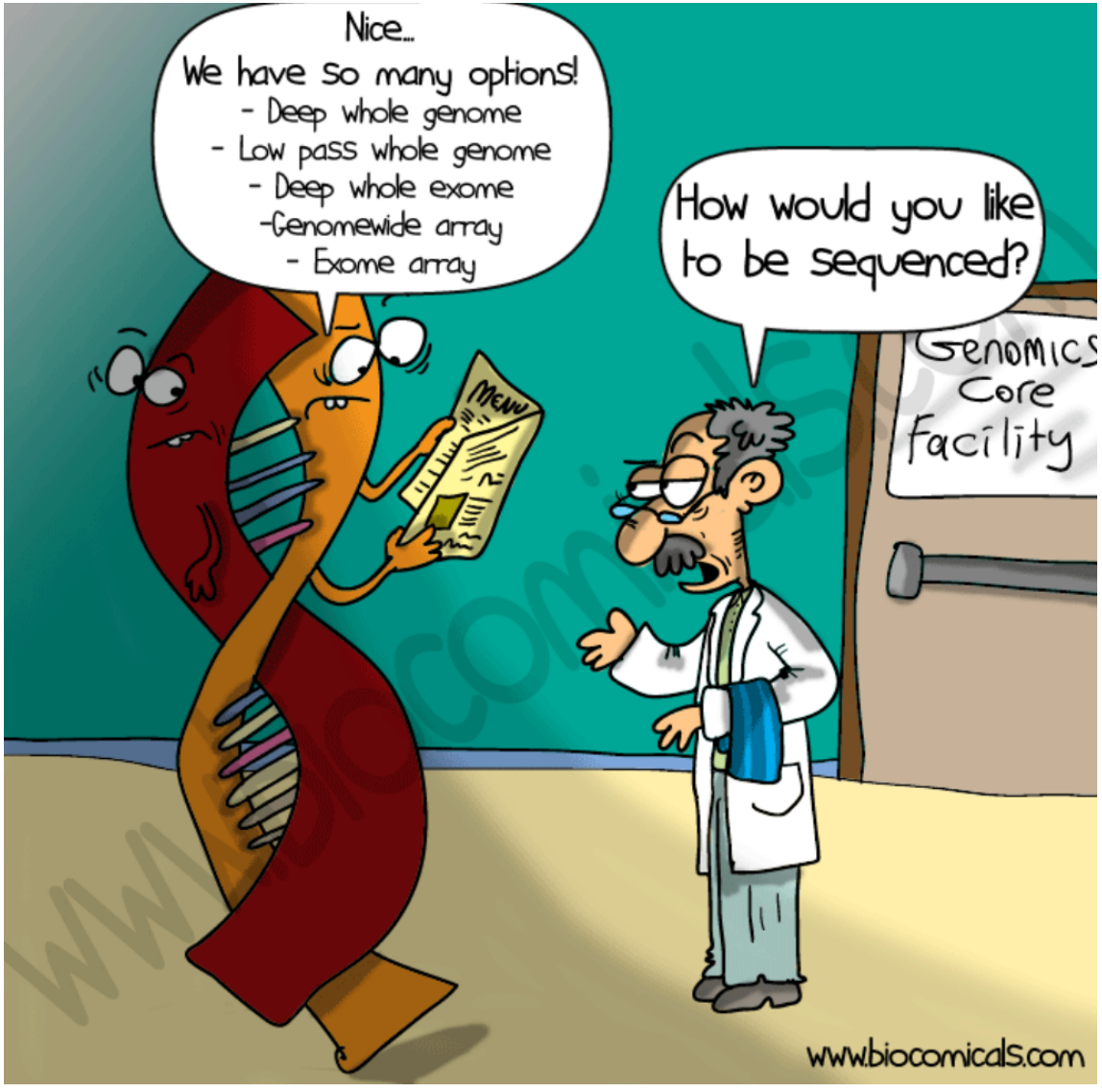
Inferring sites with recent or ongoing selection for NGS data(+admixture/population structure)

http://popgen.dk/albrecht/BAG2018/web/

Anders Albrechtsen

Sequencing types



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

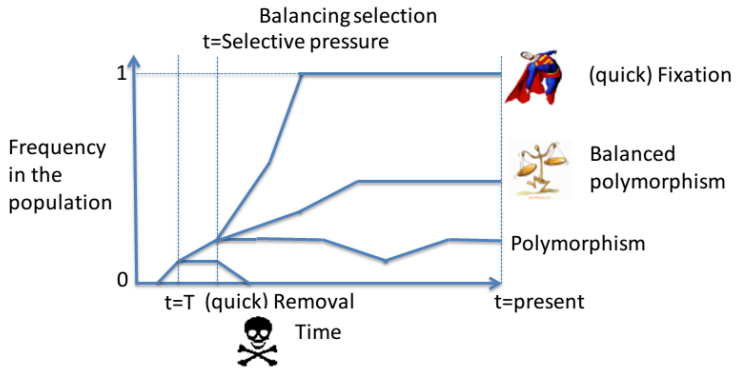
- Depth lower than 10X
- Often a financial choice
- Ancient DNA

Ultra low sequencing

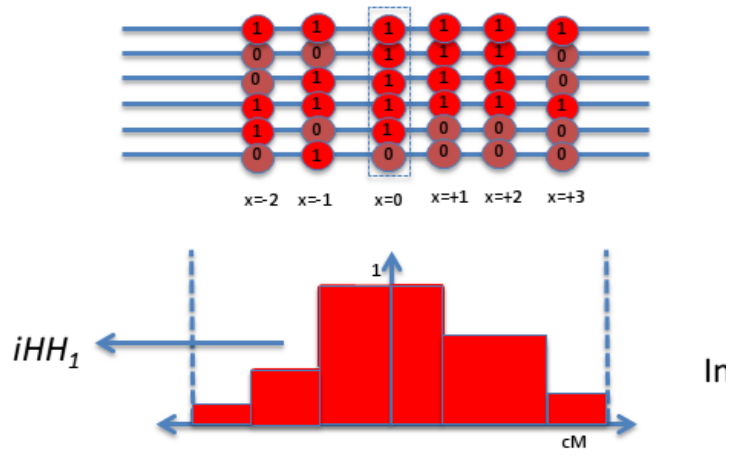
- Depth lower than 1X
- by product of capture data
-

This morning

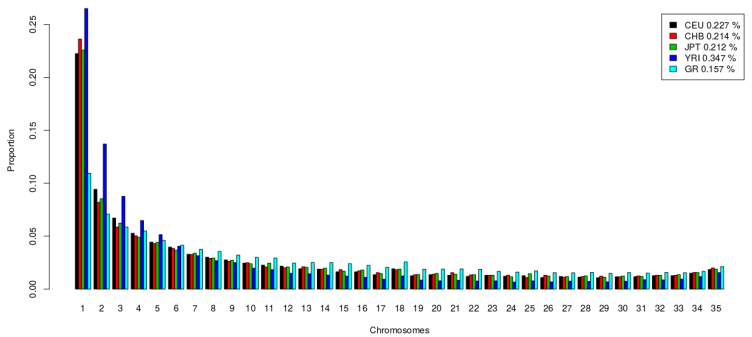
Short intro to recent selection



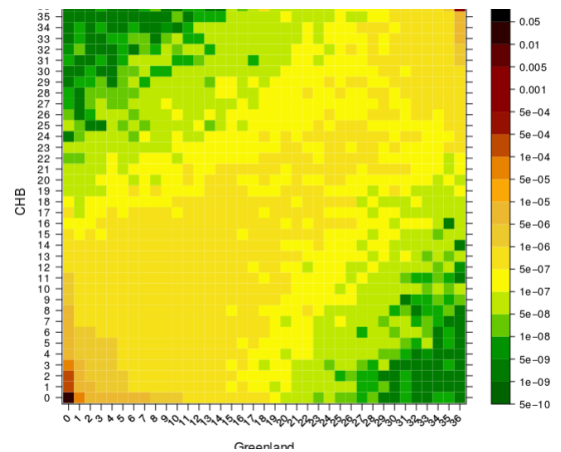
EHH



SFS for NGS

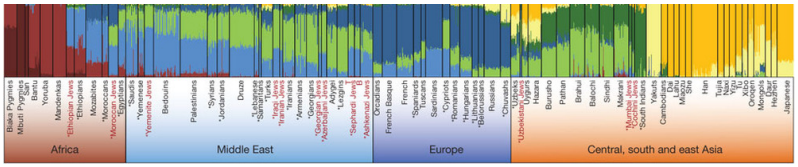


2D SFS, Fst and PBS for NGS

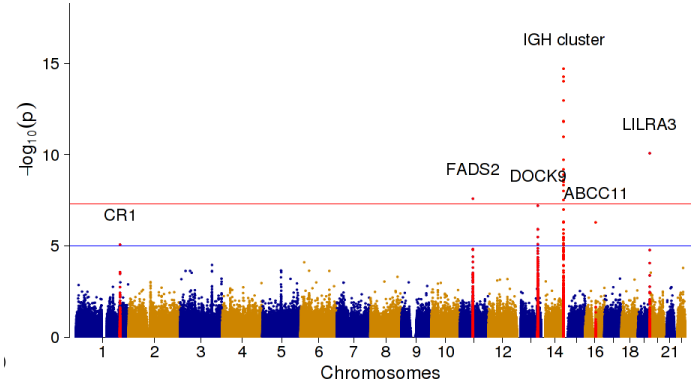


Afternoon - locus on low depth sequencing

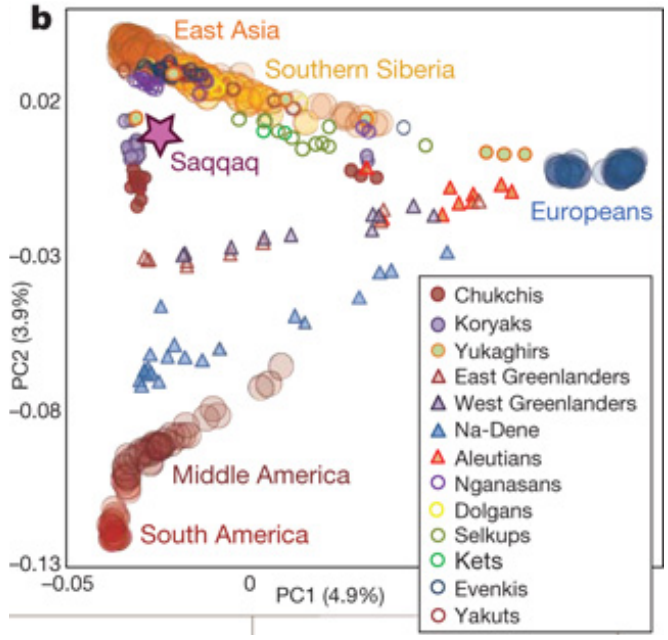
Admixture proportions



Individual allele frequencies (PCA)

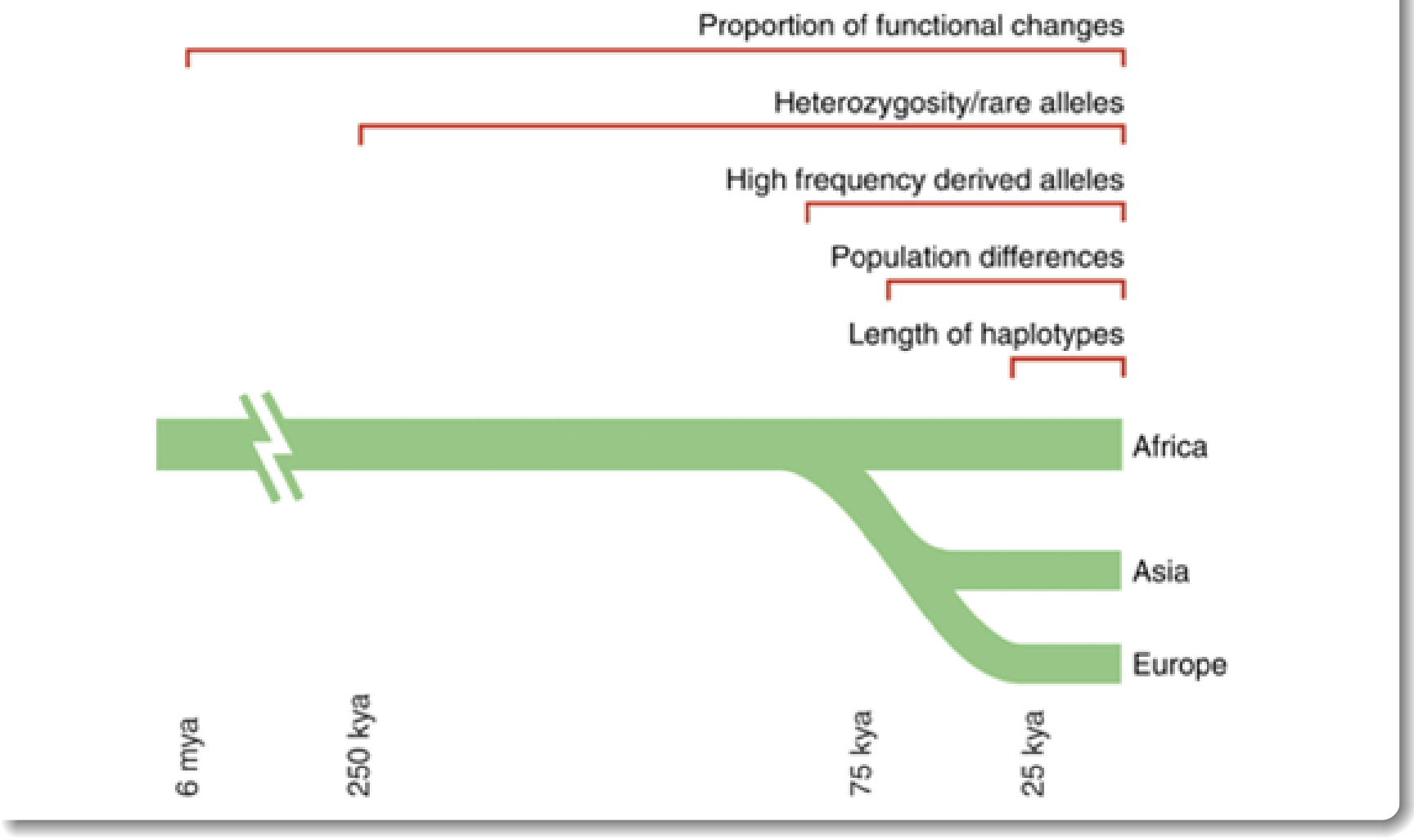


PCA



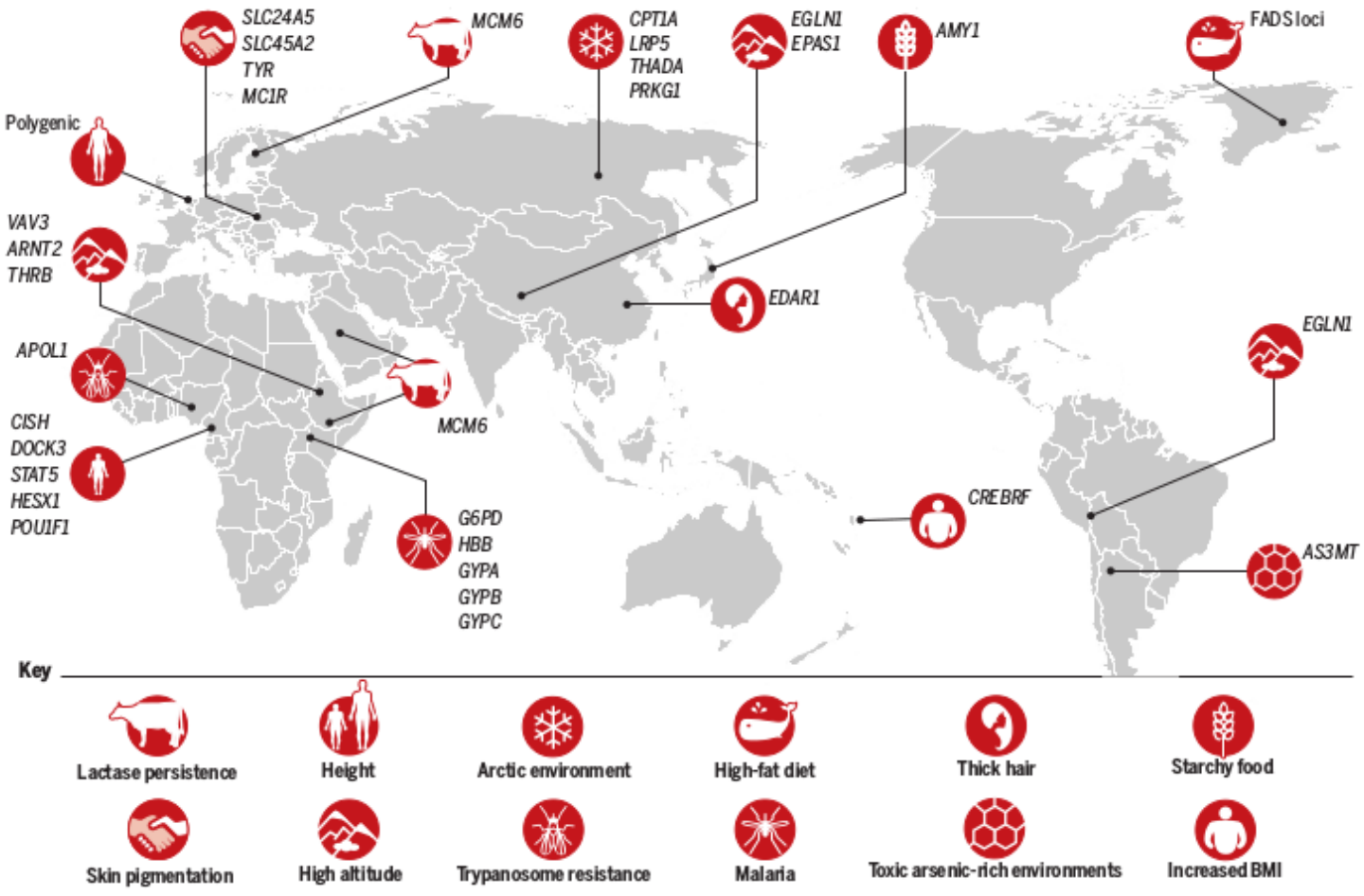
Recent selection

within species / using shared variation



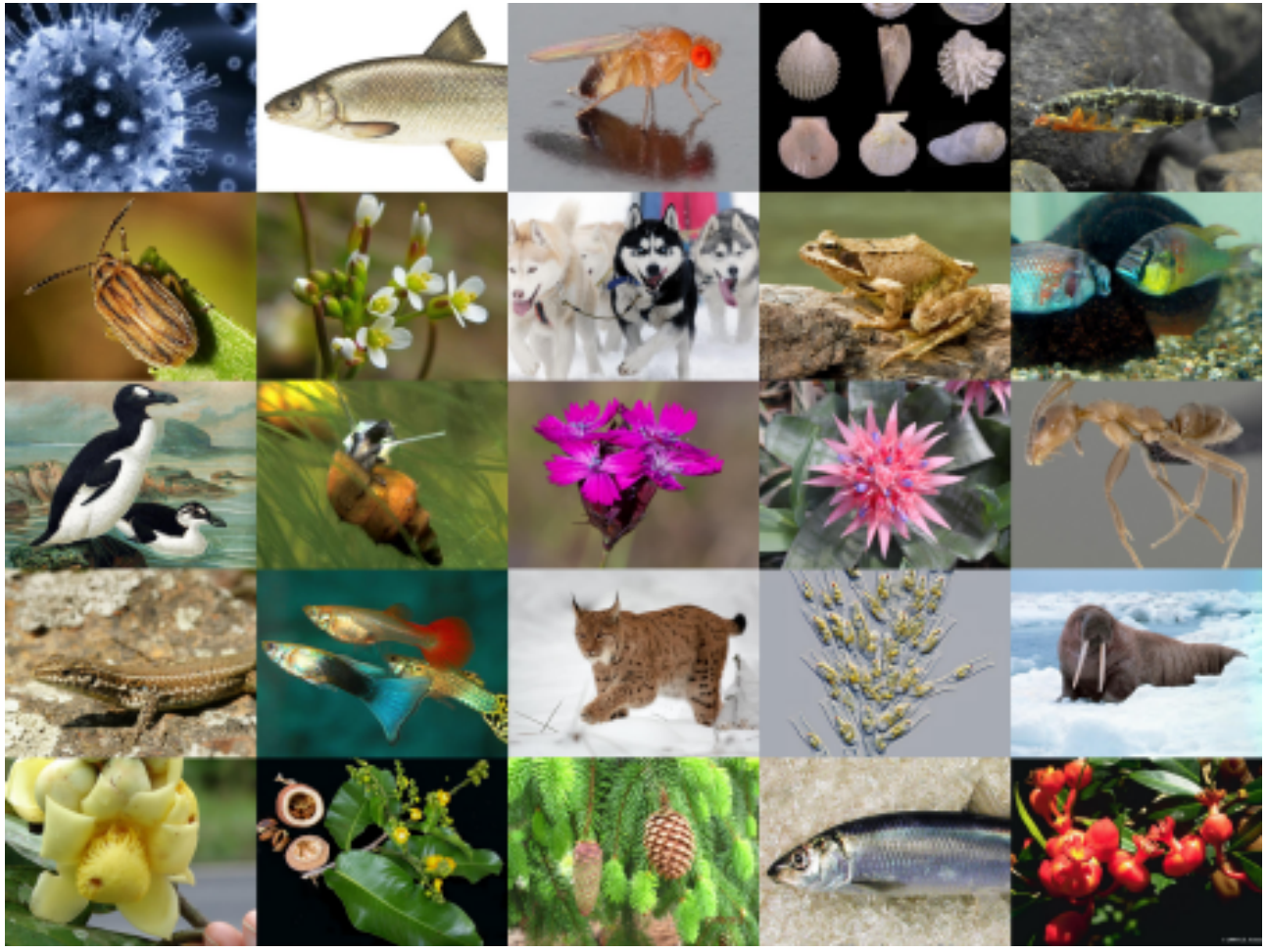
Sorry about the Human-centric talk

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

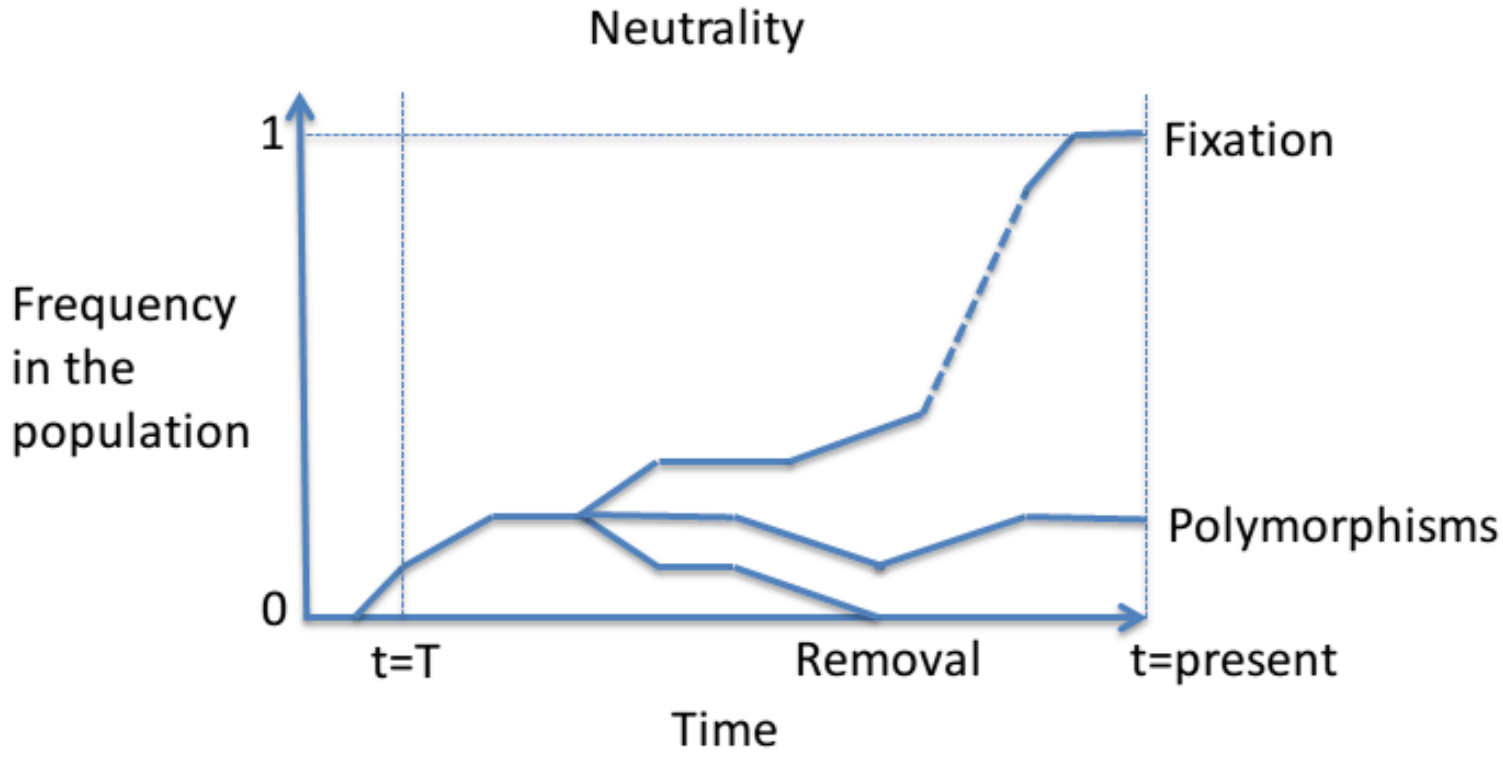
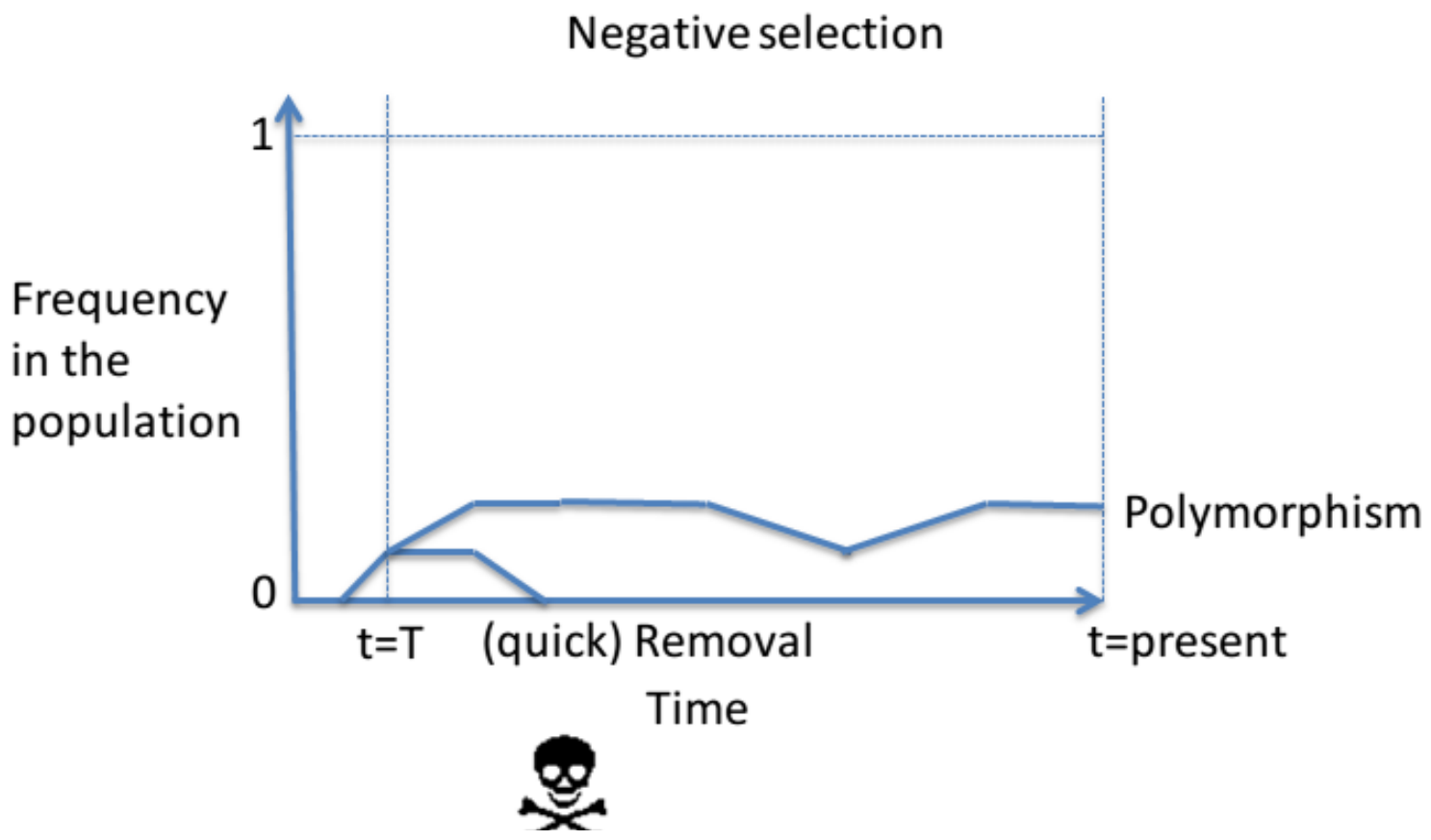


figure from Matteo Fumagalli

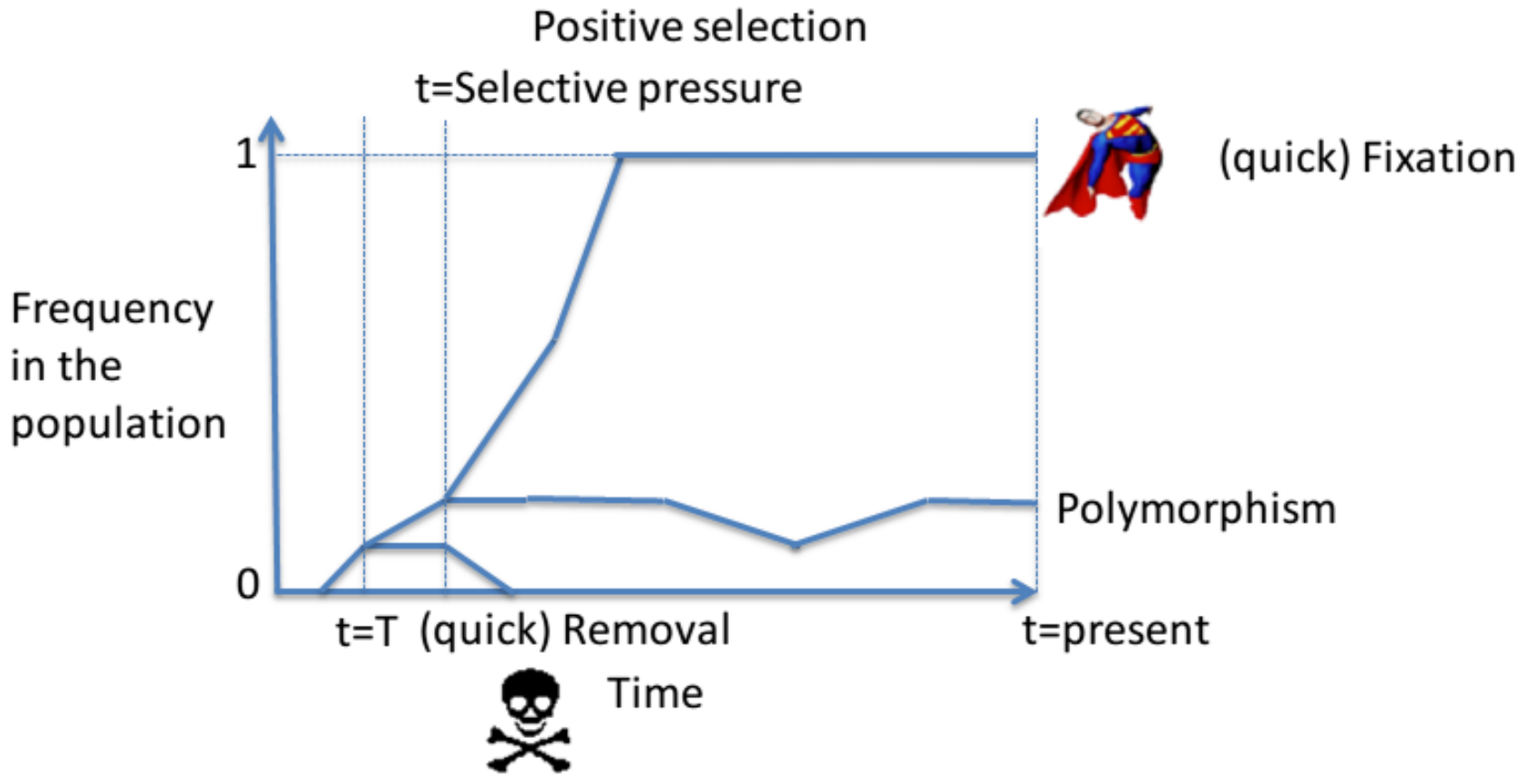
strong negative selection

alleles can be removed or be polymorphic



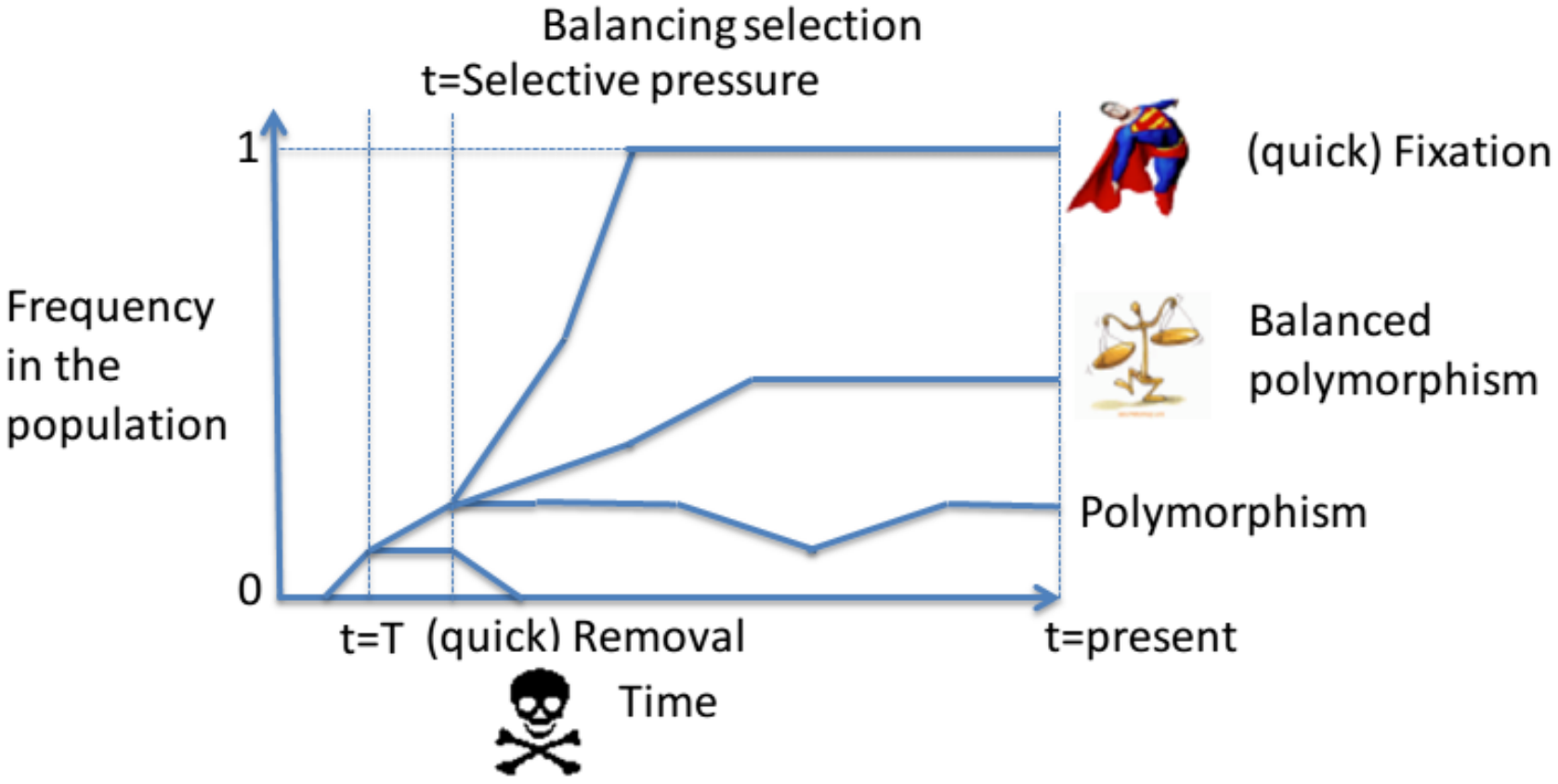
Strong positive selection

Alleles can be removed, polymorphic or fixed



Balancing selection

Alleles can be removed, polymorphic or fixed



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

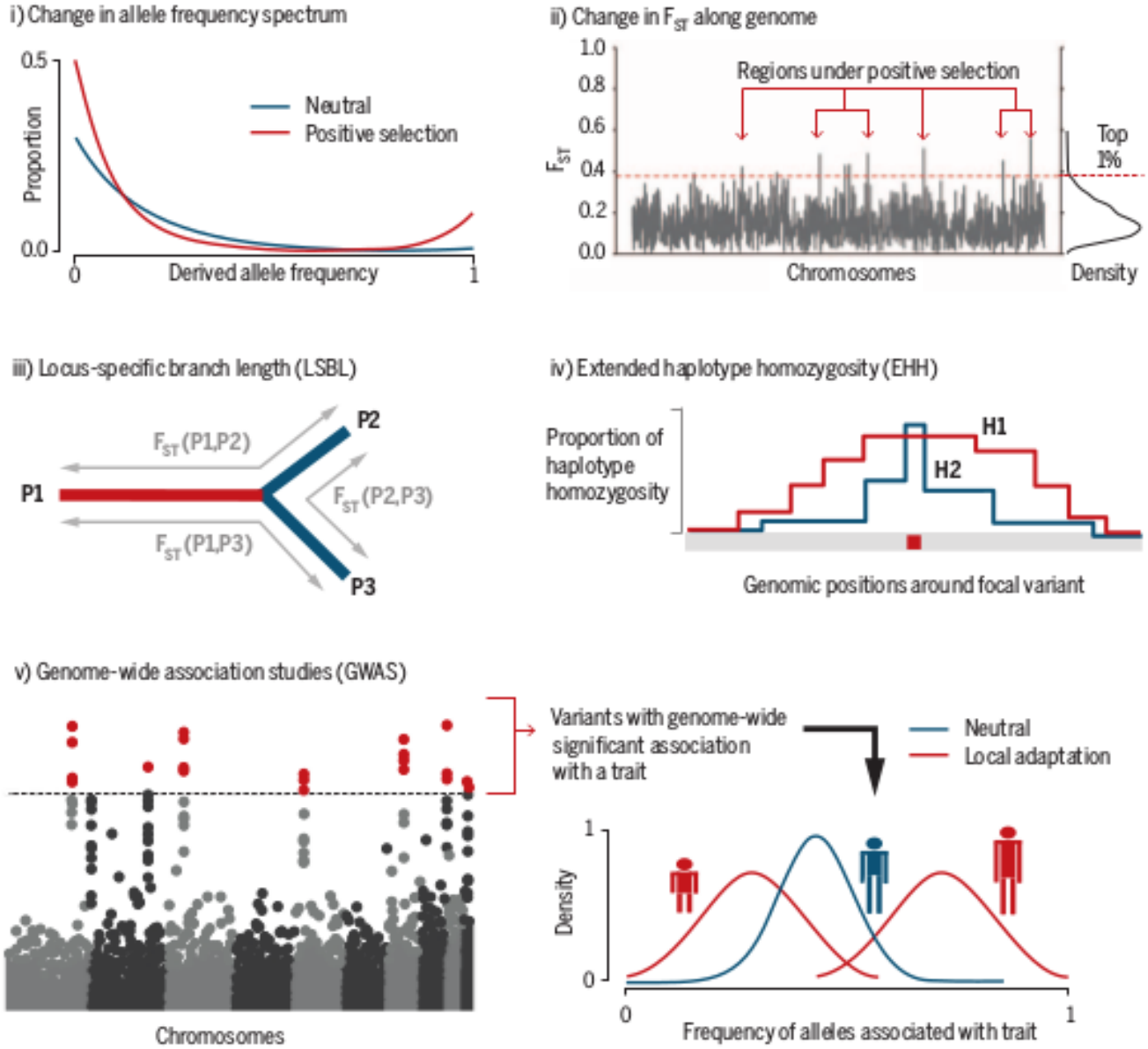
Options 3

selection experiments

External information

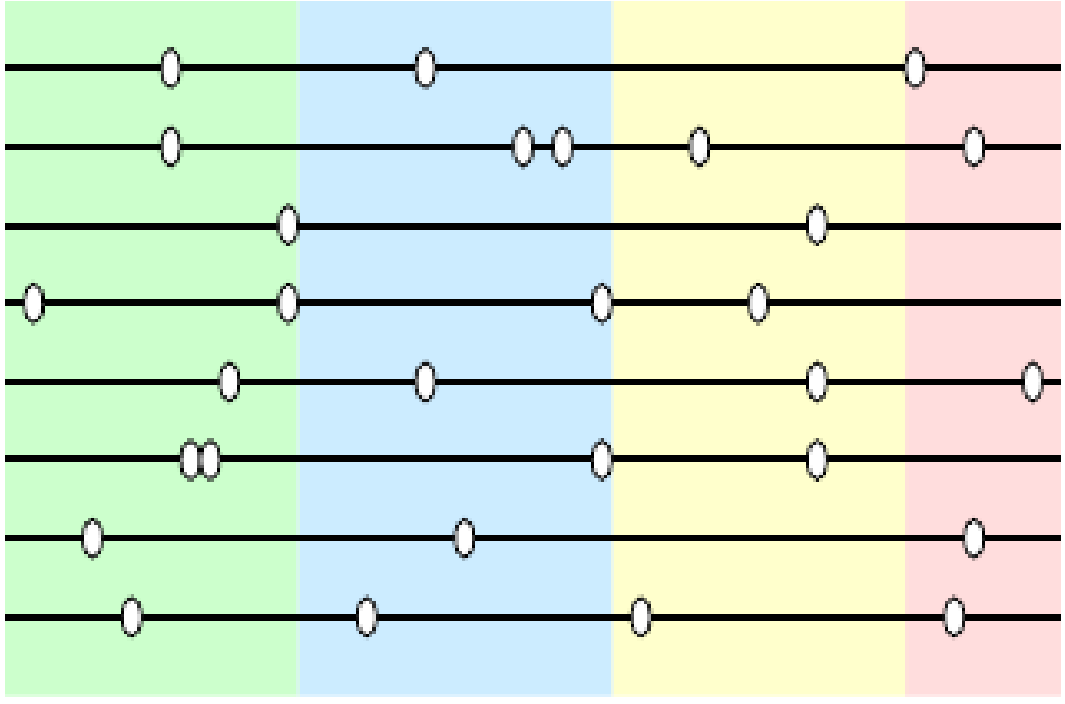
- Candidate genes/biological knowledge
- Functional categories
- Association to phenotypes

Common methods used to detect selection



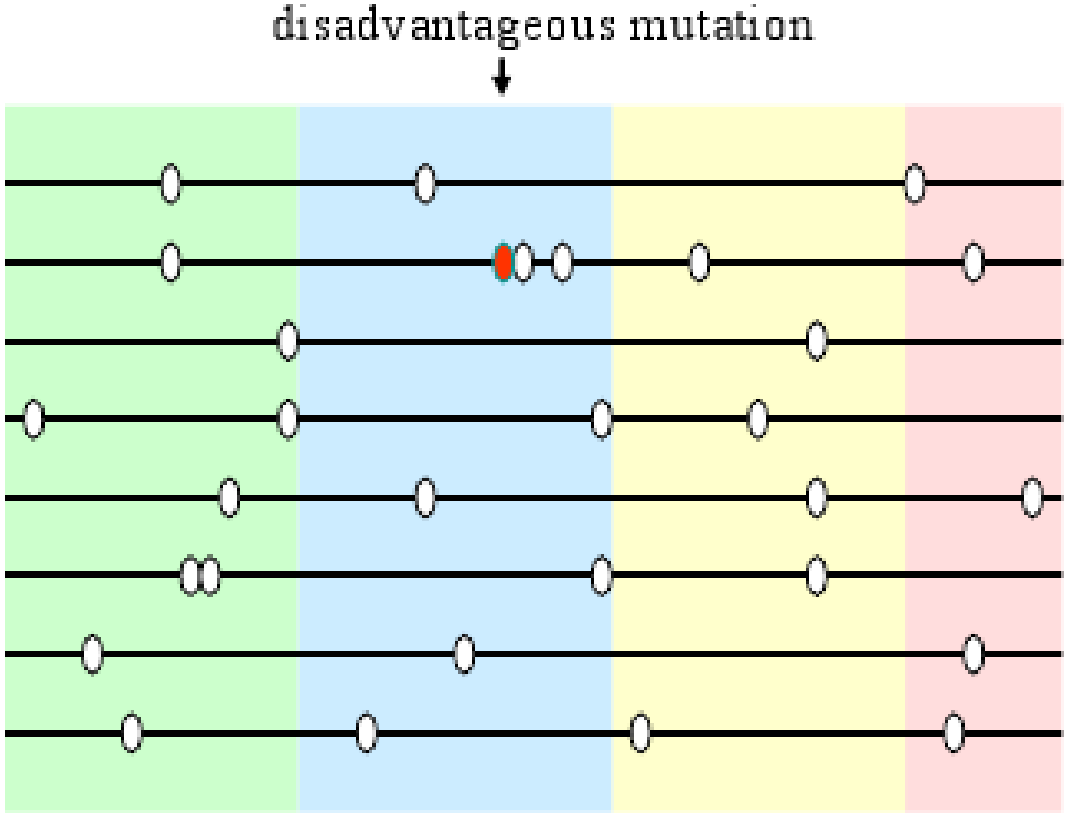
Signature of selection

- Neutral locus
- Lots of variability



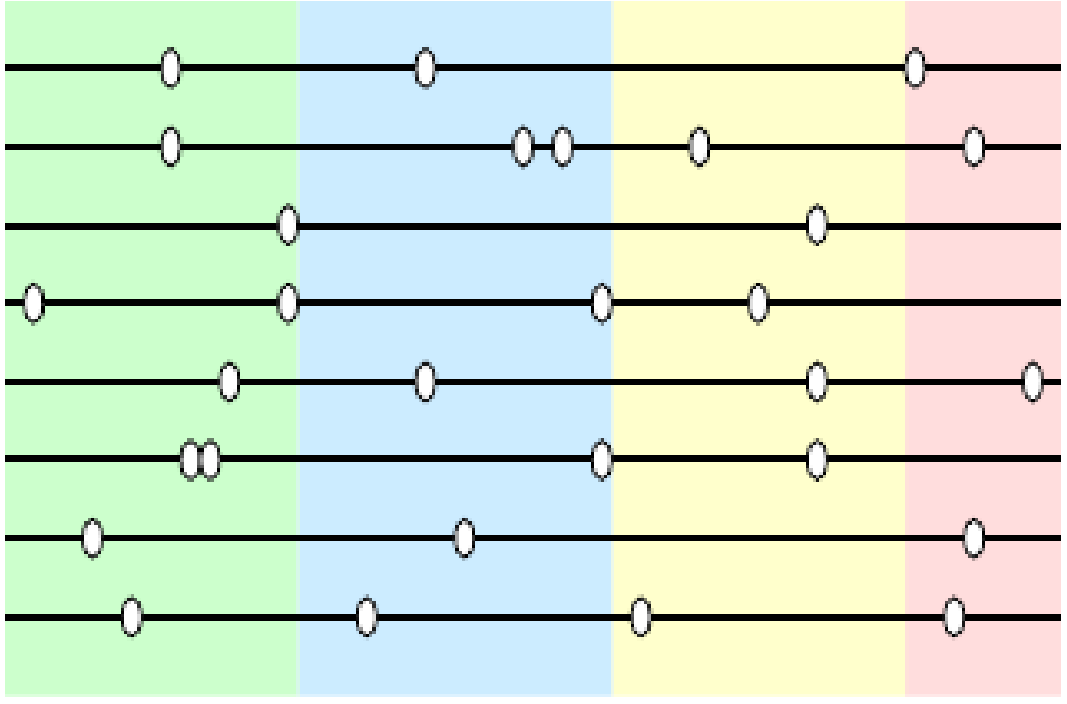
Signature of selection

- Mutation enters the population



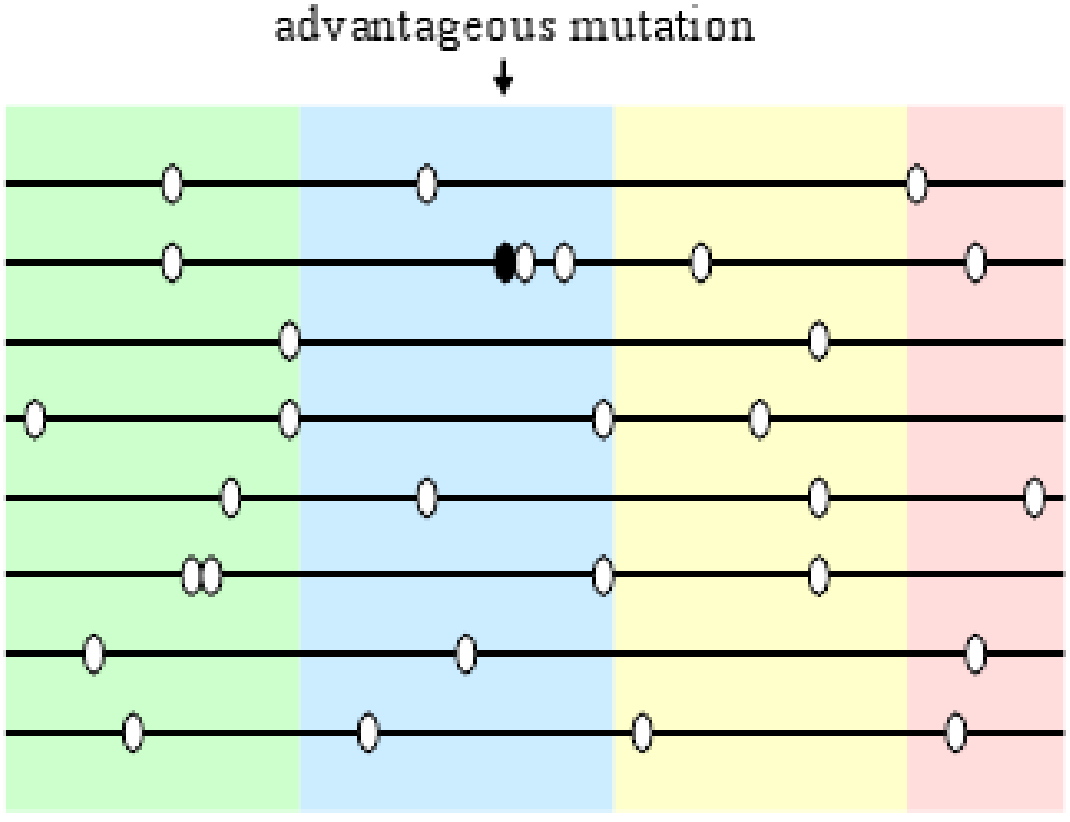
Signature of selection

- Negative selection removed the allele



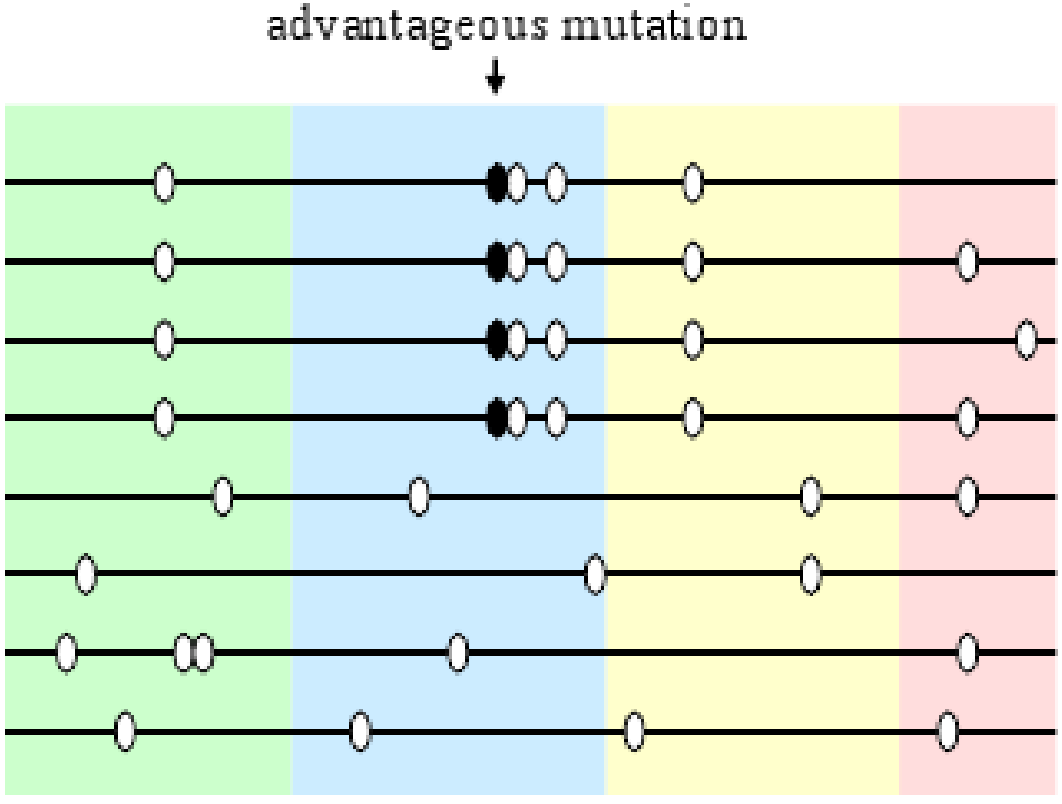
Signature of selection

- Mutation enters the population



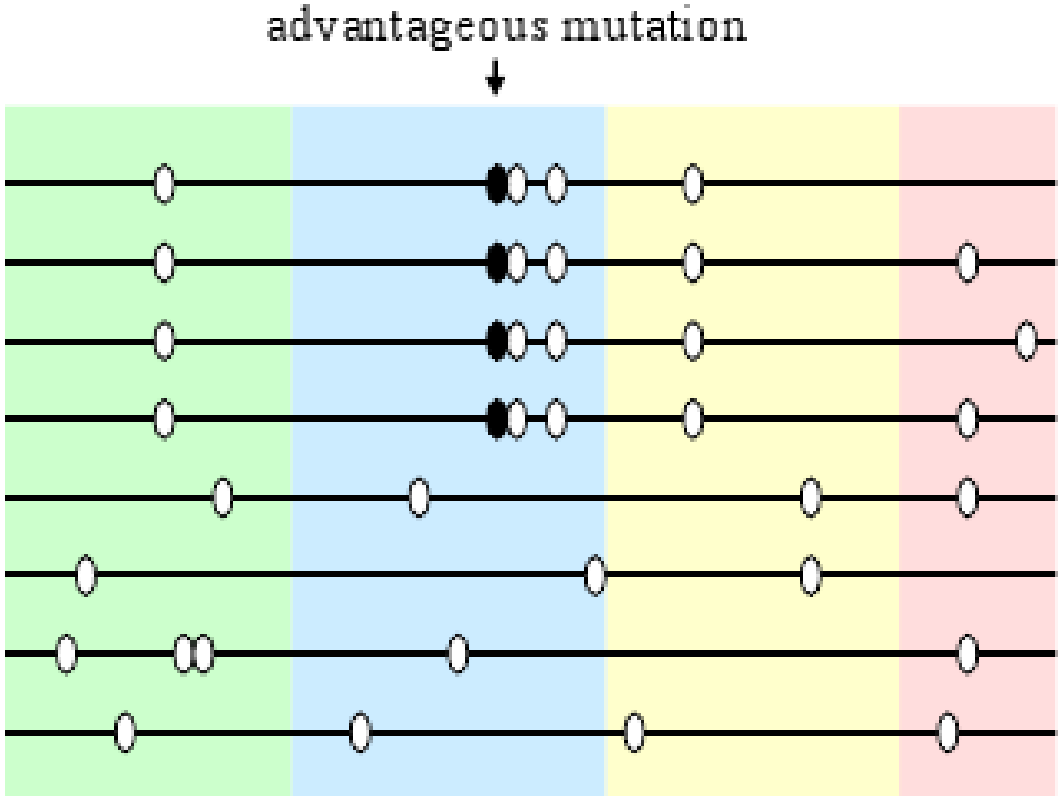
Signature of selection

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



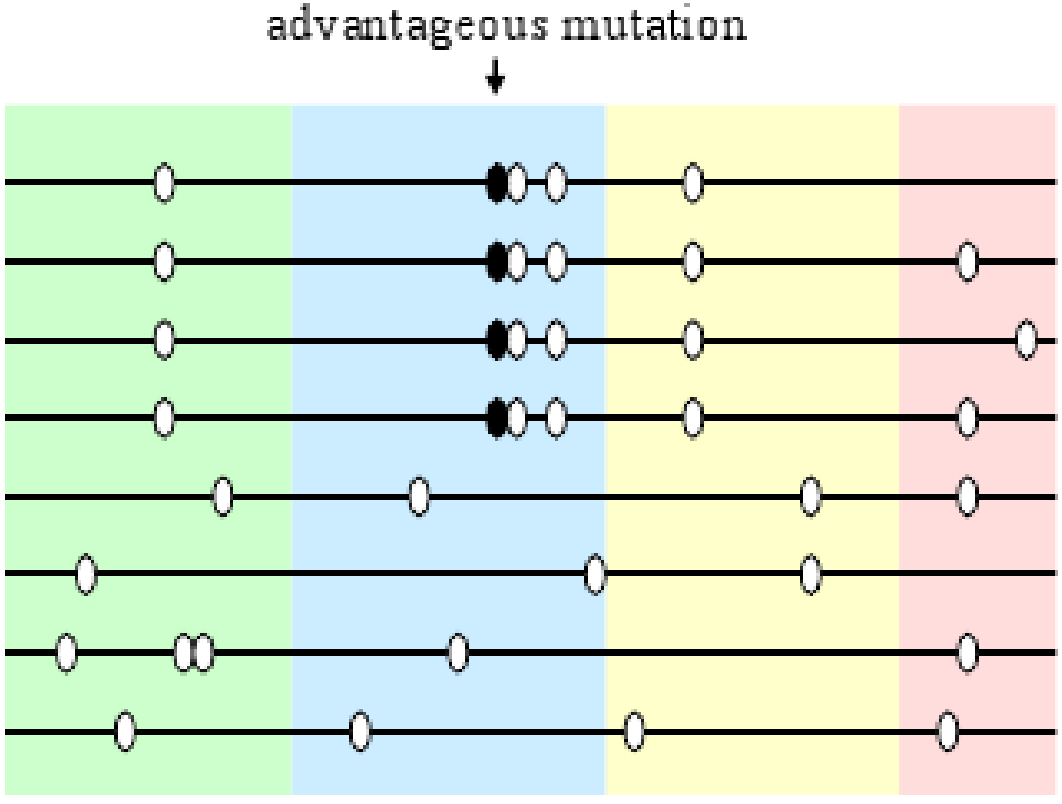
Signature of selection

- Increases LD
- Affects the variability



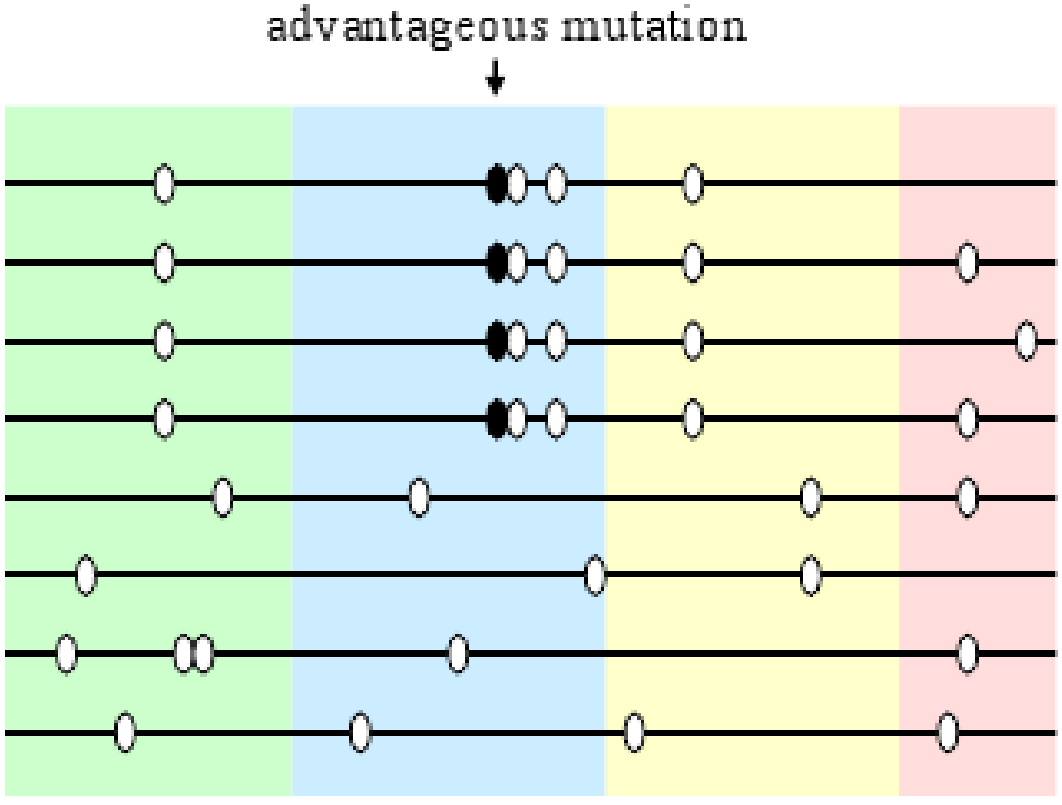
Signature of selection

- Increases haplotype similarity



Signature of selection

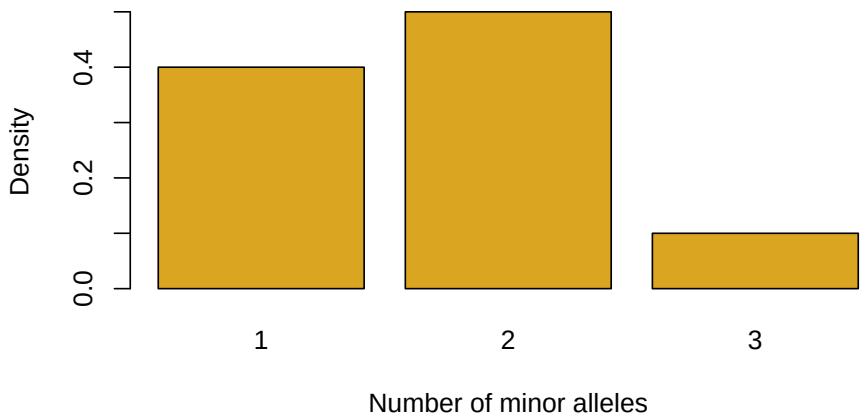
- Increases differences with other populations in the whole region



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

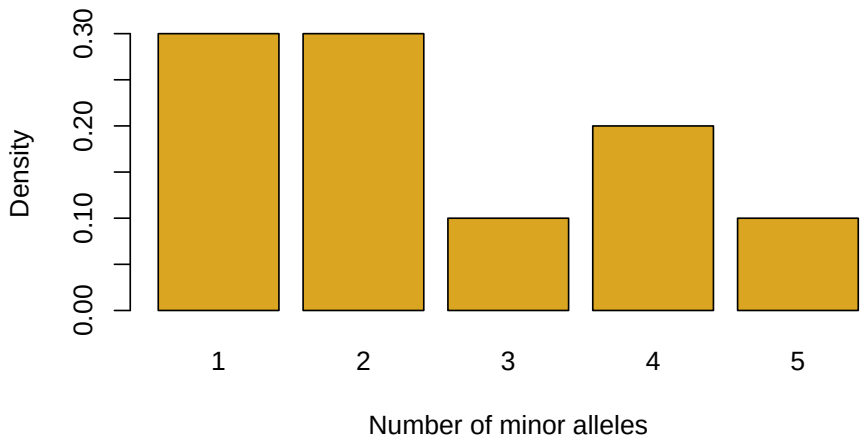
Number of minor alleles (folded) $\eta = (0.4, 0.5, 0.1)$



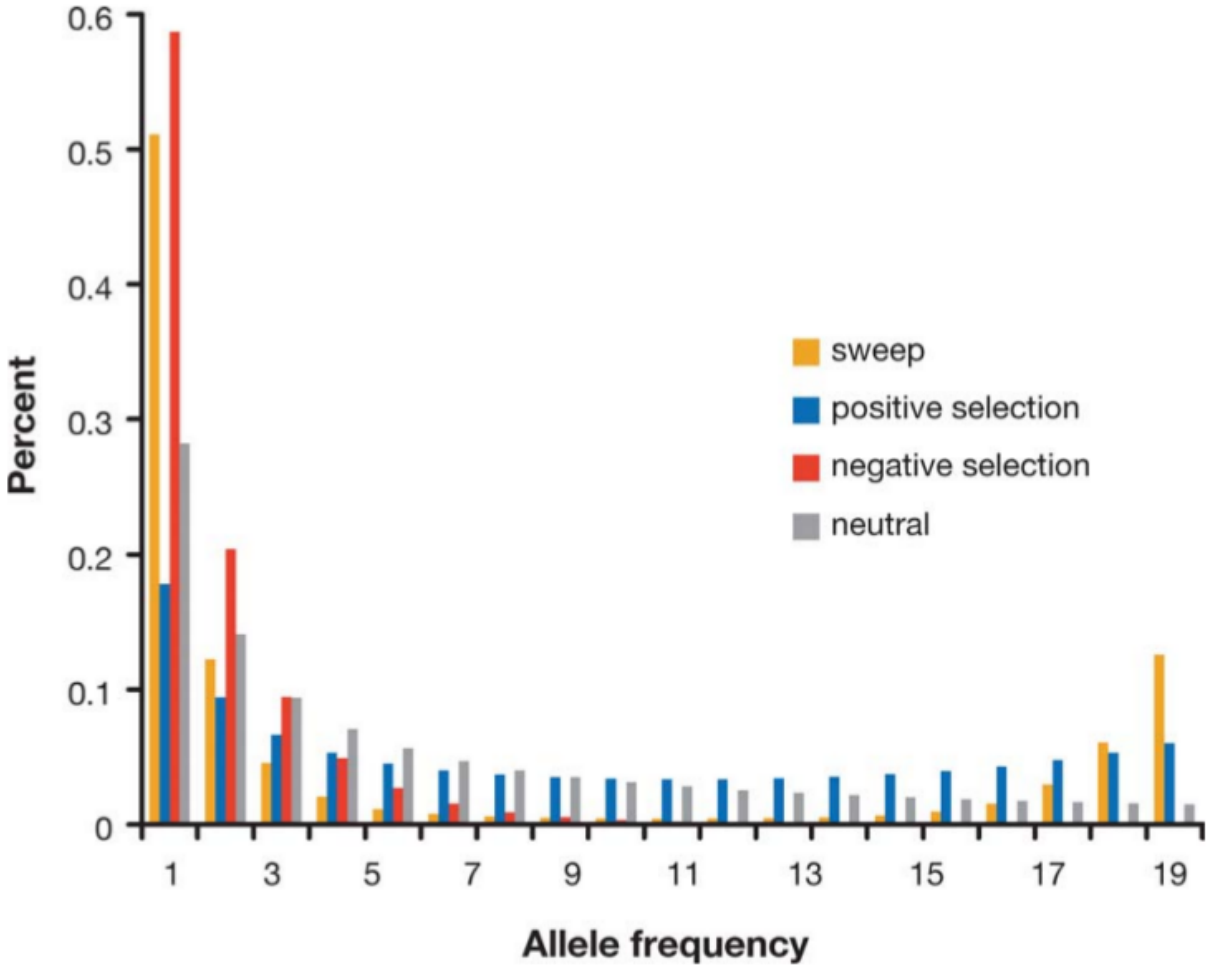
What is the site frequency spectrum

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

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

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

Tajima's D

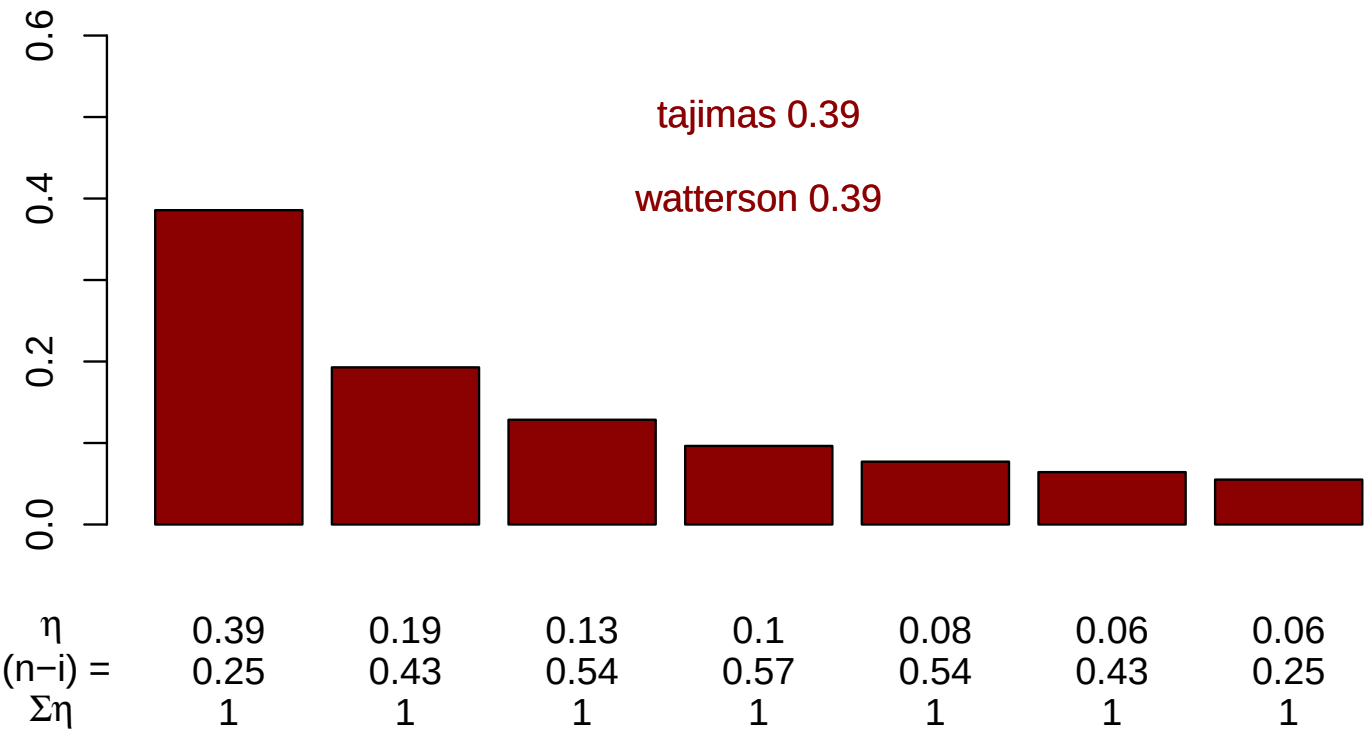
$D = \frac{\theta_T - \theta_W}{\sqrt{Var(\theta_T - \theta_W)}}$ under a neutral model* $\theta_T = \theta_W$

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

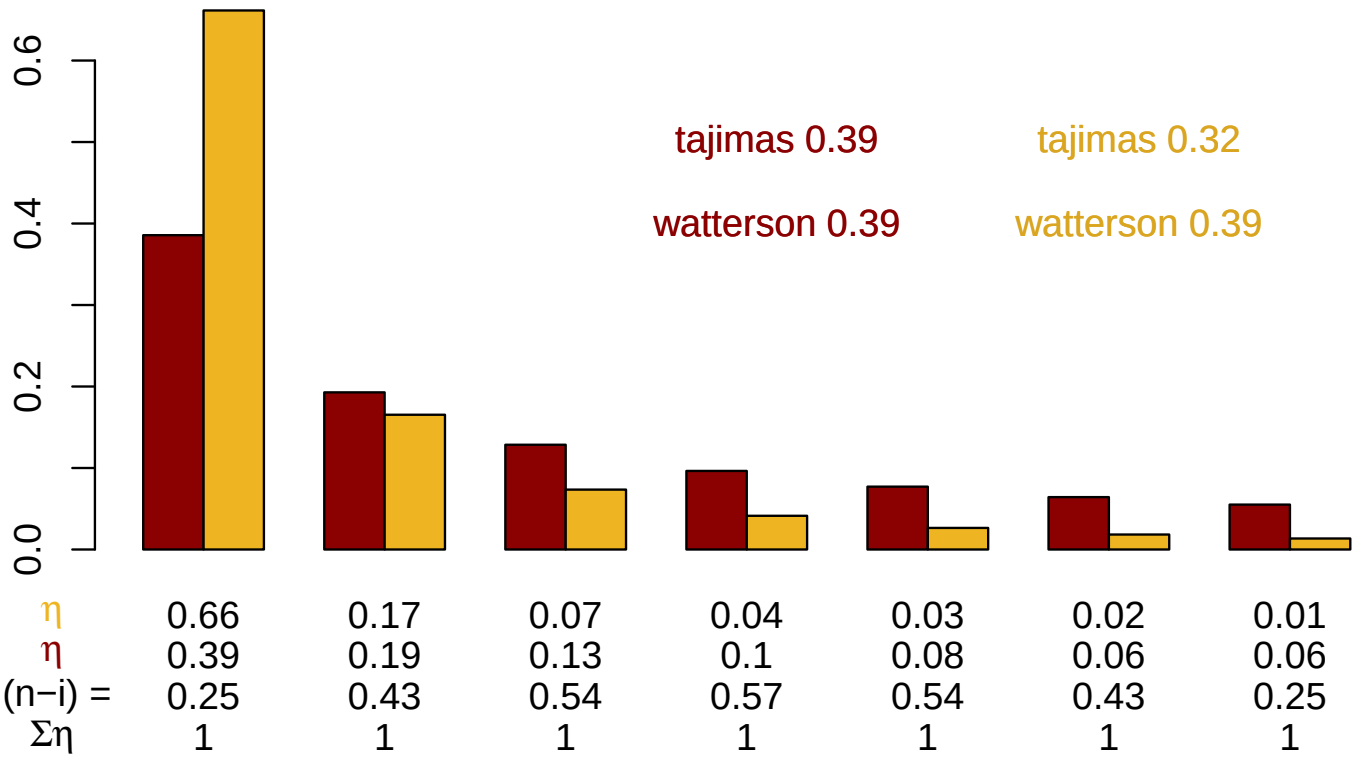


Theta are based on the frequency spectrum

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



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

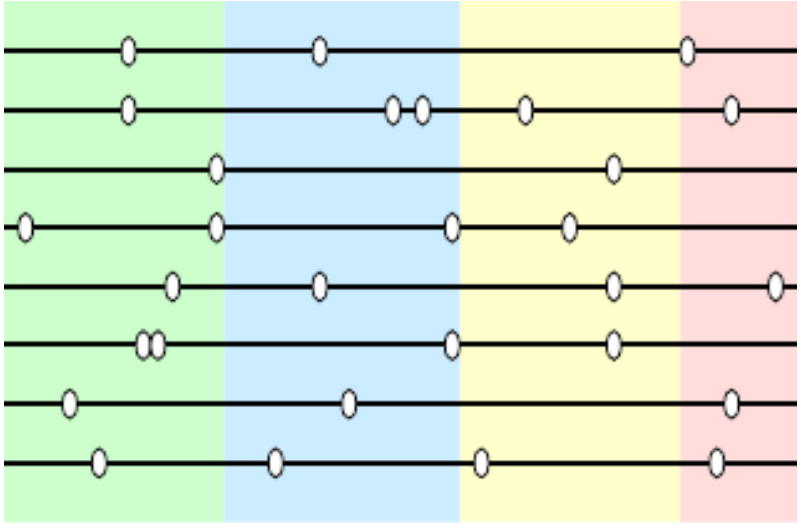
general $\hat{\theta} = \sum_{i=0}^{n} \alpha_i \eta_i$

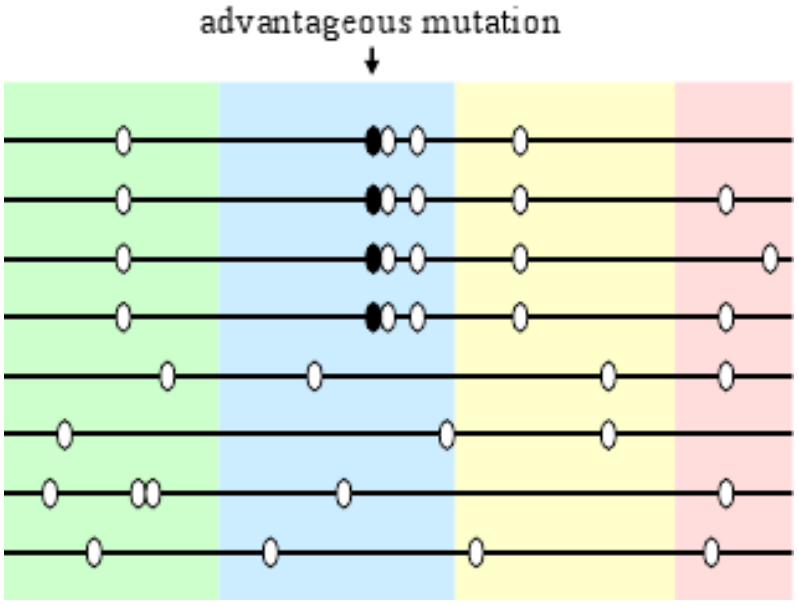
Test statistics

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

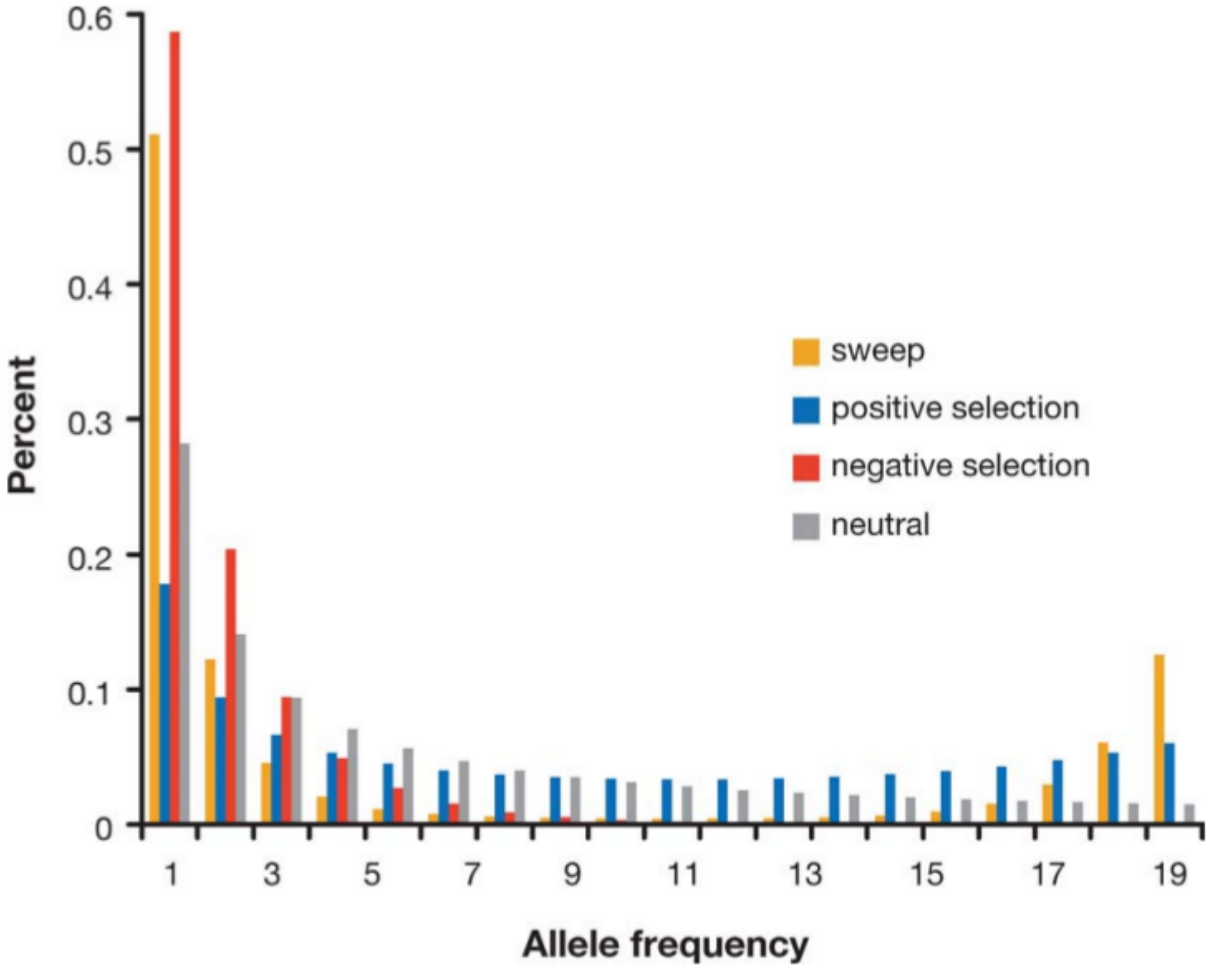
Difference weighting schemes for the SFS

Why does selection affect the SFS



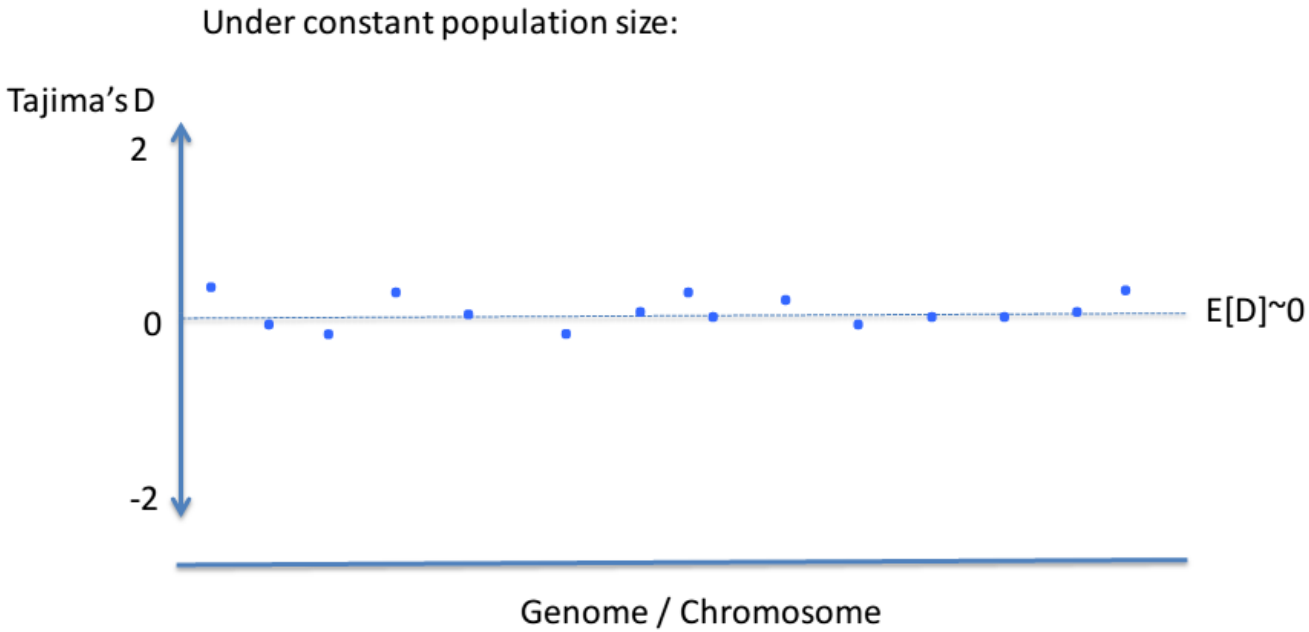


Frequency spectrum gives information about selection and demography



How to assess significance

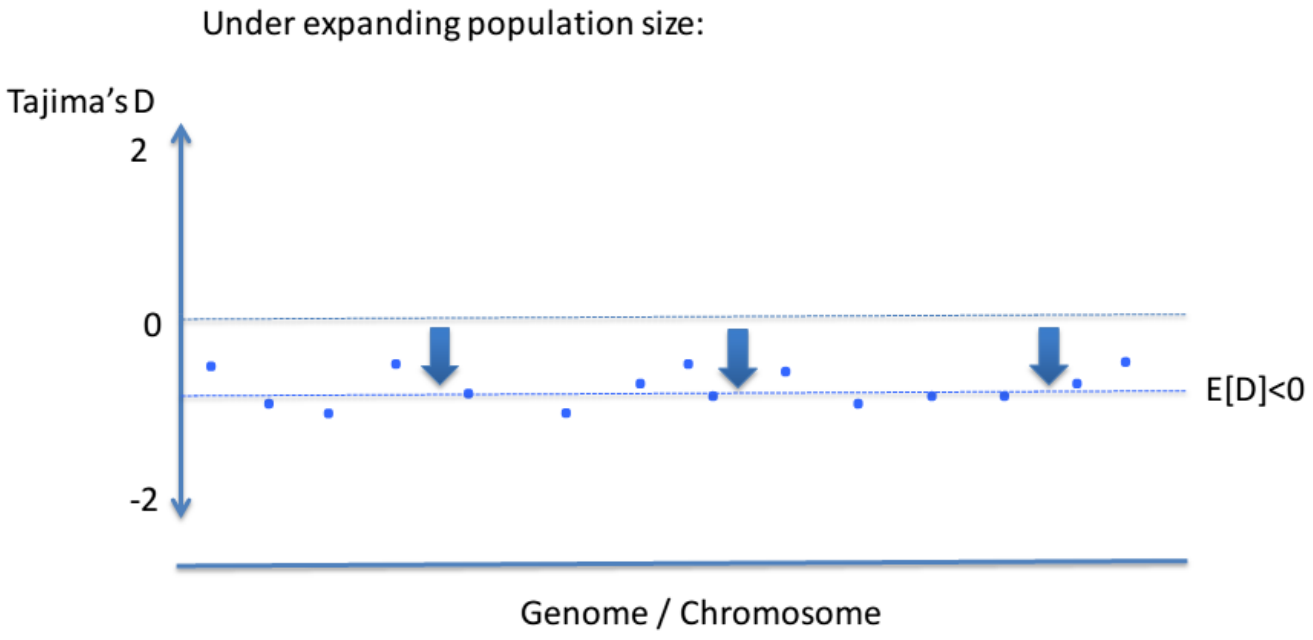
How to take neutral confounding factors into account?



slides stolen from Matteo Fumagalli

How to assess significance

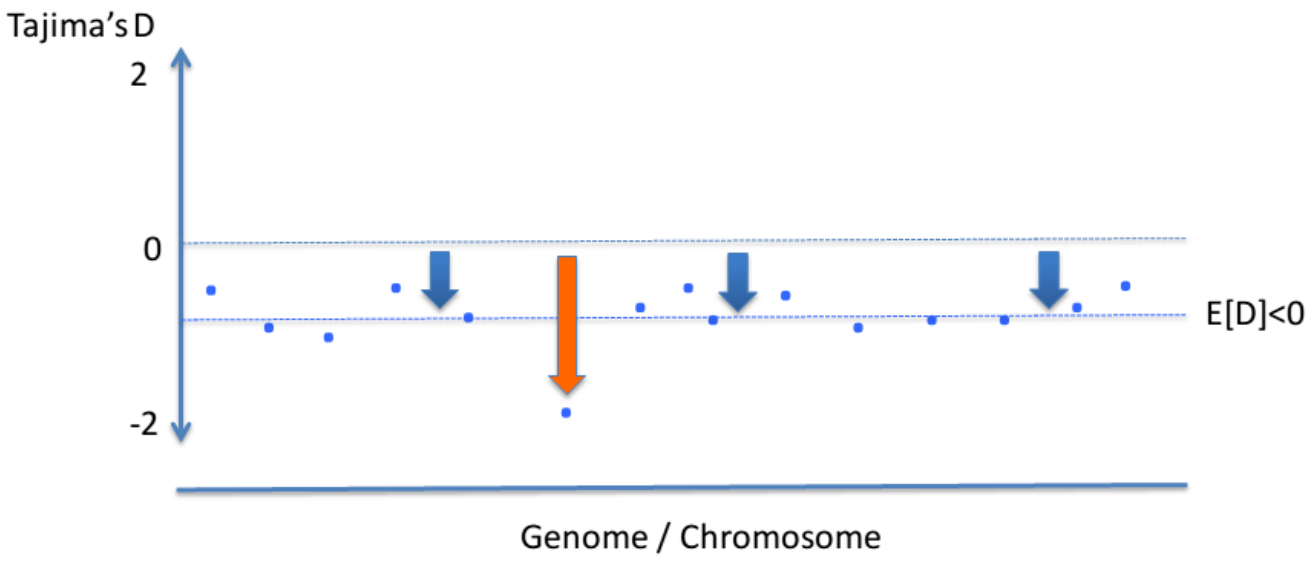
How to take neutral confounding factors into account?



How to assess significance

How to take neutral confounding factors into account?

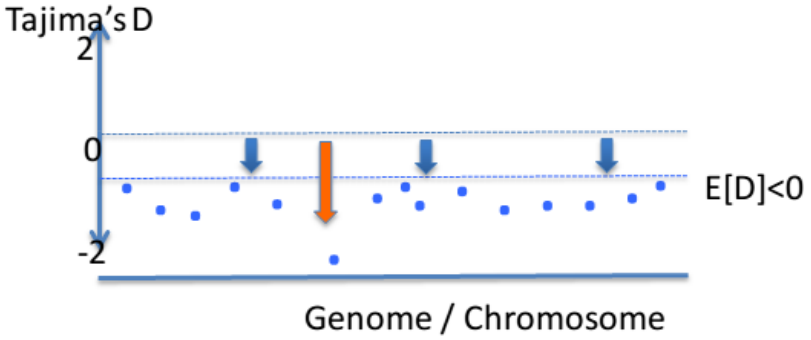
Under expanding population size and positive selection:



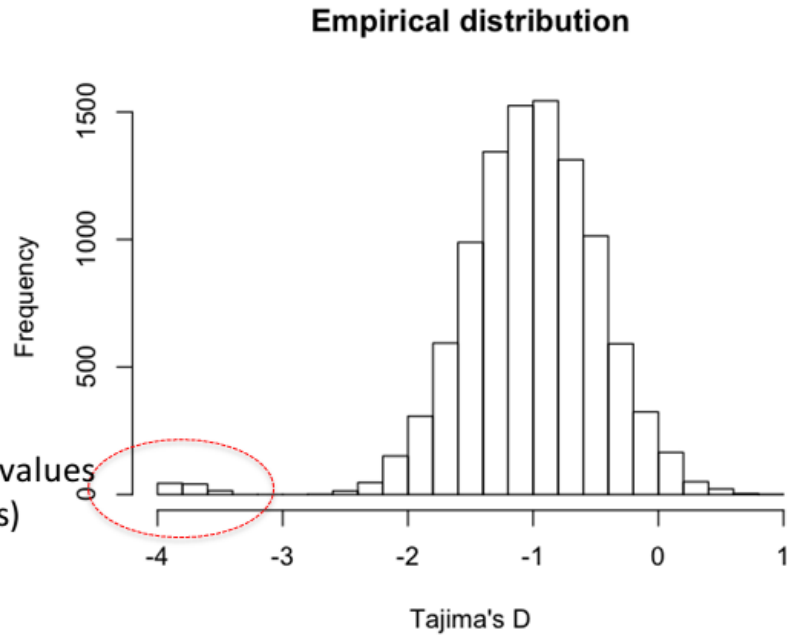
- Demography affects all loci equally, while selection changes local patterns

How to assess significance

Outlier approach

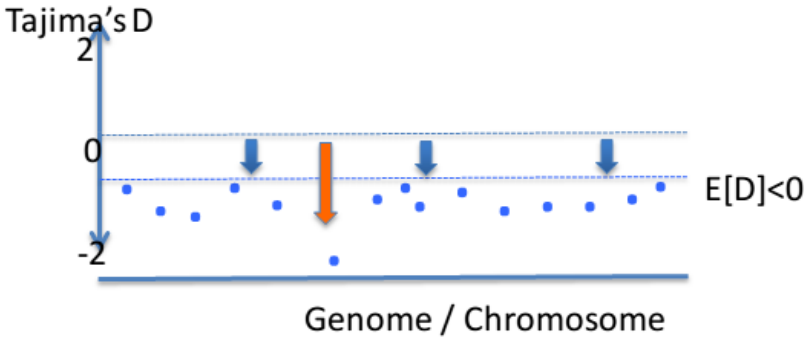


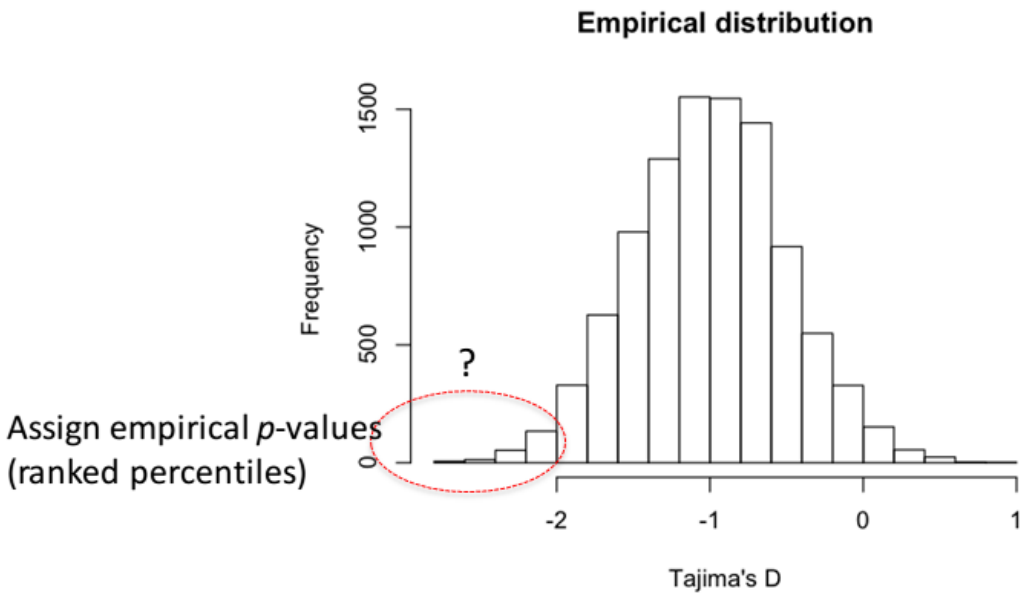
Assign empirical *p*-values (ranked percentiles)



How to assess significance

Outlier approach

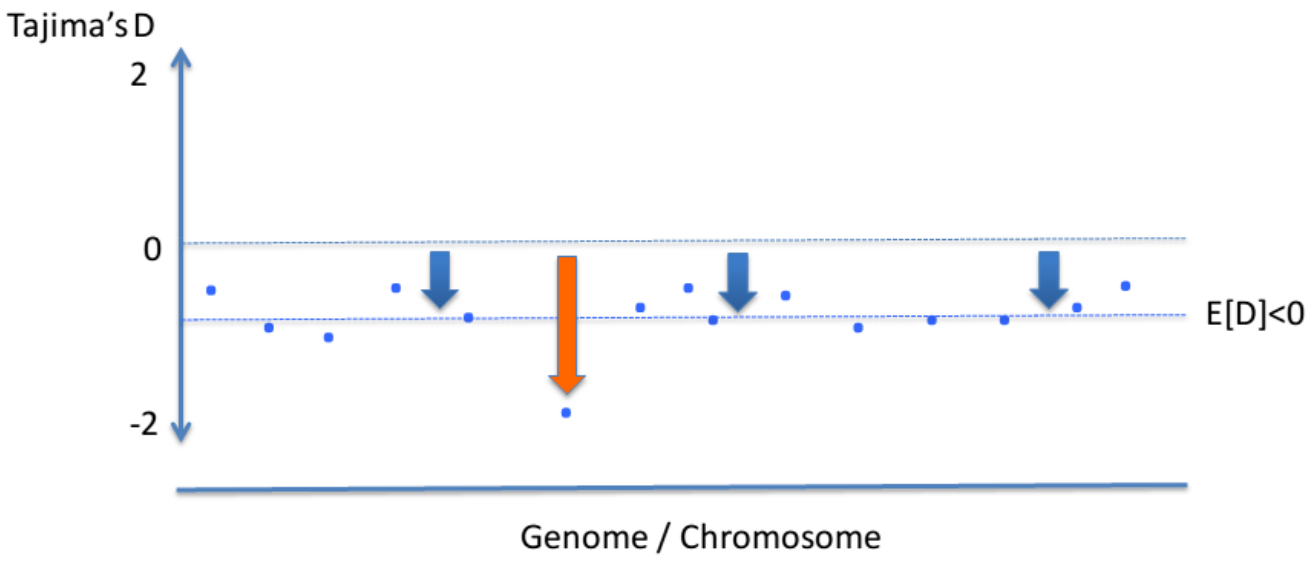




How to assess significance

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

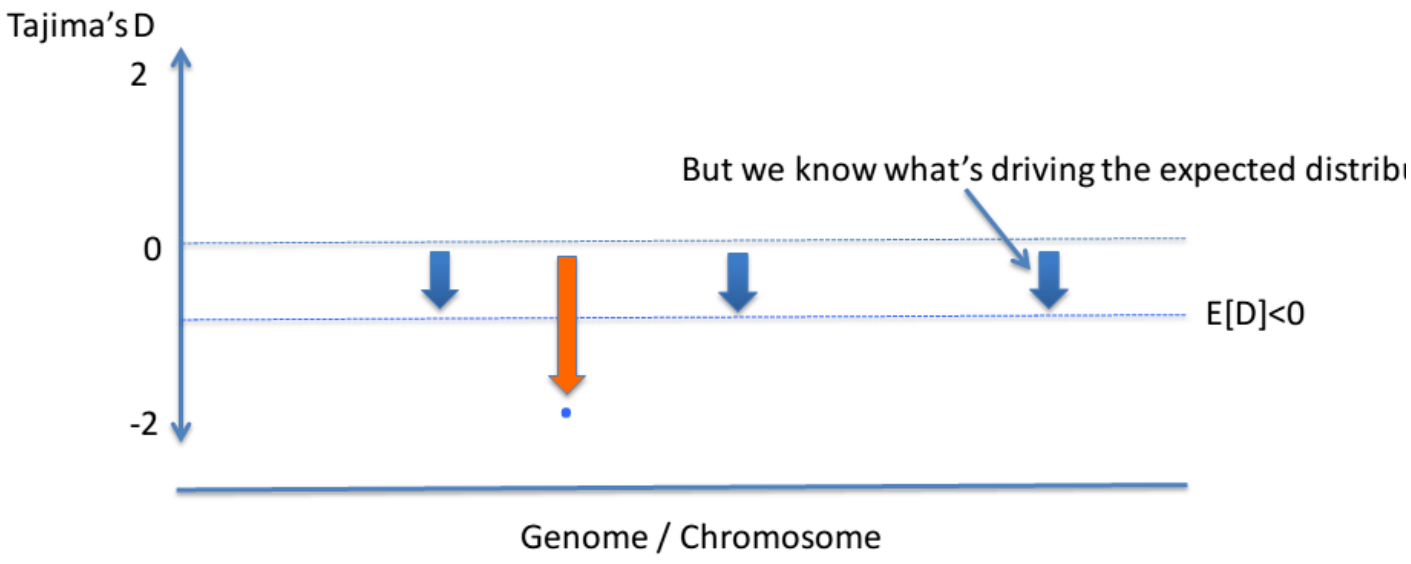


- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

How to assess significance

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

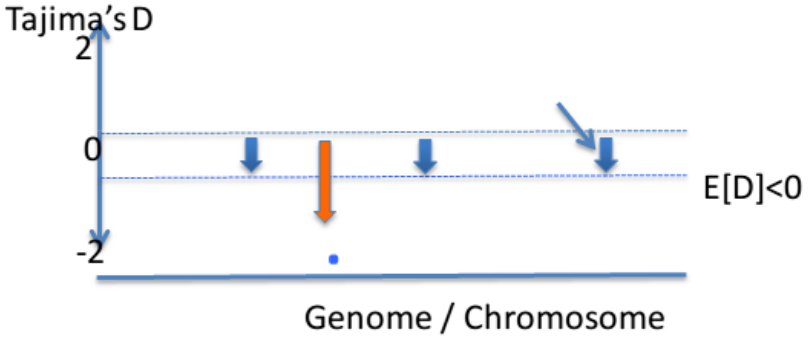


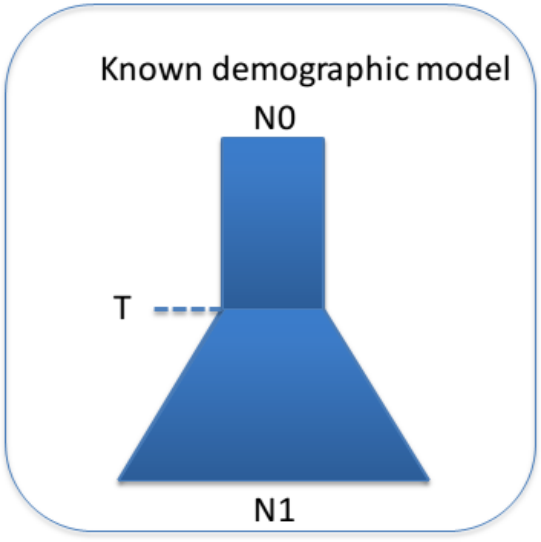
- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

How to assess significance

Simulations-based approach

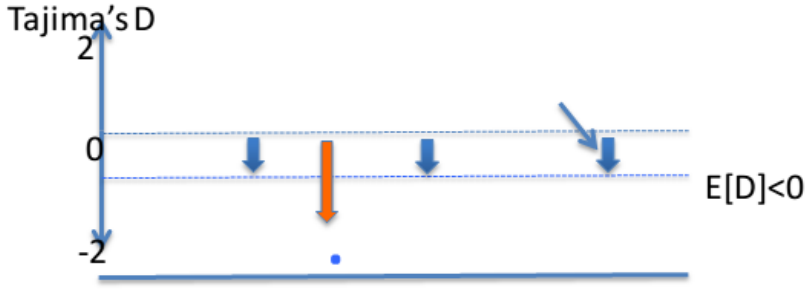
e.g. msms



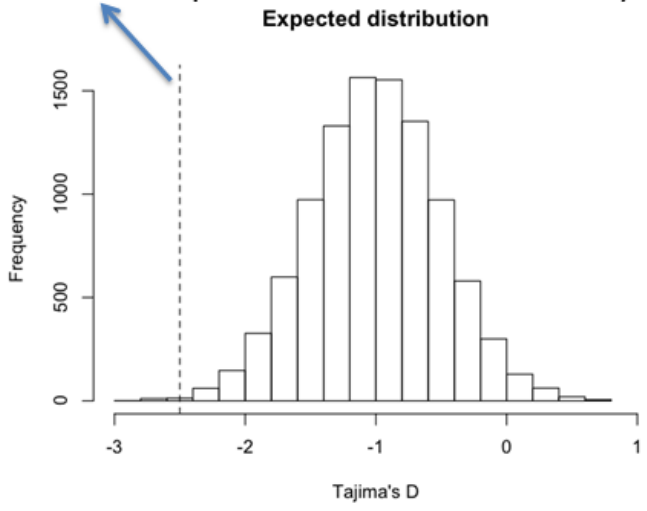


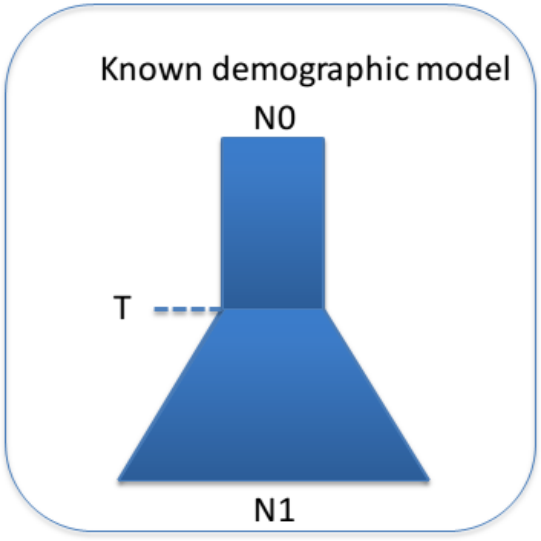
How to assess significance

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

`http://popgen.dk/albrecht/BAG2018/web/`

Graphics

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