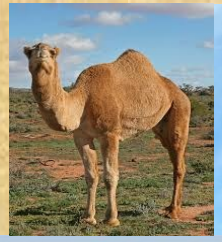
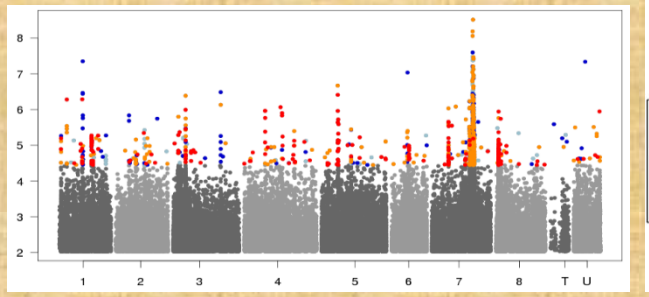
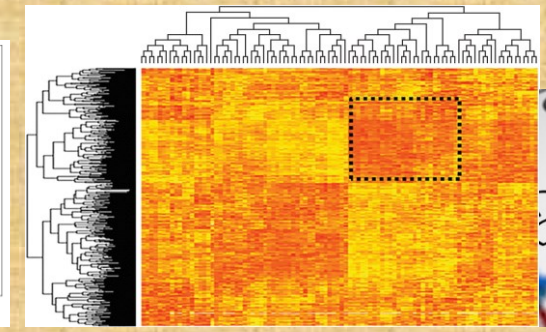


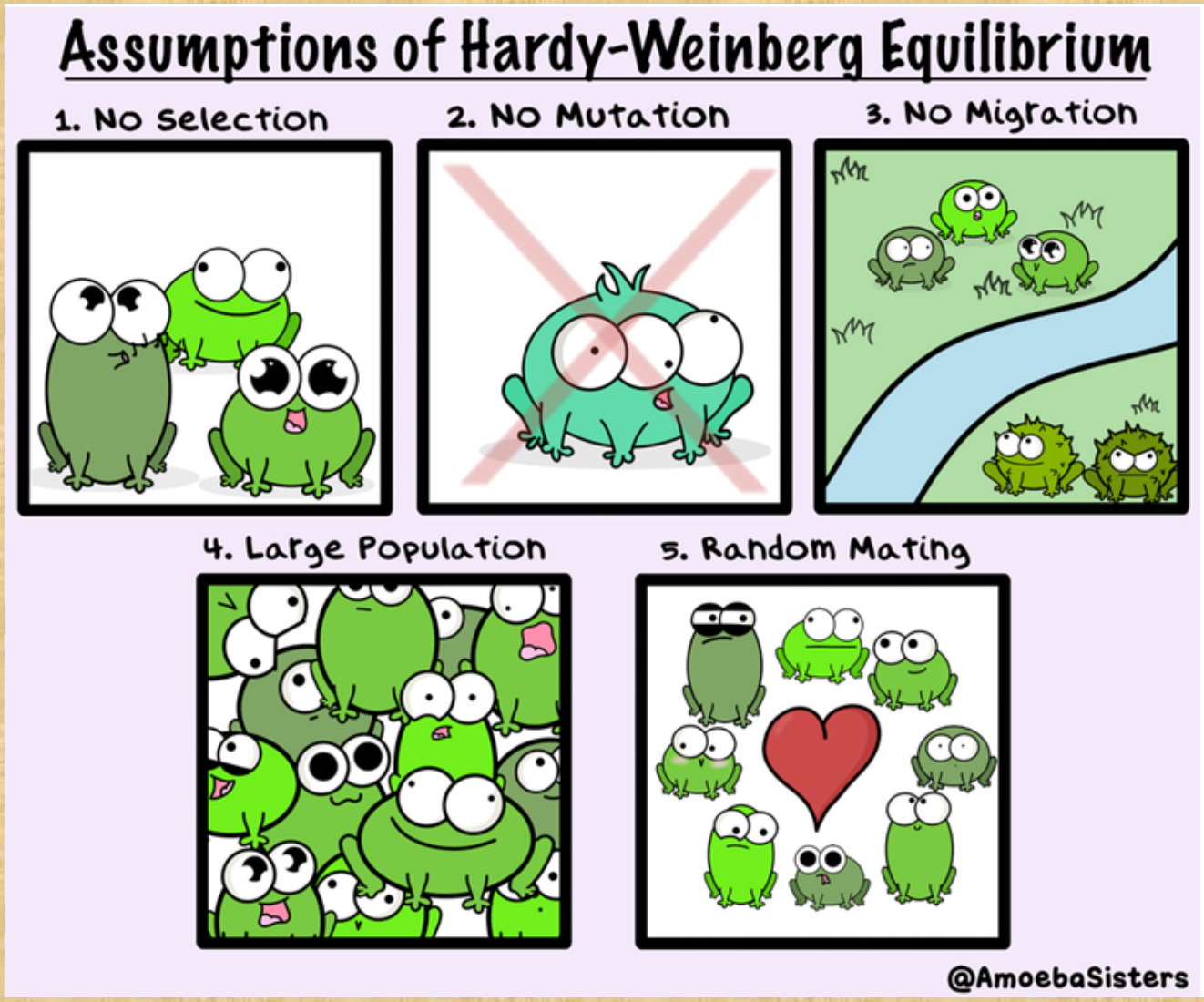
Methodologies and concepts in population genomics and adaptomics

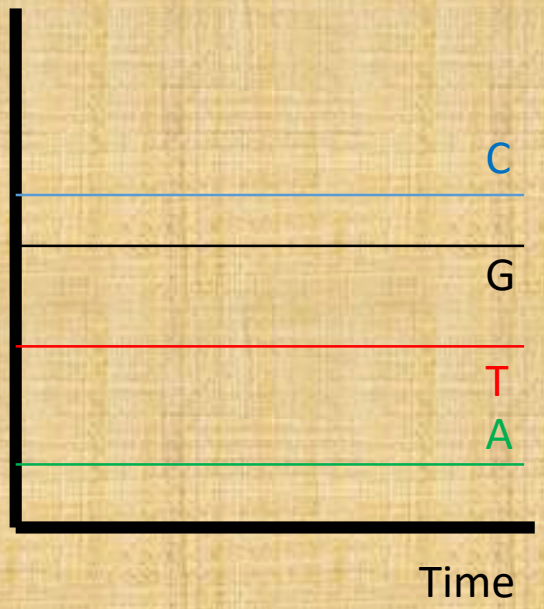
19.02.2019, Clément Lafon Placette





A *Lapalissade* model: if everything is stable, nothing changes





What if??? 😱

The neutral theory paradigm

Most variation at the molecular level does not affect fitness (not due to selection)

Genetic variation is best explained by stochastic processes

Detecting selection = testing allele frequencies against the null expectation (neutral evolution)

Duret, 2008

Plan

Demography and other factors influencing genetic variation

- Bottlenecks, migration, mating system etc
- N_e and efficacy of selection
- LD/Recombination rate

Testing for demographic cofounding factors

- Nucleotide diversity, heterozygosity
- Testing for demographic events: Tajima's D
- Estimating the "genetic load" and efficacy of selection: SFS and DFE

Detecting natural selection vs neutrality (and demography) in the genome

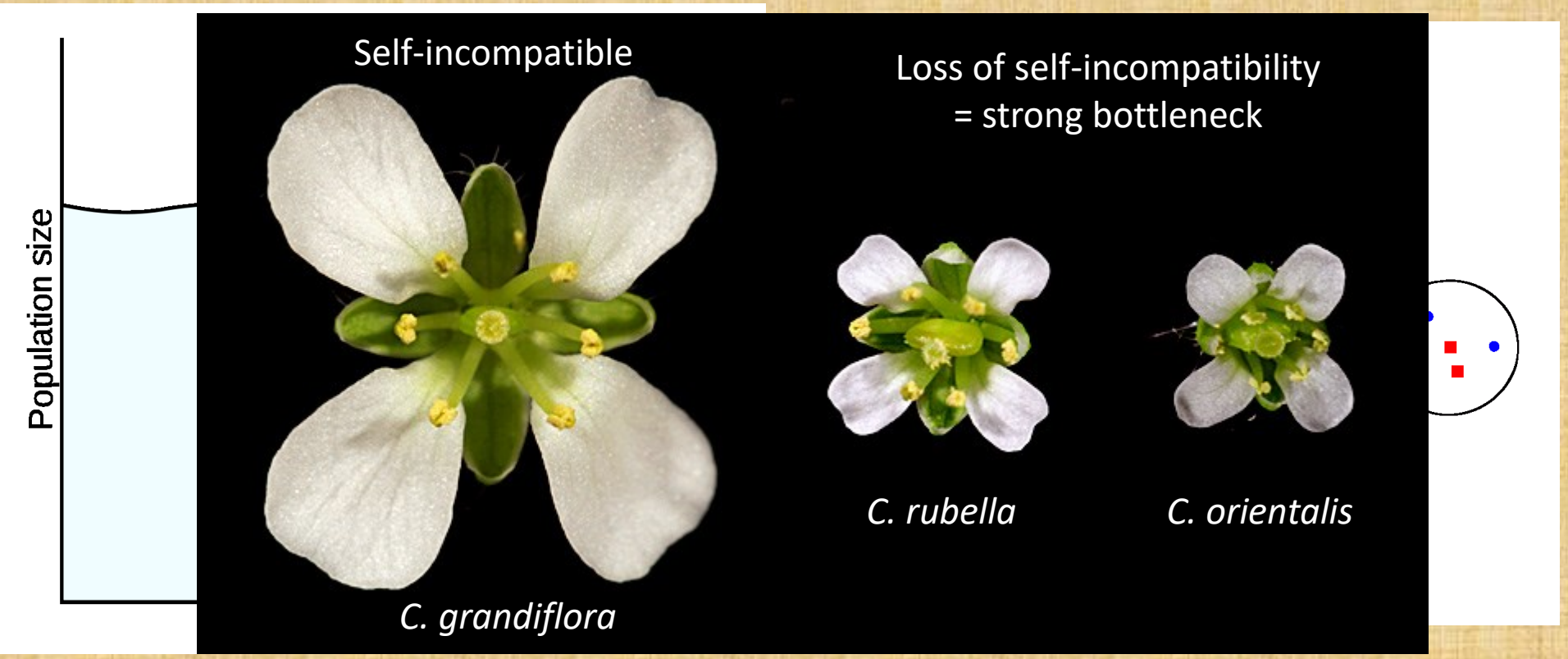
- types of selection
- selective sweeps
- The MK test
- DFE and SFS
- Fst scans

Adaptomics methods involving additional aspects than just allele frequencies

- Genetic vs phenotypic association: QTLs and GWAS
- Environmental associations

Dramatic demographic events shaping genetic variation

Bottle neck **Founder effect**



Reduced genetic diversity

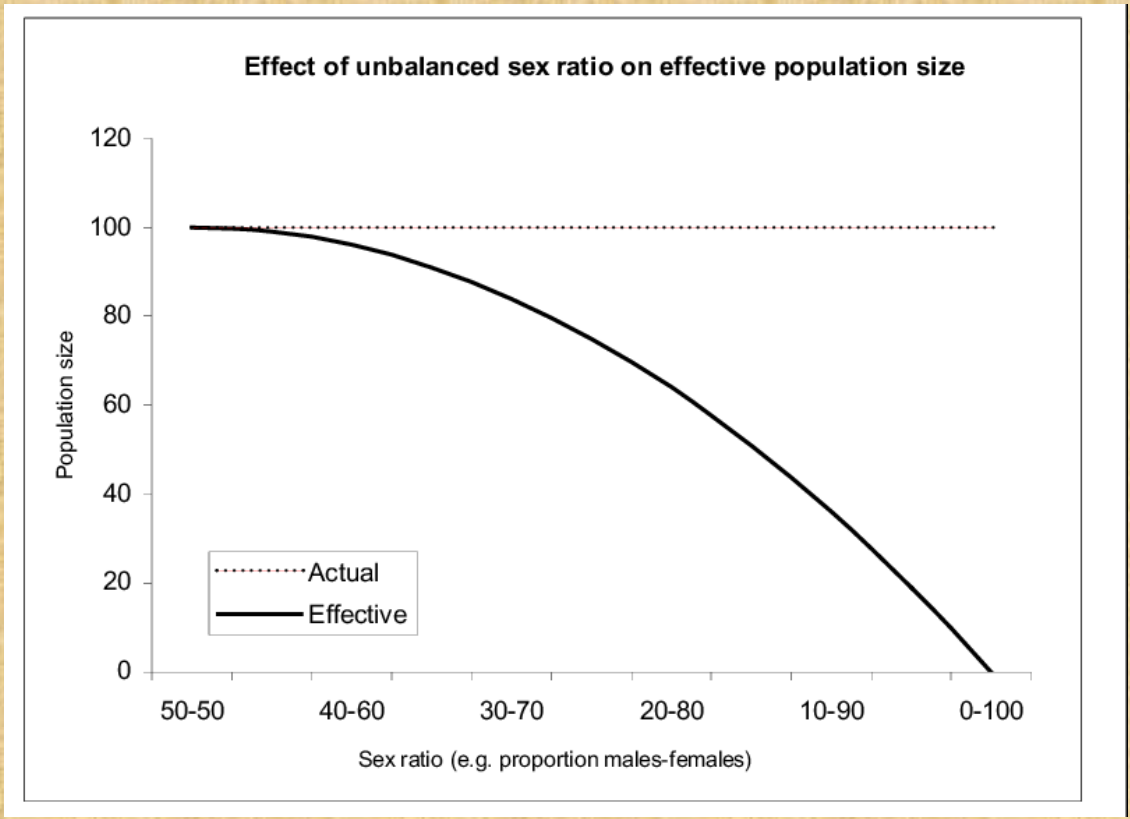
Increased homozygosity

Decreased frequency of rare alleles

Effective population size, a critical factor in pop genetics

The effective population size (N_e)
= the size of an idealised population , which would give rise to the same rate of inbreeding and the same rate of change in allele frequencies

N_e of humans?

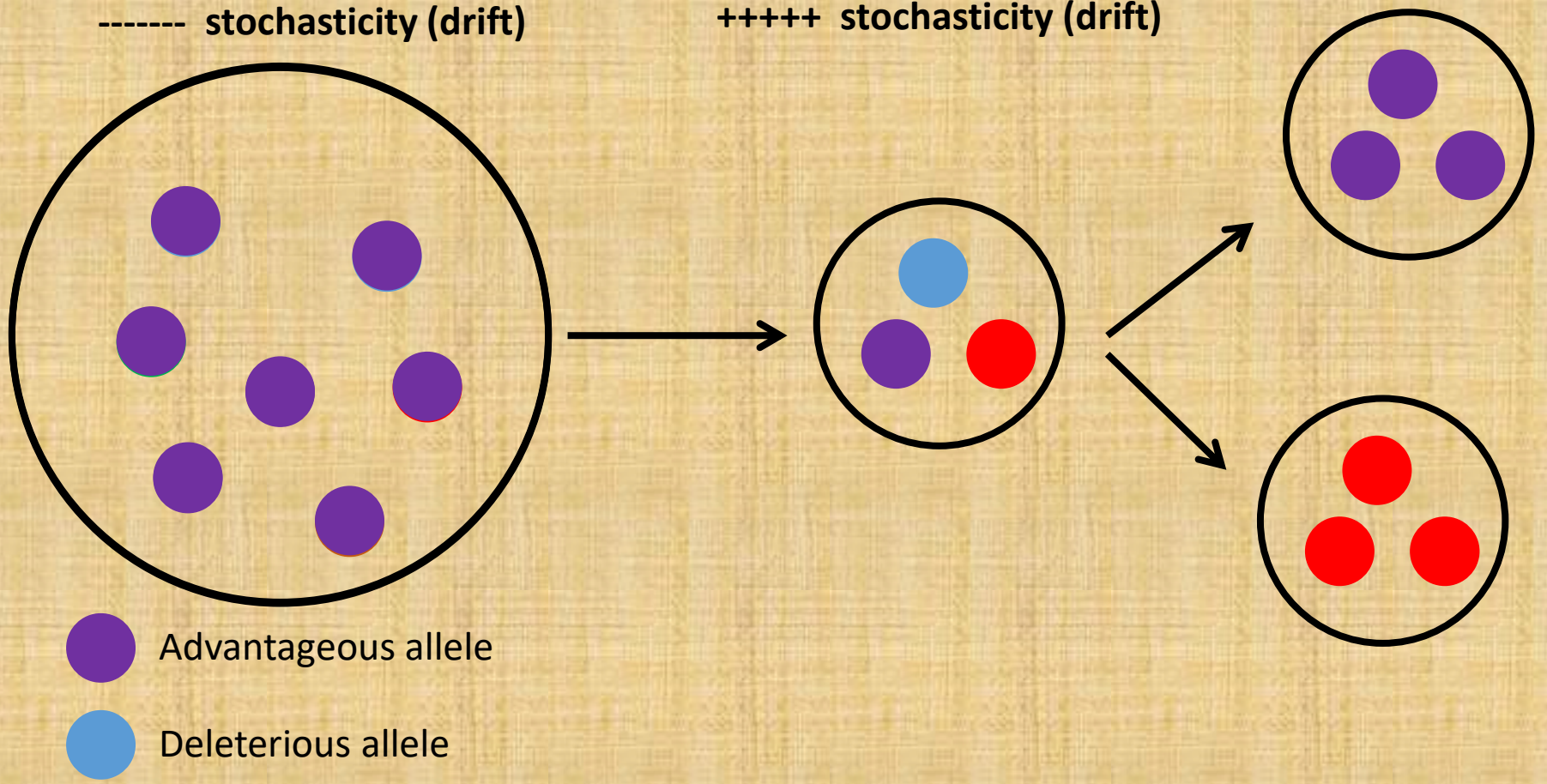


Decker, 2006

Effective population size and efficacy of selection

+++++ selection
------- stochasticity (drift)

-------- selection
+++++ stochasticity (drift)

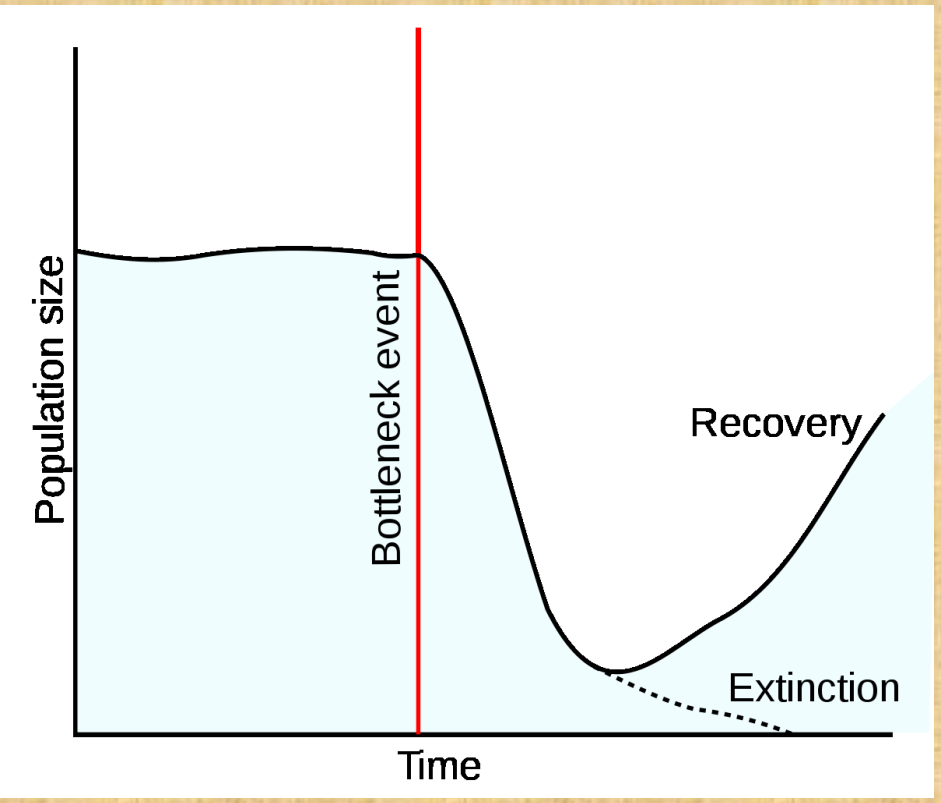


Advantageous allele

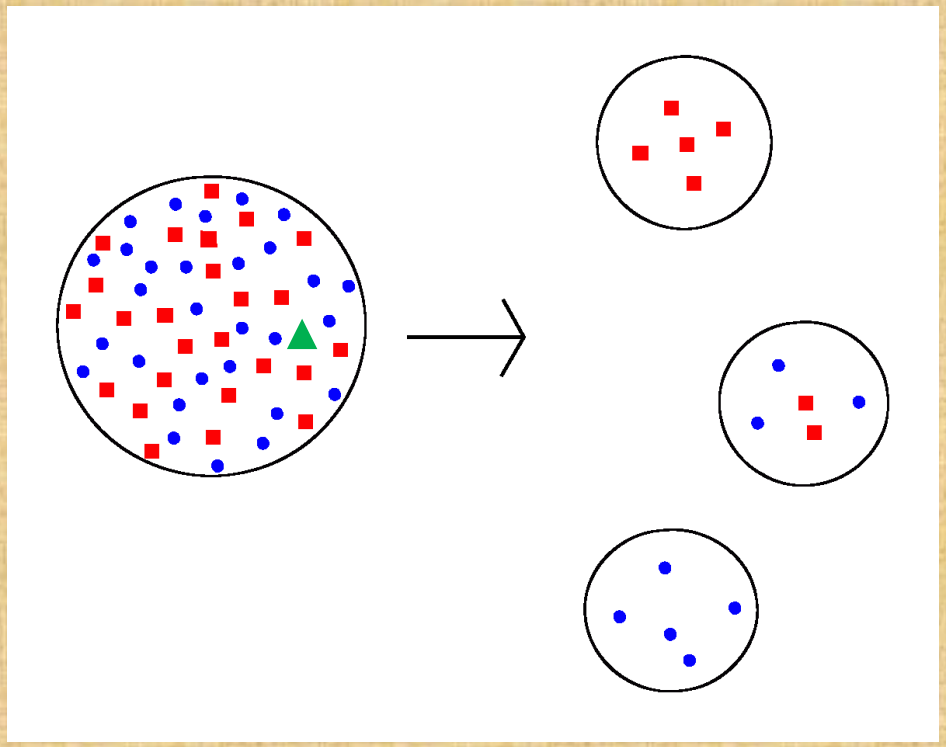
Deleterious allele

Dramatic demographic events shaping genetic variation

Bottle neck



Founder effect

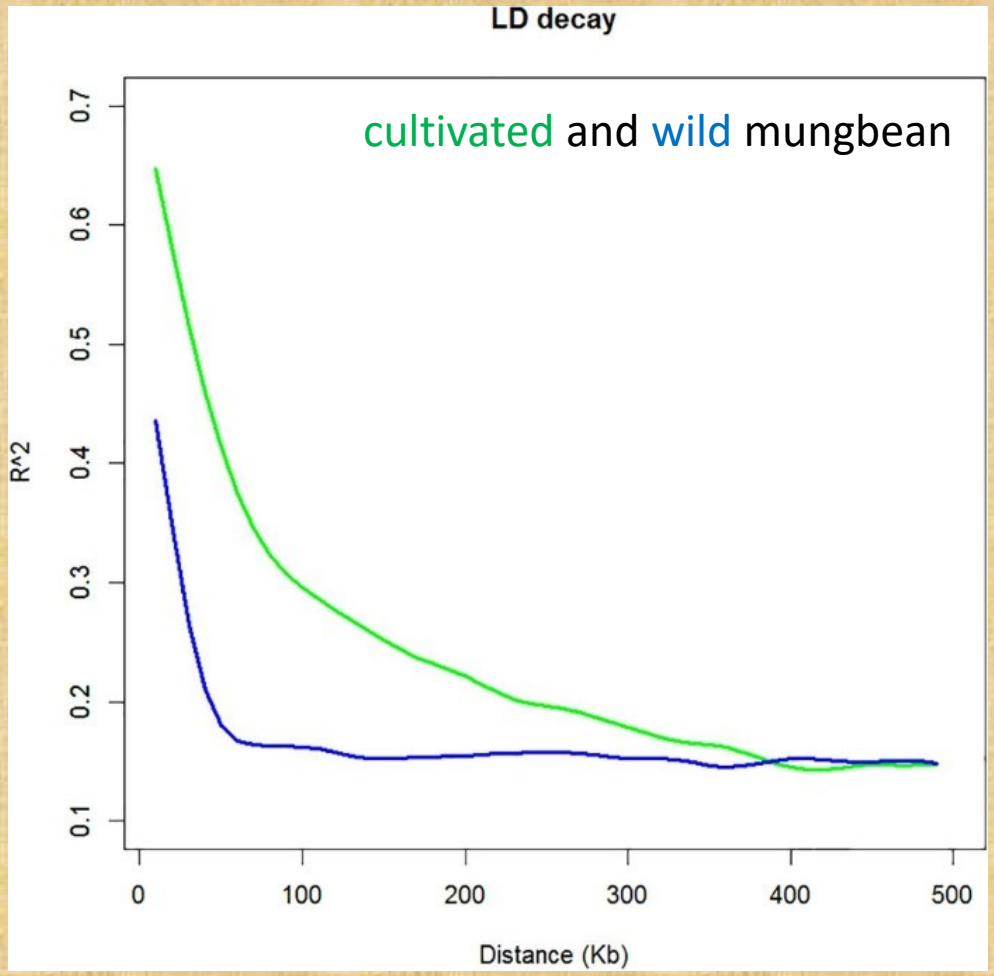


Reduced genetic diversity

Increased homozygosity

Decreased frequency of rare alleles

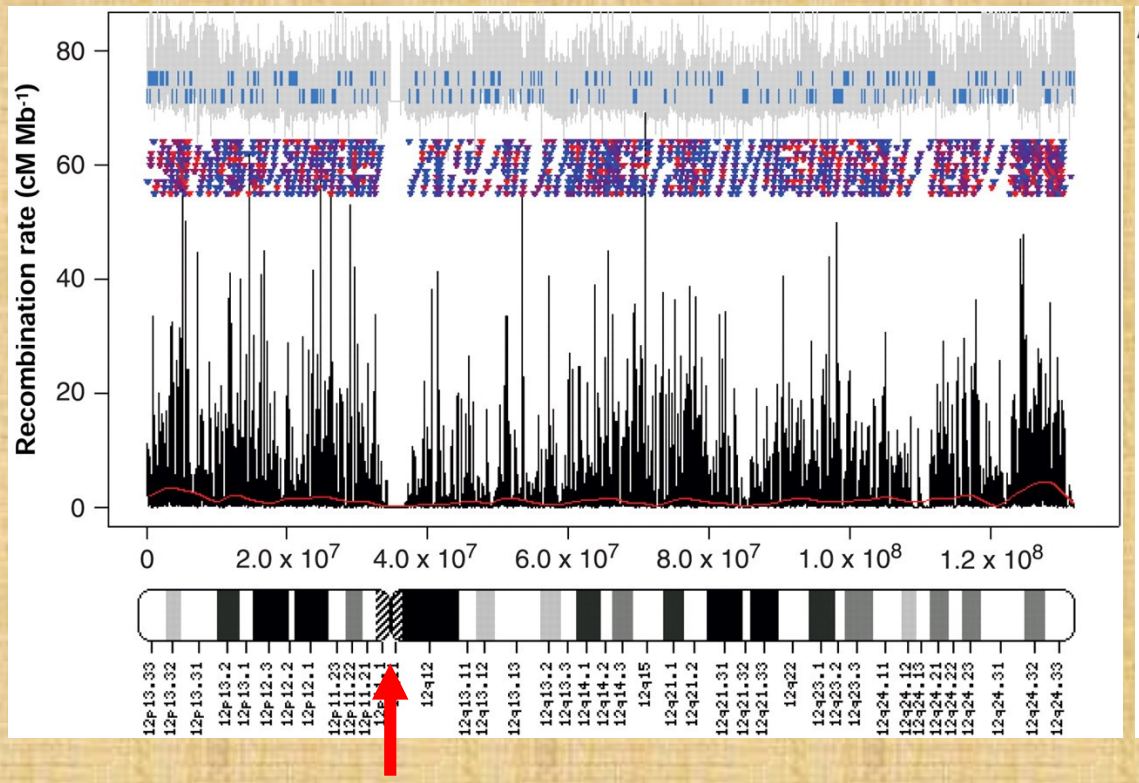
Linkage disequilibrium: a marker for effective recombination



"Effective" recombination depends on:
- the actual recombination process
- genetic diversity

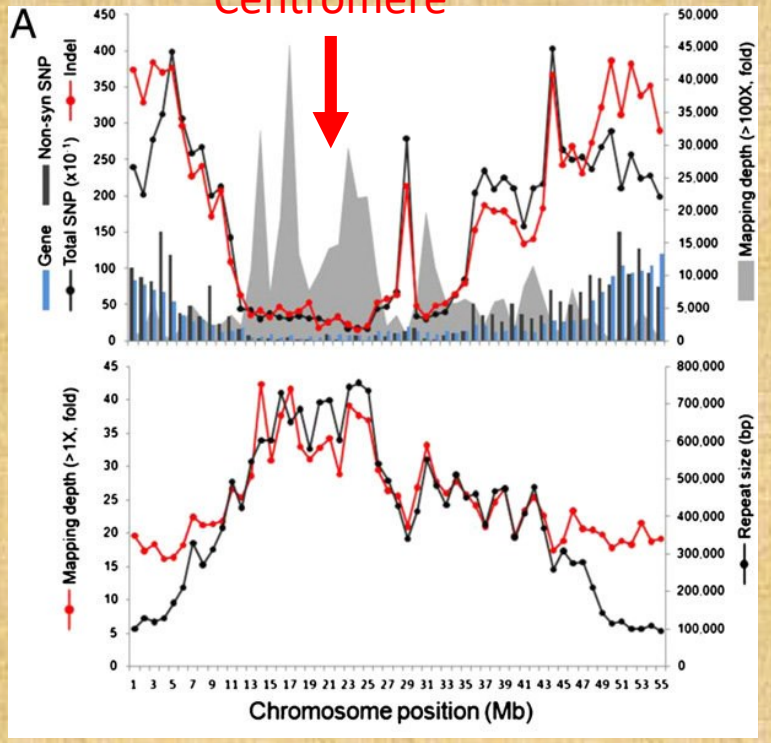
Noble et al., 2018

Recombination rate varies across genome; Influences genetic diversity



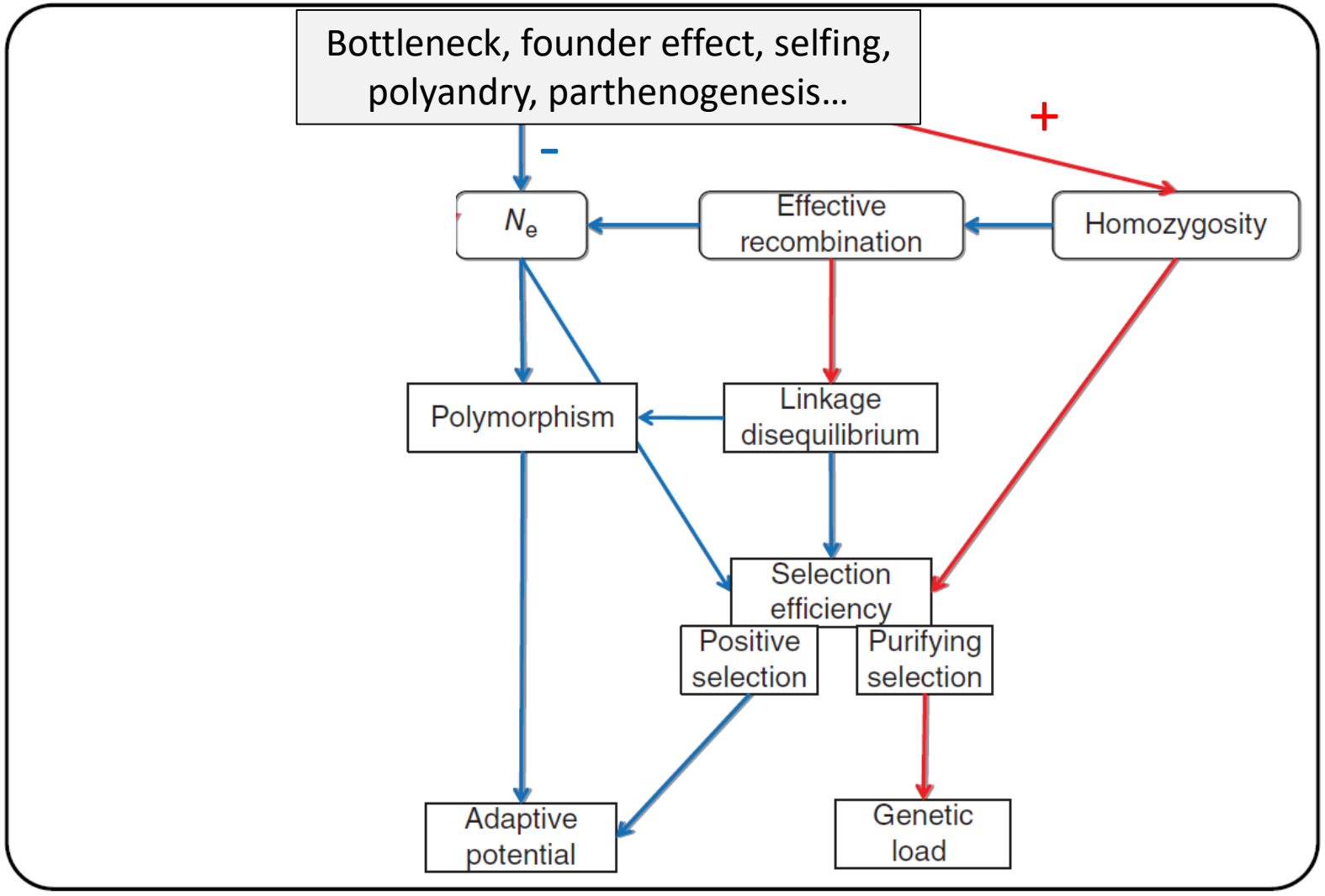
Centromere

Centromere



Myers et al., 2005; Kim et al., 2010

Demographic events play a major role in population genetics



Burgarella and Glémin, 2017

Plan

Demography and other factors influencing genetic variation
- Bottlenecks, migration, mating system etc
- N_e and efficacy of selection
- LD/Recombination rate

Testing for demographic cofounding factors
- Nucleotide diversity, heterozygosity
- Testing for demographic events: Tajima's D
- Estimating the "genetic load" and efficacy of selection: SFS and DFE

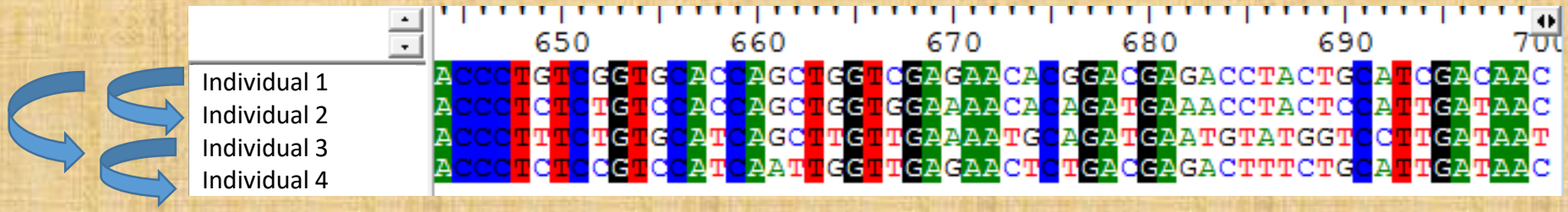
Detecting natural selection vs neutrality (and demography) in the genome
- types of selection
- selective sweeps
- The MK test
- DFE and SFS
- Fst scans

Adaptomics methods involving additional aspects than just allele frequencies
- Genetic vs phenotypic association: QTLs and GWAS
- Environmental associations

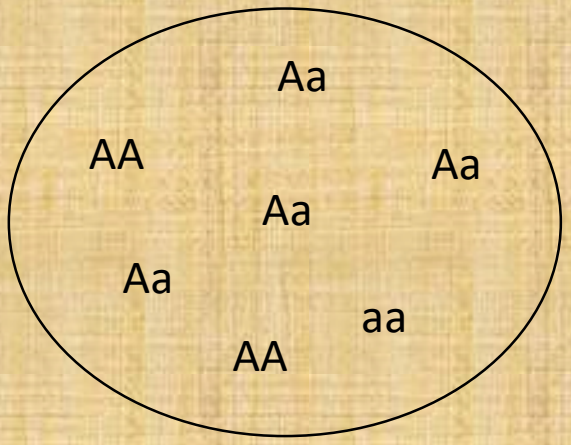
Nucleotide diversity, heterozygosity, can indicate demographic processes going on

Nucleotide diversity

Π, average pairwise differences



Heterozygosity



p = f(A); q = f(a)

$H_{observed}$ = f(Aa) = ?

$H_{expected} = 1 - \sum_{i=1}^{n} (f_i)^2$ for *n* alleles

$H_e = 1 - (p^2+q^2)$ = ?

Tajima's D

D ≈ observed genetic variation – expected genetic variation (neutrality)

D = average pairwise difference among individuals – expected number of variable sites for a given number of individuals

Sum of pairwise differences

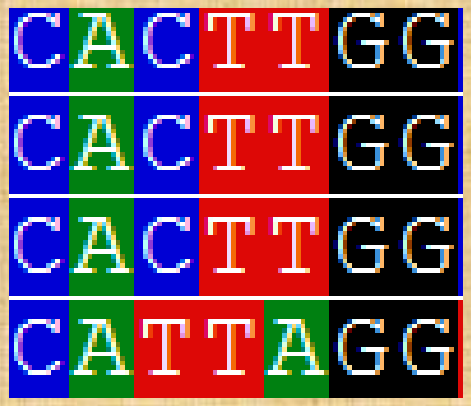
No of variable sites

$$D = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2} - \frac{S}{\sum_{i=1}^{n-1} 1/i}$$

No of individuals

n = 4

S = 2



D = [(2+2+2)/(4*3/2)] - [2/(1/1+1/2+1/3)]

D = 1 – 1.09 = -0.09 ≈ 0

Observed variation = expected

→ Population at equilibrium

D ≈ observed genetic variation – expected genetic variation (neutrality)

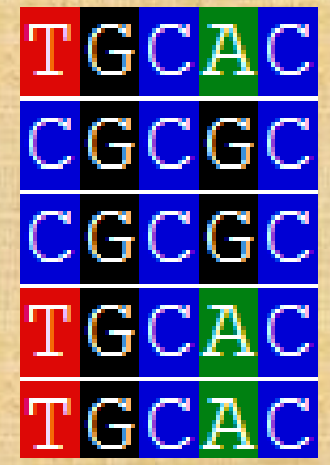
D = average pairwise difference among individuals – expected number of variable sites for a given number of individuals

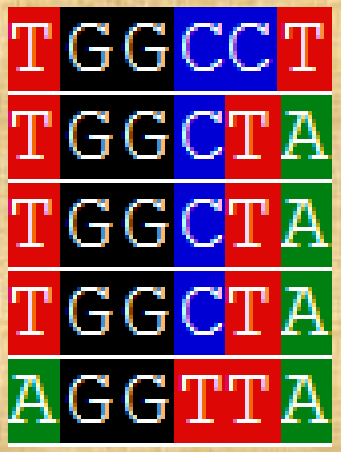
Sum of pairwise differences

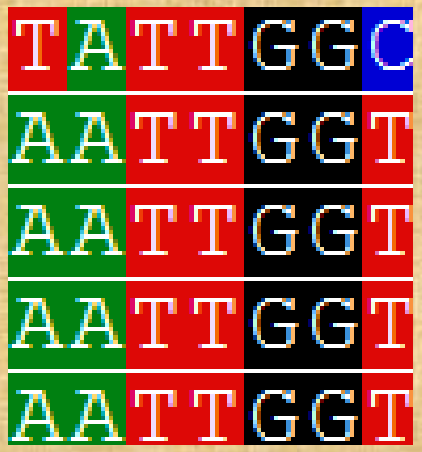
No of variable sites

$$D = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2} - \frac{S}{\sum_{i=1}^{n-1} 1/i}$$

No of individuals







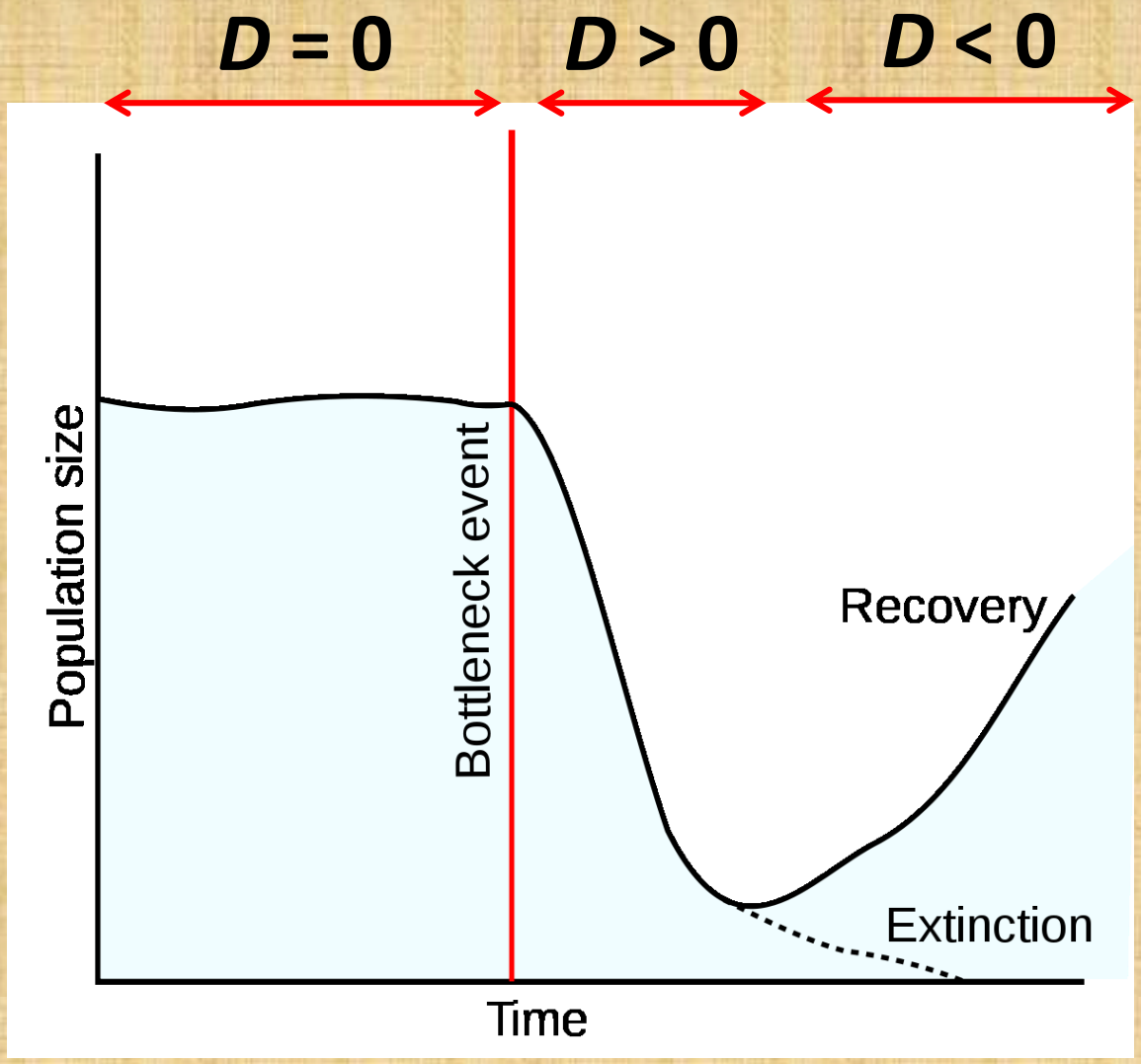
***D* > 0**

More variation than expected considering the no of individuals

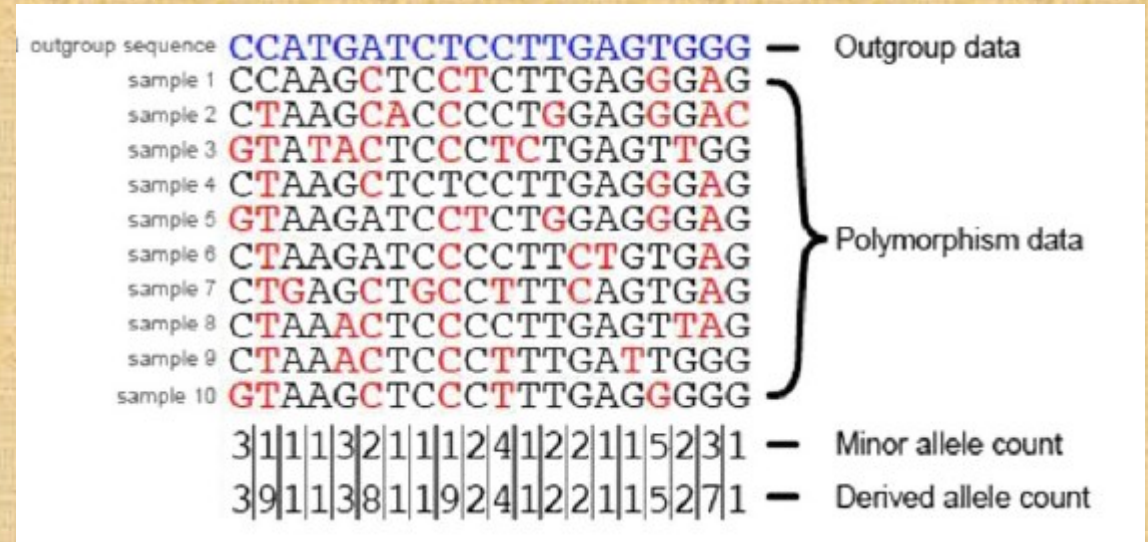
***D* < 0**

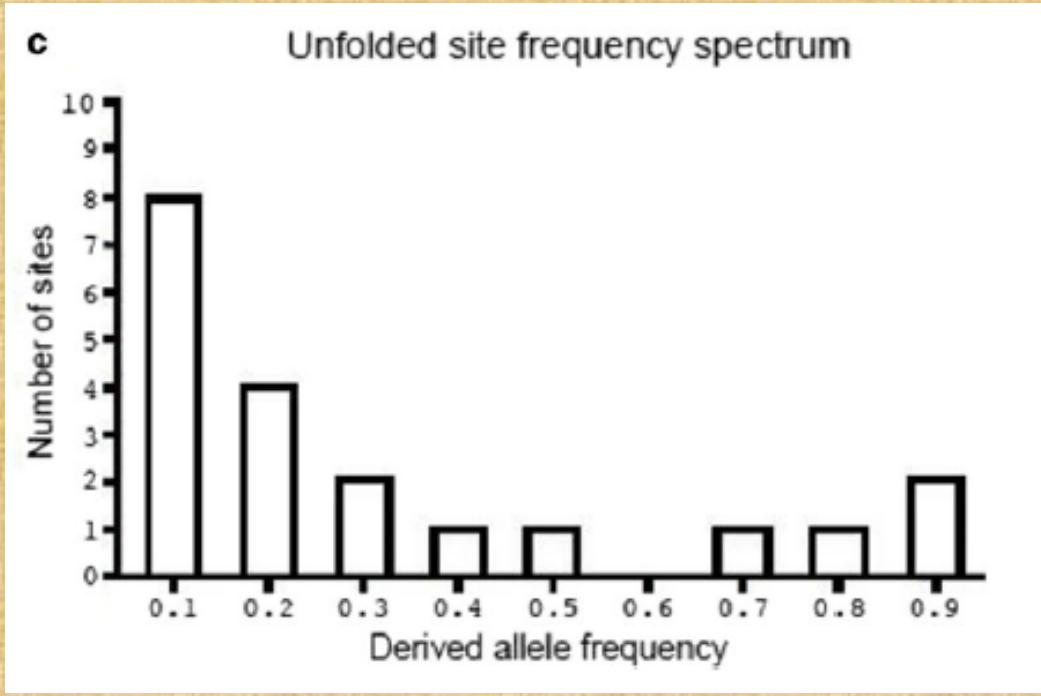
Less variation than expected considering the no of individuals

Tajima's *D* and demographic events



Estimating the Site Frequency Spectrum



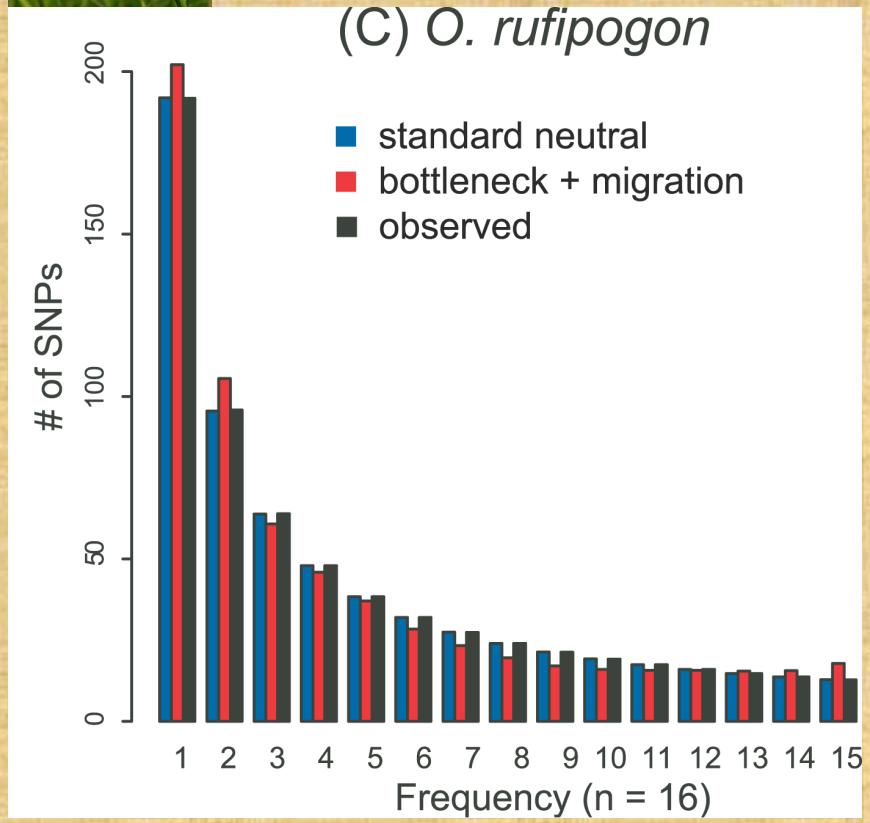


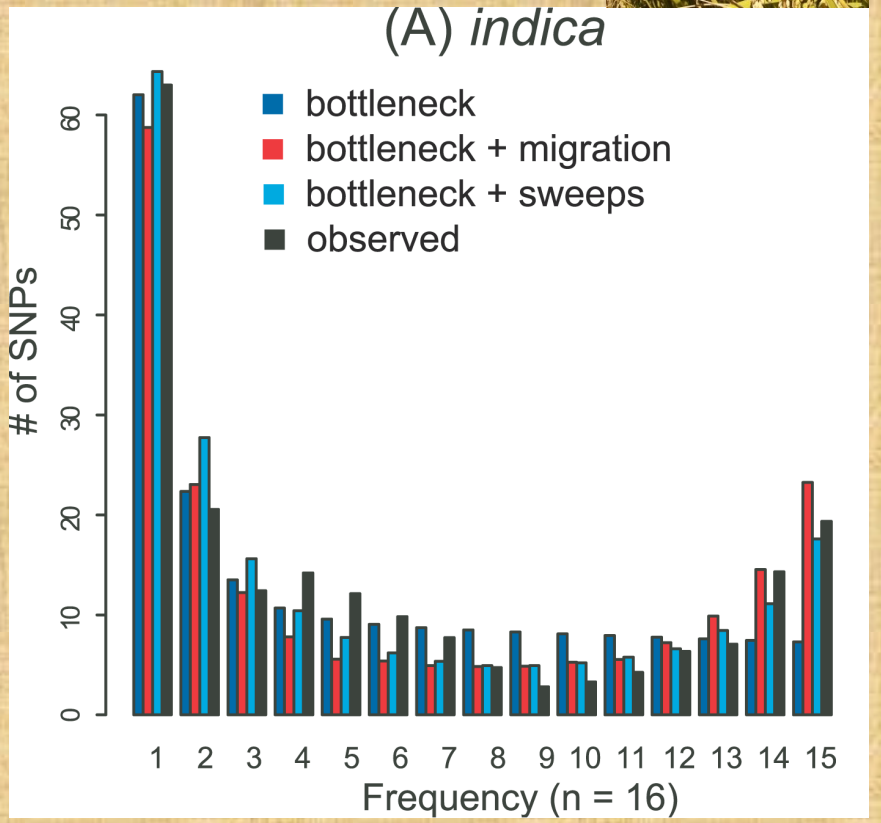
Booker et al., 2017

The SFS can tell about demographic events

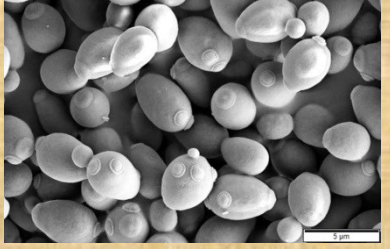
Wild rice

Cultivated rice



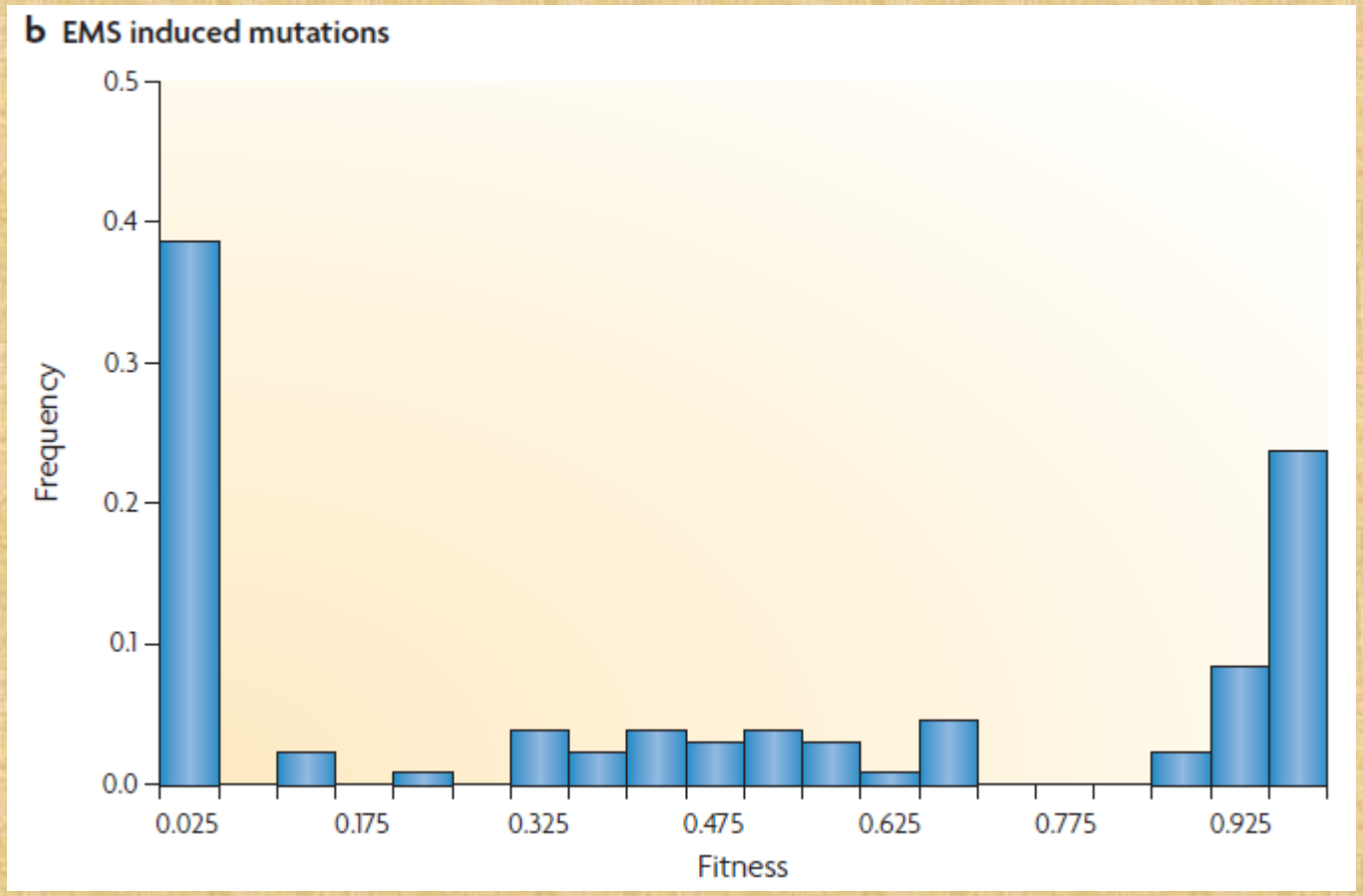


Bottleneck: few high frequency mutations



Most new mutations are deleterious

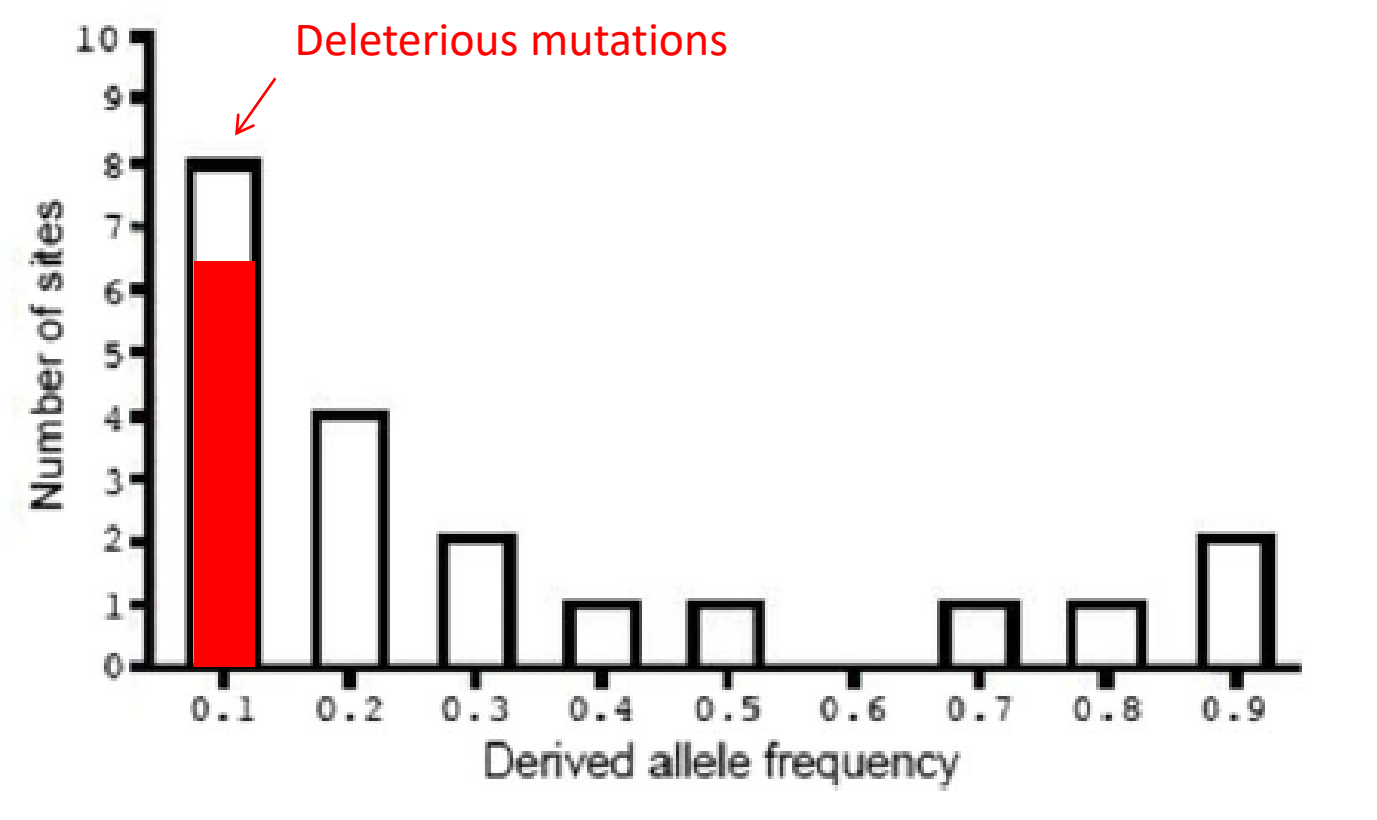
S. cerevisiae



It is not always possible to determine the effect of mutations on fitness

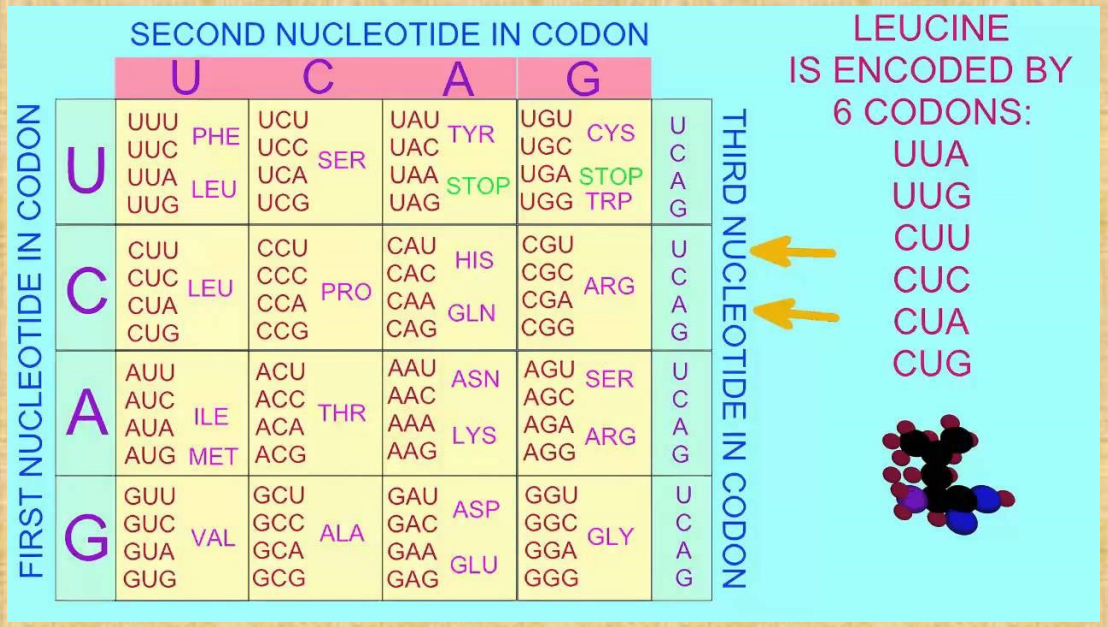
Inference from allele frequencies in pop genomics

Estimating the Site Frequency Spectrum



Booker et al., 2017

Syn vs non syn. mutations



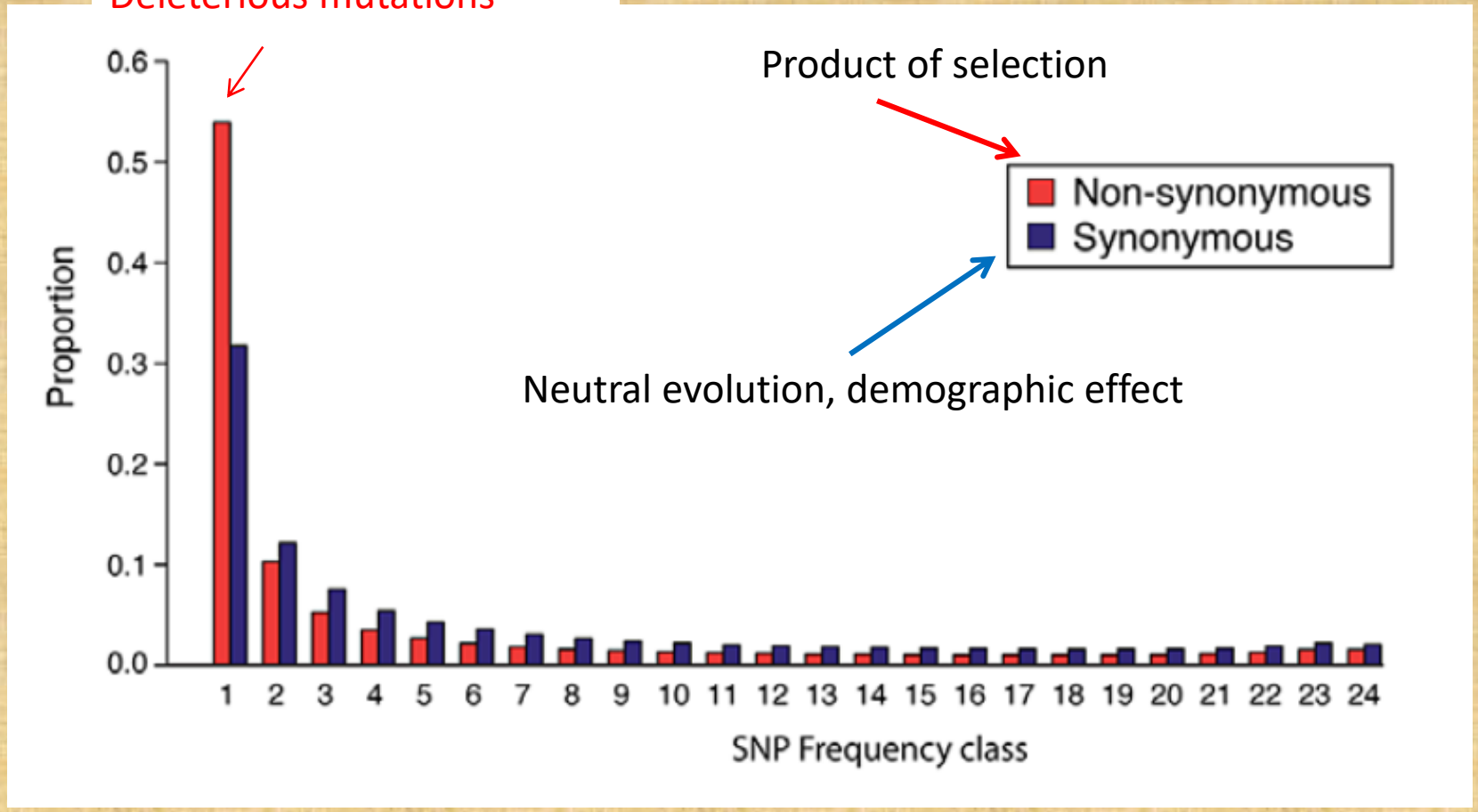
Synonymous: same AA encoded; non-synonymous: different AA

Ratio between syn and non-syn will tell about a given gene evolution:

syn = neutral mutations, due to population demography, mutation rate…

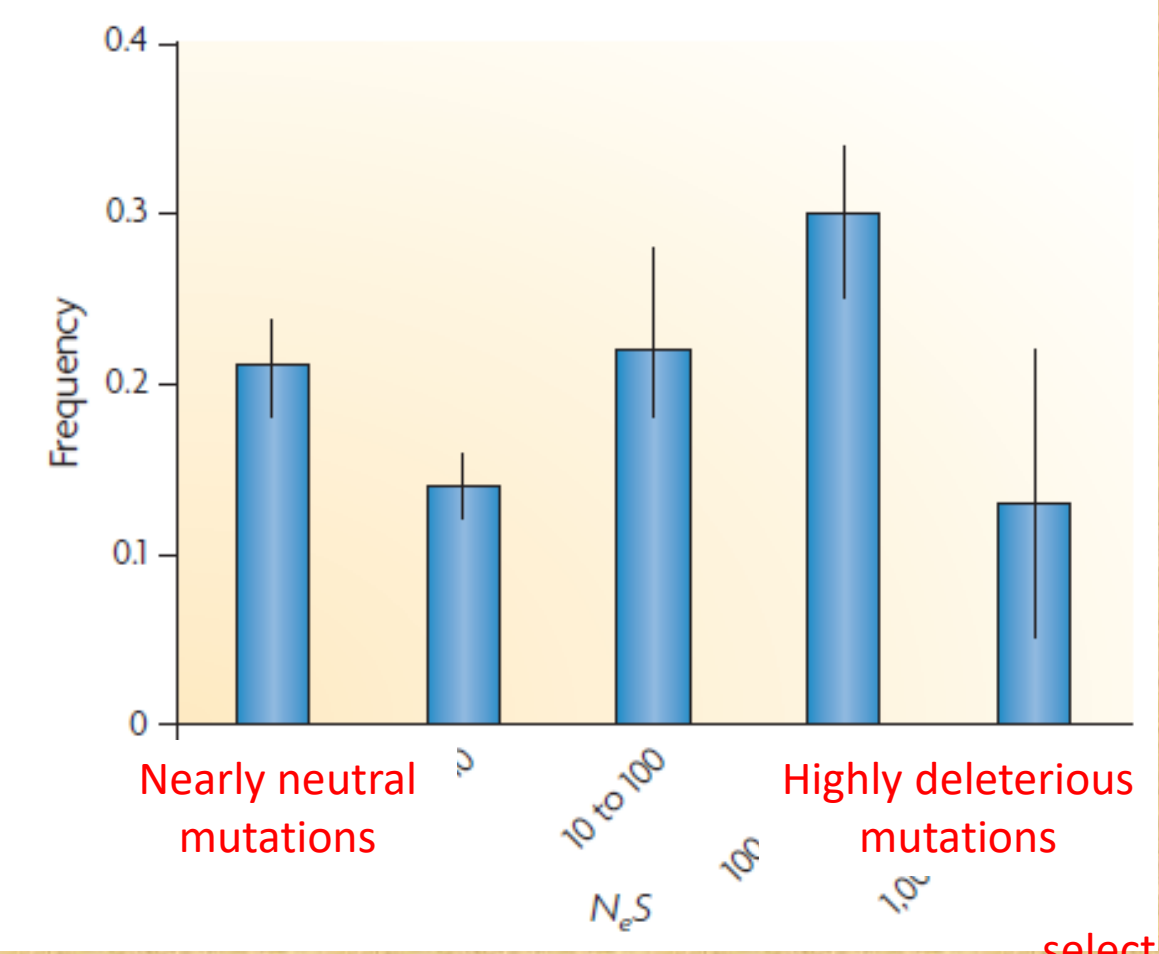
non syn = important for prot function; can be purged (if deleterious) or selected (if advantageous)

SFS to infer the frequency of deleterious mutations



Hollister et al., 2012

The distribution of fitness effect



selection coefficient

The strength of selection: N_eS

Eyre-Walker and Keightley, 2007

Plan

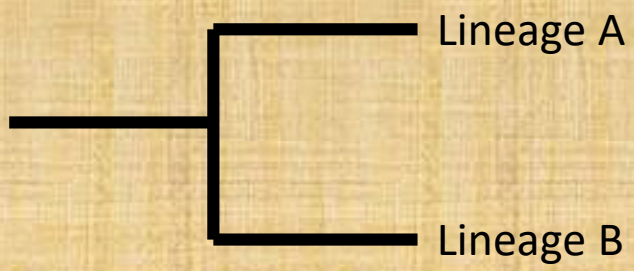
Demography and other factors influencing genetic variation
- Bottlenecks, migration, mating system etc
- N_e and efficacy of selection
- LD/Recombination rate

Testing for demographic cofounding factors
- Nucleotide diversity, heterozygosity
- Testing for demographic events: Tajima's D
- Estimating the "genetic load" and efficacy of selection: SFS and DFE

Detecting natural selection vs neutrality (and demography) in the genome
- types of selection
- selective sweeps
- the MK test
- DFE and SFS
- Fst scans

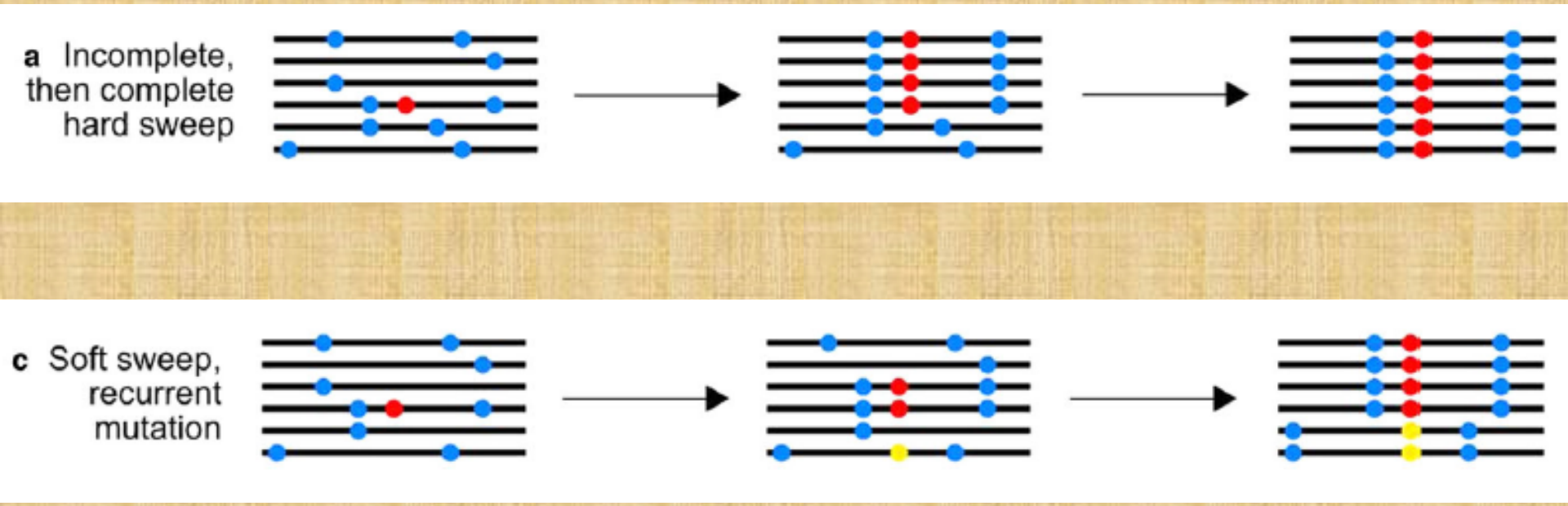
Adaptomics methods involving additional aspects than just allele frequencies
- Genetic vs phenotypic association: QTLs and GWAS
- Environmental associations

Types of selection



Type of selection	Consequence on intralineage variability	On interlineage divergence	On site frequency spectrum
Negative/Purifying	Reduced	Reduced	more low frequency derived alleles
Positive selection	Reduced	Increased	more high frequency derived alleles
Balancing selection	Increased	none	more intermediate frequency derived/ancestral alleles

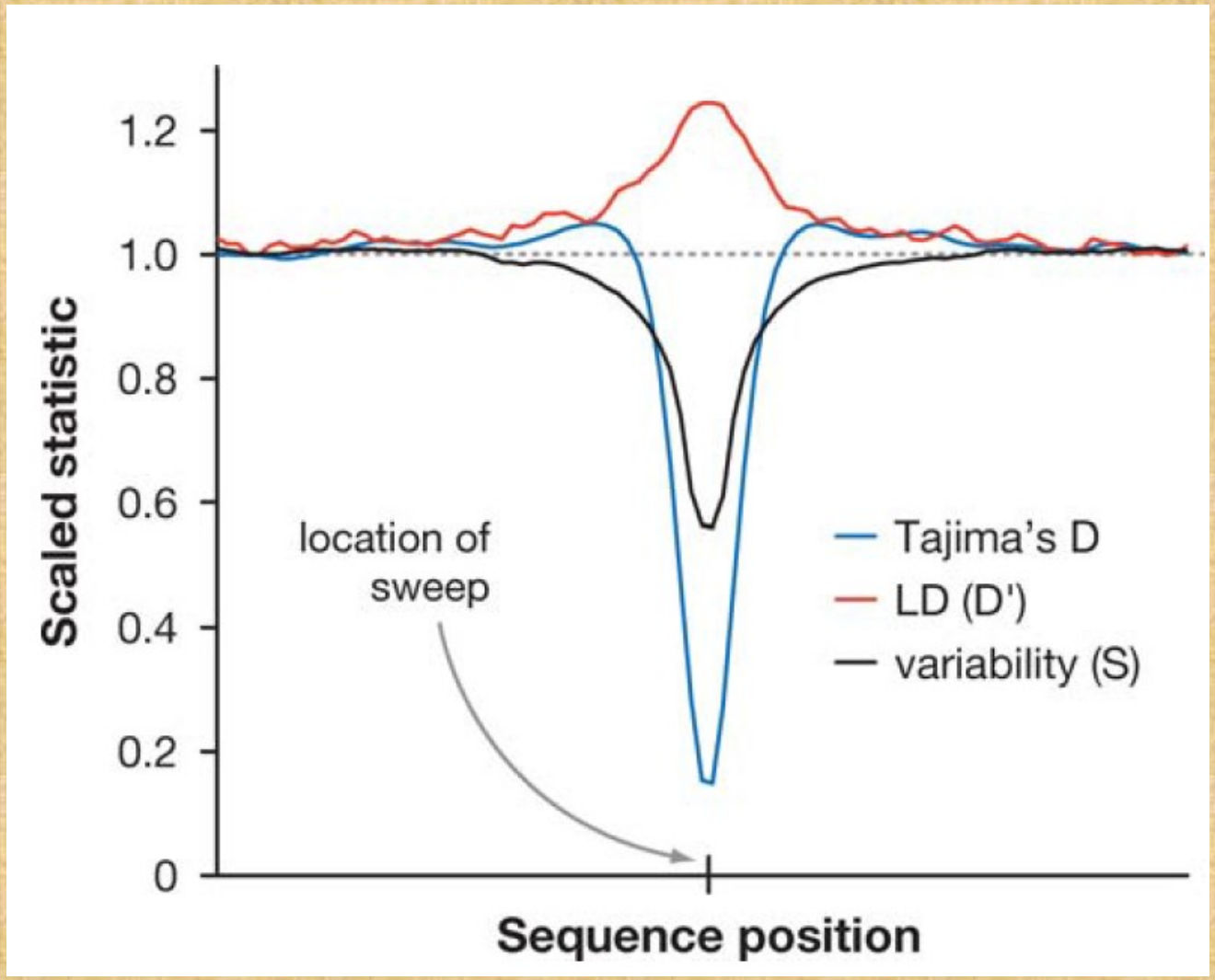
A selective sweep



Because of genetic linkage, reduction of local genetic diversity around the selected mutation

Booker et al., 2017

A typical selective sweep

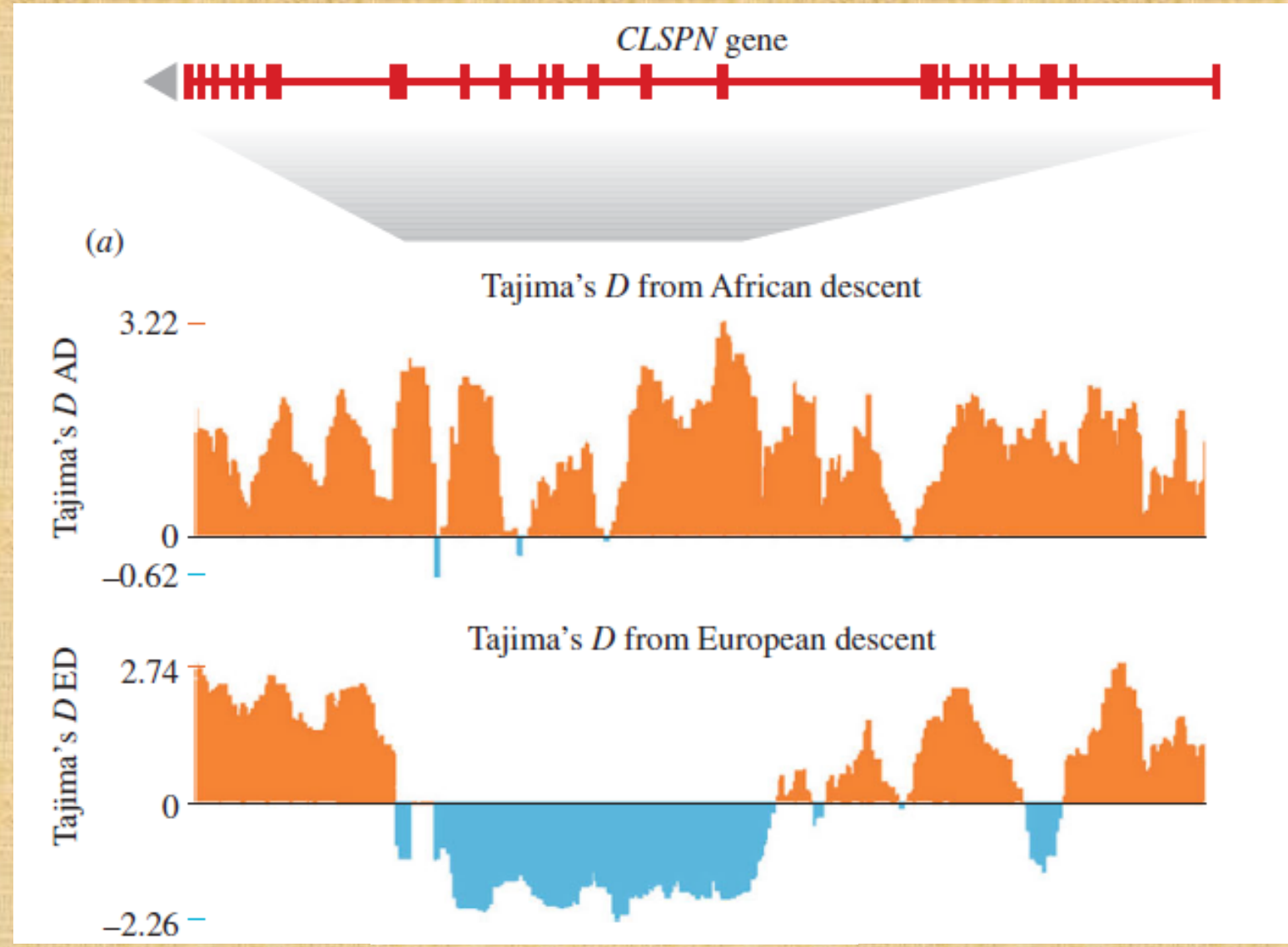


As a general rule, loci under potential selection need to be compared to the overall genome:

Demography can influence LD, Tajima's D…

Nielsen, 2005

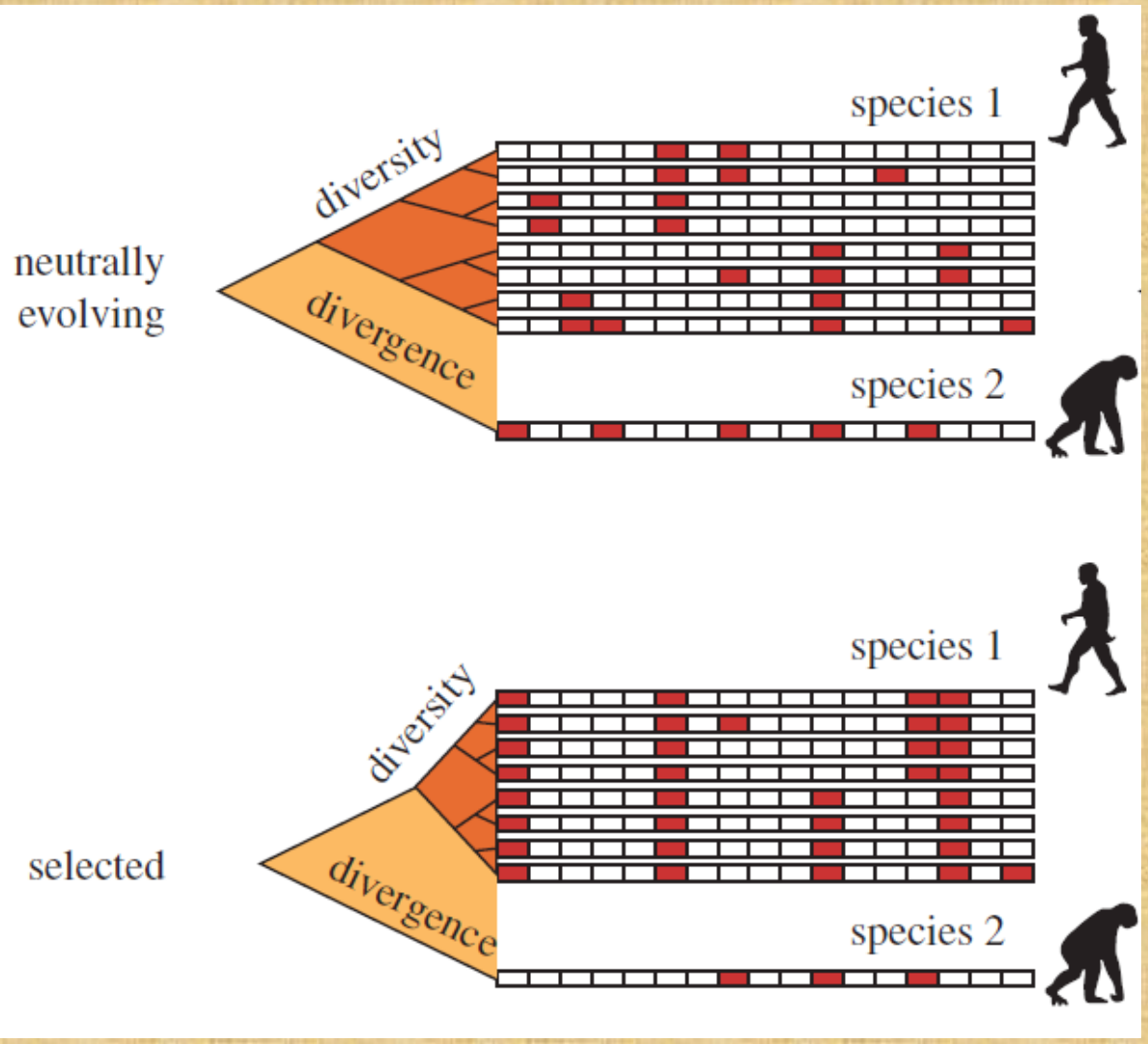
Tajima's *D* as a test for local selection

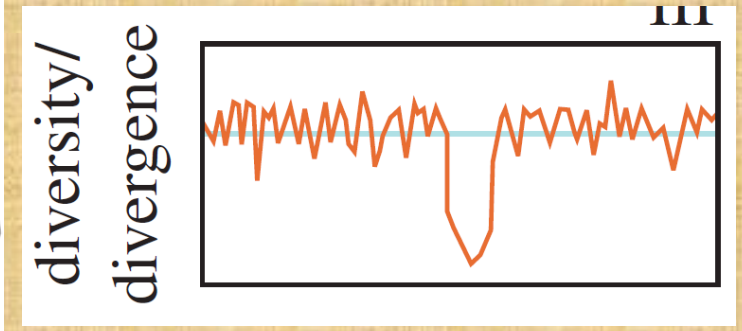


Locally, less genetic variation than expected given the number of individuals

Oleksyk et al., 2009

Testing for the ratio between interlineage divergence vs intralineage variation





Oleksyk et al., 2009

For coding genes, the rate of syn vs non syn mutations gives information on the selection process

If neutral evolution,
interlineage divergence in non synonymous site = interlineage divergence in synonymous site

$D_n / D_s = 1$

Above 1: positive selection on new non syn mutations

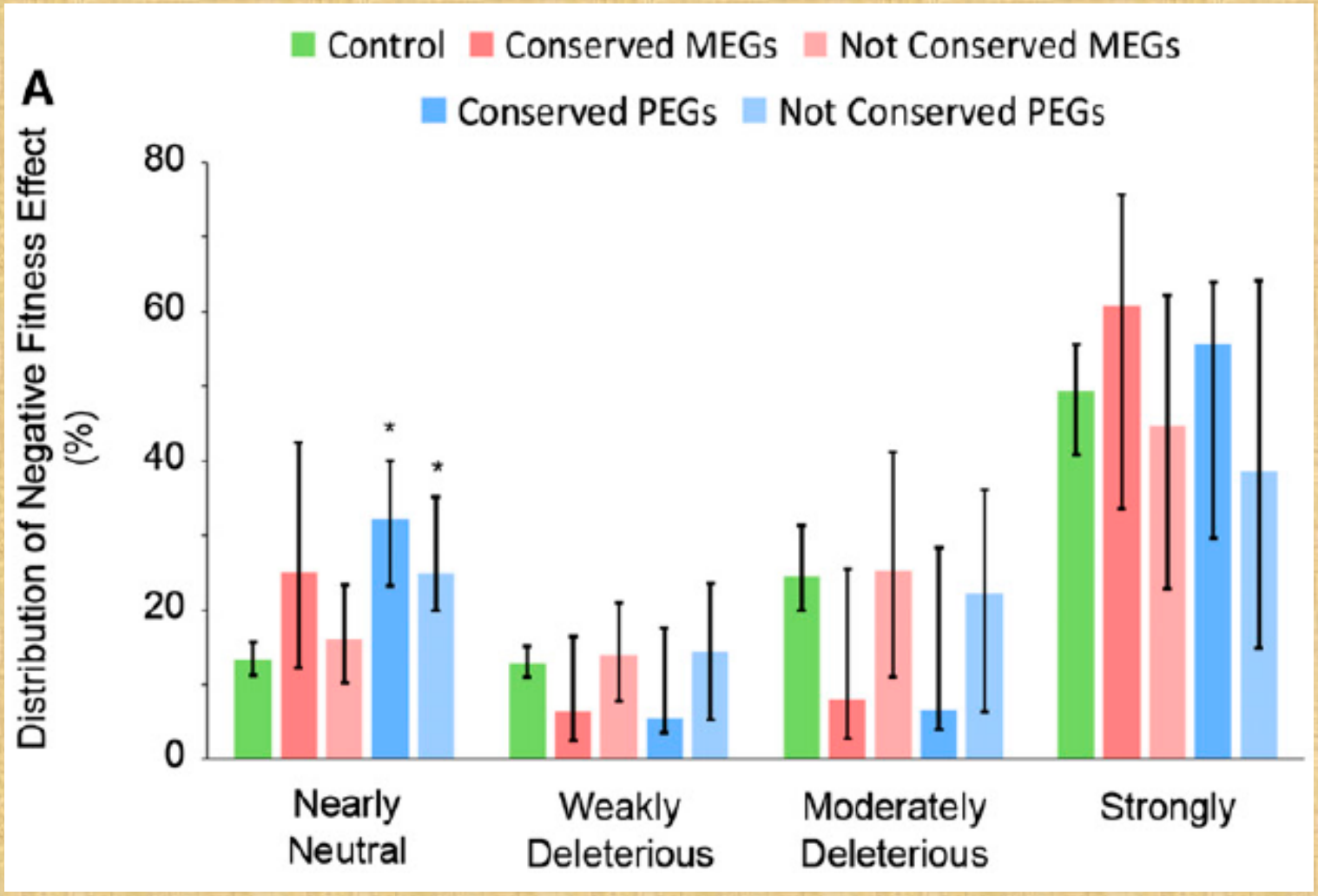
The **McDonald-Kreitman test** adds a comparison between interlineage divergence and intralineage polymorphism

Makes a ratio of P_n / P_s by D_n / D_s

$$\alpha = 1 - \frac{D_S P_n}{D_n P_s}$$

If α = 0, divergence = polymorphism = neutral
If α = 1, divergence >>> polymorphism = positive selection

The distribution of fitness effect for candidate genes



The strength of **purifying selection**: N_eS

Eyre-Walker and Keightley, 2007

F_{st} scans

Intralineage variation ⟷ Interlineage divergence

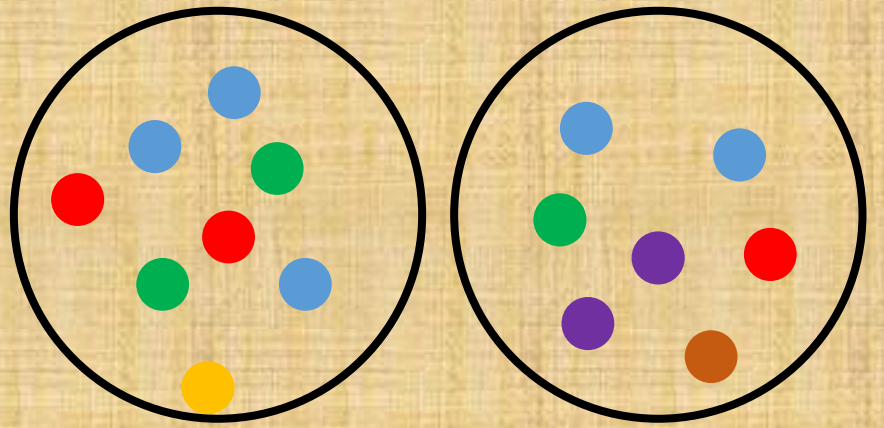
Intralineage variation > or = Interlineage divergence → LOW F_{st}

Intralineage variation < or = Interlineage divergence → HIGH F_{st}

LOW F_{st}

Lineage A

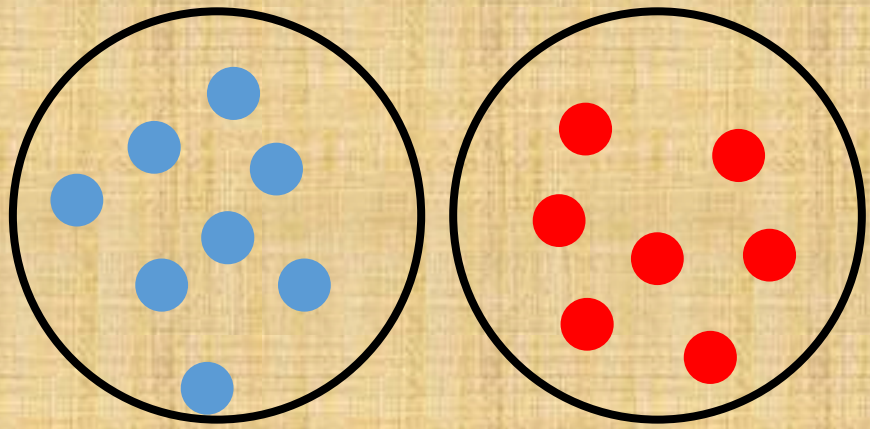
Lineage B



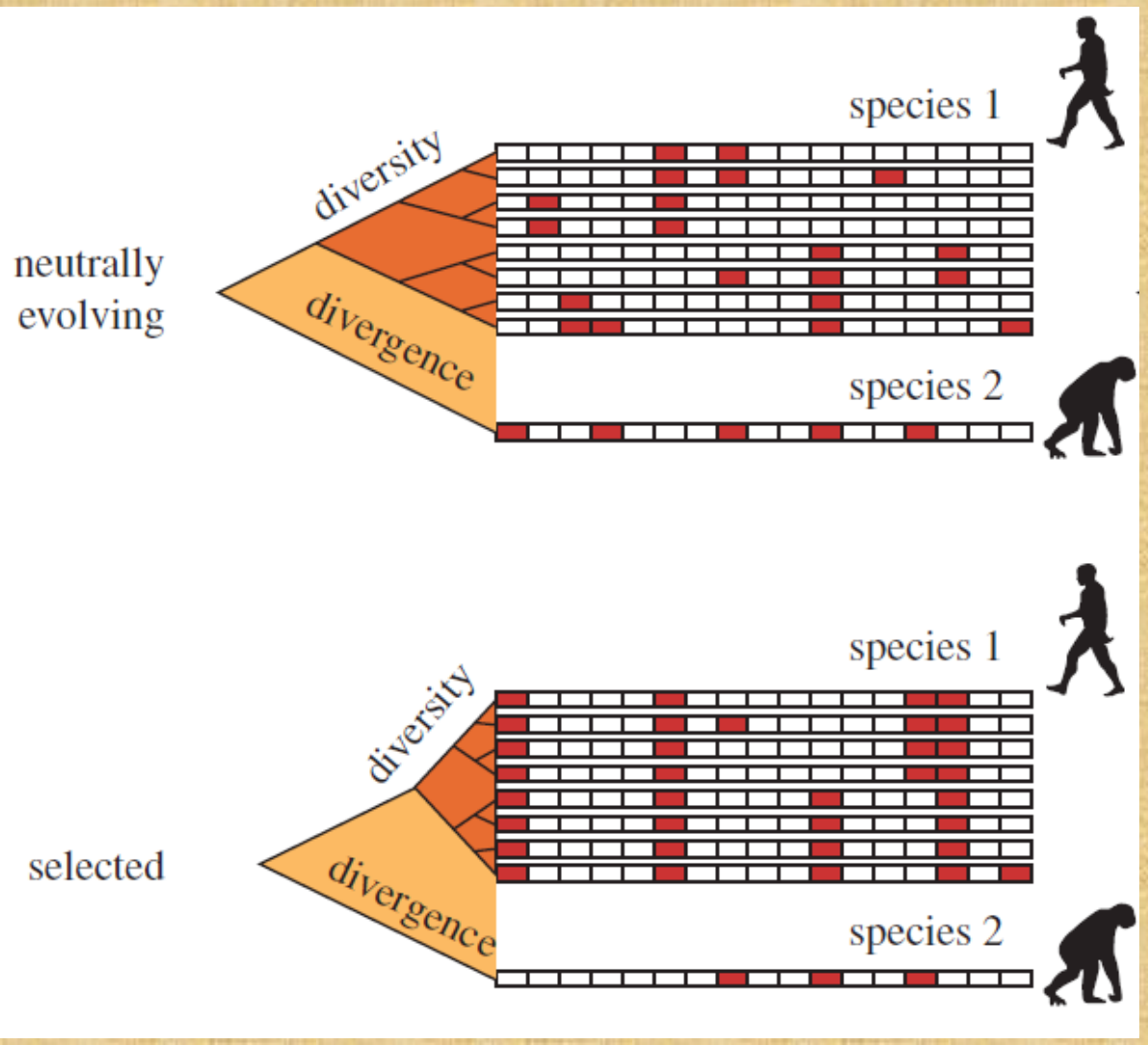
HIGH F_{st}

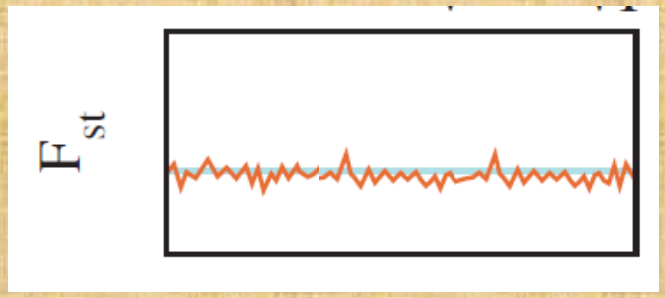
Lineage A

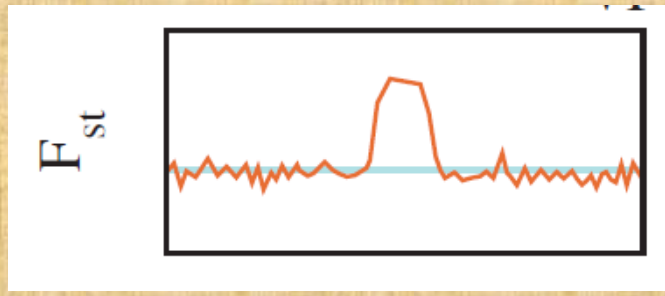
Lineage B



F_{st} scans







Oleksyk et al., 2009

Plan

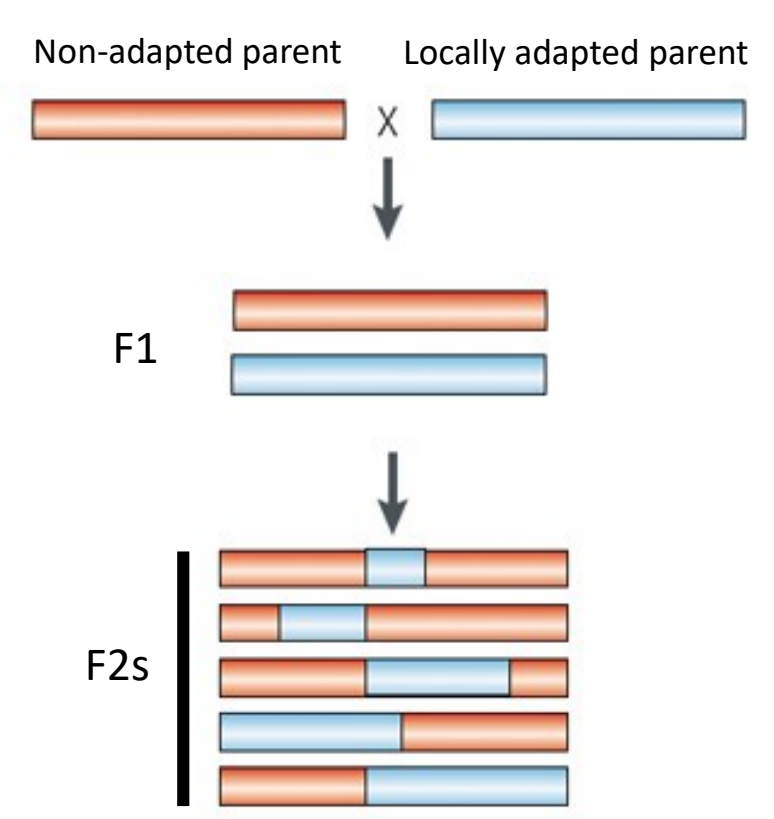
Demography and other factors influencing genetic variation
- Bottlenecks, migration, mating system etc
- N_e and efficacy of selection
- LD/Recombination rate

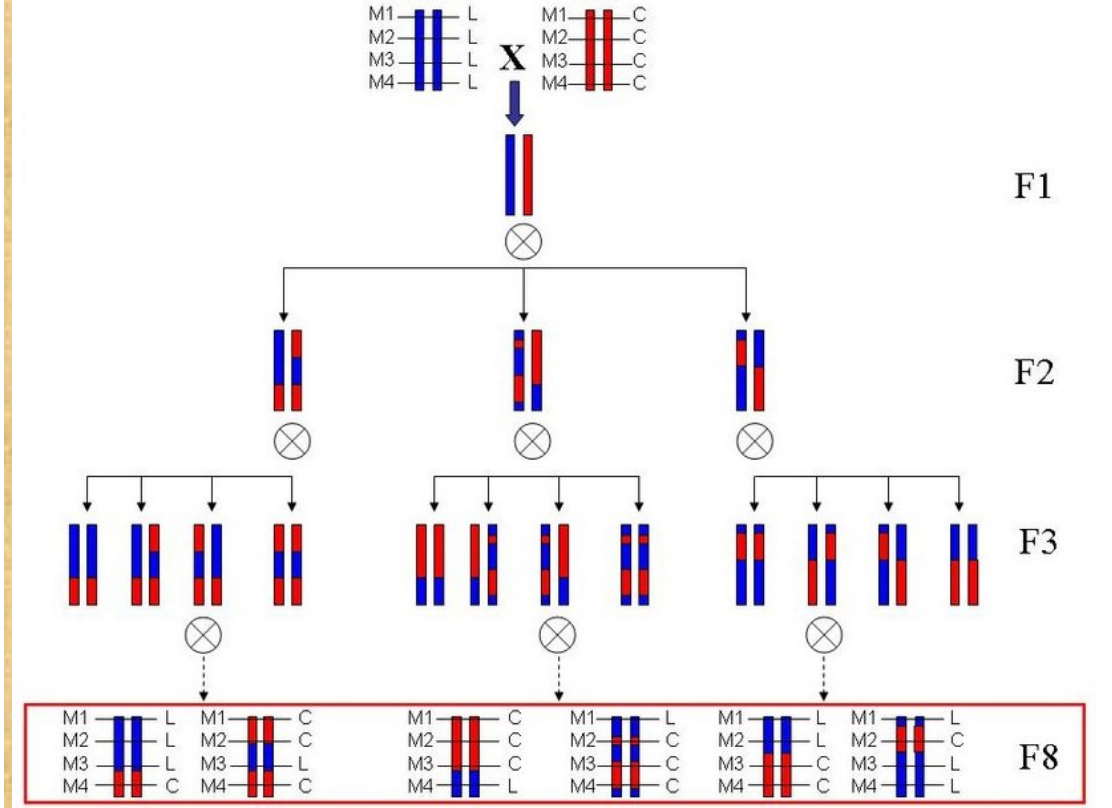
Testing for demographic cofounding factors
- Nucleotide diversity, heterozygosity
- Testing for demographic events: Tajima's D
- Estimating the "genetic load" and efficacy of selection: SFS and DFE

Detecting natural selection vs neutrality (and demography) in the genome
- types of selection
- selective sweeps
- dN/dS and PN/PS
- DFE and SFS
- Fst scans

Adaptomics methods involving additional aspects than just allele frequencies
- Genetic vs phenotypic association: QTLs and GWAS
- Environmental associations

Genetic vs phenotypic association: QTLs



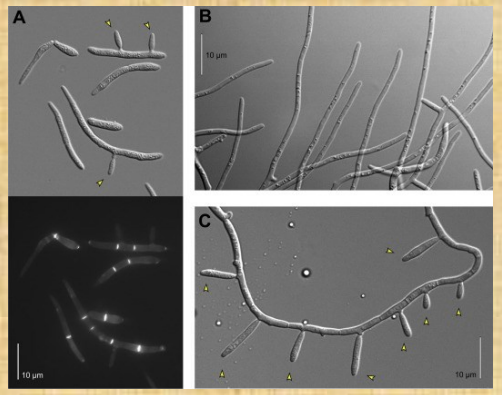


+ : fast to obtain
- : heterozygosity
- : few recombination events (one generation) → large QTLs

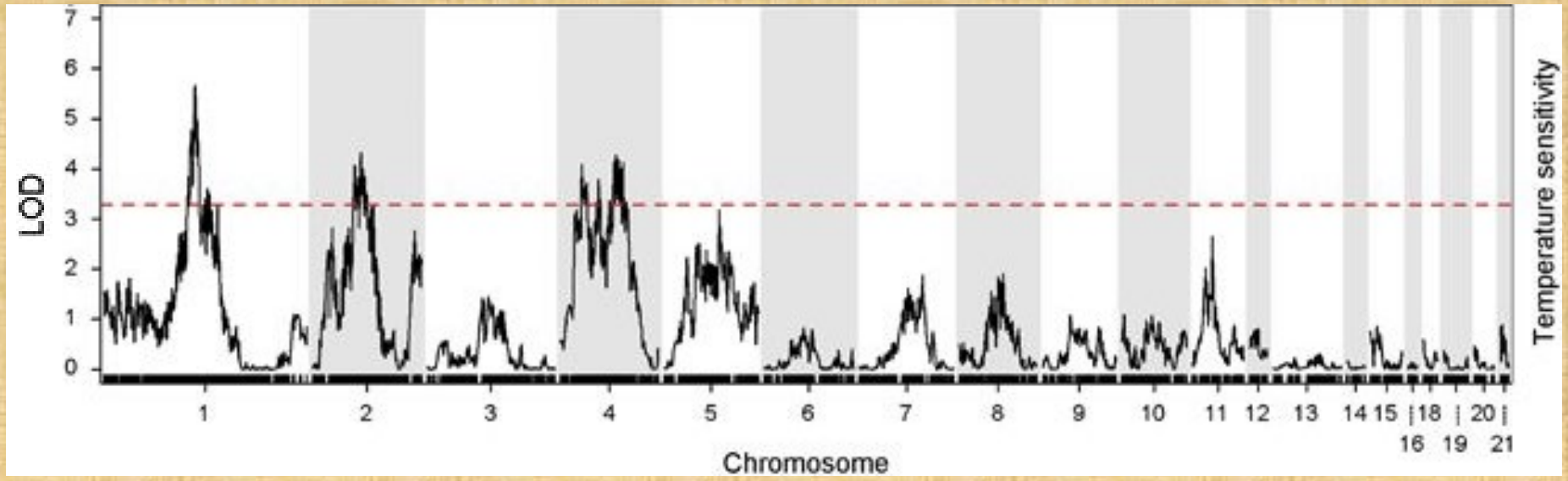
+ : more recombination events (>5 generations) = more combinations possible = narrow QTLs
+ : more homozygous = clearer phenotypes (dominance)
- : more time and effort consuming

Phenotyping of the F2s/RILs + Sequencing

Genetic vs phenotypic association: QTLs



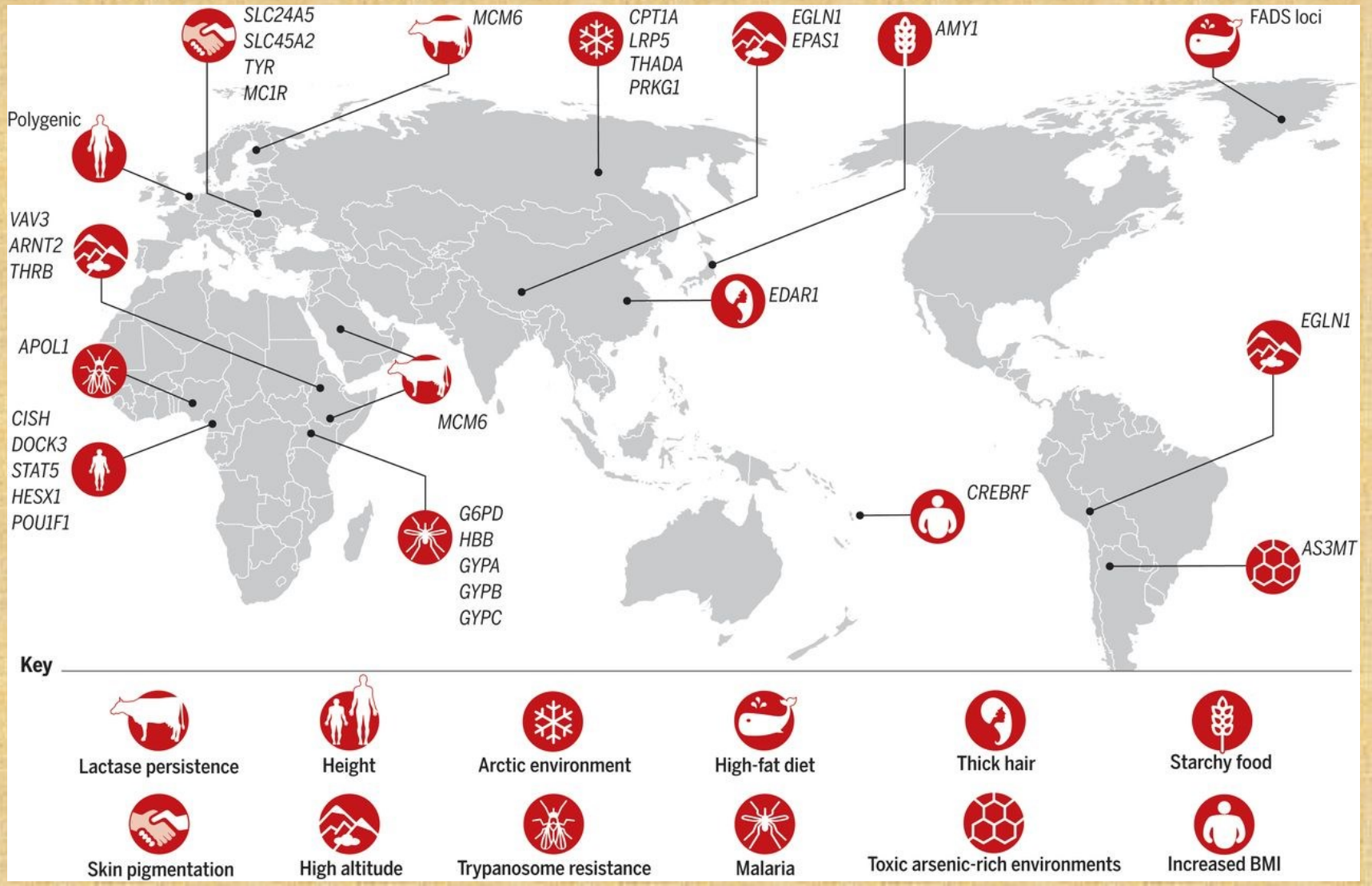
Zymoseptoria tritici



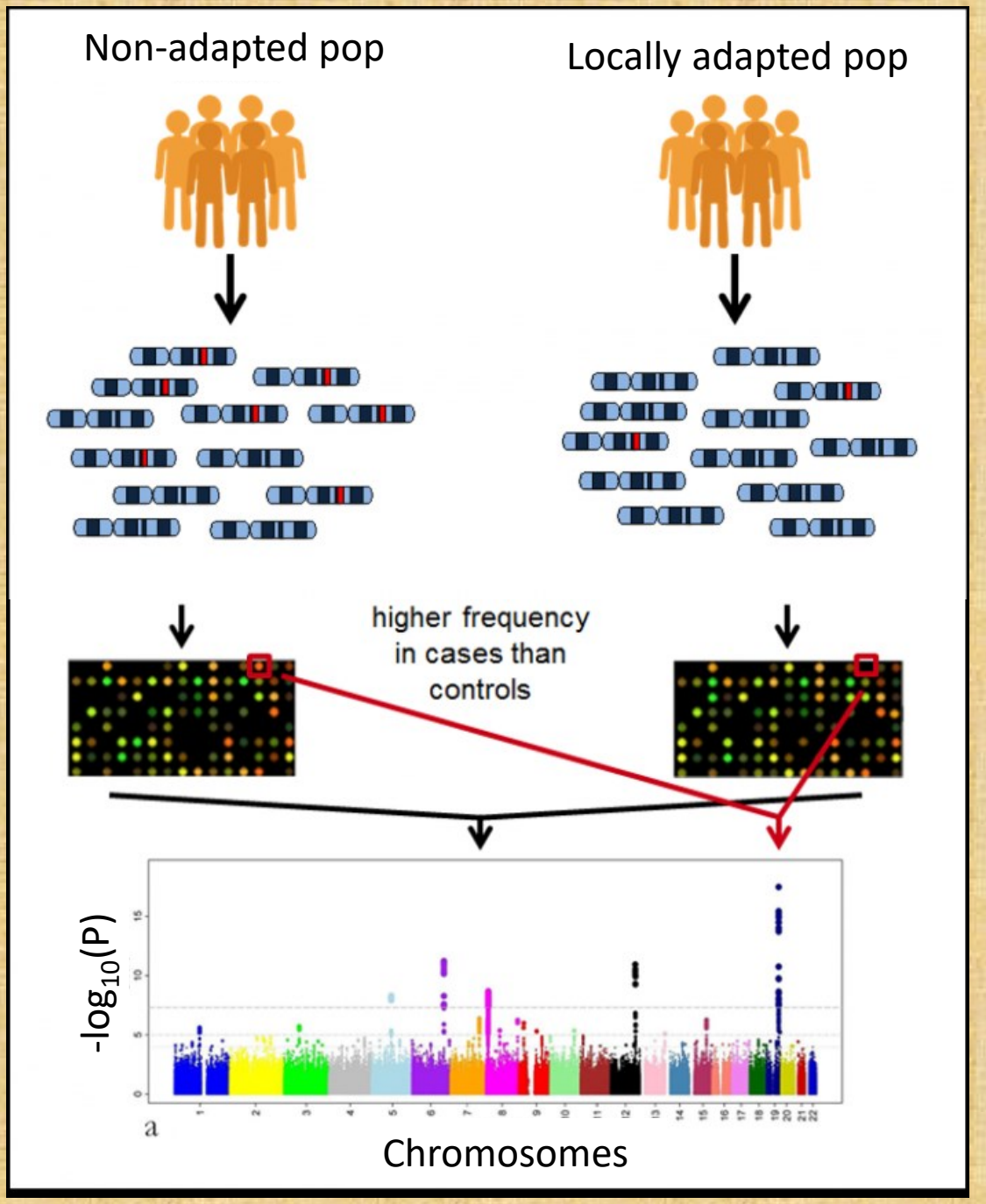
LOD score: the probably of association between phenotype and genotype

Lendenmann et al., 2016

Genetic vs phenotypic association: GWAS



Fan et al., 2016



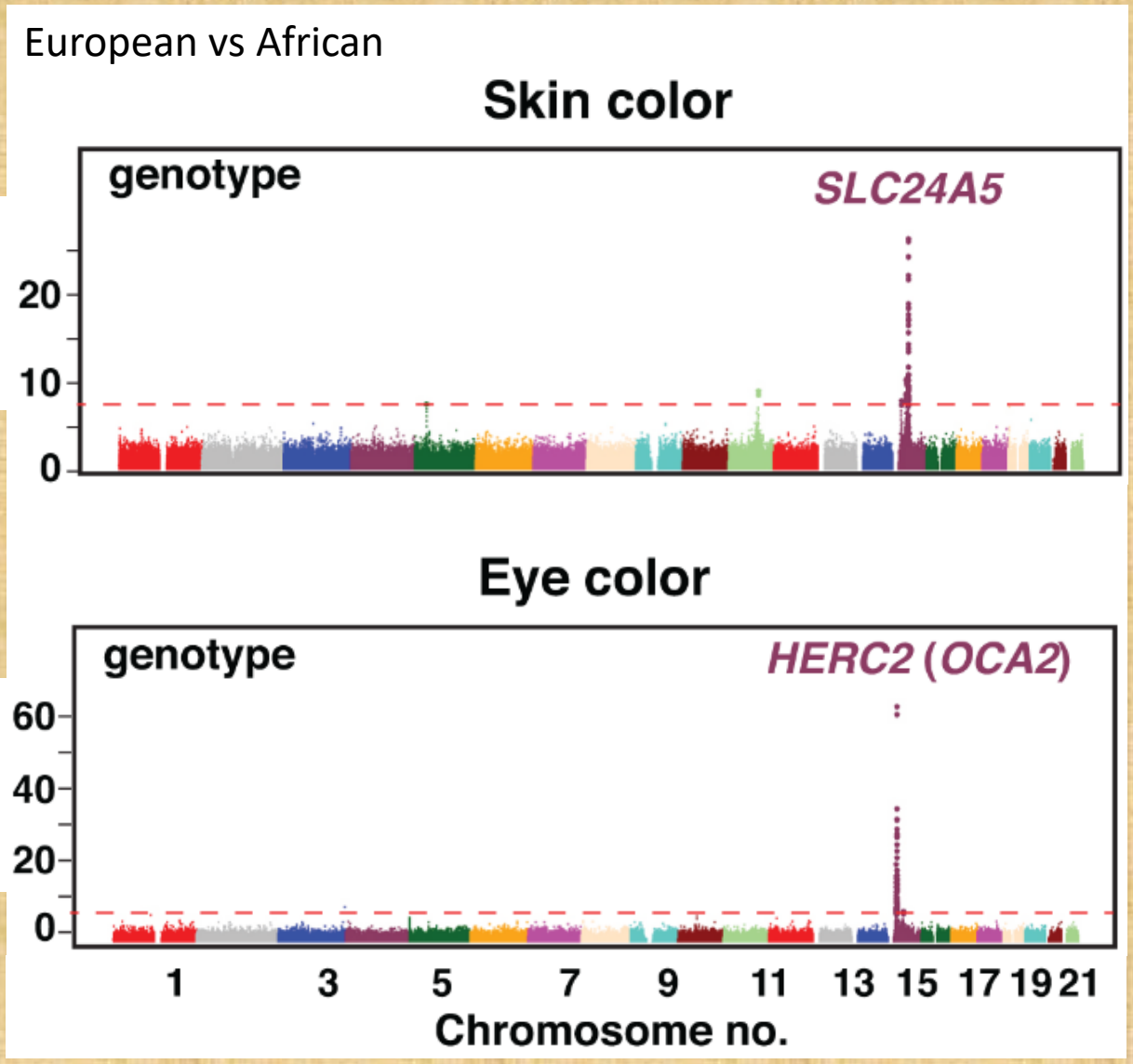
Genotype-Phenotype association

Better resolution than QTLs: one can infer the effect of a single SNP on the phenotype

Important factors to take into account: demography, especially population structure

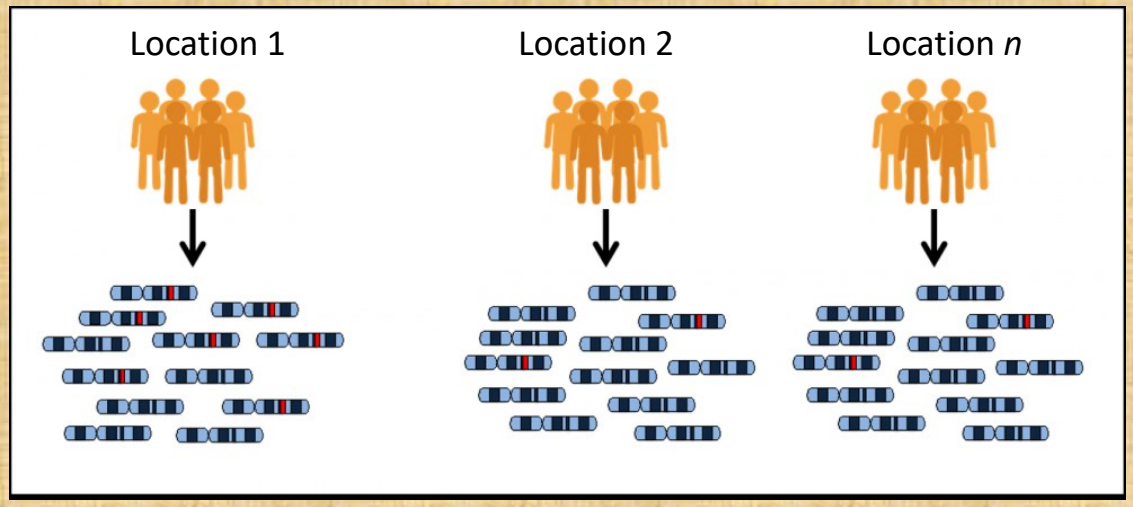
https://www.ebi.ac.uk/

Genetic vs phenotypic association: GWAS



Beleza et al., 2013

Environmental association: principle



Climatic data for Locations 1, 2… *n* (*WordClim database*…)

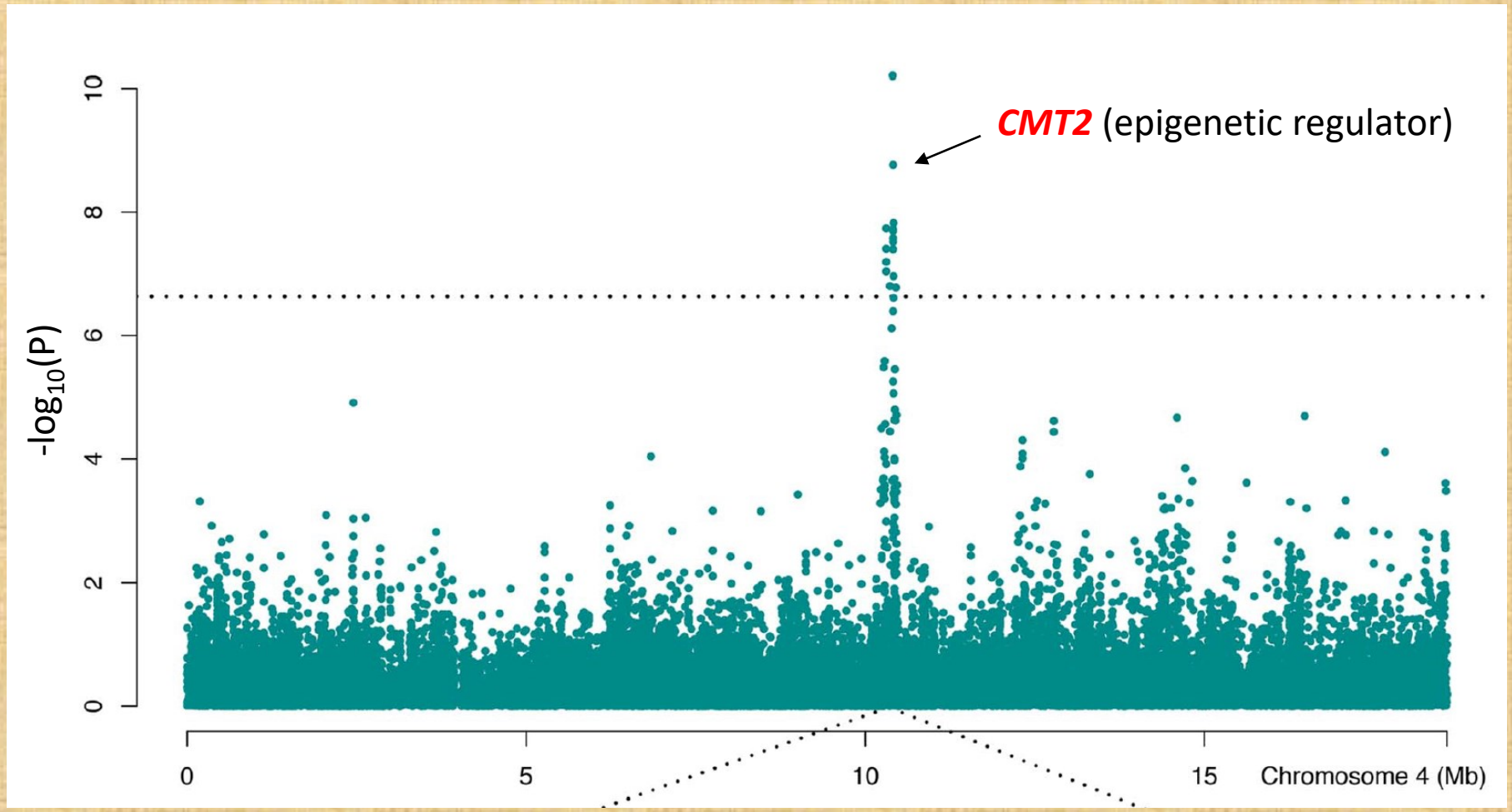
SNP-climate variable assocation

https://www.ebi.ac.uk/

Environmental association: example

A. thaliana

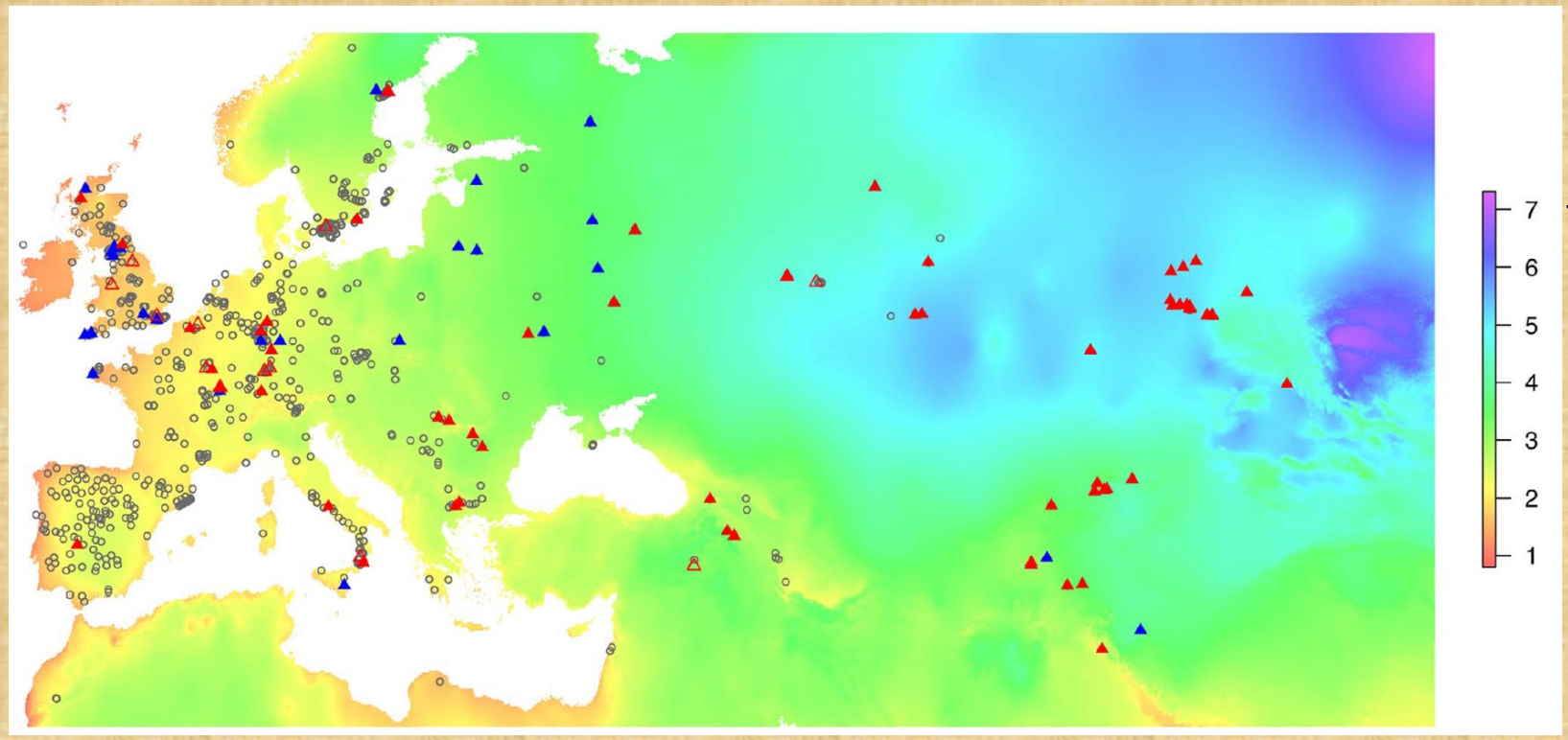
Association with temperature seasonality (difference between summer and winter)



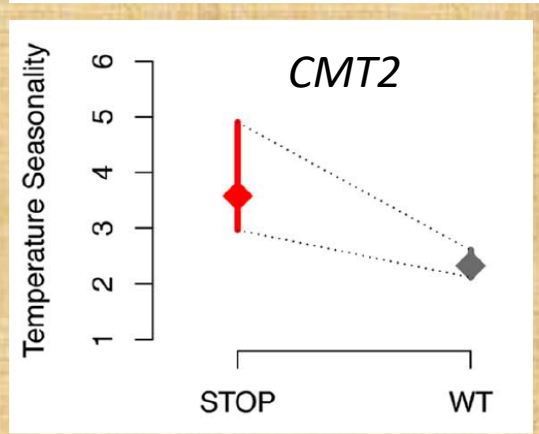
Shen et al., 2014

Environmental association: example

○ Functional allele of *CMT2* ▲/▲ Stop-codon allele



Temperature seasonality



Shen et al., 2014

Conclusions

Neutral molecular evolution = null hypothesis

Demography can have large effects on molecular evolution

It is necessary to test demographic effects before searching for selection:
when searching for loci under selection, always determine the genomic pattern

Different methods to search for selection, need to be adapted to the model study