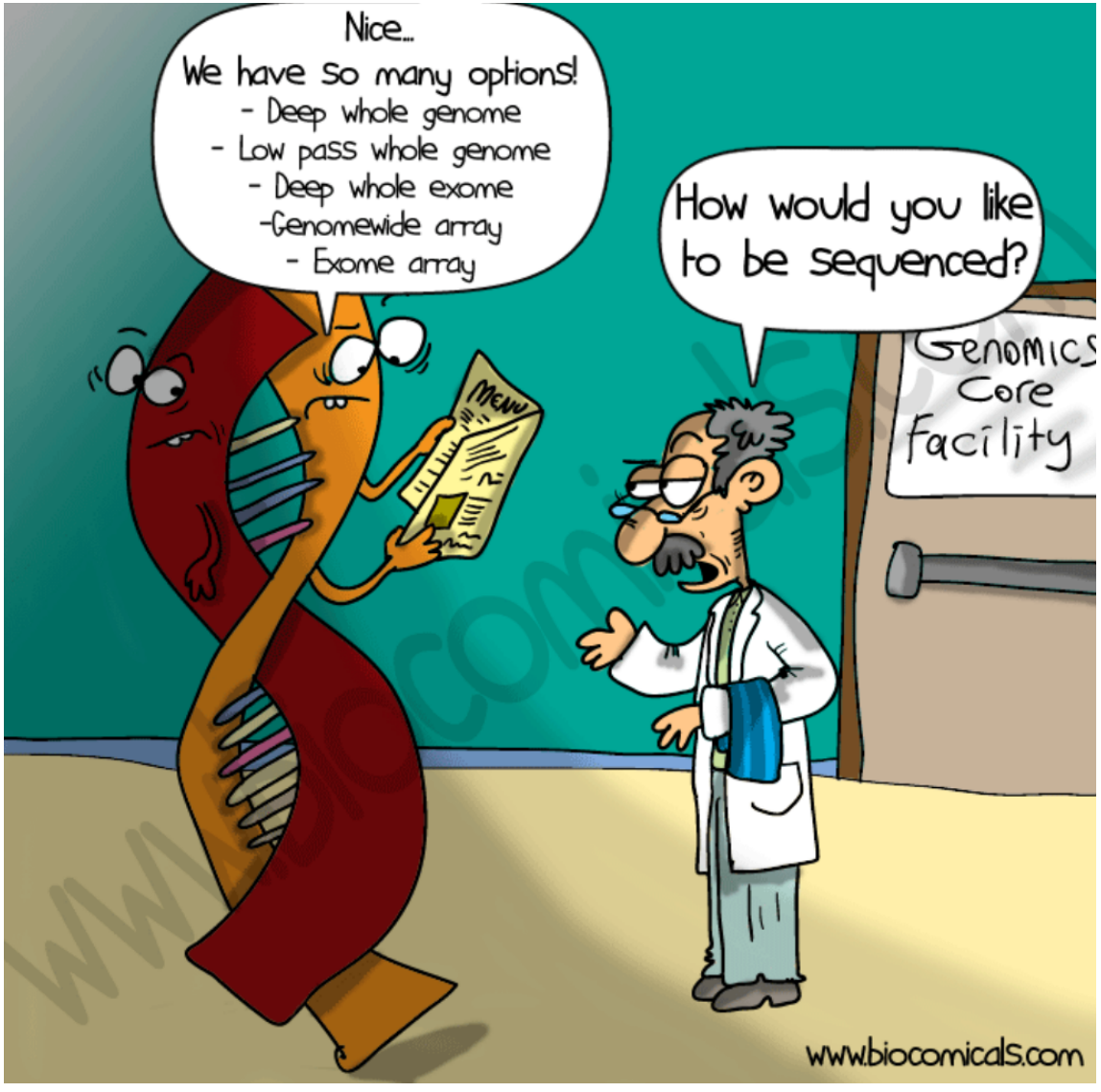
Inferring sites with recent or ongoing selection for SNP chip or NGS data

`github.com/aalbrechtsen/embo2022`

Anders Albrechtsen

Genetic data



What kind of data are you working on?

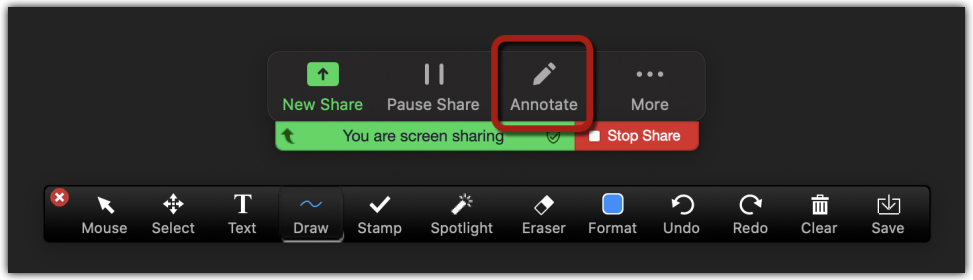


Figure: Use the stamp tool

whole genome shotgun:

SNP chips:

Capture/GBS/RADseq

High depth:

low depth:

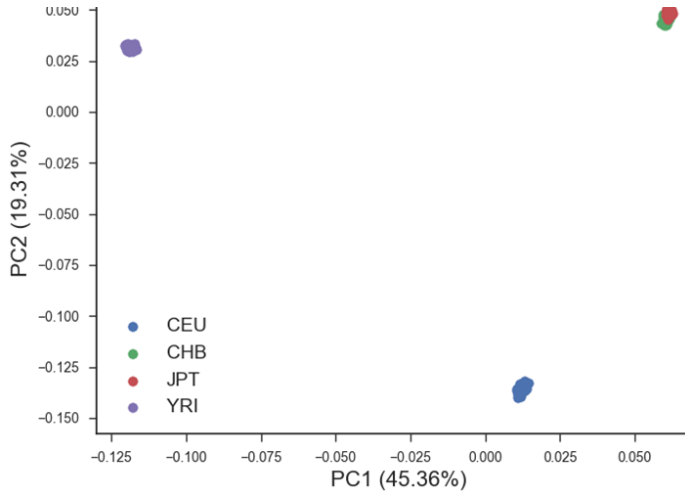
Other (use text)

Data, structure and possibilities

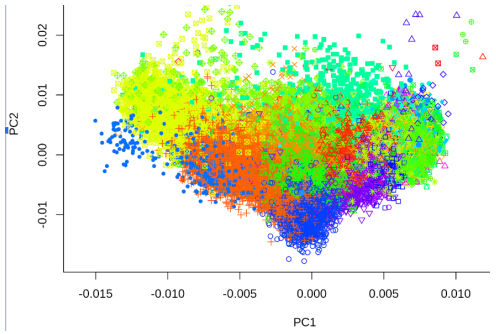
What kind of Structure?

Single population

Discrete populations



Structured populations



What kind of data?

SNP chip/SNP capture

- Autosomal
- sex chromosomes (with recomb)

Sequencing data

- High depth
- low depth
- Whole genome
- capture/GBS

What is low depth sequencing - my take on it

medium/high depth vs. ultra low depth

Medium depth sequencing

Ultra low depth sequencing

medium/low

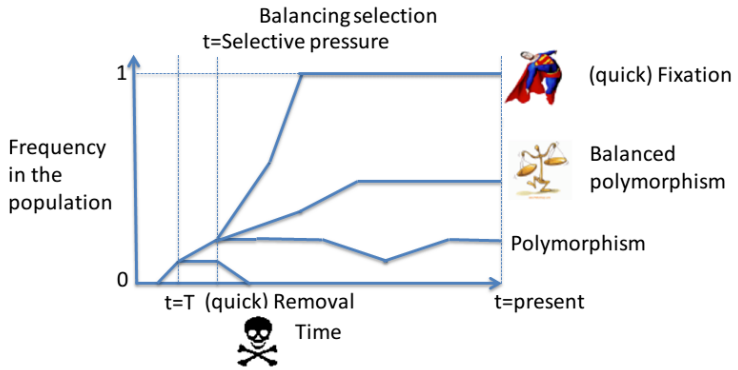
- Depth lower than 10X
- Often a financial choice
- Ancient DNA

Ultra low sequencing

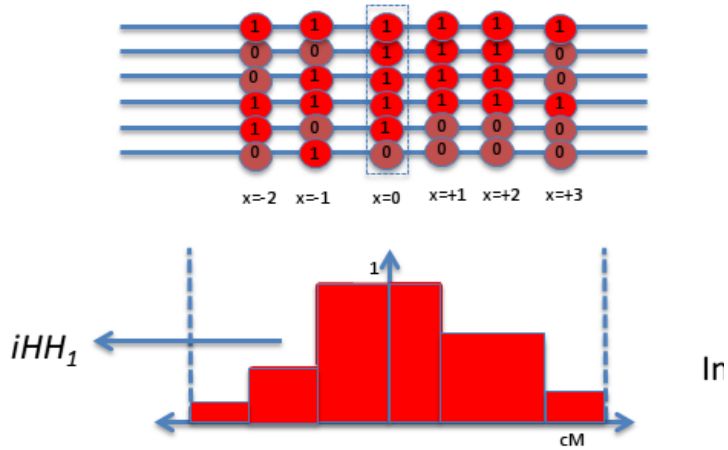
- Depth lower than 1X
- by product of capture data
- ancient DNA

This afternoon

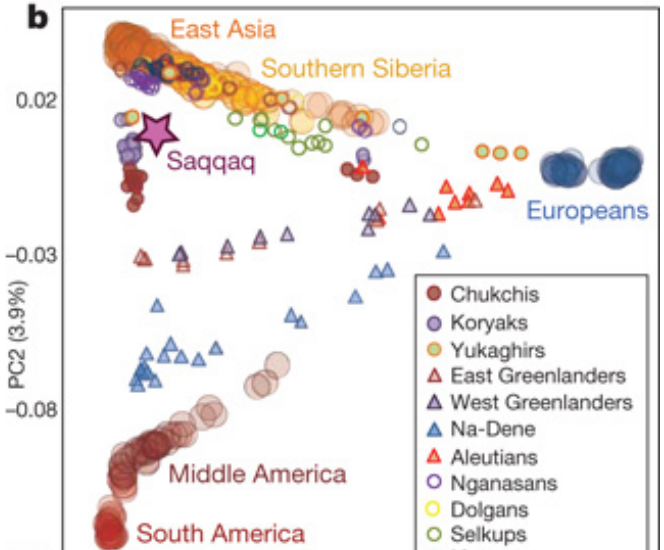
Short intro to recent selection



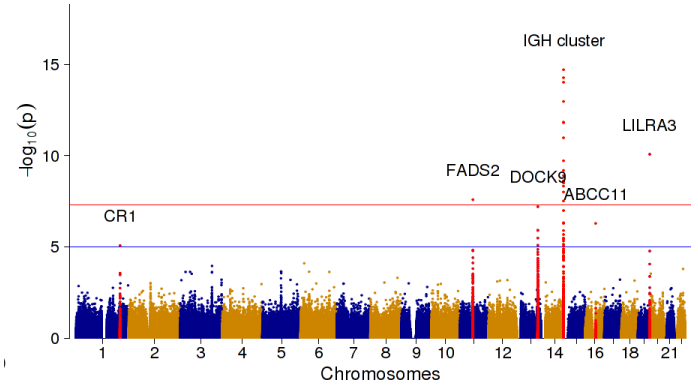
EHH



PCA

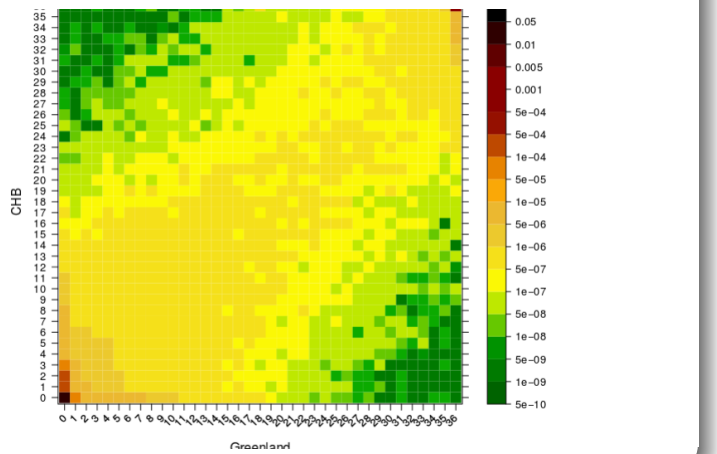


Individual allele frequencies (PCA)

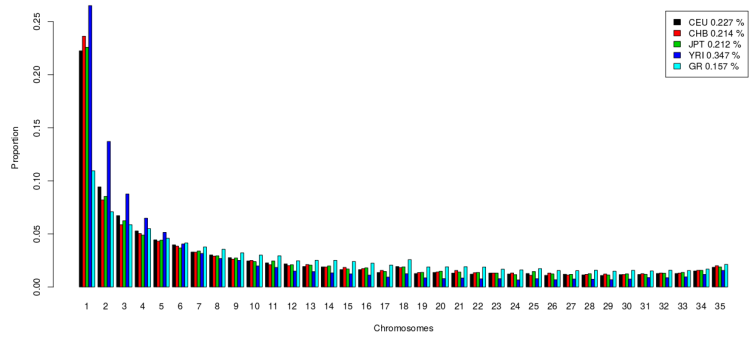


Tomorrow - focus on low depth sequencing and structure

2-3D SFS, Fst and PBS for NGS

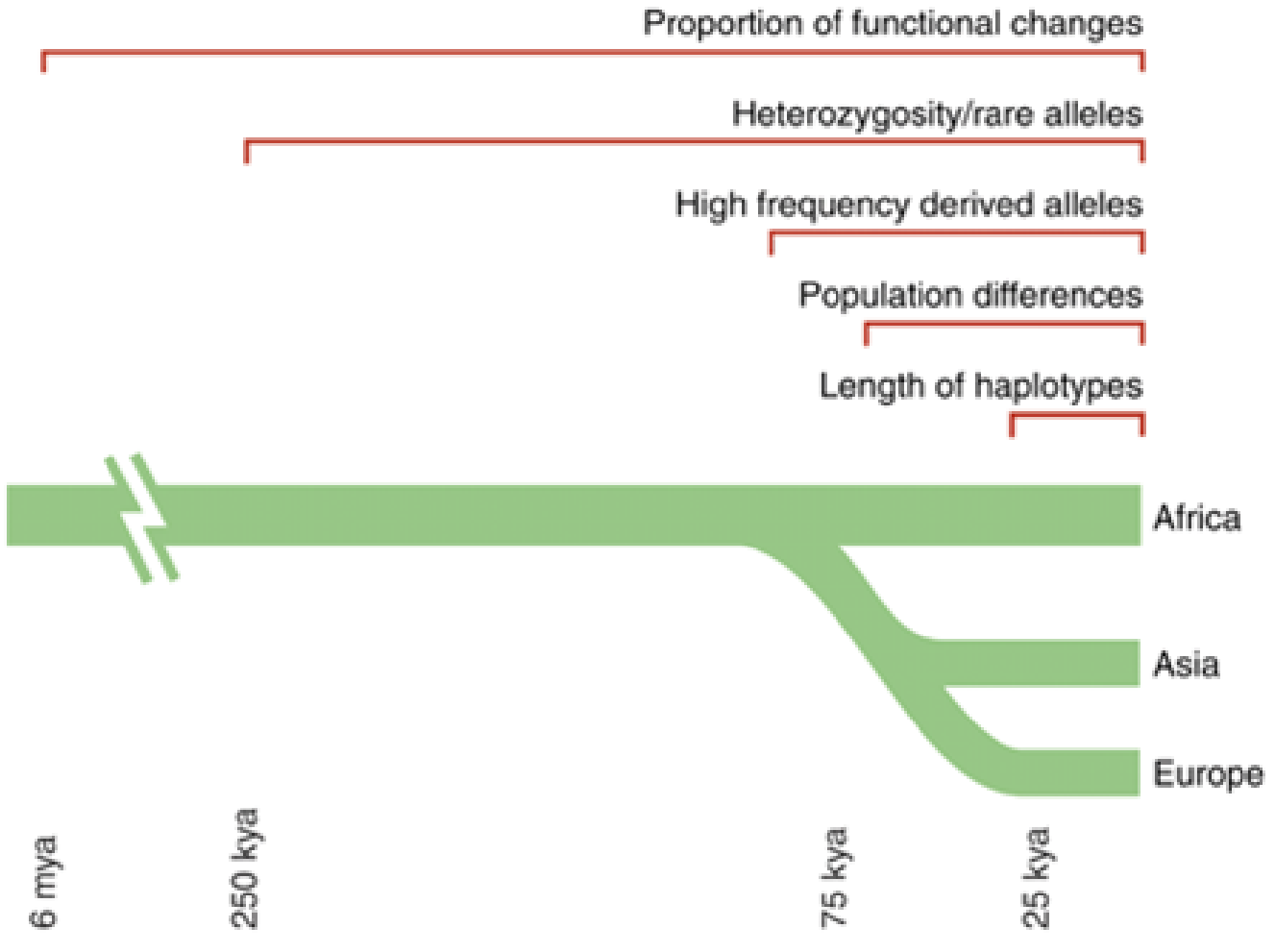


SFS for NGS



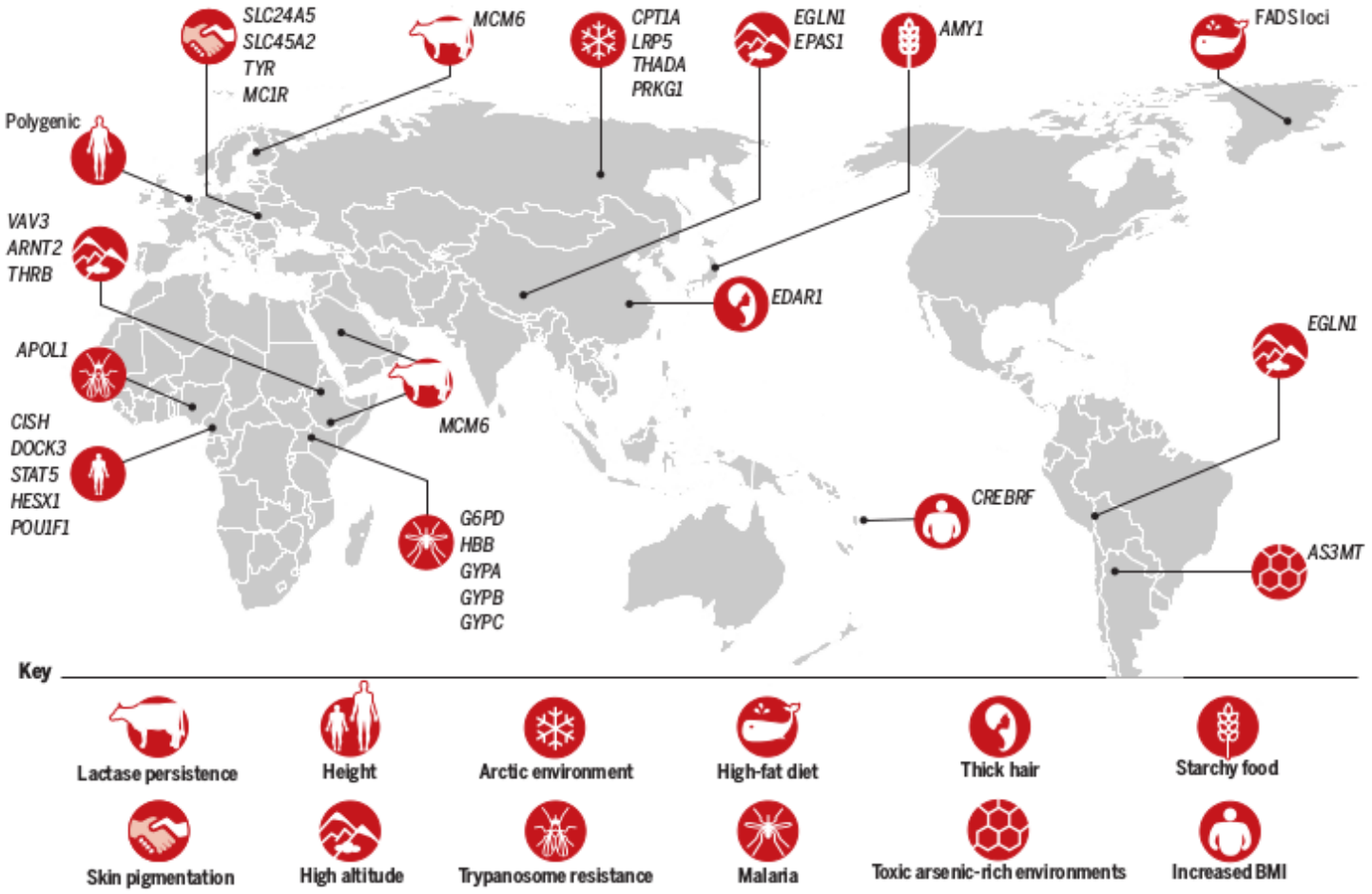
Recent selection

within species / using shared variation



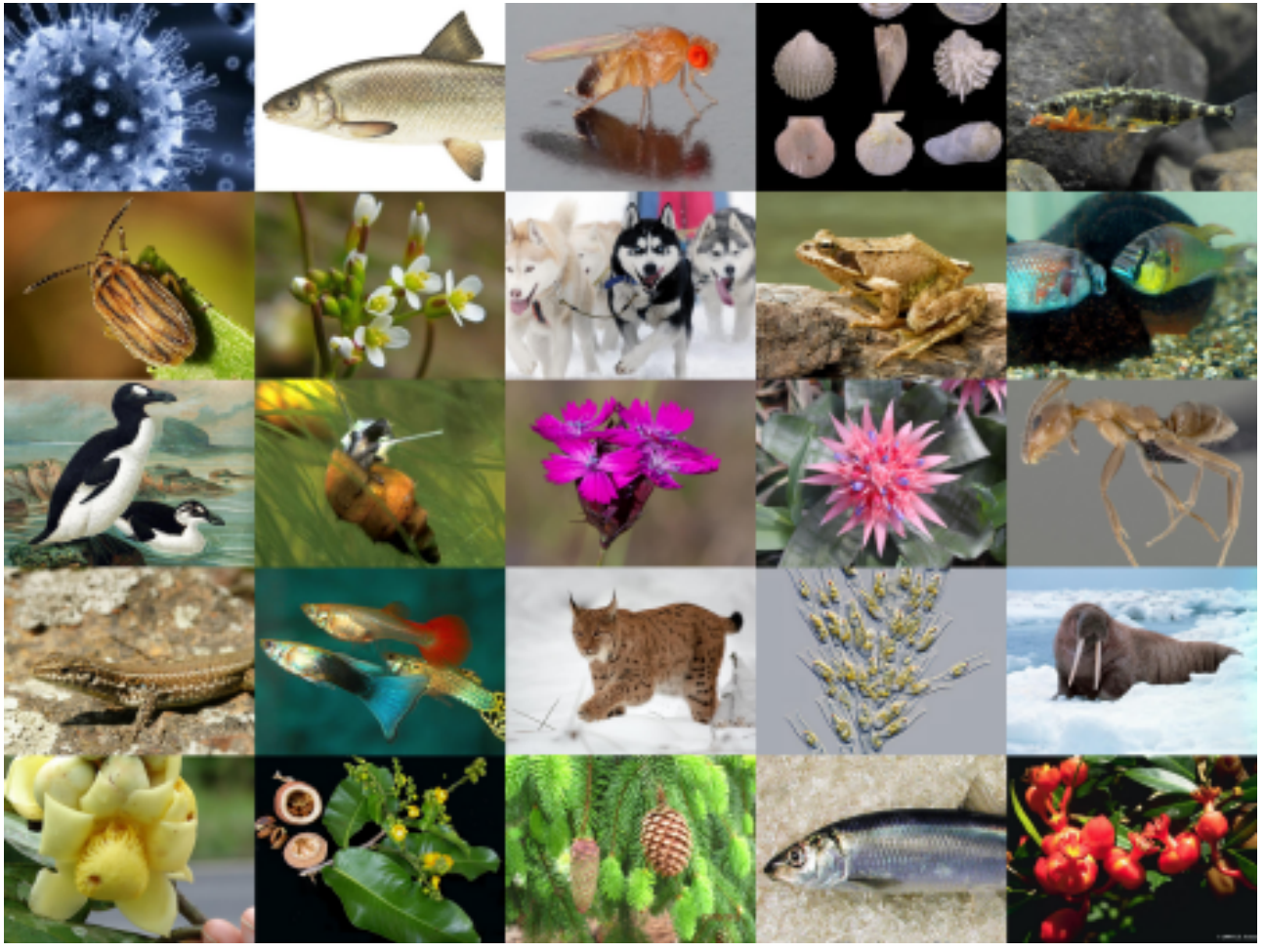
Sorry about the Human-centric talk

Good candidates for genes under recent selection



Methods is applicable for most organisms

Examples of organisms with DNA



Neutral selection

Alleles can be removed,polymorphic or fixed

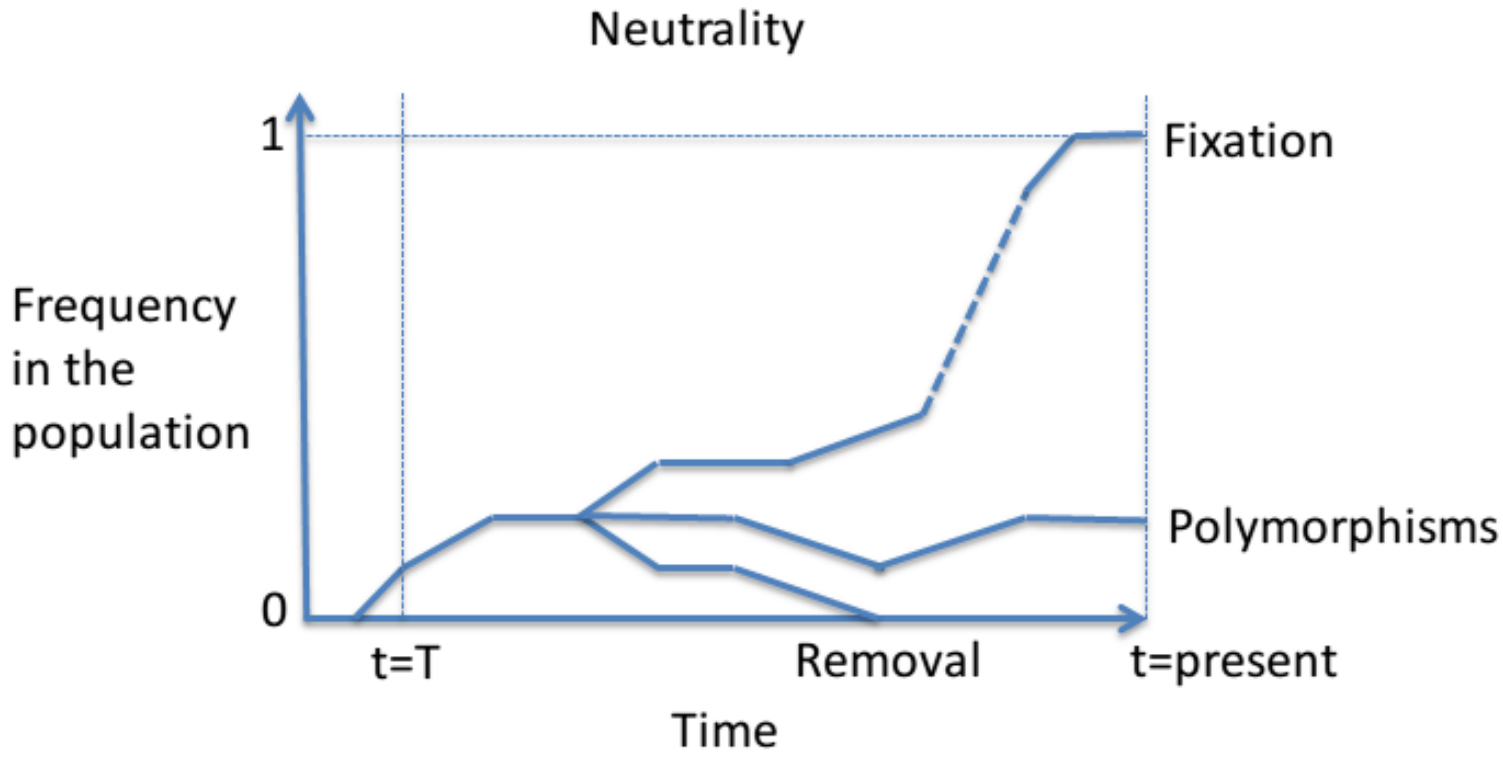
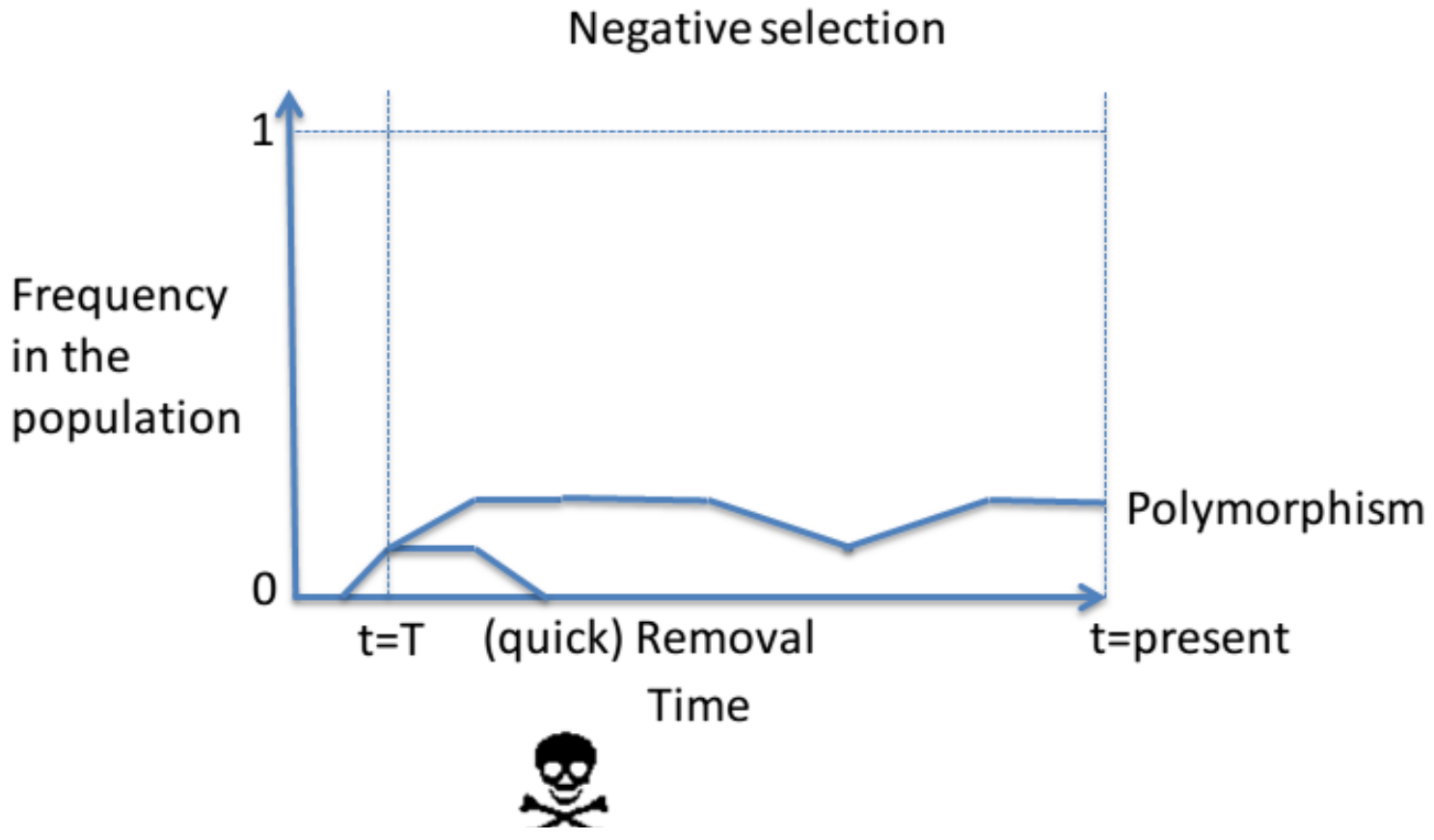


figure from Matteo Fumagalli

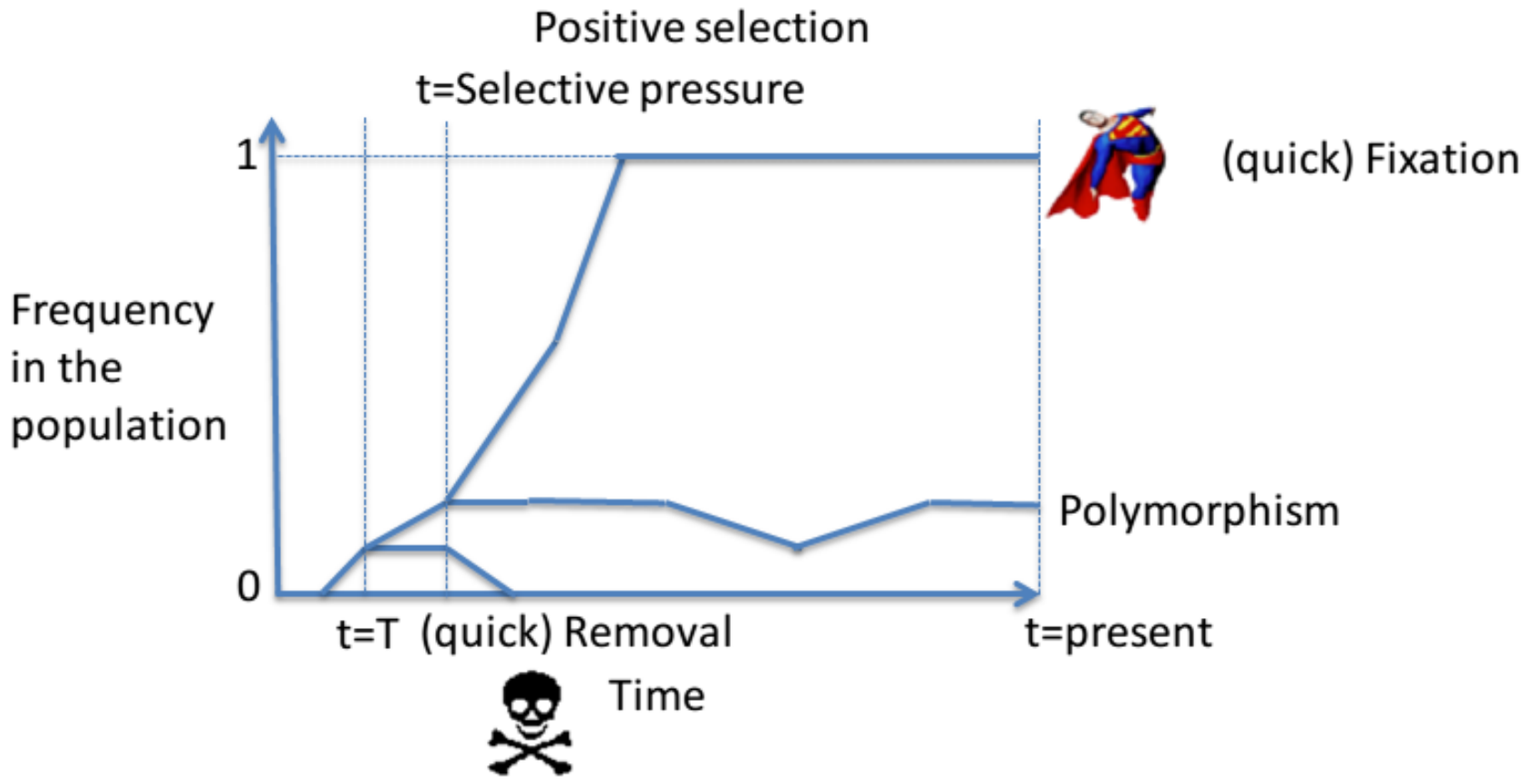
strong negative selection

alleles can be removed or be polymorphic



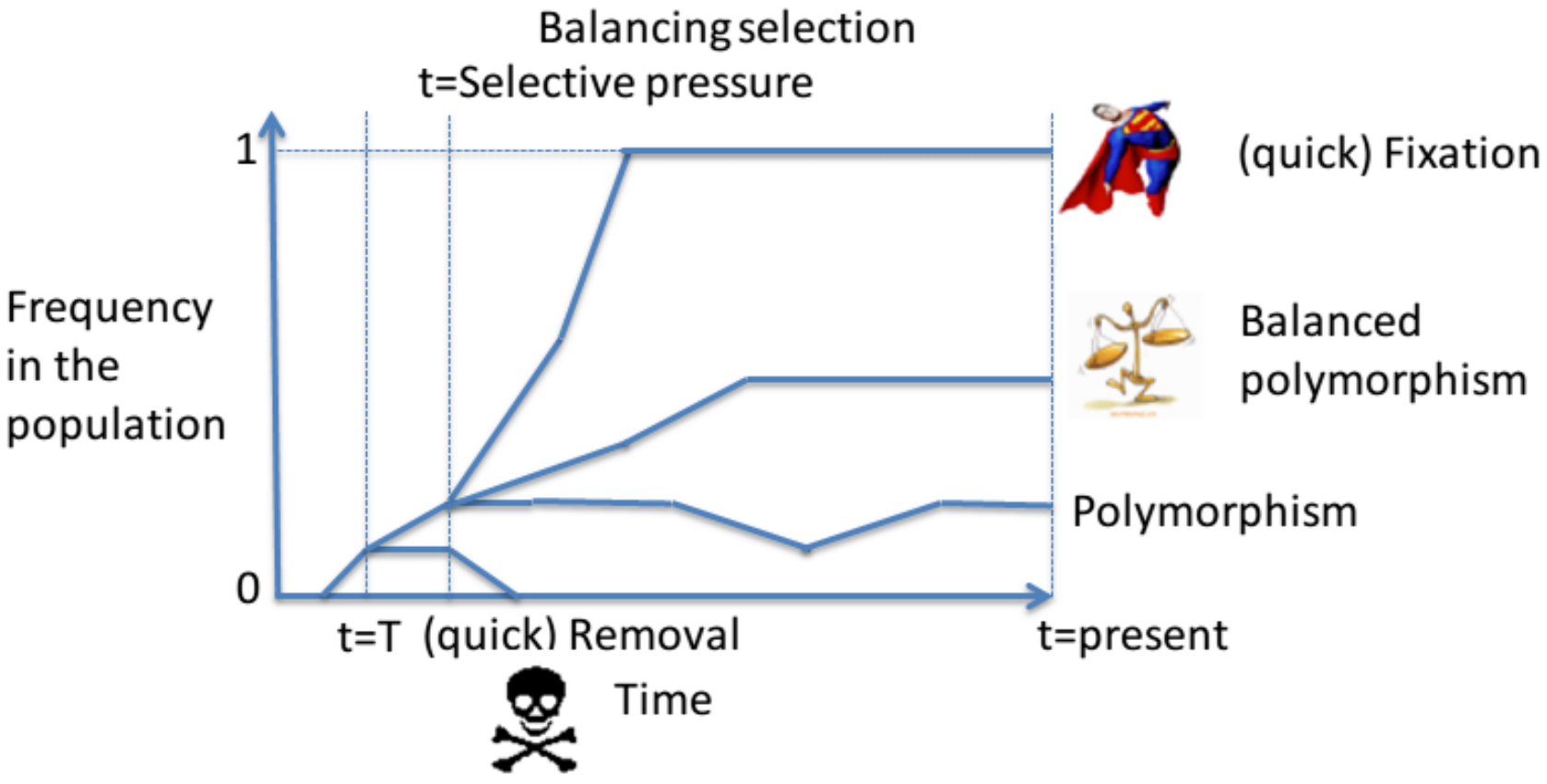
Strong positive selection

Alleles can be removed, polymorphic or fixed



Balancing selection

Alleles can be removed, polymorphic or fixed



strong positive selection

7 of 50 simulations reach fixation

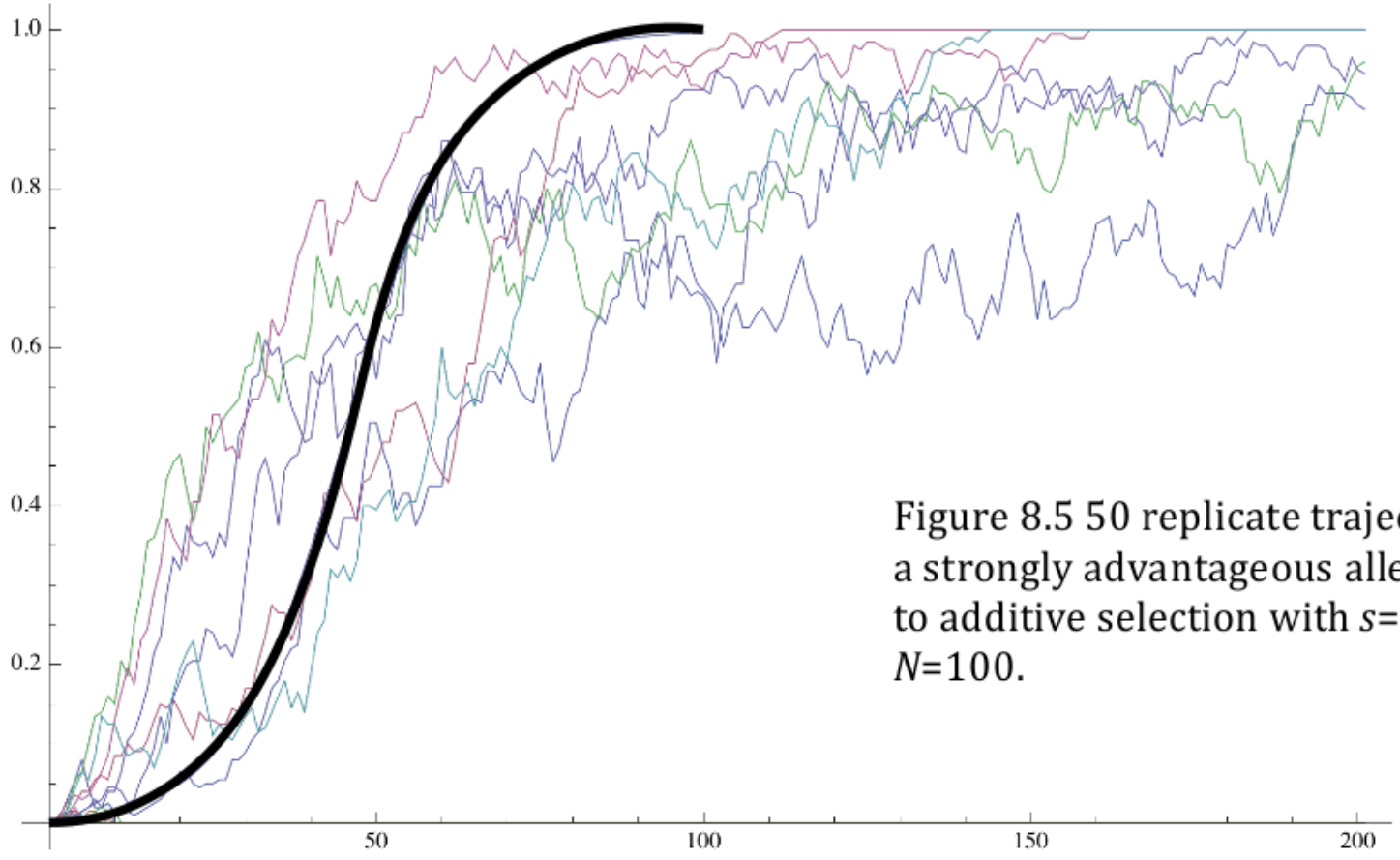
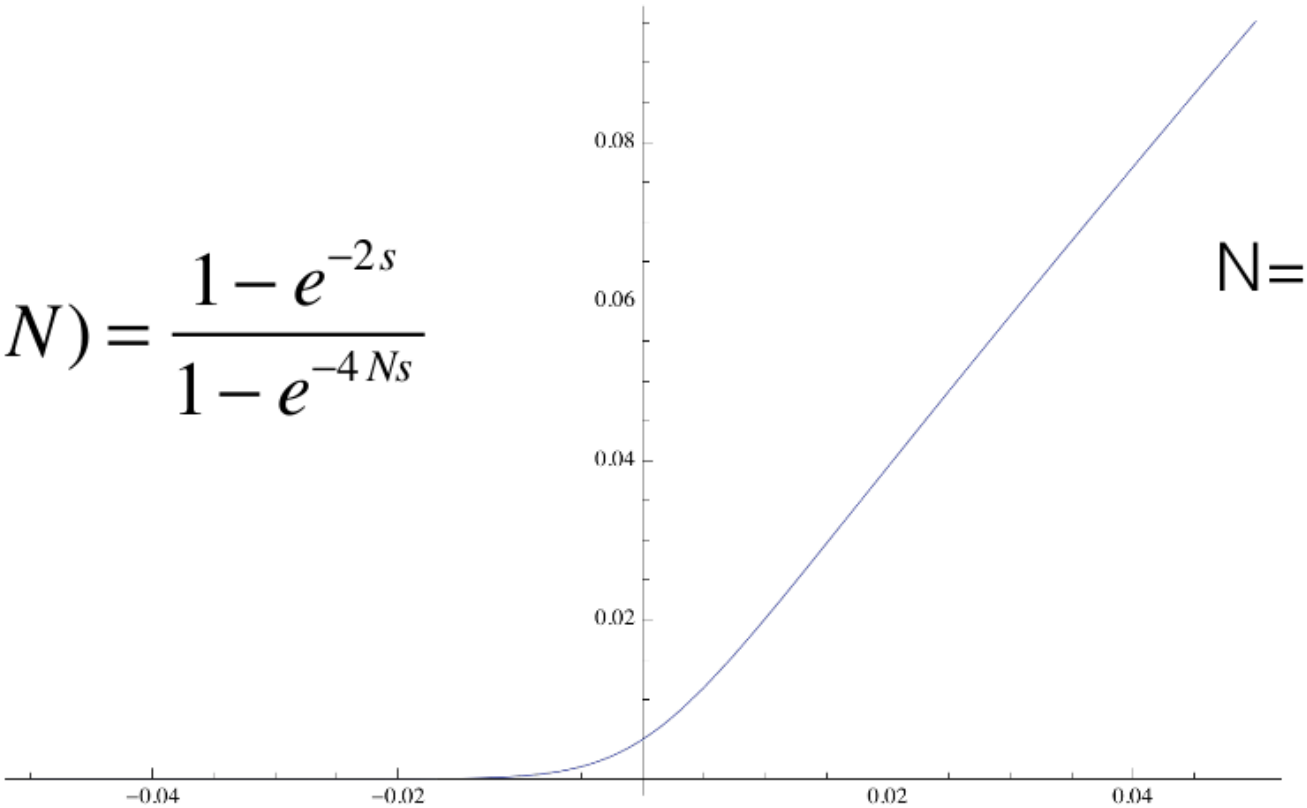


Figure 8.5 50 replicate trajectories for a strongly advantageous allele subject to additive selection with s=0.1 and N=100.

Probability of fixation

$$u(s,N) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}$$



N=100

Strongly deleterious
2Ns < 1

Nearly neutral

Strongly advantageous
2Ns > 1

Summary of allele frequencies and selection

allele removed	allele fixed	allele polymophic
strong positive:	strong positive:	strong positive:
strong negative:	strong negative:	strong negative:
balancing:	balancing:	balancing:
neutral:	neutral:	neutral:

Summary of allele frequency changes

selections effect on alleles

Neutral/weak removed, polymorphic or fixed

Strong negative removed or polymorphic

Strong positive removed, polymorphic or fixed

Balacing removed, polymorphic or fixed

Strong selection

Depends on the population size

Conclusion

Allele frequency is (almost always) not enough to determine selection

Need for additional information

Option 1

use information from the genomic region

Option 2

Use information from mulitple species/populations

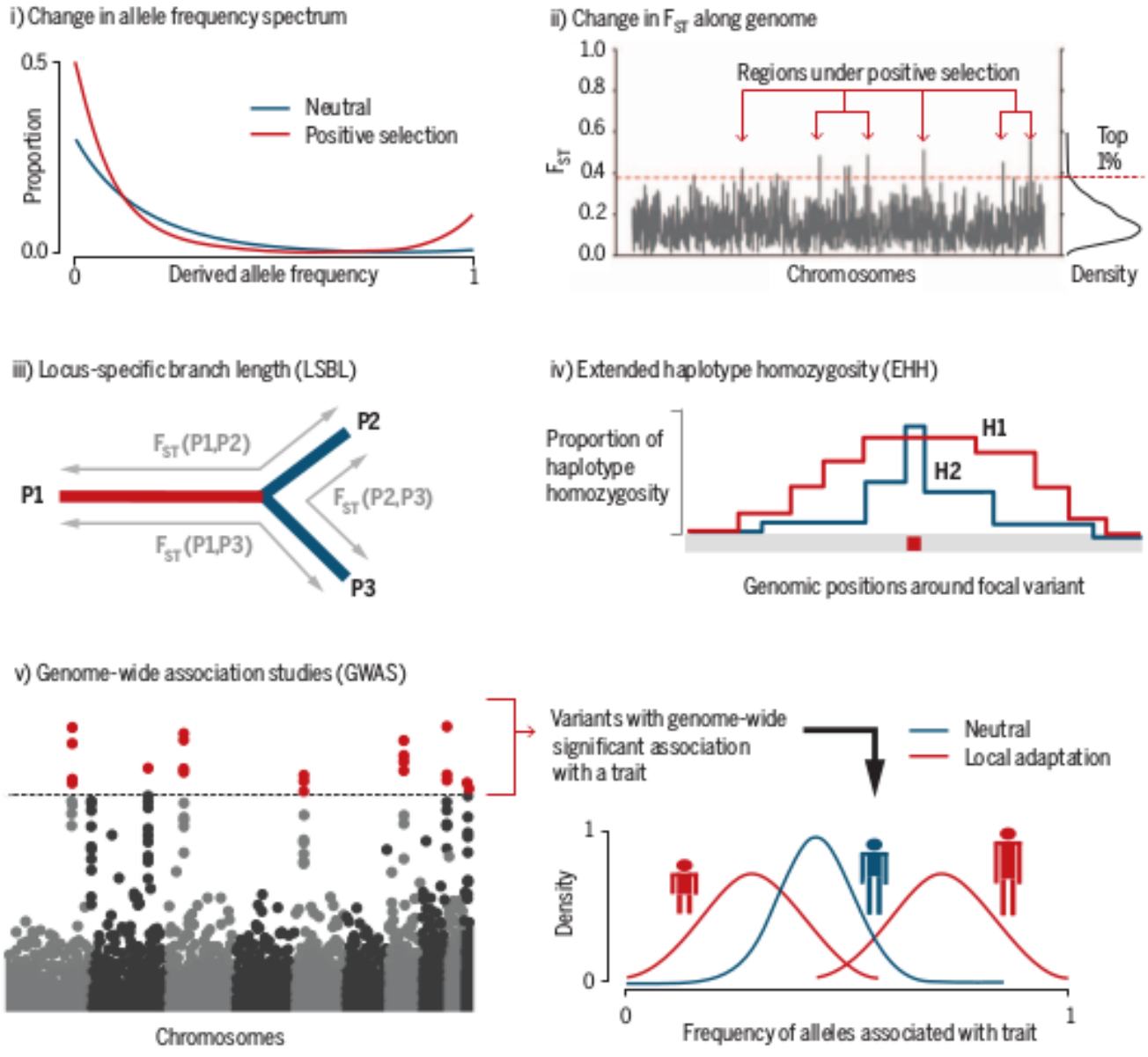
Options 3

selection experiments

External information

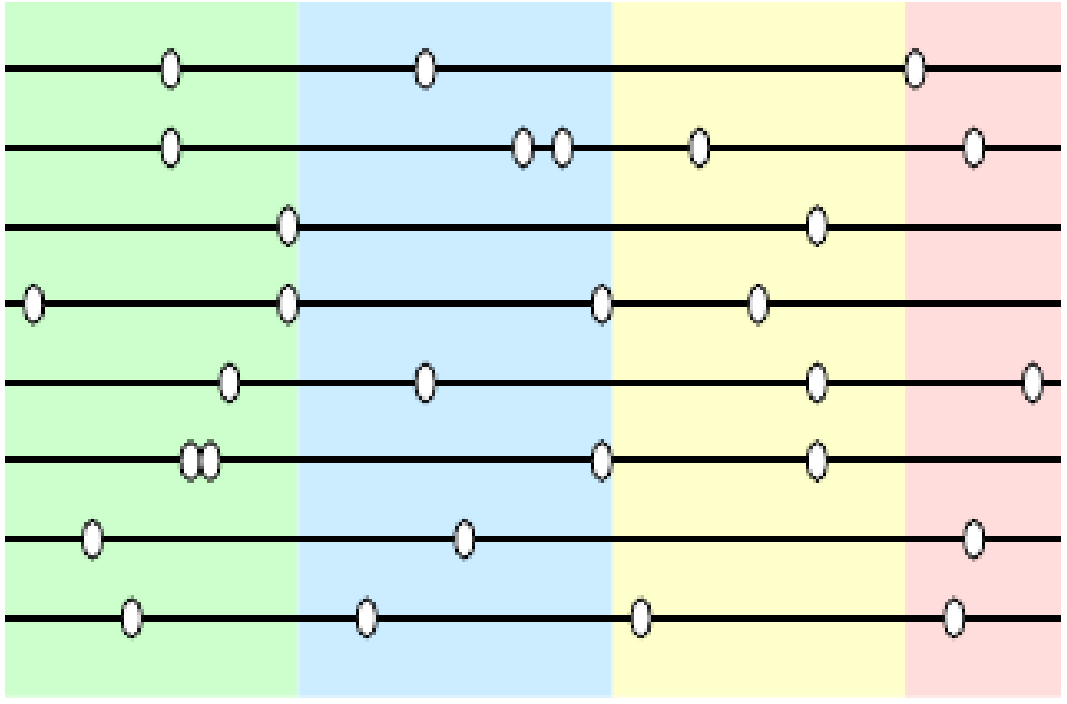
- Candidate genes/biological knowledge
- Functional categories
- Association to phenotypes

Common methods used to detect selection



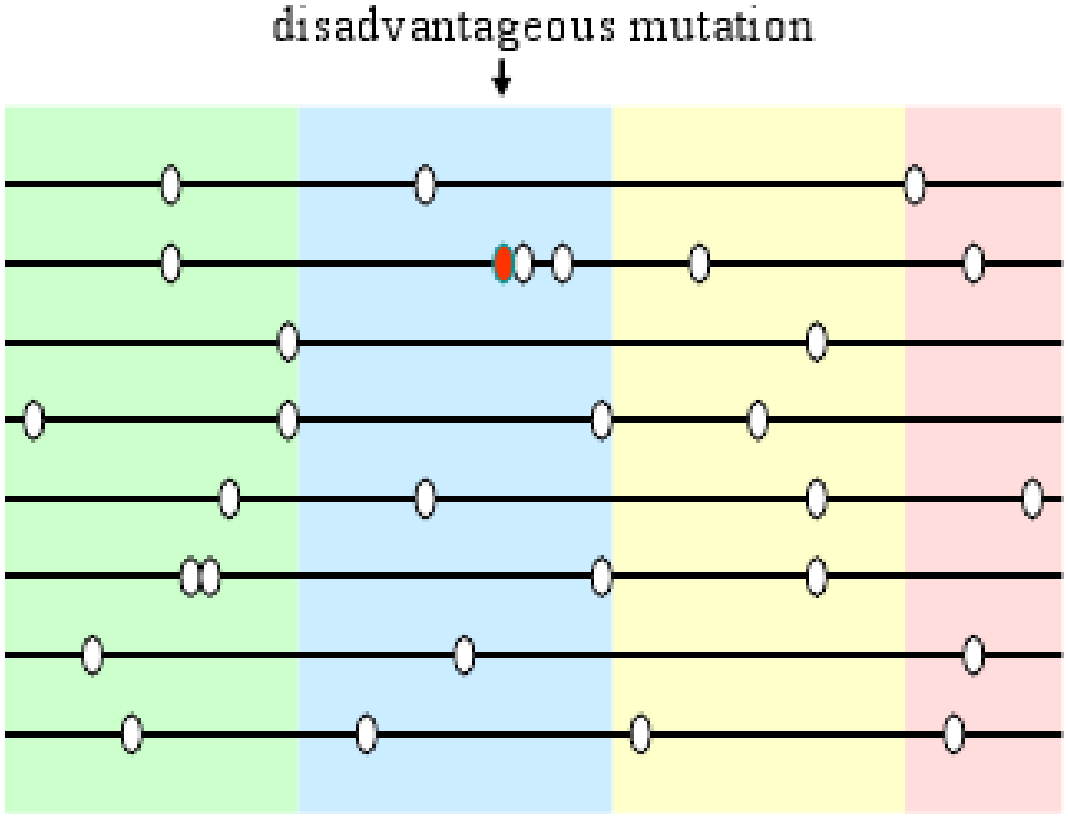
Signature of selection

- Neutral locus
- Lots of variability



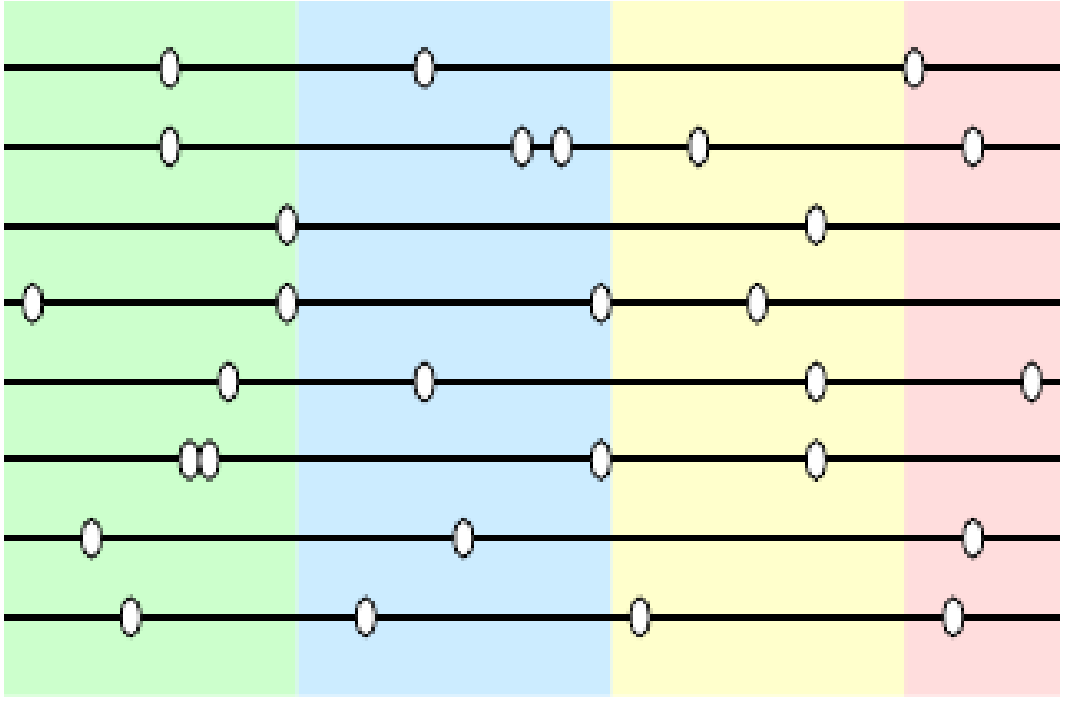
Signature of selection

- Mutation enters the population



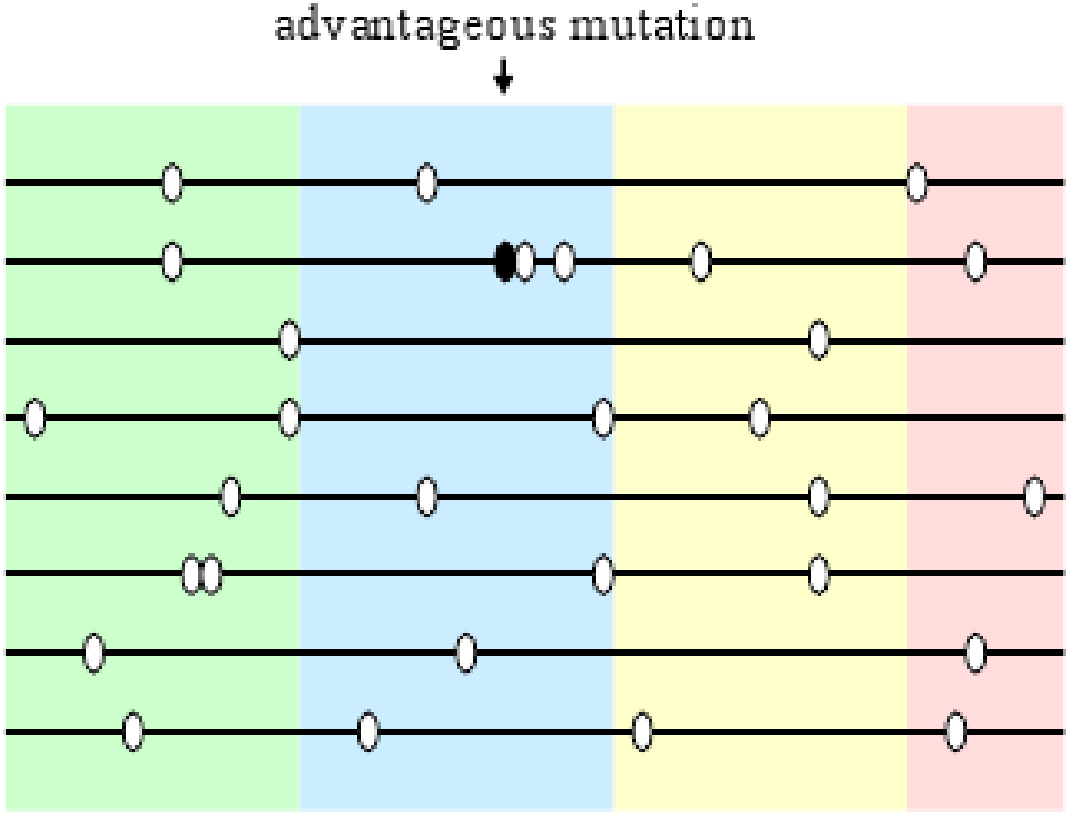
Signature of selection

- Negative selection removed the allele



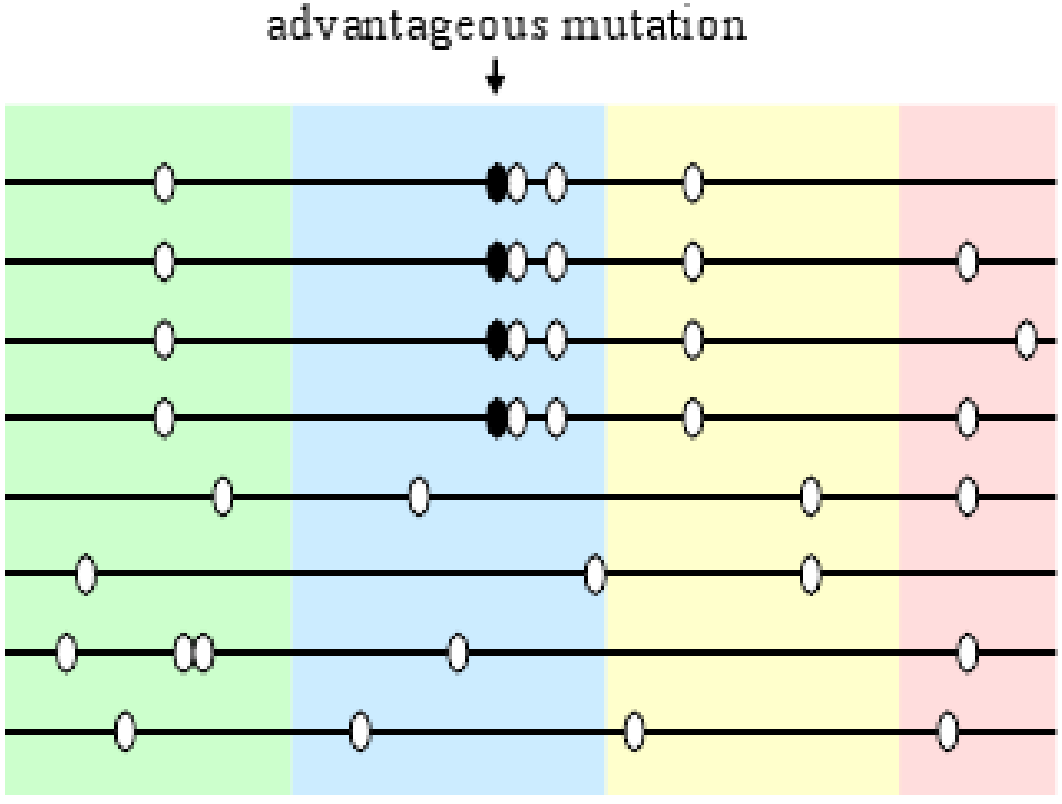
Signature of selection

- Mutation enters the population



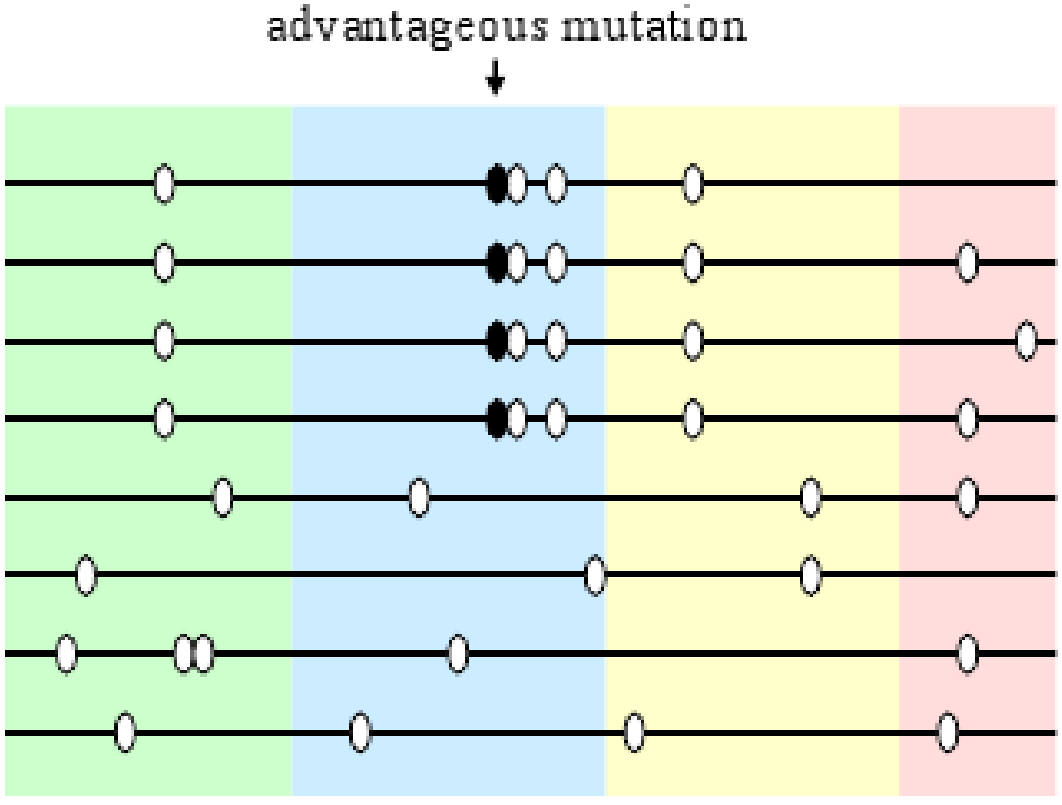
Signature of selection

- Mutation enters the population
- Mutation increases in frequency due to positive selection



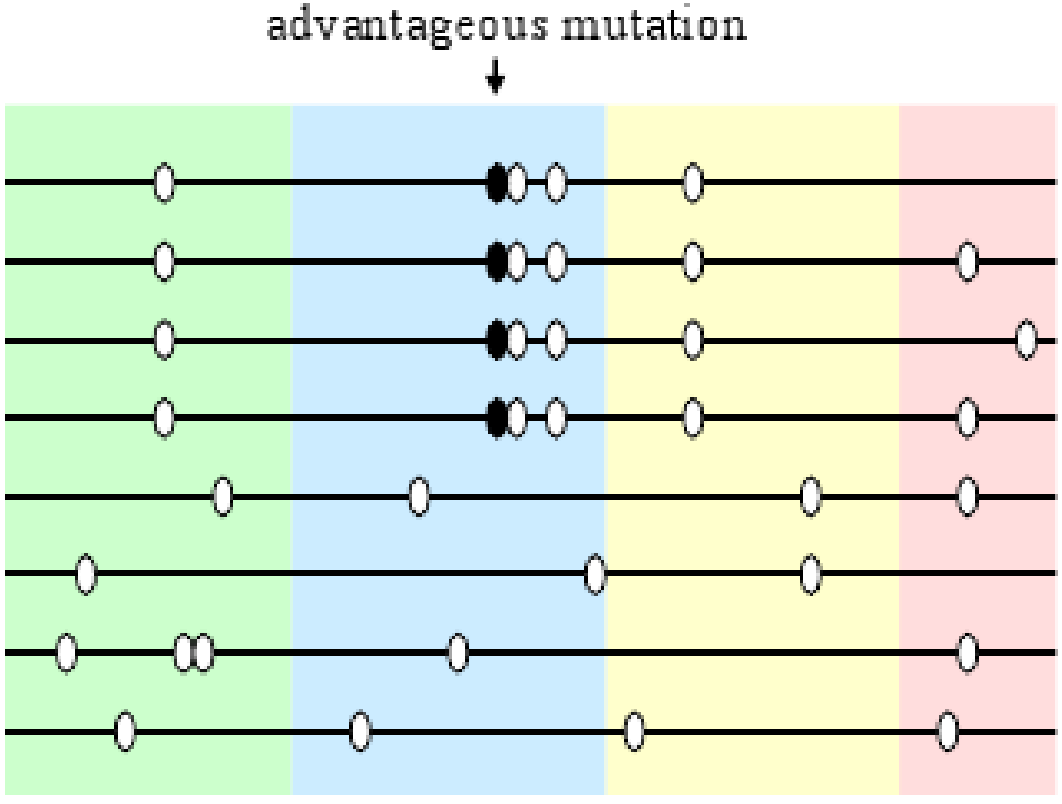
Signature of selection

- Increases LD
- Affects the variability

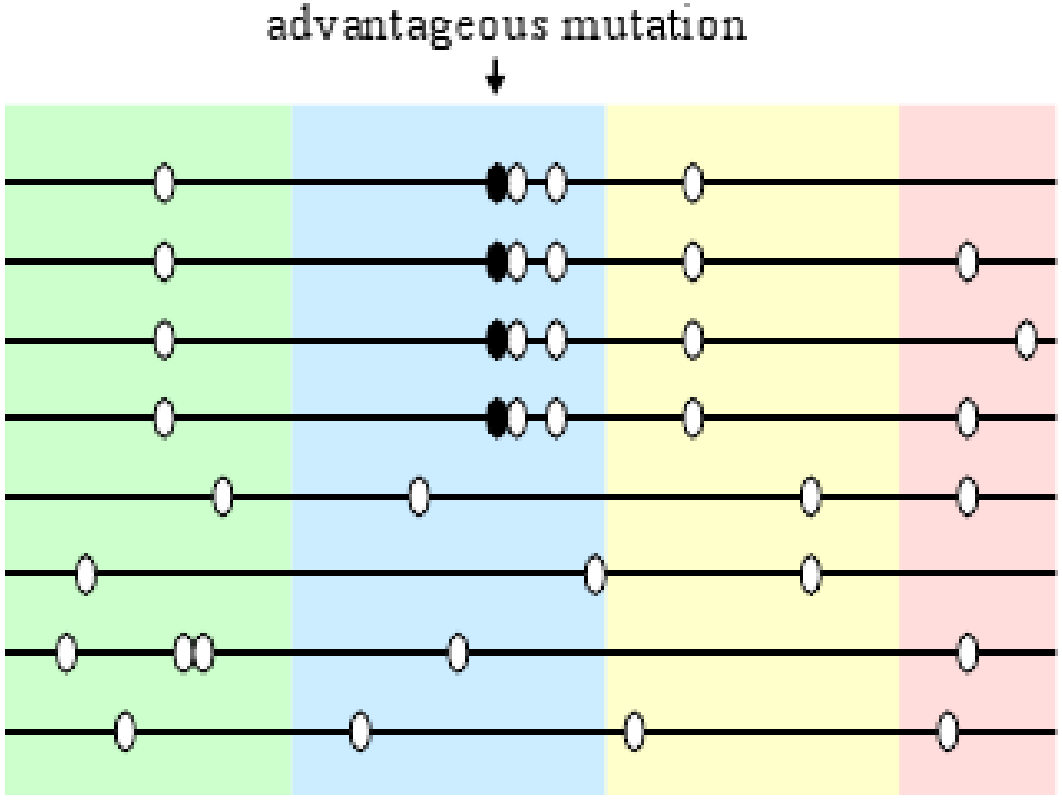


Signature of selection

- Increases haplotype similarity



Signature of selection

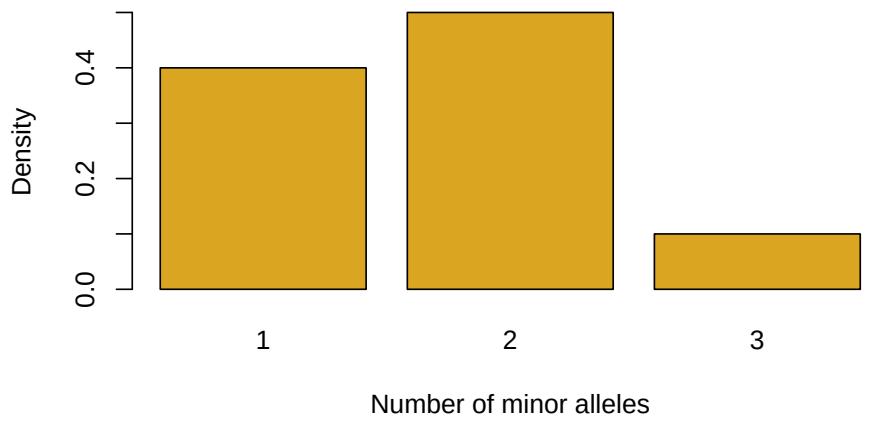


- Increases differences with other populations in the whole region

What is the site frequency spectrum

Ind										
1_1	T	C	G	T	C	T	C	A	A	T
1_2	T	C	G	T	C	T	C	C	A	G
2_1	A	G	G	T	C	G	C	C	A	T
2_2	A	C	G	T	G	G	T	C	A	T
3_1	A	C	T	A	G	G	C	C	T	T
3_2	A	C	T	A	G	G	T	C	A	T
# Minor	2	1	2	2	3	2	2	1	1	1

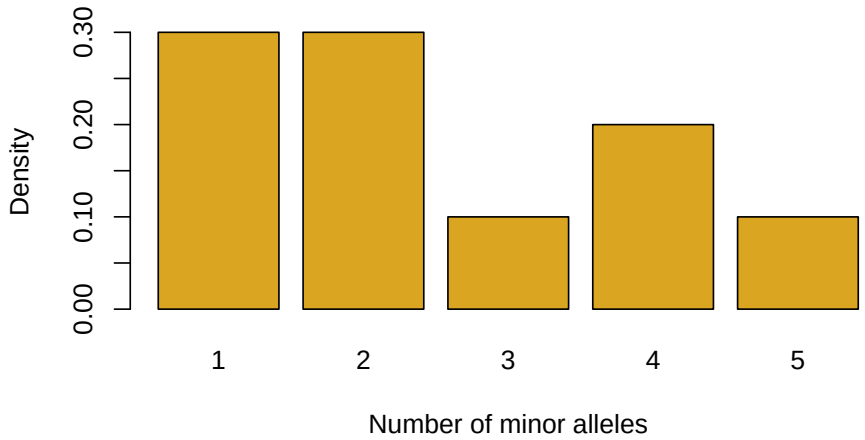
Number of minor alleles (folded) $\eta = (0.4, 0.5, 0.1)$



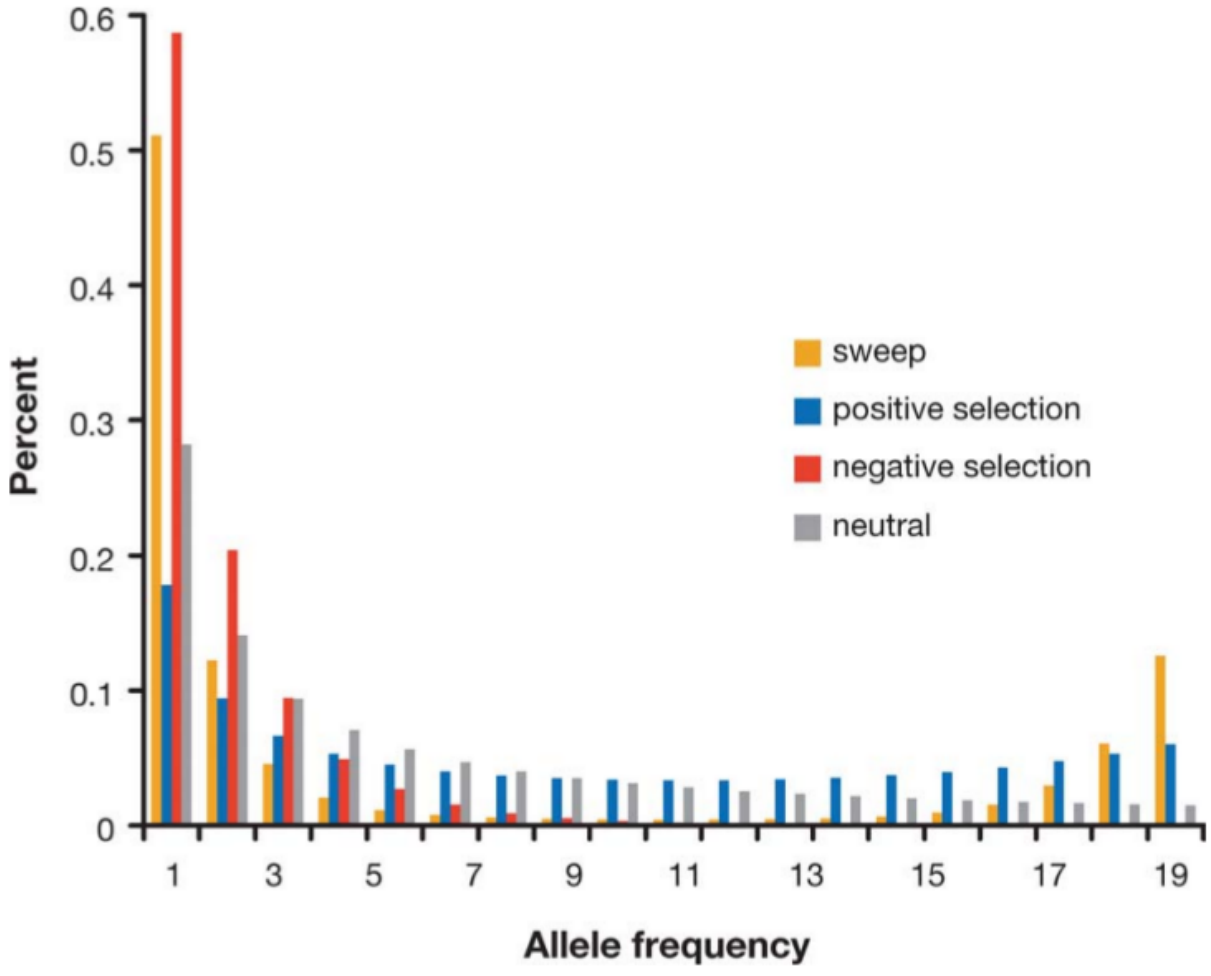
What is the site frequency spectrum

Ind										
1_1	T	C	G	T	C	T	C	A	A	T
1_2	T	C	G	T	C	T	C	C	A	G
2_1	A	G	G	T	C	G	C	C	A	T
2_2	A	C	G	T	G	G	T	C	A	T
3_1	A	C	T	A	G	G	C	C	T	T
3_2	A	C	T	A	G	G	T	C	A	T
Outgroup	A	C	T	T	C	T	C	C	A	G
# Derived	2	1	4	2	3	4	2	1	1	5

polarized SFS (unfolded) $\eta = (0.3, 0.3, 0.1, 0.2, 0.1)$



Frequency spectrum gives information about selection and demography



Thetas are based on the frequency spectrum

Relative thetas

Watterson $\theta_W = a^{-1} \underbrace{\sum_{i=1}^{n-1} \eta_i}_{\text{Segregating sites}}$, where $a = \sum_{i=1}^{n-1} 1/i$

Tajima $\theta_T = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)\eta_i$

Tajima's D

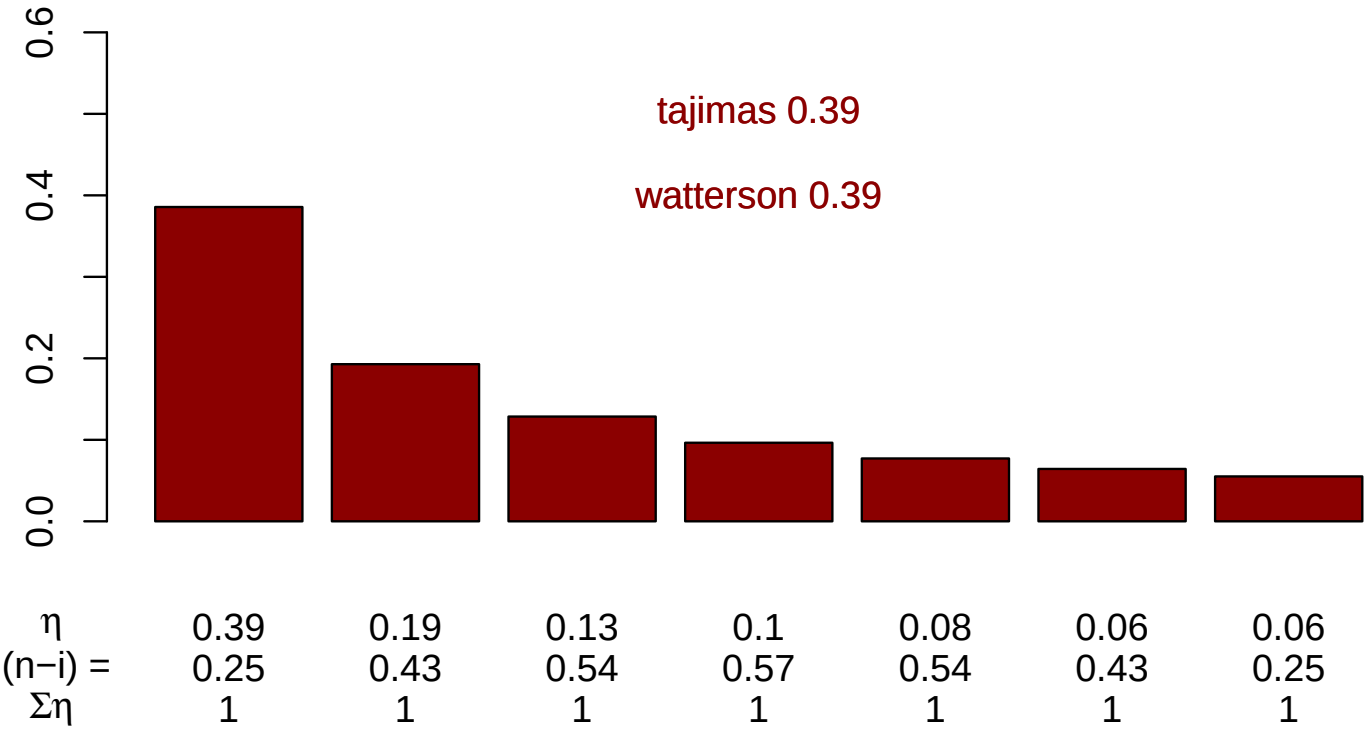
$D = \frac{\theta_T - \theta_W}{\sqrt{Var(\theta_T - \theta_W)}}$ under a neutral model* $\theta_T = \theta_W$

Theta are based on the frequency spectrum

Watterson $\theta_W = a^{-1} \sum_{i=1}^{n-1} \eta_i$, where $a = \sum_{i=1}^{n-1} 1/i$

Tajima $\theta_T = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)\eta_i$

4 diploid individuals - excluding non-variable sites

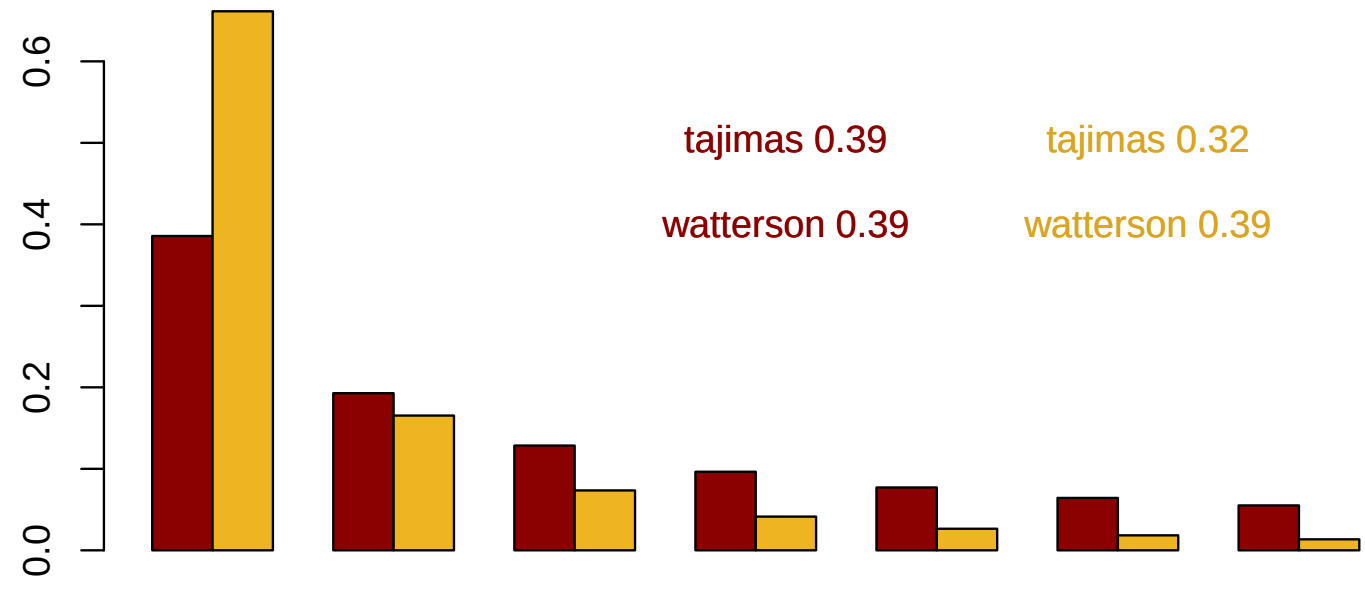


Theta are based on the frequency spectrum

Watterson $\theta_W = a^{-1} \sum_{i=1}^{n-1} \eta_i$, where $a = \sum_{i=1}^{n-1} 1/i$

Tajima $\pi = \theta_T = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)\eta_i$

4 diploid individuals



η	0.66	0.17	0.07	0.04	0.03	0.02	0.01
η	0.39	0.19	0.13	0.1	0.08	0.06	0.06
i(n−i) =	0.25	0.43	0.54	0.57	0.54	0.43	0.25
Ση	1	1	1	1	1	1	1

Thetas are based on the frequency spectrum

Watterson $\theta_W = a^{-1} \sum_{i=1}^{n-1} \eta_i$, where $a = \sum_{i=1}^{n-1} 1/i$

Tajima $\pi = \theta_T = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)\eta_i$

Fu & Li $\theta_{FL} = \eta_1$

Fay & Wu $\theta_H = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i^2 \eta_i$

Zeng, Fu,Shi and Wu $\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\eta_i$

general $\hat{\theta} = \sum_{i=0}^{n} \alpha_i \eta_i$

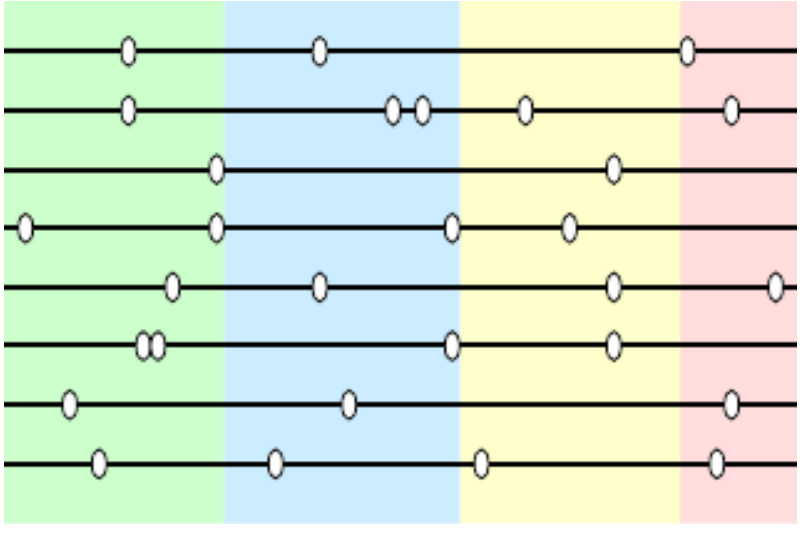
Test statistics

$D = \frac{\theta_1 - \theta_2}{\sqrt{Var(\theta_1 - \theta_2)}}$ under a neutral model* $\theta_1 = \theta_2$

Difference weighting schemes for the SFS

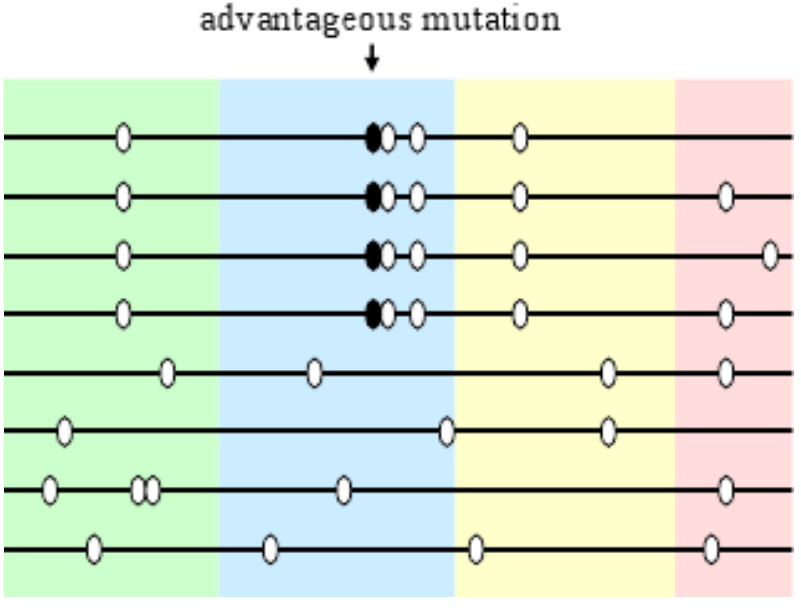
Why does selection affect the SFS

neutral



high pairwise distance

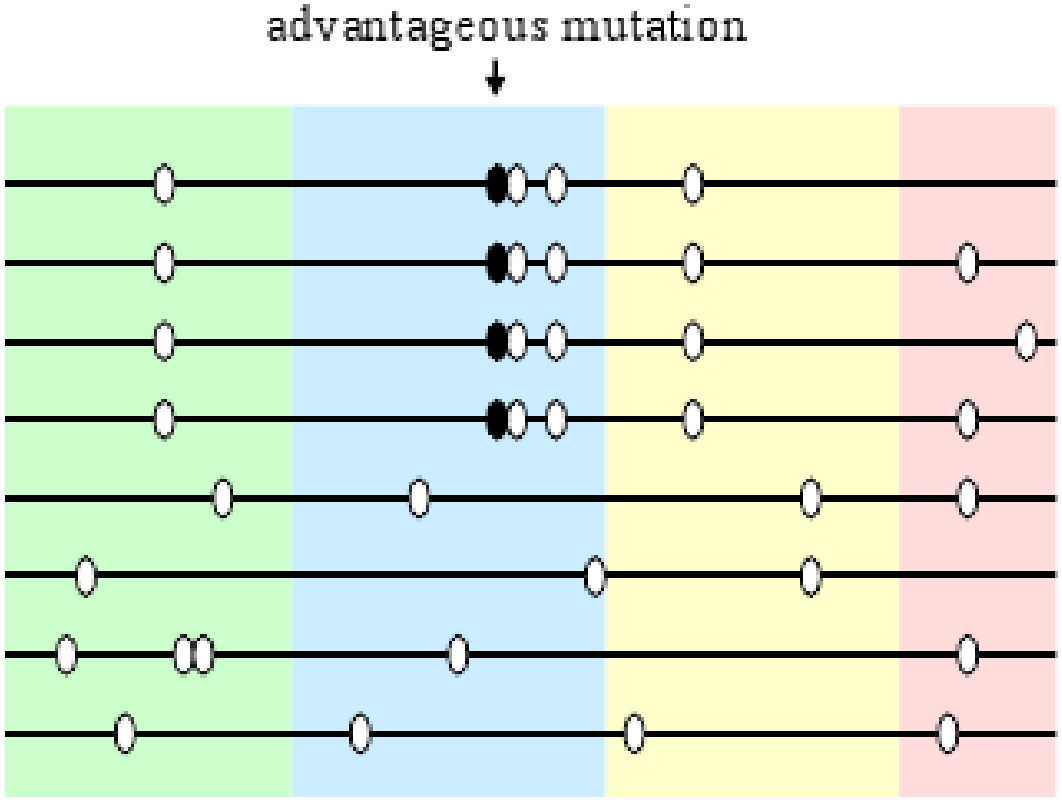
Sweep



low pairwise distance

Signature of selection

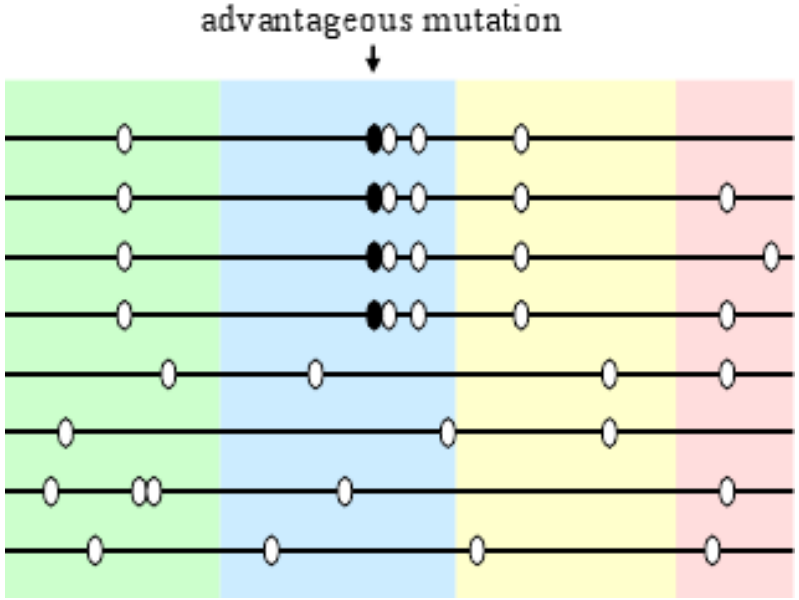
- Mutation enters the population
- Mutation increases in frequency due to positive selection
- **Increases LD**
- Affects the variability
- **Increases haplotype similarity**
- Increases differences with other populations in the whole region



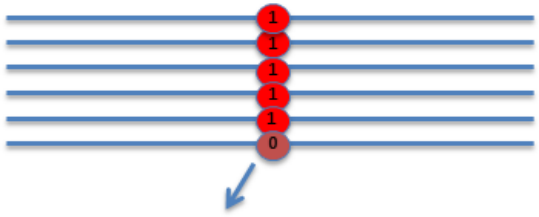
EHH - Extended Haplotype Homozygosity

What is EHH?

Extended haplotype homozygosity (EHH): EHH at distance x from the core region is the probability that two randomly chosen chromosomes carry a tested core haplotype are homozygous at all SNPs for the entire interval from the core region to the distance x .



Extended Haplotype Homozygosity



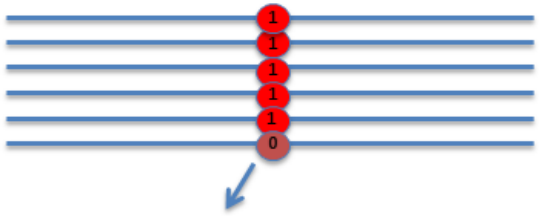
Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Core SNP

slides stolen from Matteo Fumagalli

Extended Haplotype Homozygosity

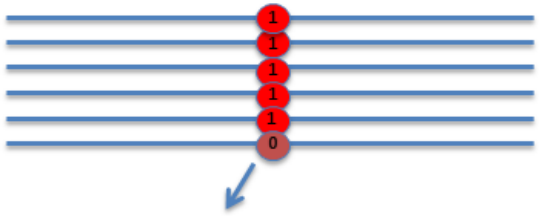


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

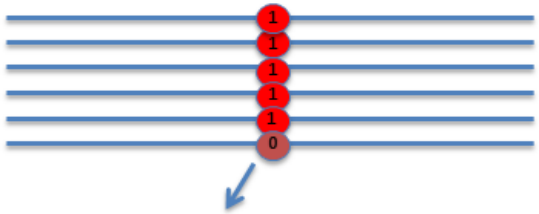


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Sum across all unique haplotypes carrying the core SNP

Extended Haplotype Homozygosity



Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

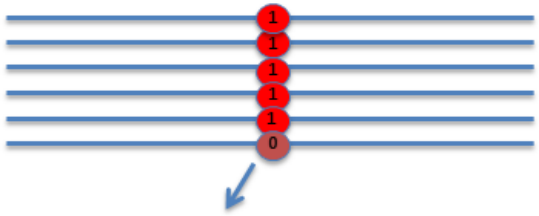
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n_h is haplotype frequency of h

n_h is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

Extended Haplotype Homozygosity



Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

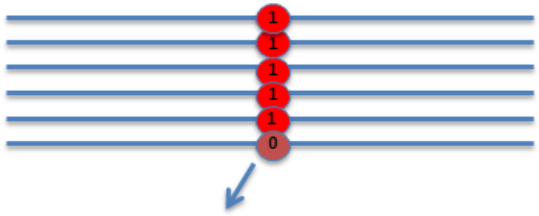
n_h is haplotype frequency of h

n_h is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

$$EHH_c(x_i = 0) = ?$$

Extended Haplotype Homozygosity



Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

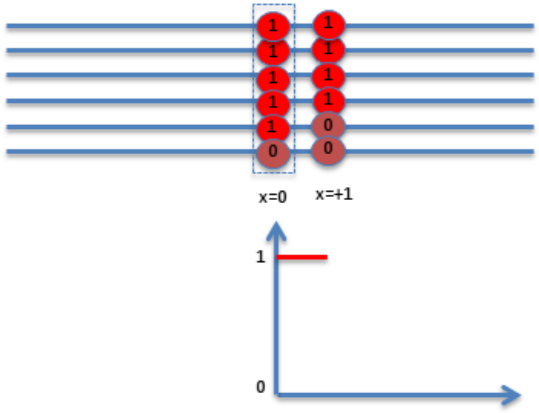
n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

$$EHH_c(x_i = 0) = \frac{\binom{5}{2}}{\binom{5}{2}} = 1$$

Extended Haplotype Homozygosity

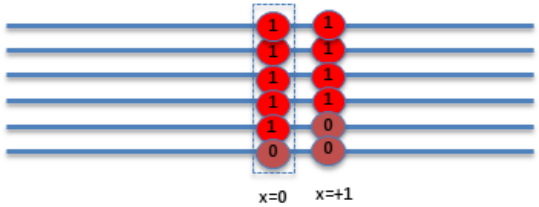


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +1) = ?$$

How many unique haplotypes carrying the core SNP?
What is their frequency?

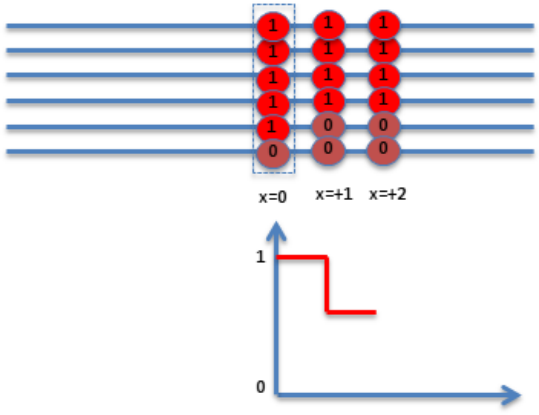
Extended Haplotype Homozygosity



$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +1) = \frac{\binom{4}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{6+0}{10} = 0.60$$

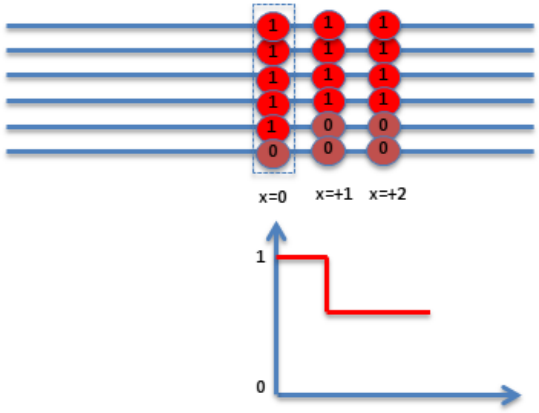
Extended Haplotype Homozygosity



$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +2) = ?$$

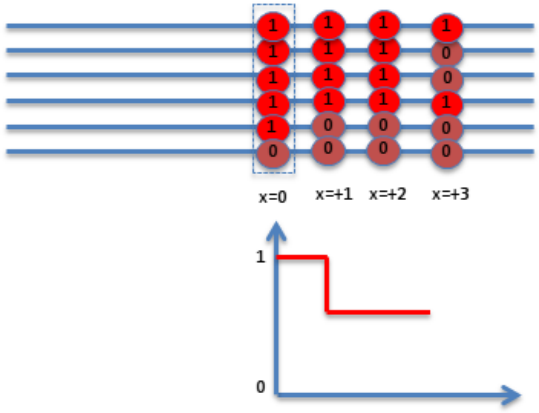
Extended Haplotype Homozygosity



$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +2) = EHH_c(x_i = +1) = 0.60$$

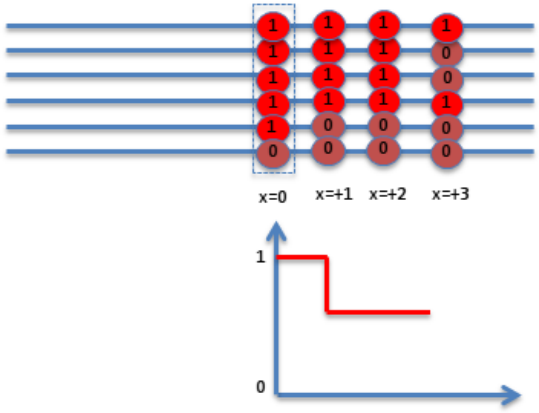
Extended Haplotype Homozygosity



$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?

Extended Haplotype Homozygosity

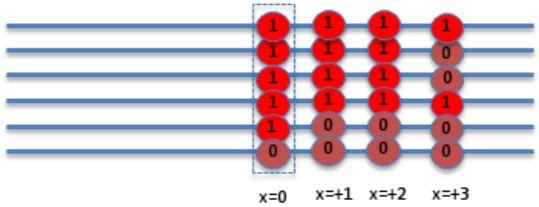


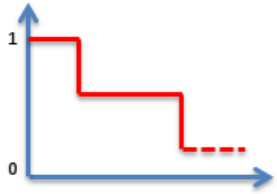
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?
1111 with freq=2
1110 with freq=2
1000 with freq=1

$$EHH_c(x_i = +3) = ?$$

Extended Haplotype Homozygosity



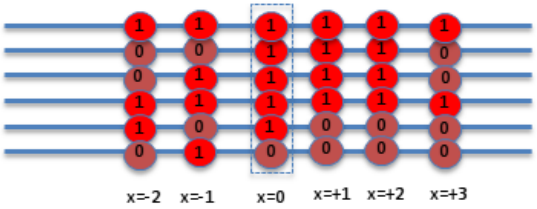


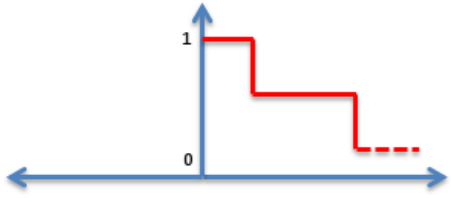
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?
1111 with freq=2
1110 with freq=2
1000 with freq=1

$$EHH_c(x_i = +3) = \frac{\binom{2}{2} + \binom{2}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+1+0}{10} = 0.20$$

Extended Haplotype Homozygosity





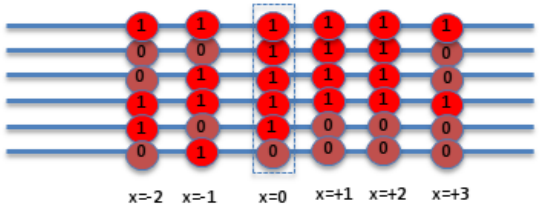
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n	*n choose 2*
1	0
2	1
3	3
4	6
5	10
6	15

$$EHH_c(x_i = -1) = ?$$
$$EHH_c(x_i = -2) = ?$$

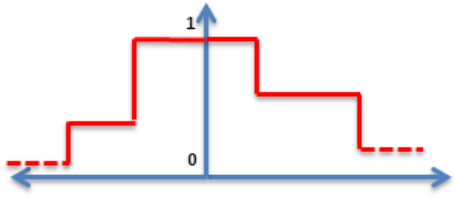
Comment on differences (if any) between *EHH(x=+2)* and *EHH(x=-2)*.

Extended Haplotype Homozygosity



$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n	*n choose 2*
1	0
2	1
3	3
4	6
5	10
6	15

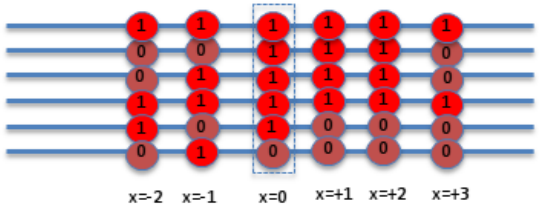


$$EHH_c(x_i = -1) = \frac{\binom{3}{2} + \binom{2}{2}}{\binom{5}{2}} = \frac{3+1}{10} = 0.4$$

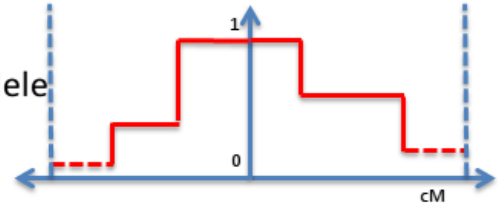
$$EHH_c(x_i = -2) = \frac{\binom{2}{2} + \binom{1}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+0+0}{10} = 0.1$$

Comment on differences (if any) between *EHH(x=+2)* and *EHH(x=-2)*?

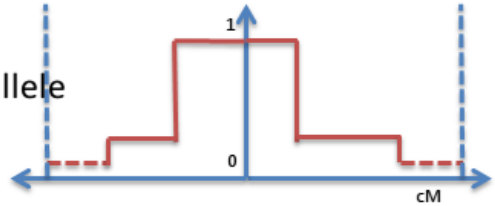
Integrated Haplotype Score



For the derived allele

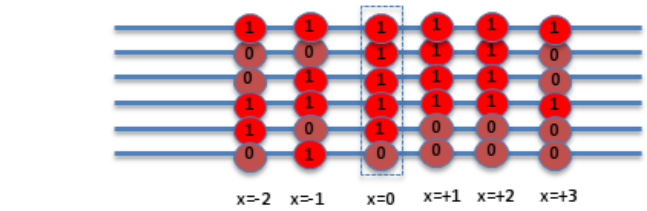


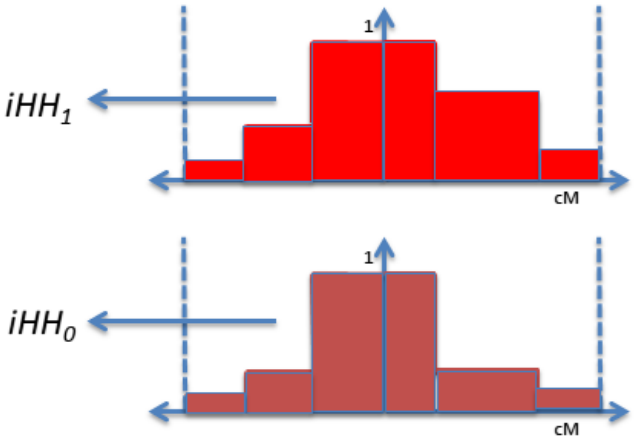
For the ancestral allele



iHs

Integrated Haplotype Score

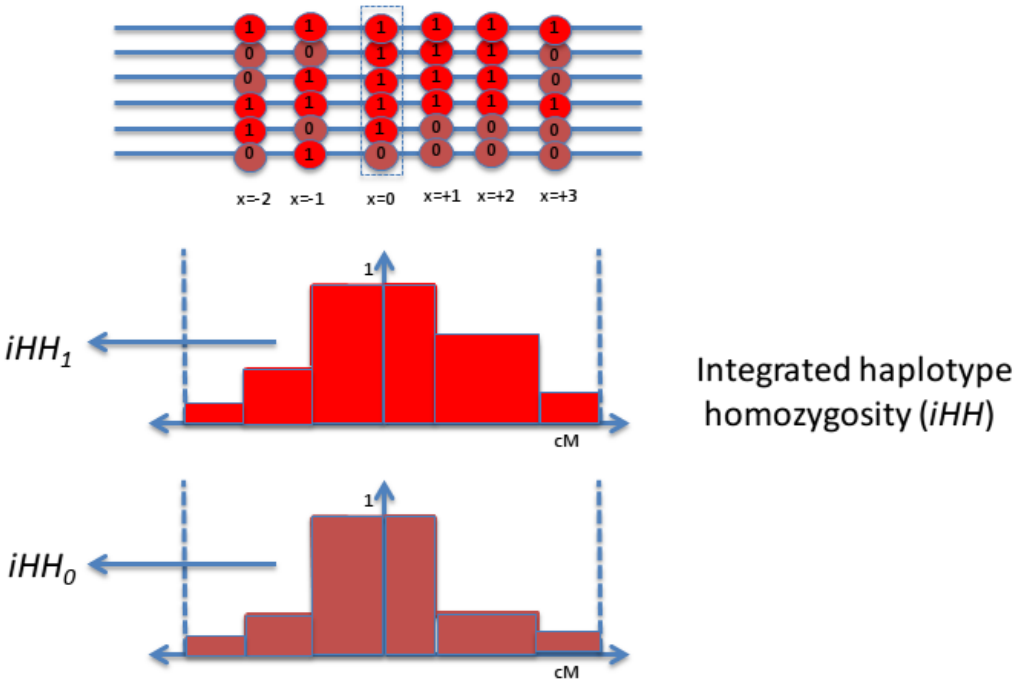




Integrated haplotype homozygosity (

iHs

Integrated Haplotype Score

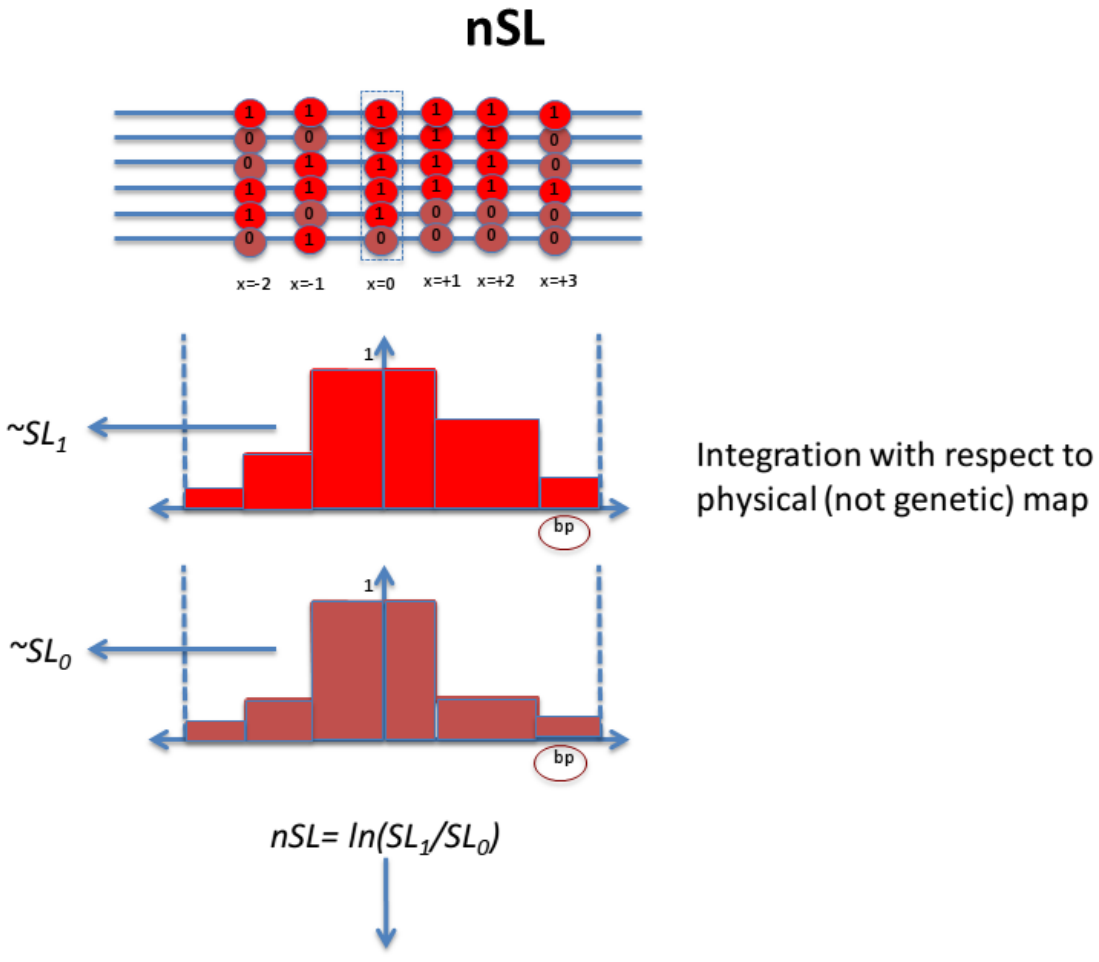


Integrated haplotype score:

$iHs = ln(iHH_1/iHH_0)$

Genome-wide normalization in frequency bins (to mean=0 and sd=1)

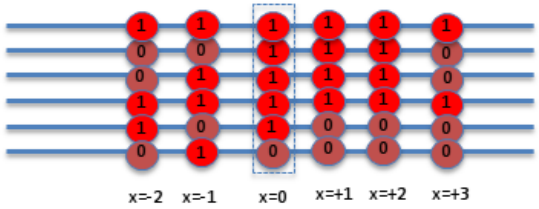
often $|iHs|$ is used

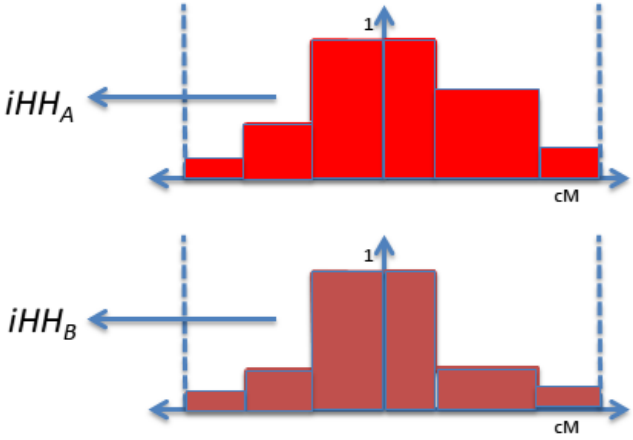


Integration with respect to physical (not genetic) map

Genome-wide normalization in frequency bins (to mean=0 and sd=1)

Cross-population Extended Haplotype Homozygosity





Integrated haplotype homozygosity (*iHH*) for **populations** A and B

Integrated haplotype score:

$XP\text{-}EHH = ln(iHH_A/iHH_B)$

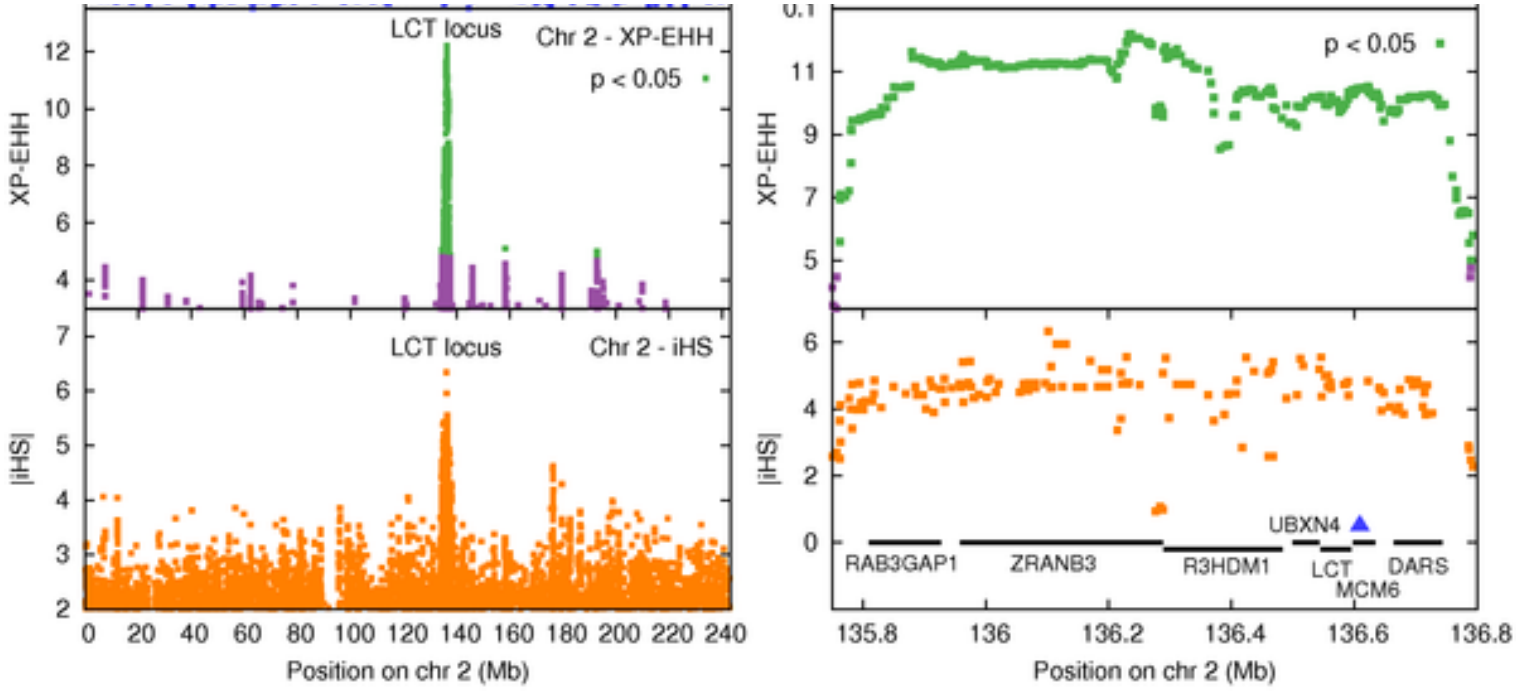
Genome-wide normalization in frequency bins (to mean=0 and sd=1)

Exercises

Let see how the haplotype methods works on famous examples of human adaptation (LCT).

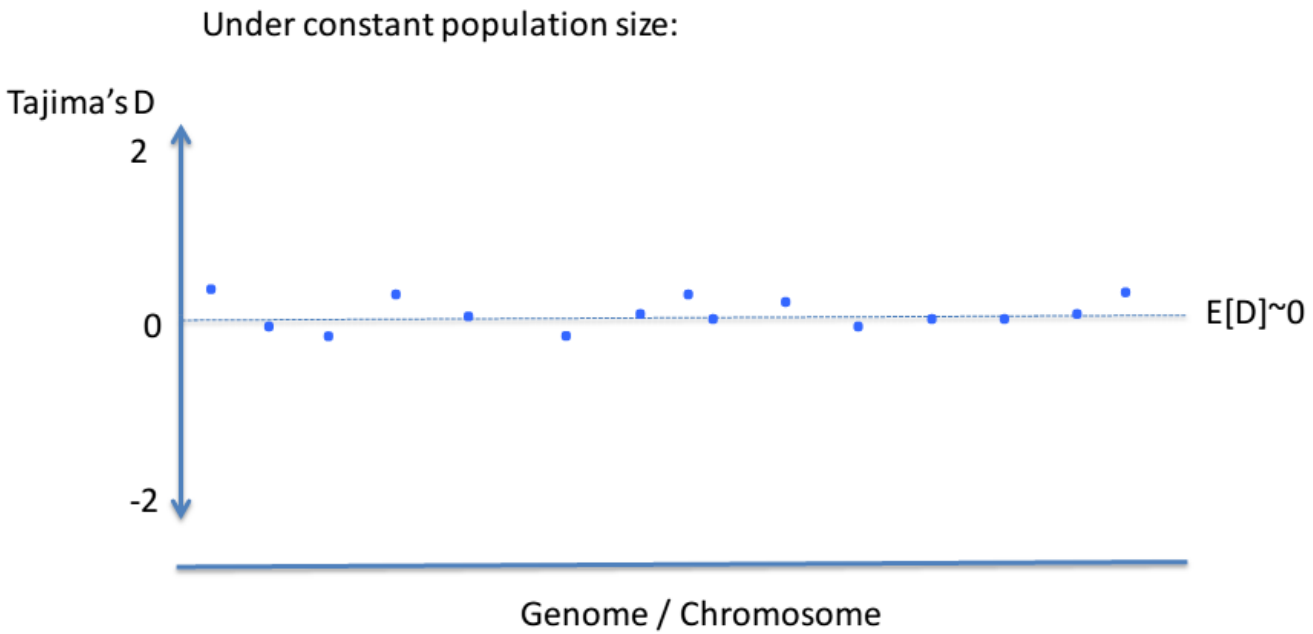
go to

`http://popgen.dk/albrecht/EMBO2021/`


LCT locus
Chr 2 - XP-EHH
p < 0.05
XP-EHH
LCT locus
Chr 2 - iHS
|iHS|
Position on chr 2 (Mb)
p < 0.05
UBXN4
RAB3GAP1
ZRANB3
R3HDM1
LCT
MCM6
DARS
Position on chr 2 (Mb)

How to assess significance

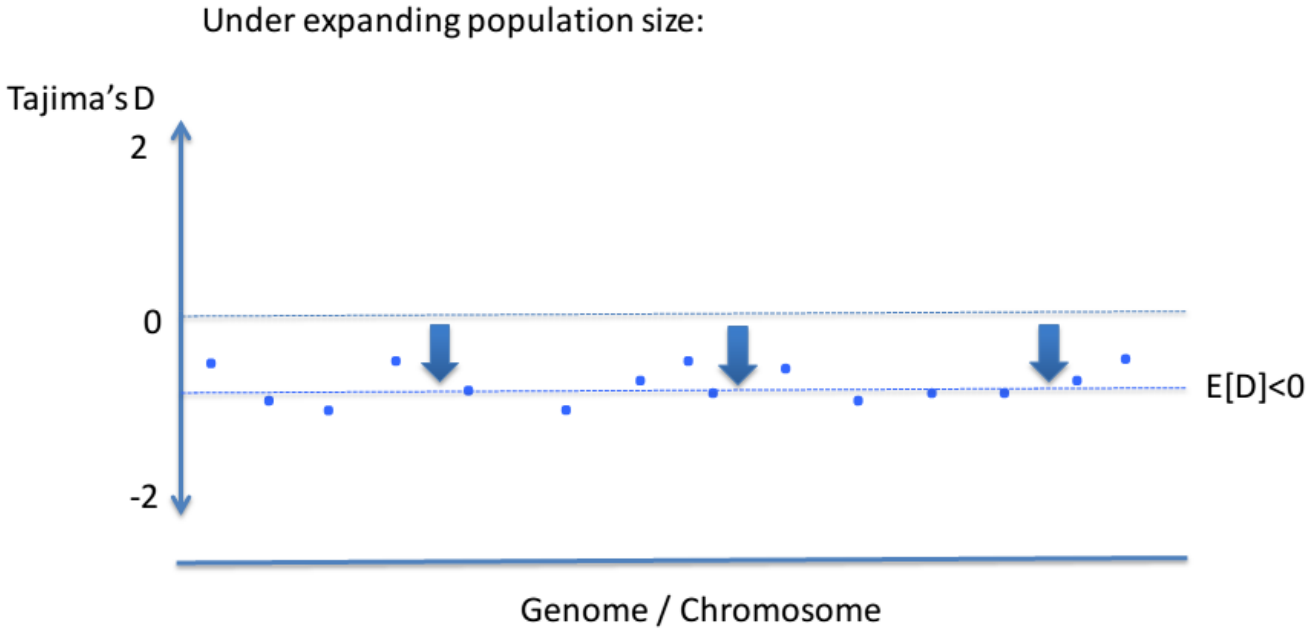
How to take neutral confounding factors into account?



slides stolen from Matteo Fumagalli

How to assess significance

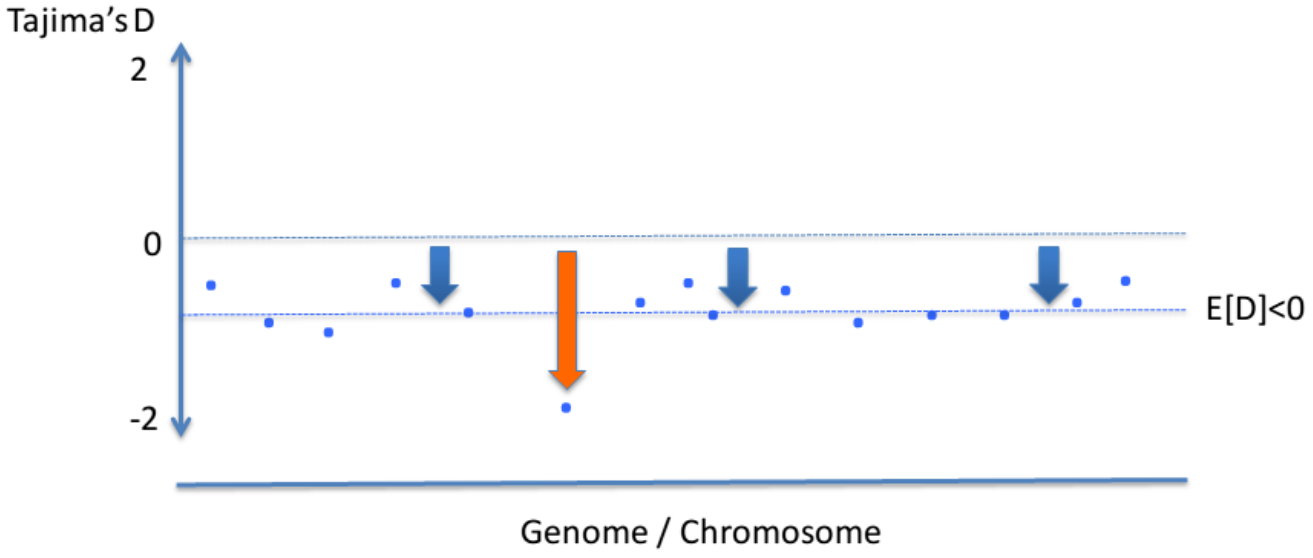
How to take neutral confounding factors into account?



How to assess significance

How to take neutral confounding factors into account?

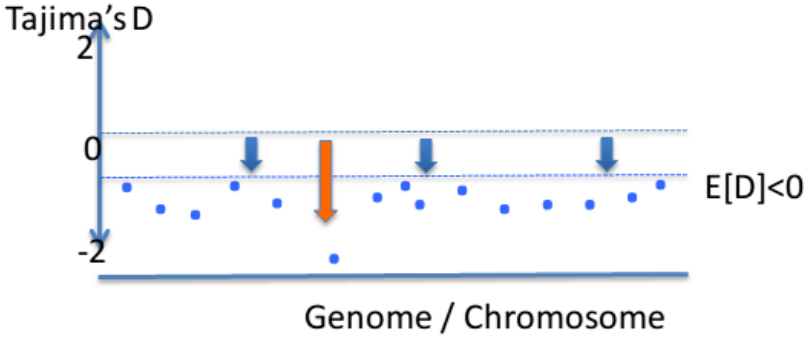
Under expanding population size and positive selection:



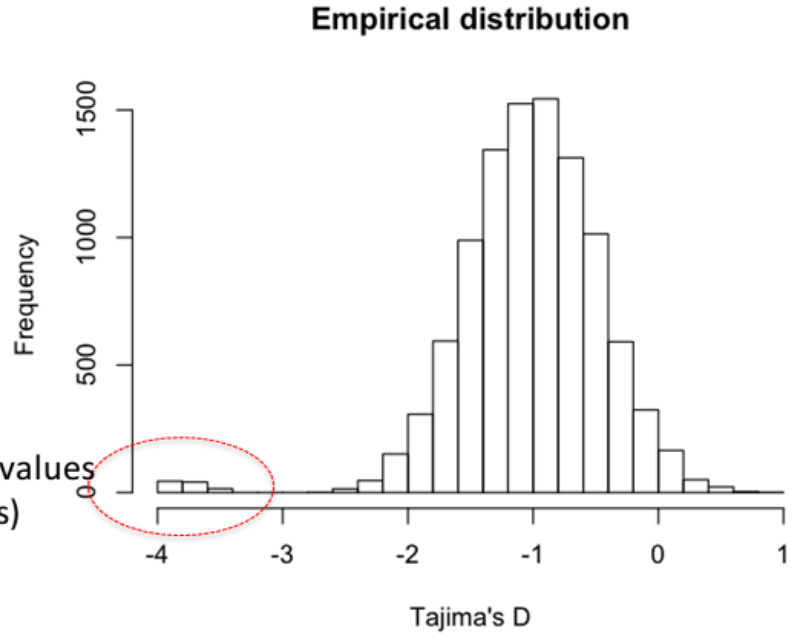
- Demography affects all loci equally, while selection changes local patterns

How to assess significance

Outlier approach

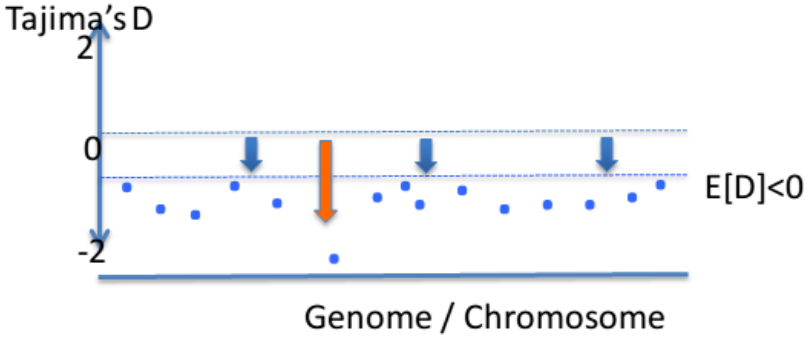


Assign empirical *p*-values (ranked percentiles)

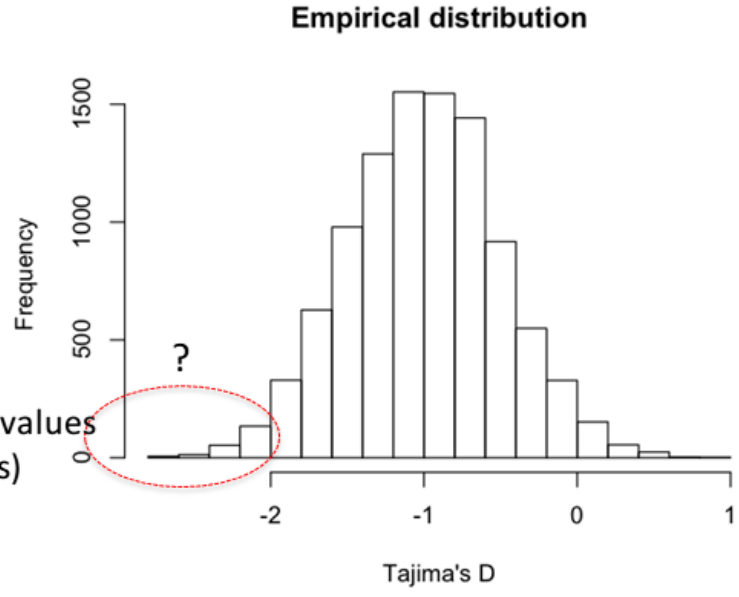


How to assess significance

Outlier approach



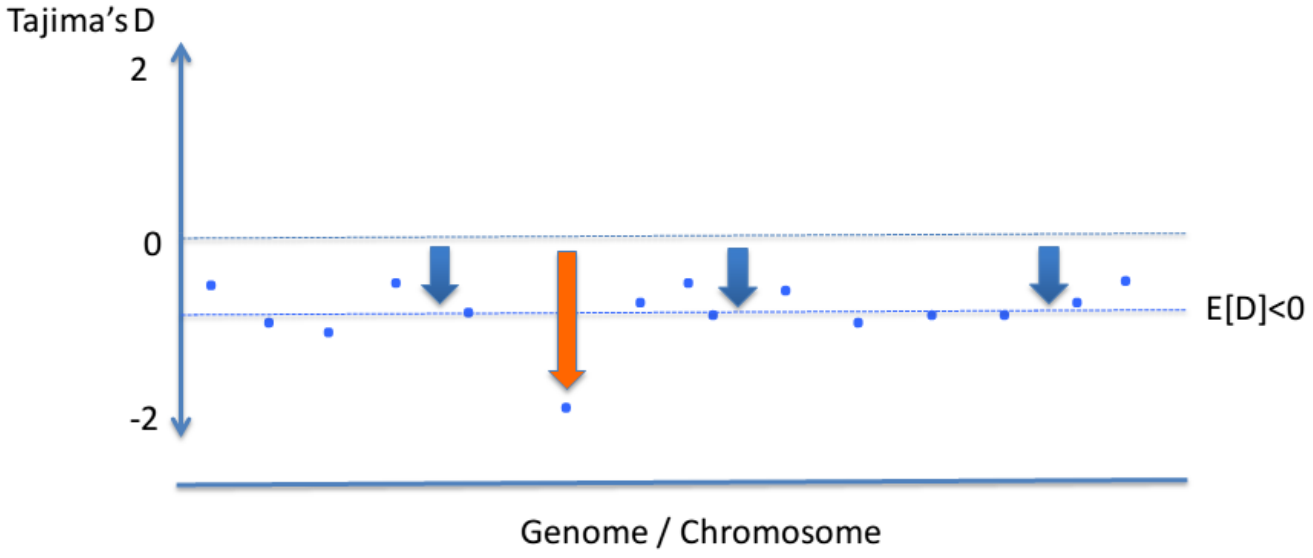
Assign empirical *p*-values (ranked percentiles)



How to assess significance

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

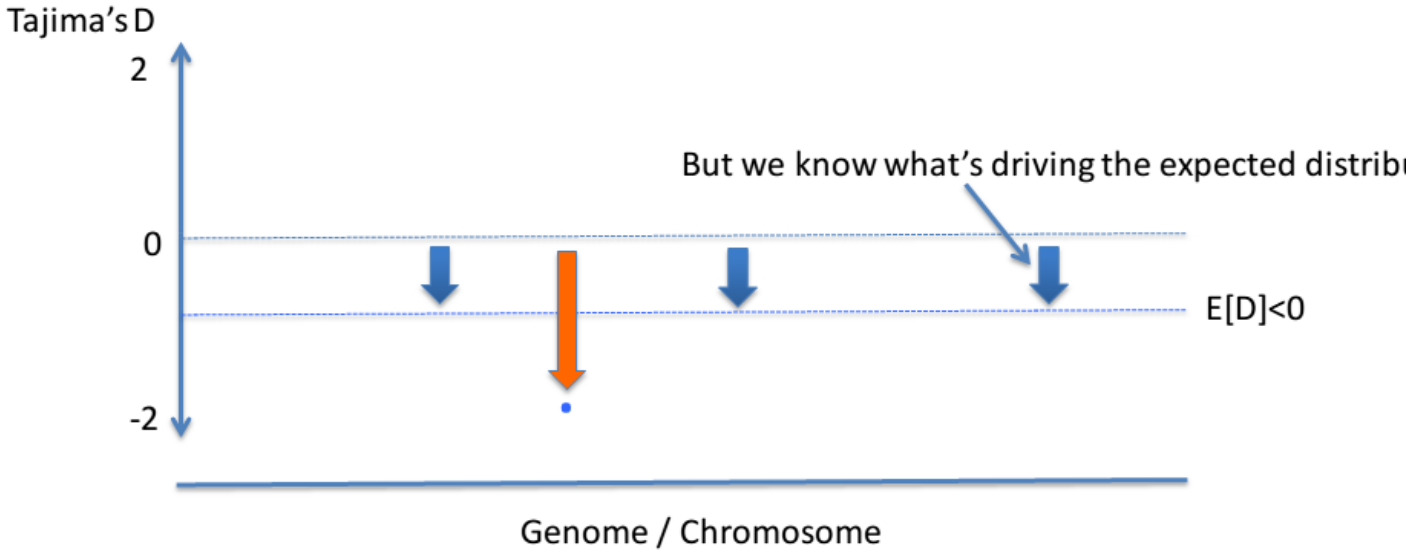


- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

How to assess significance

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

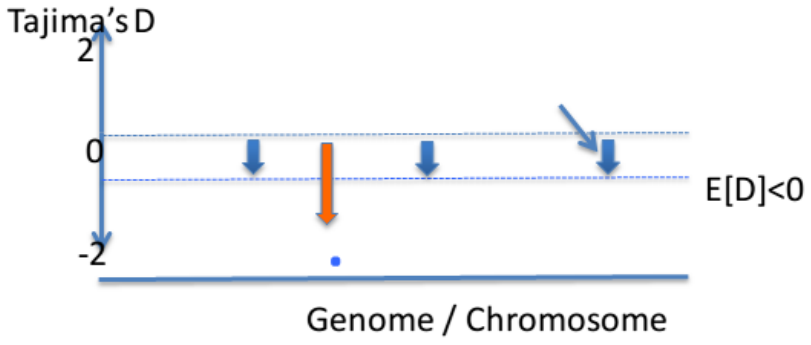


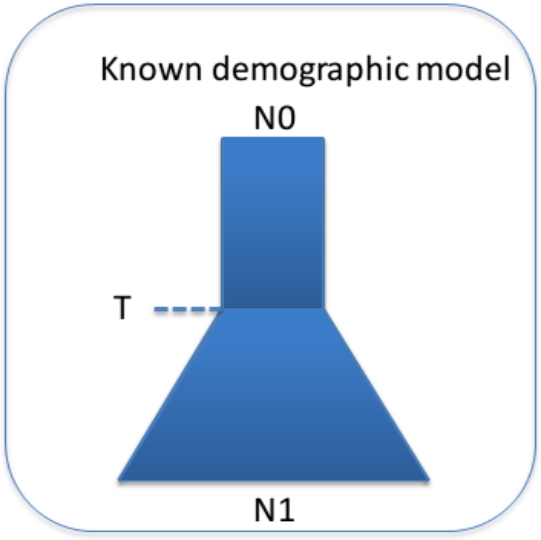
- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

How to assess significance

Simulations-based approach

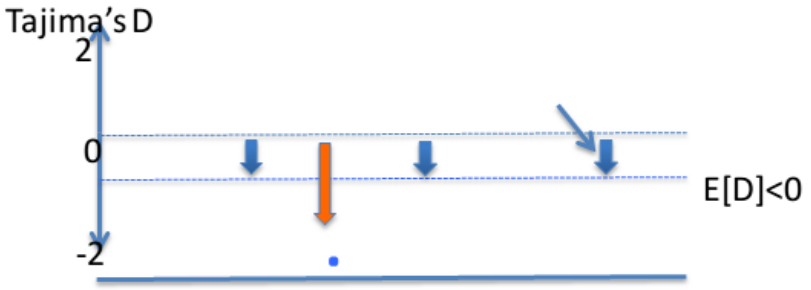
e.g. msms



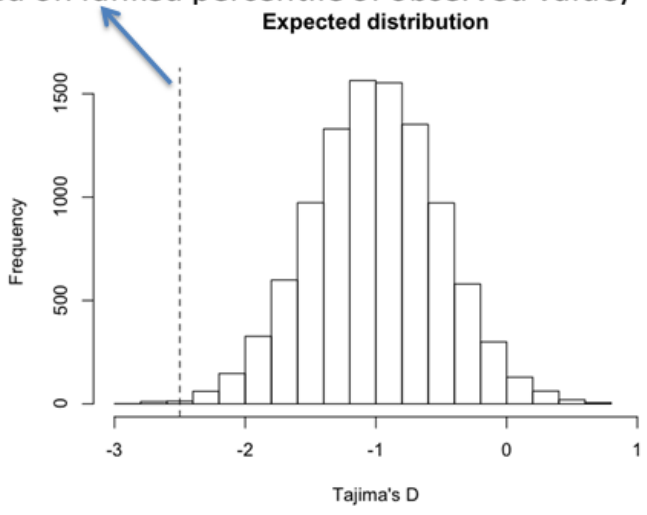


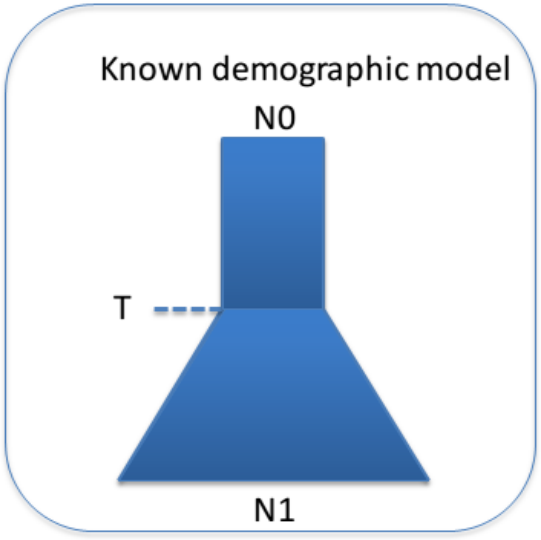
How to assess significance

Simulations-based approach



Assign *p*-values
(based on ranked percentile of observed value)





Exercises

Let see how variability π and Tajimas D performs on famous examples of human adaptation.

go to

`https://github.com/aalbrechtsen/embo2022`

Graphics

When you will run analysis on the server you will need graphic (see above link)