Inferring natural selection from NGS

Detecting genomic regions under (positive) selection

Cindy G. Santander
22/08/2024

Thanks to

Anders Albrechtsen

and

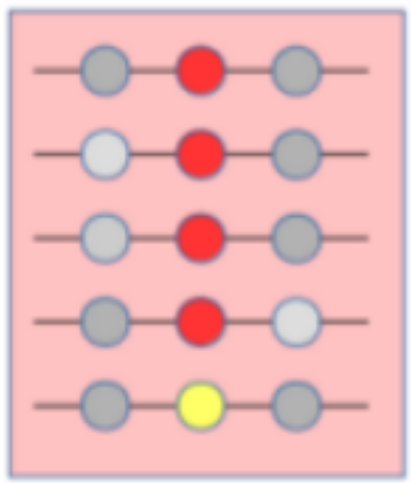
Matteo Fumagalli

for allowing me to shamelessly steal their slides

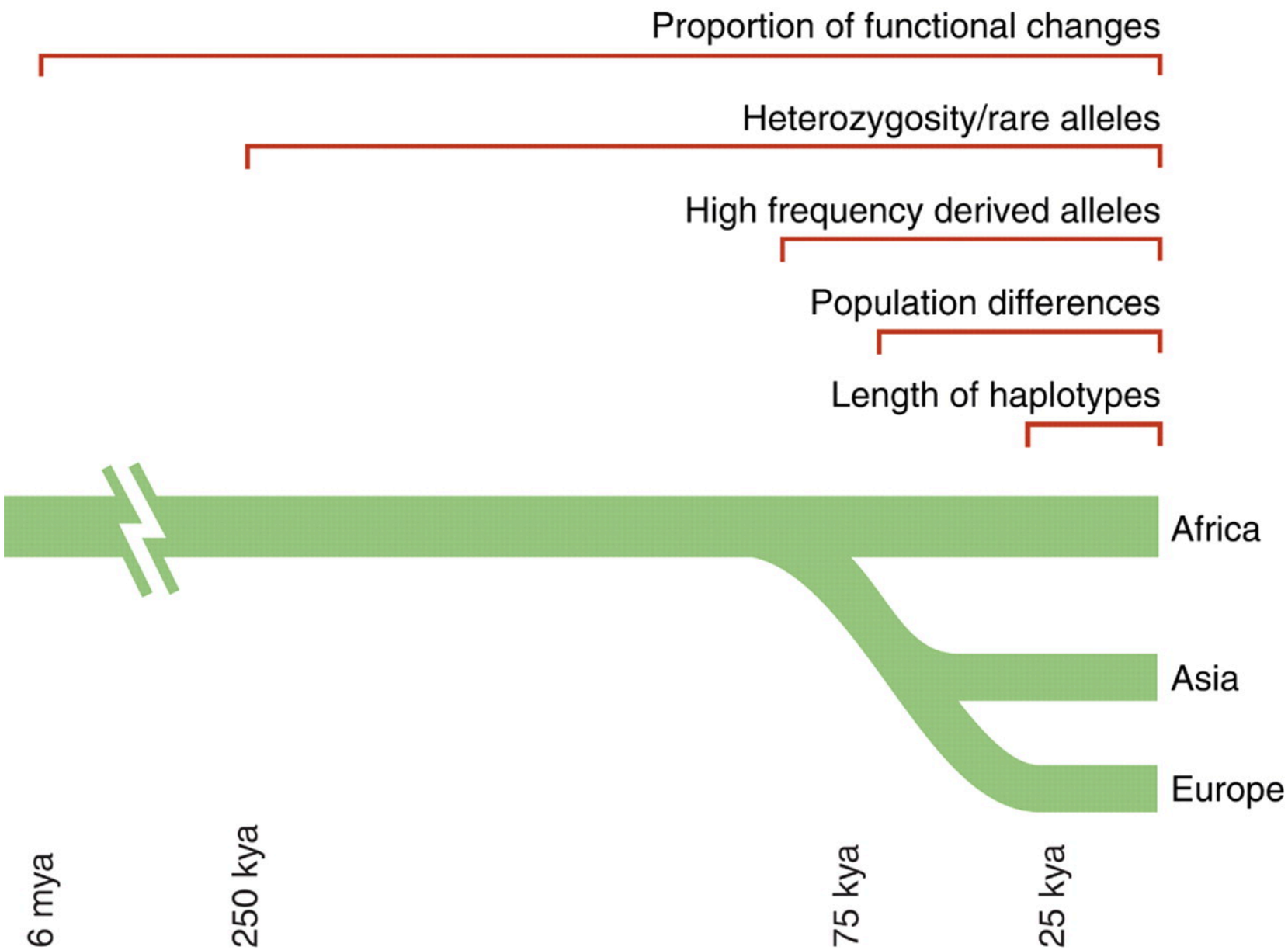
Demographic history

Whole genome sequencing

Natural selection



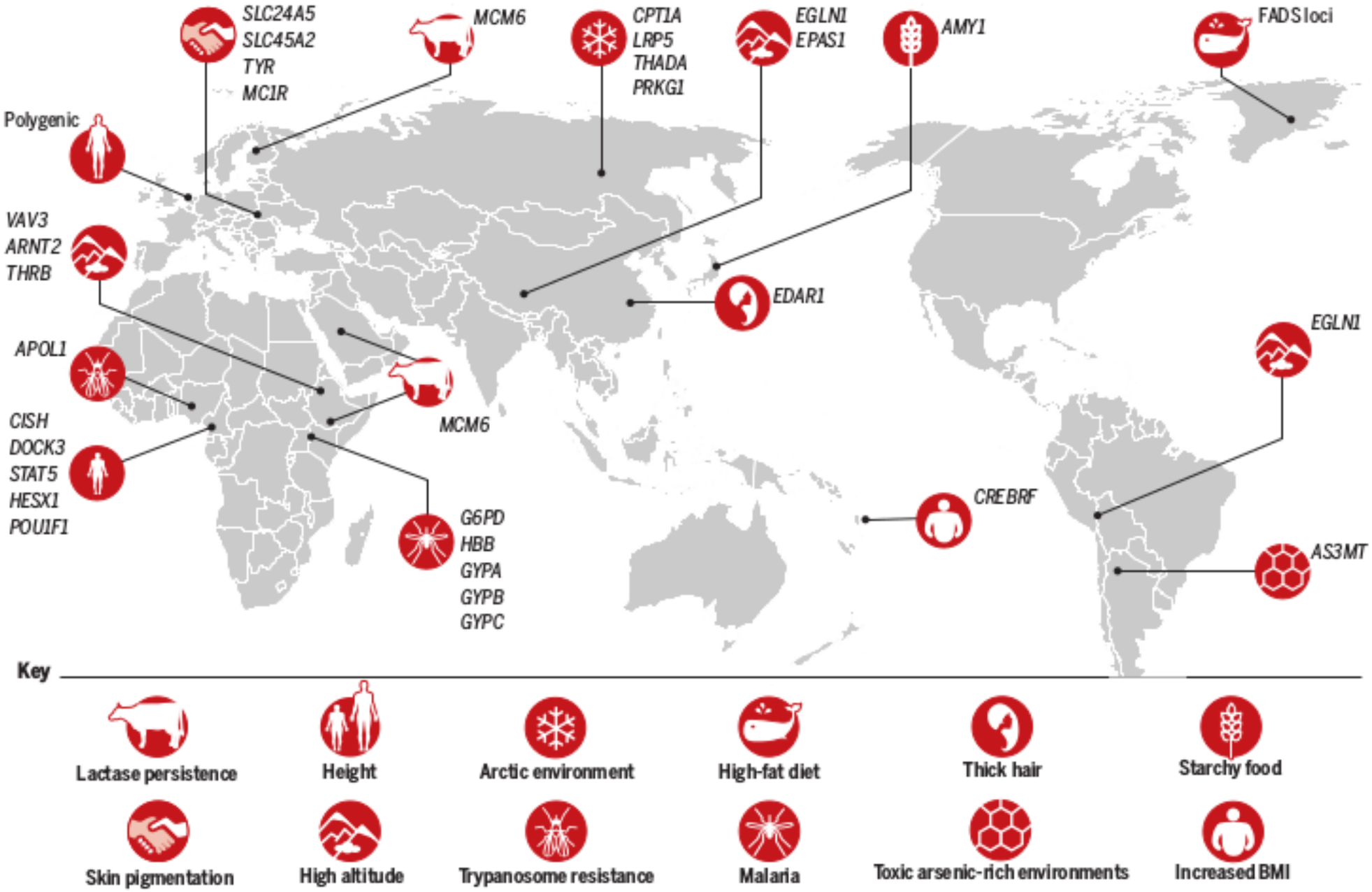
Detecting recent selection



Sabeti et al. 2006, Science

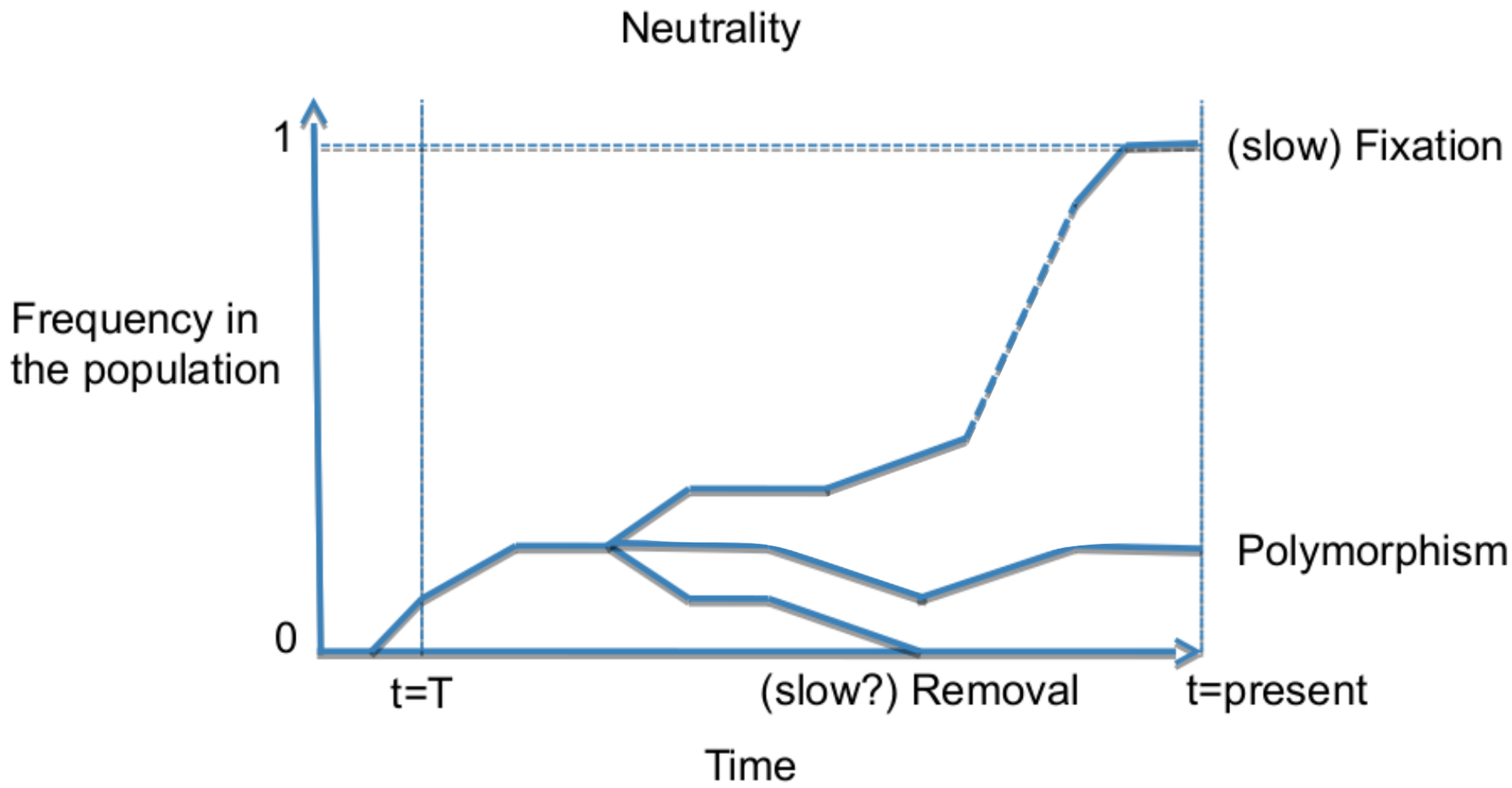
Good candidates for genes under recent selection

^ HUMAN



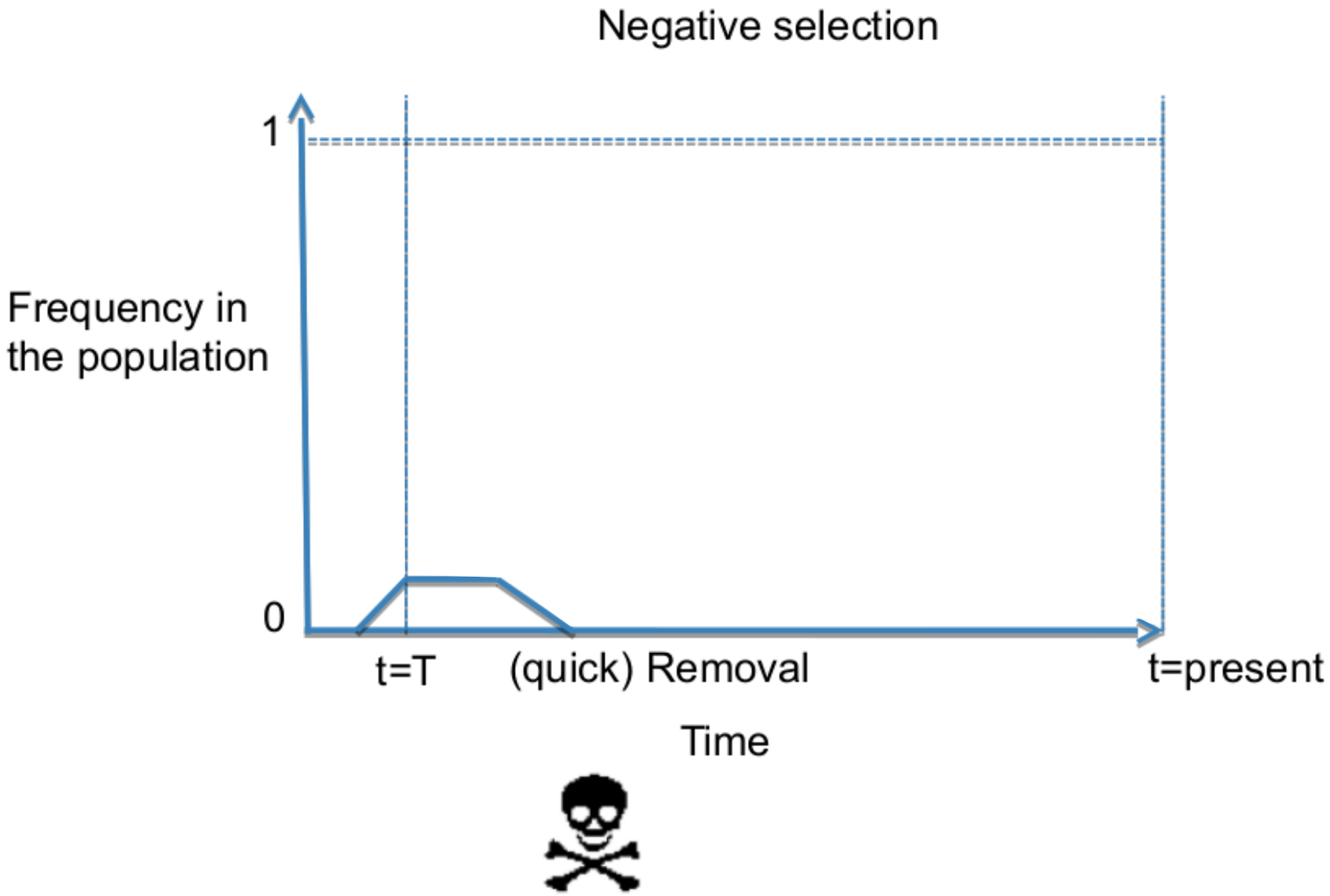
Allele Frequency Trajectory

Neutrality



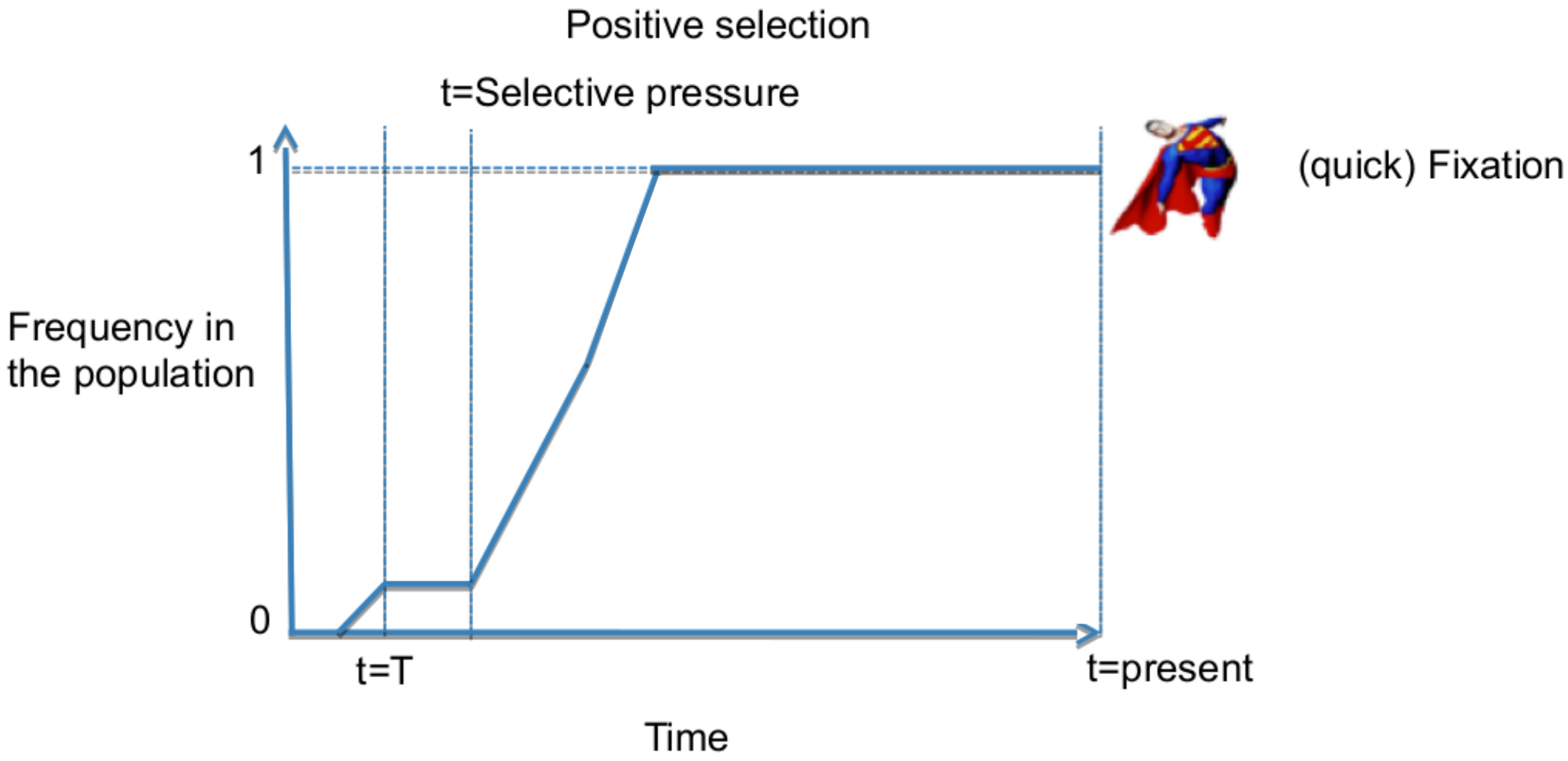
Allele Frequency Trajectory

Negative



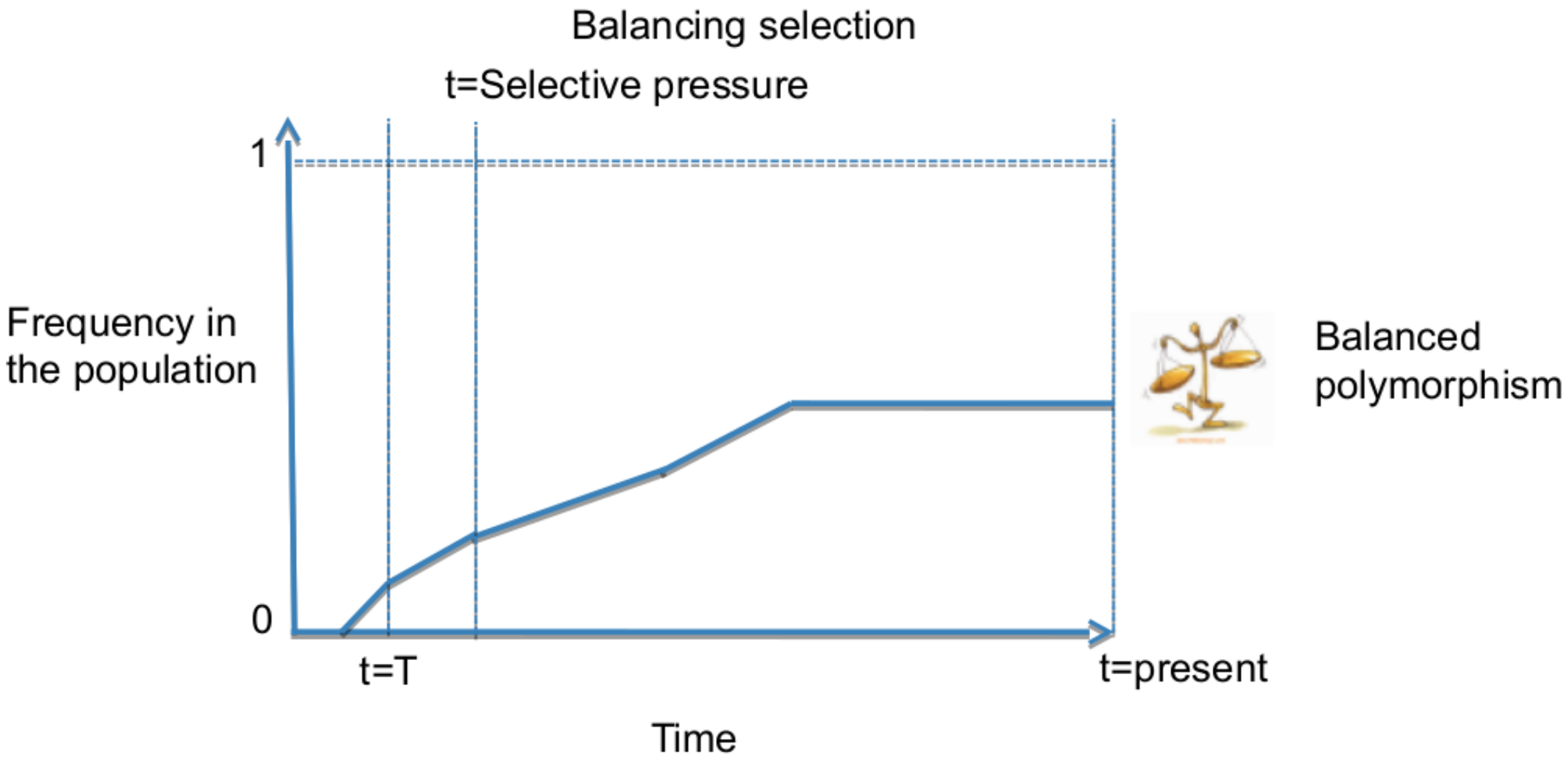
Allele Frequency Trajectory

Positive



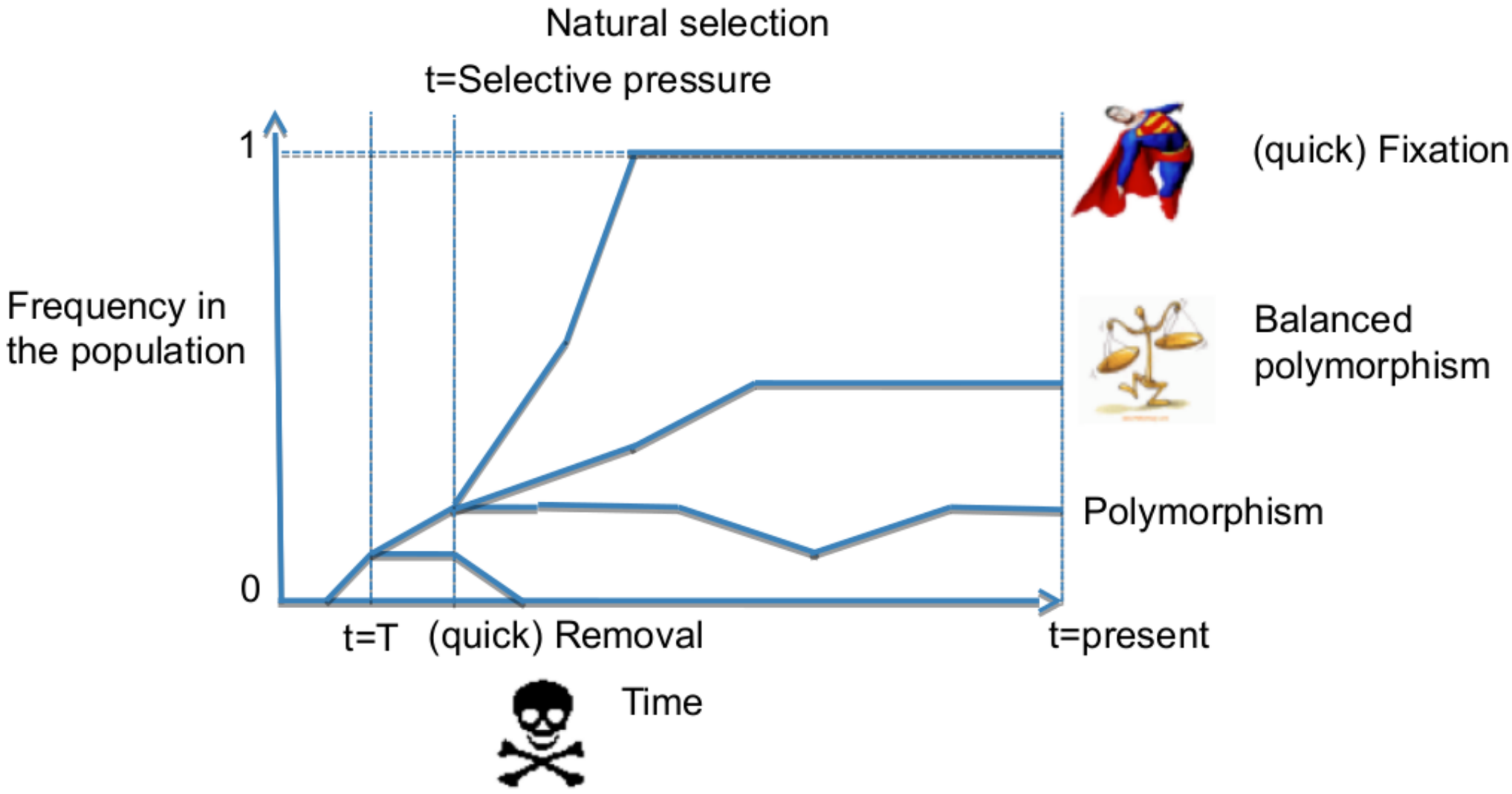
Allele Frequency Trajectory

Balancing



Allele Frequency Trajectory

Natural Selection



Allele Frequency Trajectory

Summary

Effect of selection on alleles:

- Neutral/weak: removed, polymorphic or fixed
- Strong negative: removed or polymorphic
- Strong positive: removed, polymorphic or fixed
- Balancing: removed, polymorphic or fixed

What is “strong” selection?
It depends on the effective population size.

Thus, allele frequency is (almost always) not enough to determine selection.

Testing for natural selection

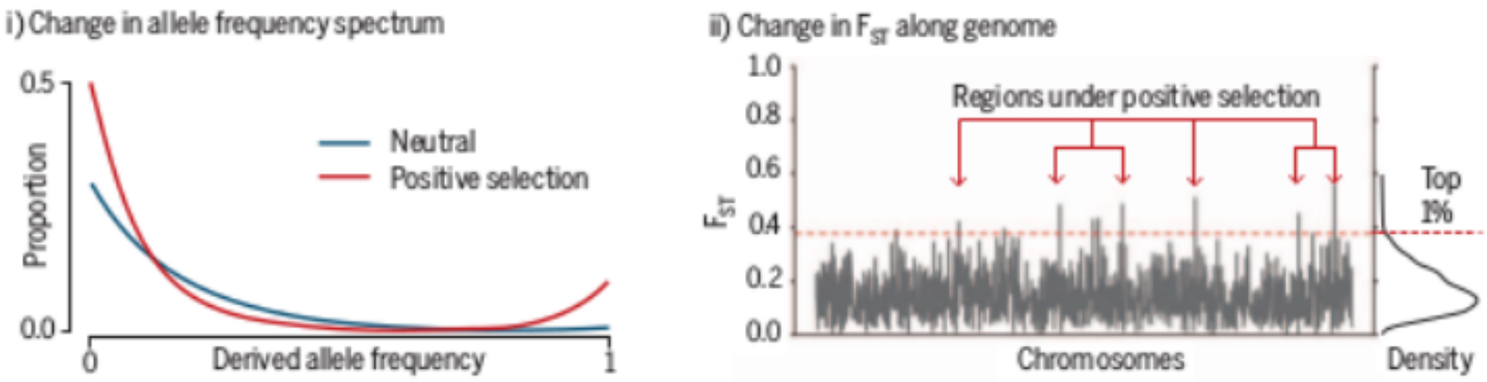
If the simple observation of allele frequencies is not sufficient, what else can we do to detect signals of natural selection?

Testing for natural selection

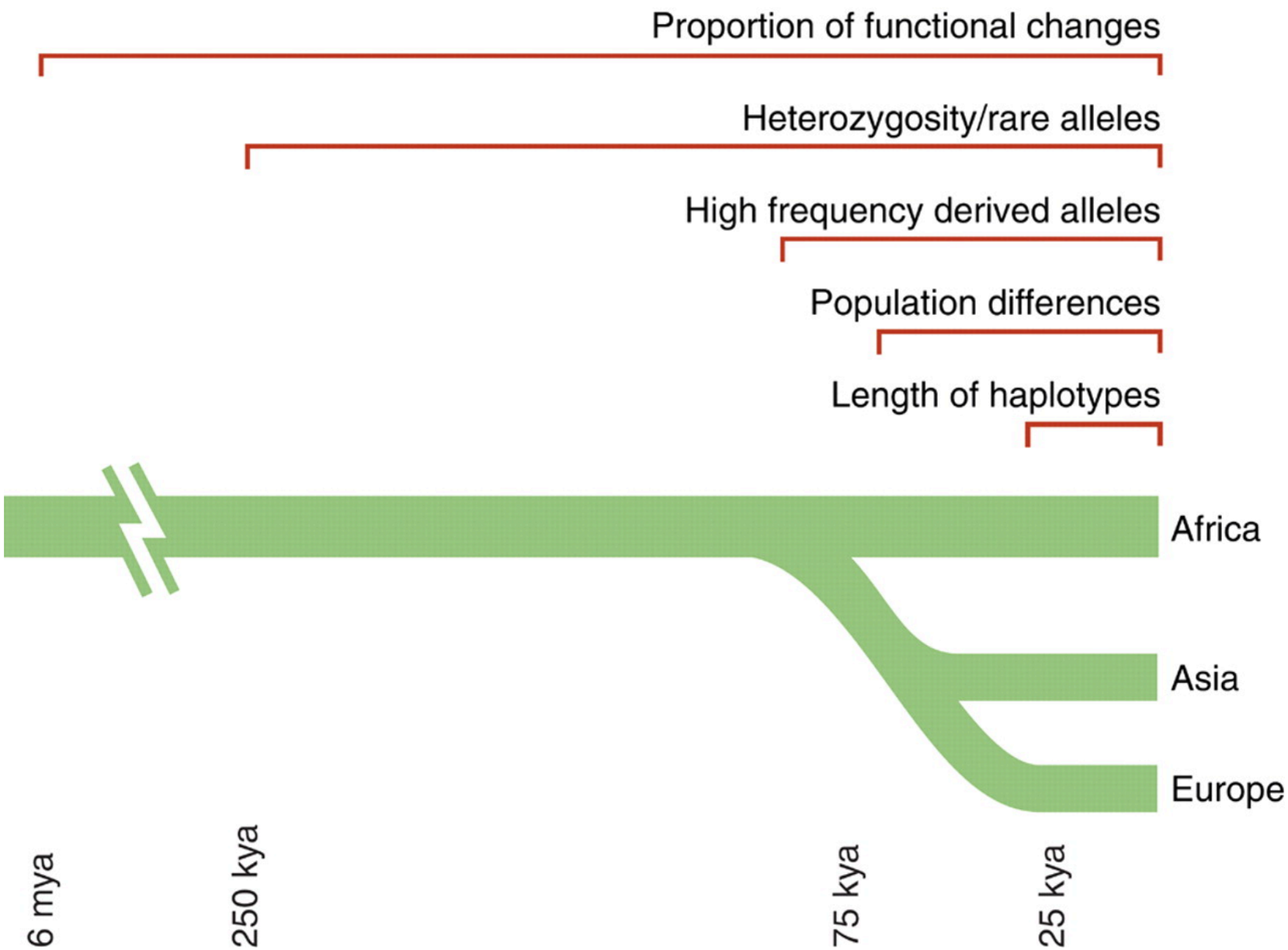
If the simple observation of allele frequencies is not sufficient, what else can we do to detect signals of natural selection?

- Option 1: Use information from the genomic region
- Option 2: Use information from multiple species/populations
- Option 3: Selection experiments
- Use external information:
  - Candidate genes/biological knowledge
  - Functional categories
  - Association to phenotypes

Common methods used to detect selection

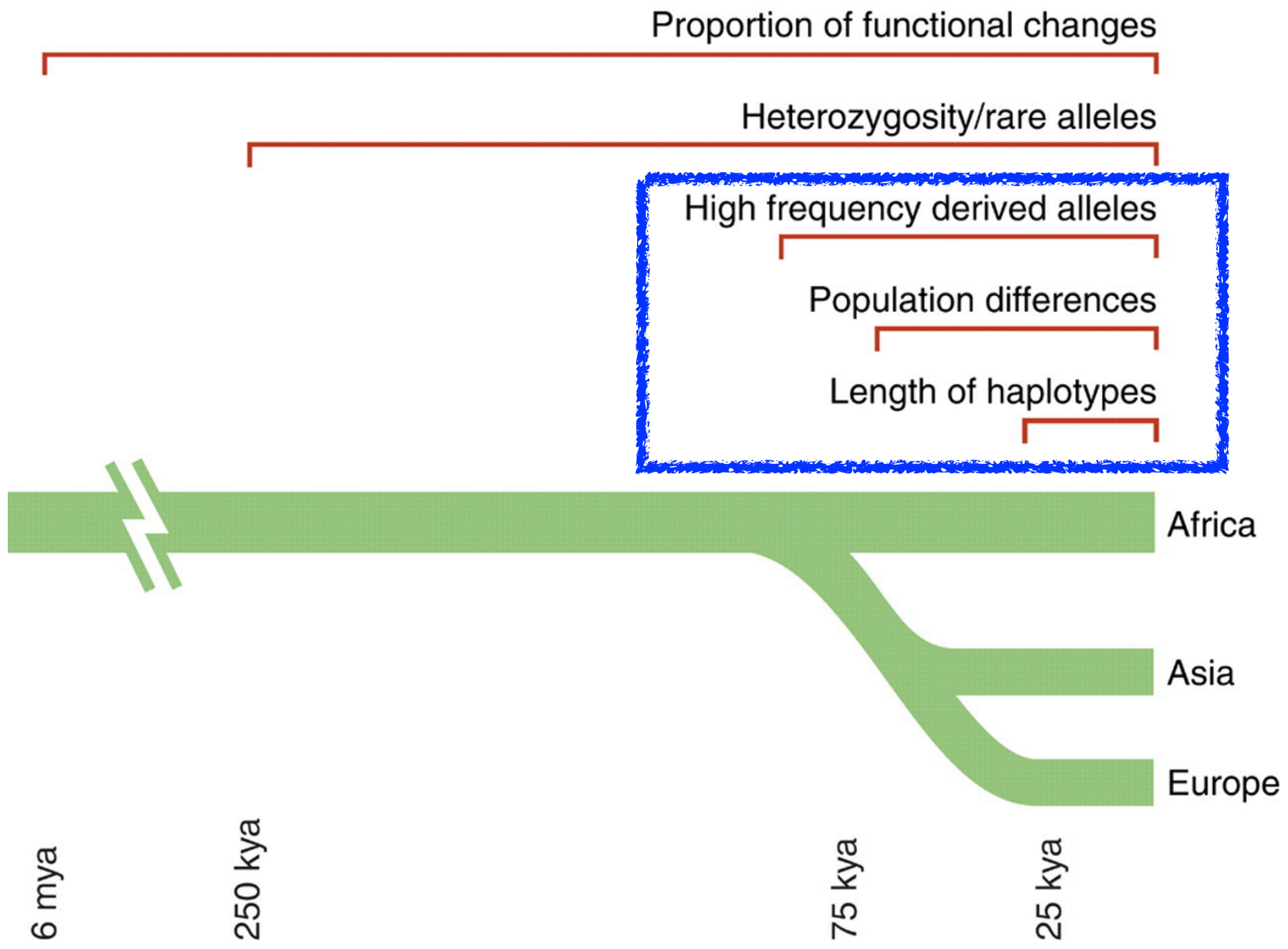


Detecting recent selection



Sabeti et al. 2006, Science

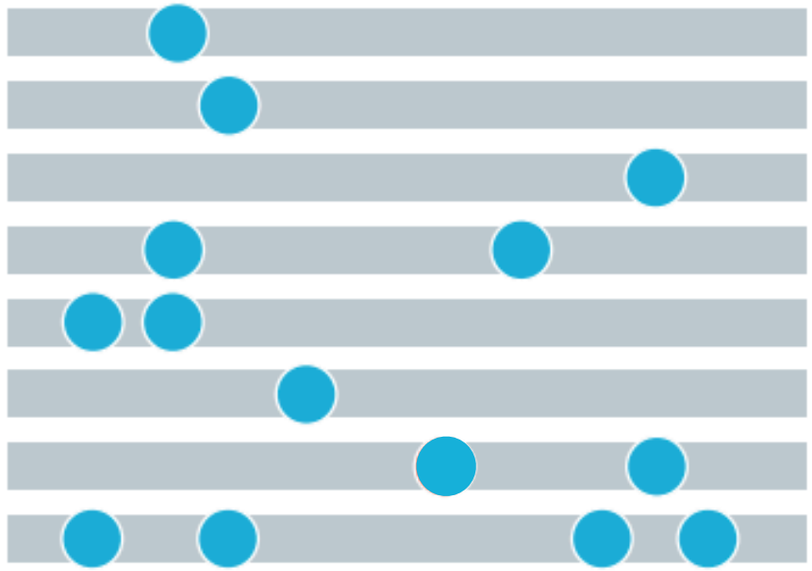
Detecting recent selection



Sabeti et al. 2006, Science

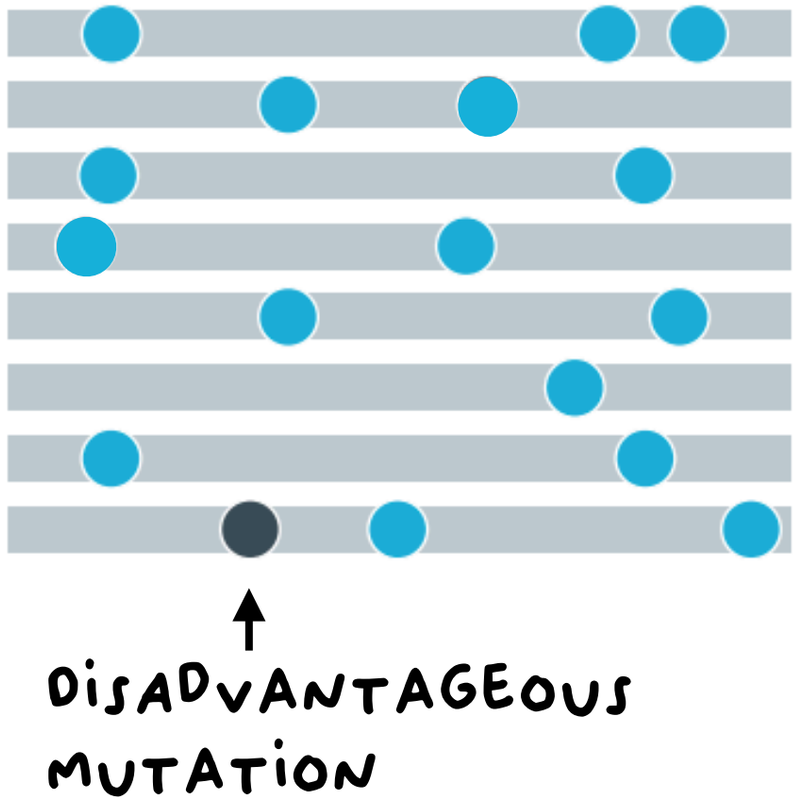
Signature of selection

- Neutral locus
- Lots of variability



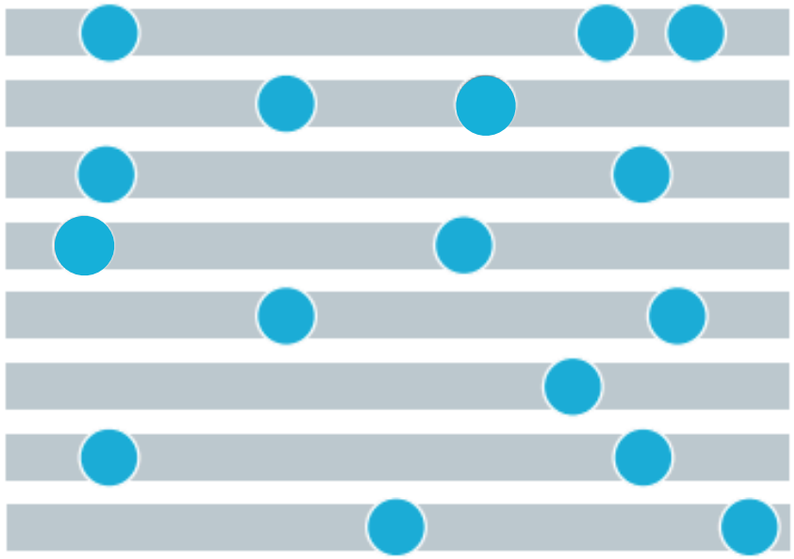
Signature of selection

- Mutations enters the population



Signature of selection

- Negative selection removed the allele



Signature of selection

- Mutation enters the population

Signature of selection

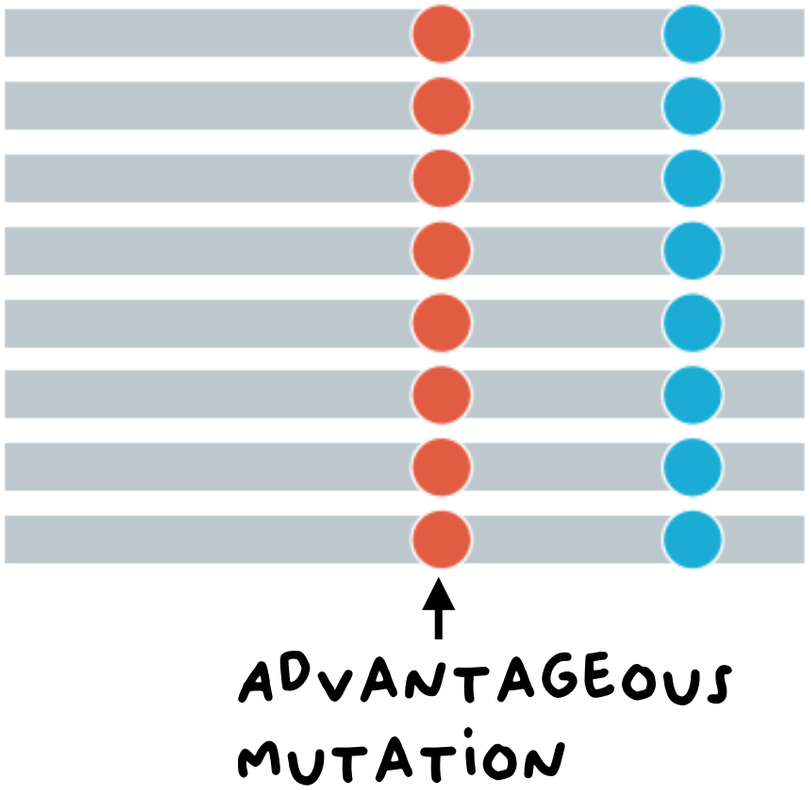
- Mutation enters the population
- Mutation increases in frequency due to positive selection

Signature of selection

- Increases LD
- Affects the variability

Signature of selection

- Increases haplotype similarity

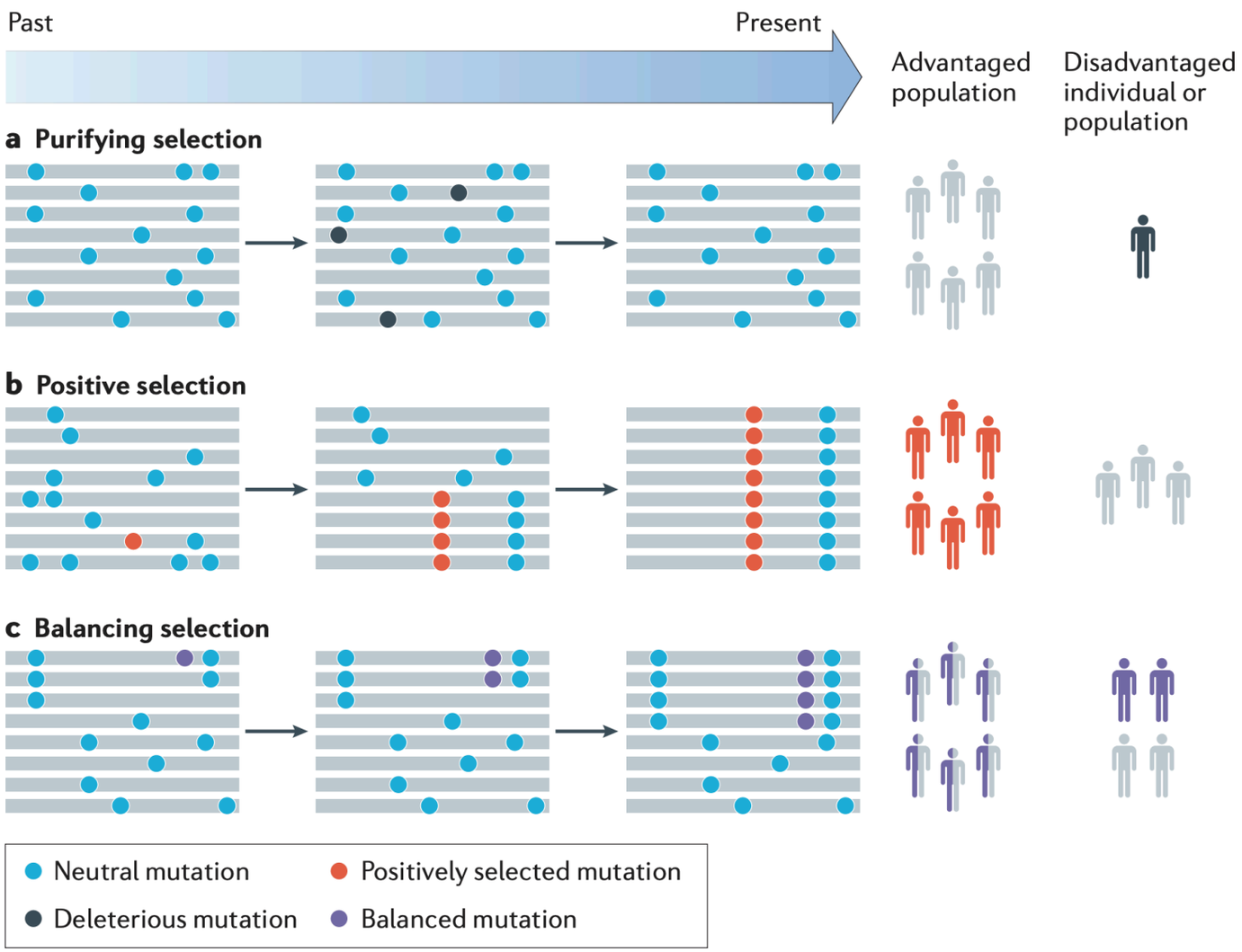


Signature of selection

- Increases differences with other populations in the whole region

Natural Selection

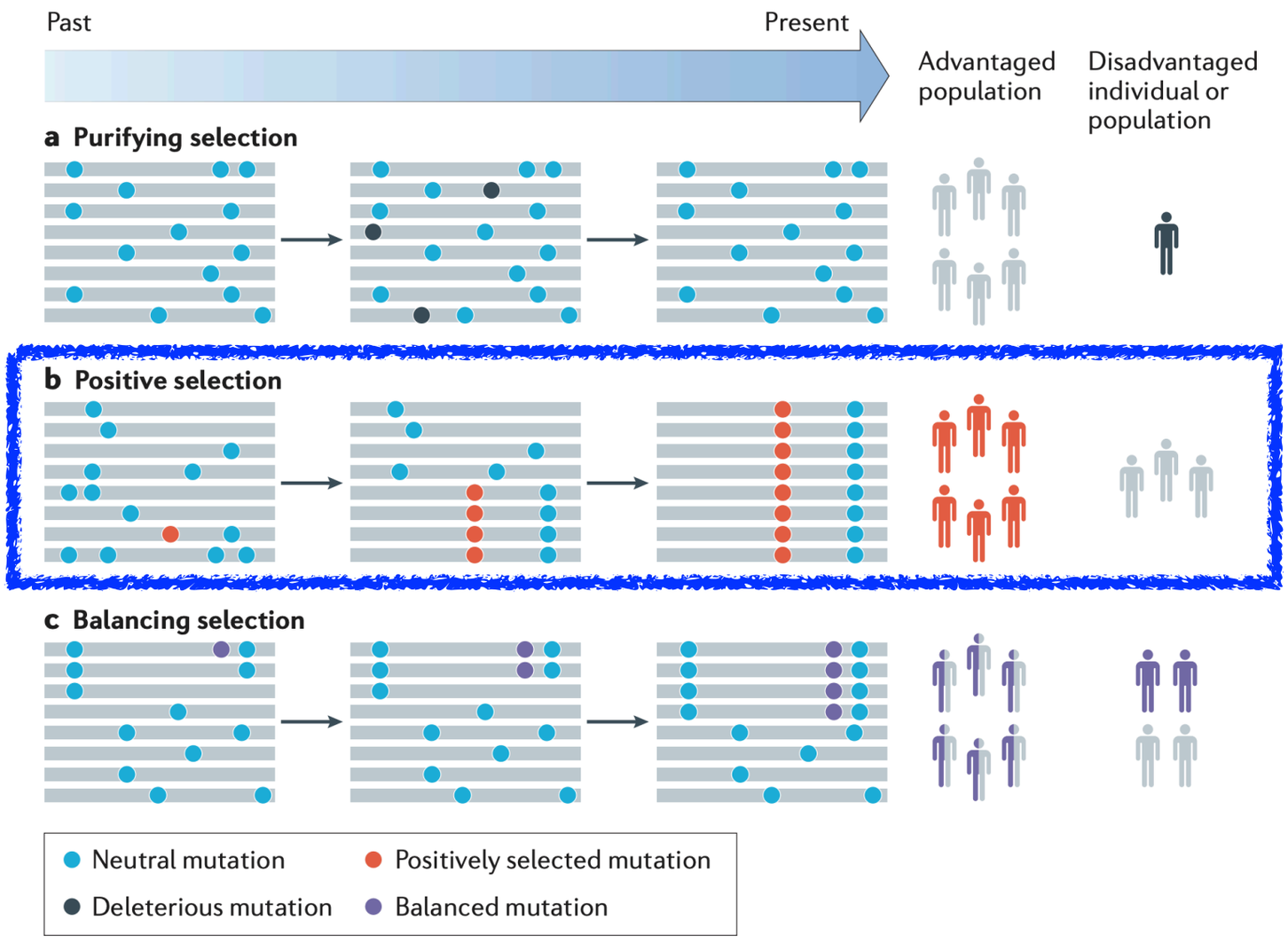
Types of Selection



Quintana-Murci, L., Clark, A. Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol* **13,** 280–293 (2013). https://doi.org/10.1038/nri3421

Natural Selection

Types of Selection



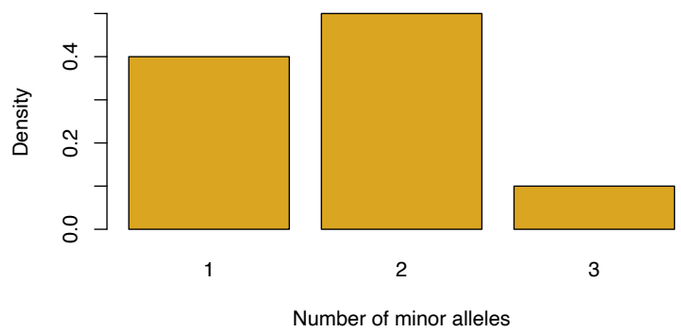
Quintana-Murci, L., Clark, A. Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol* **13,** 280–293 (2013). https://doi.org/10.1038/nri3421

Frequency-based methods

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

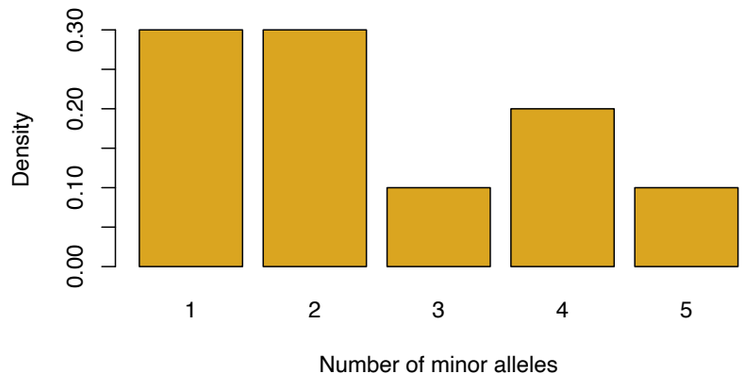
Number of minor alleles (folded) $\eta = (0.4, 0.5, 0.1)$



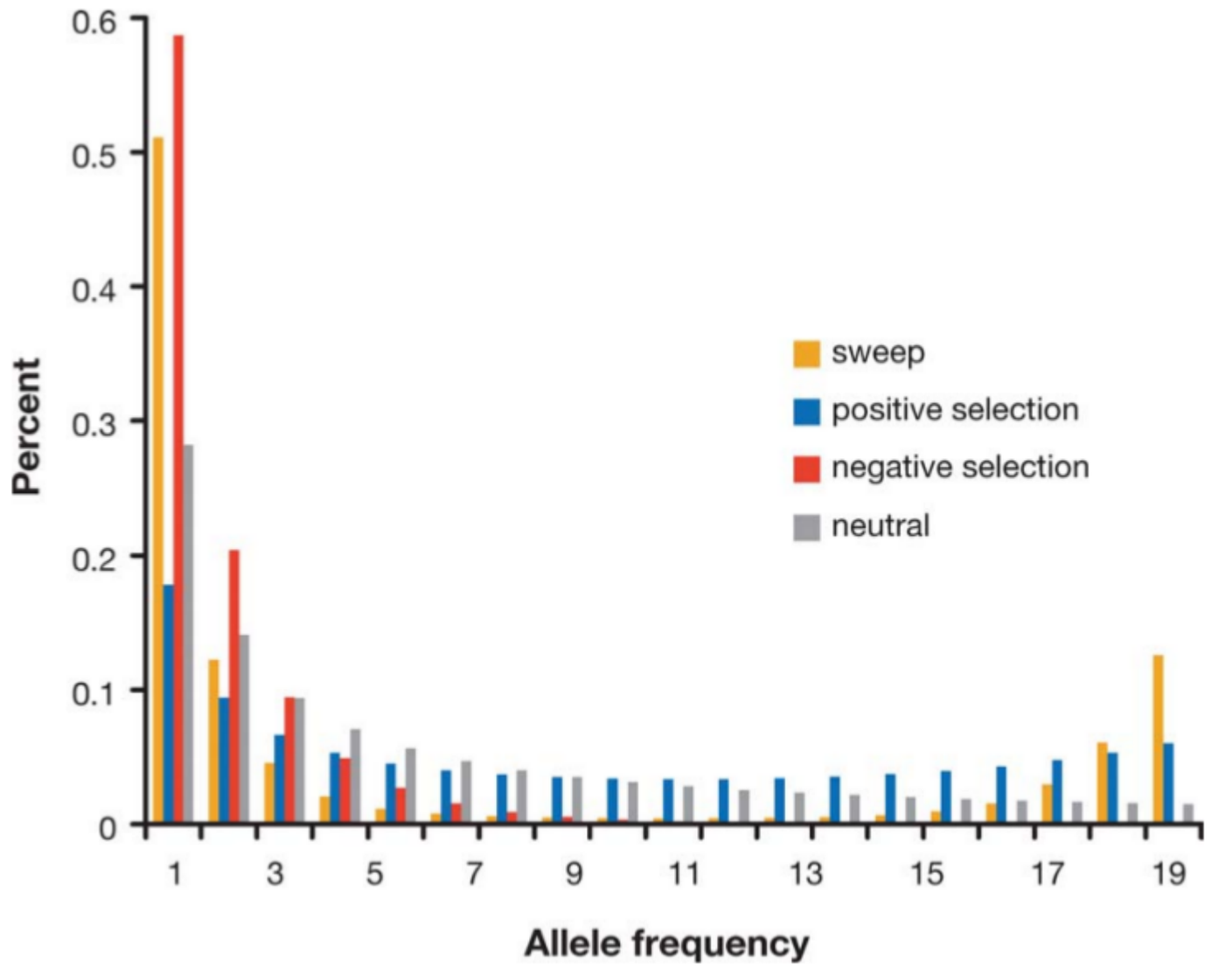
What is the site frequency spectrum?

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* $\theta_T = \theta_W$

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

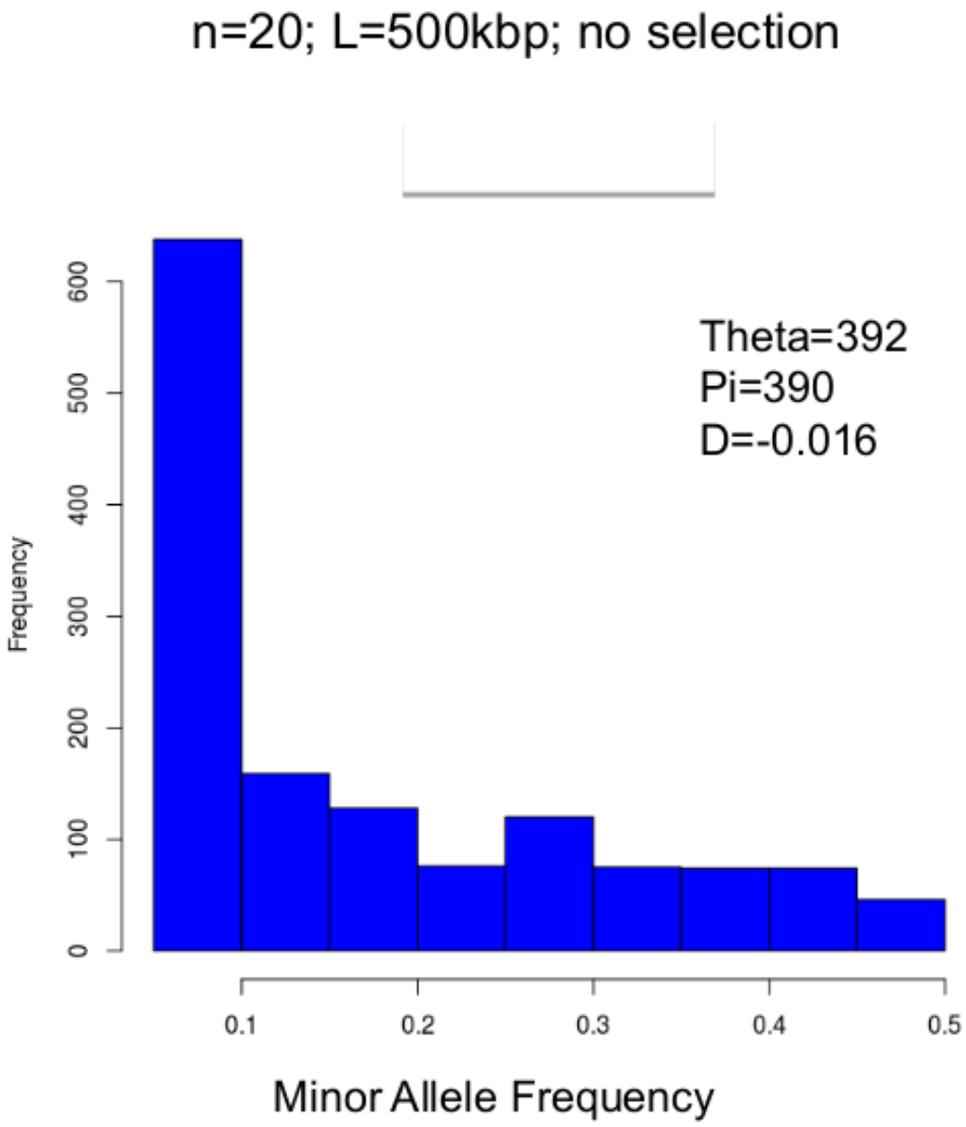
general $\hat{\theta} = \sum_{i=0}^{n} \alpha_i \eta_i$

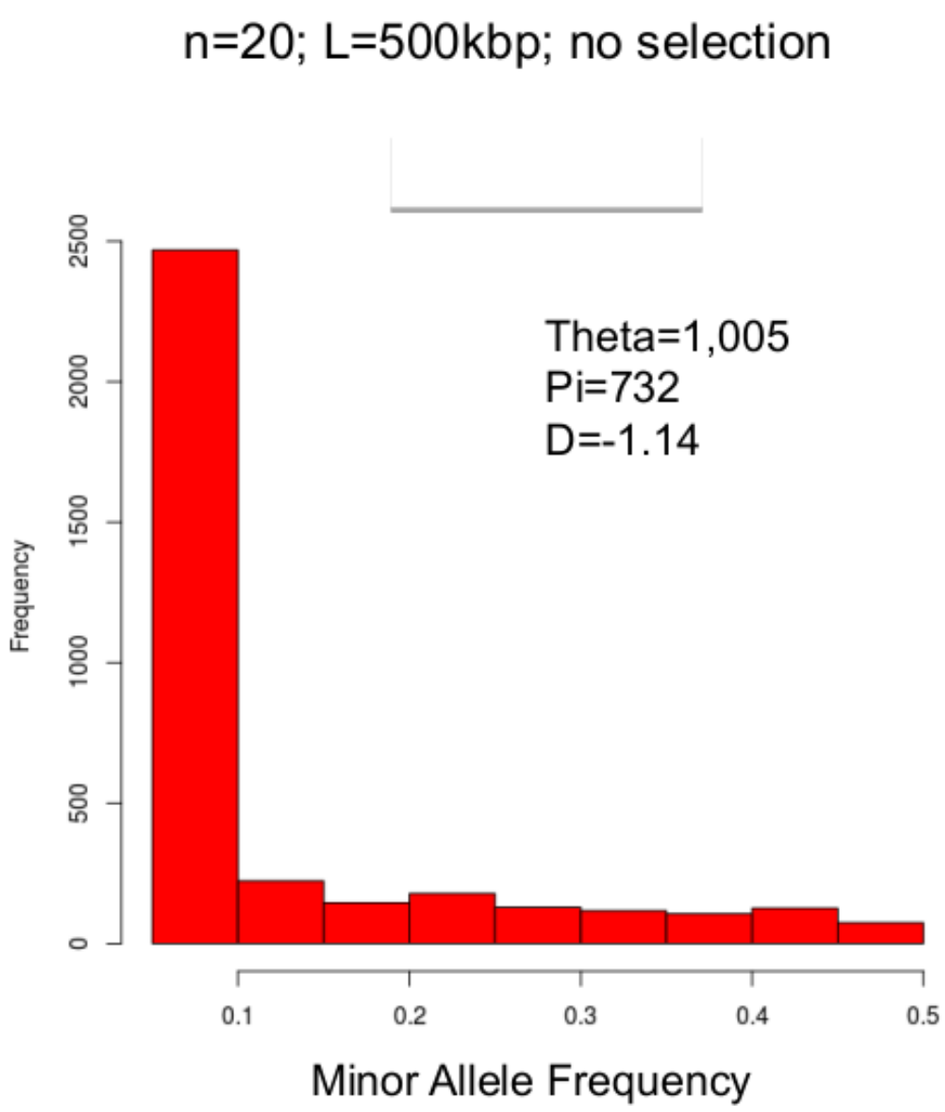
Test statistics

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

Difference weighting schemes for the SFS

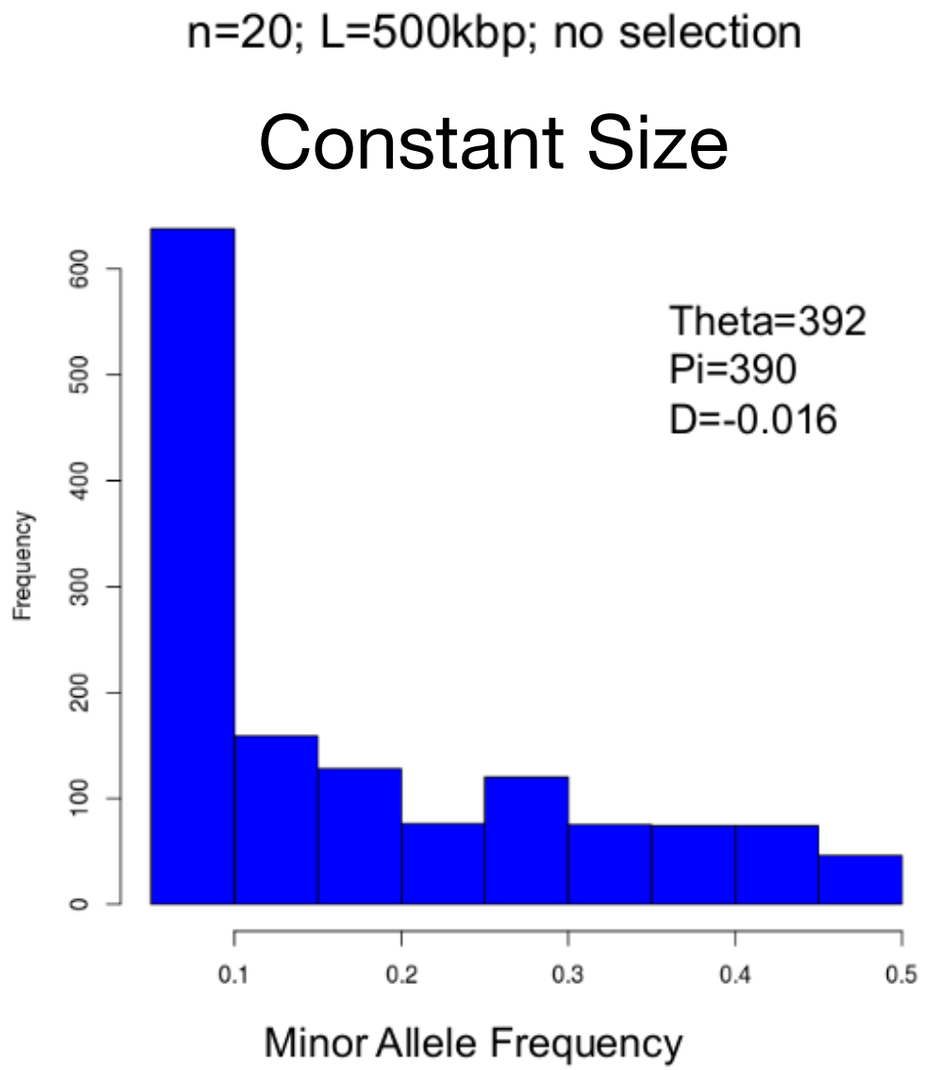
Confounding factors

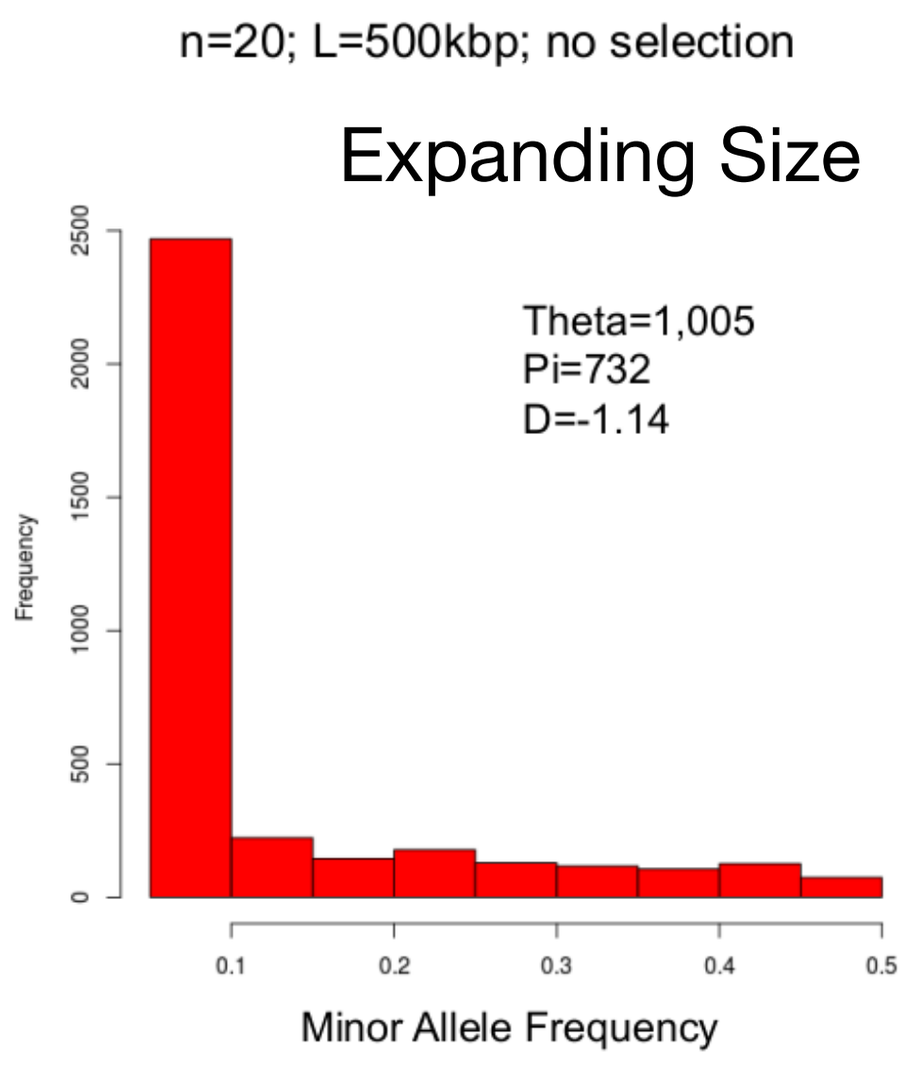




Confounding factors

Demography matters

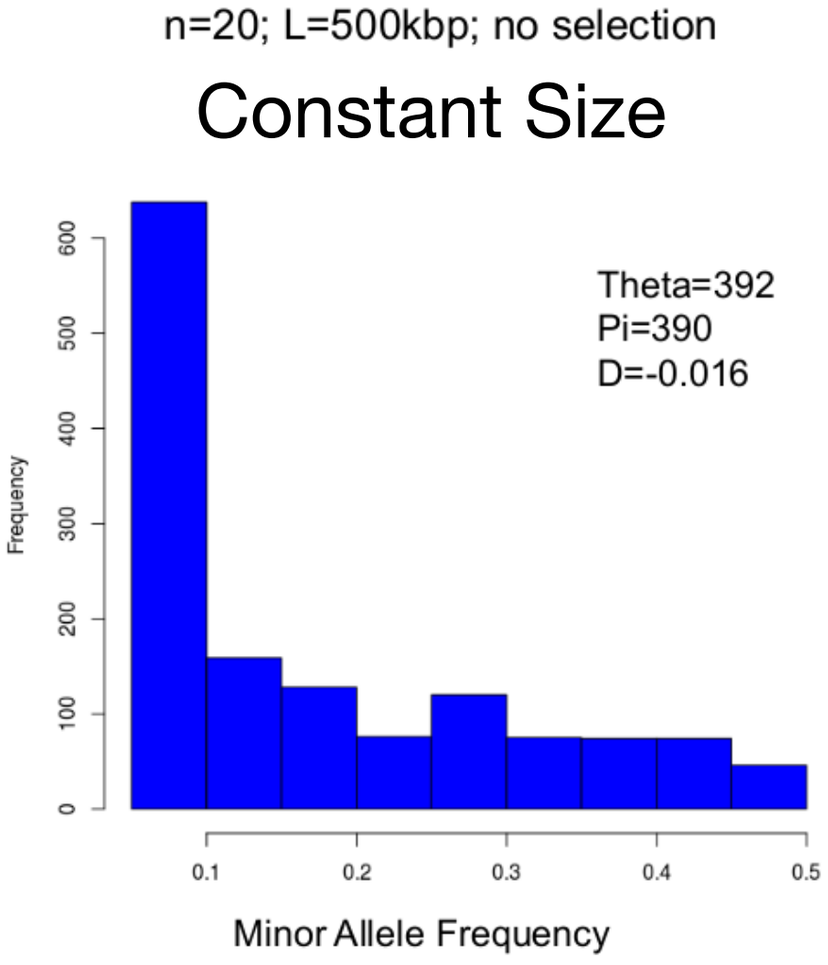


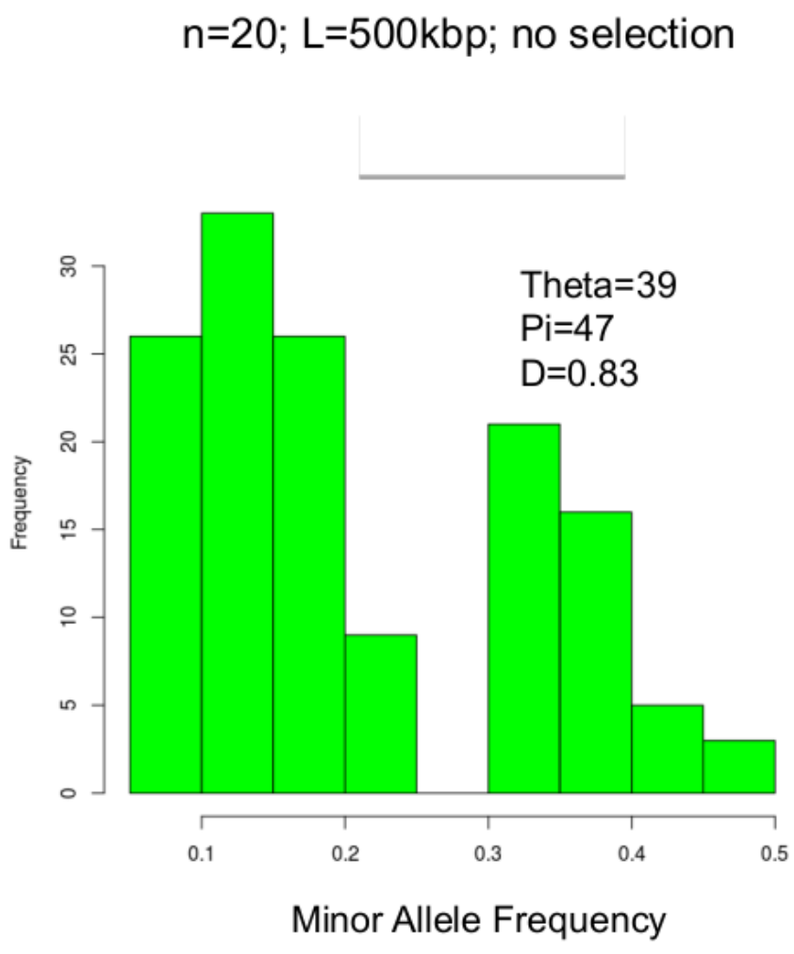


- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

Confounding factors

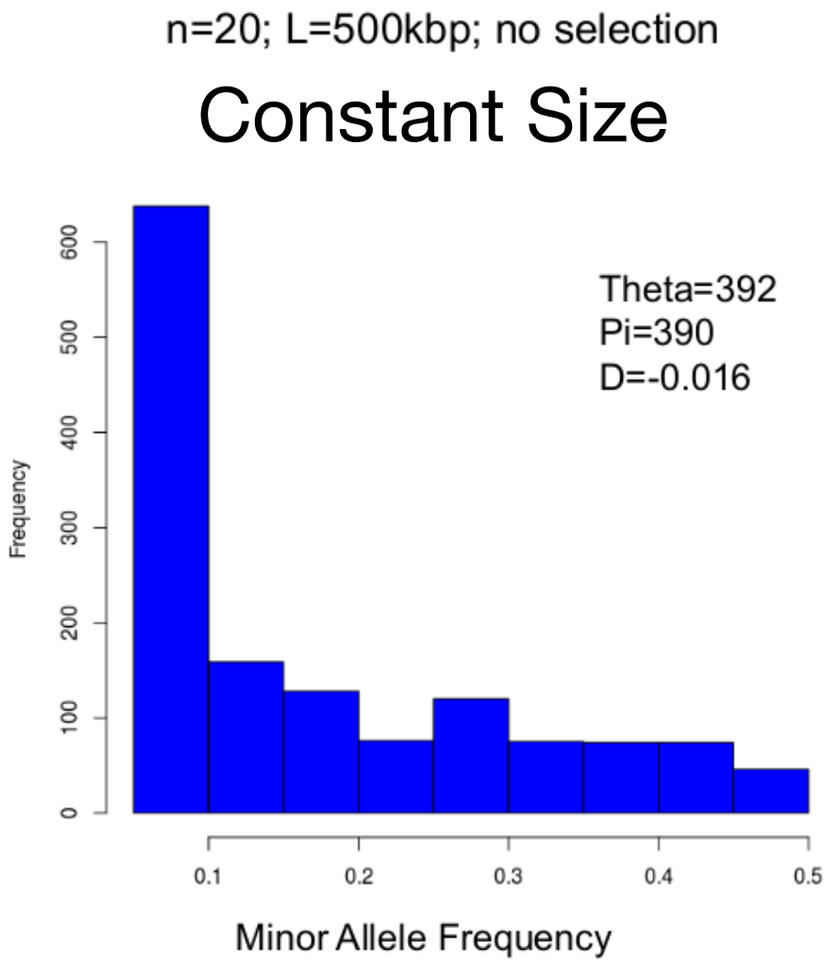
Demography matters

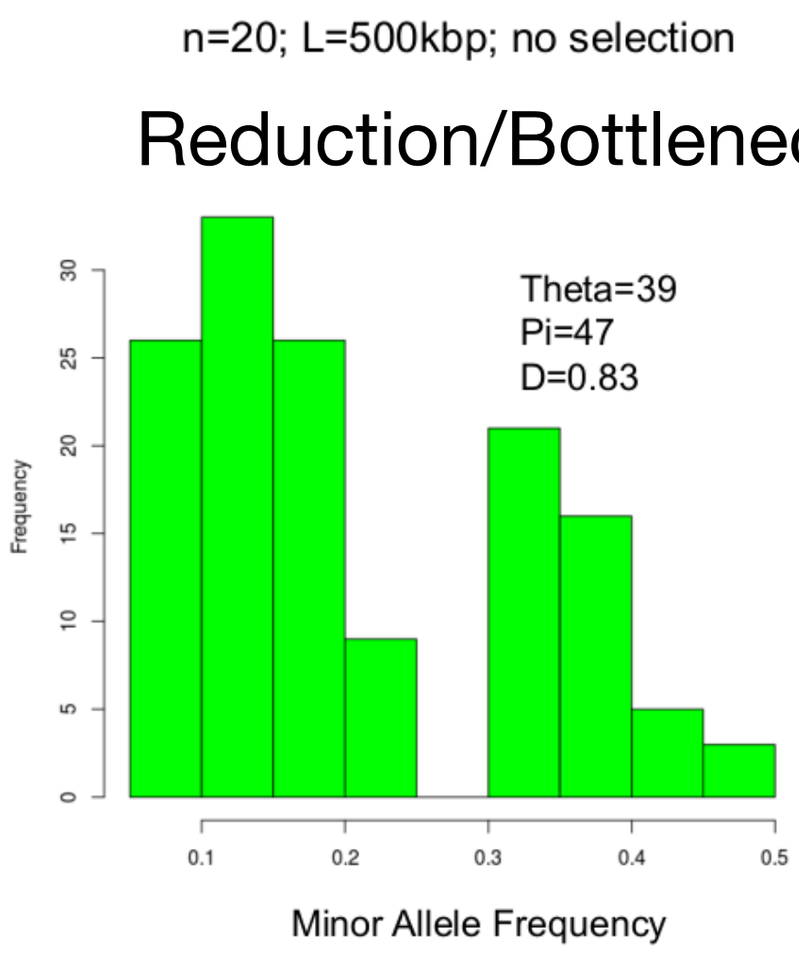




Confounding factors

Demography matters

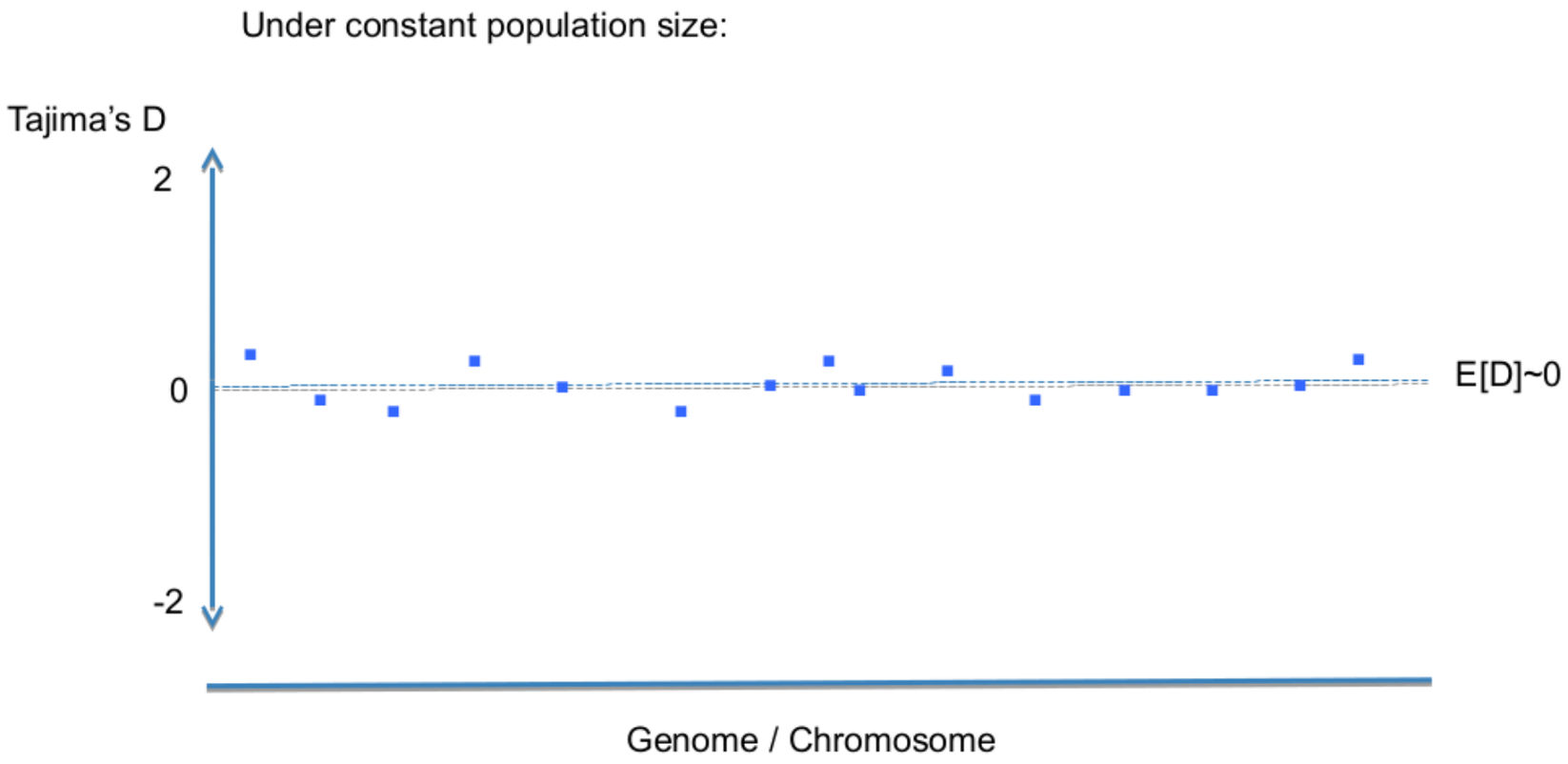




- Depletion of segregating sites
- Excess of intermediate-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

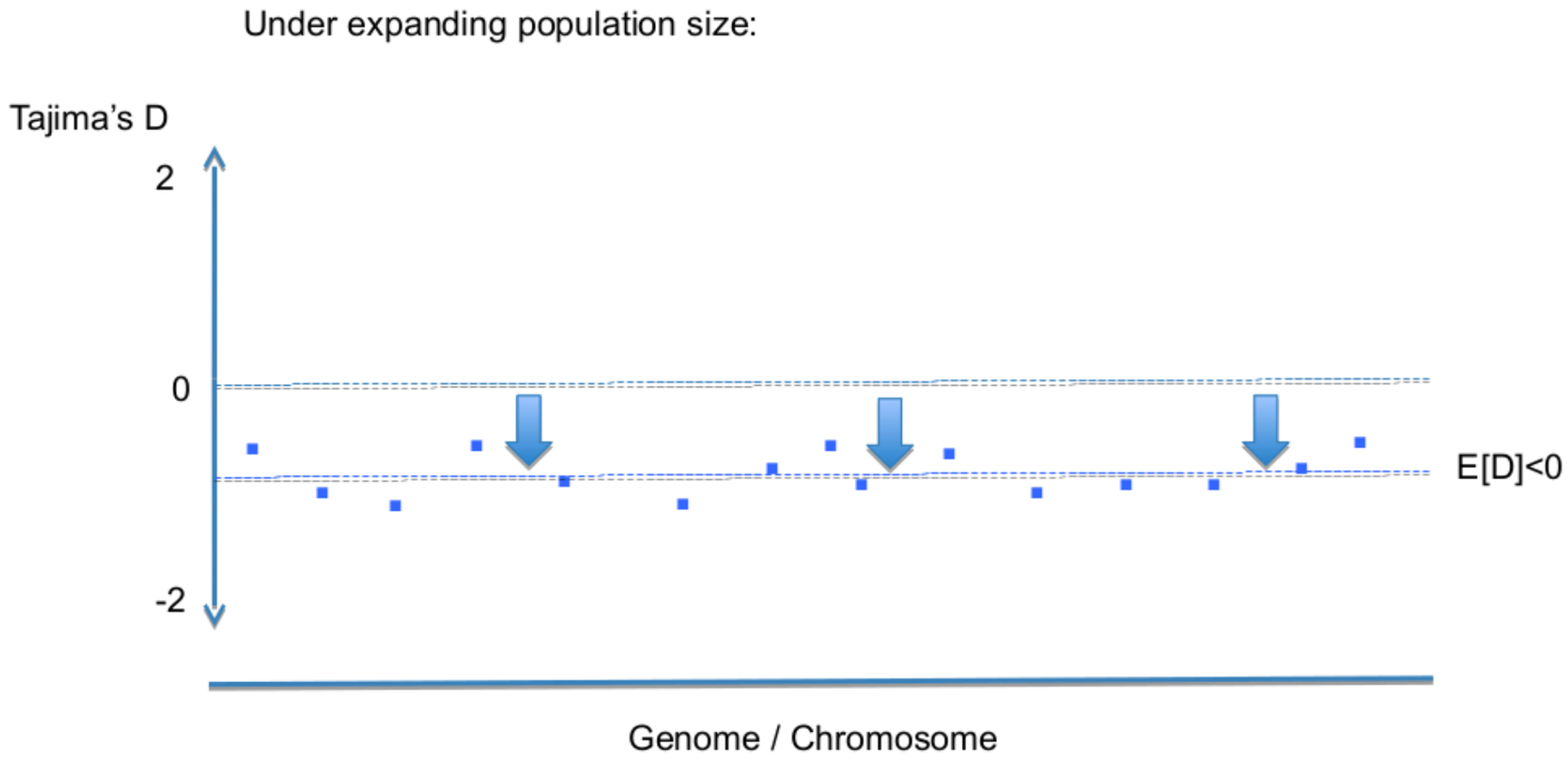
Accounting for confounding factors

How to assess significance



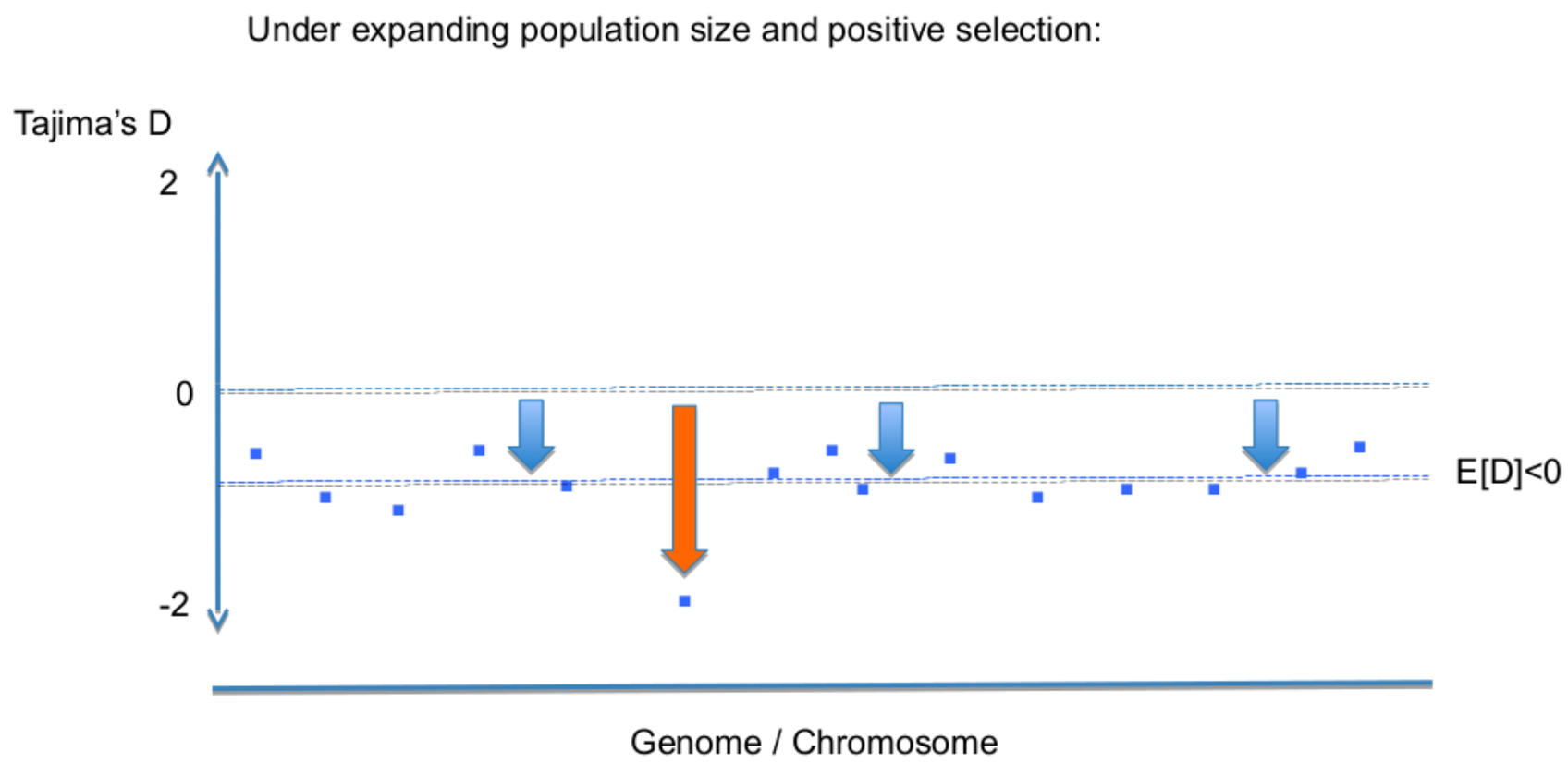
Accounting for confounding factors

How to assess significance



Accounting for confounding factors

How to assess significance

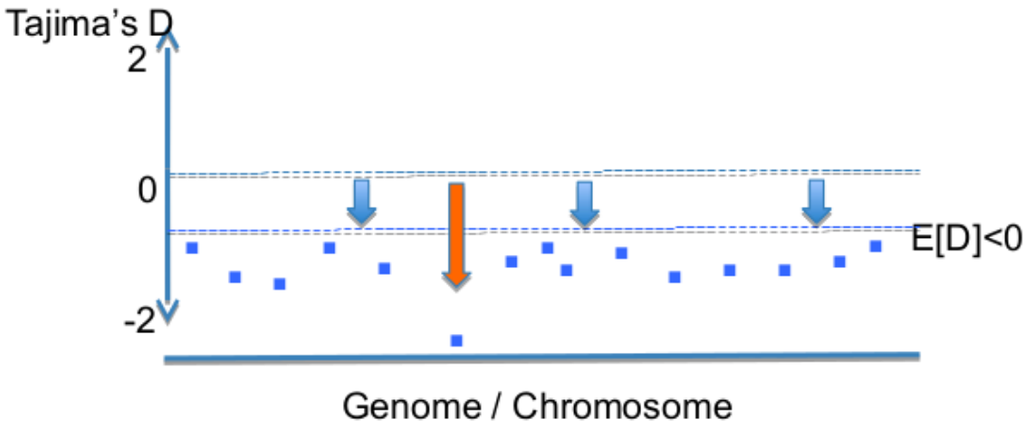


- Demography affects all loci equally, while selection changes local patterns

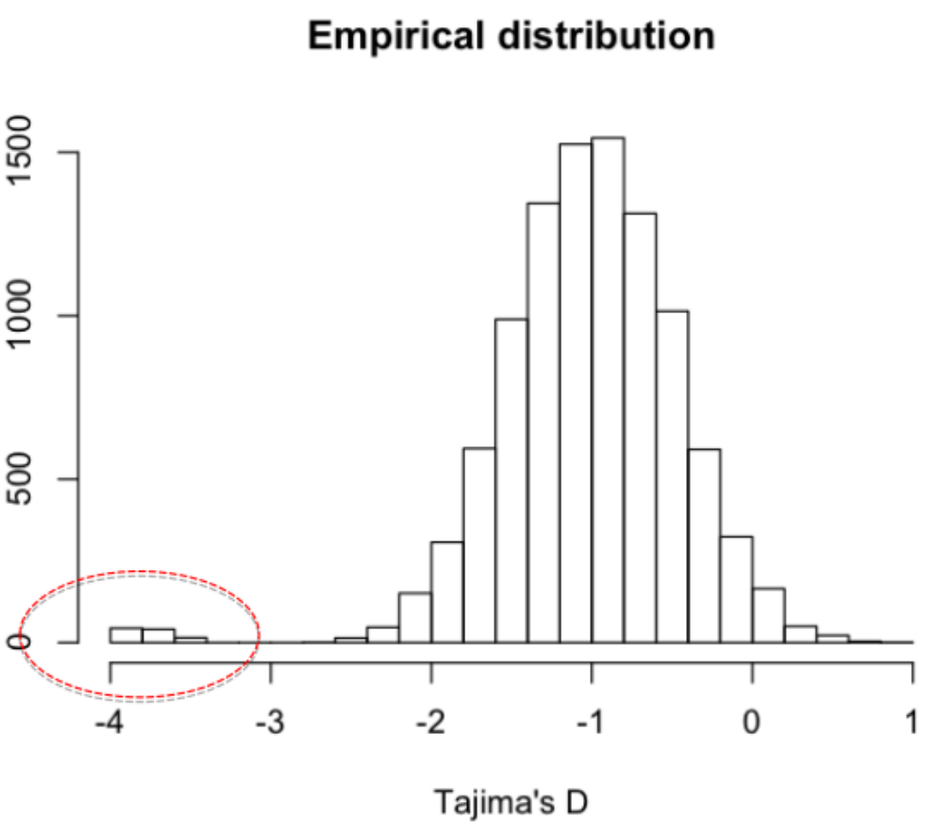
Accounting for confounding factors

How to assess significance

Outlier approach

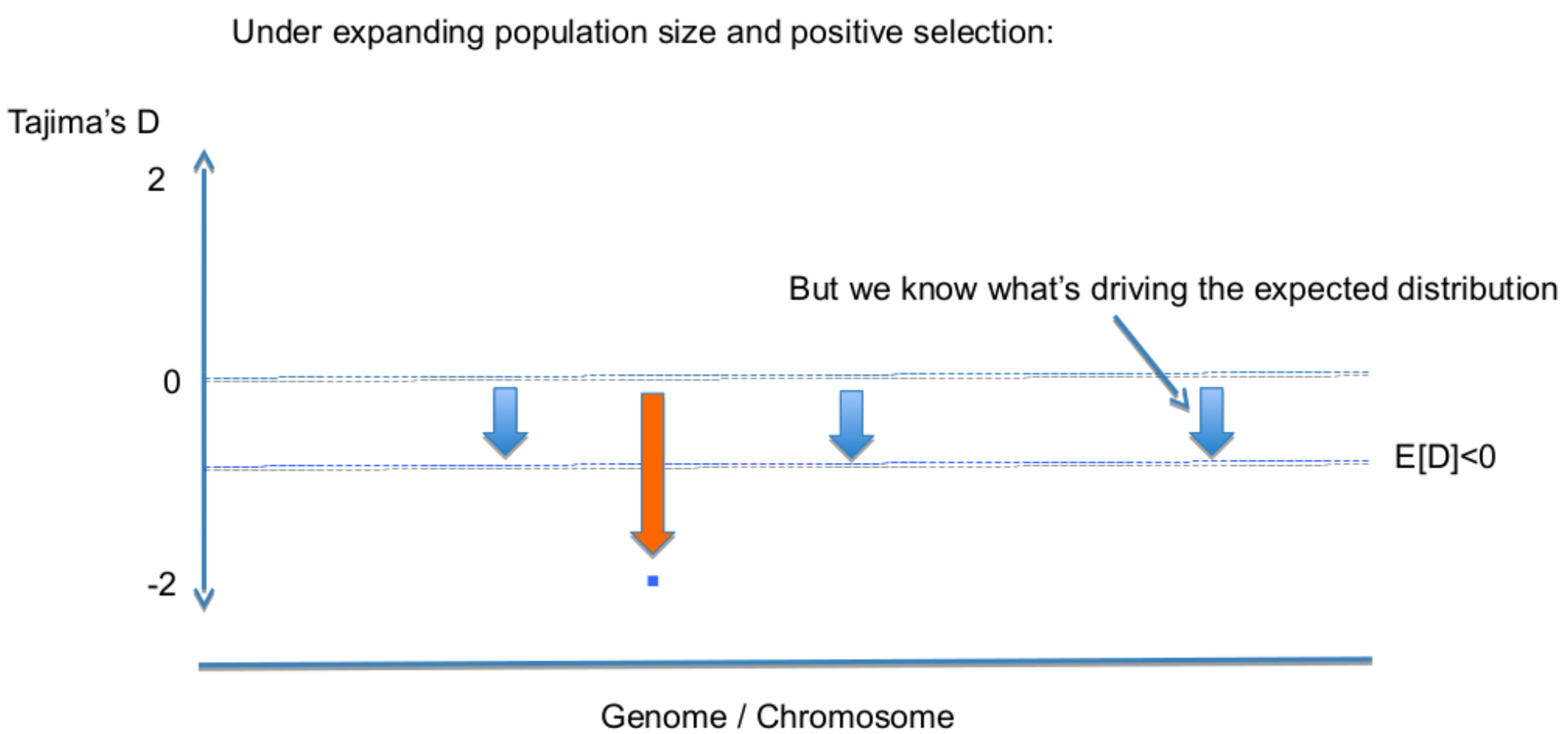


Assign empirical *p*-values
(ranked percentiles)



Accounting for confounding factors

How to assess significance

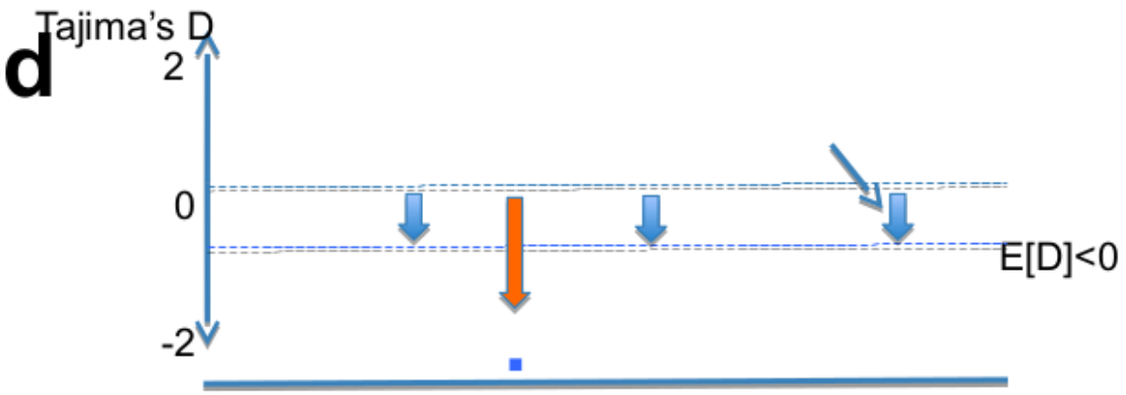


- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

Accounting for confounding factors

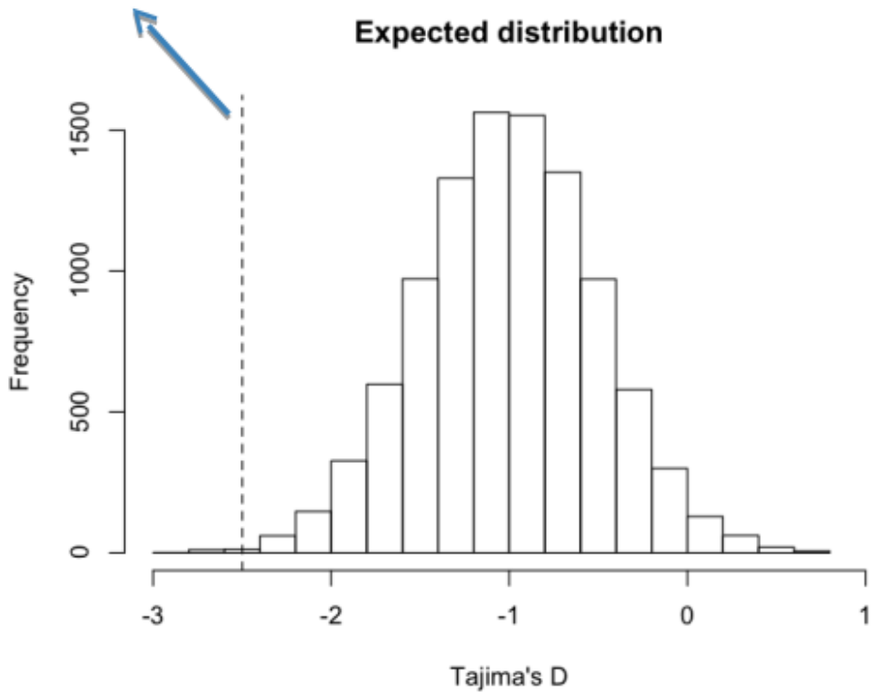
How to assess significance

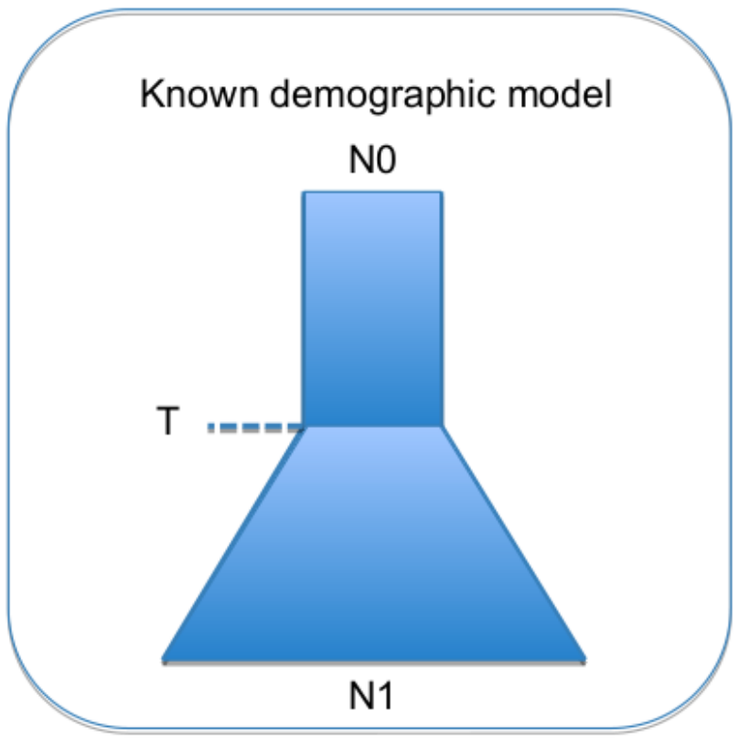
Simulations-based approach



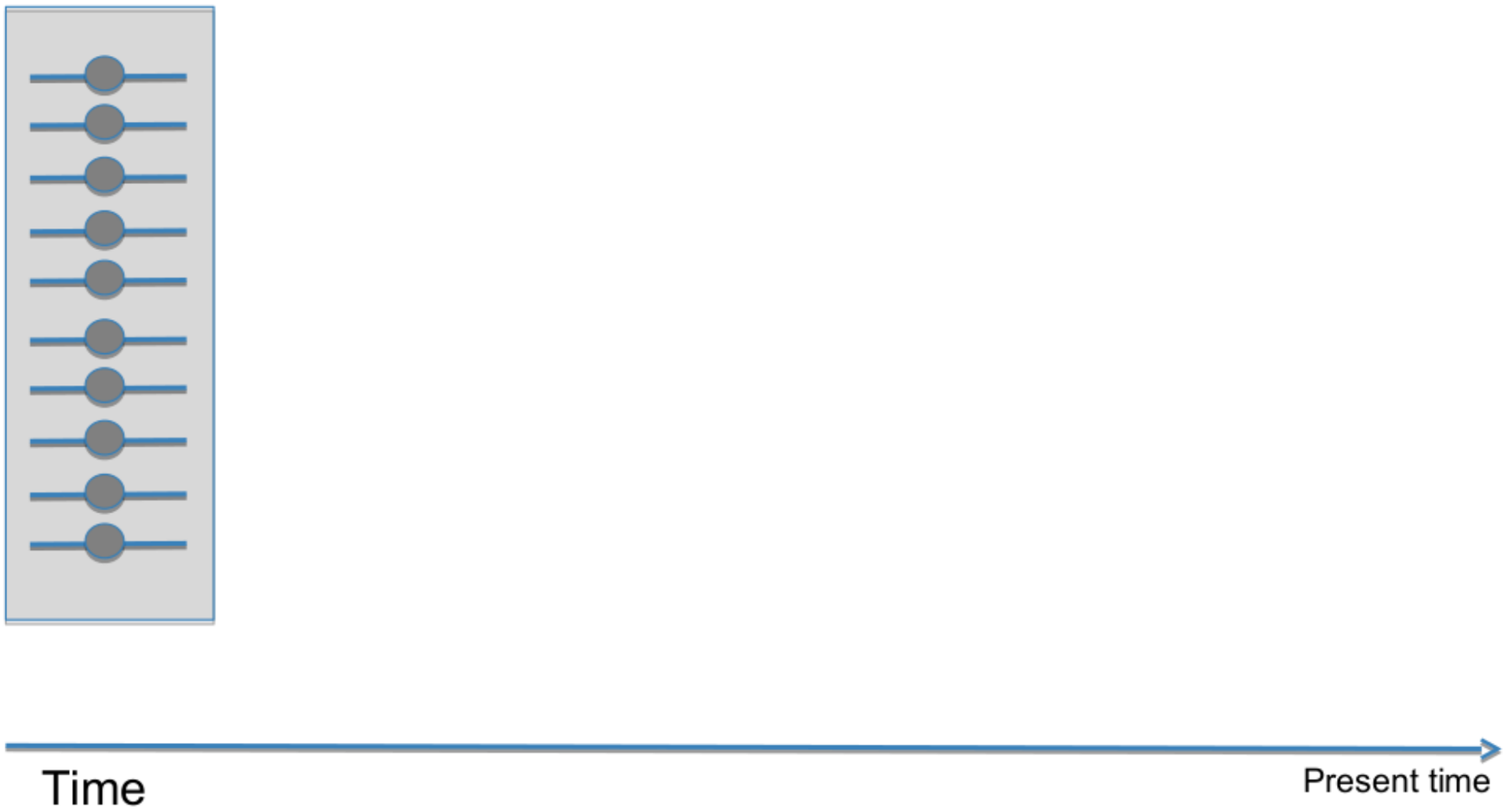
Genome / Chromosome

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

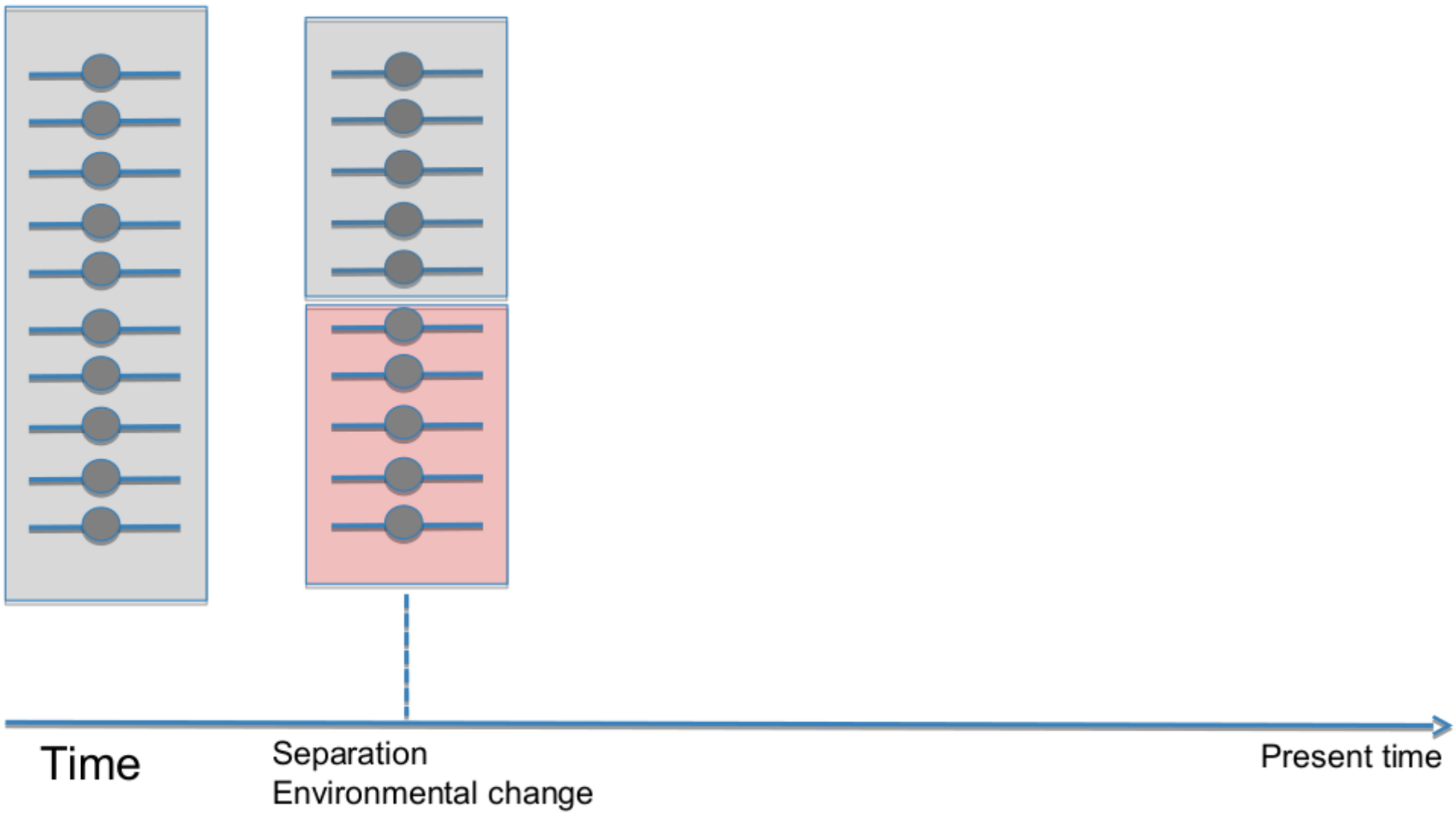




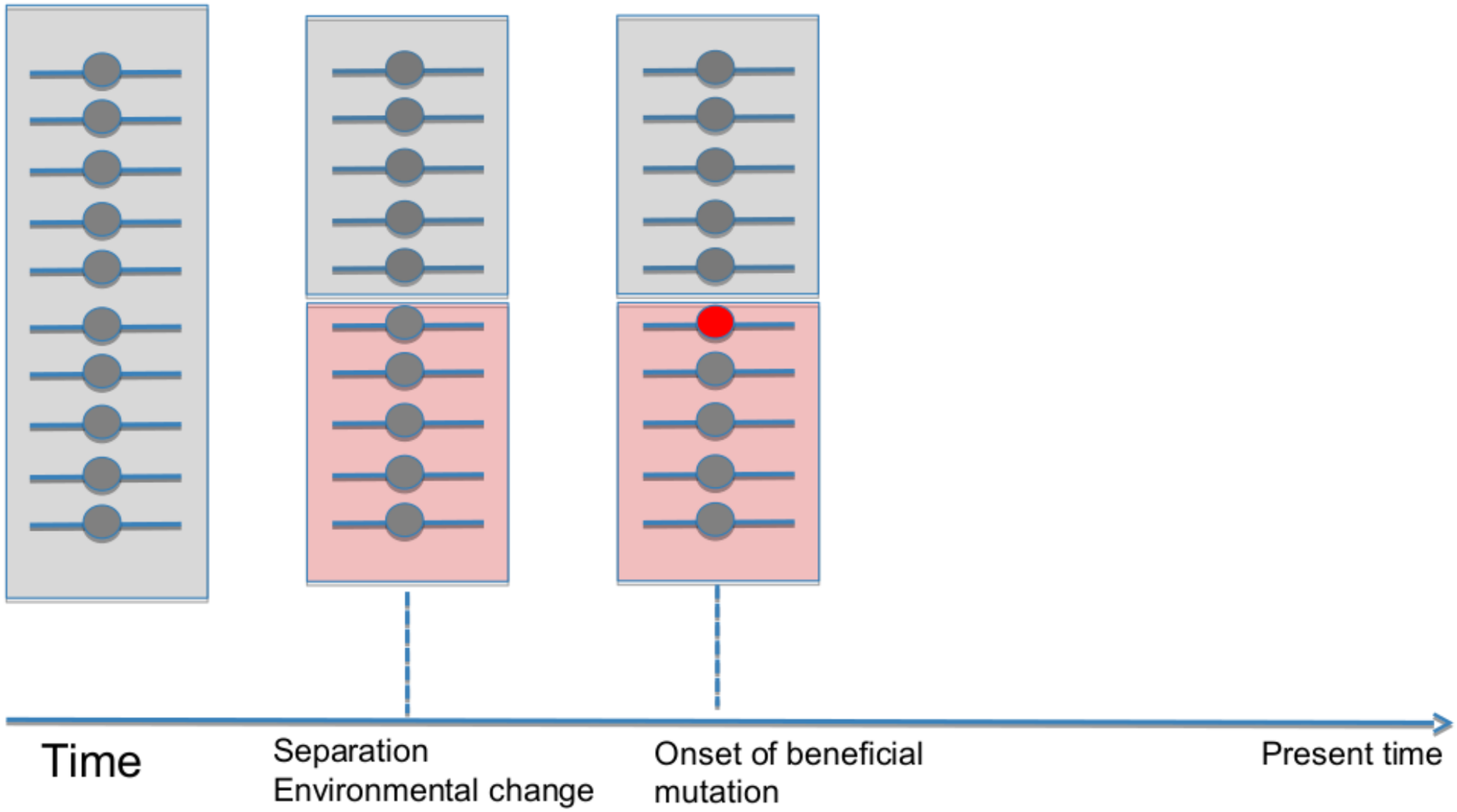
Allele Frequency Differentiation



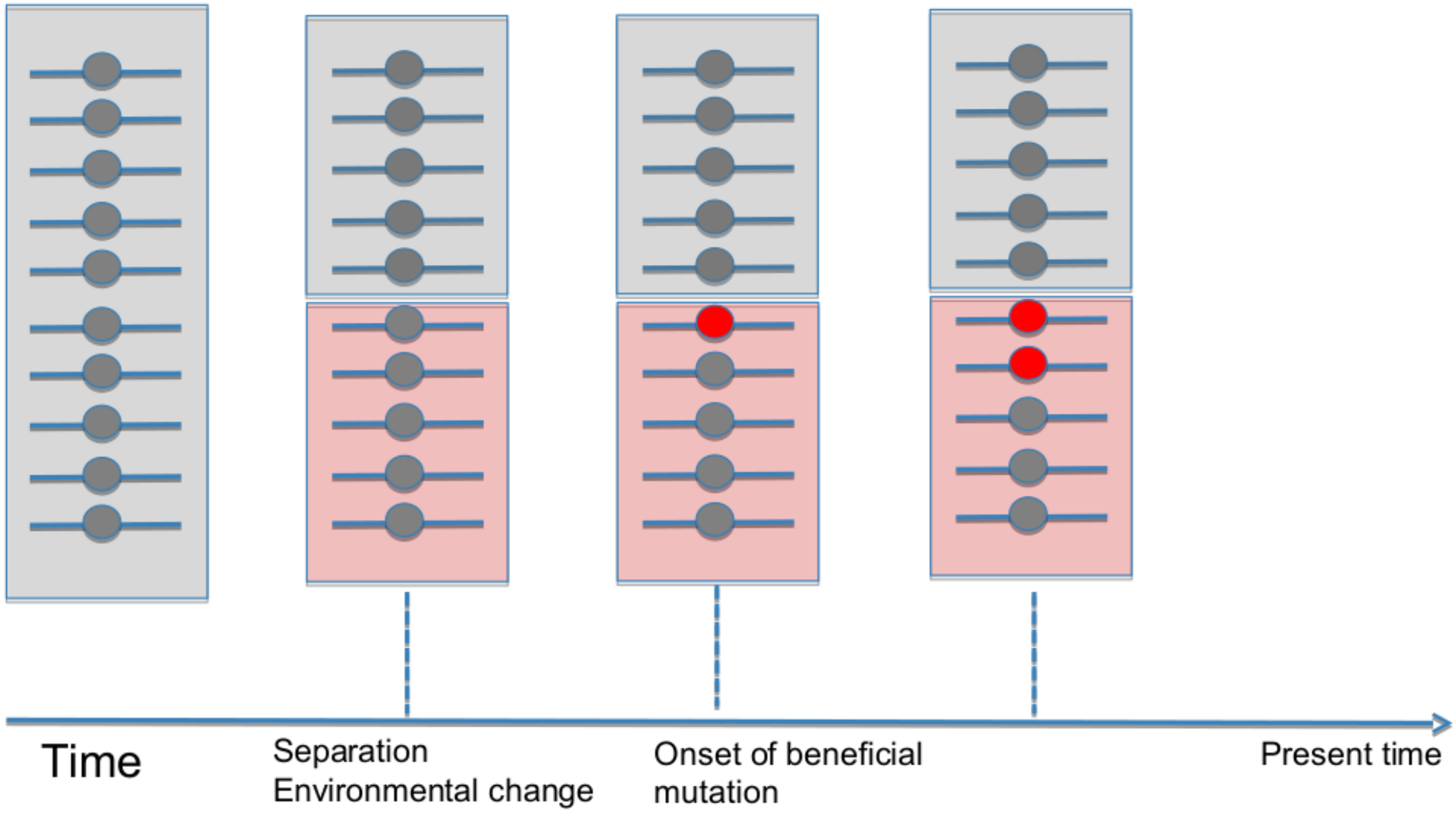
Allele Frequency Differentiation



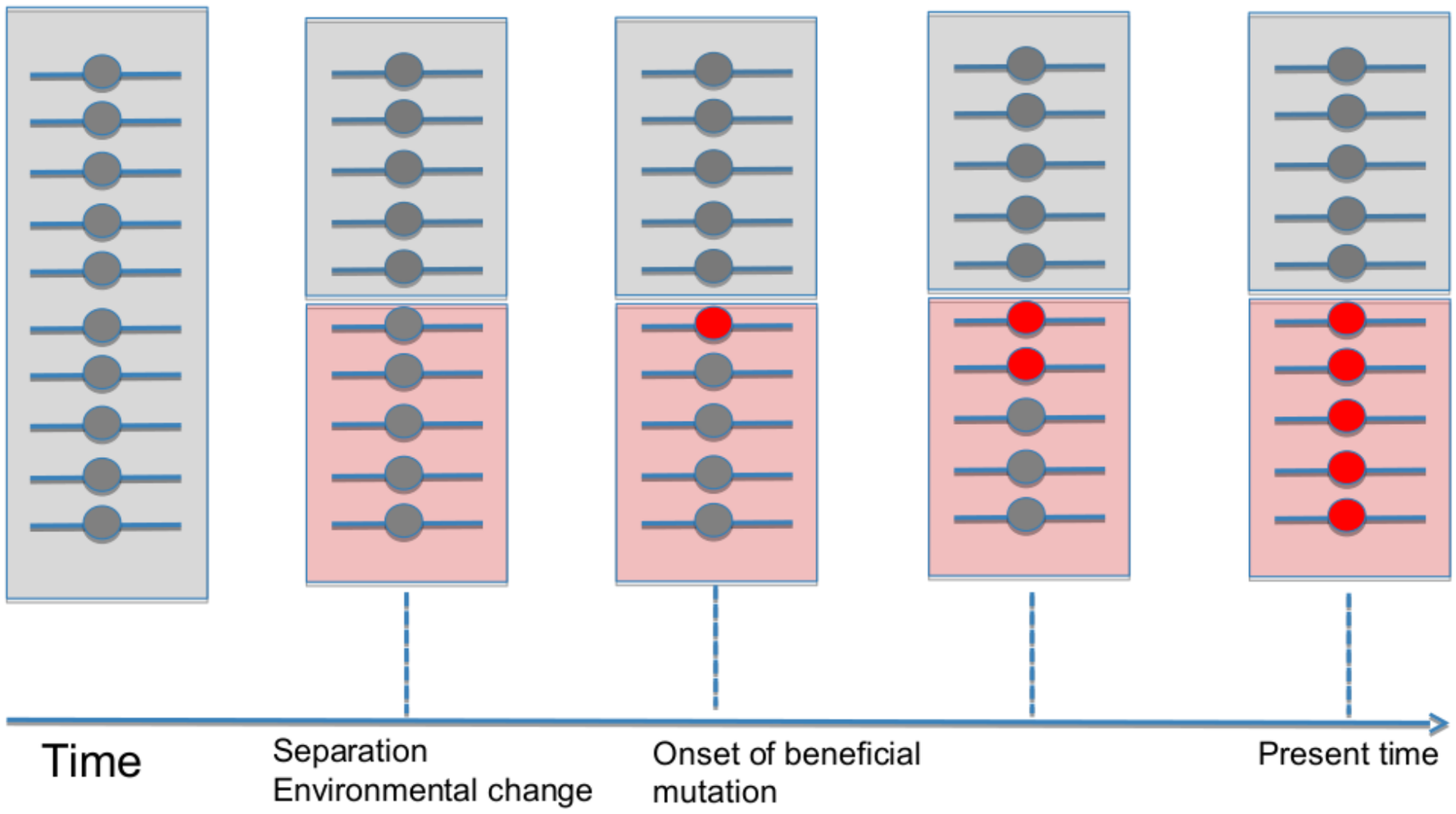
Allele Frequency Differentiation



Allele Frequency Differentiation



Allele Frequency Differentiation



Altitude adaptation in Tibet

Yi et al. 2010

- Low oxygen has a large effect on fitness
- People living in high altitude have high risk birth complications

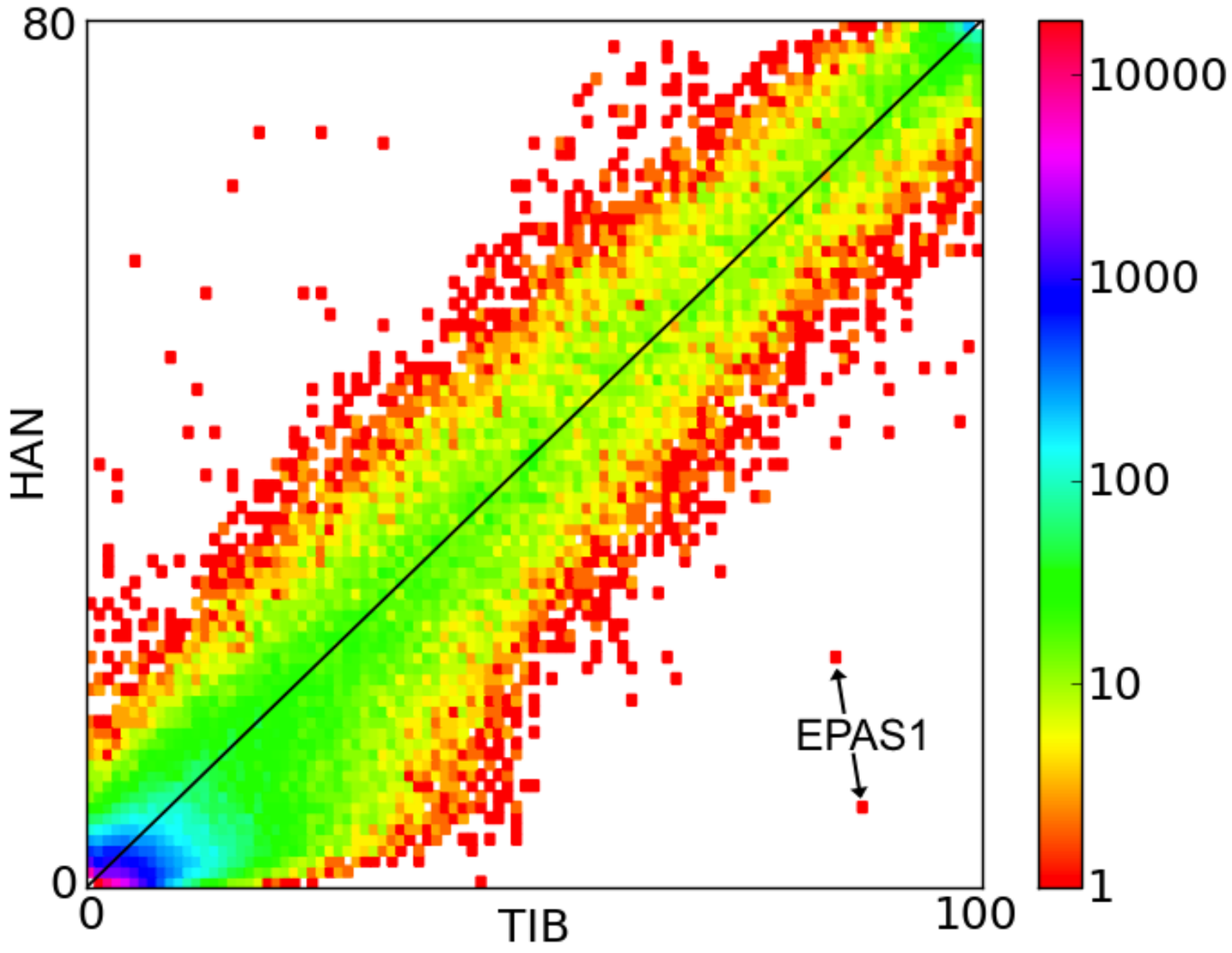
Altitude adaptation in Tibet

Yi et al. 2010

- Full exams of 50 Tibetan individuals: Average ~18X coverage
- Compared to Han Chinese individuals sequenced at ~6X (1000G)
- Estimated joint allele frequencies for each SNP

Altitude adaptation in Tibet

Joint Site Frequency Spectrum



F_{ST}

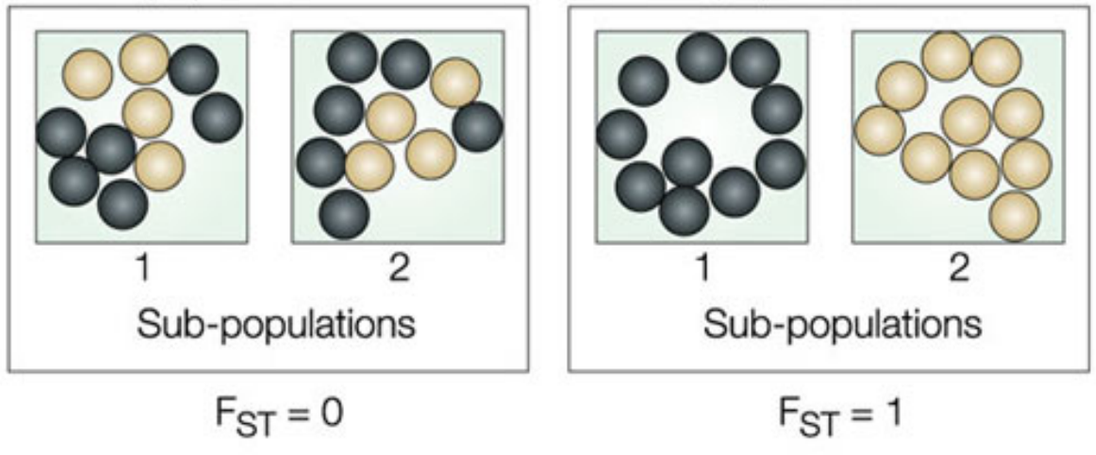
Population genetic differentiation

Common measure for ***quantifying*** populations subdivision.

- Can be thought of as measuring the variations that is **between** subpopulations normalised by the **total** amount of variation

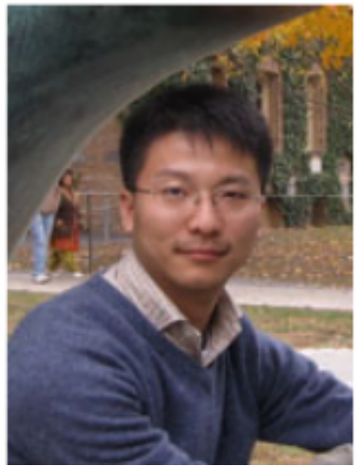
Fst from heterozygosity

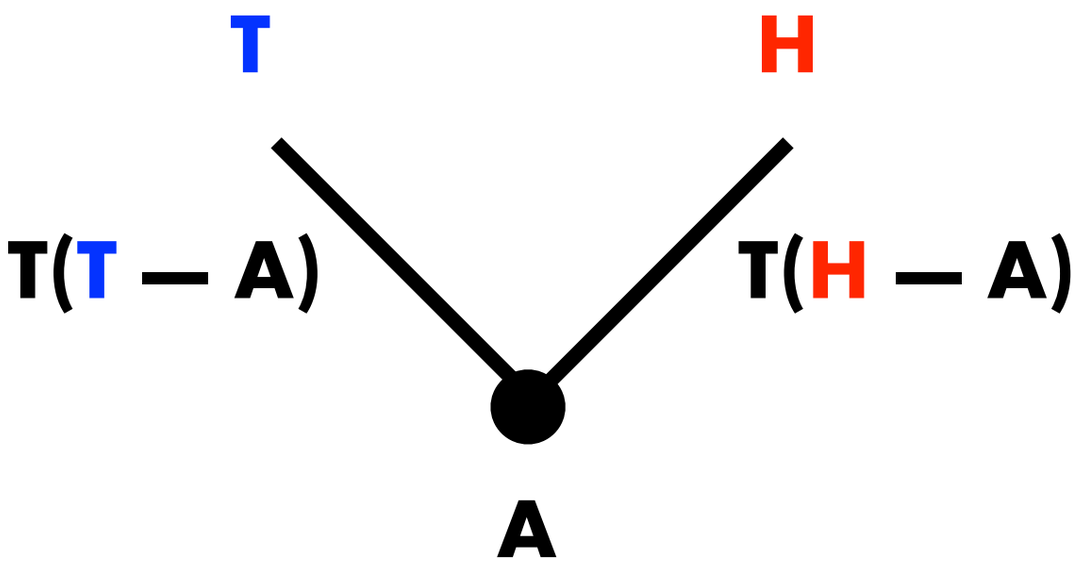
$$F_{st} = \frac{\sigma_B}{\sigma_T} = \frac{H_{total} - H_{subpolulations}}{H_{total}}$$



F_{ST}

Population genetic differentiation





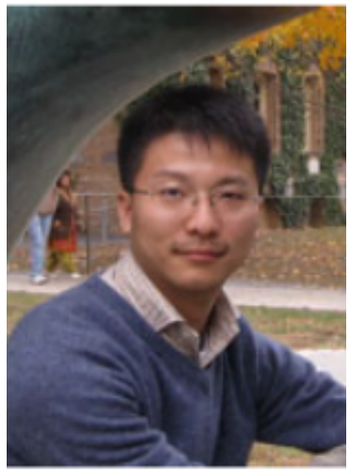
F_{ST} (T — H) ~ T(T — A — H)

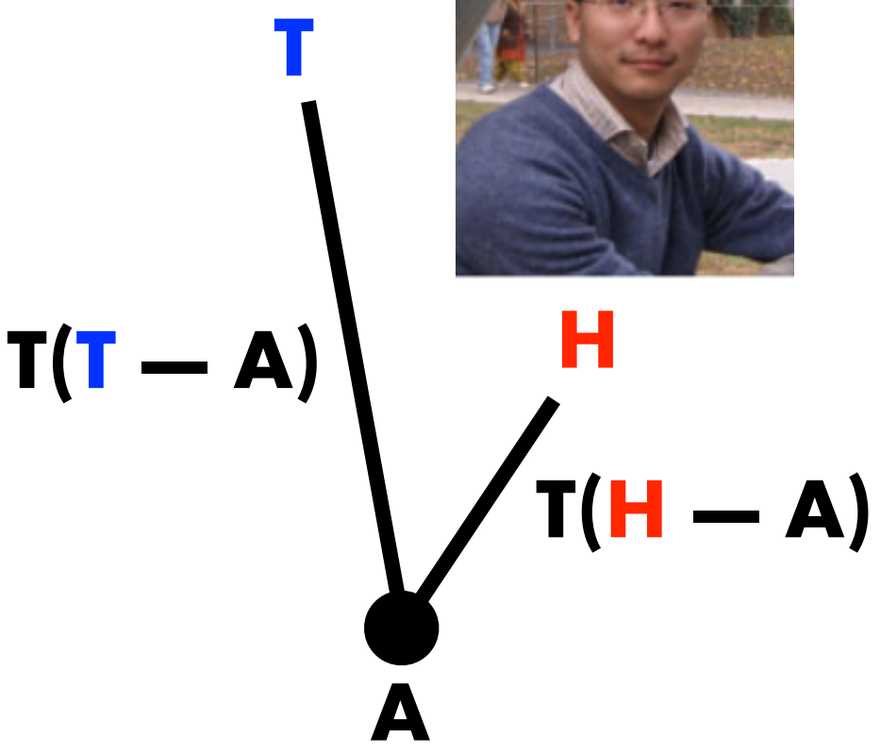
F_{ST}

Population genetic differentiation

$$F_{ST}\,(\mathbf{T} - \mathbf{H}) \sim \mathbf{T}(\mathbf{T} - \mathbf{A} - \mathbf{H})$$

?

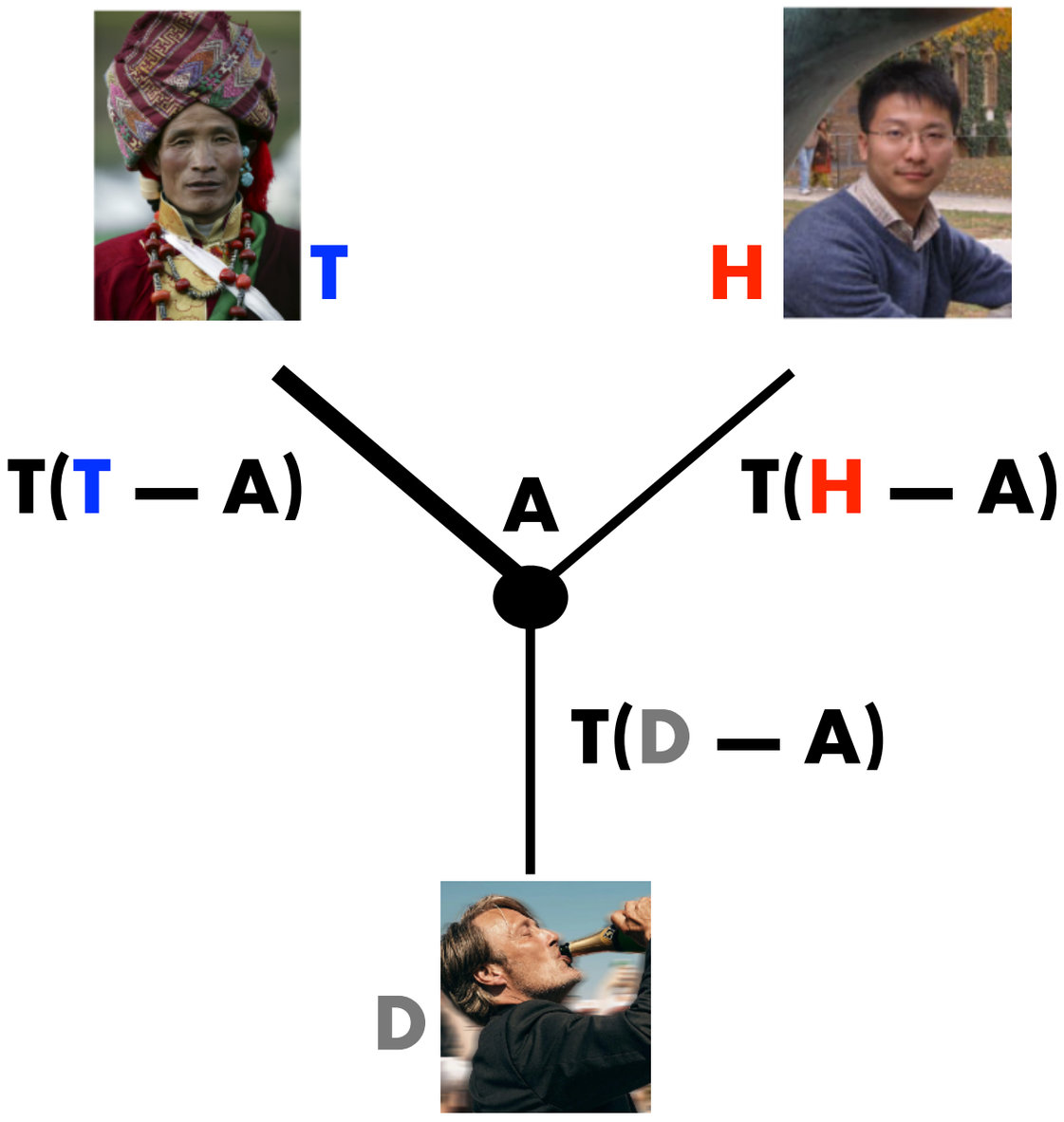




Population Branch Statistic (PBS)

Population genetic differentiation

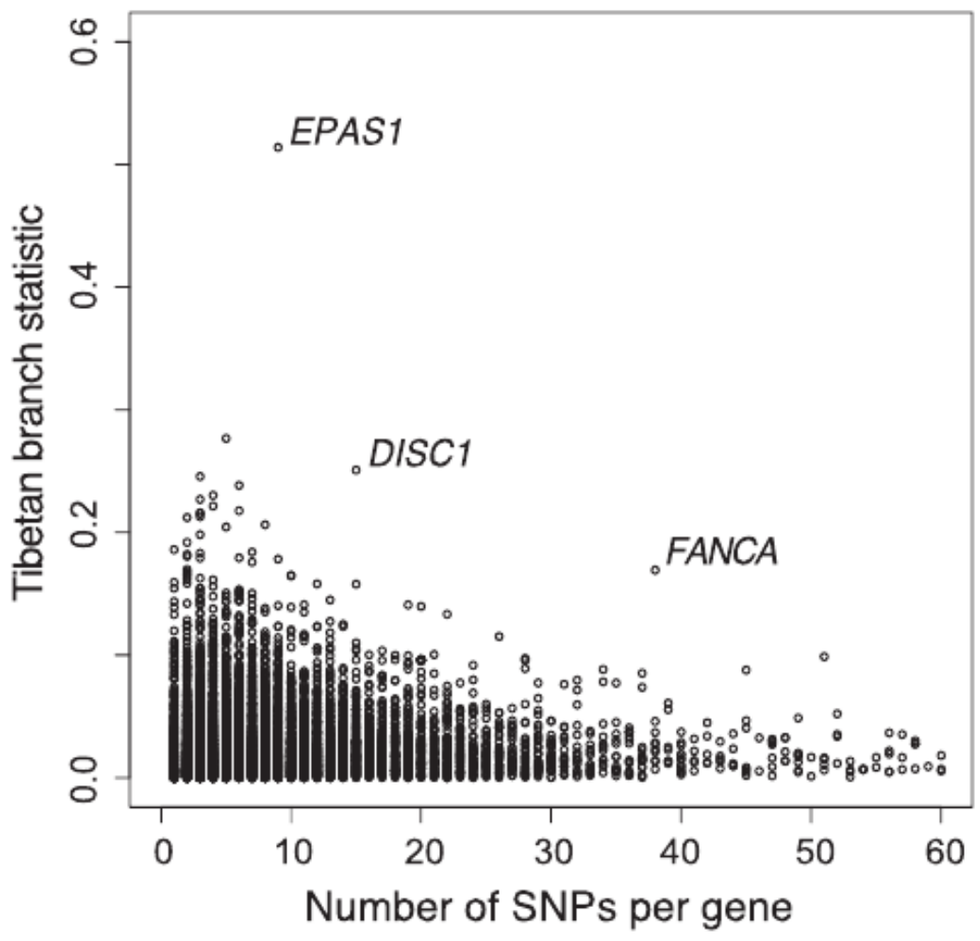
$$PBS = TBS = (T^{TH} + T^{TD} - T^{HD})/2, \qquad T^{AB} = -log(1 - F_{st}^{AB})$$

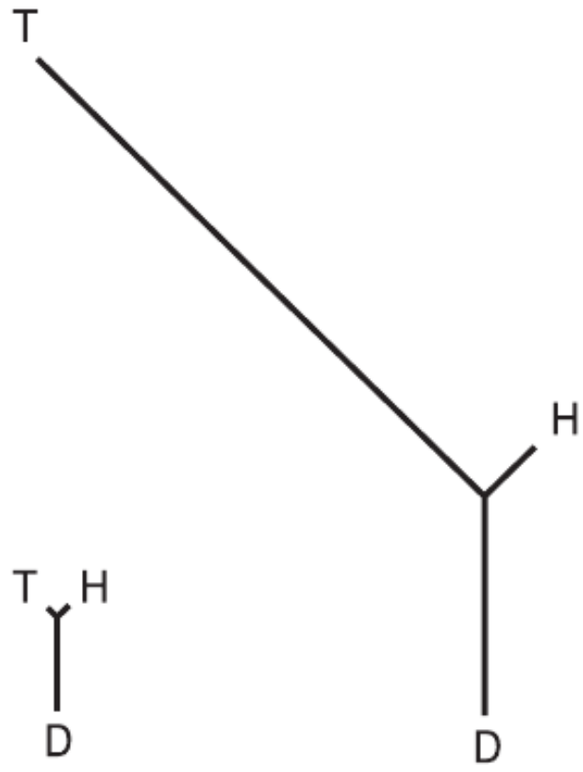


Population Branch Statistic (PBS)

Population genetic differentiation

$$PBS = TBS = (T^{TH} + T^{TD} - T^{HD})/2, \qquad T^{AB} = -log(1 - F_{st}^{AB})$$





Population Branch Statistic (PBS)

Population genetic differentiation

EPAS1 SNP allele frequencies

Allele	Tibetan	Han	Danish
C	0.13	0.9125	1
G	0.87	0.0875	0

Altitude adaptation in Tibet

EPAS1's peak performance in Tibetans (:

EPAS1

- Type of hypoxia-inducible factor
- Active under low oxygen

- Variant of gene confers increased athletic performance (i.e. “super athlete gene”)

Altitude adaptation in Tibet

EPAS1'S PEAK PERFORMANCE IN TIBETANS (:

Conclusion

- Tibetans have adapted to life in high altitude
- A locus in *EPAS1* was found that has undergone strong adaptive selection
- The locus associated with haemoglobin concentrations and erythrocyte counts
- The mutations were introduced by Denisovan introgression
- First example of adaptive introgression in humans

Exercise I

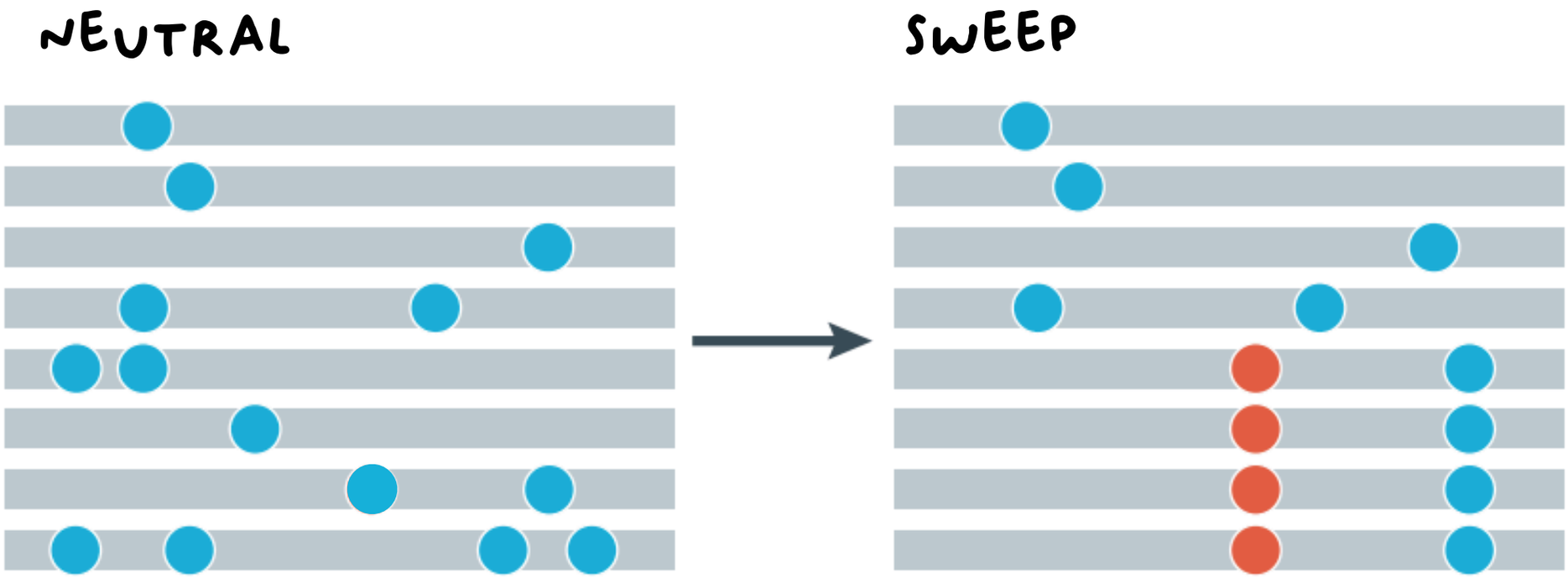
Frequency-based methods

- Let's see how frequency-based methods work on the famous *LCT* example of human adaptation.
    - How do π and Tajima's *D* compare to Fst and PBS?
    - Which method gives the most convincing results to make a case for selection in *MCM6/LCT*?

Haplotype-based methods

Signature of selection

Why does selection affect the SFS?



High pairwise distance

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

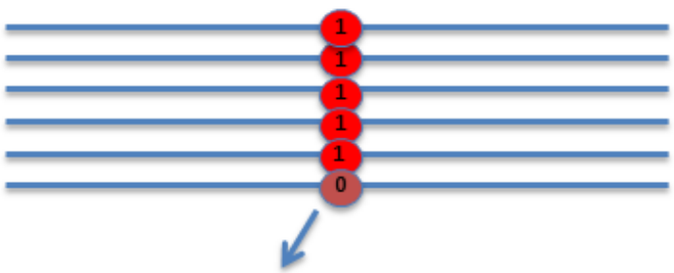
Extended Haplotype Homozygosity

What is EHH?

Extended haplotype homozygosity (EHH):

- Measures the probability that 2 randomly chosen chromosomes carry identical core haplotypes over a certain distance.
- i.e. homozygous at all SNPs for the entire interval from the core region to *x* distance.

Extended Haplotype Homozygosity

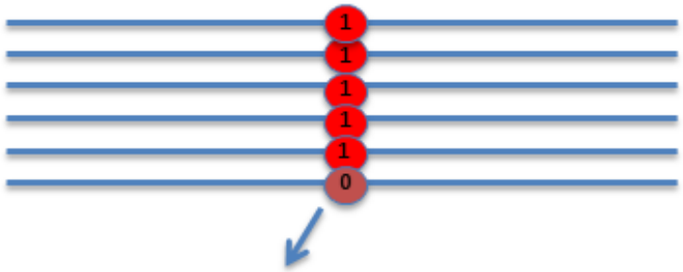


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele

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

Core SNP

Extended Haplotype Homozygosity

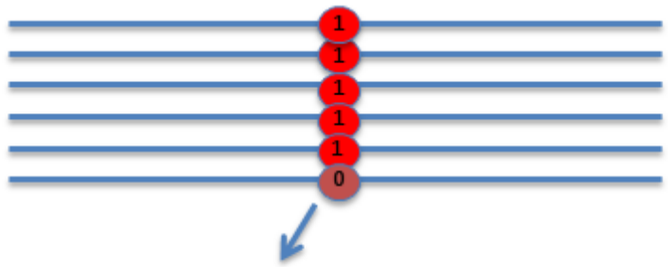


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

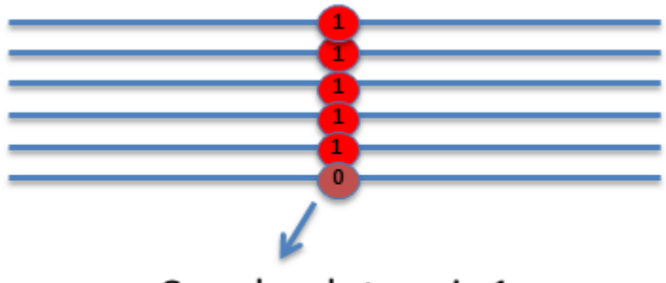


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Sum across all unique haplotypes carrying the core SNP

Extended Haplotype Homozygosity



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

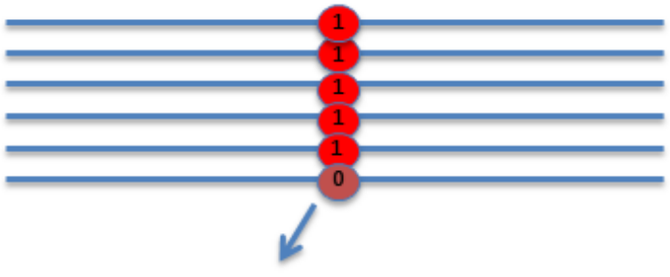
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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)

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

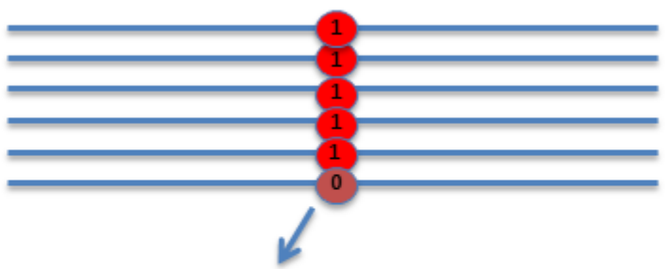
n_h is haplotype frequency of *h*

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)

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

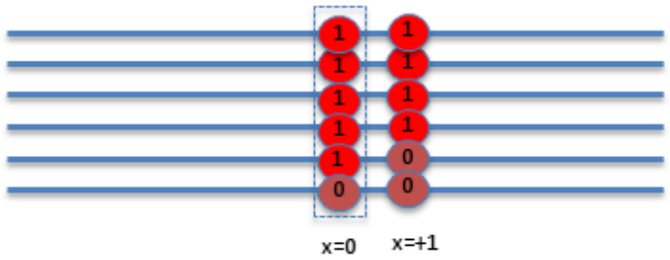
n_h is haplotype frequency of *h*

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

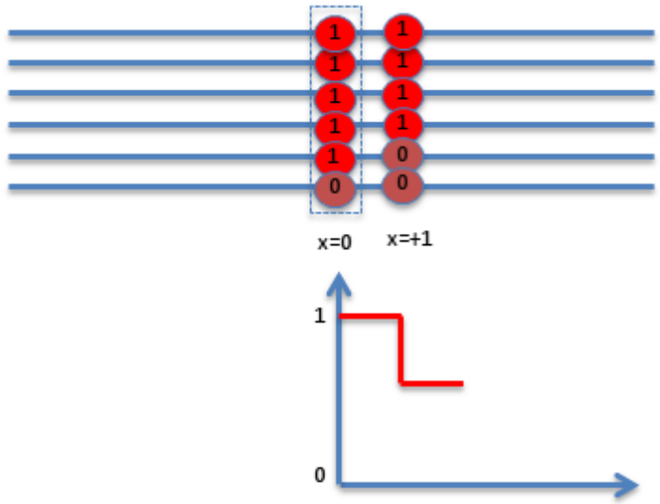


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

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

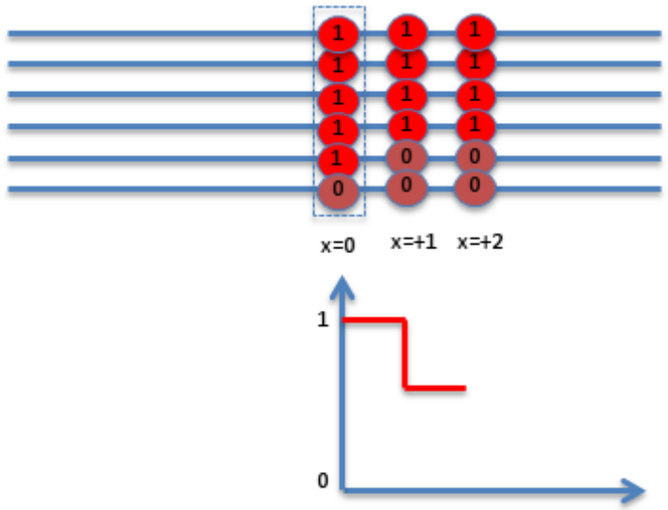
Extended Haplotype Homozygosity



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

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

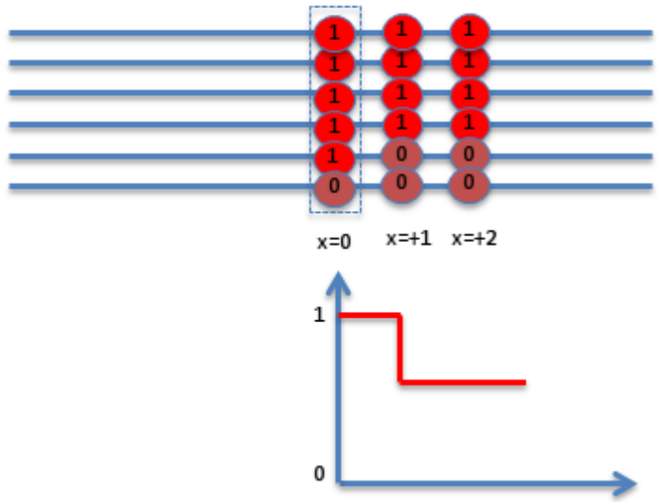
Extended Haplotype Homozygosity



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

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

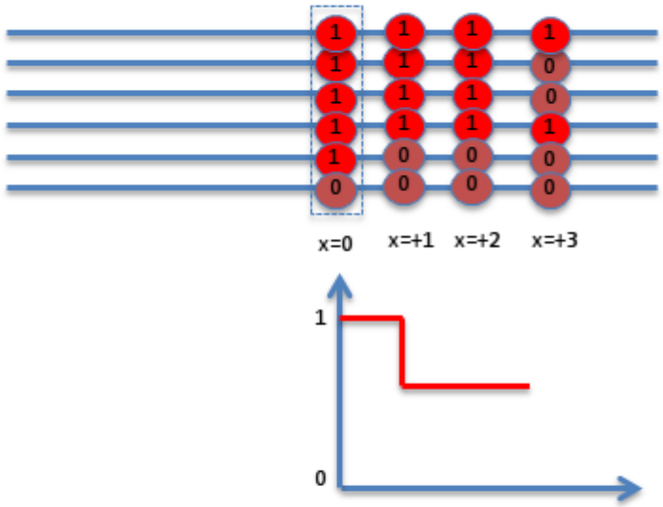
Extended Haplotype Homozygosity



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

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

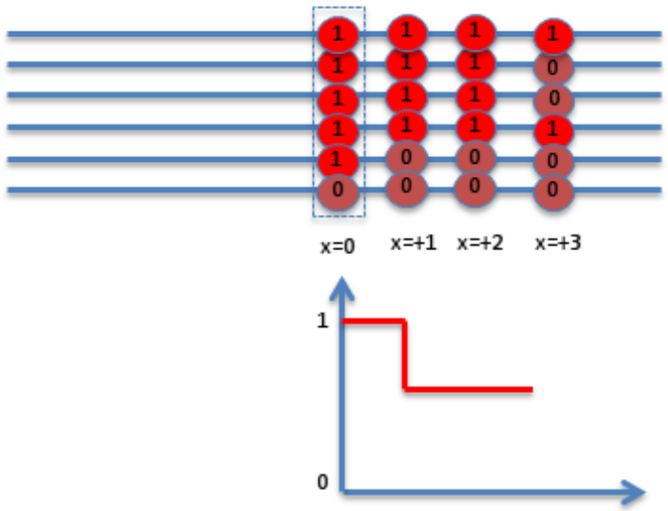
Extended Haplotype Homozygosity



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

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

Extended Haplotype Homozygosity

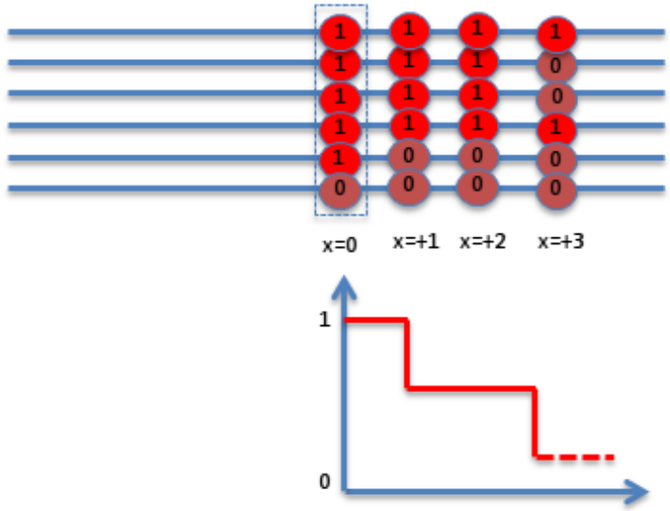


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

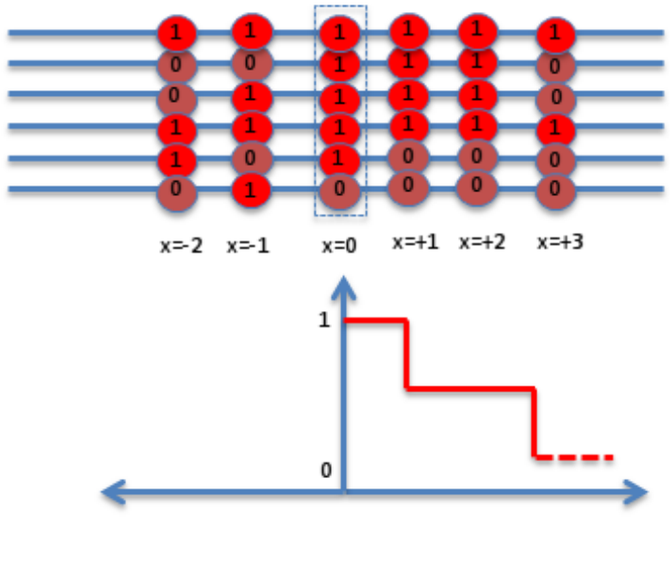


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

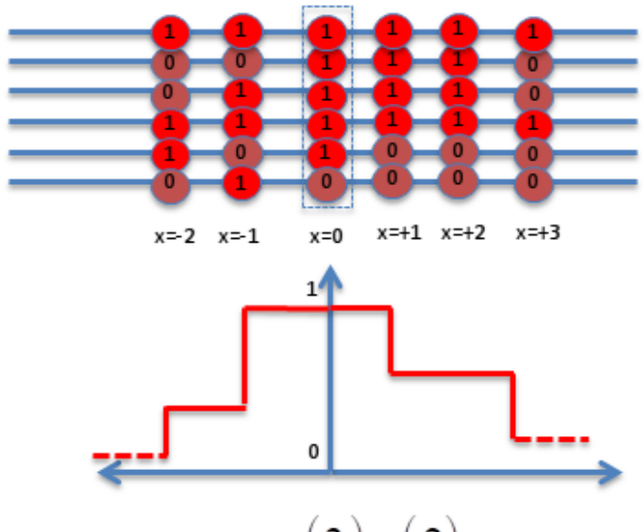


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

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

Extended Haplotype Homozygosity



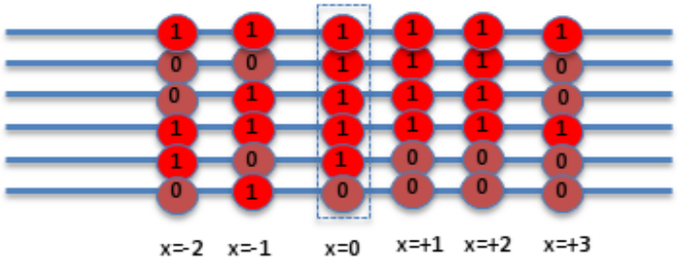
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

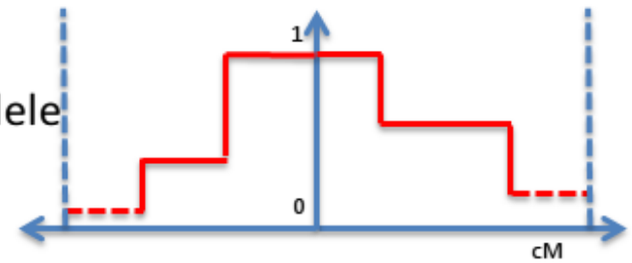
$$EHH_c(x_i = -2) = \frac{\binom{2}{2} + \binom{1}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+0+0}{10} = 0.1$$

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

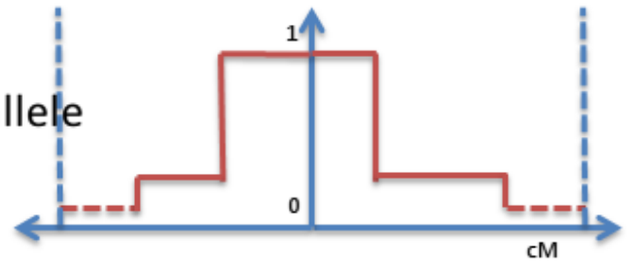
Integrated Haplotype Score (iHS)



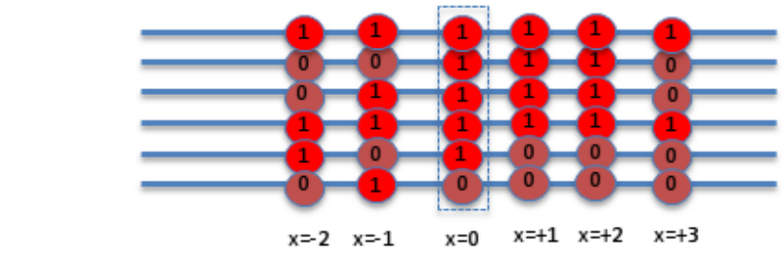
For the derived allele

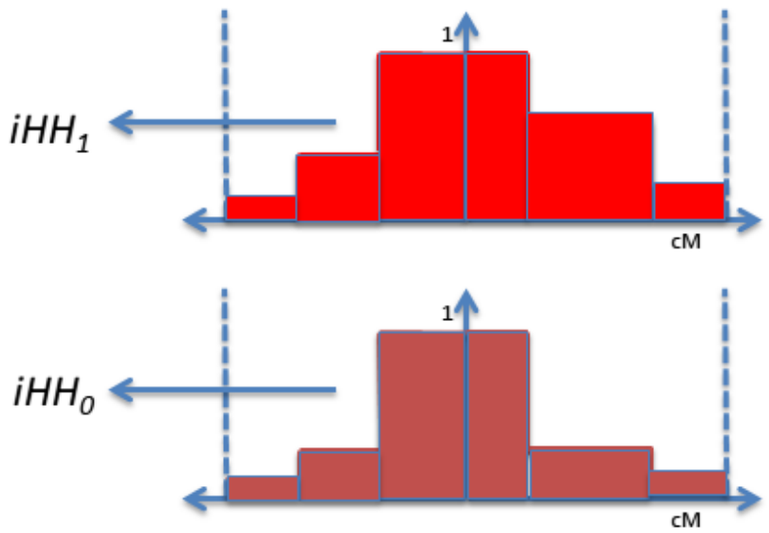


For the ancestral allele



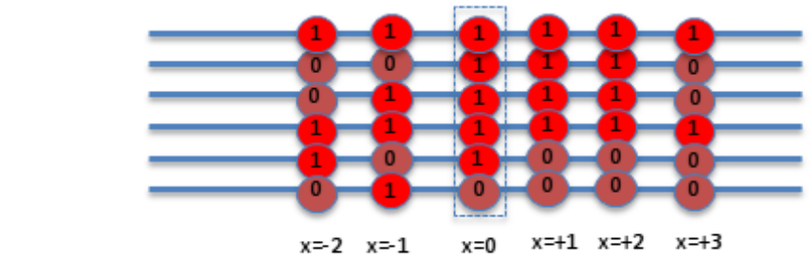
iHS

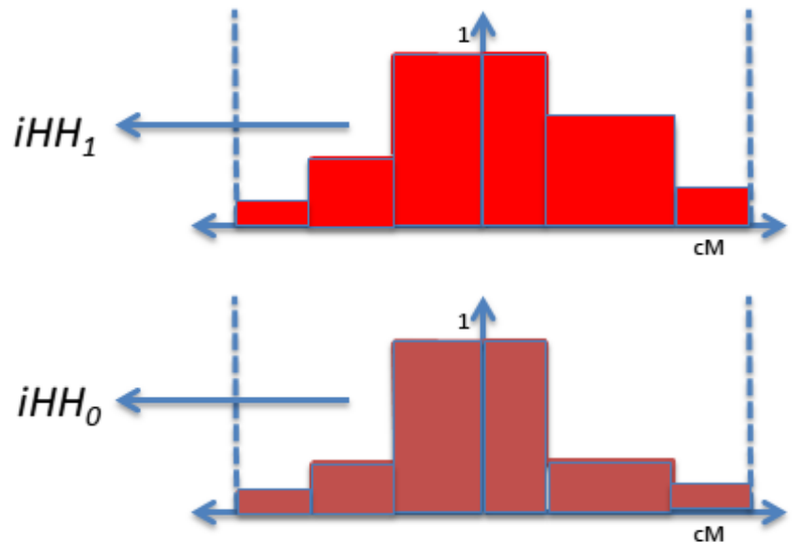




Integrated haplotype homozygosity

iHS





Integrated haplotype homozygosity (*iHH*)

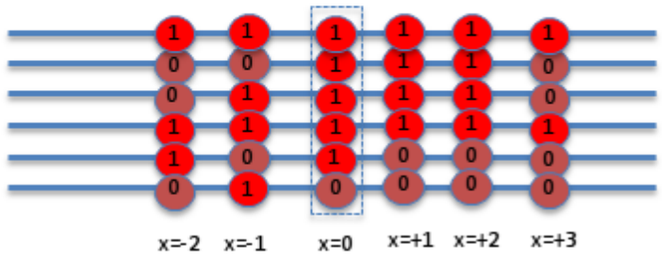
Integrated haplotype score:

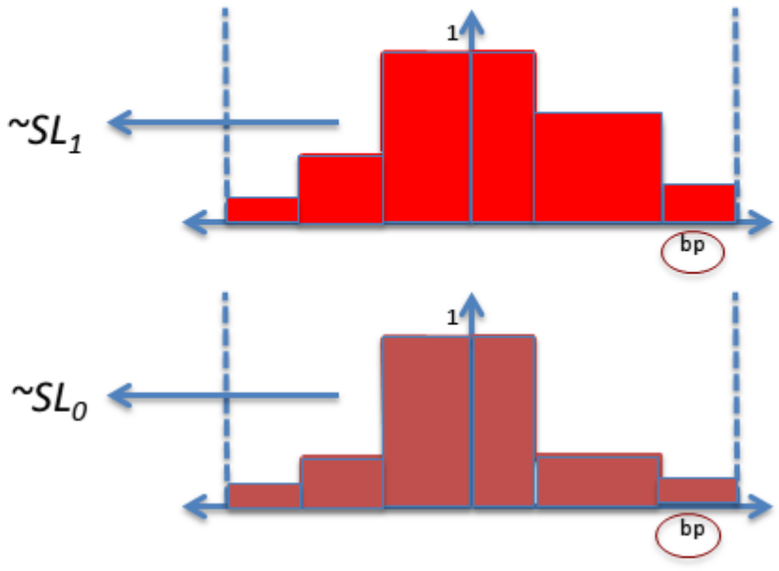
$iHs = ln(iHH_1/iHH_0)$

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

often $|iHs|$ is used

nSL



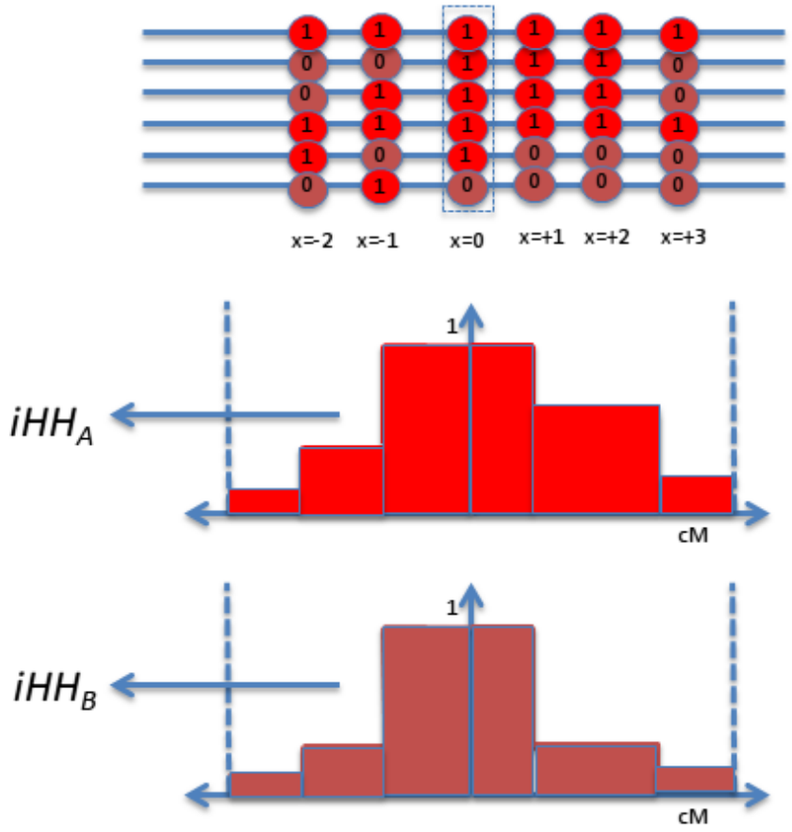


Integration with respect to physical (not genetic) map

$nSL = ln(SL_1/SL_0)$

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

Cross-population EHH (XP-EHH)



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)

Exercise II

Haplotype-based methods

- Let's see how haplotype methods work on the same *LCT* example.
    - How does iHS compare to XP-EHH?
    - Which of these methods (frequency or haplotype-based) render the most convincing results?

Take home messages

Selection Methods

	Frequency-based	Haplotype-based
Single-population	• Tajima's D • Tajima's π • Watterson Theta	• iHS • nSL
Comparative	• jSFS/2D-SFS • F_{ST} • PBS	• XP-EHH • XP-nSL

Take home messages

Selection Methods

	Frequency-based	Haplotype-based
S		
pop		
Com		

Take home messages

Conclusions

FEBS Letters

Review Article | Free Access

How well do we understand the basis of classic selective sweeps in humans?

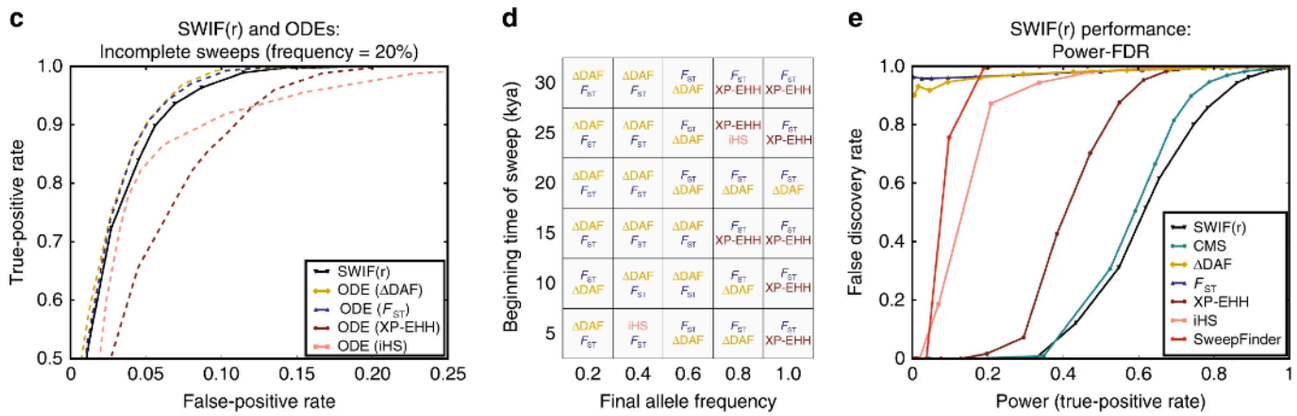
Michał Szpak, Yali Xue, Qasim Ayub, Chris Tyler-Smith

First published: 22 May 2019 | https://doi.org/10.1002/1873-3468.13447 | Citations: 15

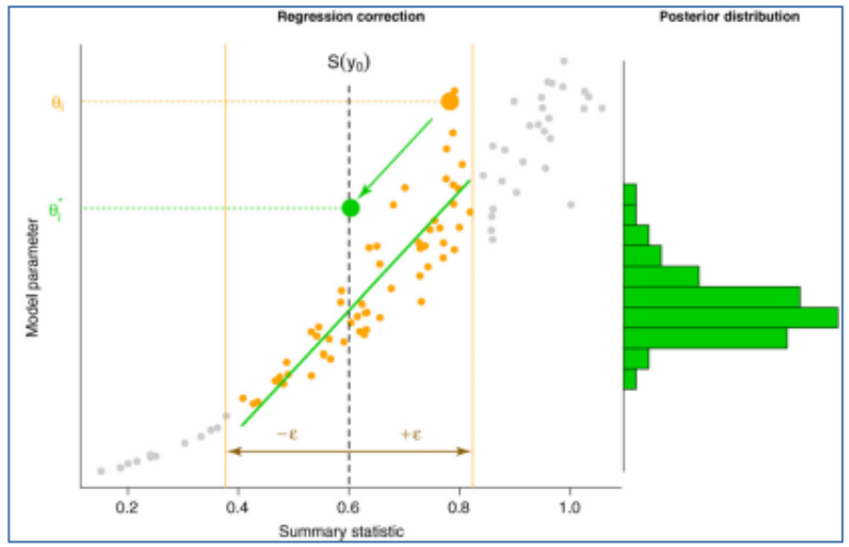
Edited by Claudia Bank

Recent advances to detect selection

1. Composite scores (Sugden et al. 2018)

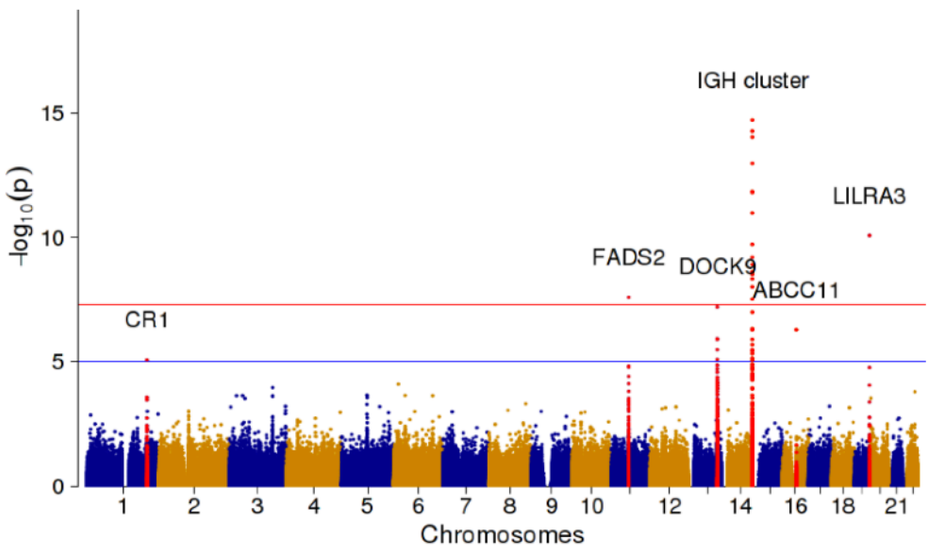


2. Simulations-based (rejection, ABC)



3. Unsupervised machine learning

(Meisner et al. 2021)



4. Supervised machine learning

(SVM, Schrider & Kern 2018)