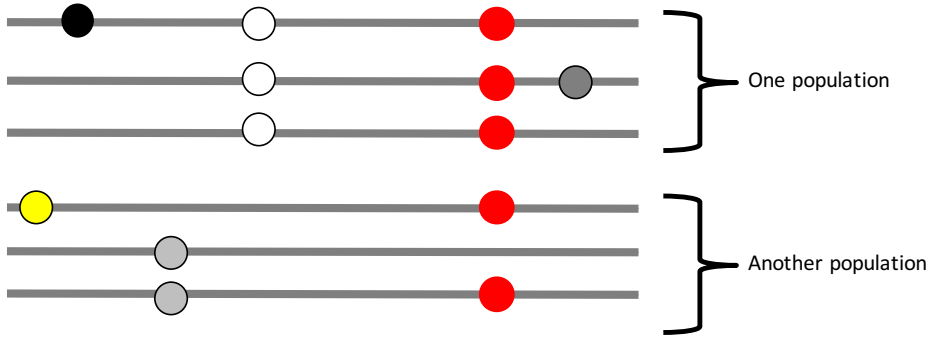
EACCR2 NID Node Training Course: Genomics

Lecture 4

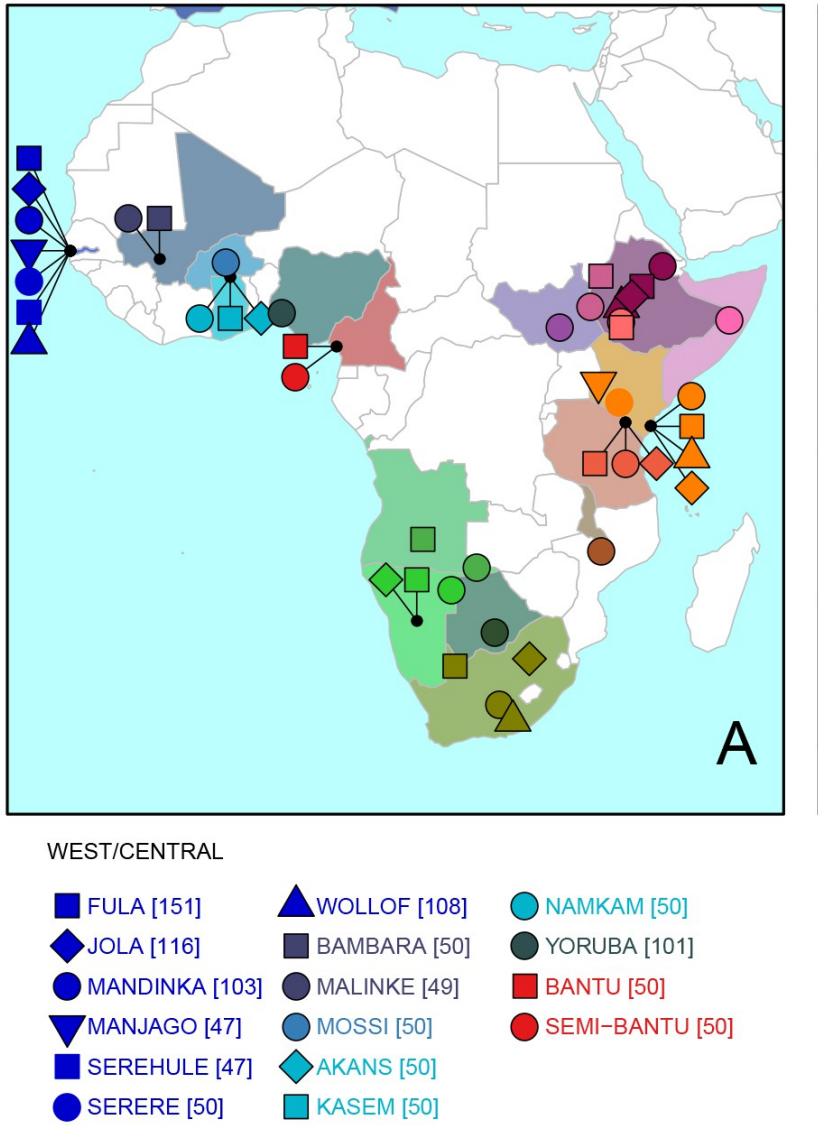
Principles of population genomics

Population genomics

Variants tend to remain within their original population

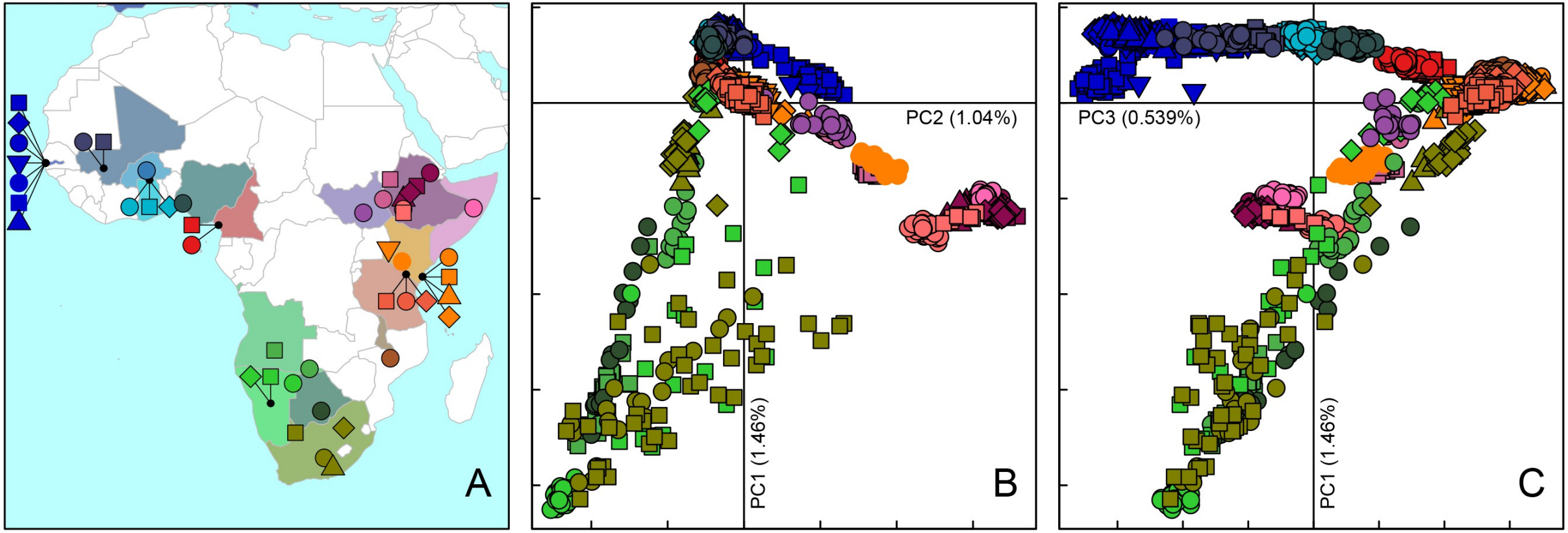


- When people, animals, plants or microbes travel, they carry their DNA with them.
- Every individual carries its history within its DNA
- Because we know the ‘rules’ (mutation, drift, recombination) population movements can be modelled
- Because populations can have small contributions from one/other populations, this can’t always be drawn on a phylogenetic tree.



Busby 2016

Population genomics



WEST/CENTRAL

- FULA [151]
- JOLA [116]
- MANDINKA [103]
- MANJAGO [47]
- SEREHULE [47]
- SERERE [50]
- WOLLOF [108]
- BAMBARA [50]
- MALINKE [49]
- MOSSI [50]
- AKANS [50]
- KASEM [50]
- NAMKAM [50]
- YORUBA [101]
- BANTU [50]
- SEMI−BANTU [50]

EASTERN

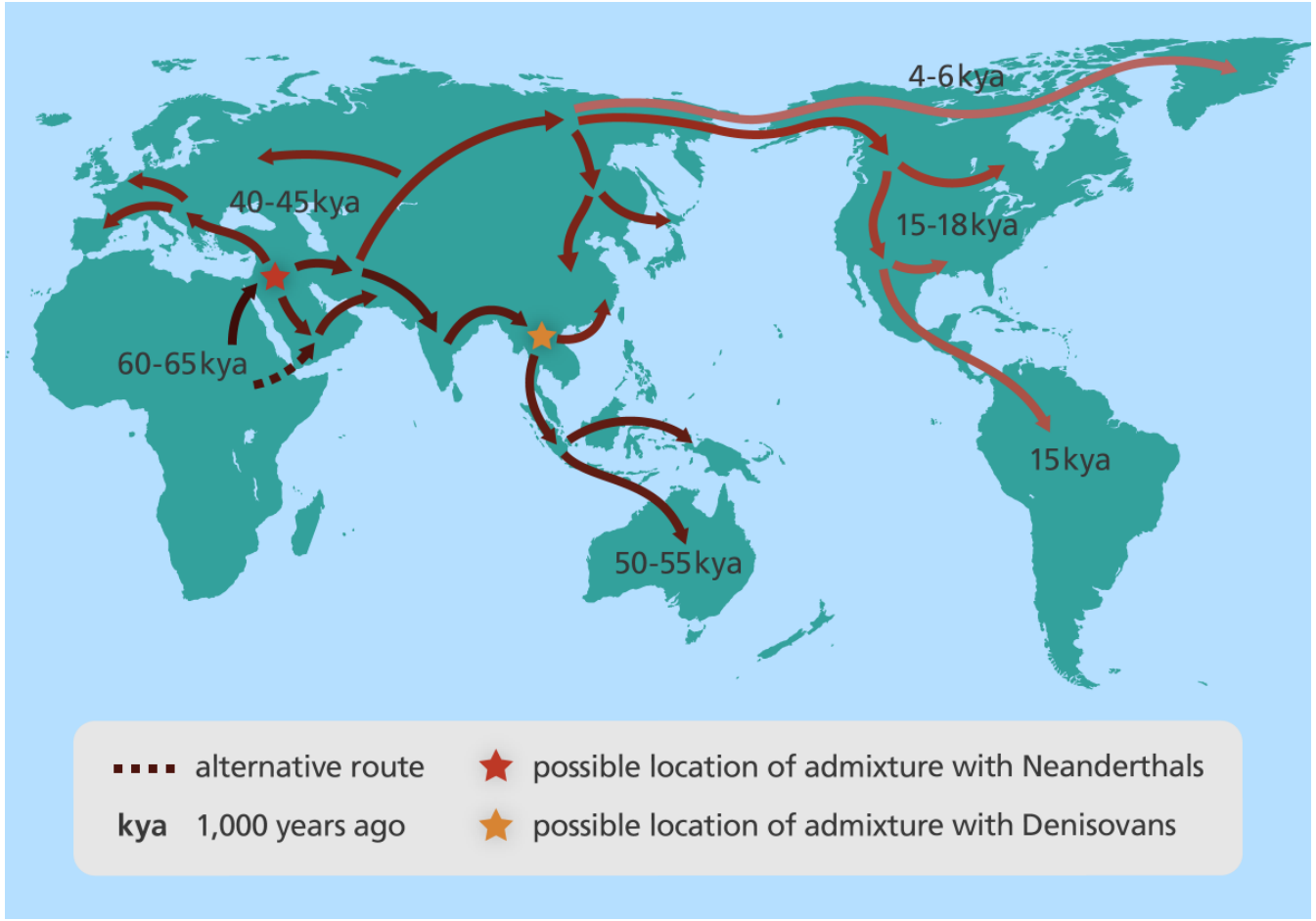
- SUDANESE [24]
- AFAR [12]
- AMHARA [26]
- OROMO [21]
- TIGRAY [21]
- ANUAK [23]
- GUMUZ [19]
- ARI [41]
- WOLAYTA [8]
- SOMALI [40]
- CHONYI [47]
- GIRIAMA [46]
- KAMBE [42]
- KAUMA [97]
- LUHYA [91]
- MAASAI [31]

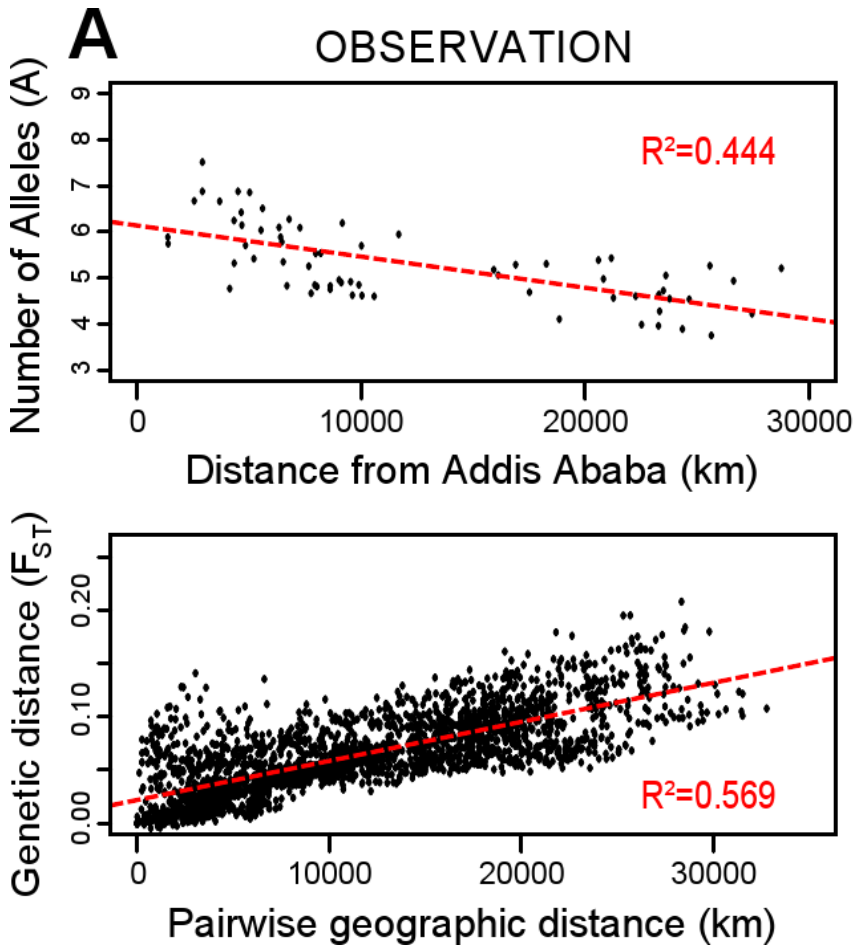
SOUTHERN

- MZIGUA [50]
- WABONDEI [48]
- WASAMBAA [50]
- MALAWI [100]
- AMAXHOSA [15]
- KARRETJIE [20]
- =KHOMANI [39]
- SEBANTU [20]
- HERERO [12]
- JU/'HOANSI [37]
- NAMA [20]
- /GUI//GANA [15]
- KHWE [17]
- !XUN [33]

Busby 2016

Population genomics





We are, right now, close to our original home.

What is population genomics

- Population **genetics** is the study of genetic variation within species.
- Population **genomics** expands the data to study variation within species using whole genome data.
- **Population genomics:**
  - Is more challenging data to gather, more expensive, more challenging to analyse
  - Genome-scale data produces a more comprehensive picture
- Demography (population size, migration, population structuring)
- Natural selection (purifying, adaptive, balancing)

Demography, selection and quantitative genetics are intimately related.

Collecting population genomics data

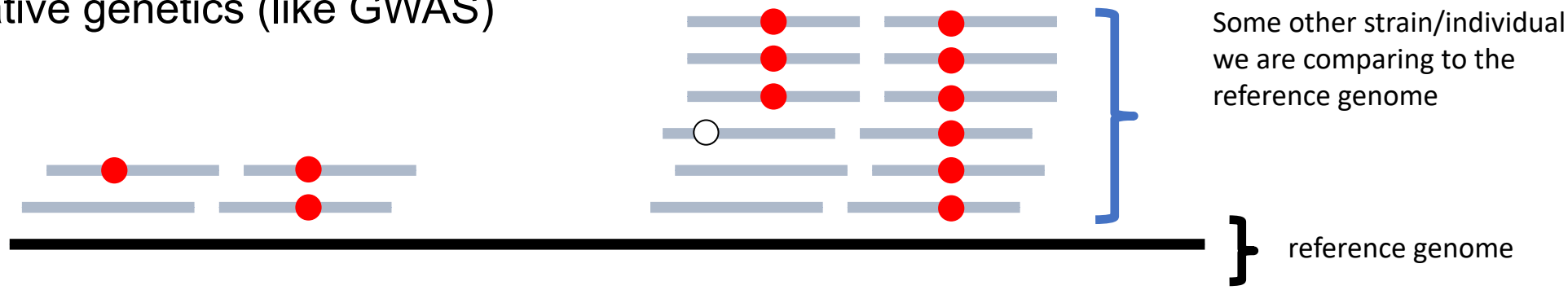
1. **Hypothesis/query**
2. **Sample collection and DNA extraction**
   a) Choose geographic/habitat region of interest
   b) Gather hundreds to thousands of individuals (strains/subjects) from within a species (sometimes tens of thousands)
   c) Extract genomic DNA from each individual
3. **Genome sequencing**
   a) Aim: to obtain sequence data covering the entire genome 5x to 40x coverage

   Coverage: how many reads per site
   b) Sequence genome using 'short read' technology (usually Illumina)
   c) Main issue: cost/base

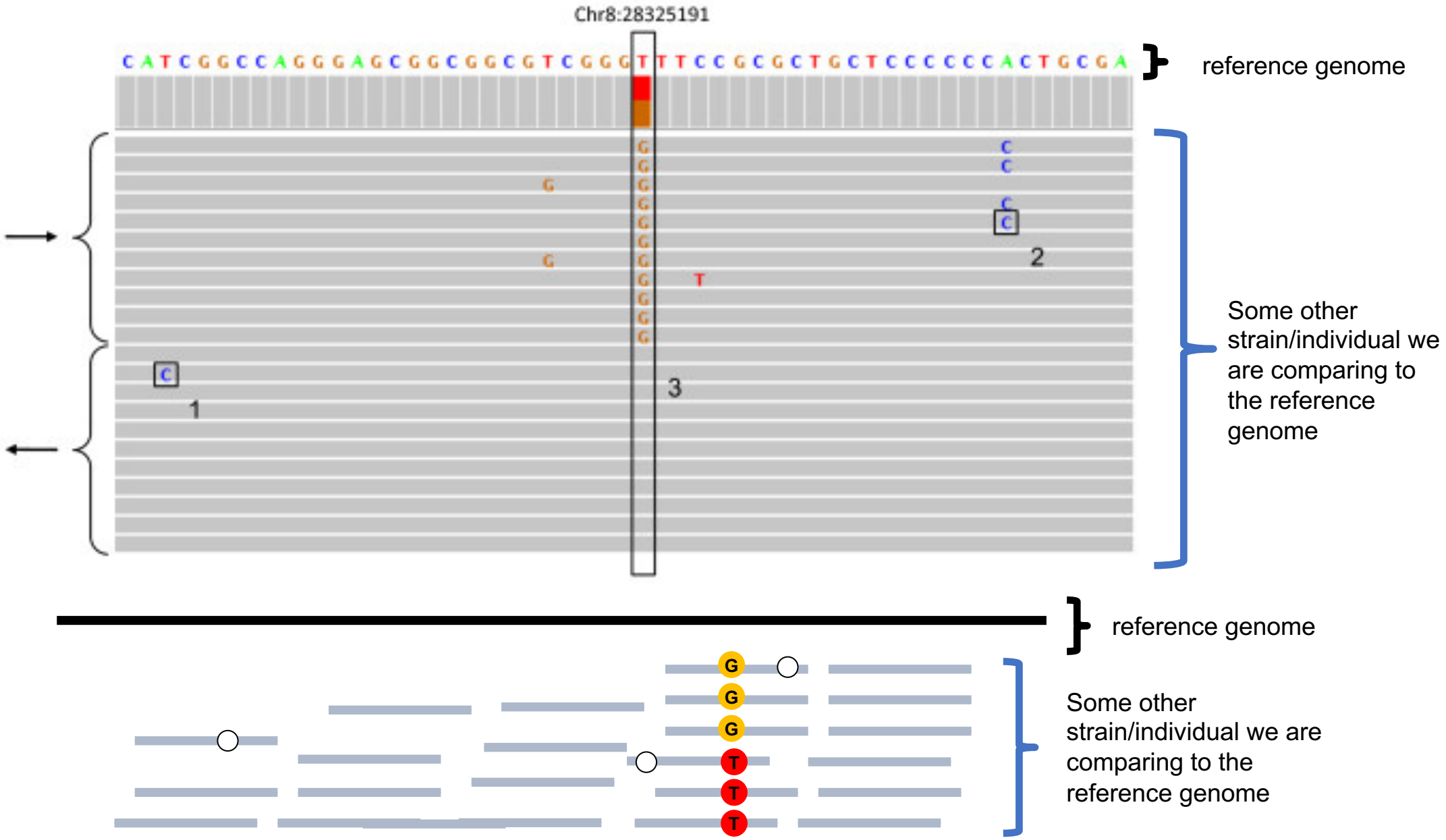
Reads from one individual, aligned to a reference genome

Collecting population genomics data

4. **Read mapping and 'variant calling'**
   a) Locate genetic variants (sites the genome that differ between individuals, polymorphisms)
      a) eg: SNPs, indels etc
   b) By mapping (aligning) sequence reads to a reference genome, and identifying sites the genome that differ
5. **Segregating genetic variants are (usually) the final data set**
   a) A list of positions that vary. Alleles/polymorphisms/variants.
6. **Analysis:**
   a) Describing demography
   b) Detecting selection
   c) Quantitative genetics (like GWAS)



Reads from one individual, aligned to a reference genome

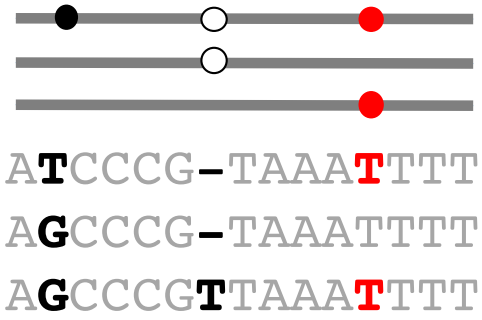


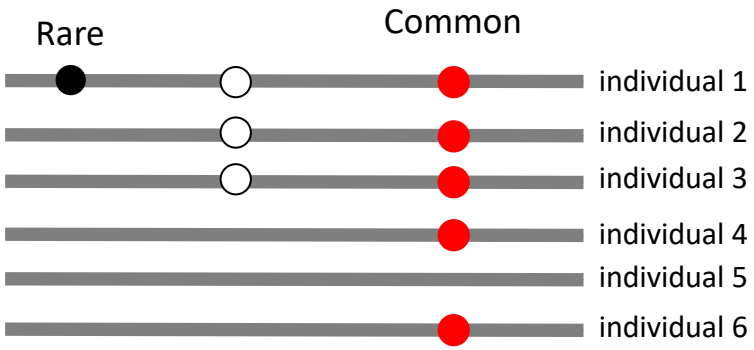
Reads from one individual, aligned to a reference genome

Concepts in population genomics

Genetic diversity

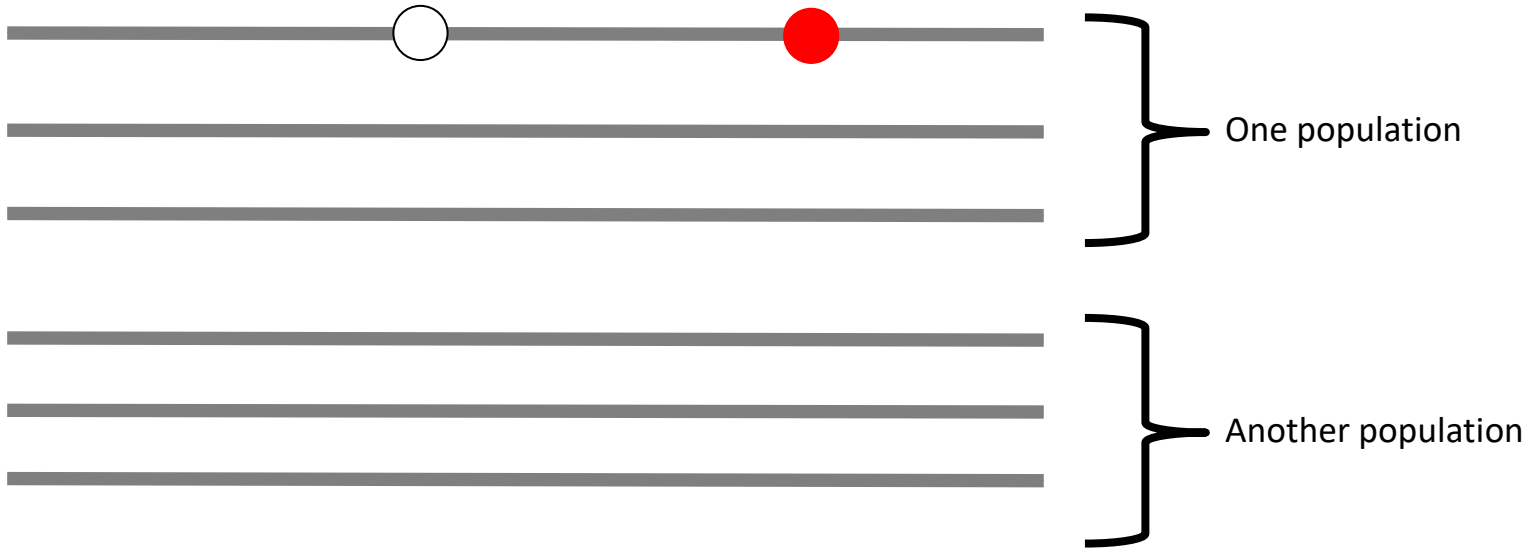
- Polymorphisms/alleles/variants: sites in a genome that differ between individuals of a species
  - Single nucleotide polymorphisms (SNPs)
  - Small insertion/deletions (indels)
  - Transposon insertions
  - ‘Structural’ variants: duplications, rearrangements, large insertions/deletions
- Initial origin: a mutation in one individual
  - **All polymorphisms begin their existence in just one individual**
- Polymorphisms then move through space and time, within the population
- Their frequency in the population will change





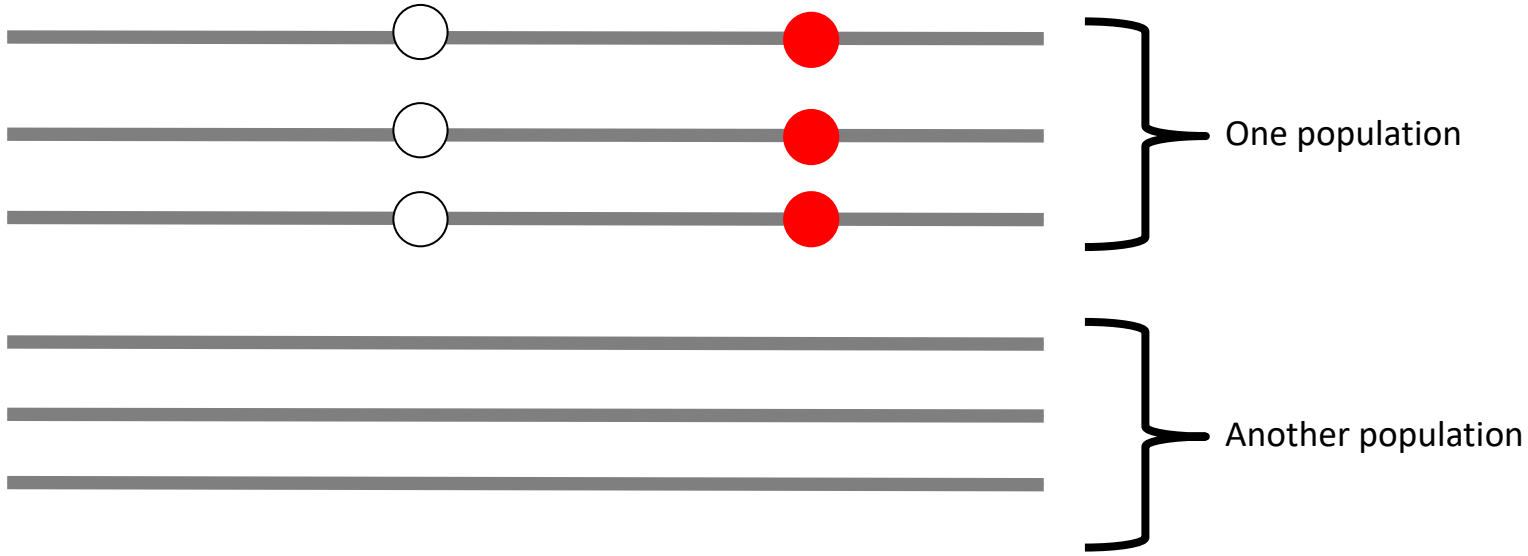
A genomic site/gene from many individuals within a species

Genetic diversity: the life of a SNP

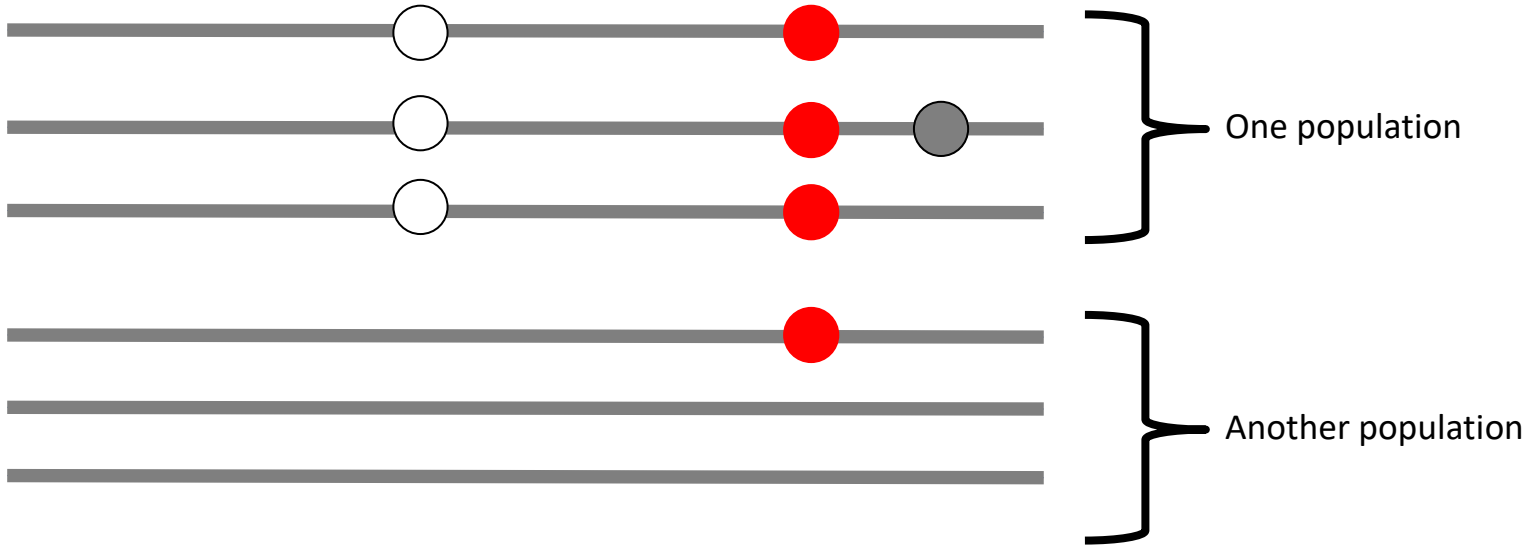


A genomic site/gene from many individuals within a species

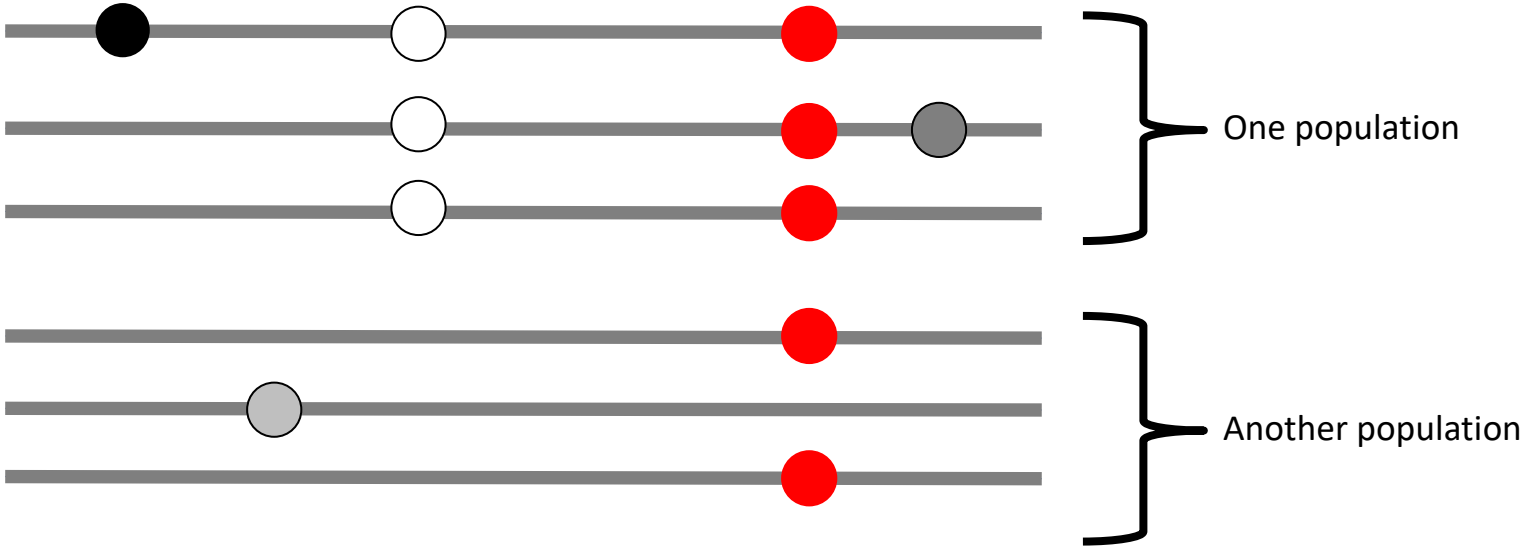
Genetic diversity: the life of a SNP



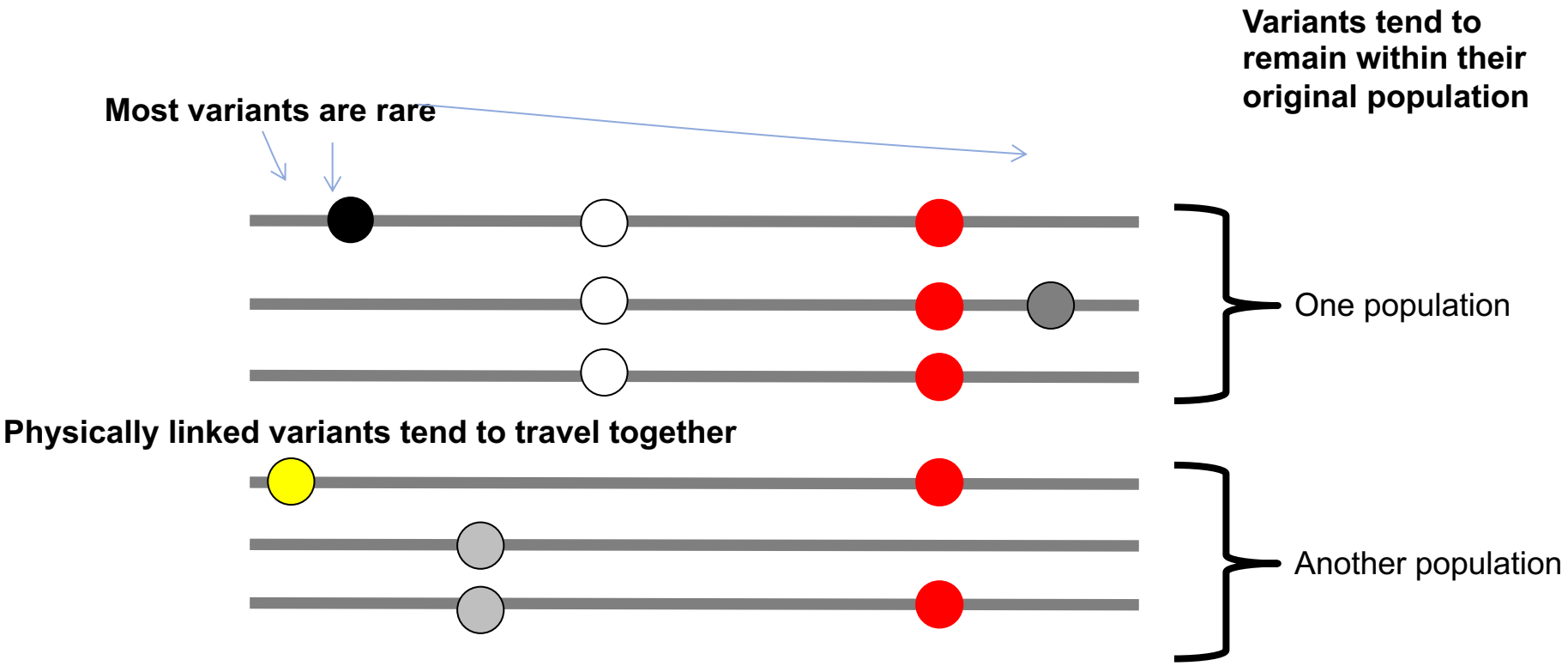
Genetic diversity: the life of a SNP



Genetic diversity: the life of a SNP



Genetic diversity: the life of a SNP



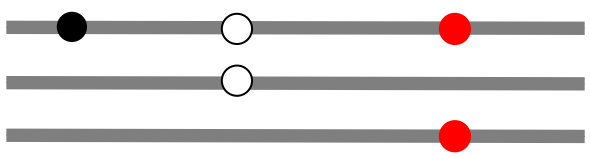
Many signals can be found within polymorphism data.

The patterns of polymorphism data are complex.

Hence: summary statistics

Genetic diversity summary statistics

- Average pairwise similarity ('diversity', π):
  - If we compare every sequence to every other, what is the average number of differences?
- The number of segregating sites (S):
- Allele frequencies:
  - Minor allele frequency (MAF)
  - Derived allele frequency (DAF)
- Each of these statistics can be described as:
  - A histogram (a distribution)
  - Plots along chromosomes



A genomic site/gene

from many individuals within a species

```
ATCCCG-TAAATTTT
AGCCCG-TAAATTTT
AGCCCGTTAAATTTT

AGCCCGTTAAATTTT
```

Ancestral state

See:

Nucleotide diversity (**π**) on Wikipedia: https://en.wikipedia.org/wiki/Nucleotide_diversity
Watterson estimator of **Θ** on Wikipedia https://en.wikipedia.org/wiki/Watterson_estimator

Tajima's D

Tajima's D uses summary statistics to detect selection or demography

- The expected number of segregating sites, $\Theta \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$
- In a neutral site $\pi = \Theta$
- Fumio Tajima worked out that $\pi \neq \Theta$ in certain circumstances
- Tajima's D is approximately = $\pi - \Theta$
- **When D is negative:**
  - more rare alleles that expected
  - selective sweep (or expanding population)
- **When D is positive:** (too many rare alleles)
  - Balancing selection or shrinking population or population structure

$$\frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

S is the number of segregating sites

N is the number of sequences compared

$$\sum_{i=1}^{n-1} \frac{1}{i}$$

For six sequences:

Sum of: ½ + 1/3 +1/4 +1/5

In theory, $\Theta = \pi = 4N\mu$
Where N = population size
μ = mutation rate

So π can give us an estimate of population size.

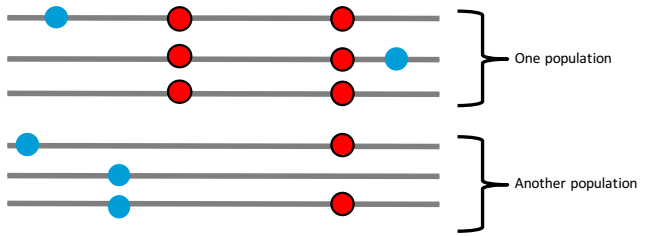
Tajima's D on Wikipedia https://en.wikipedia.org/wiki/Tajima%27s_D

Excellent explanation of π, Θ and Tajima's D on Youtube:
https://www.youtube.com/watch?v=wiyay4YMq2A

Tajima's D

Tajima's D = 0

Neutrally evolving, stable population

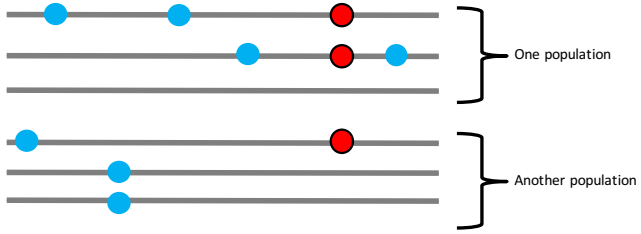


When D is negative:

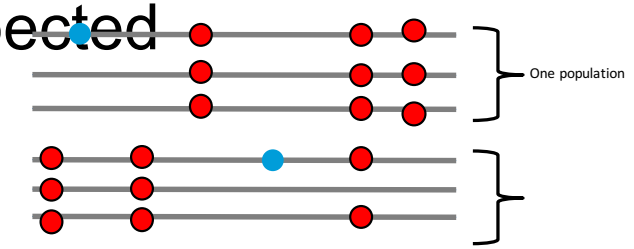
more rare alleles that expected

selective sweep

or expanding population



- **When D is positive:**
  - **more common alleles** than expected
  - balancing selection
  - shrinking population
  - population structure



Tajima's D on Wikipedia https://en.wikipedia.org/wiki/Tajima%27s_D

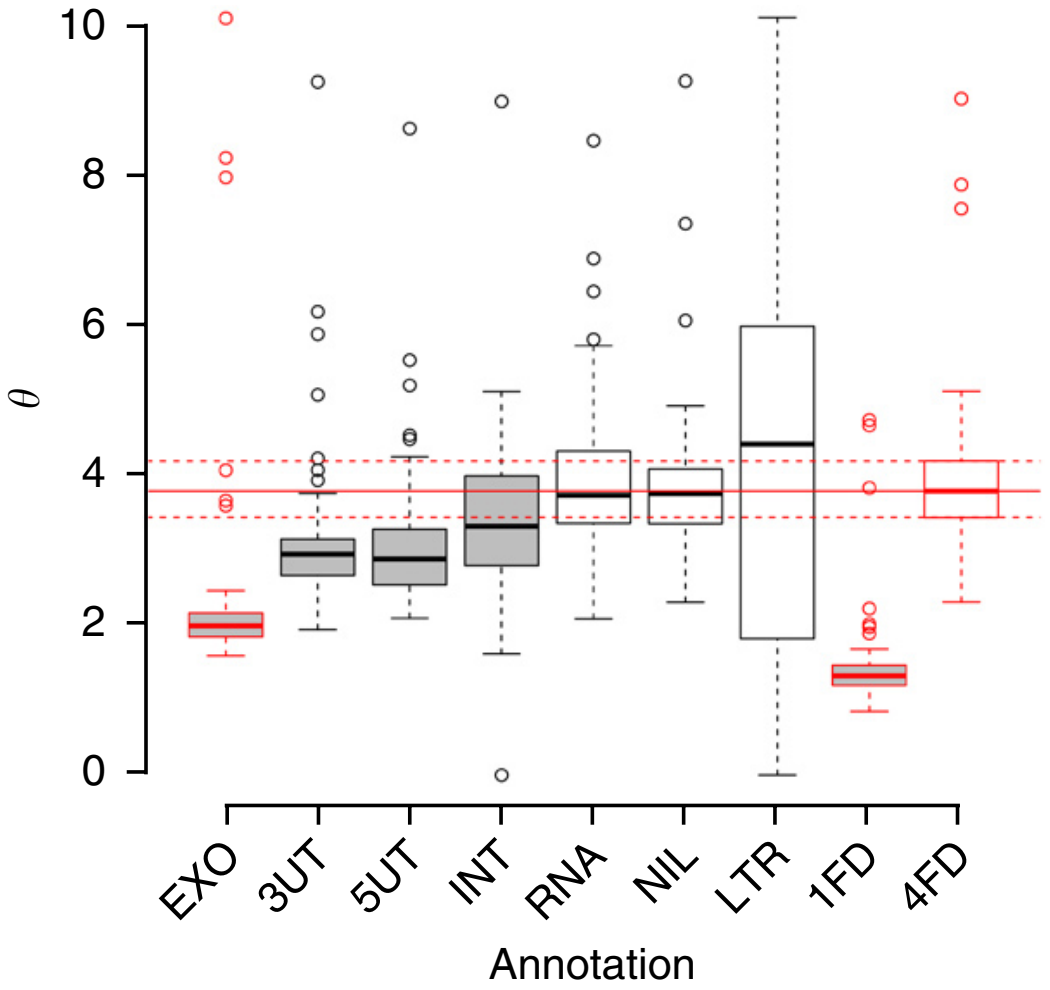
Excellent explanation of π, Θ and Tajima's D on Youtube:

Purifying selection (negative selection)

Purifying is the loss of deleterious (harmful) variants

This process will

- Reduce diversity in regions that are important
- Increase the proportion of rare alleles
- Cause negative Tajima's D
- Purifying selection is expected to be a **common event**



Jeffares 2015

Positive selection (adaptive evolution)

Adaptive evolution is the the increase in frequency of an adaptive (helpful) variant

This process will

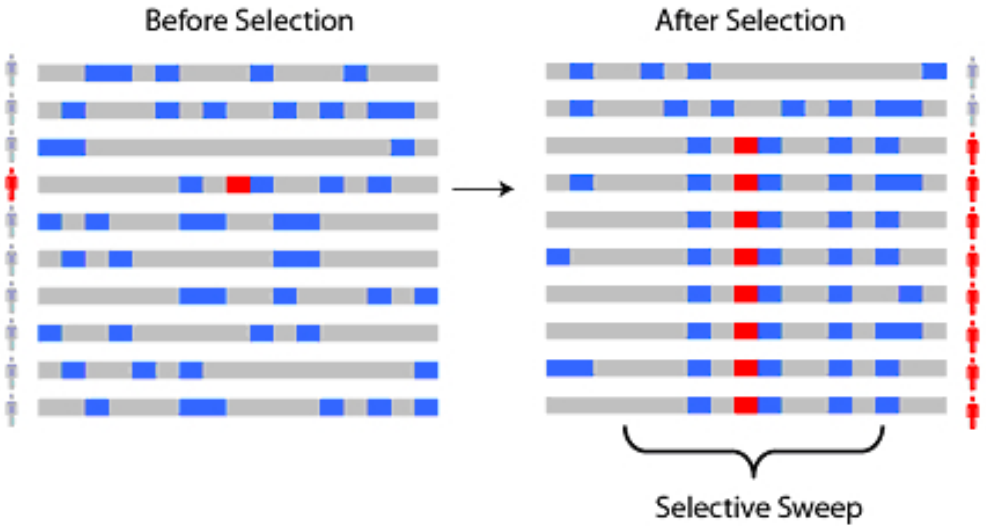
- Reduce diversity around the beneficial allele
- Increase rare alleles
- Cause negative Tajima's D
- Strong adaptive evolution is expected to be a **rare event**

Linkage

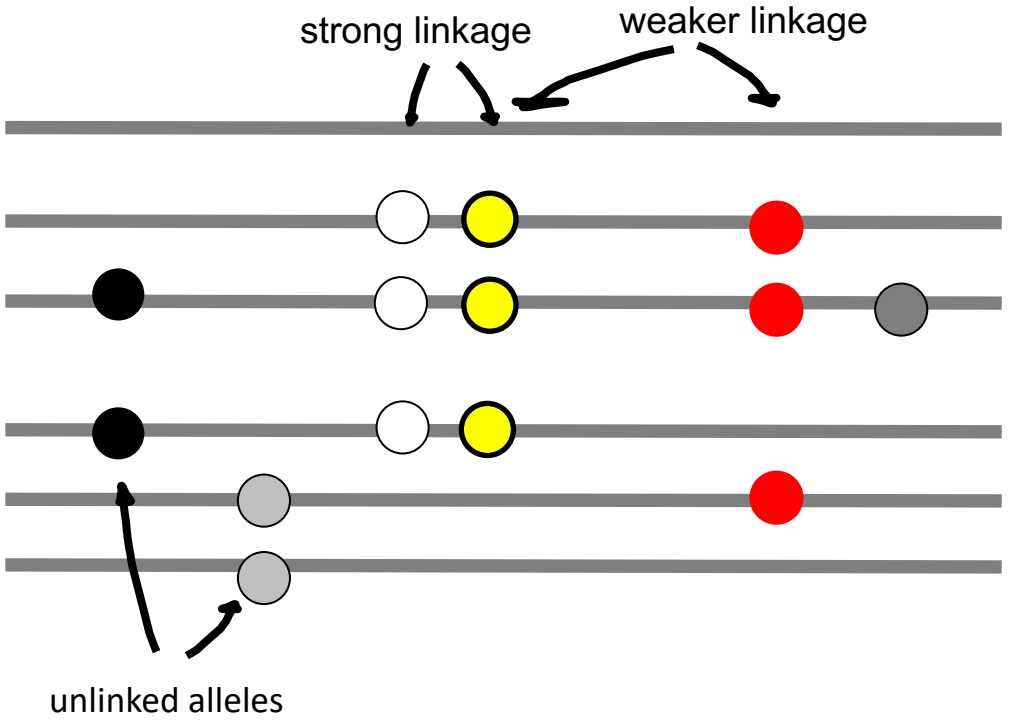
When a strongly beneficial allele arises it will ‘sweep’ through the population.

This will cause strong genetic signatures in the genome:

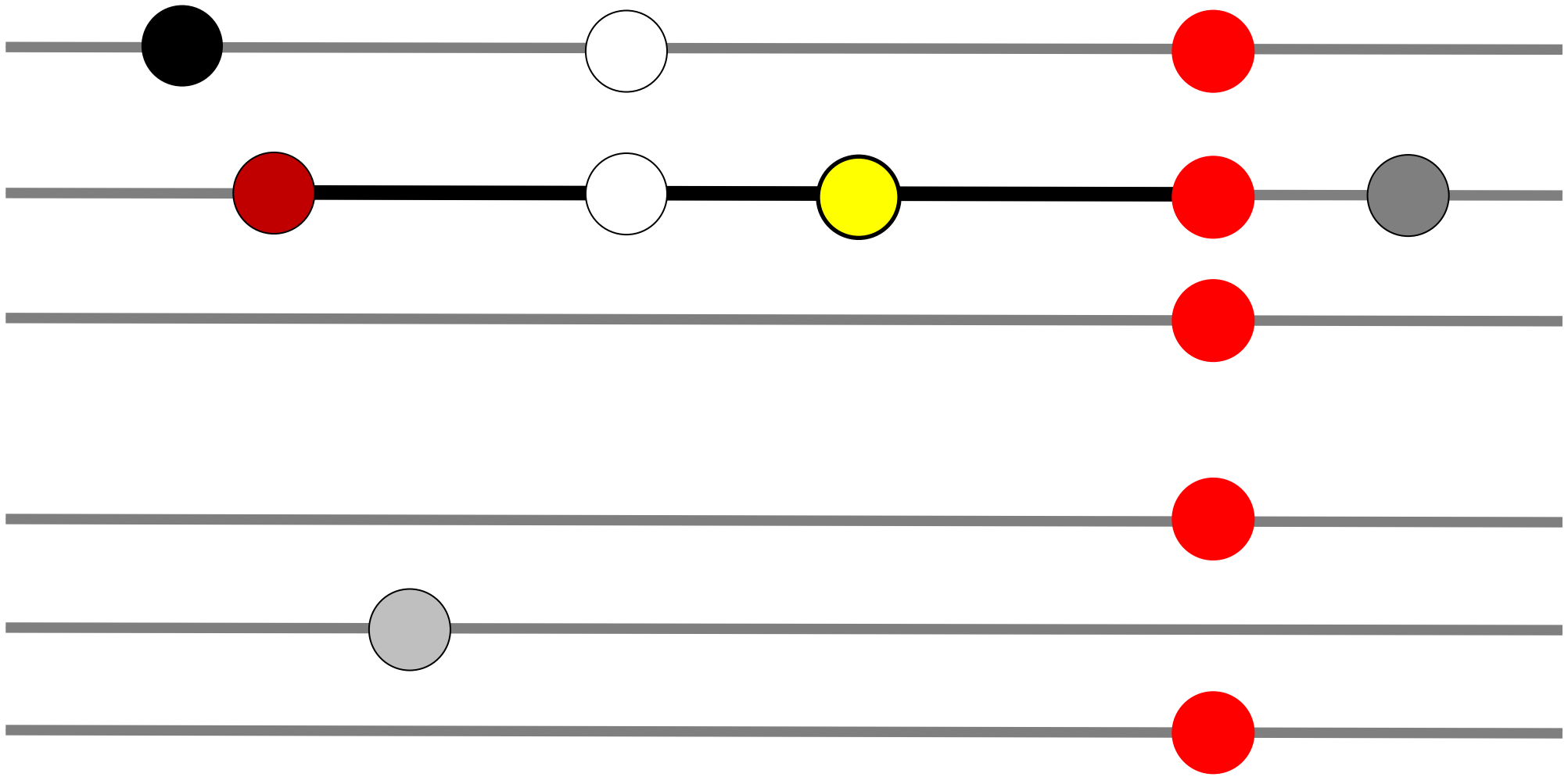
- loss of diversity around the sweep (asses by looking at π)
- increase in linkage (look at linkage)



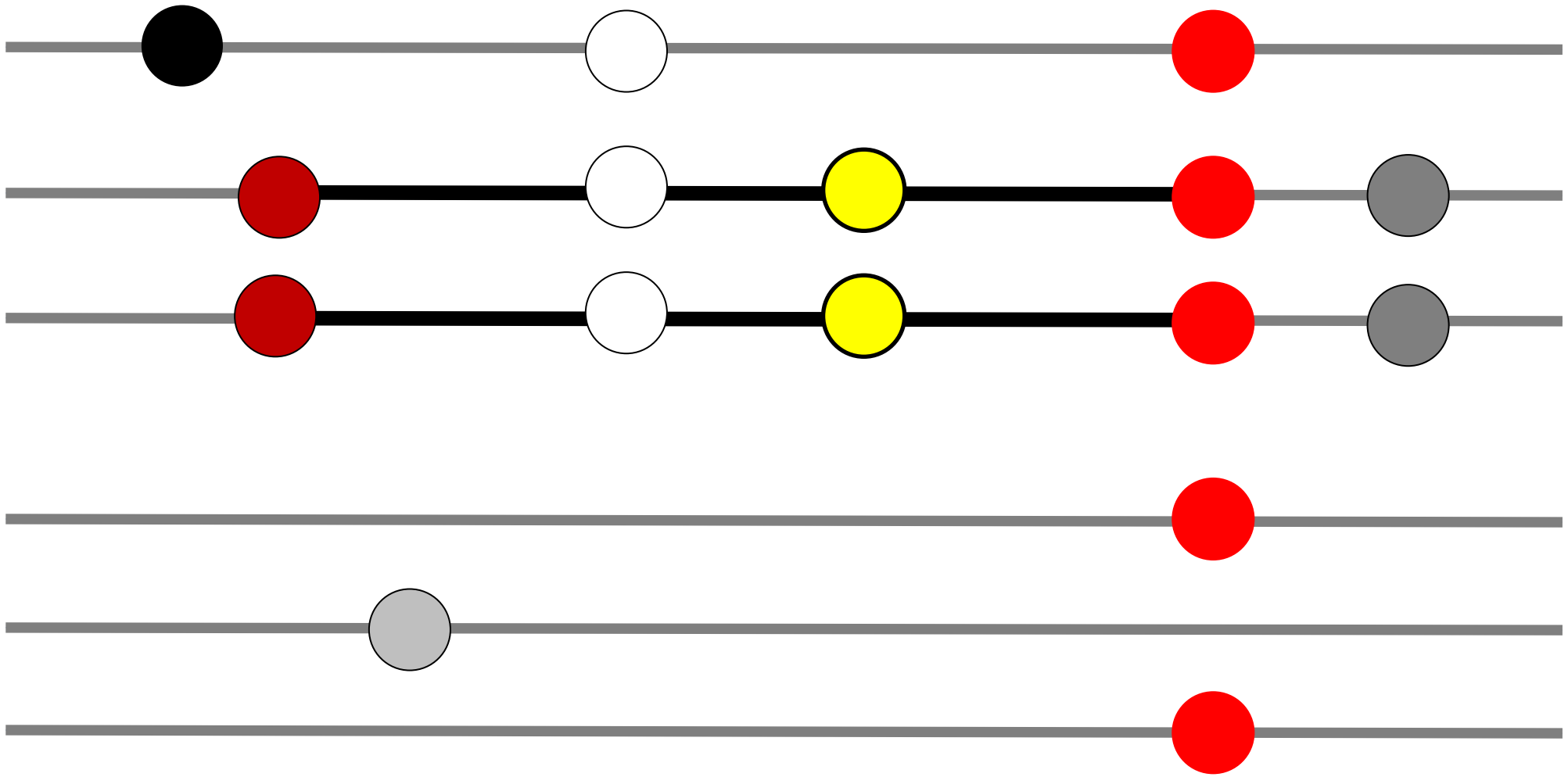
Alleles are referred to as ‘linked’ when they are often found together
linkage is strongest for alleles that site close on the same chromosome



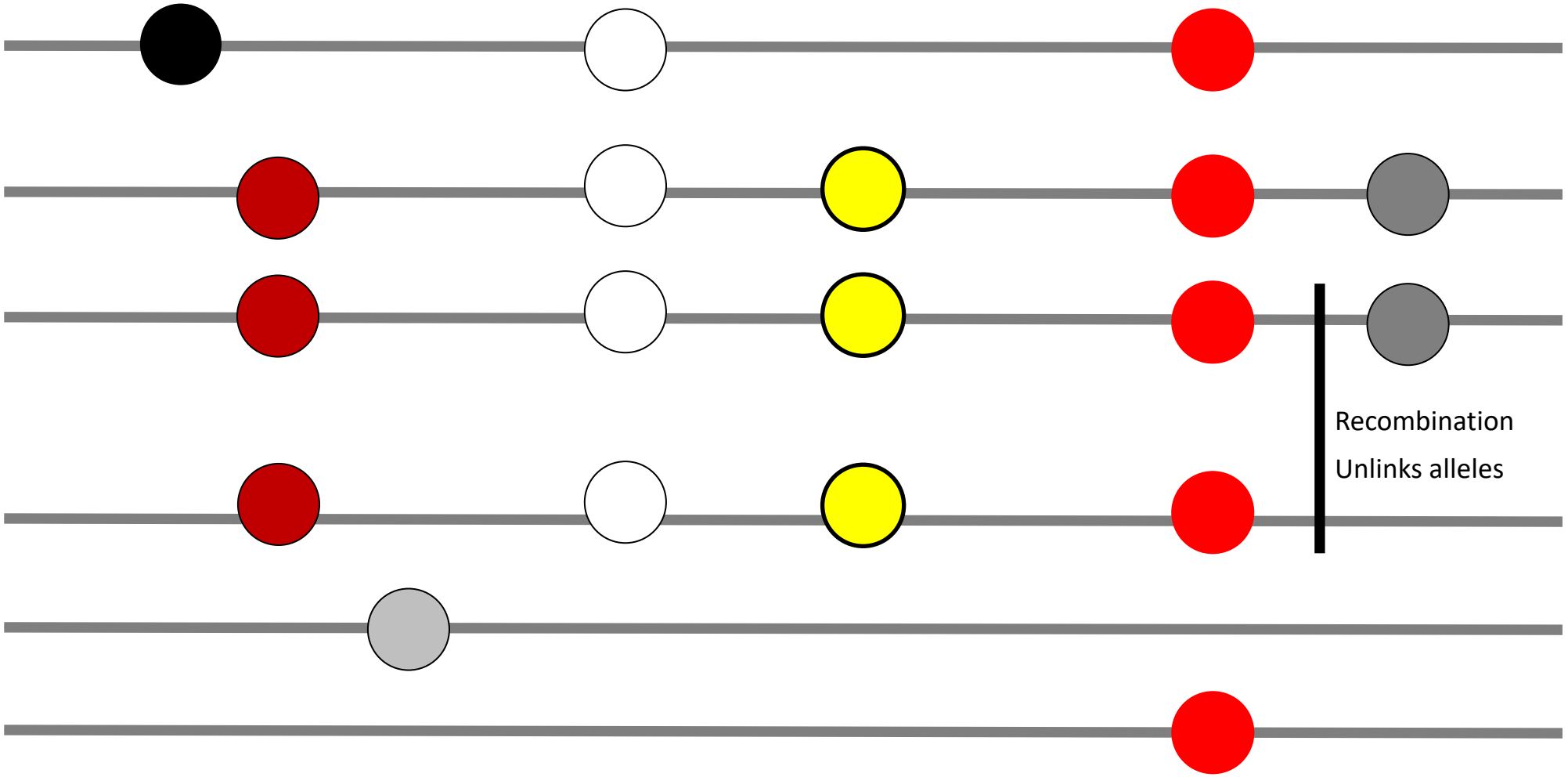
See: https://en.wikipedia.org/wiki/Selective_sweep



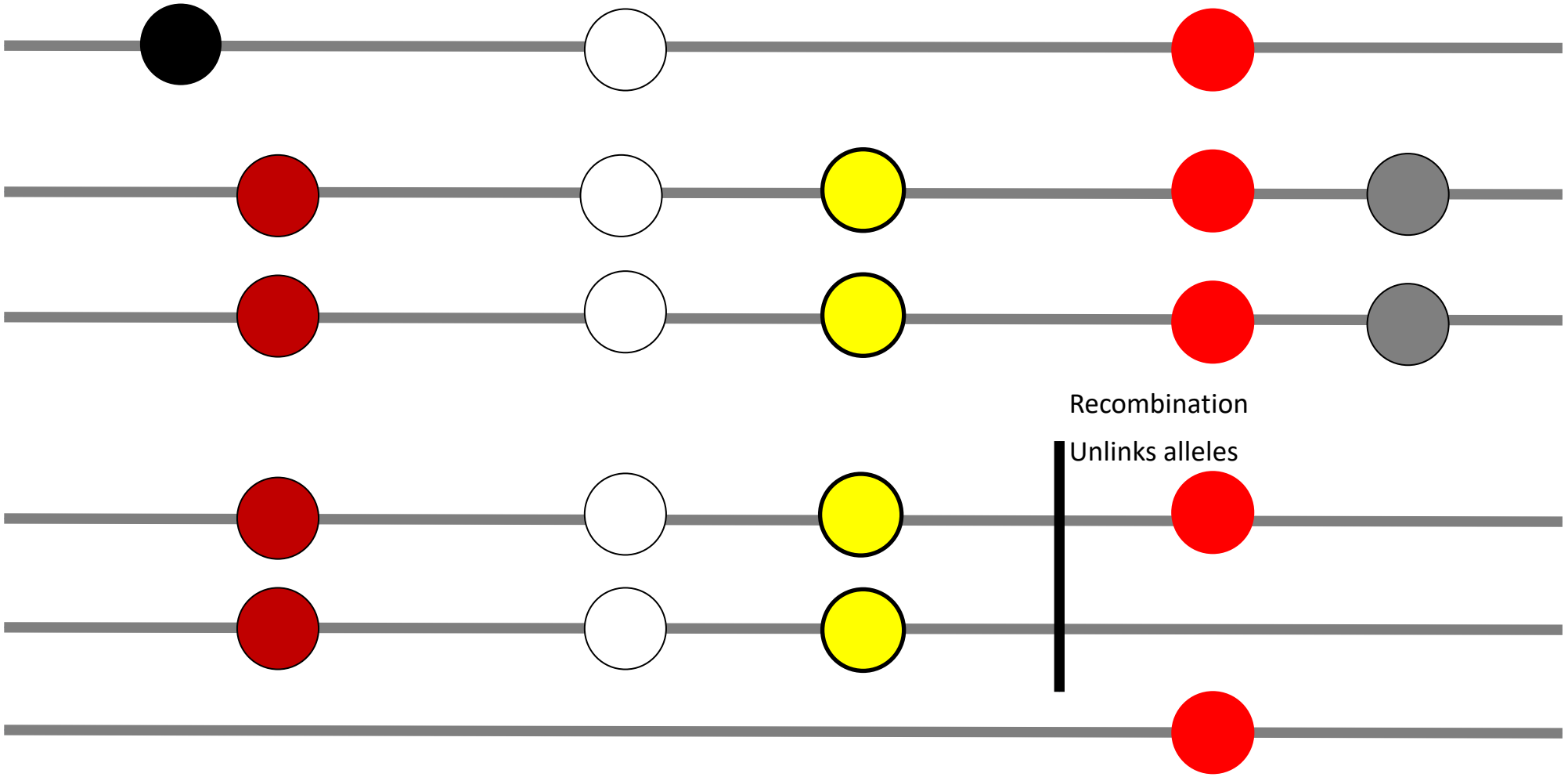
Affect of linkage in a sweep



Affect of linkage in a sweep

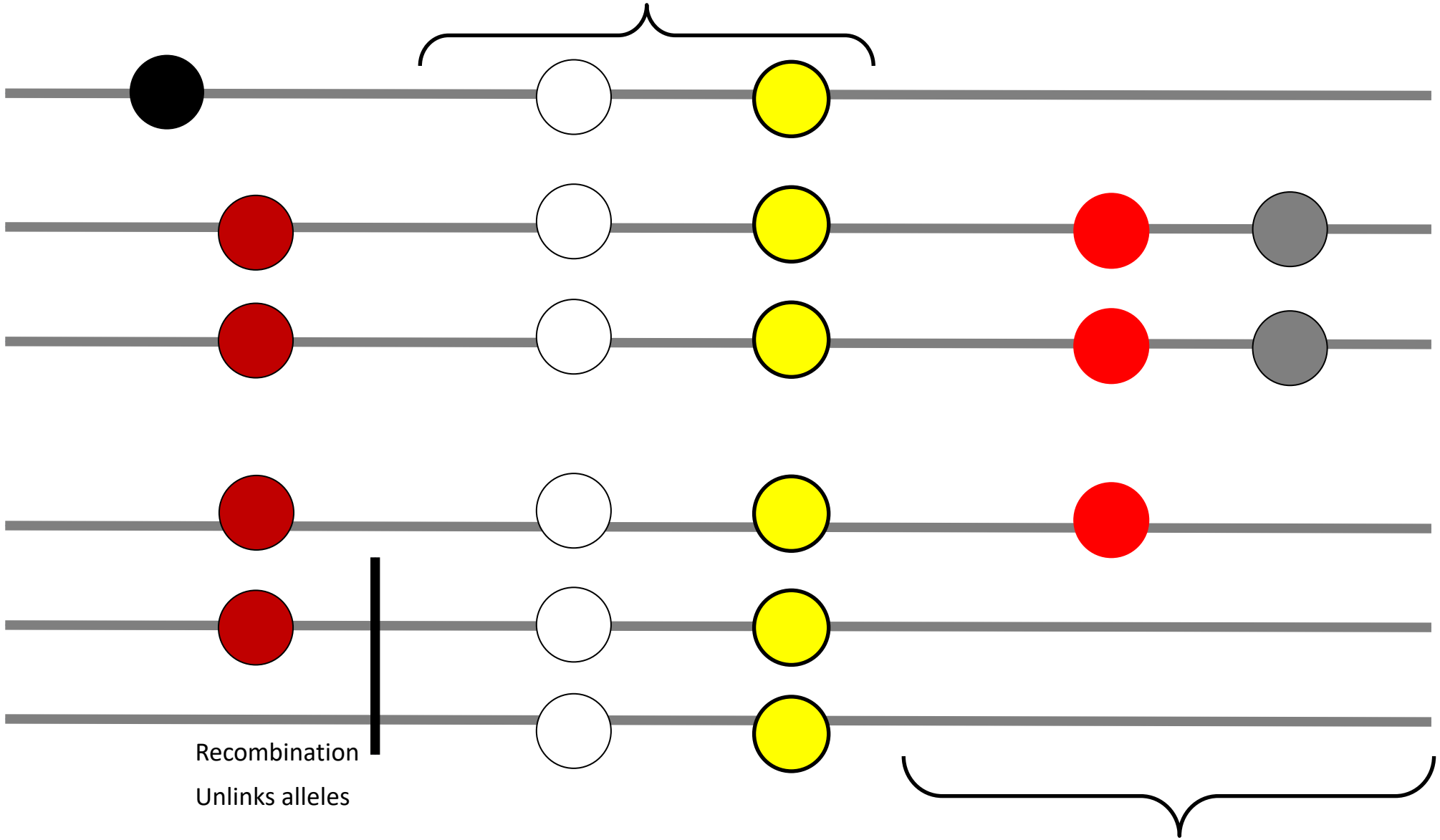


Affect of linkage in a sweep



Affect of linkage in a sweep

Very close to advantageous allele: very low diversity (none), all alleles linked



Affect of linkage in a sweep

Near to advantageous allele:
some loss of diversity, increased linkage

Questions?