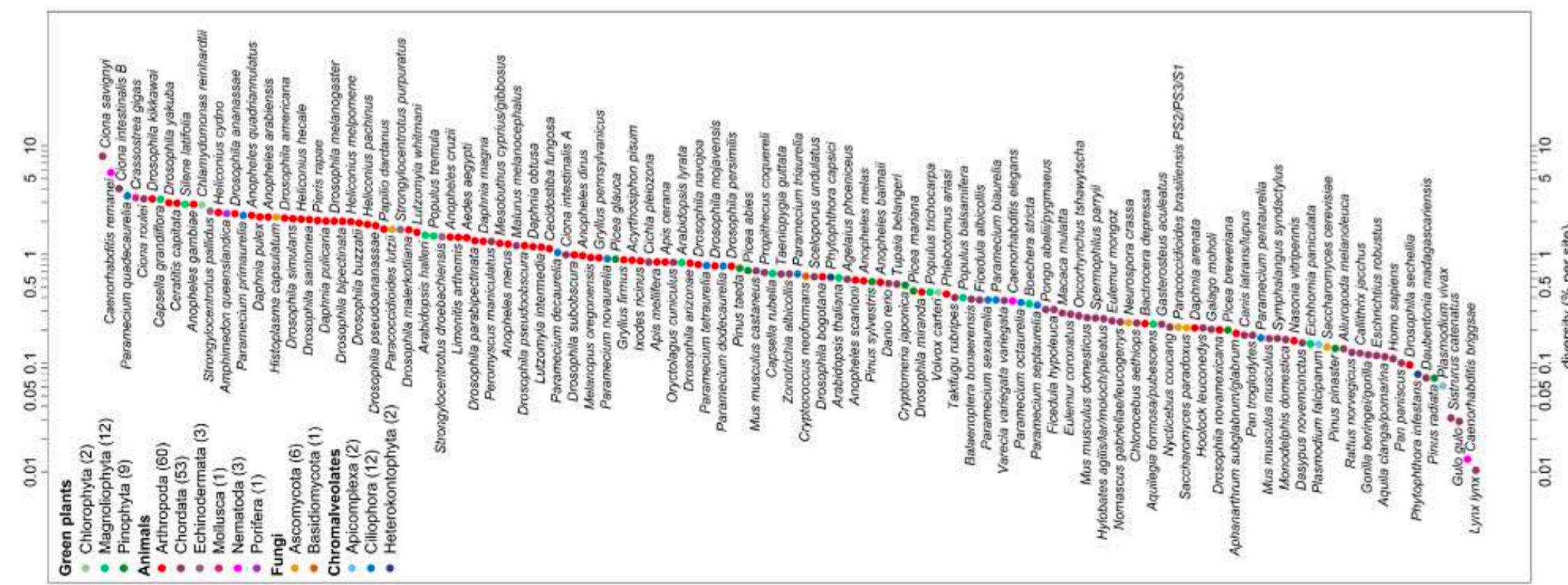
Figure 13.11: F_{ST} across the stickleback genome, with colored bars indicating significantly elevated ($p \leq 10^{-5}$, blue; $p \leq 10^{-7}$, red) and reduced ($p \leq 10^{-5}$, green) values. The alternating white and grey panels indicate different linkage groups. **A)** F_{ST} between two oceanic populations **B)** Average F_{ST} between a freshwater population and the two marine populations. Figure and caption text from HOHENLOHE *et al.* (2010), licensed under CC BY 4.0.

Soft Sweeps from multiple mutations and standing variation.

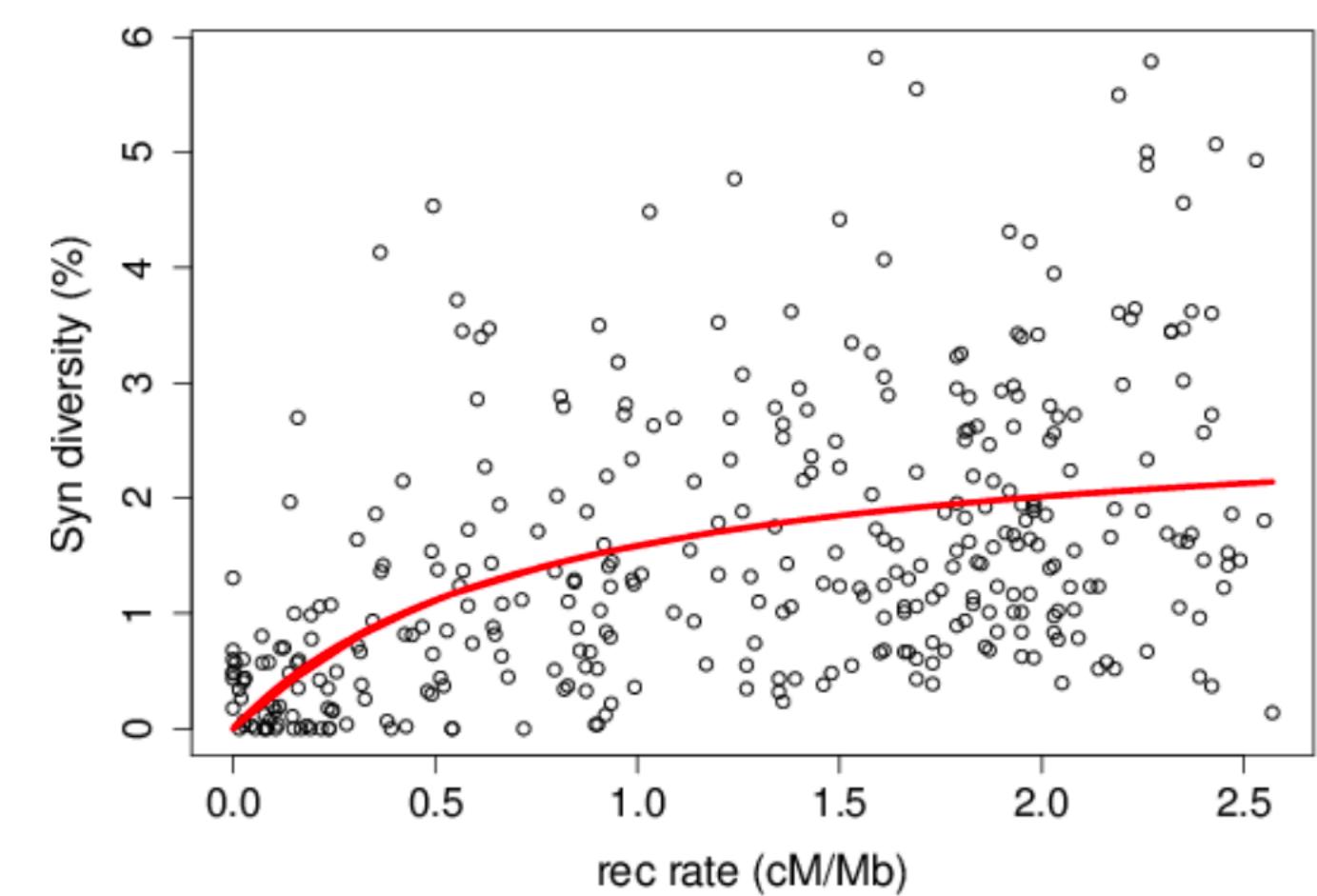


The genome-wide effects of linked selection

- Puzzle:
 - relatively narrow range of polymorphism across species with vastly different census sizes
- Hypothesis:
 - In large populations, selective sweeps and other forms of linked selection may come to dominate over genetic drift as a source of stochasticity in allele frequencies, potentially establishing an upper limit to levels of diversity



- How do selective sweeps contribute to broader patterns of variation?
- Positive correlation between putatively neutral diversity and recombination
- Could be the signature of of selective sweeps happening recurrently along the genome.



Recurrent model of selective sweeps

- A pair of neutral alleles c distance away from a locus where sweeps are initiated within the population at some very low rate ν per generation.
- The mean waiting time between sweeps at our locus $1/\nu$
- $\tau \ll 1/\nu$
- Per generation rate of coalescence due to sweep
 - $\nu \times p_{NR}^2 = \nu e^{-c\tau}$

Recurrent model of selective sweeps

- Per generation rate of coalescence due to sweep

- $\nu \times p_{NR}^2 = \nu e^{-c\tau}$

- Per generation rate of coalescence due to drift

- $1/2N$

- $\mathbb{E}(T_2) = \frac{1}{\nu e^{-c\tau} + 1/2N}$

Recurrent model of selective sweeps

- $\mathbb{E}(T_2) = \frac{1}{\nu e^{-c\tau} + 1/2N}$
- Suppose sweep occurs uniformly at random across our genome
- at a very low rate of ν_{BP} per basepair per generation.
- rate of coalescence due to sweeps at a locus ℓ basepairs away from our neutral loci
- $2\nu_{BP}e^{-c_{BP}\ell\tau}$

Recurrent model of selective sweeps

- Rate of coalescence due to sweeps at a locus ℓ basepairs away from our neutral loci

- $2\nu_{BP}e^{-c_{BP}\ell\tau}$

- If our neutral locus is in the middle of a chromosome that stretches L basepairs in either direction, the total rate of sweeps per generation that could force our pair of lineages to coalesce is

- $2 \int_0^L \nu_{BP}e^{-c_{BP}\ell\tau} d\ell = \frac{2\nu_{BP}}{c_{BP}\tau} (1 - e^{-c_{BP}\tau L})$

- Rate of coalescence per generation due to sweeps

- $2\nu_{BP}/c_{BP}\tau$

- Average time until a pair of lineages coalesce is

$$\bullet \quad \mathbb{E}(T_2) = \frac{1}{2\nu_{BP}/c_{BP}\tau + 1/2N} = \frac{c_{BP}2N}{4N\nu_{BP}/\tau + c_{BP}}$$

- Expected diversity

$$\bullet \quad \mathbb{E}(\pi) = \pi_0 \frac{c_{BP}}{4N\nu_{BP}/\tau + c_{BP}}$$

- Fitting to data
- Assume $4N\nu_{BP}/\tau$ is constant
- We can fit in π across regions that vary in their recombination rate (cBP) to estimate a population's rate of recurrent sweeps ν_{BP}
- Best fitting $4N\nu_{BP}/\tau = 7 \times 10^{-9}$
- Assume $N_e = 10^6$
- $\tau \approx 1000$
- $\nu_{BP} \approx 10^{-12}$
- one every megabase every million generations !

