Queen Mary
University of London

Inference of natural selection from sequencing data

(at the intra-species level)

Matteo Fumagalli

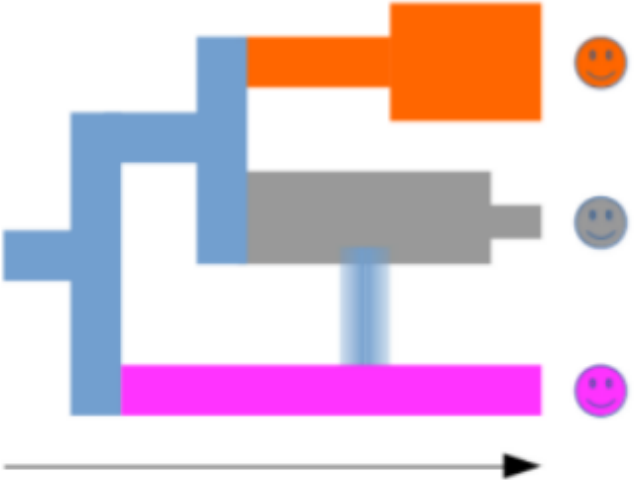
Intended Learning Outcomes

At the end of this session you will be able to:

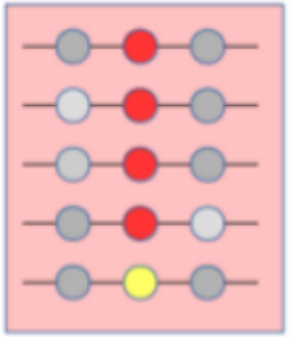
- list commonly used methods to detect selection
- calculate various summary statistics
- understand main confounding factors to neutrality tests
- assess statistical significance of tests

Whole genome sequencing

Demographic history



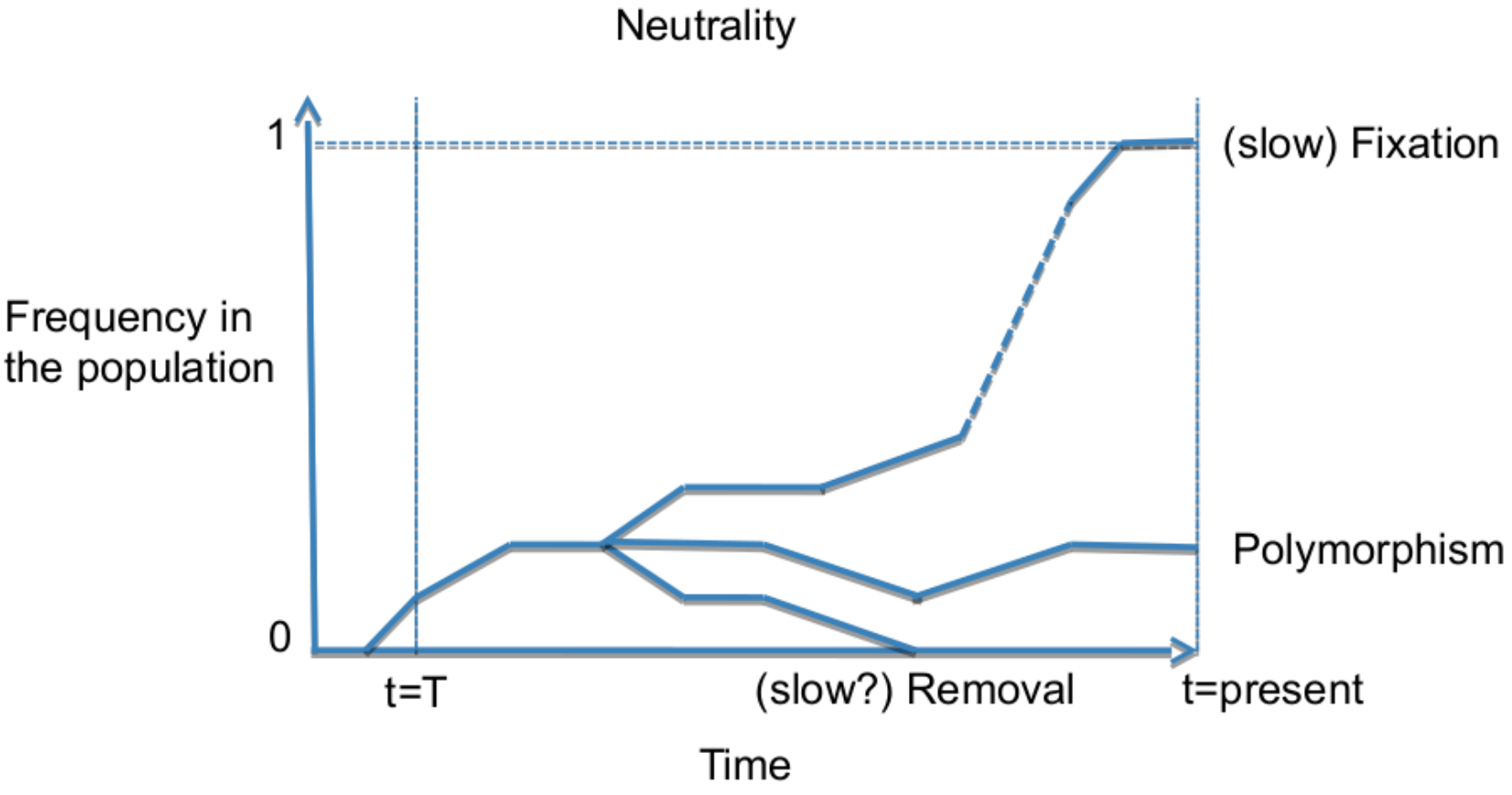
Natural selection



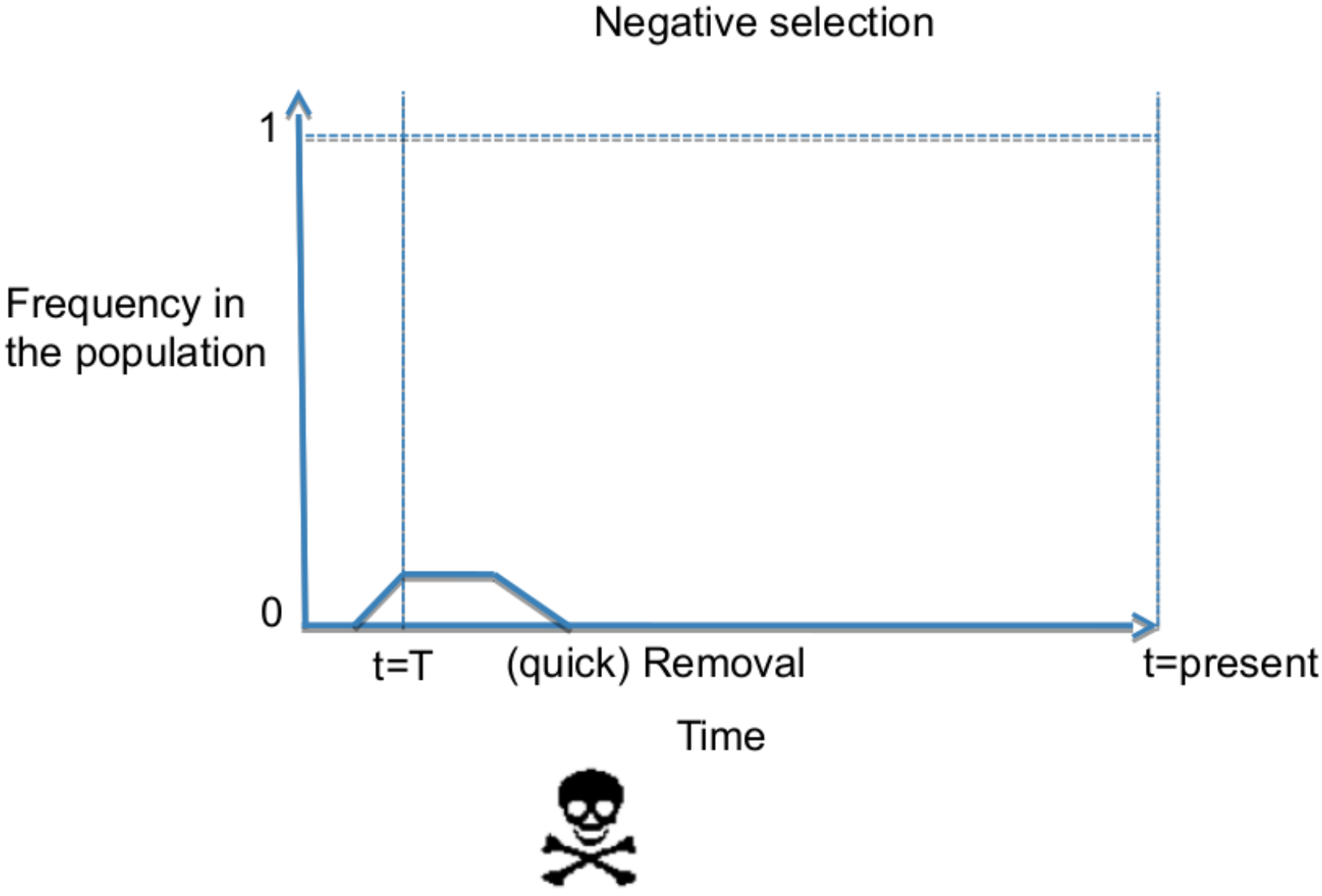
Natural selection

- Heritable traits that increase the fitness of the become more common.
- Sites targeted by natural selection are likely to harbour functionality.
- Mutations arise randomly and evolve according to their effect on the fitness of the carrier.

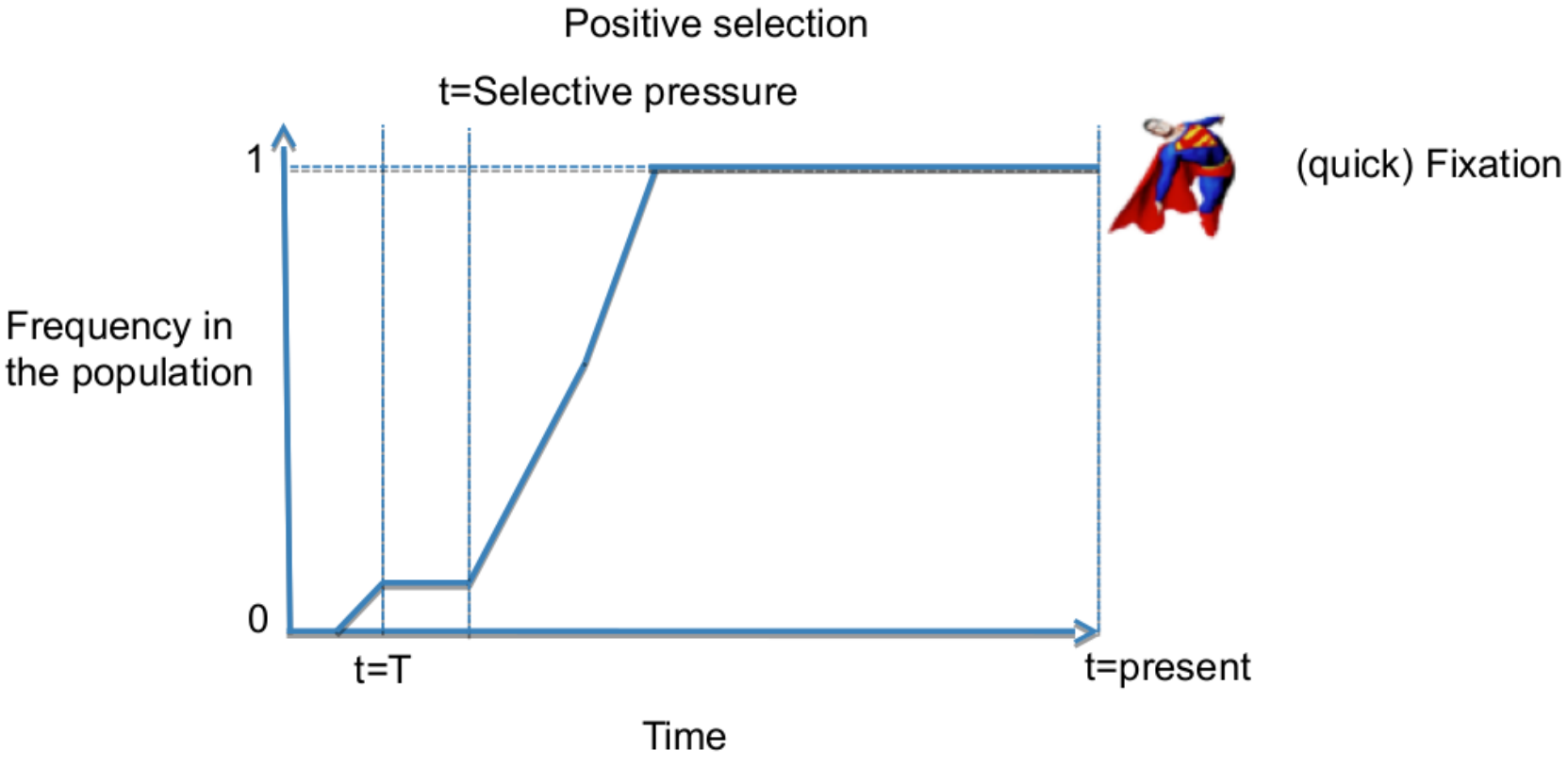
Allele frequency trajectory



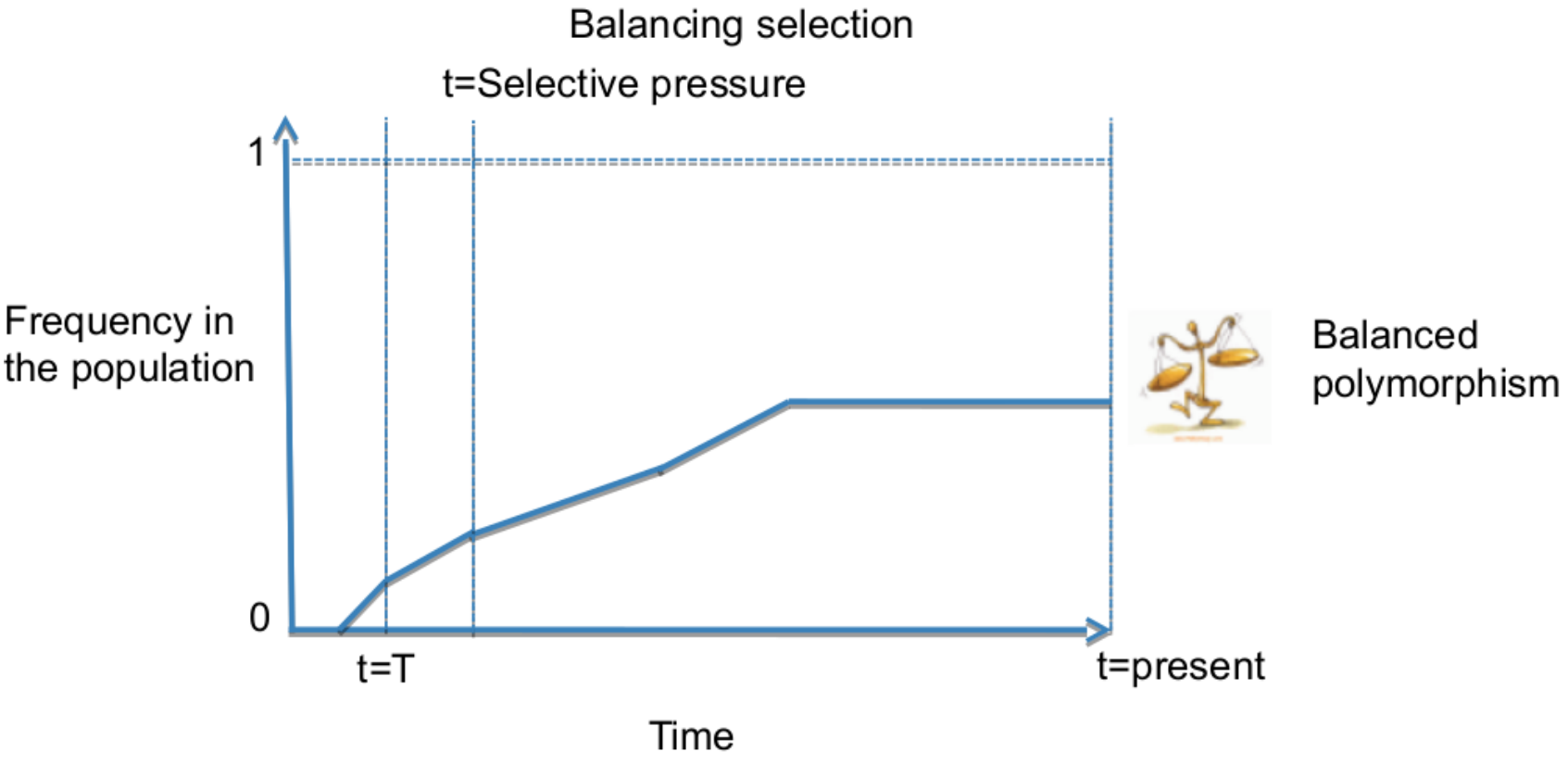
Allele frequency trajectory



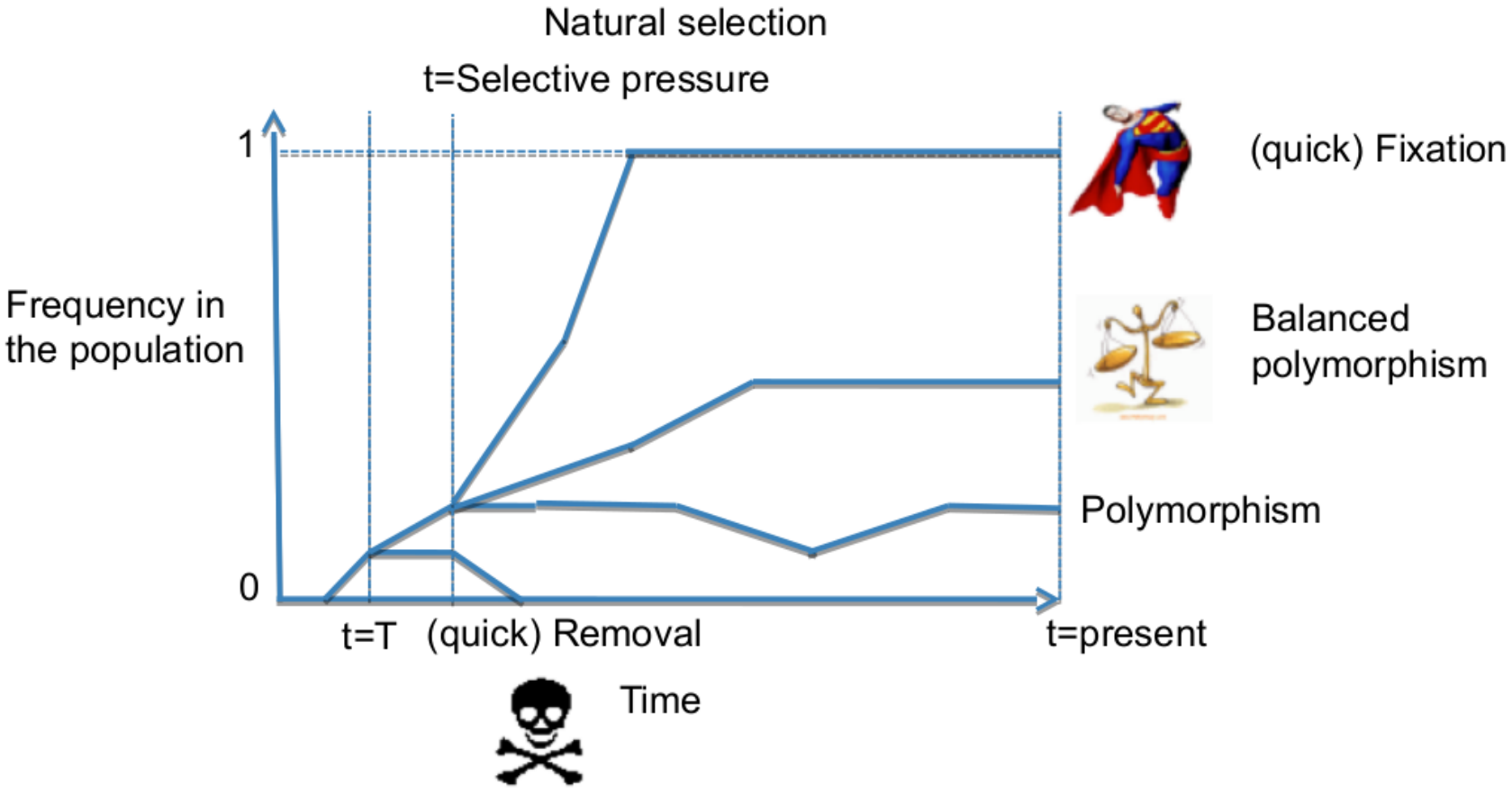
Allele frequency trajectory



Allele frequency trajectory



Allele frequency trajectory



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.

(slide from Anders)

Testing for natural selection

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

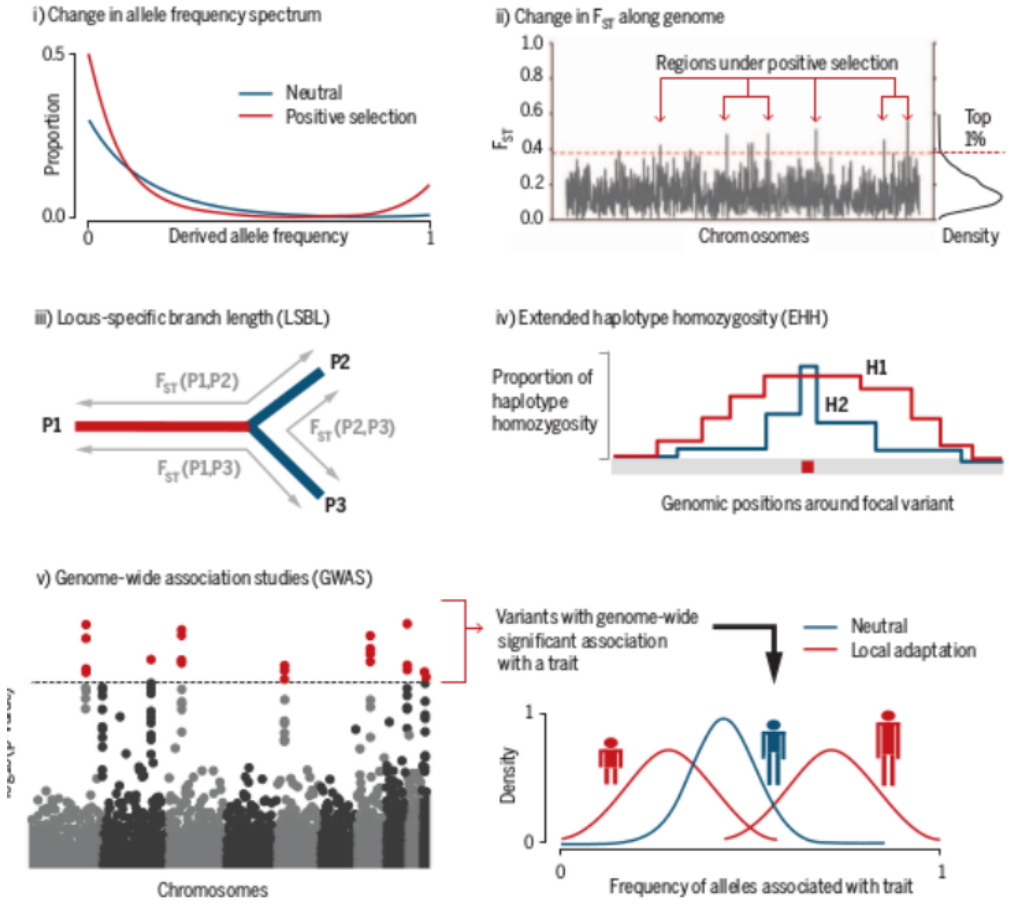
Testing for natural selection

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

- use information from the surrounding genomic region
- use information from multiple species/populations
- perform selection experiments
- use external information: candidate genes/biological knowledge,functional categories, association to phenotypes

(slide from Anders)

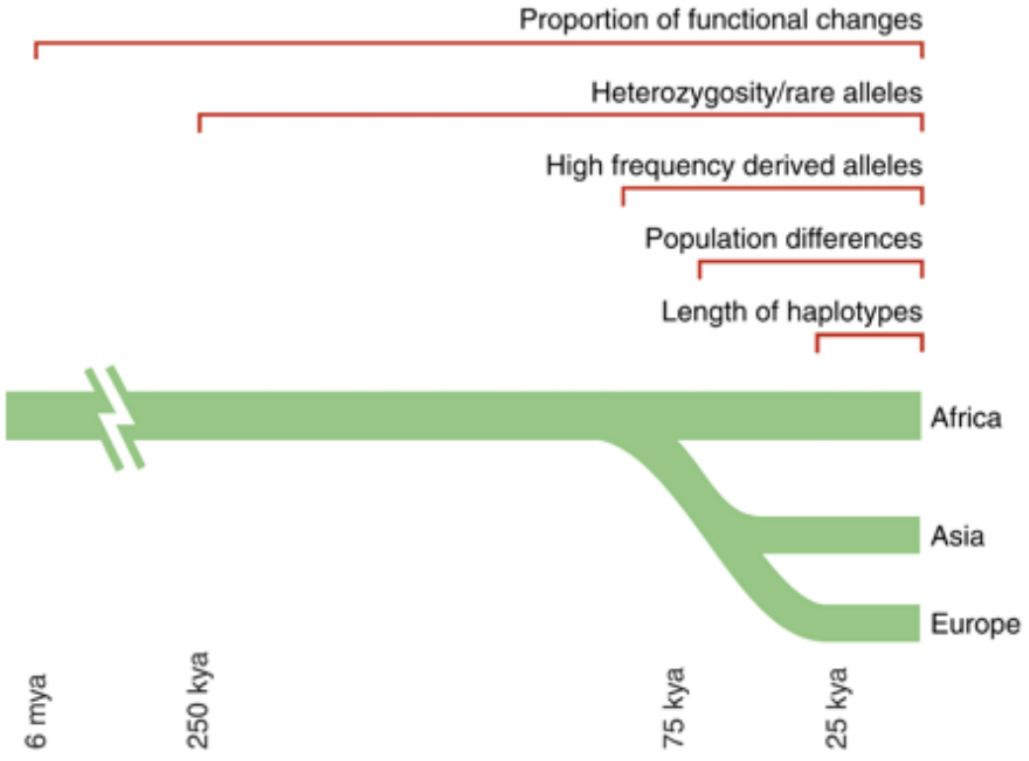
Common methods to detect selection



(slide from Anders)

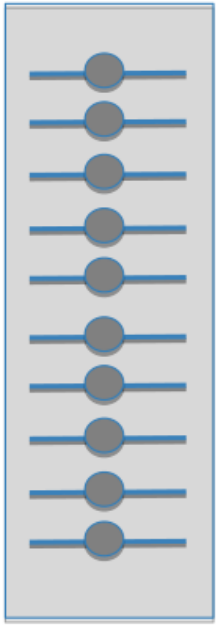
Detect recent selection

within species / using shared variation

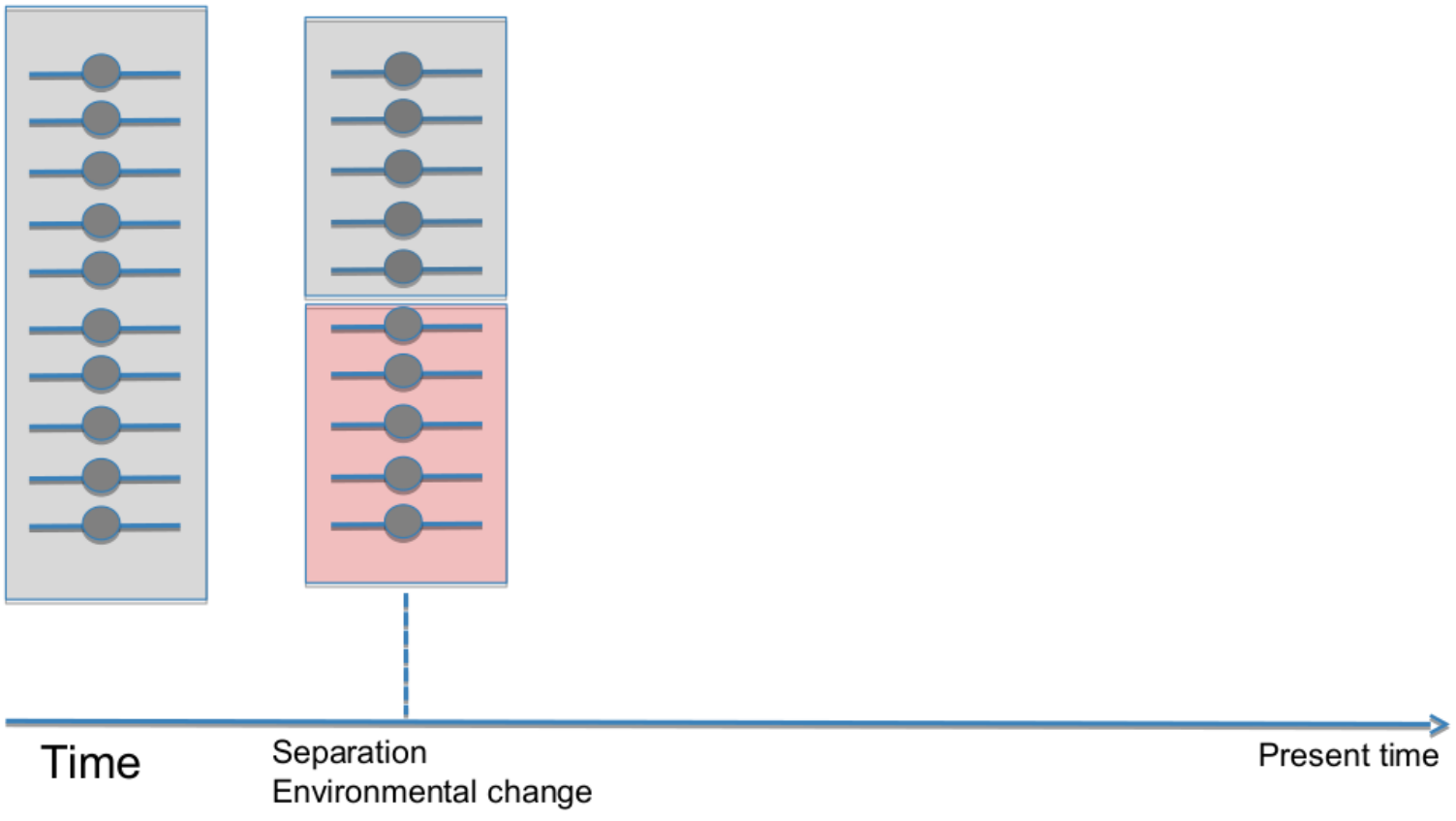


Sabeti et al. 2006 Science

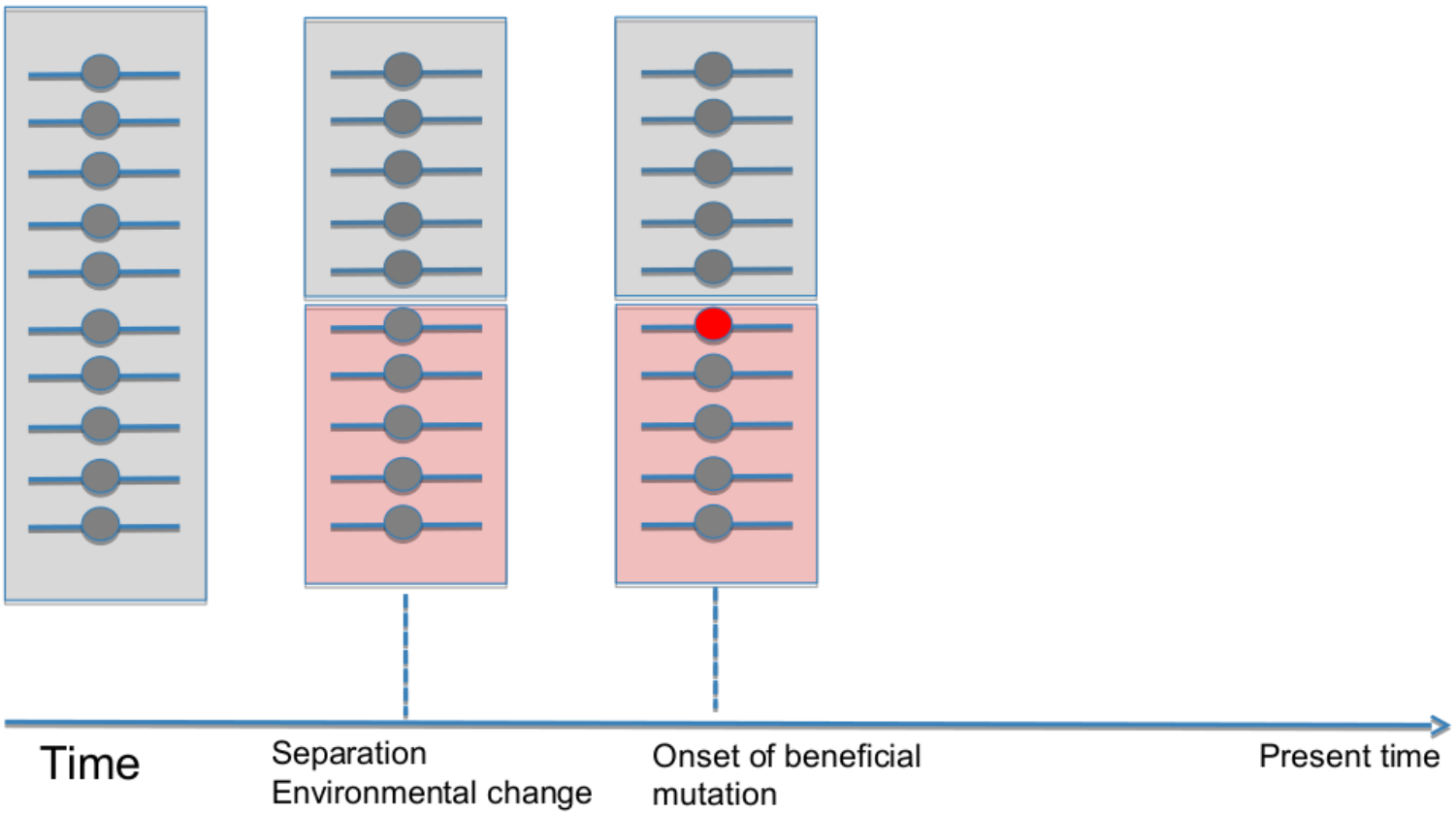
Allele frequency differentiation



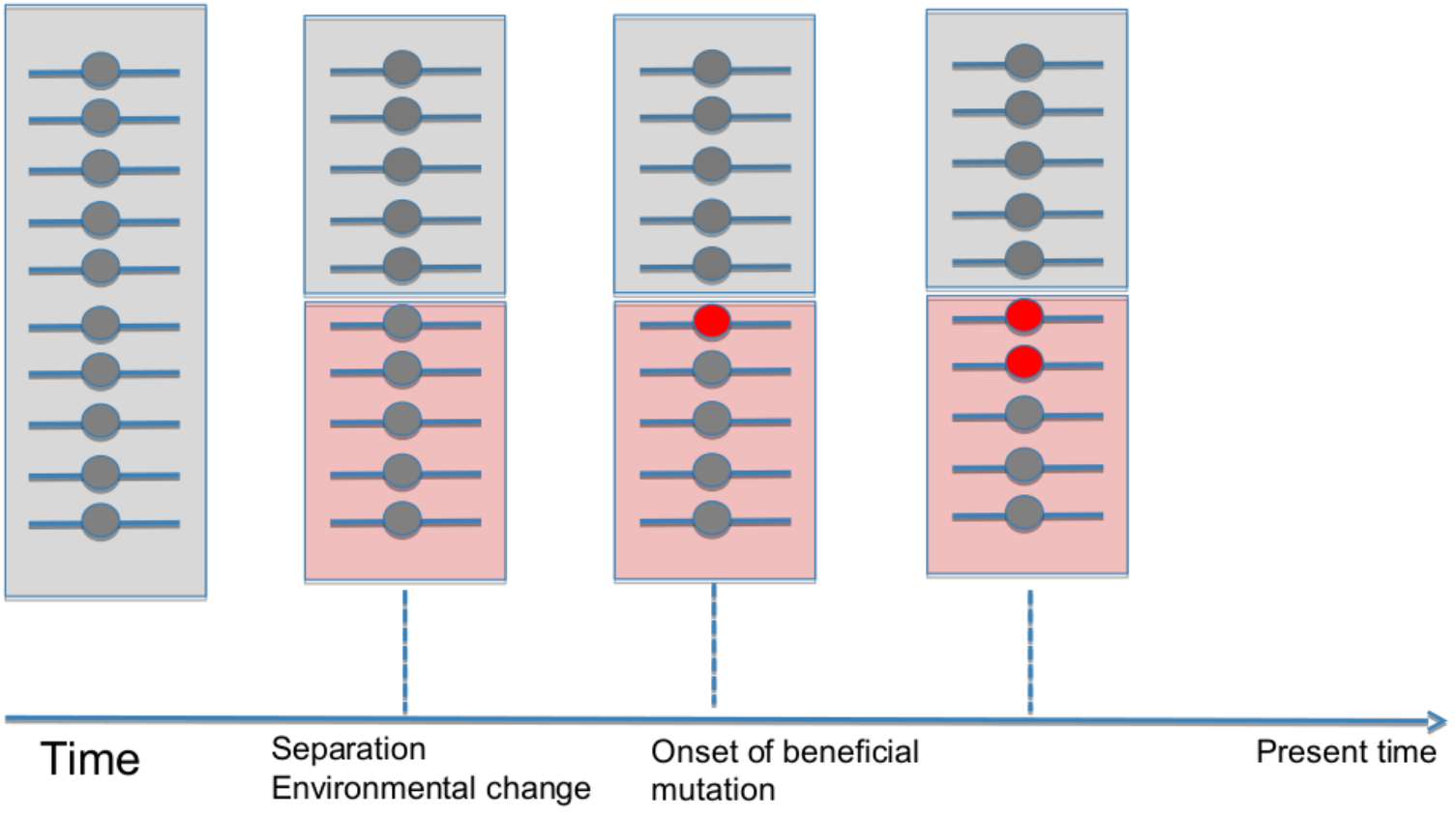
Allele frequency differentiation



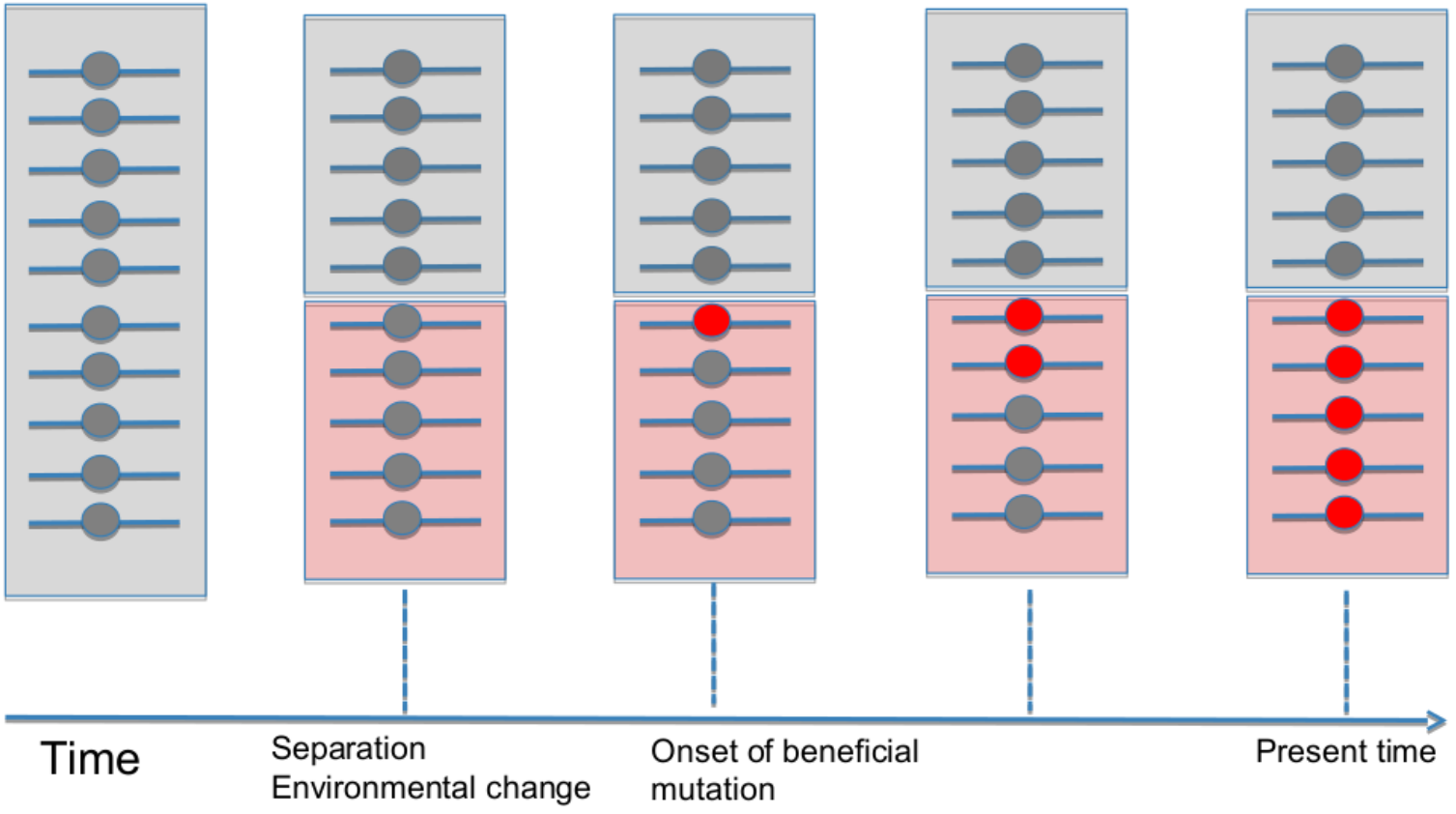
Allele frequency differentiation



Allele frequency differentiation



Allele frequency differentiation



F_{ST}

Common measure for quantifying population subdivision.

$$F_{ST} = H_B / (H_W + H_B)$$

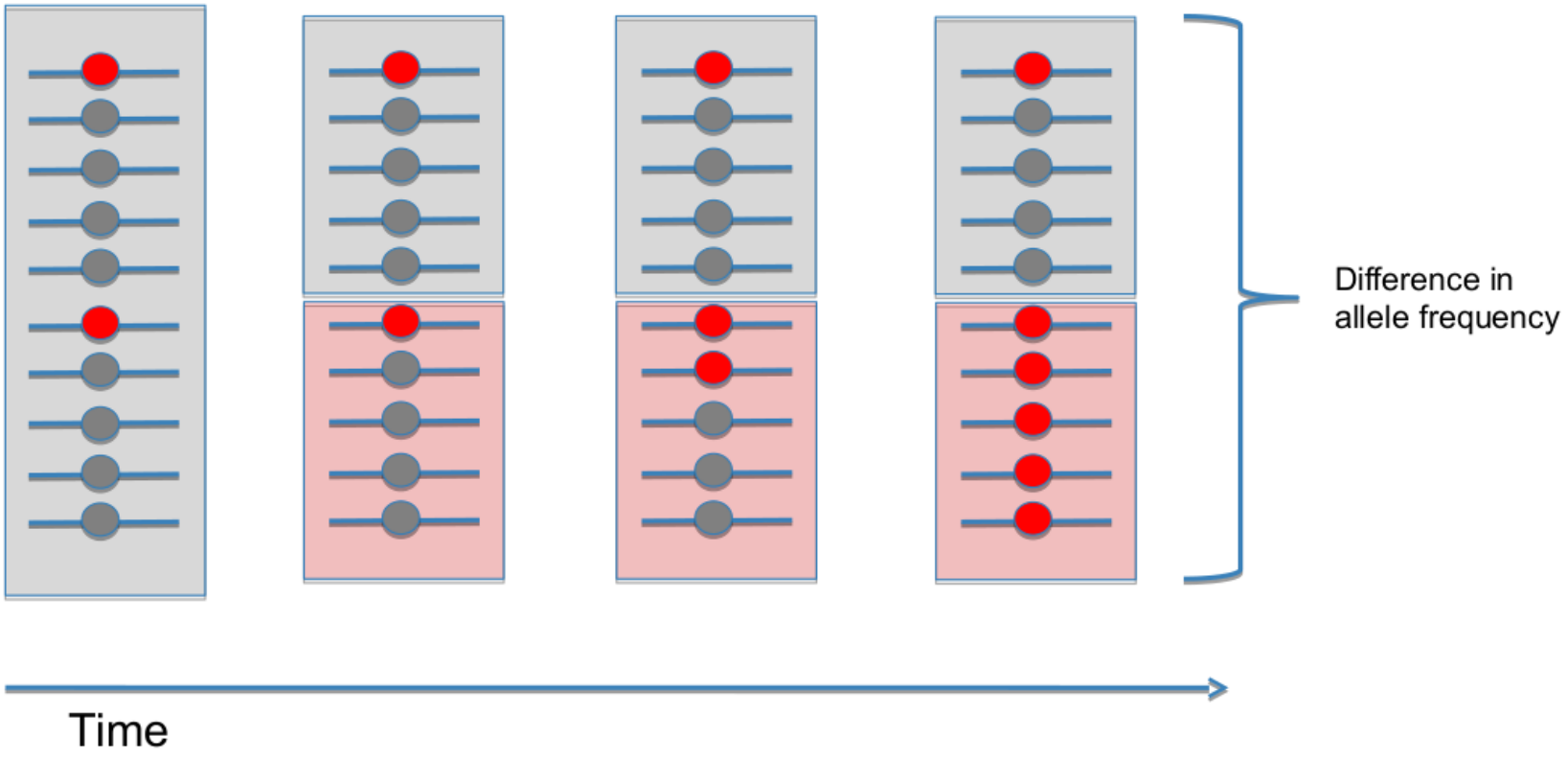
H_B: between populations

H_W: average within populations

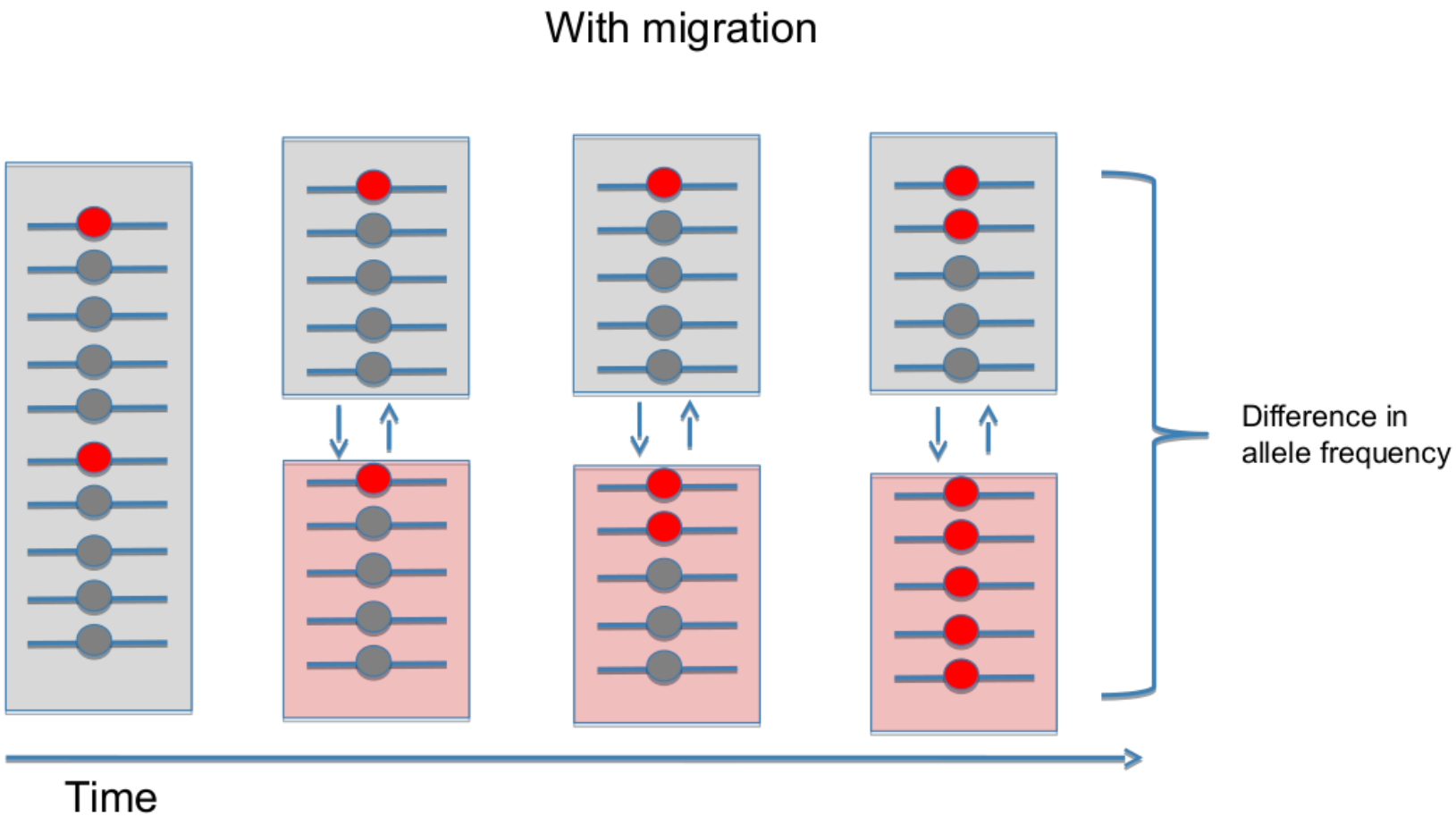
- if $H_W << H_B 0$ then $F_{ST} \sim 1$
- if $H_B = 0$ then $F_{ST} = 0$

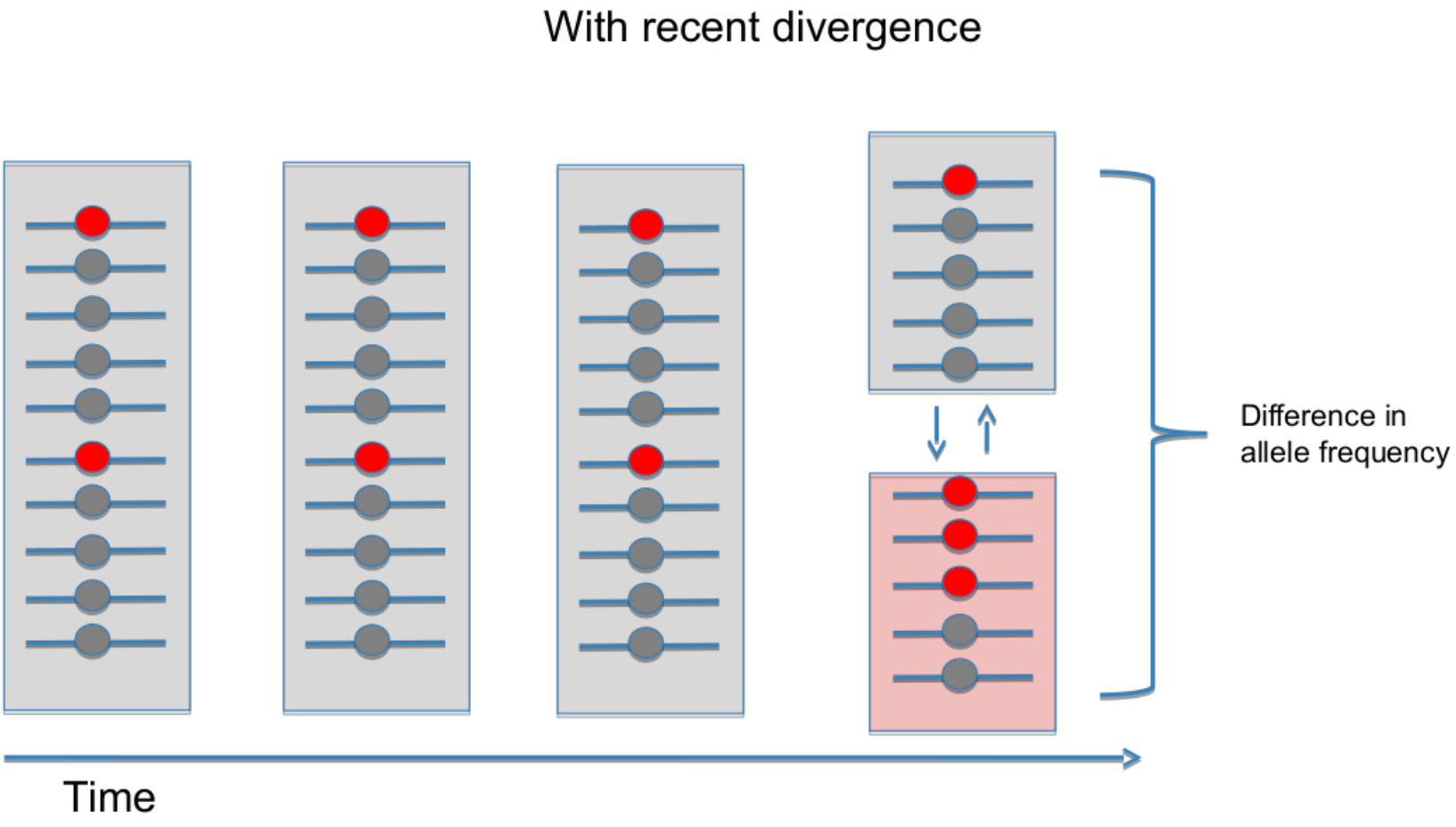
Allele frequency differentiation

From standing variation

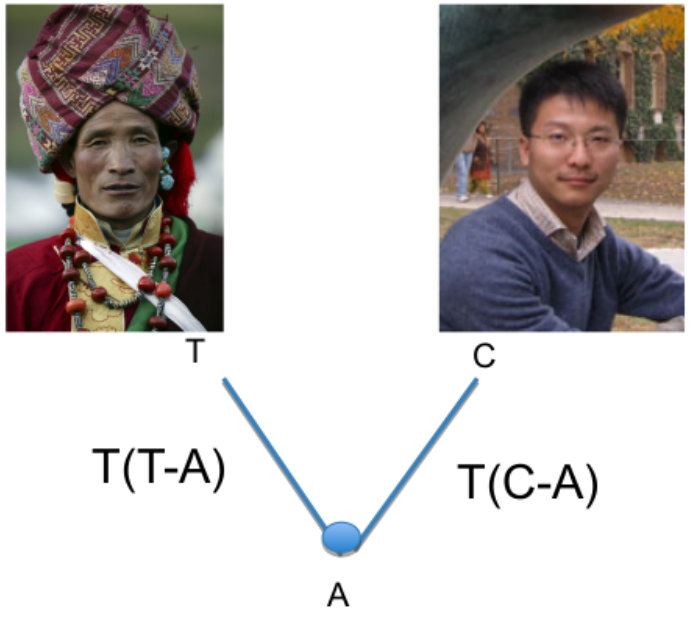


Allele frequency differentiation




With recent divergence
Difference in allele frequency
Time

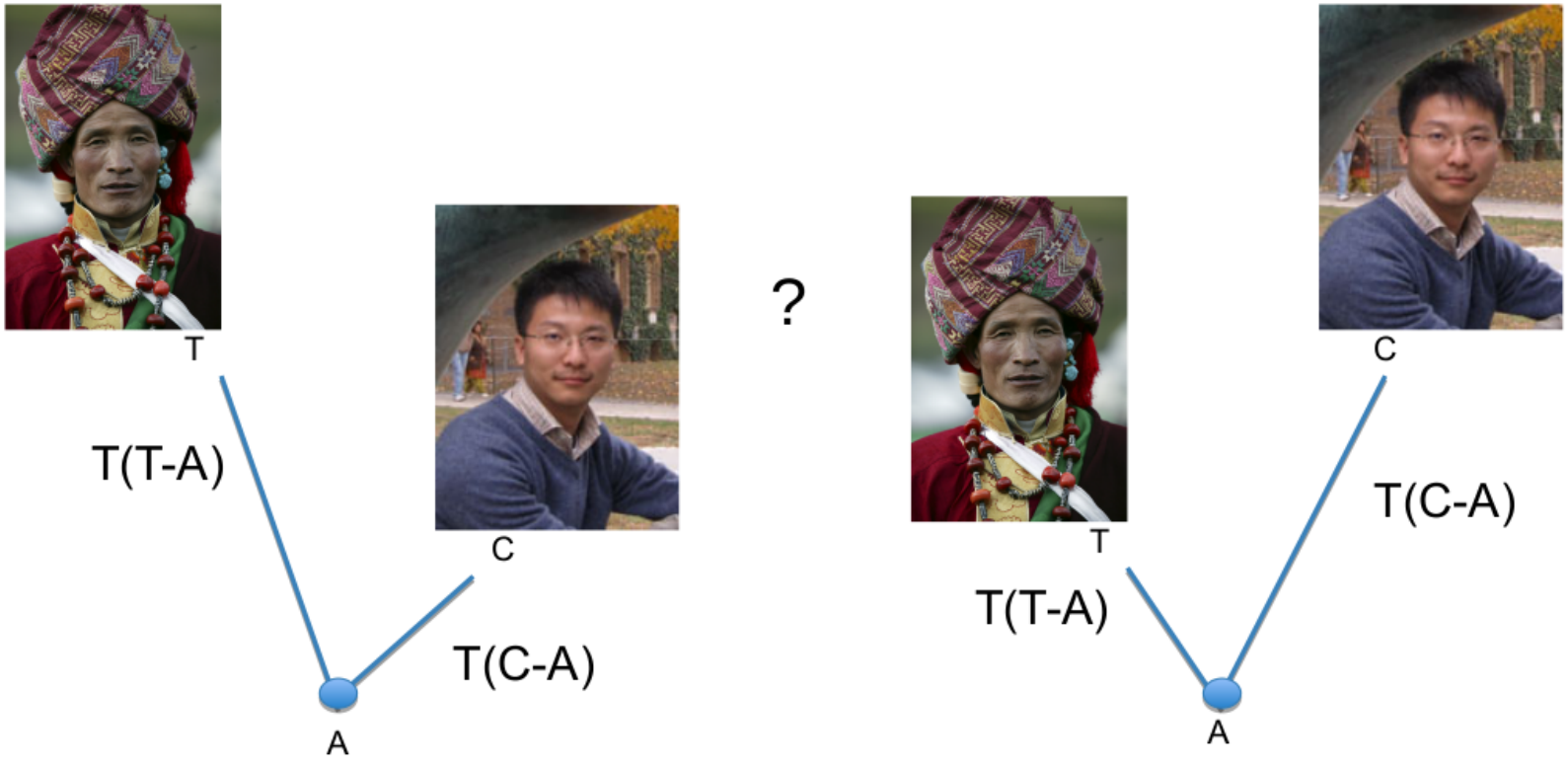
Population genetic differentiation



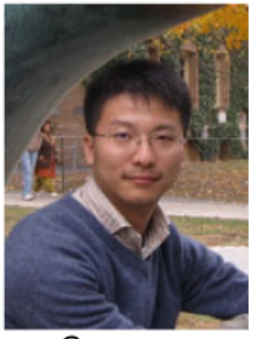
F_{ST}(T-C)~T(T-A-C)

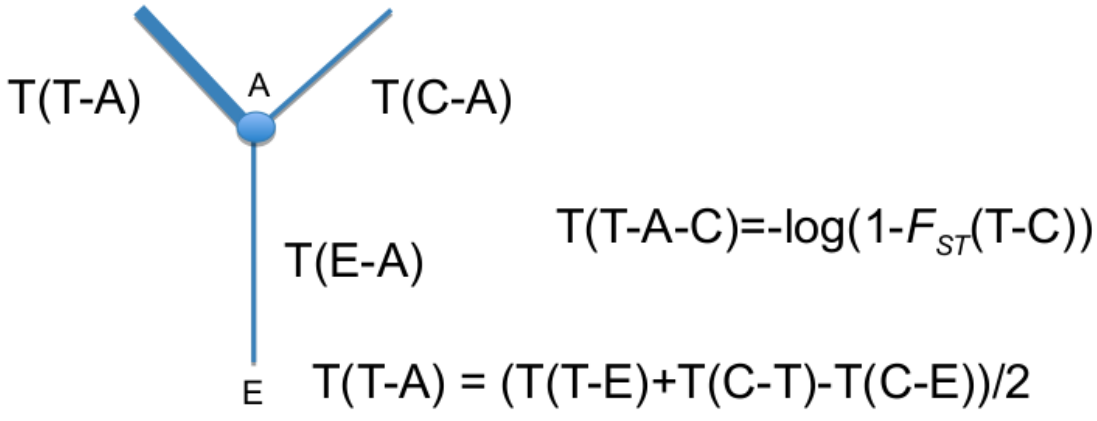
Population genetic differentiation

F_{ST}(T-C)~T(T-A-C)



Population genetic differentiation





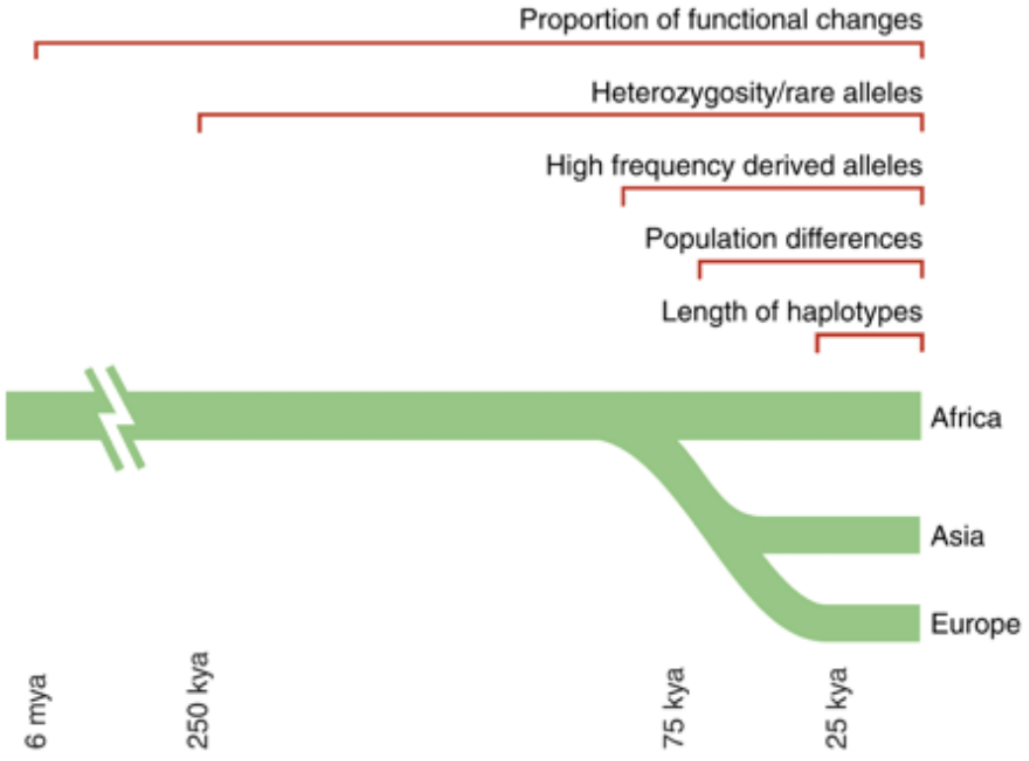
$$T(T\text{-}A\text{-}C) = -\log(1 - F_{ST}(T\text{-}C))$$

$$T(T\text{-}A) = (T(T\text{-}E) + T(C\text{-}T) - T(C\text{-}E))/2$$

Population branch statistic (PBS) (see practical)

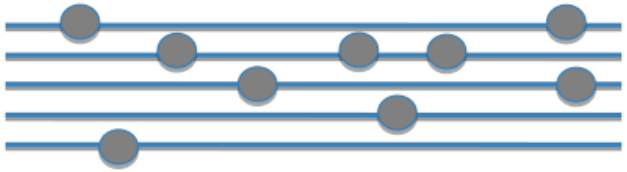
Detect recent selection

within species / using shared variation



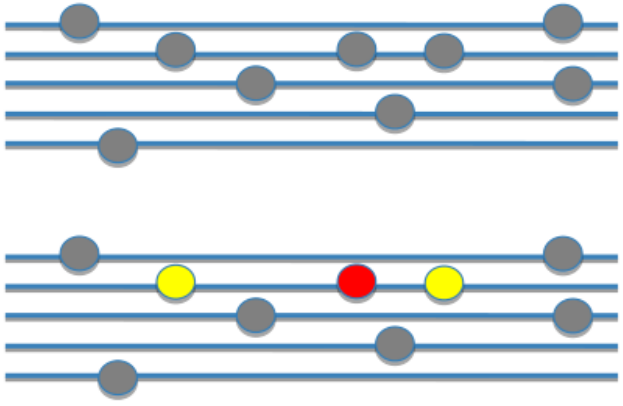
Sabeti et al. 2006 Science

Positive selection: effect on haplotypes



$t<T_{sel}$

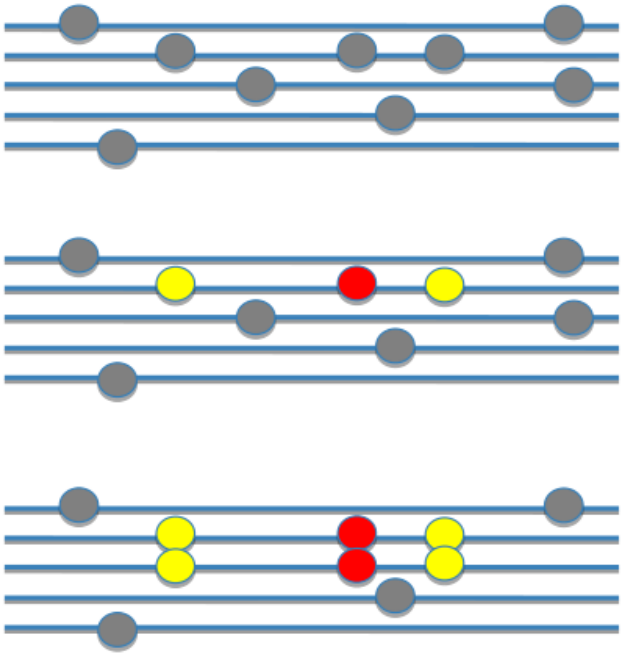
Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

Positive selection: effect on haplotypes

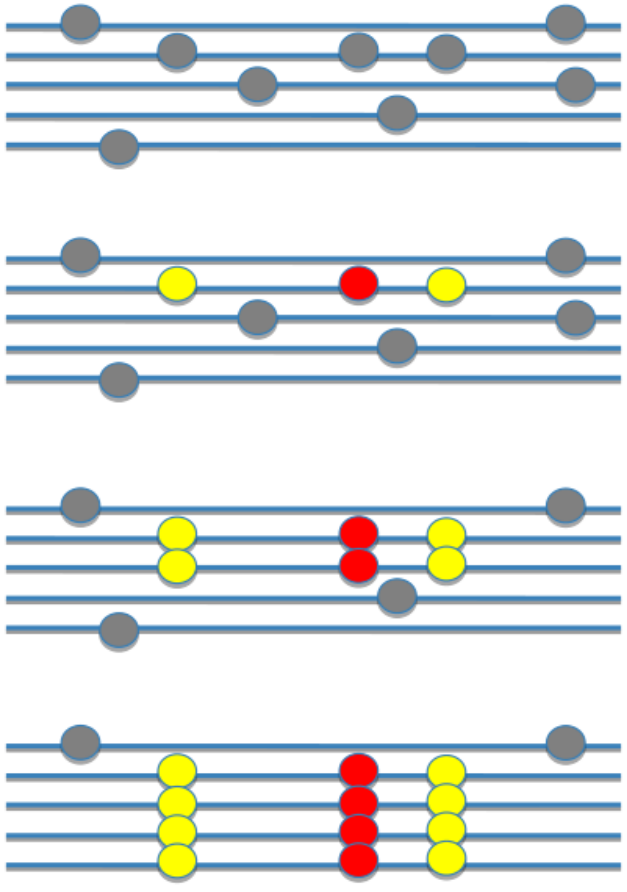


$t<T_{sel}$

$t=T_{sel}$

$t>T_{sel}$

Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

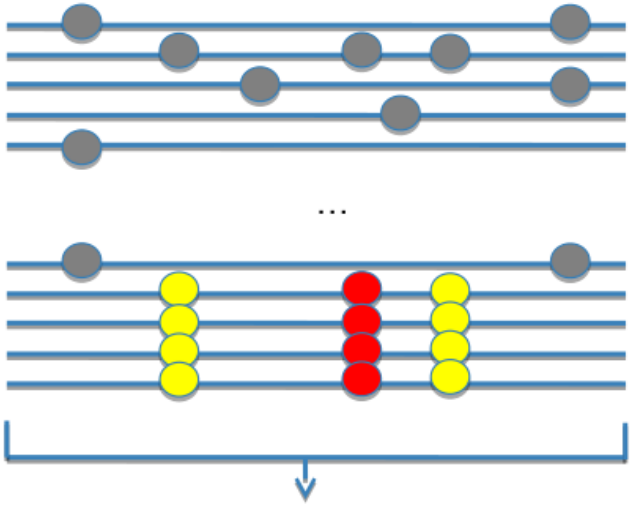
$t>T_{sel}$

$t>>T_{sel}$

Selective sweep

Genetic hitch-hiking

Positive selection



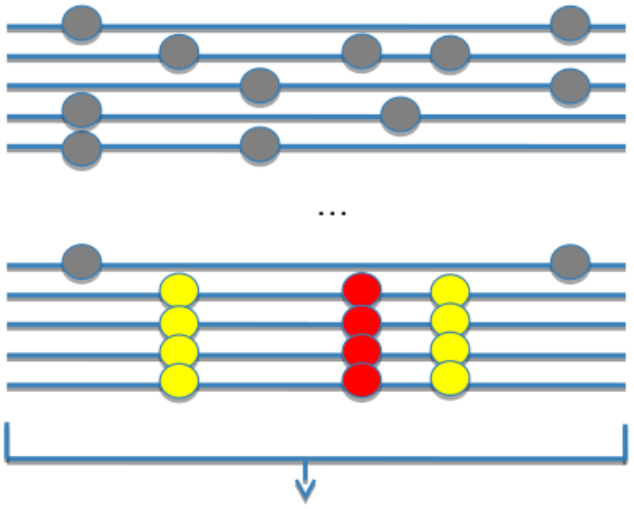
- Reduction of polymorphisms levels
(e.g. from 7 to 5 SNPs)

Nucleotide diversity index: Watterson's Theta
with K SNPs and n chromosomes

$$\theta_W = \frac{K}{a_n}$$

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

Positive selection

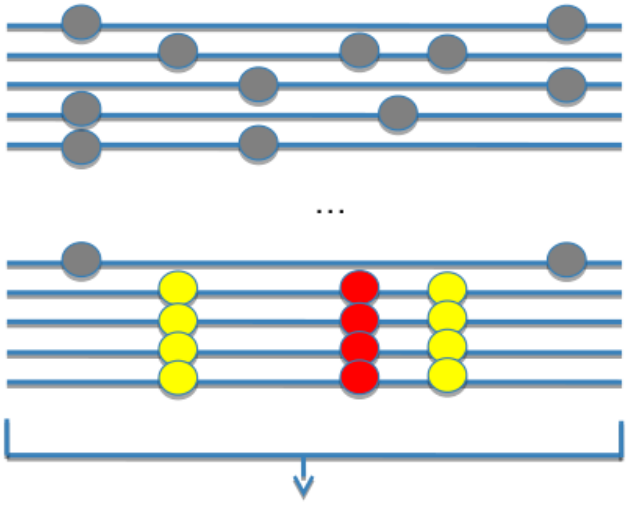


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

Nucleotide diversity index: average pairwise nucleotide differences (Pi)
with $k_{i,j}$ equal to the number of nucleotide differences between sequences *i* and *j*

$$\pi = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{i,j}}{\binom{n}{2}}$$

Positive selection



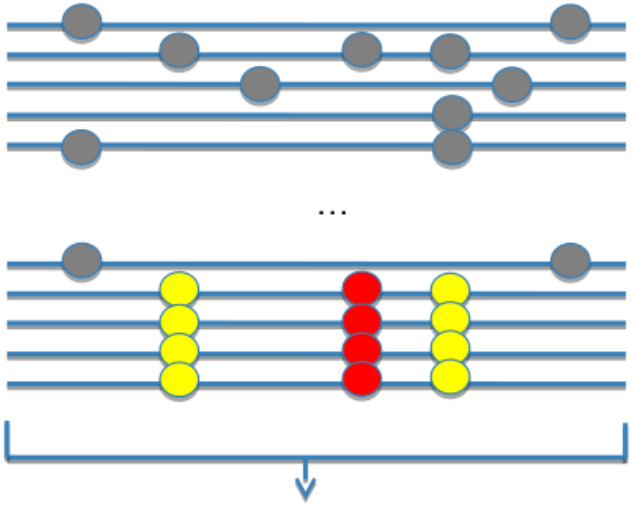
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

Under neutrality, Theta and Pi are expected to be the same.
Tajima's D measures their difference.

$$D = \frac{\pi - \theta_W}{\sqrt{\hat{V}(\pi - \theta_W)}}$$

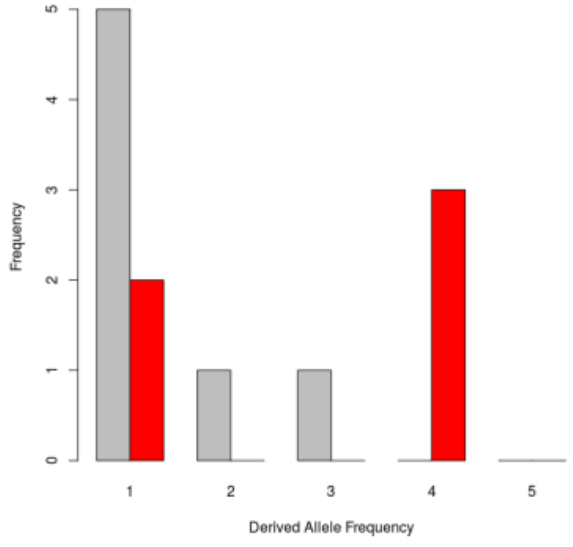
D<0 is suggestive of an excess of low-frequency variants

The Site Frequency Spectrum



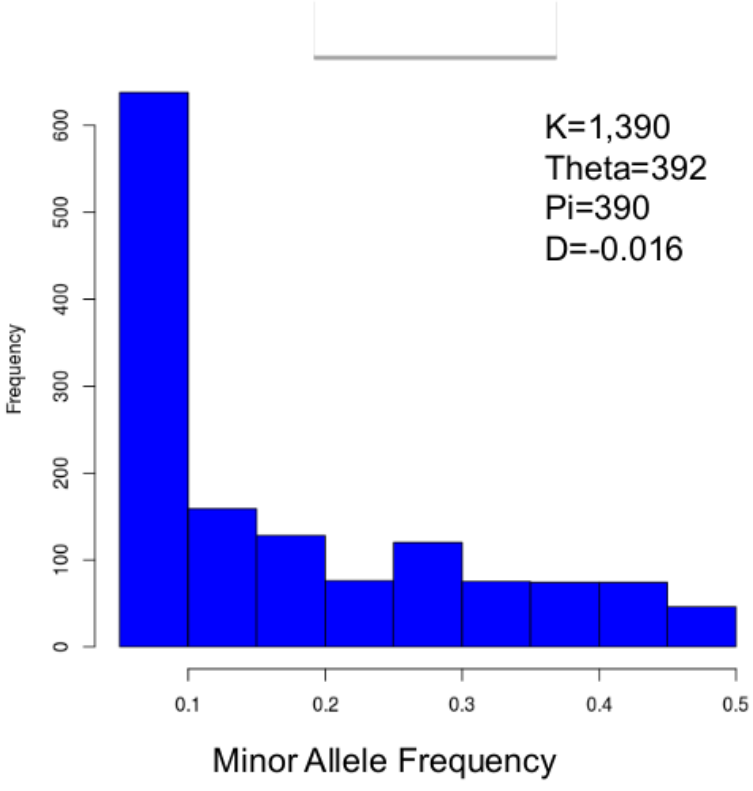
$t<T_{sel}$

$t>>T_{sel}$

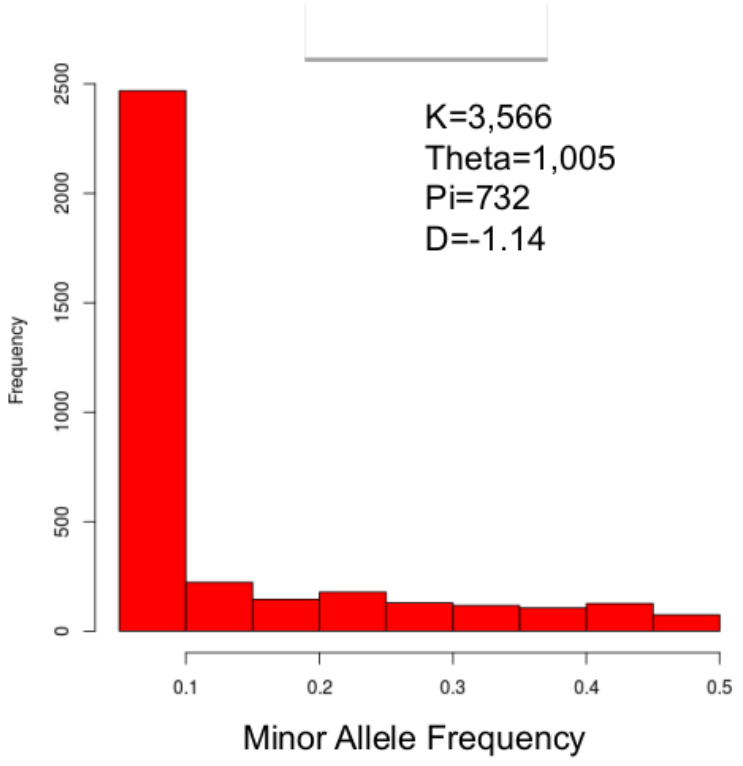


Confounding factor

n=20; L=500kbp; no selection

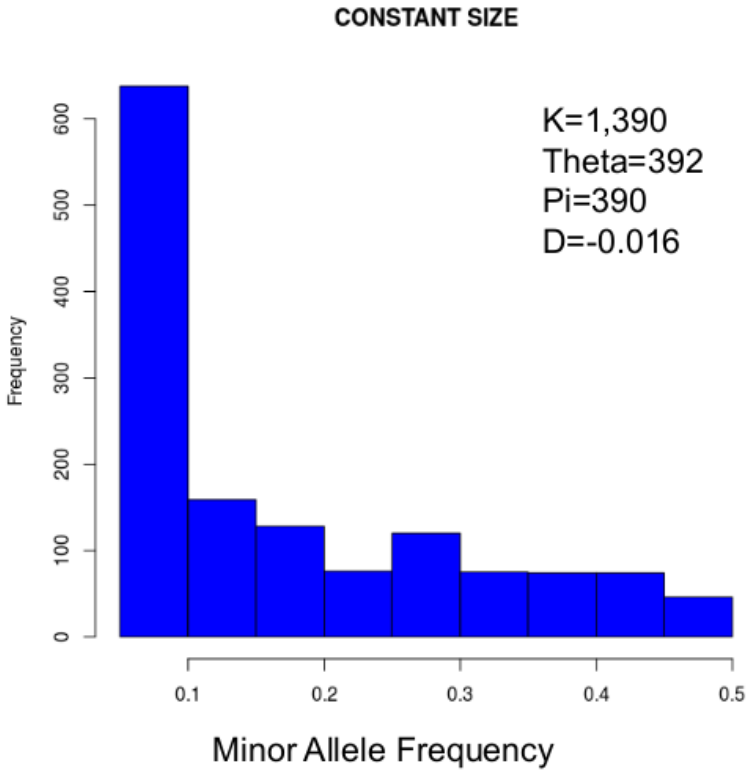


n=20; L=500kbp; no selection

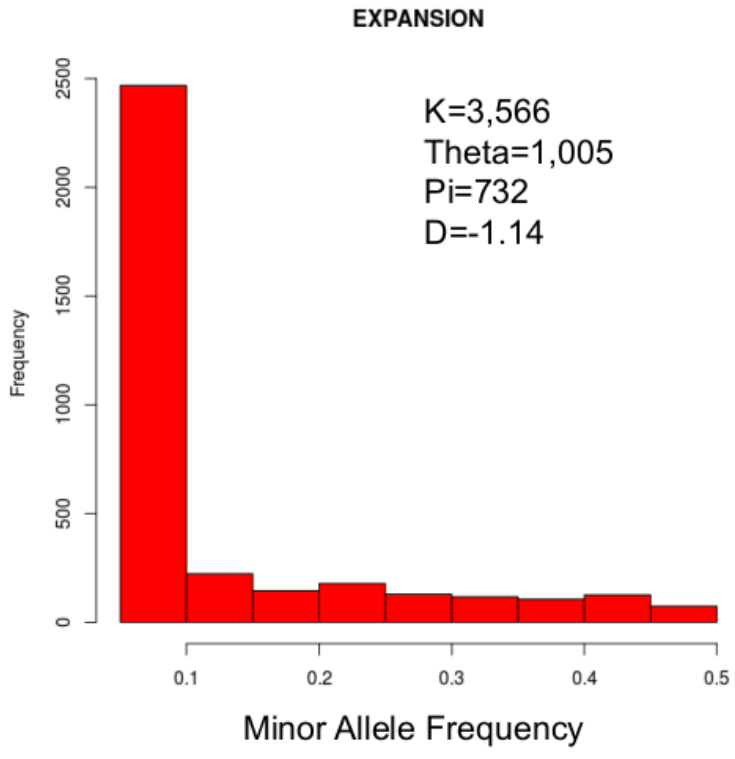


Demography matters!

n=20; L=500kbp; no selection



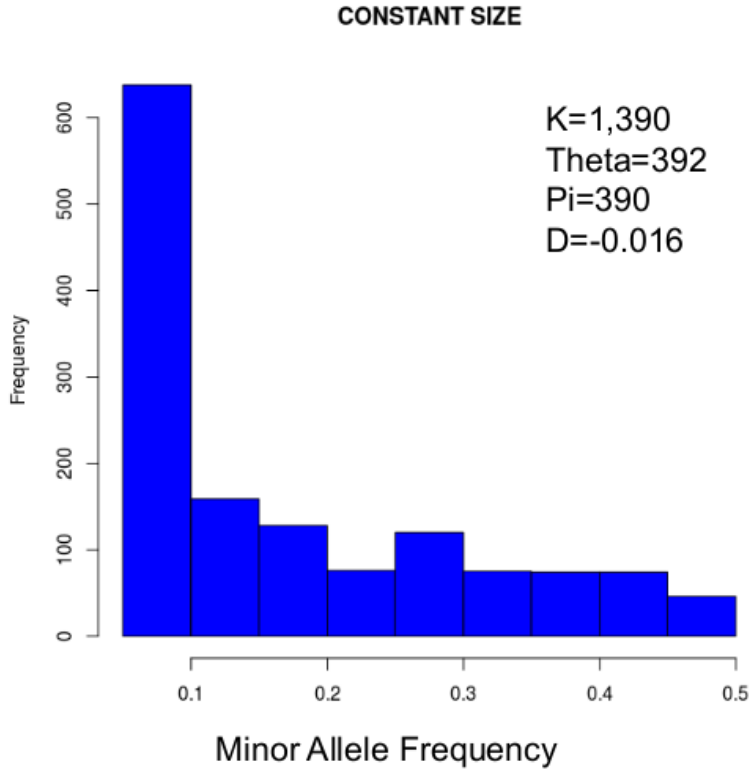
n=20; L=500kbp; no selection



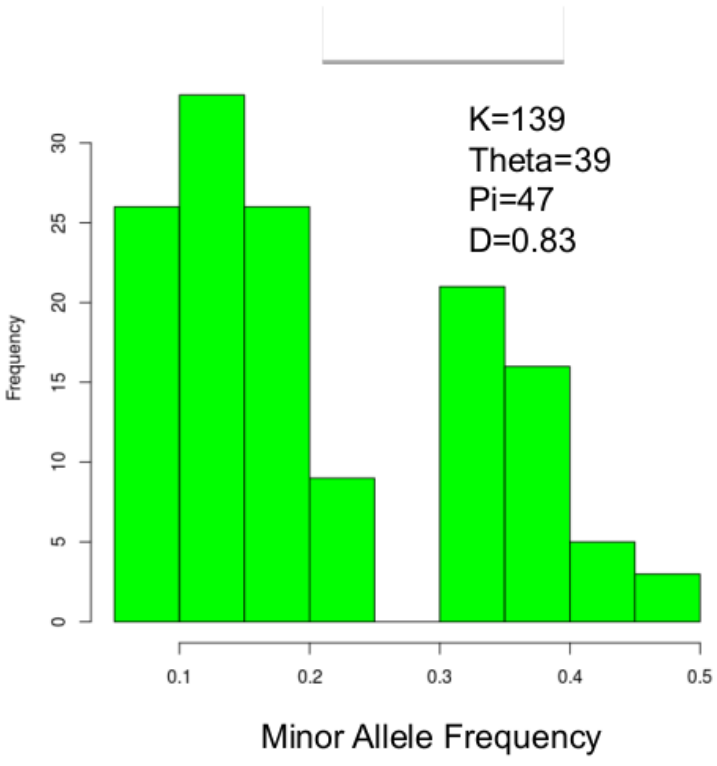
- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

Demography matters?

n=20; L=500kbp; no selection

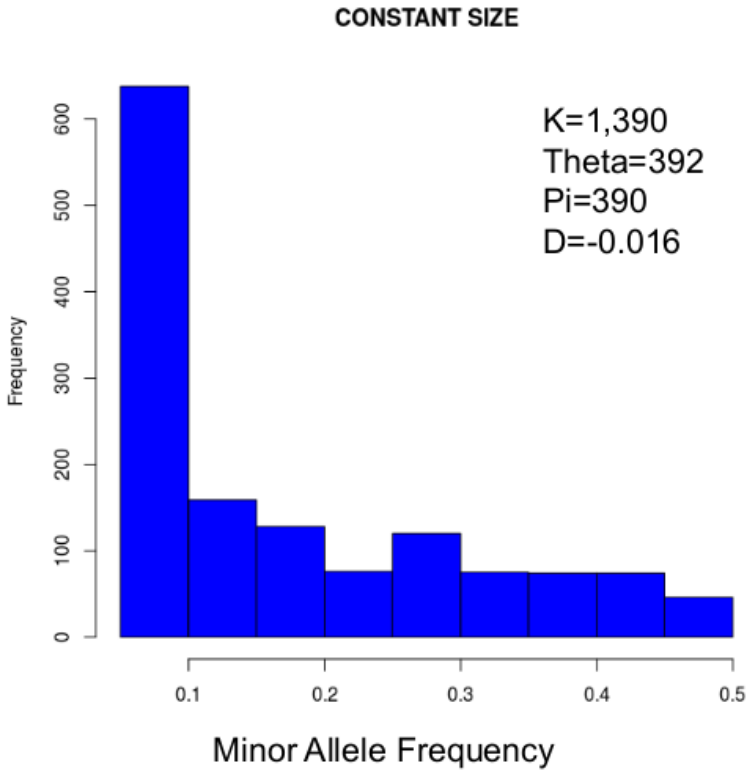


n=20; L=500kbp; no selection

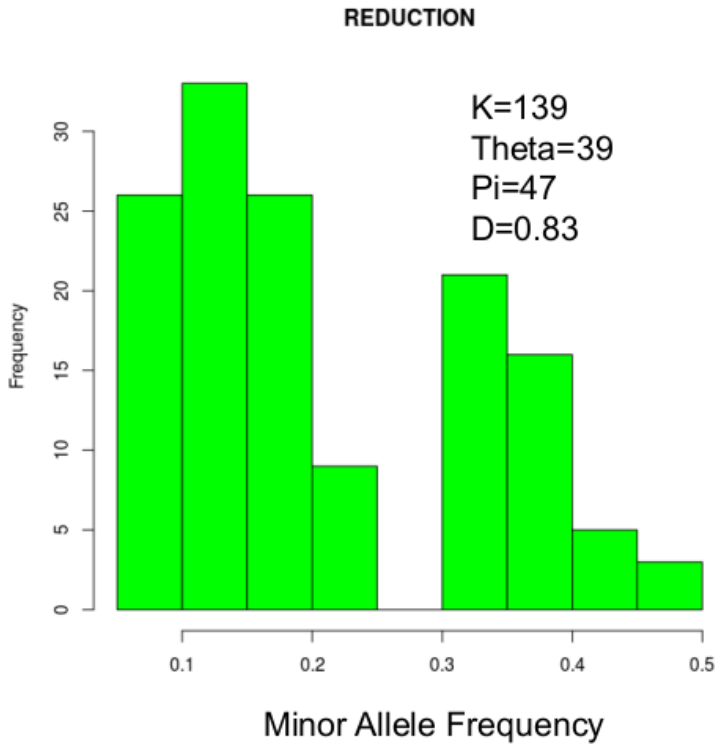


Demography matters!

n=20; L=500kbp; no selection

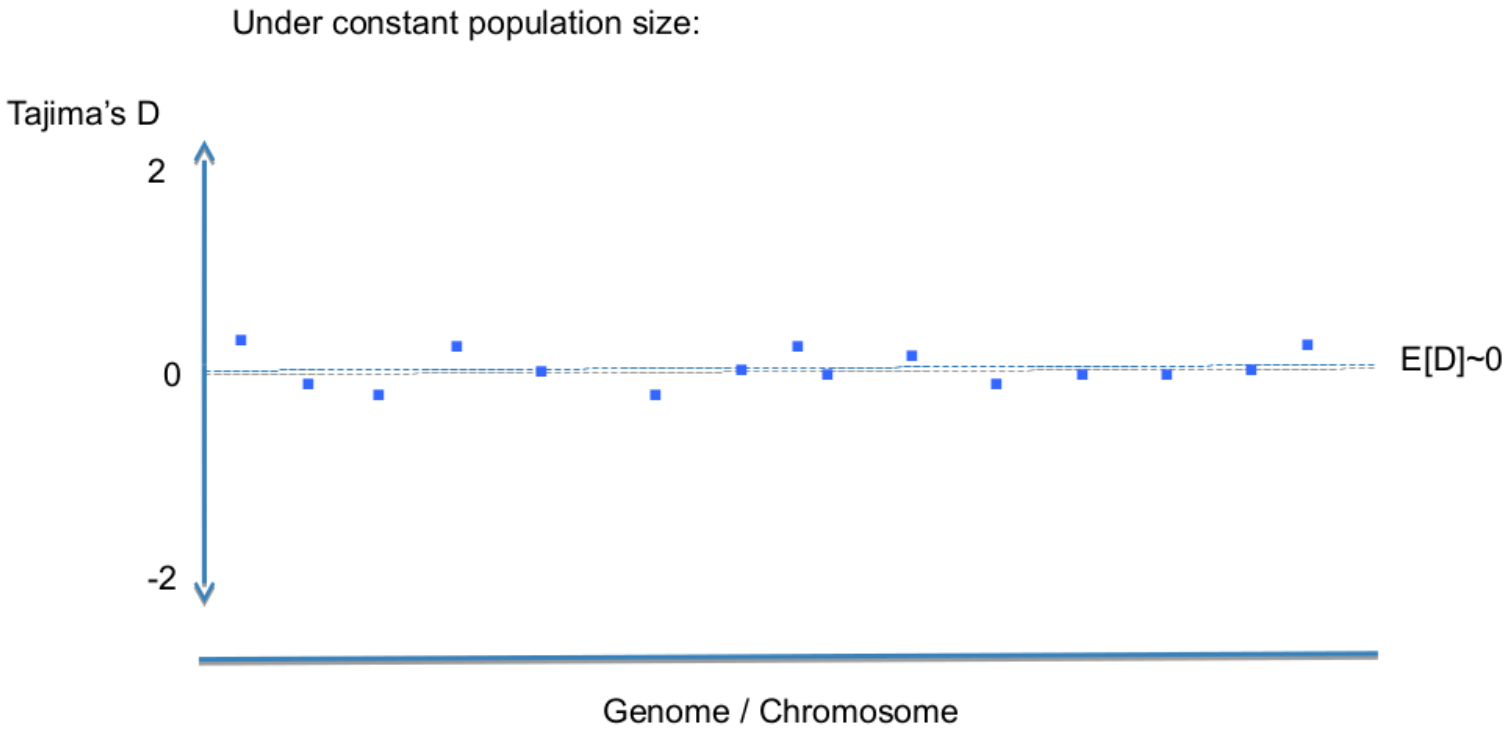


n=20; L=500kbp; no selection

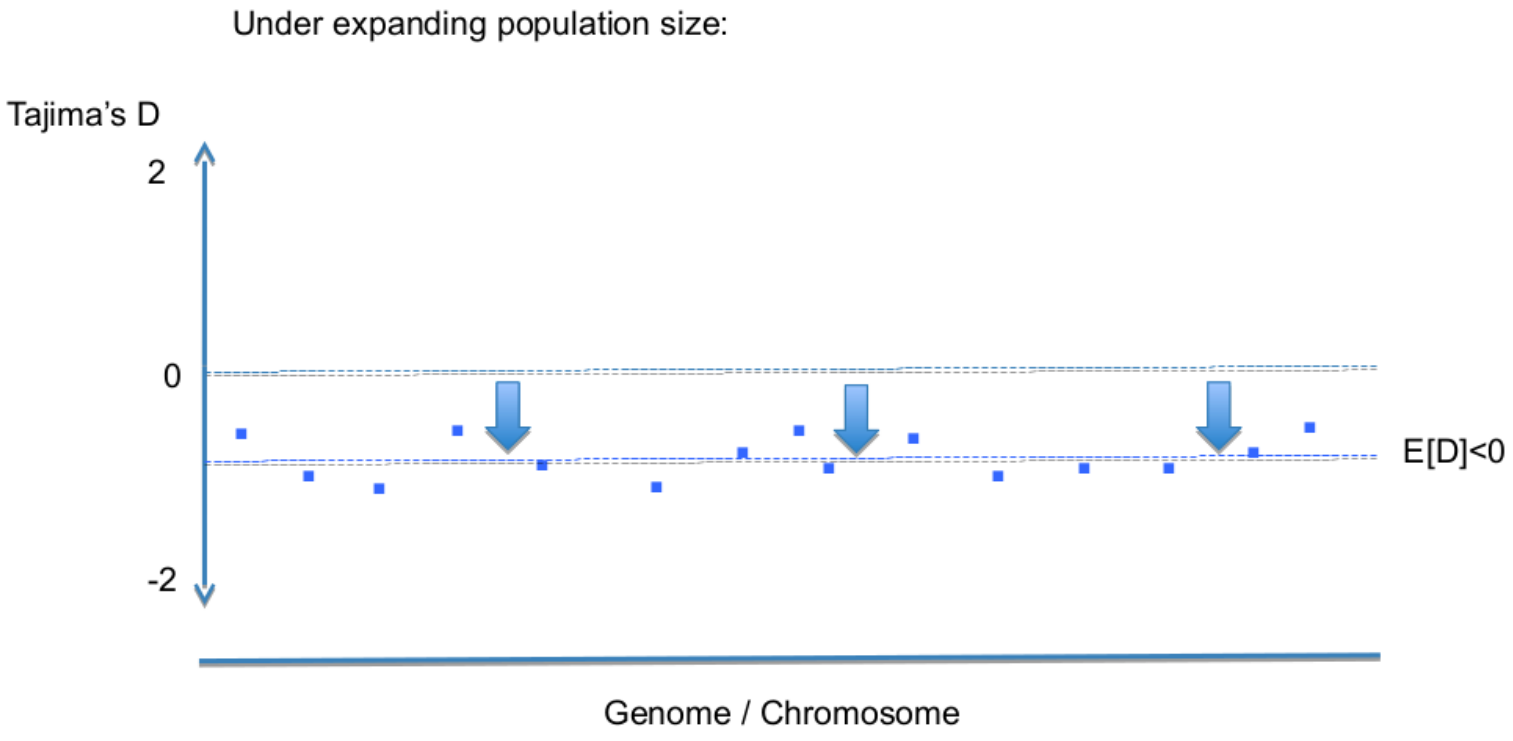


- Depletion of segregating sites
- Excess of intermediate-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

How to take neutral confounding factors into account?

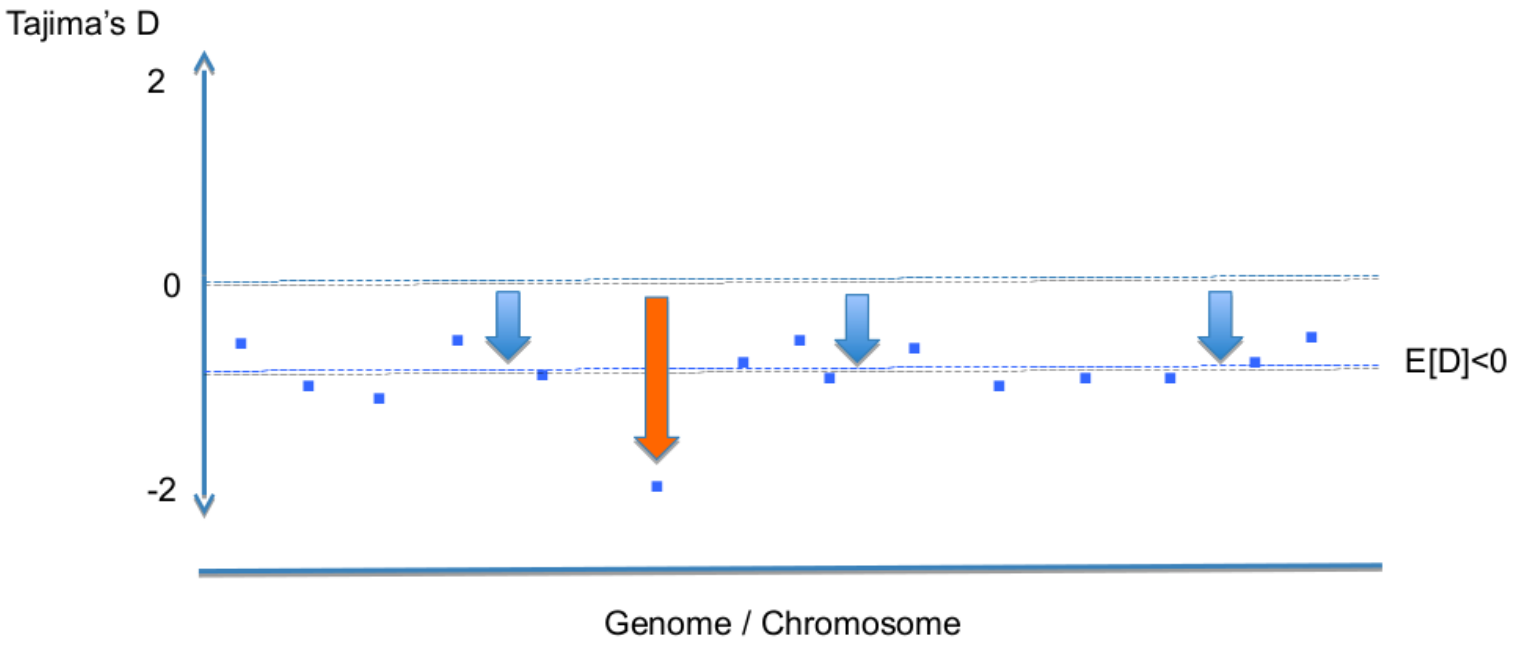


How to take neutral confounding factors into account?



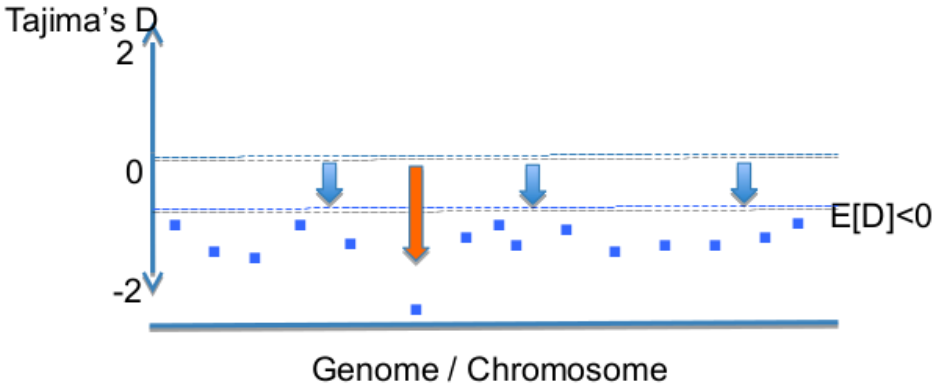
How to take neutral confounding factors into account?

Under expanding population size and positive selection:

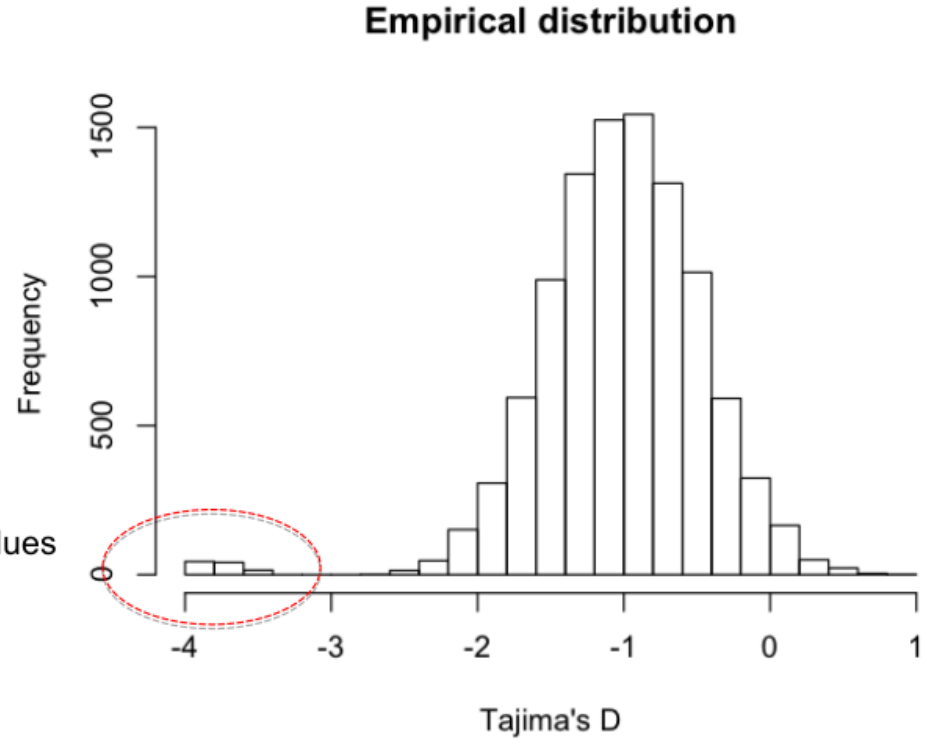


- Demography affects all loci equally, while selection changes local patterns

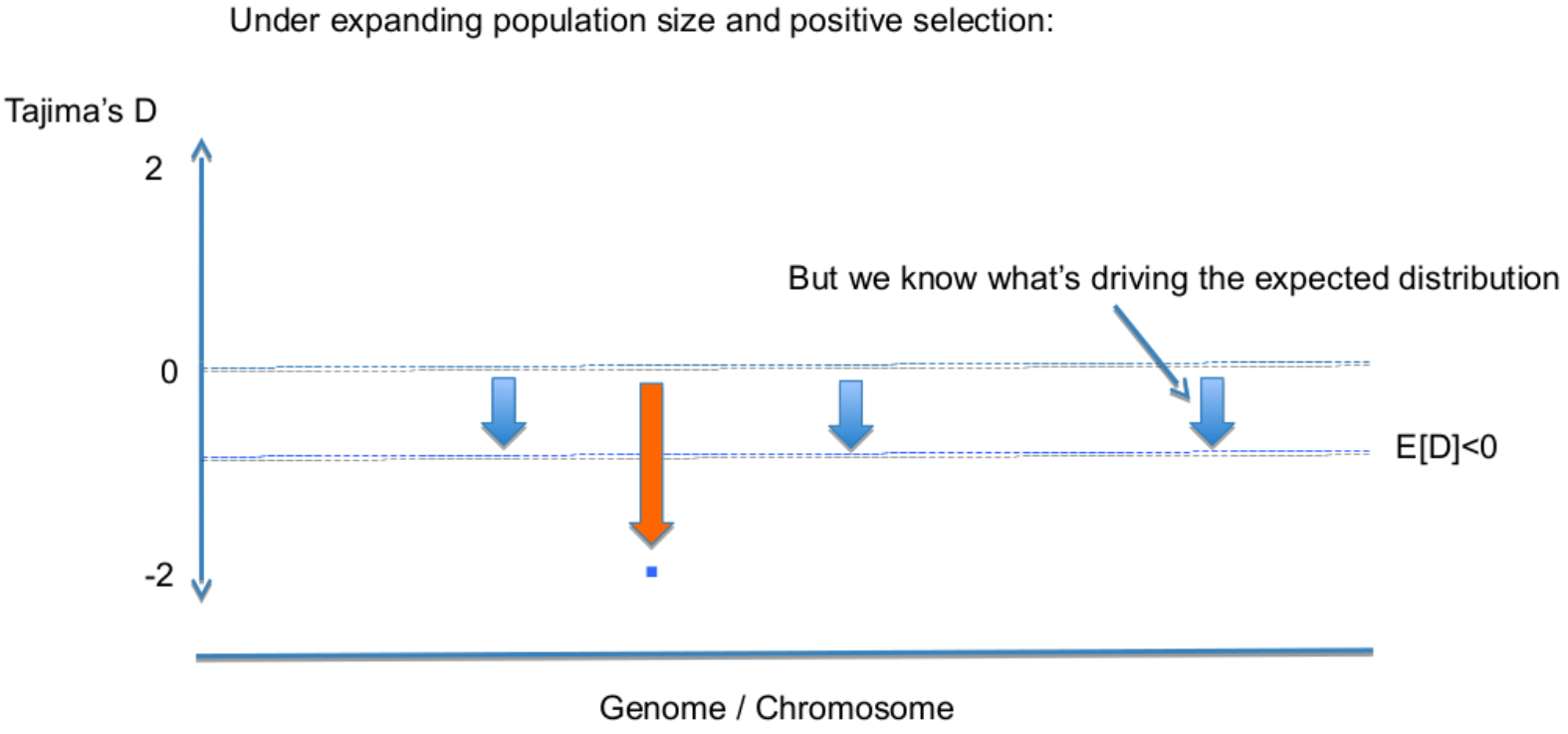
Outlier approach



Assign empirical *p*-values
(ranked percentiles)

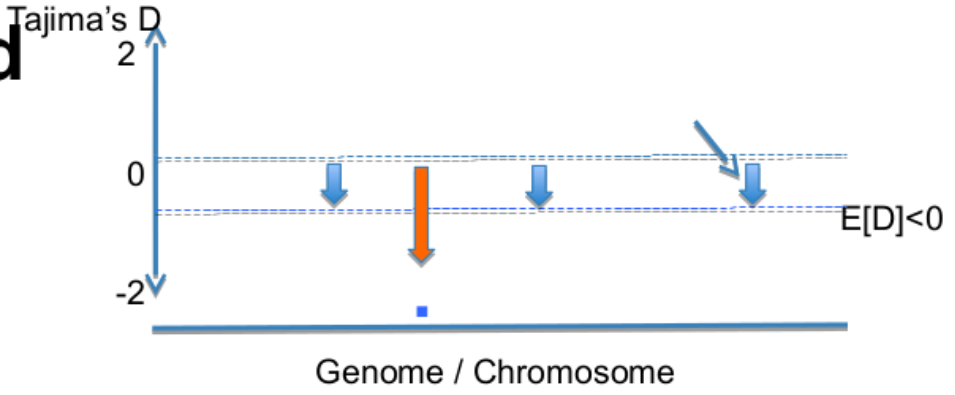


How to take neutral confounding factors into account?

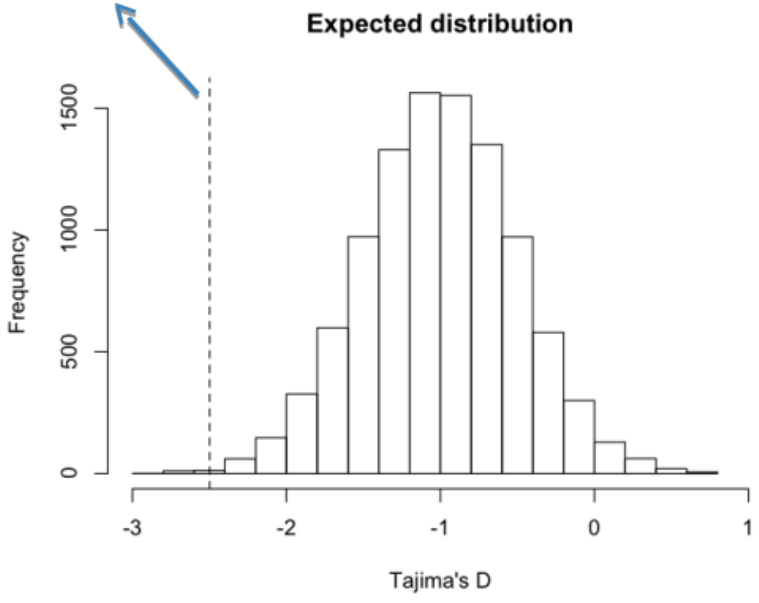


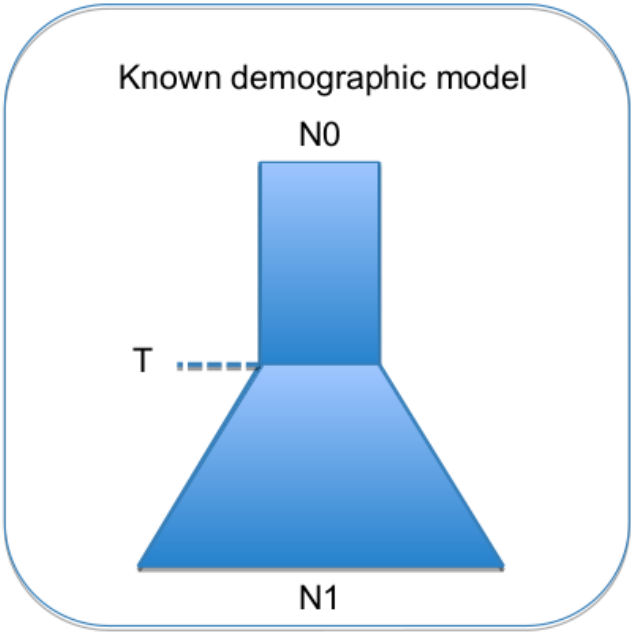
- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

Simulations-based approach



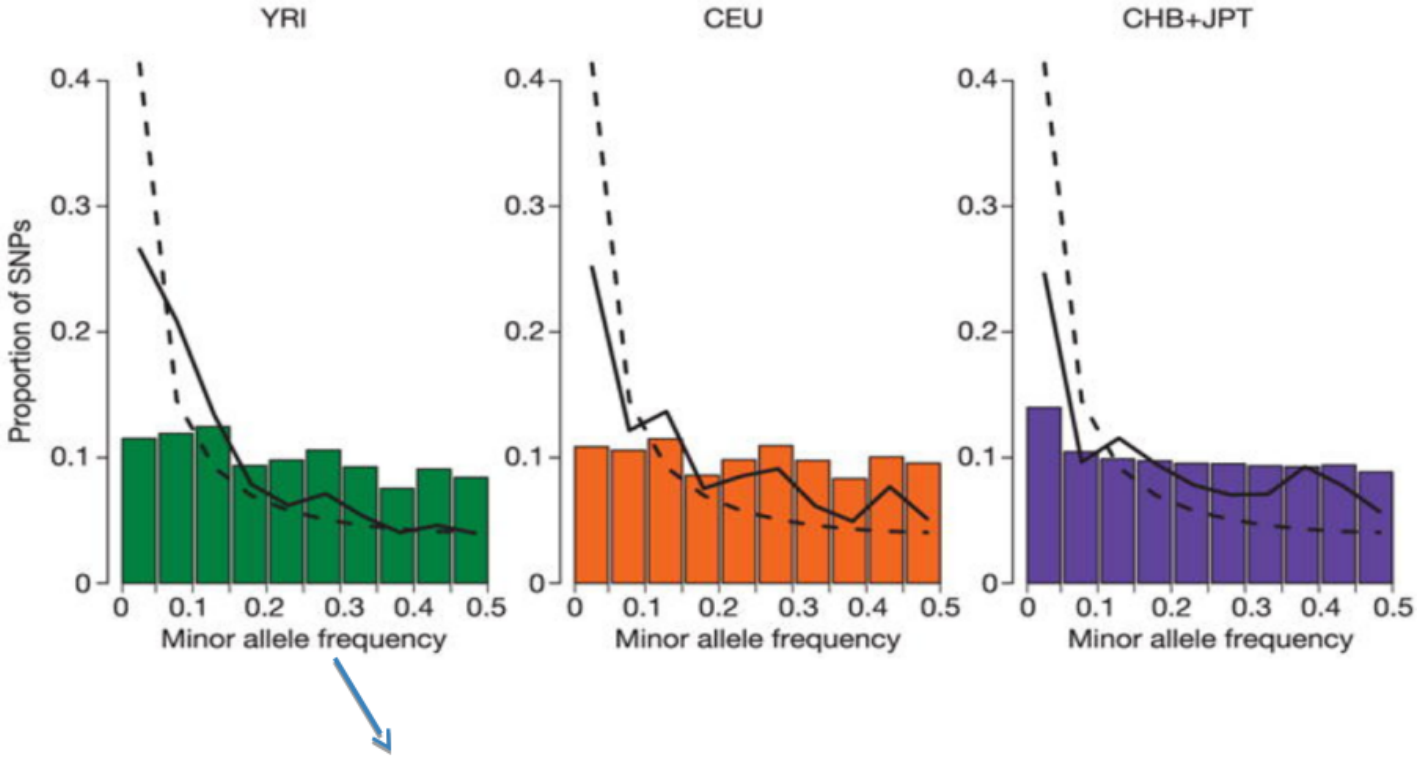
Assign *p*-values
(based on ranked percentile of observed value)





Experimental design matters?

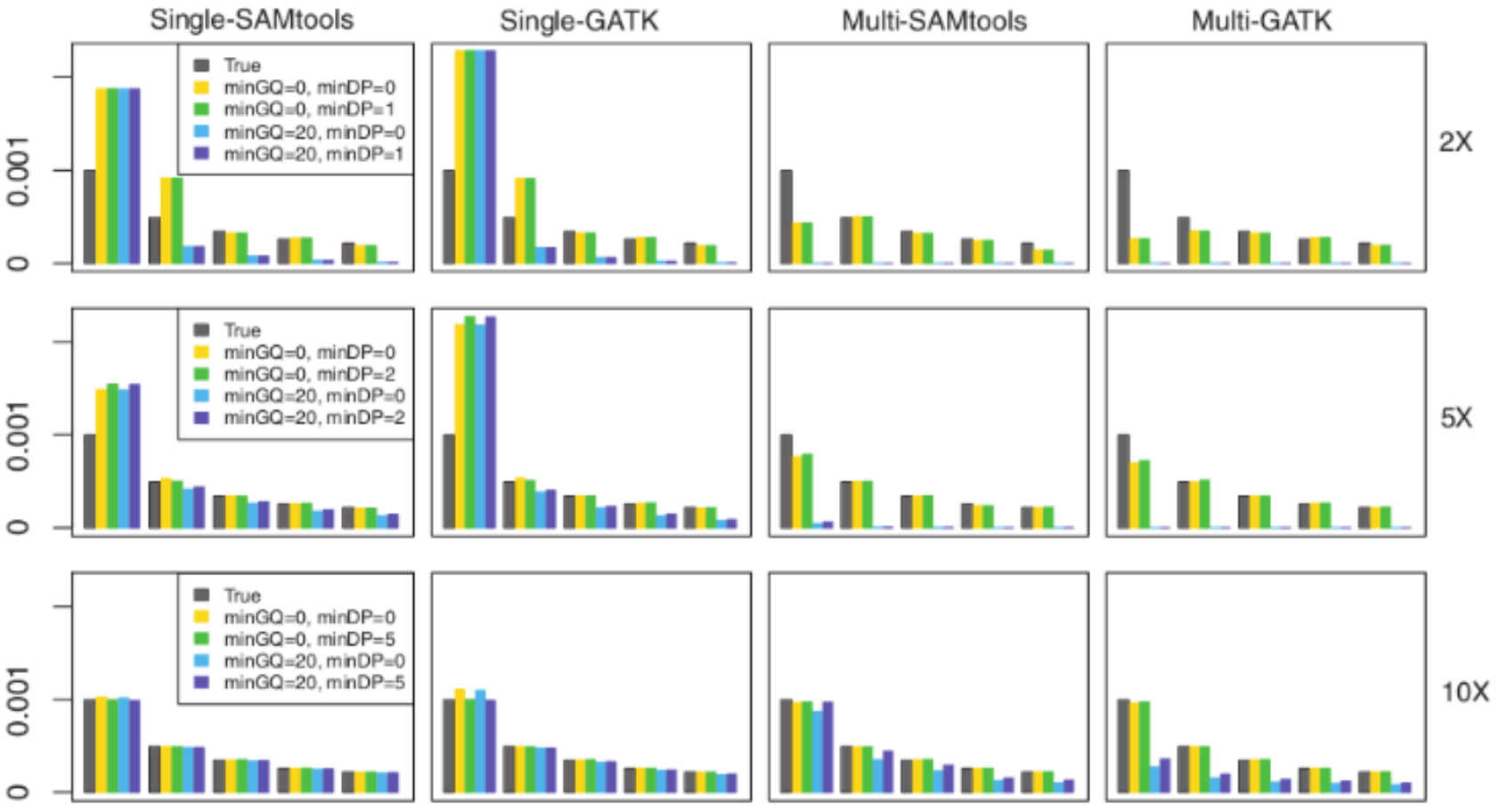
The effect of ascertainment bias



Deficiency of low-frequency variants

HapMap Consortium. Nature 2005

Effect of low-depth sequencing



(slides stolen from Anders)

Sample allele frequency posterior probabilities

With 6 chromosomes (3 diploids)

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = \ ?$$

$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Sample allele frequency posterior probabilities

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = 1 - p(S=0) - p(S=2k) \quad = 0.90$$

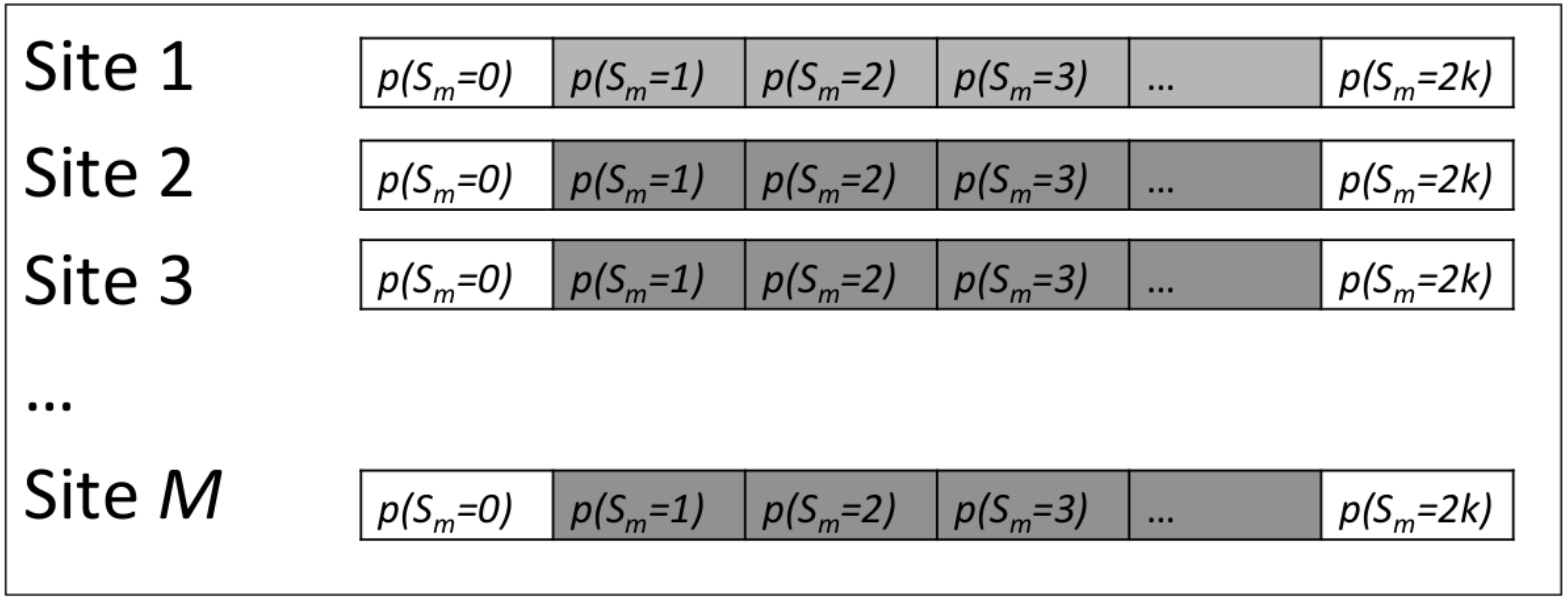
$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

Nr of segregating sites



Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[S] = \sum_{m=1}^{M} p_{\text{var}}^{(m)} = \sum_{m=1}^{M} (1 - p(S_m = 0) - p(S_m = 2k))$$

Nucleotide diversity

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$D = 2f(1-f)$$

$$E[D] =$$

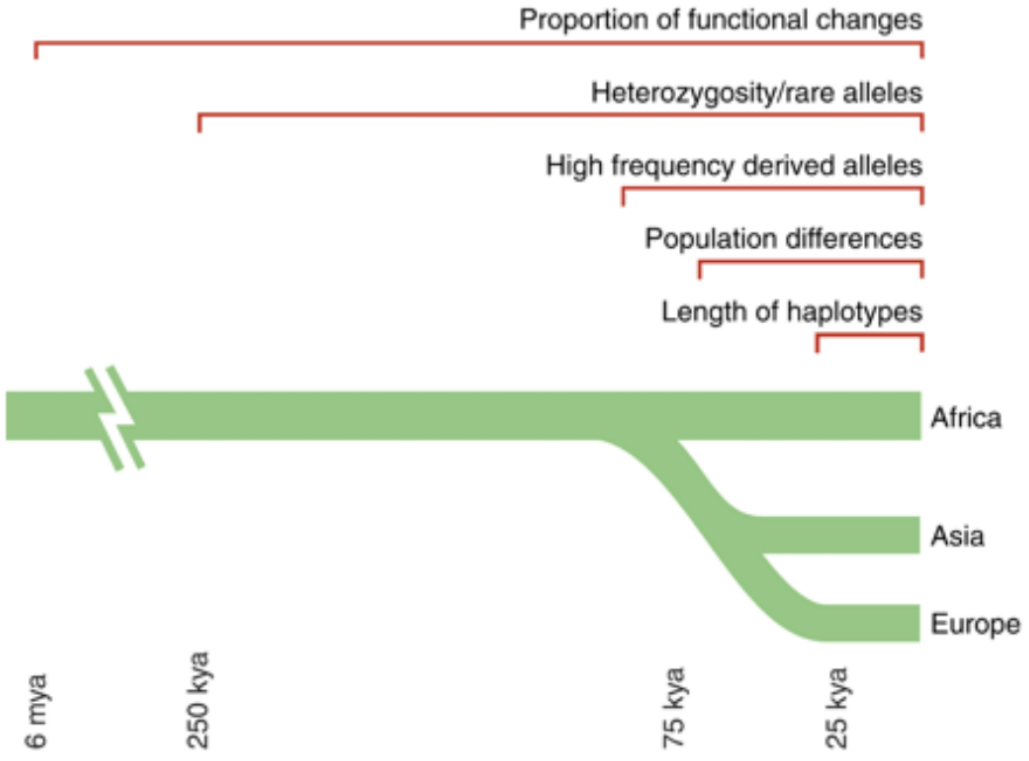
Nucleotide diversity

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[D]=\sum_{m=1}^{M}\sum_{j=0}^{2k}2\left(\frac{i}{2k}\right)\left(\frac{2k-i}{2k}\right)p(S_m=i)$$

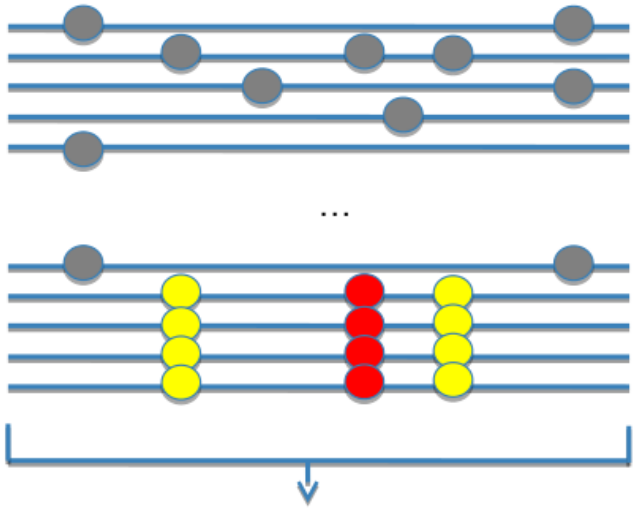
Detect recent selection

within species / using shared variation



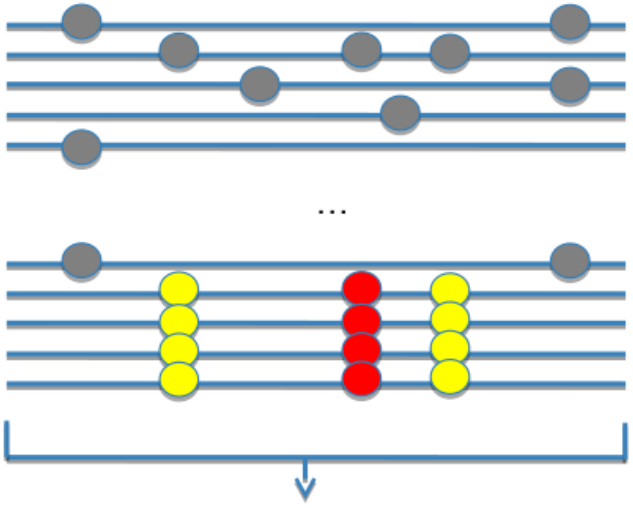
Sabeti et al. 2006 Science

Positive selection



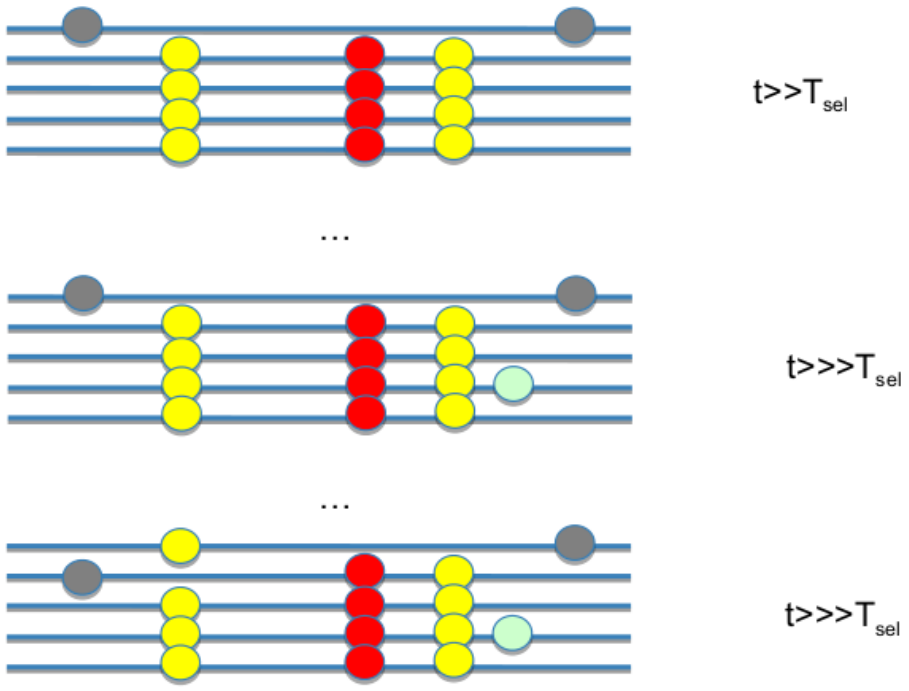
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- ?

Positive selection



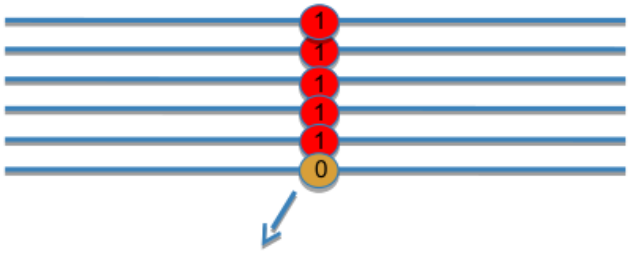
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- Extended haplotype homozygosity / Extended LD

Extended Haplotype Homozygosity



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

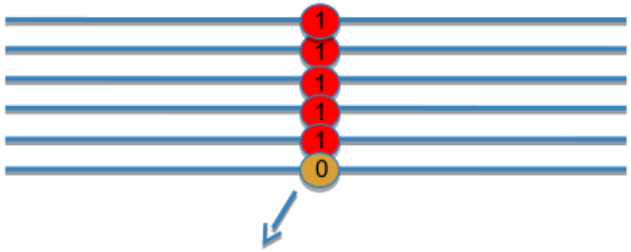


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Core SNP

Extended Haplotype Homozygosity

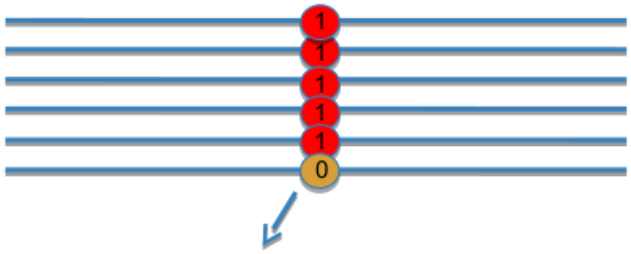


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

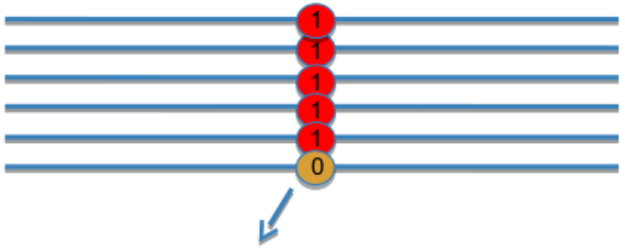


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Sum across all unique haplotypes
carrying the core SNP

Extended Haplotype Homozygosity



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

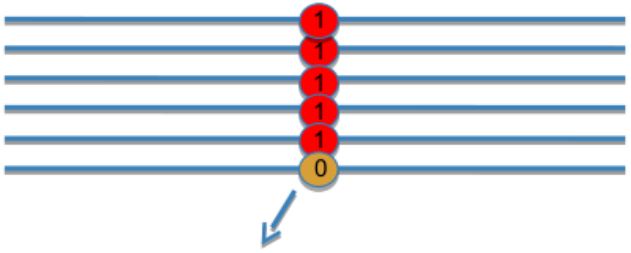
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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)

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

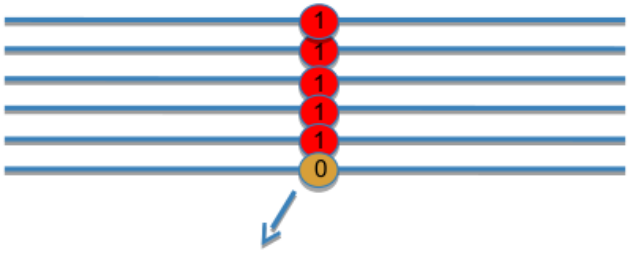
n_h is haplotype frequency of h

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)

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

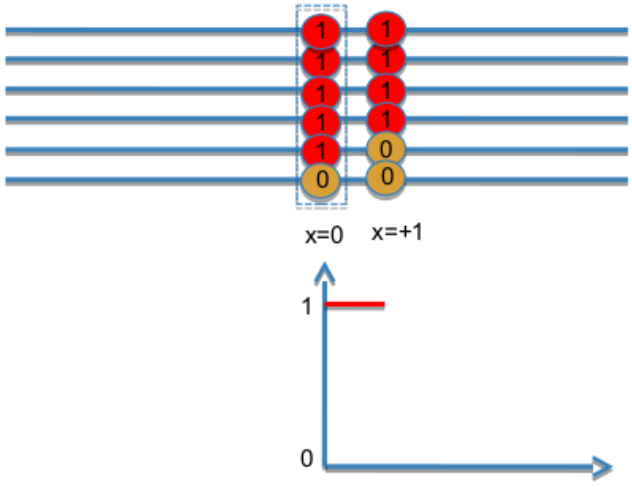
n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

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

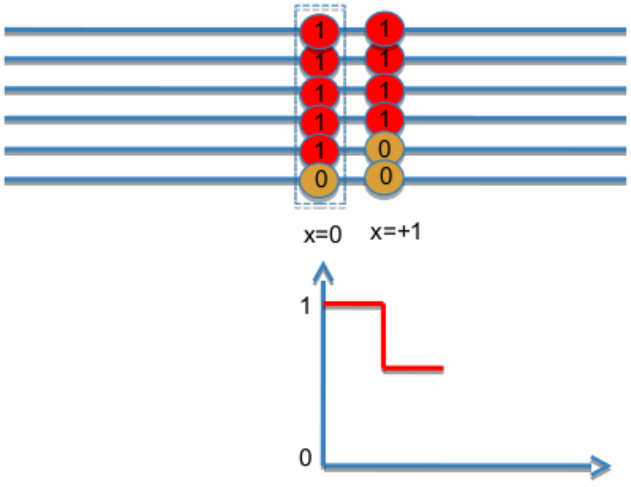
Extended Haplotype Homozygosity



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

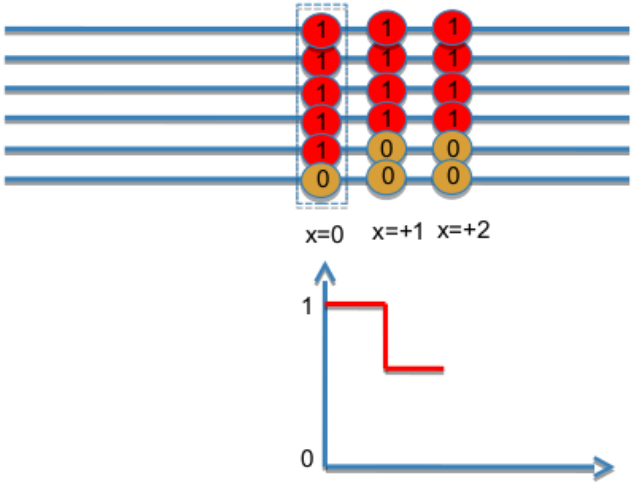
How many unique haplotypes carrying the core SNP?
What is their frequency?

Extended Haplotype Homozygosity



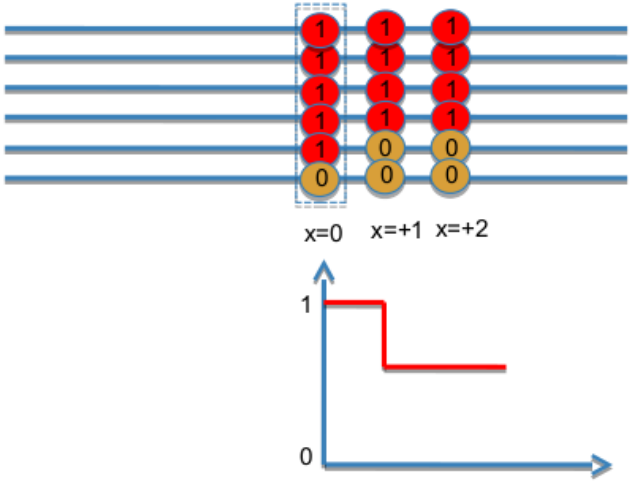
$$EHH_c(X_i = +1) = \frac{\binom{4}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{6+0}{10} = 0.60$$

Extended Haplotype Homozygosity



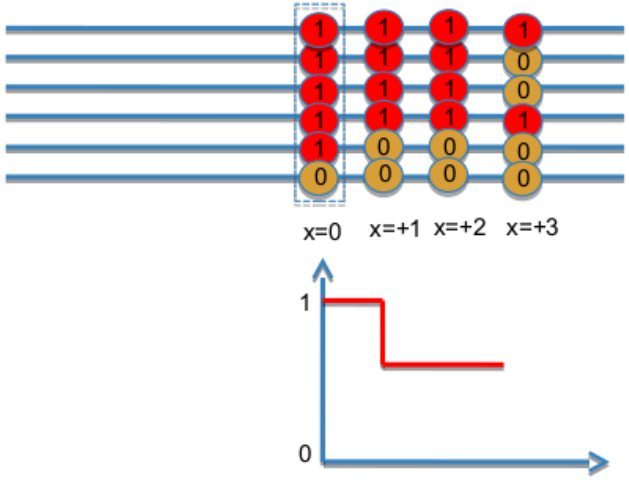
$$EHH_c(x_i = +2) = ?$$

Extended Haplotype Homozygosity



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

Extended Haplotype Homozygosity

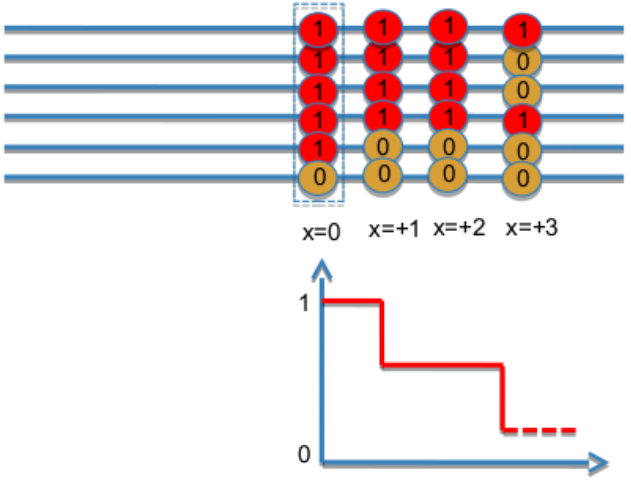


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

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

Extended Haplotype Homozygosity

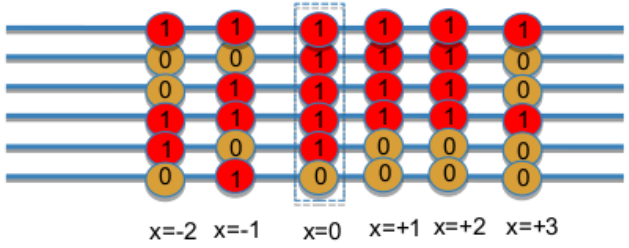


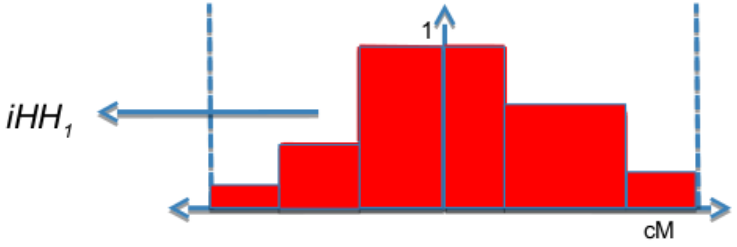
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

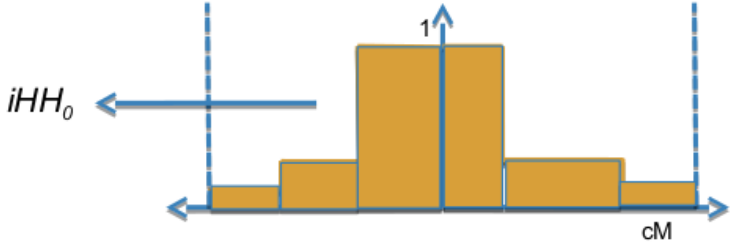
$$EHH_c(x_i = +3) = \frac{\binom{2}{2} + \binom{2}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+1+0}{10} = 0.20$$

Integrated Haplotype Score





Integrated haplotype homozygosity (*iHH*)

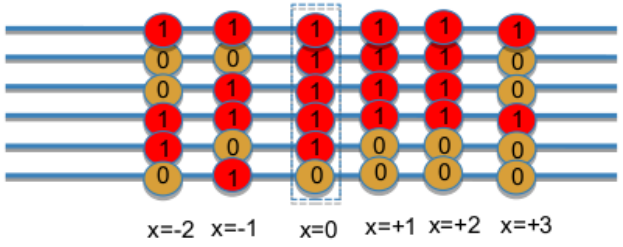


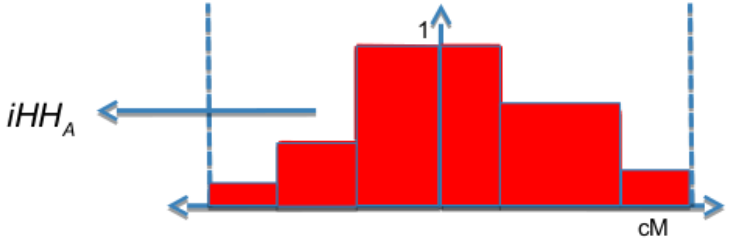
Integrated haplotype score:

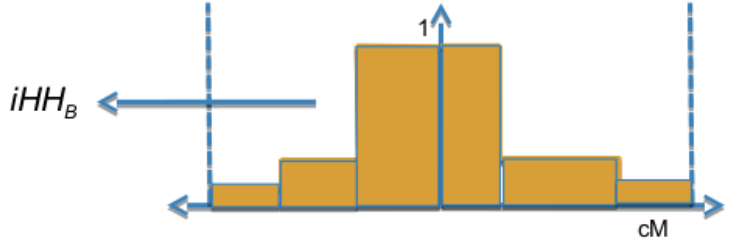
$iHs = ln(iHH_1/iHH_0)$

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

Cross-population Extended Haplotype Homozygosity







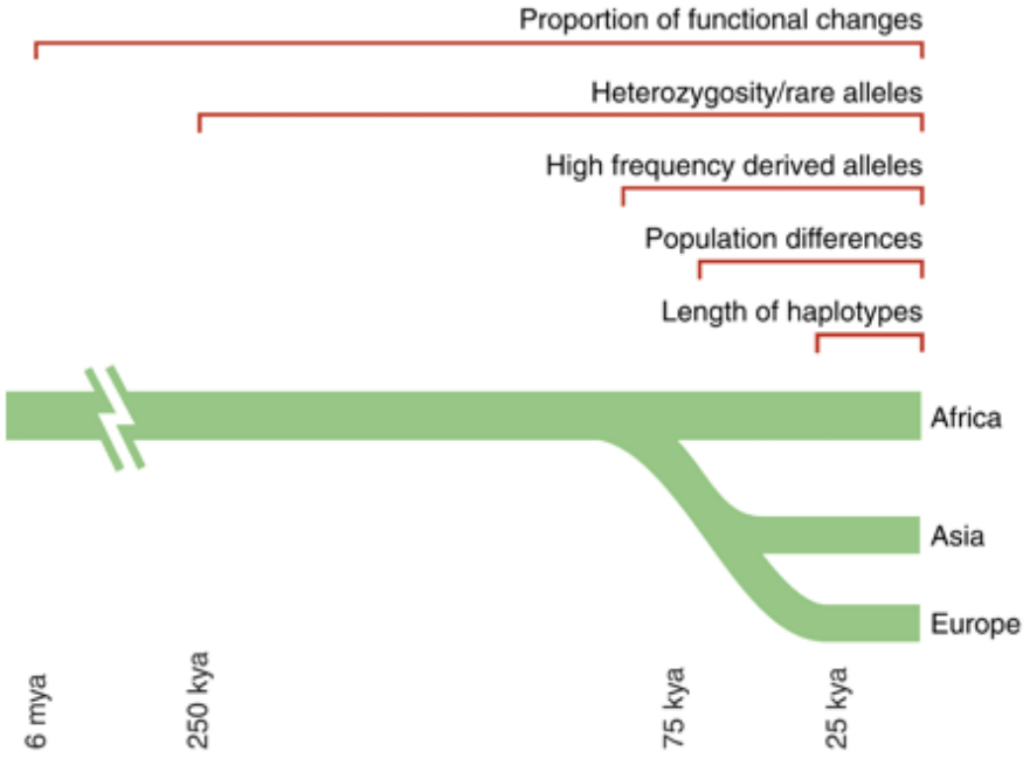
Integrated haplotype homozygosity (*iHH*) for **populations** A and B

Integrated haplotype score:
$XP\text{-}EHH = ln(iHH_A/iHH_B)$

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

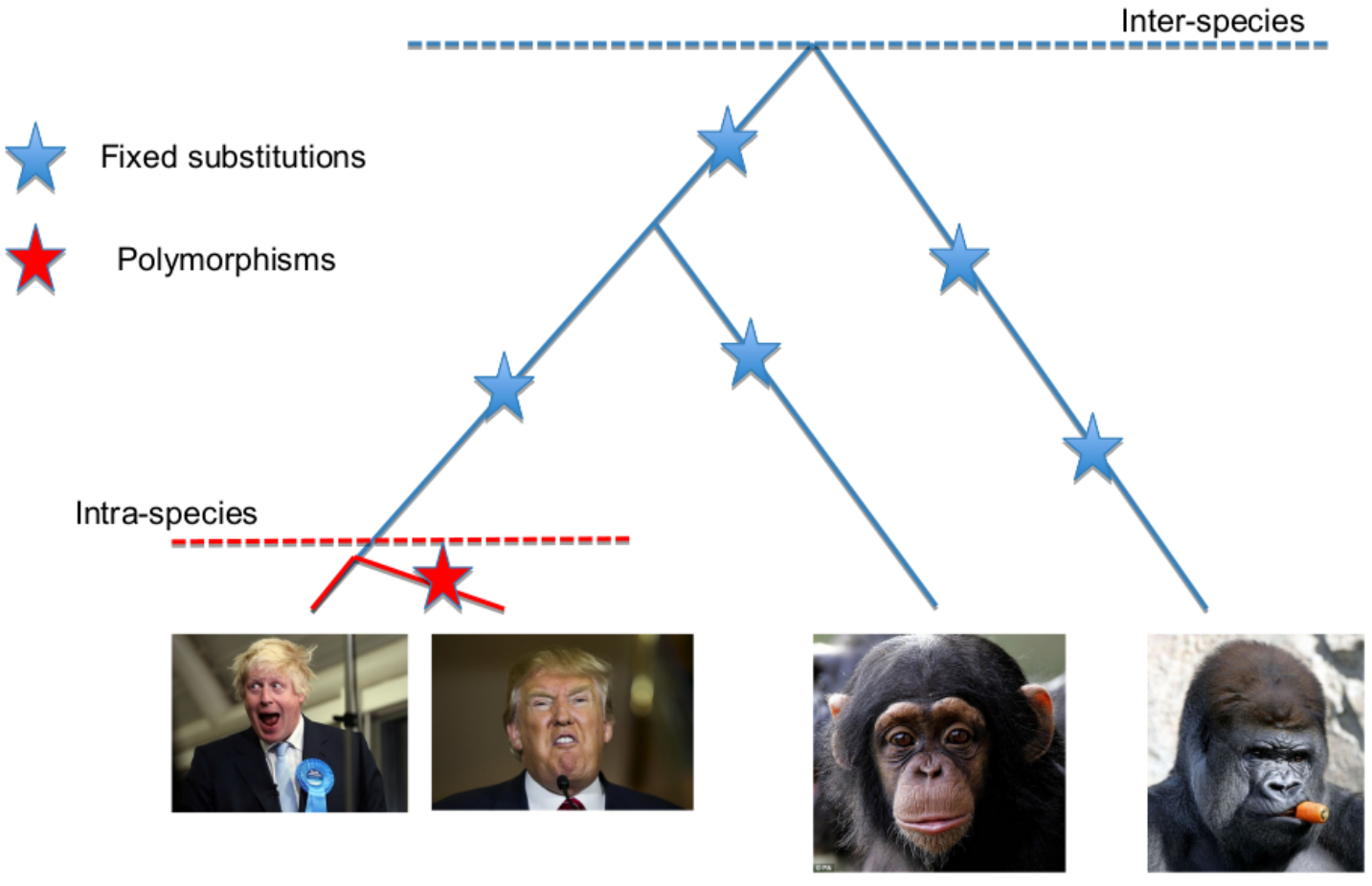
Detect recent selection

within species / using shared variation



Sabeti et al. 2006 Science

Infer inter-species selection



Recent advances to detect selection

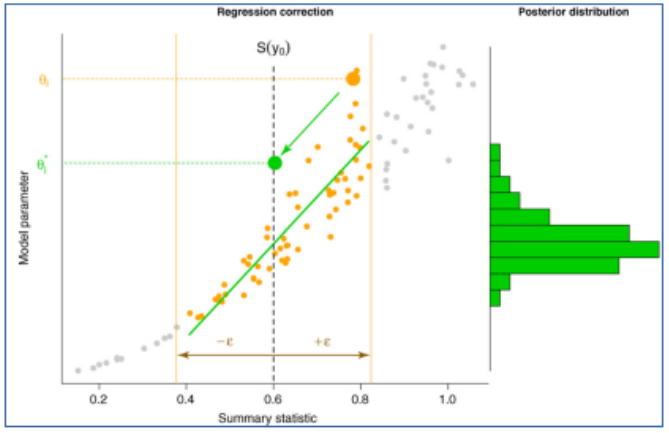
1. Composite scores (Grossman et al. 2013)

$$BF_t = \frac{P(v_t \in bin_{t,k} | selected)}{P(v_t \in bin_{t,k} | unselected)}$$

and defined the composite score as the product of the Bayes factor of each test:

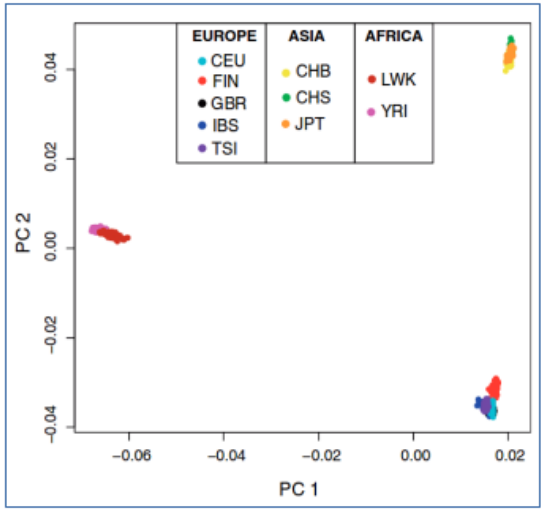
$$CMS_{GW} = \prod_{t \in tests} BF_t$$

2. Simulations-based (rejection, ABC)



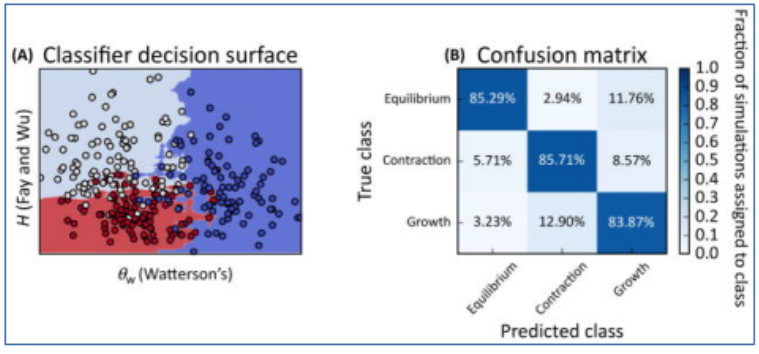
3. Unsupervised machine learning

(PCA, Duforet-Frebourg et al. 2016)



4. Supervised machine learning

(SVM, Schrider & Kern 2018)



Intended Learning Outcomes

At the end of this session you are now be able to:

- list commonly used methods to detect selection
- calculate various summary statistics
- understand main confounding factors to neutrality tests
- assess statistical significance of tests