Detecting natural selection from genomic data

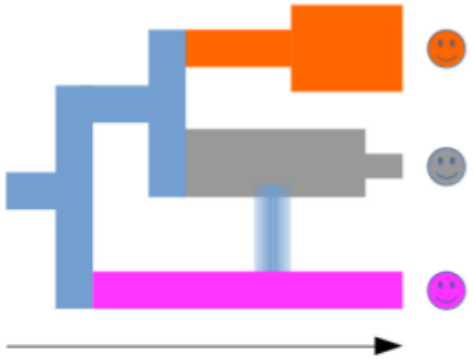
Matteo Fumagalli

2019

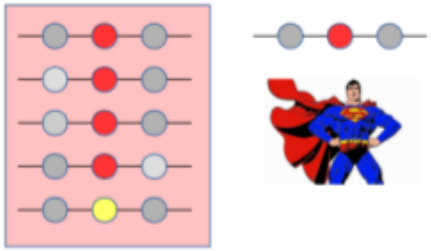
Motivation

Whole genome sequencing

Demographic history



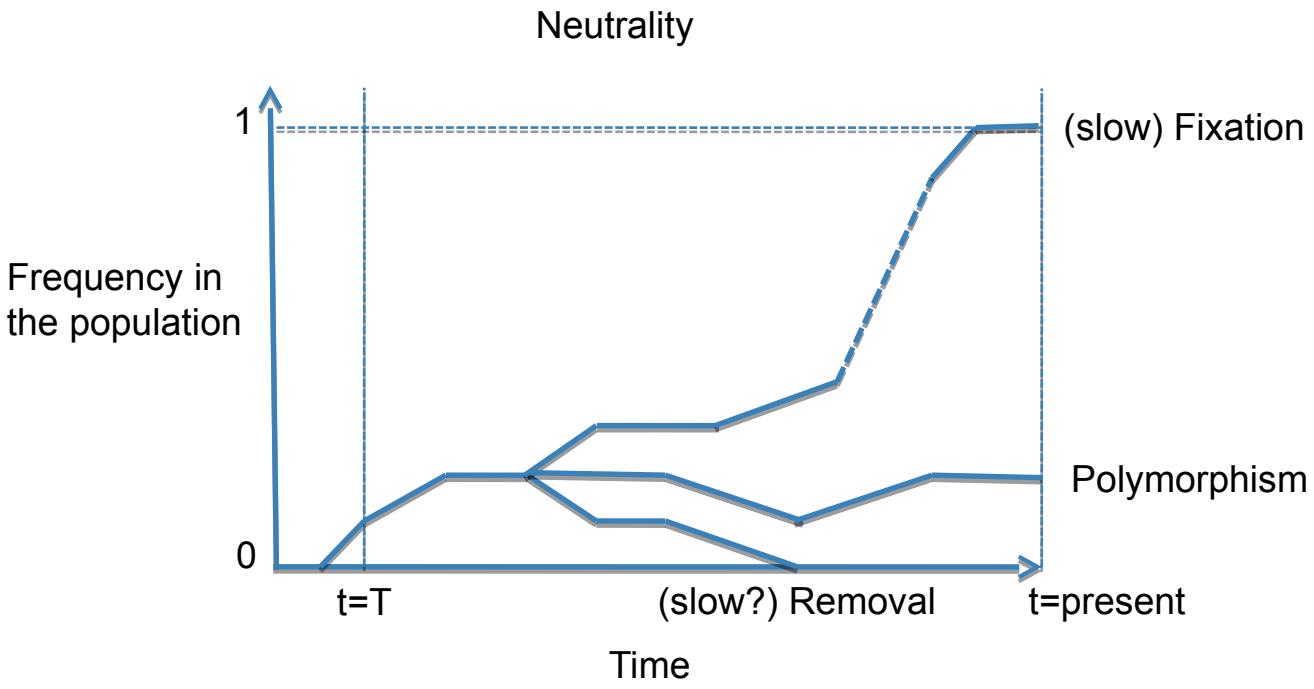
Natural selection



Natural selection

Heritable traits that increase the fitness of the become more common.

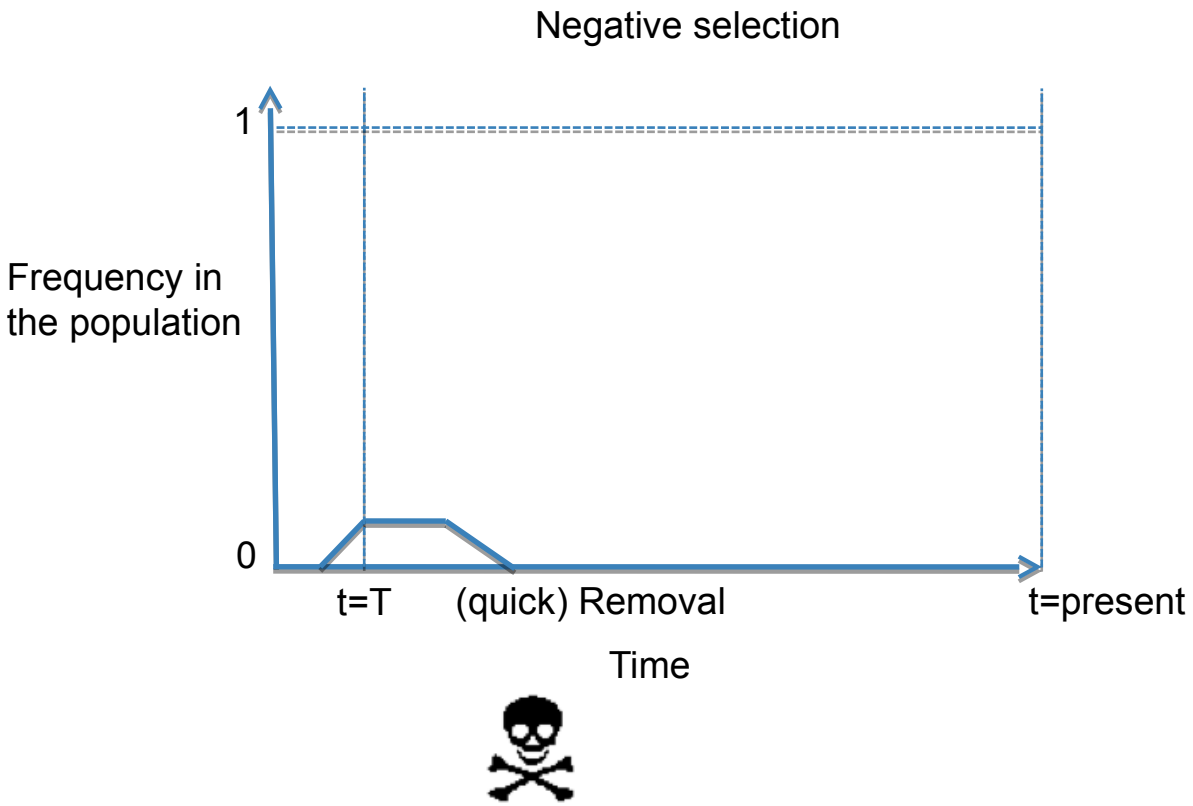
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

Heritable traits that increase the fitness of the become more common.

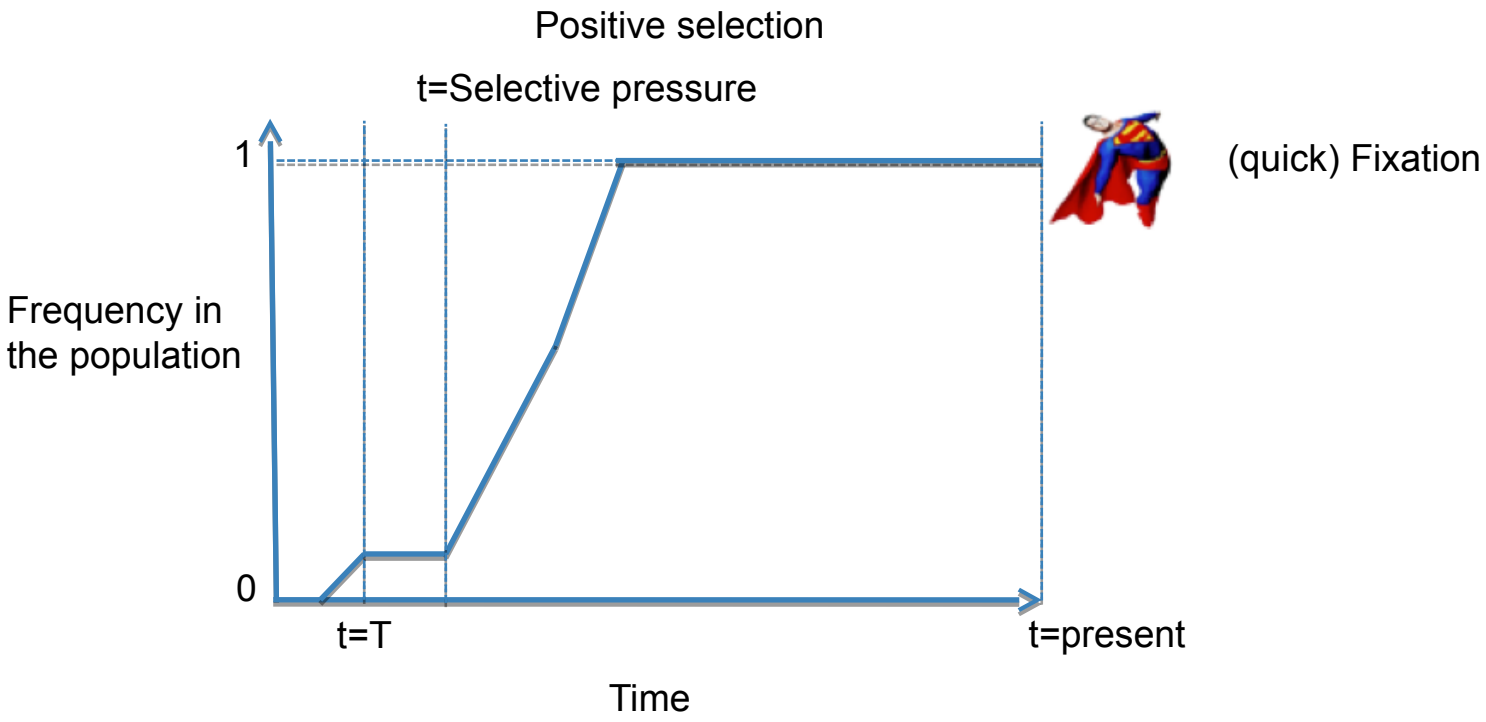
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

Heritable traits that increase the fitness of the become more common.

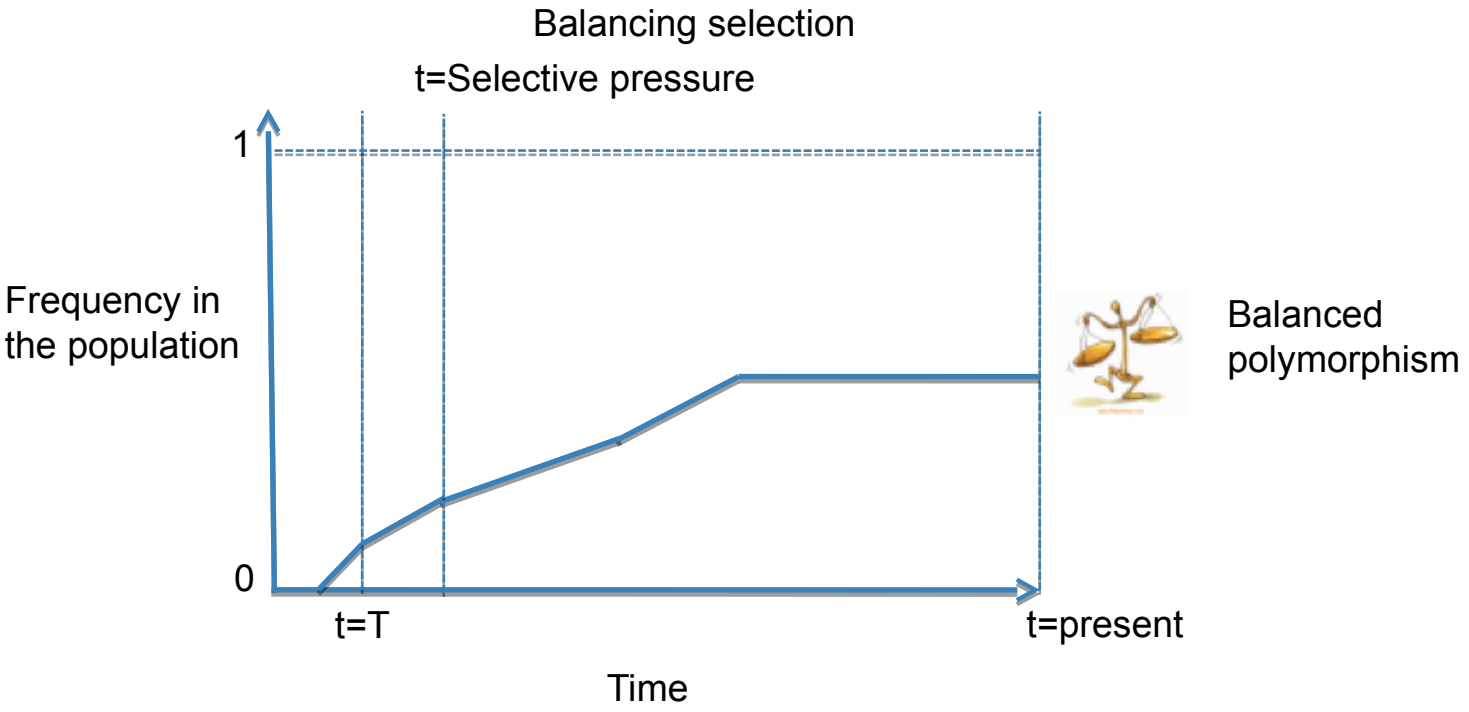
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

Heritable traits that increase the fitness of the become more common.

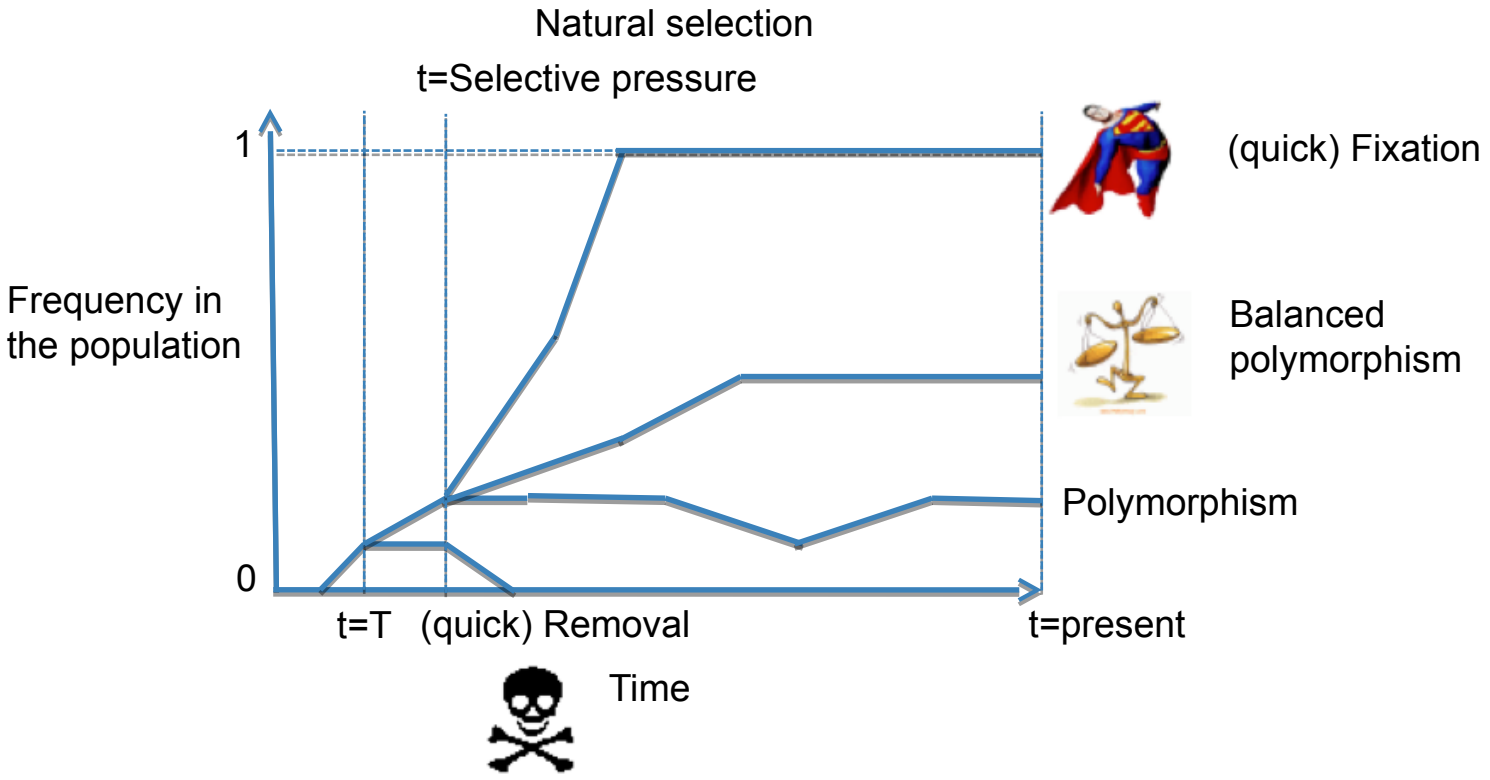
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

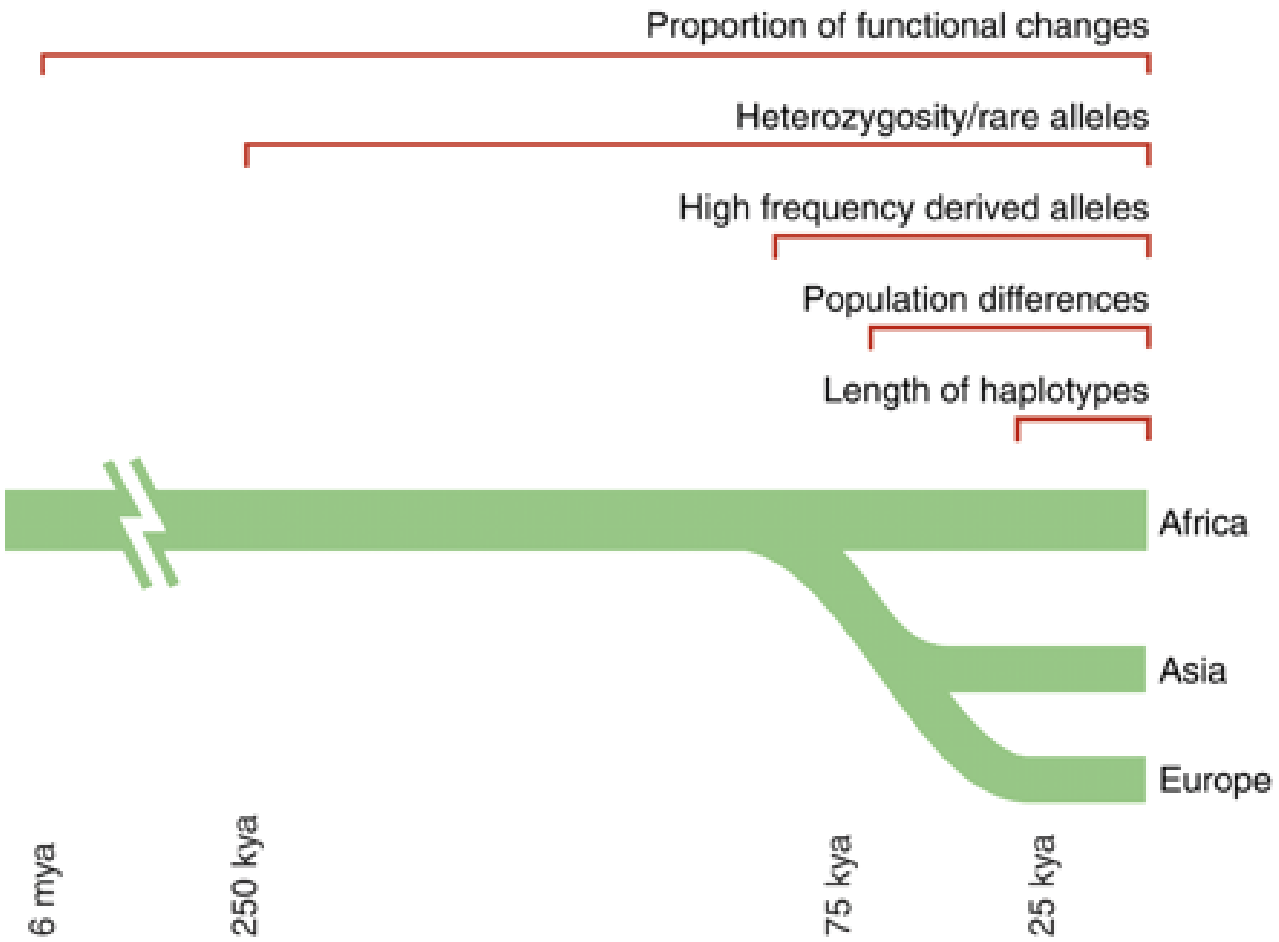
Heritable traits that increase the fitness of the become more common.

1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



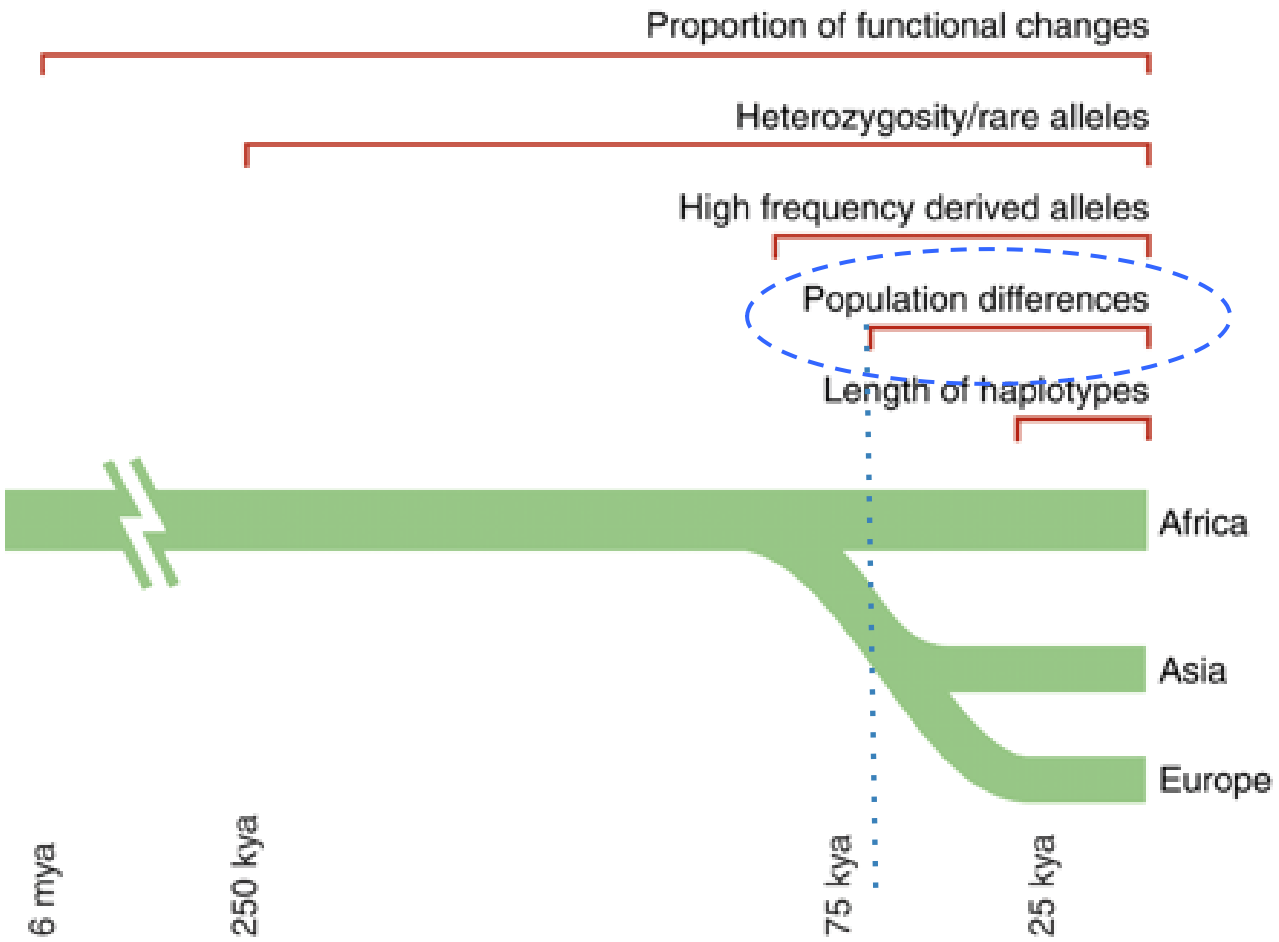
2) Sites targeted by natural selection are likely to harbour **functionality**

Methods to infer recent selection



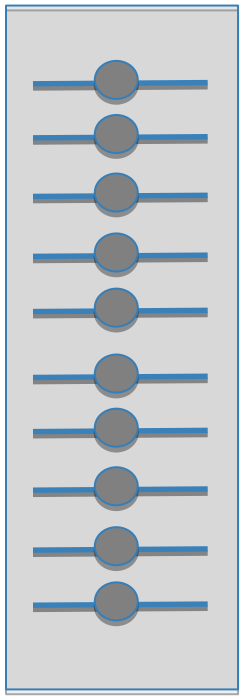
Sabeti et al. 2006 Science

Methods to infer recent selection

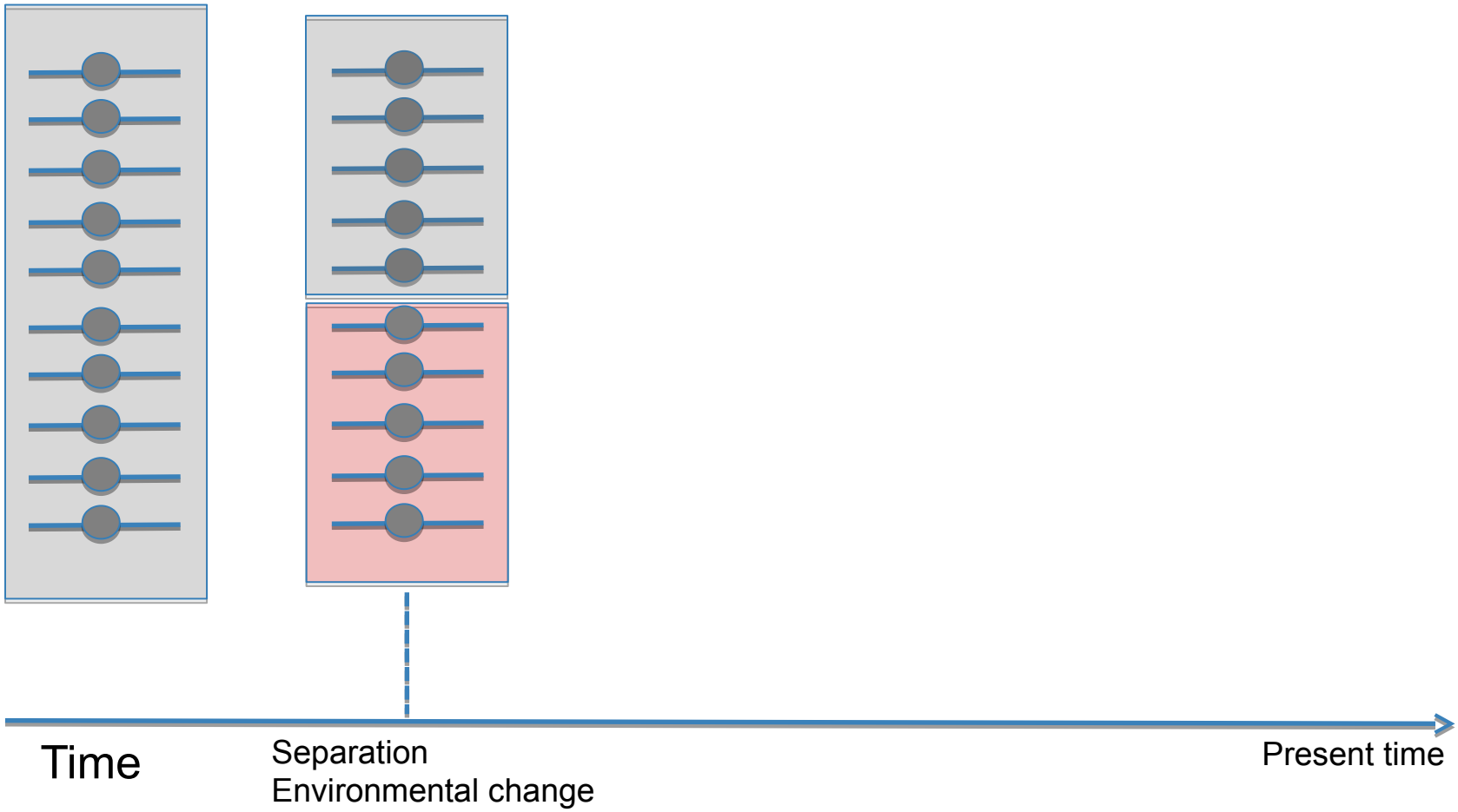


Sabeti et al. 2006 Science

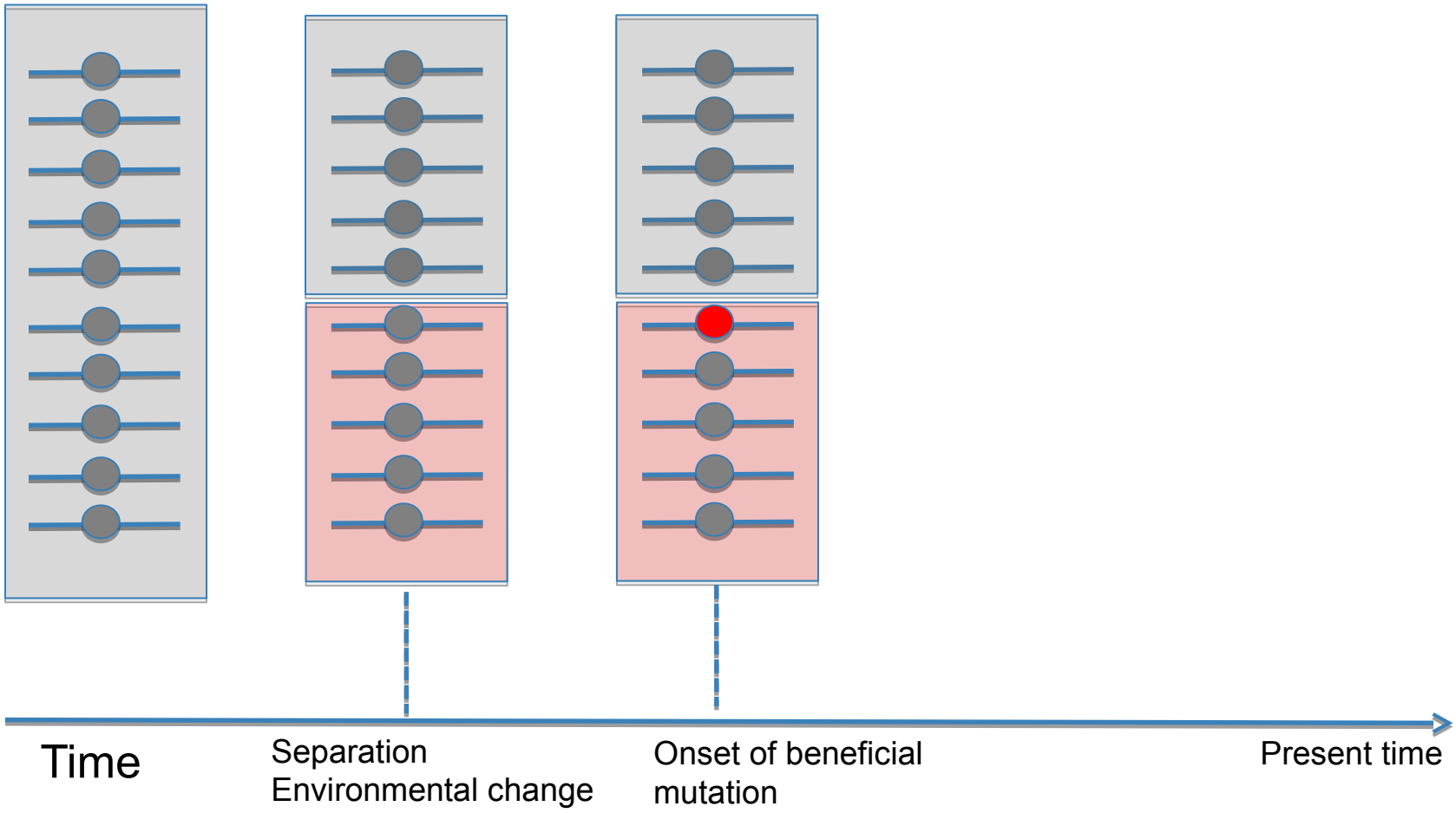
Allele frequency differentiation



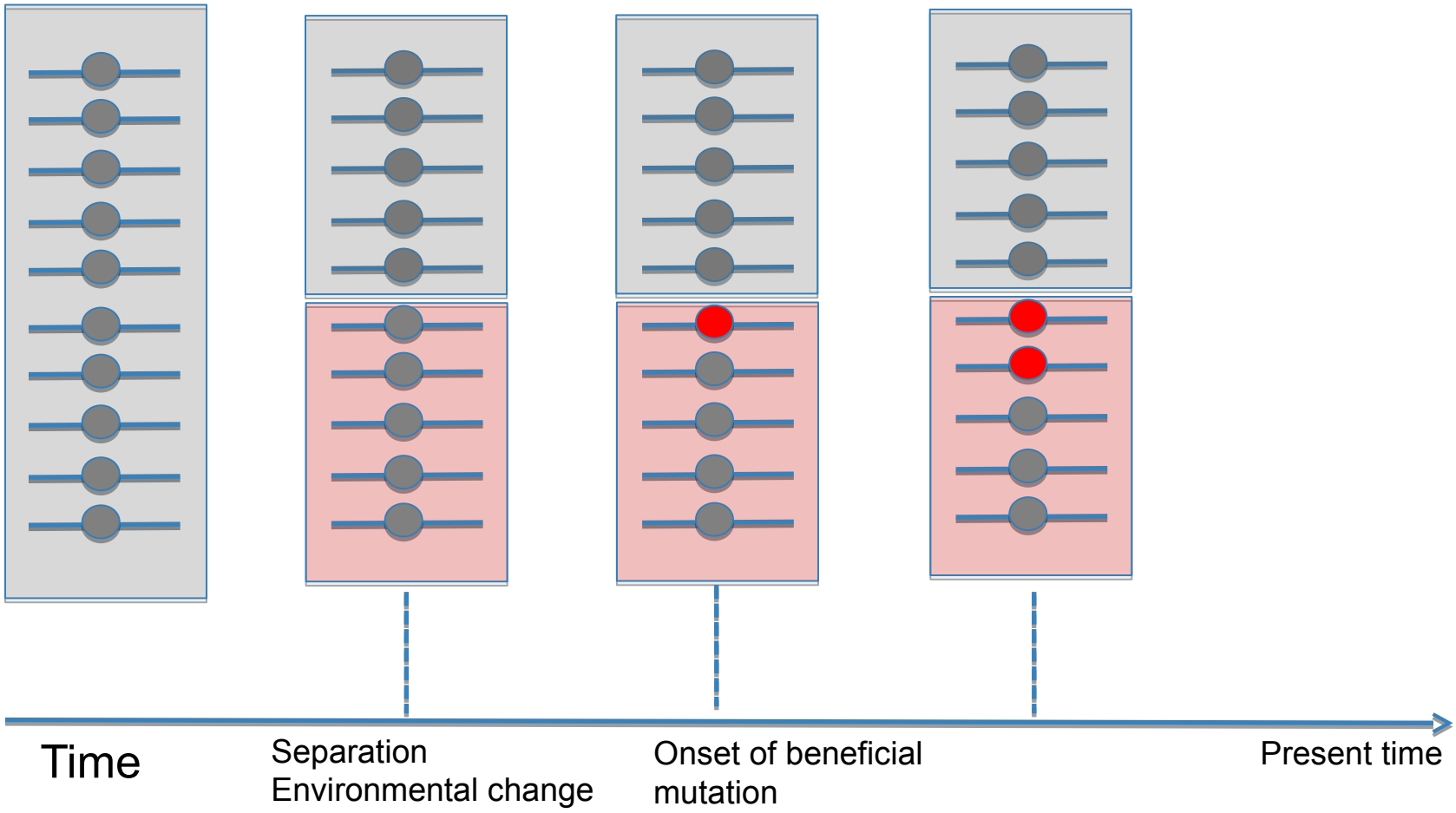
Allele frequency differentiation



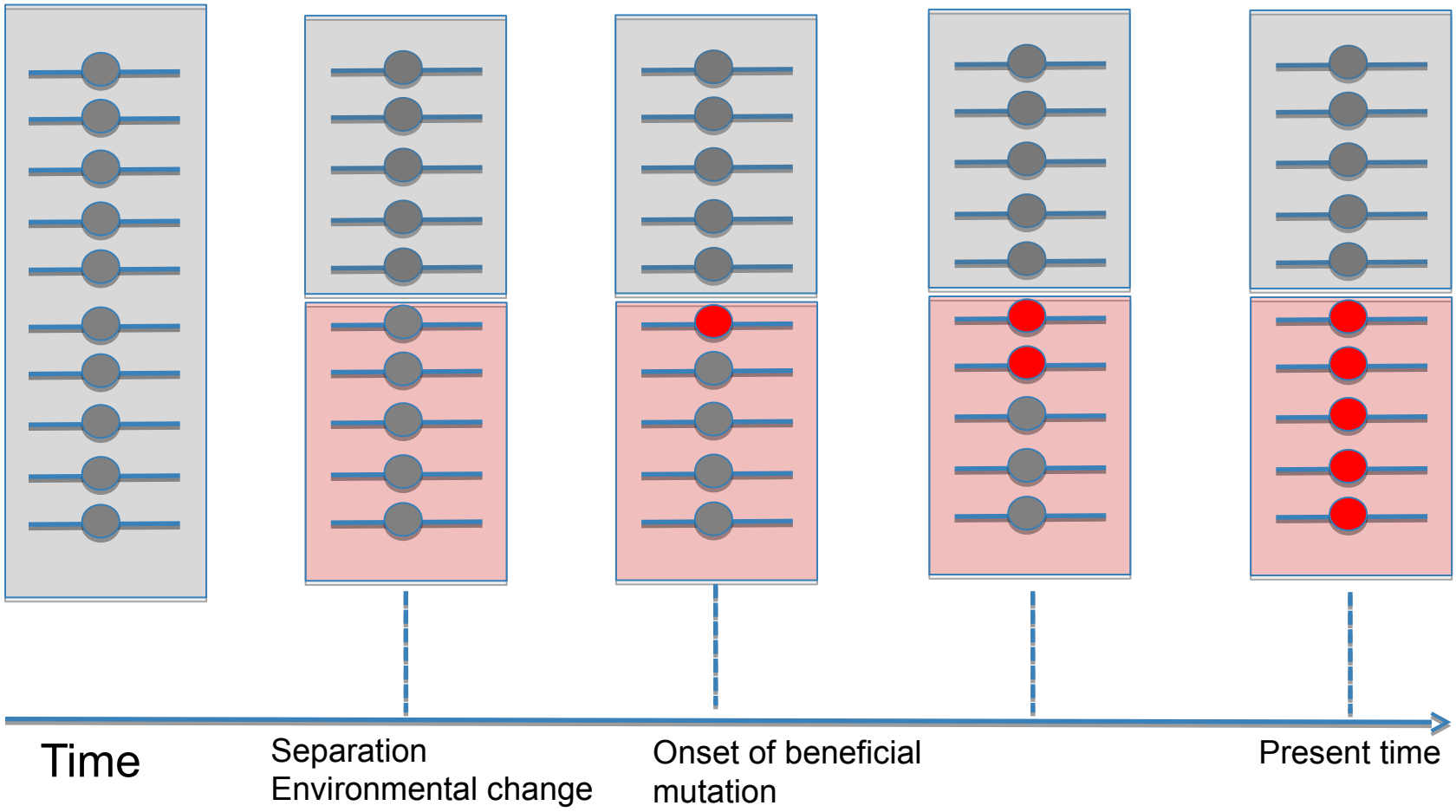
Allele frequency differentiation



Allele frequency differentiation



Allele frequency differentiation



F_{ST}

Common measure for quantifying population subdivision.

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

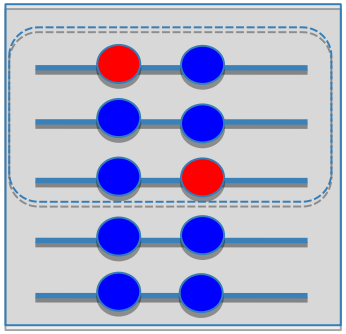
H_B: between populations

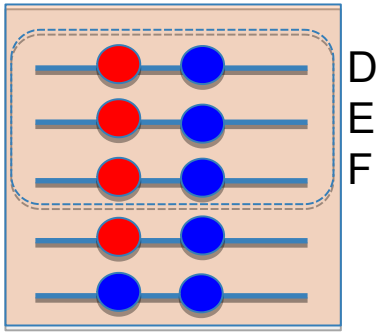
H_W: average within populations

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

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations





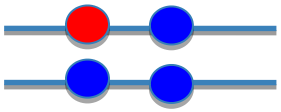
$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations

What is the variation within populations?

e.g. A vs B



The differ by 1 site

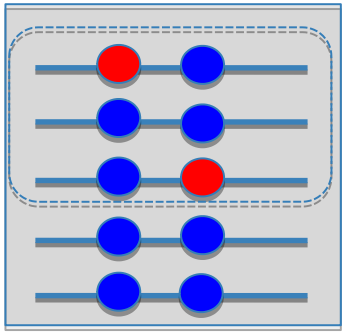
As defined by Eq. 3 in Hudson, Slatkin and Maddison 1992

Haplotype-based F_{ST}

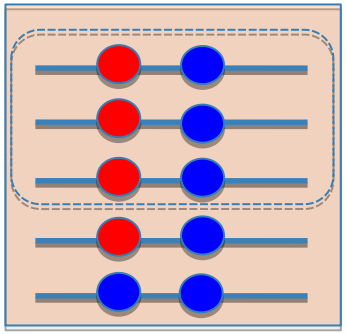
$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations



A
B
C



D
E
F

What is the variation within populations?

A	B
A	C
B	C

Mean=?

D	E
D	F
E	F

Mean=?

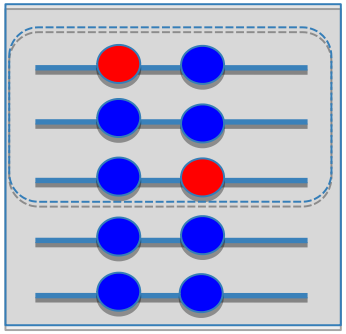
H_W is the average within-populations: ?

Haplotype-based F_{ST}

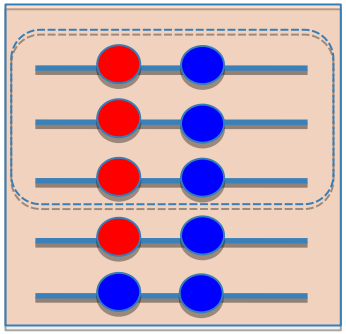
$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations



A
B
C



D
E
F

What is the variation within populations?

A	B	1
A	C	2
B	C	1

Mean=4/3

D	E	0
D	F	0
E	F	0

Mean=0/3

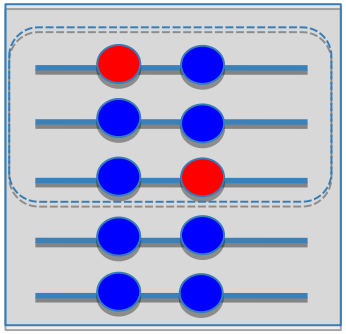
H_W is the average within-populations: (4/3+0/3)/2=2/3

Haplotype-based F_{ST}

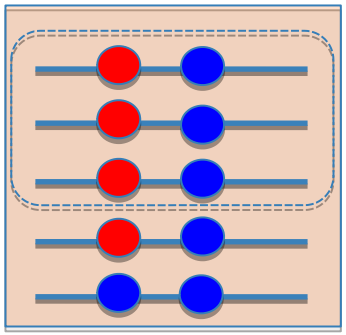
$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations



A
B
C



D
E
F

What is the variation between populations?

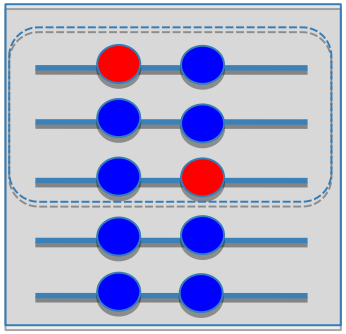
A	D	0
A	E	0
A	F	0
B	D	1
B	E	1
B	F	1
C	D	2
C	E	2
C	F	2

Mean=9/9

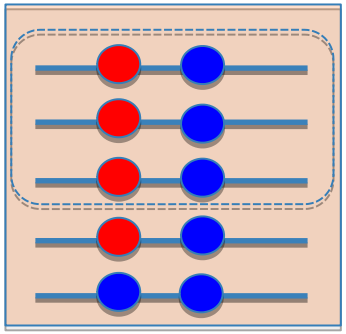
H_B is the average between-populations: 9/9=1

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations



A
B
C

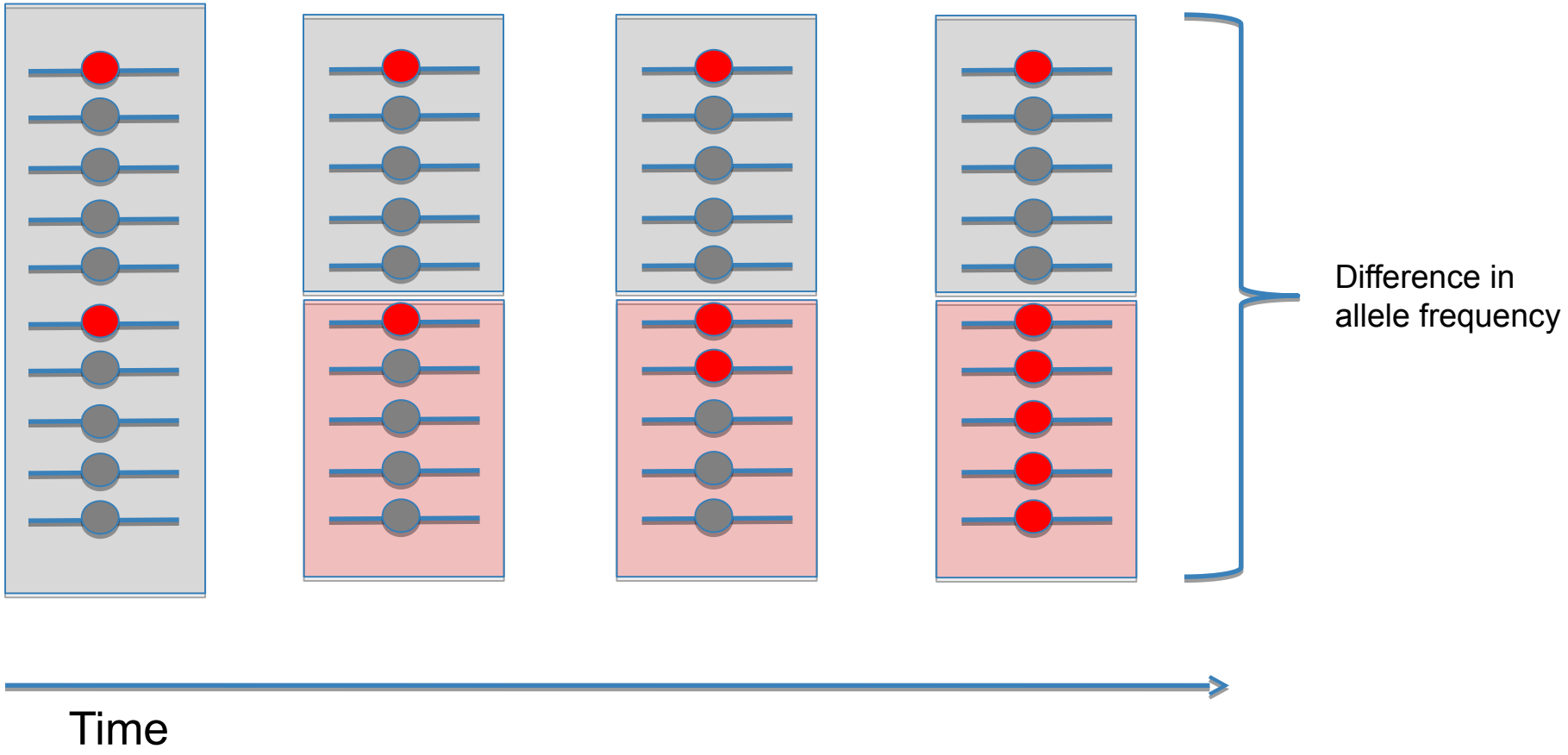


D
E
F

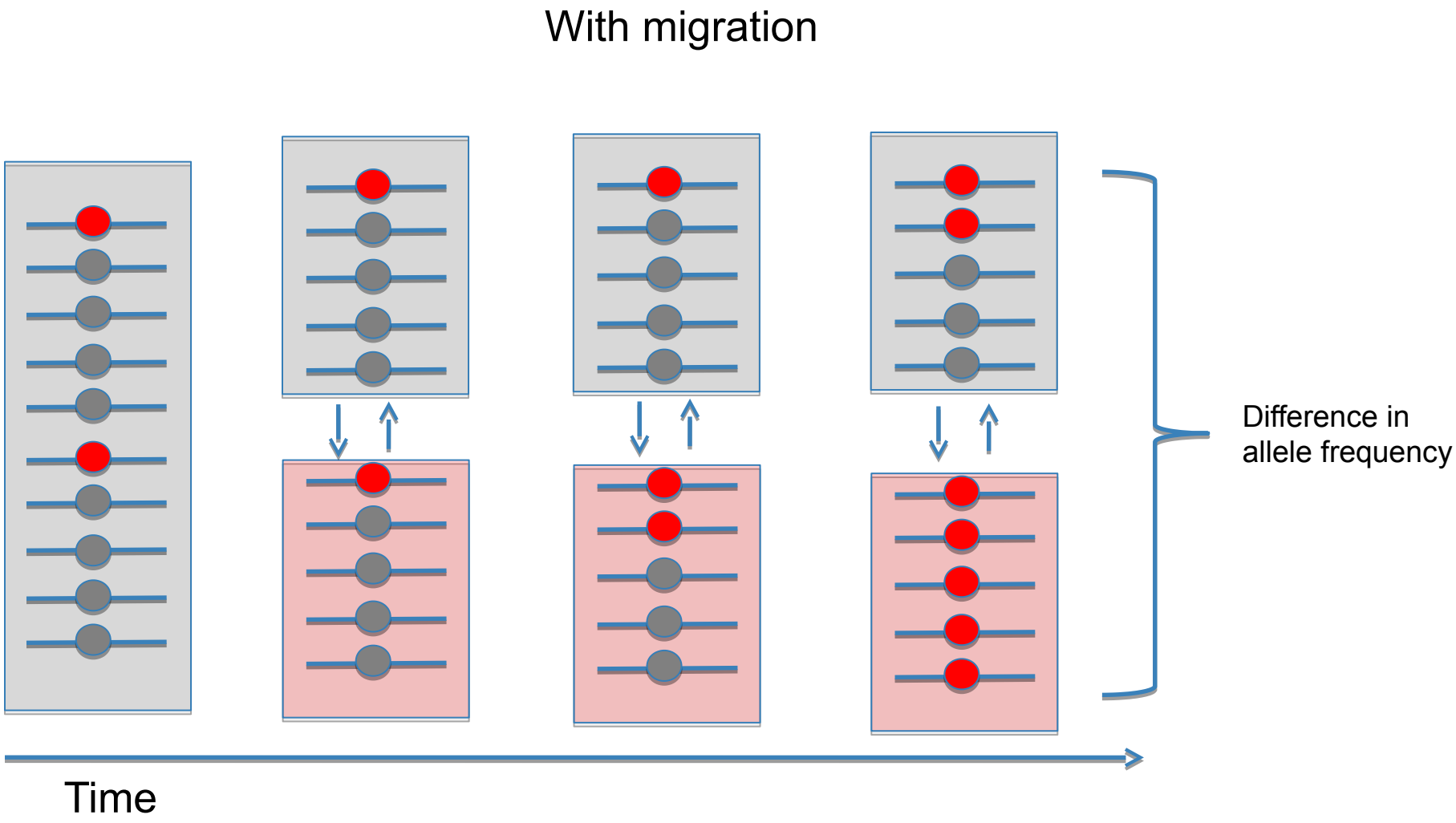
$F_{ST} = 1 - (H_W / H_B) = 1 - ((2/3)/1) = 1/3 \sim 0.33$

Allele frequency differentiation

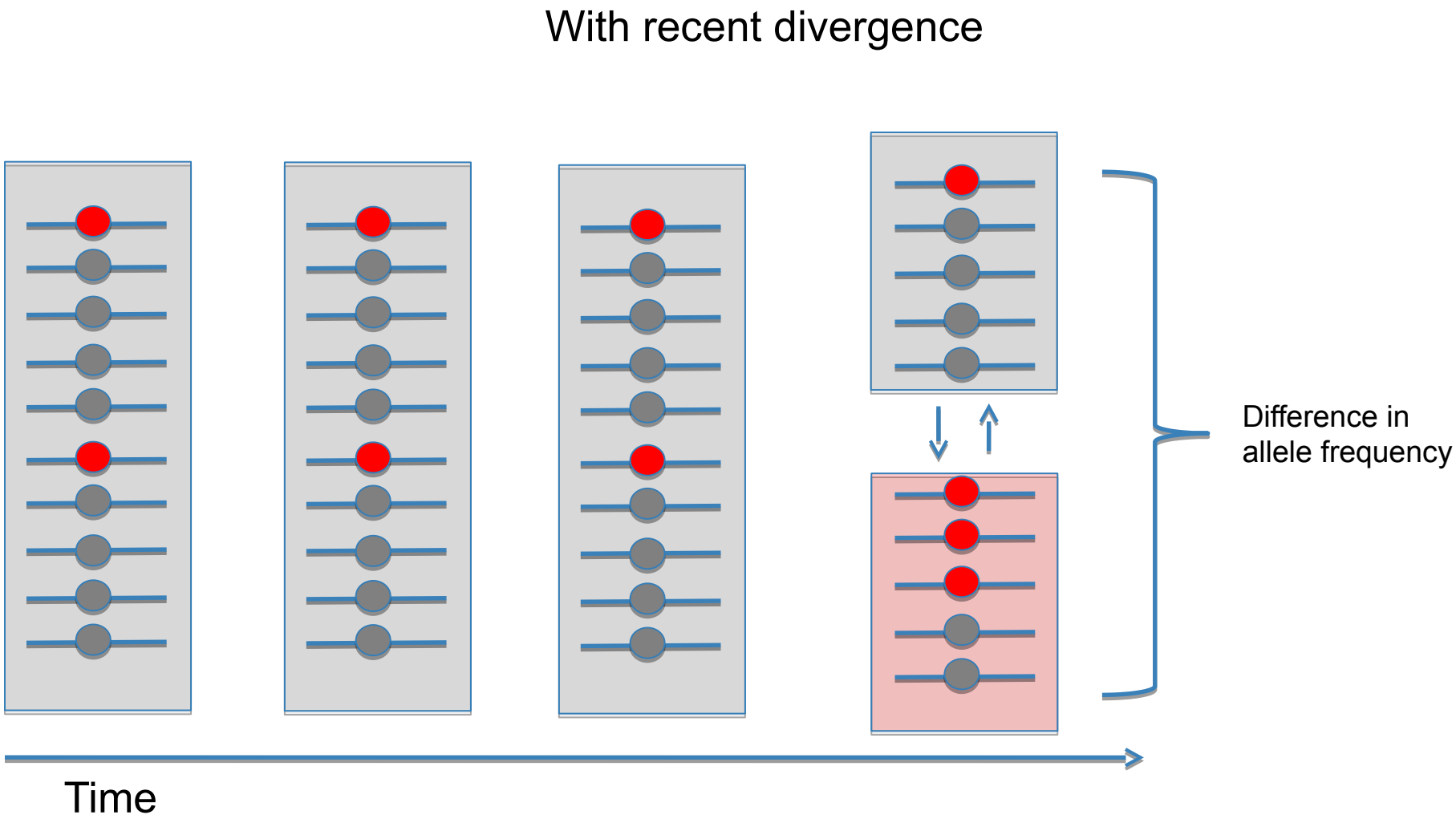
From standing variation



Allele frequency differentiation

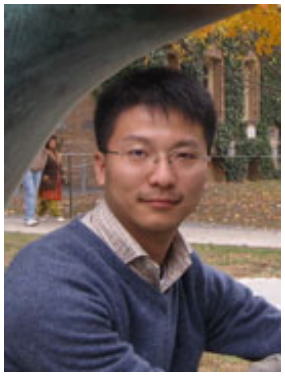


Allele frequency differentiation



Population genetic differentiation

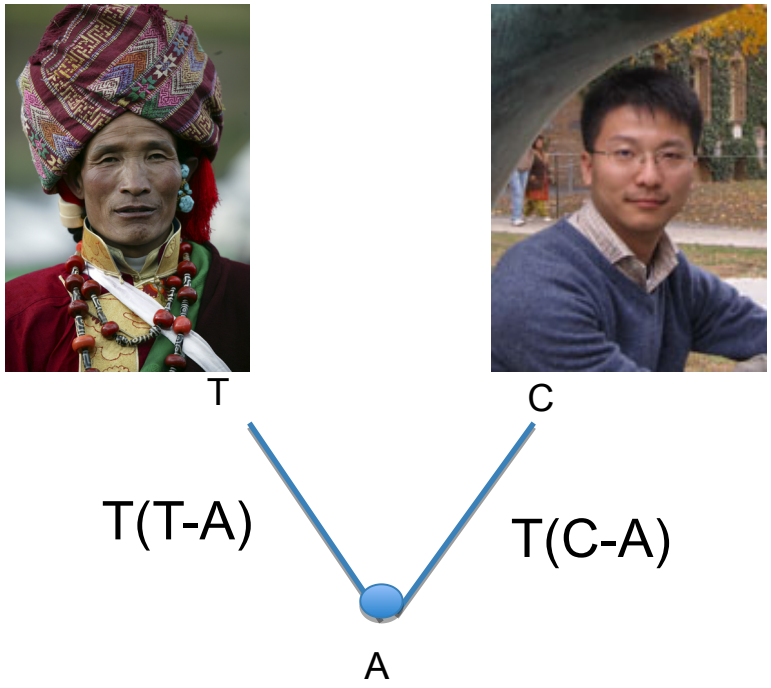
T



C

F_{ST}(T-C)

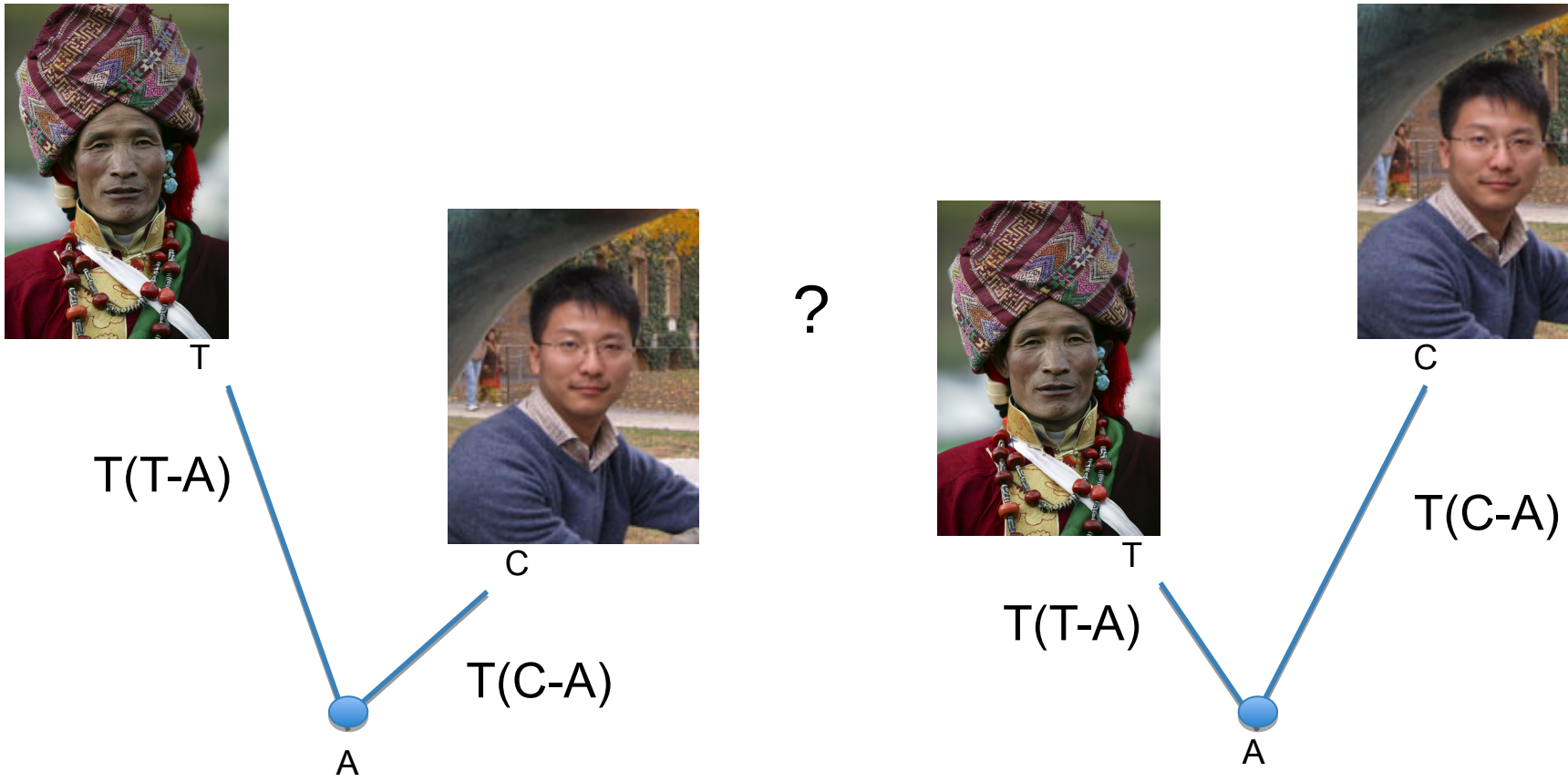
Population genetic differentiation



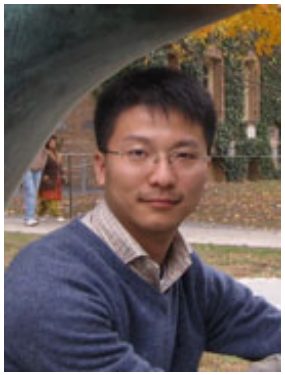
F_{ST}(T-C)~T(T-A-C)

Population genetic differentiation

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



Population genetic differentiation



T

C

T(T-A)

A

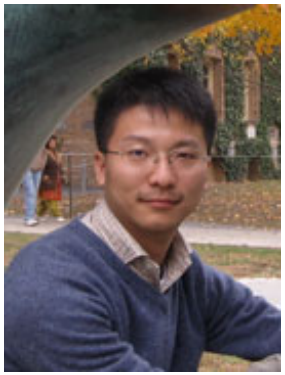
T(C-A)

T(E-A)

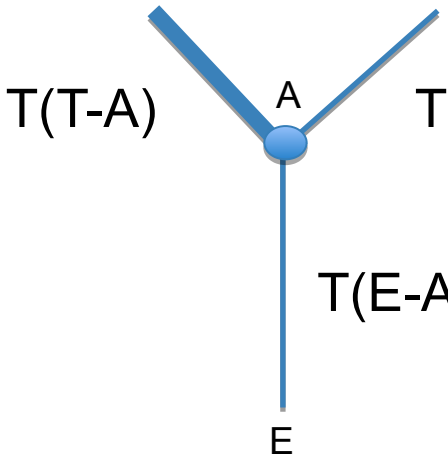
E

Population genetic differentiation

T



C

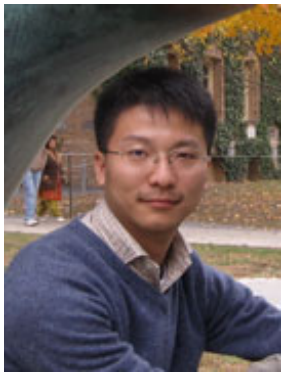


$T(T\text{-}A\text{-}C)=-\log(1-F_{ST}(T\text{-}C))$

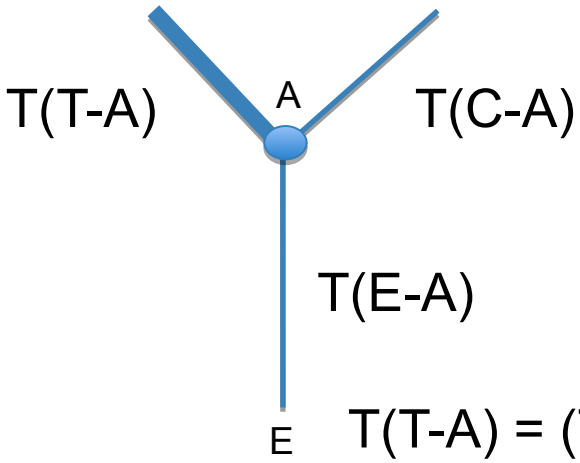
T(T-A)?

Population genetic differentiation

T



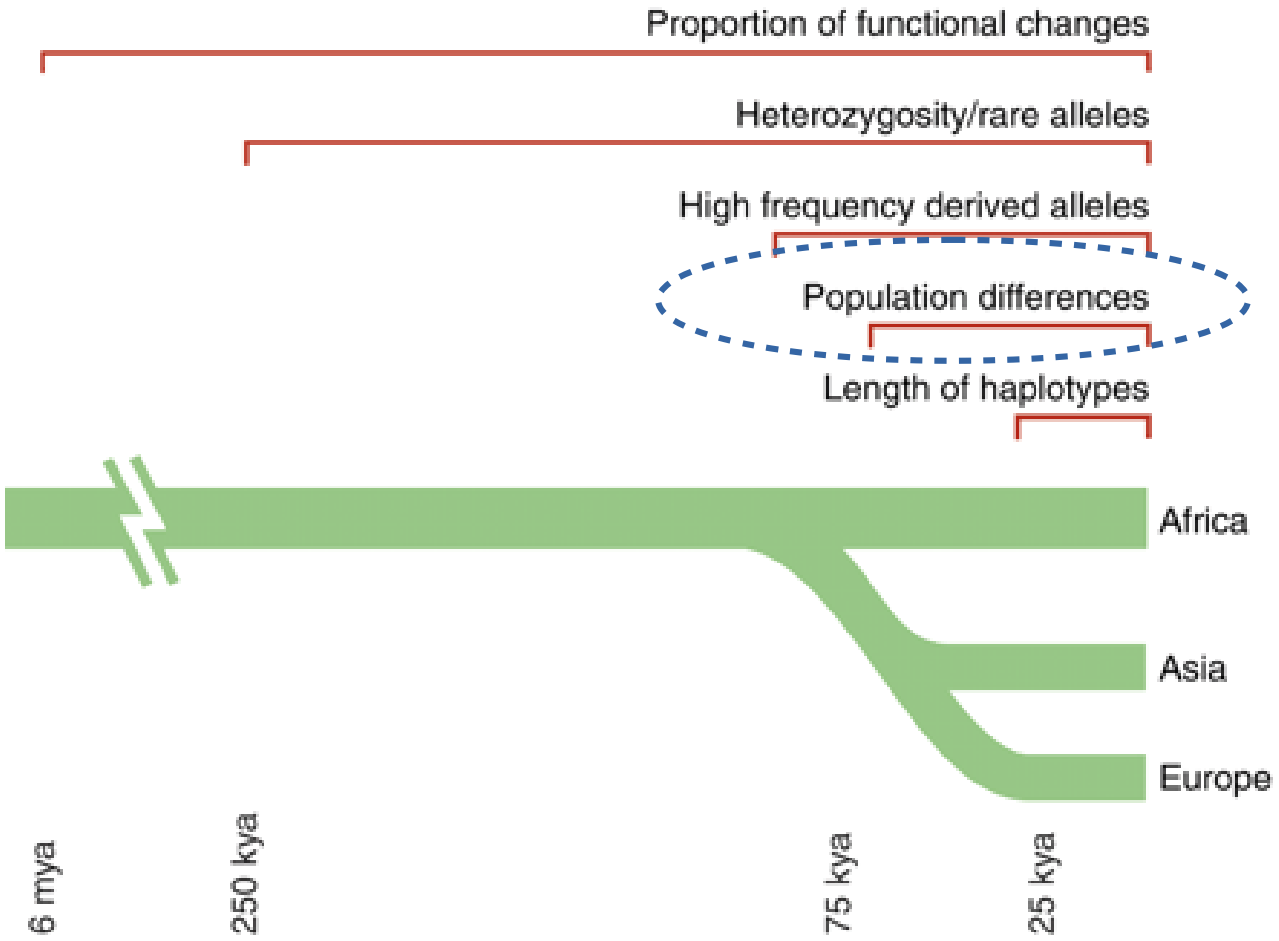
C



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

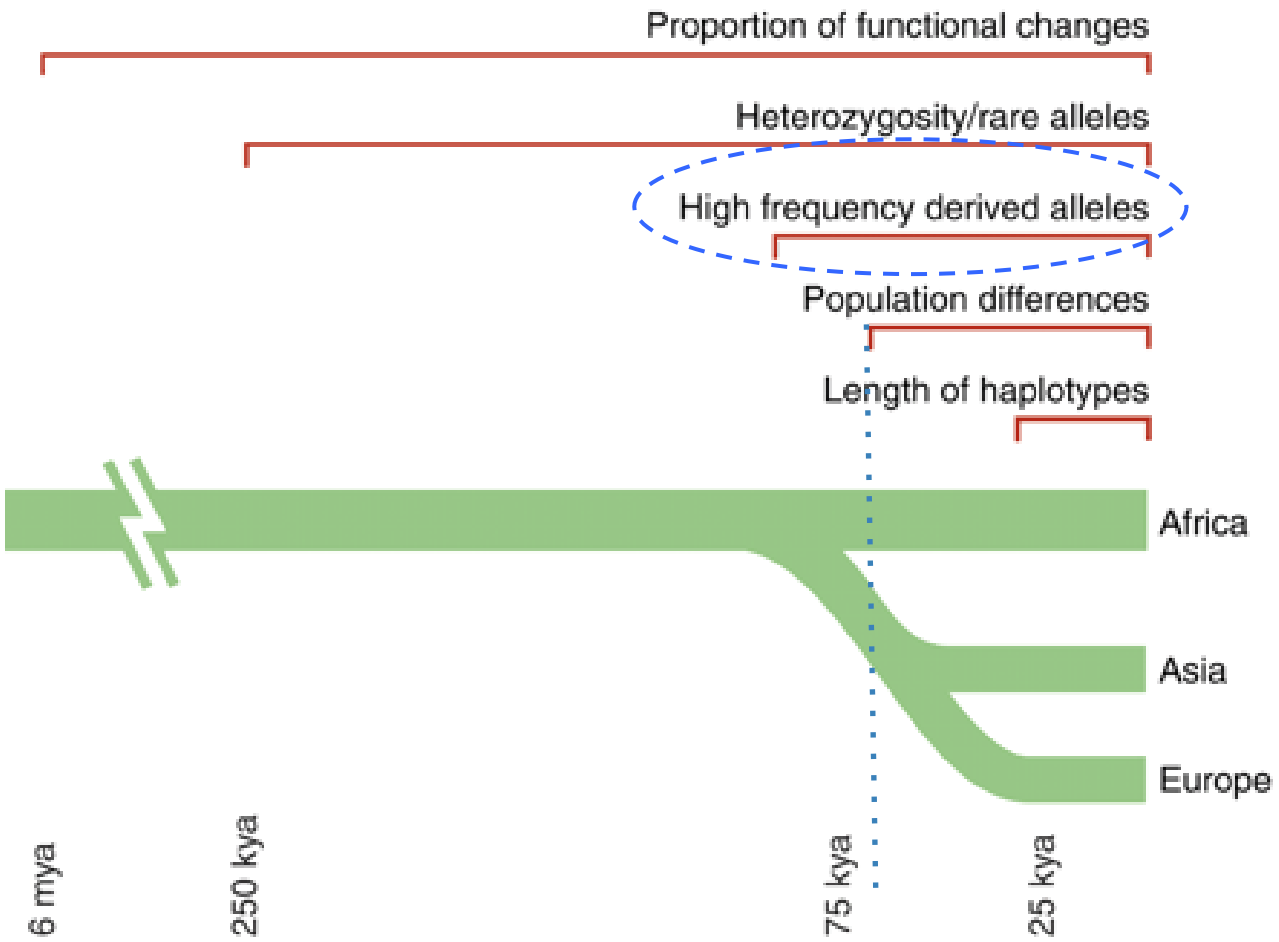
$T(T\text{-}A) = (T(T\text{-}E)+T(C\text{-}T)-T(C\text{-}E))/2$

Inference of positive selection



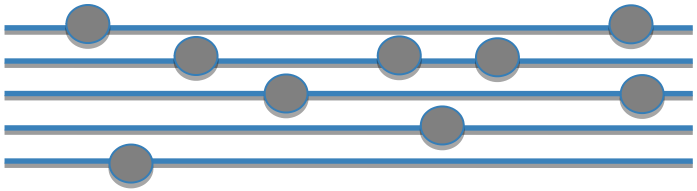
How can we calculate these summary statistics from low-depth sequencing data?

Methods to infer selection



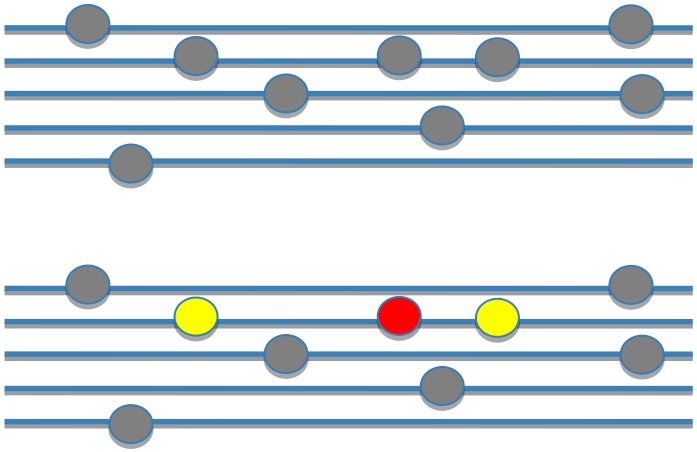
Sabeti et al. 2006 Science

Positive selection: effect on haplotypes



$t<T_{sel}$

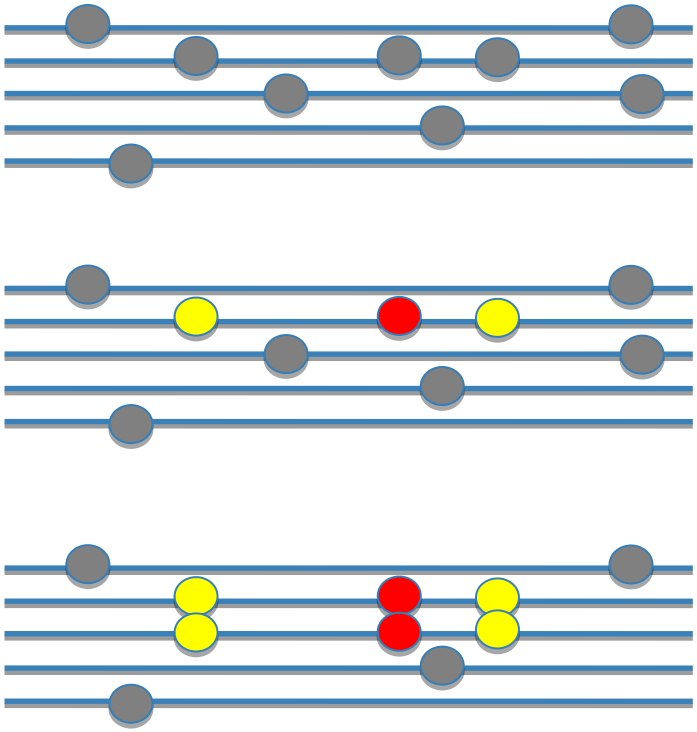
Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

Positive selection: effect on haplotypes

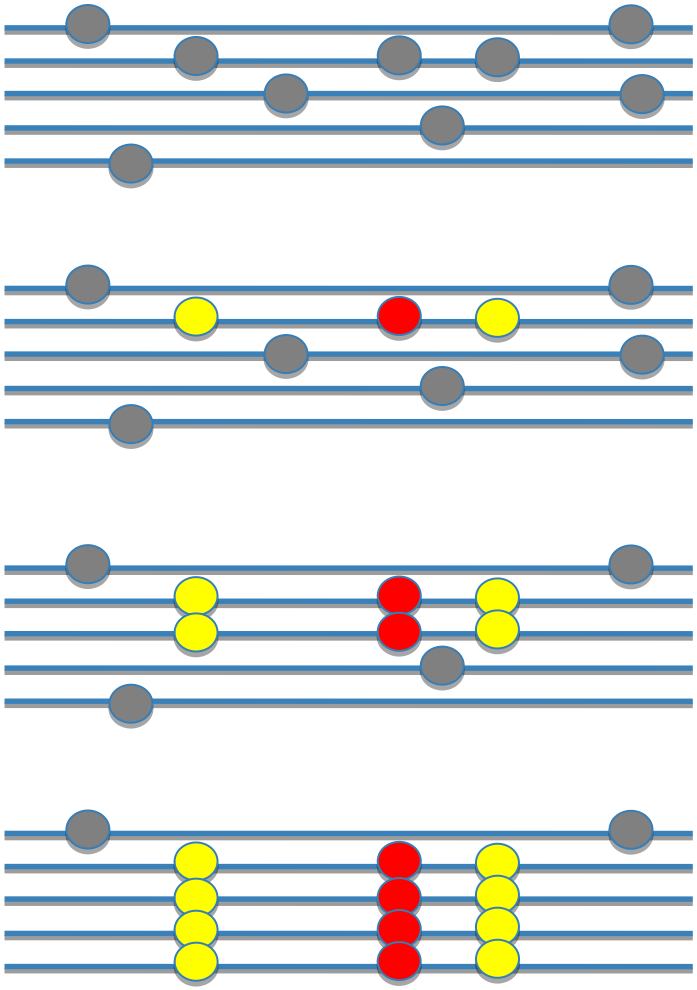


$t<T_{sel}$

$t=T_{sel}$

$t>T_{sel}$

Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

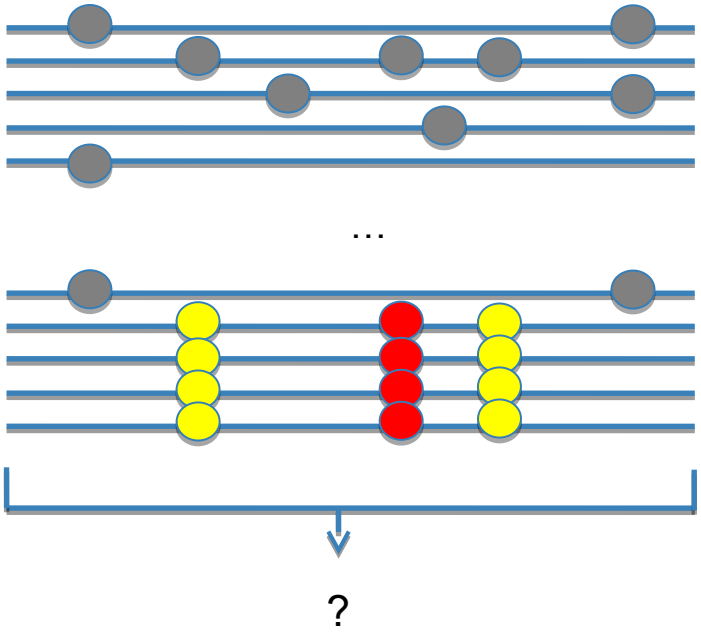
$t>T_{sel}$

$t>>T_{sel}$

Selective sweep

Genetic hitch-hiking

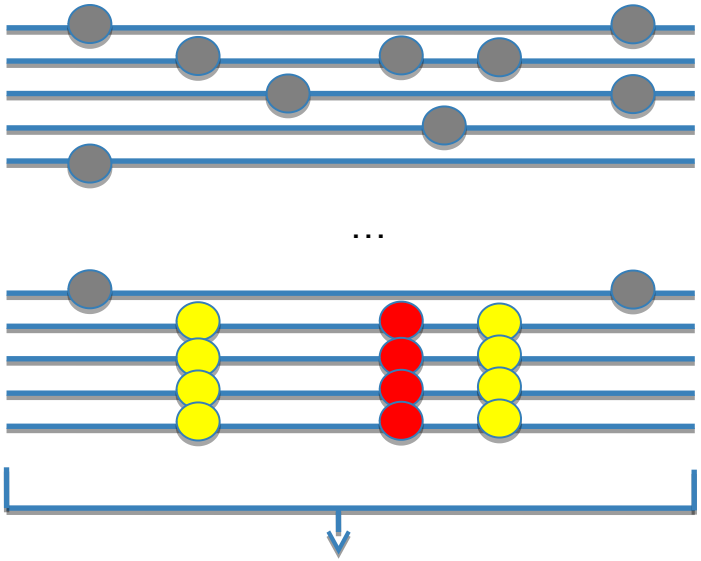
Positive selection



$t<T_{sel}$

$t>>T_{sel}$

Positive selection



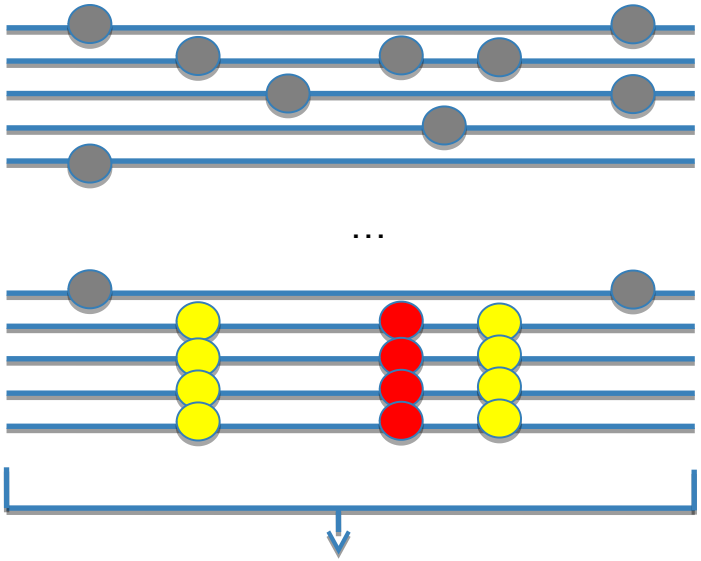
$t<T_{sel}$

…

$t>>T_{sel}$

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

Positive selection



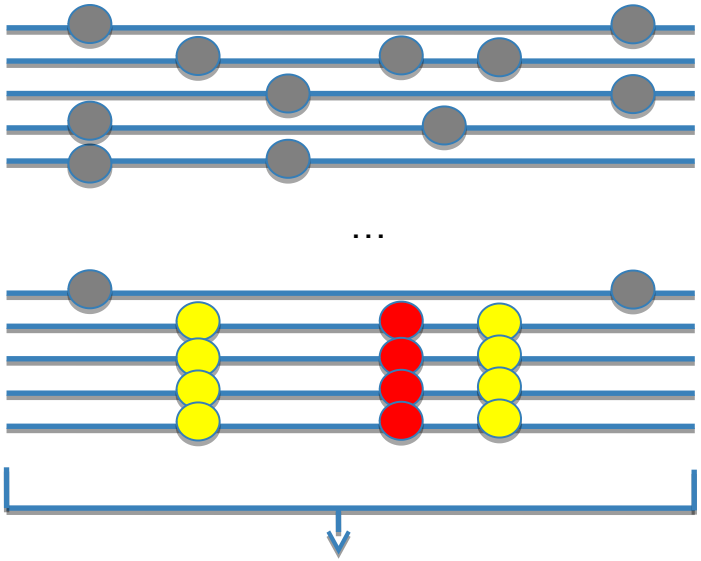
- Reduction of polymorphisms levels
(e.g. from 7 to 5 SNPs)

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

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

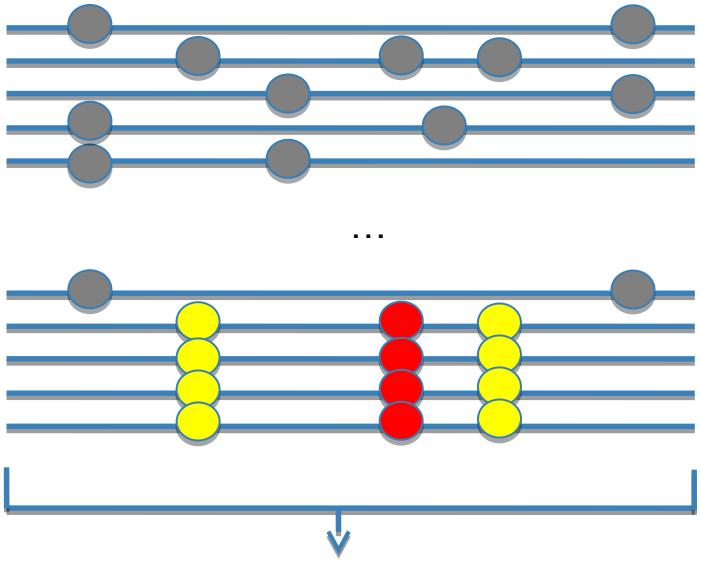
$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Positive selection



- Reduction of polymorphisms levels (Theta)
- ?

Positive selection

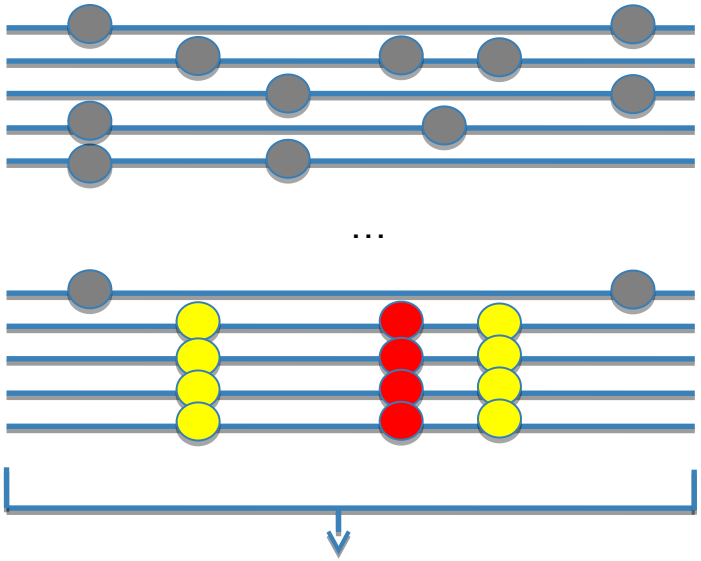


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

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



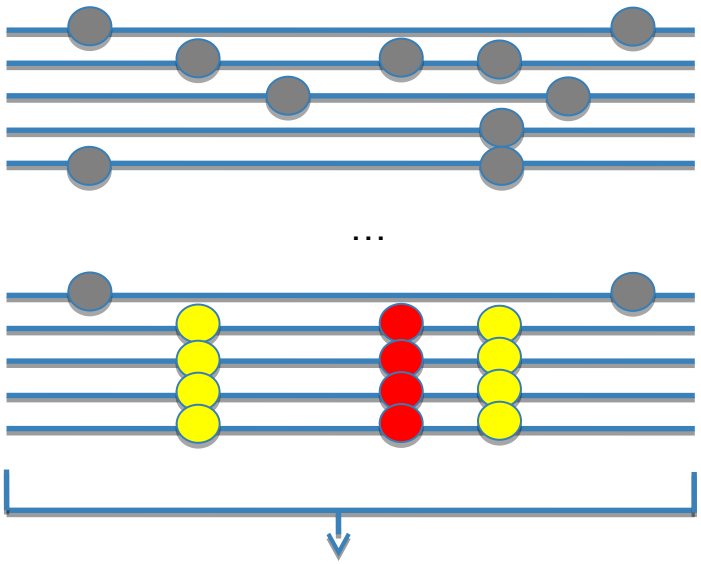
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

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

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

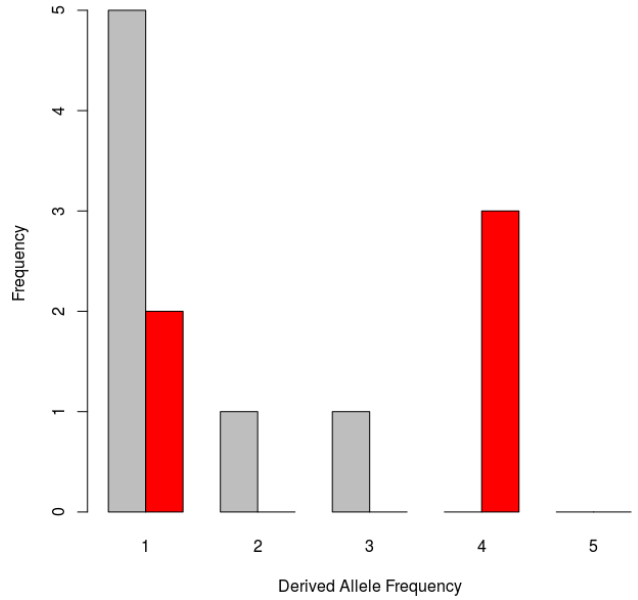
D<0 is suggestive of an excess of low-frequency variants

The Site Frequency Spectrum

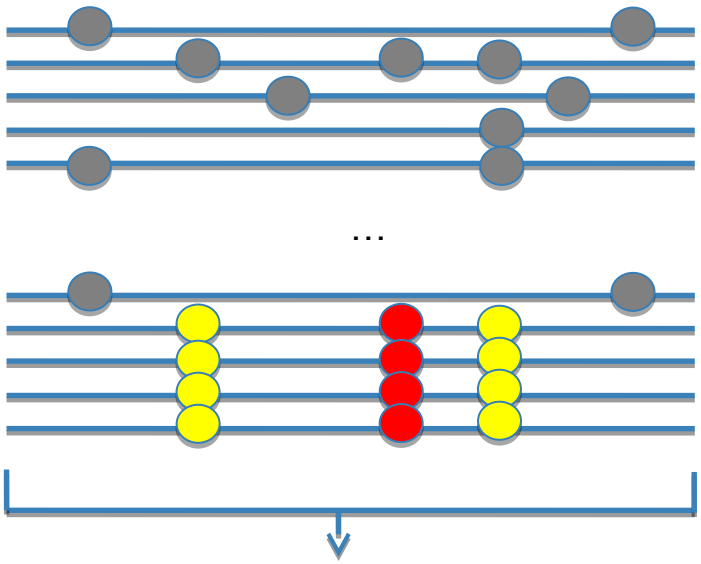


$t<T_{sel}$

$t>>T_{sel}$



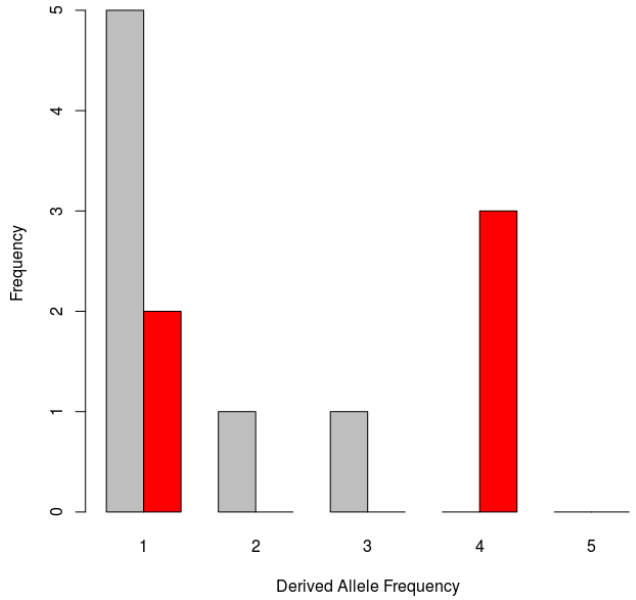
The Site Frequency Spectrum



t<T_{sel}

...

t>>T_{sel}



Tajima's D?

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

~0.33=1/3

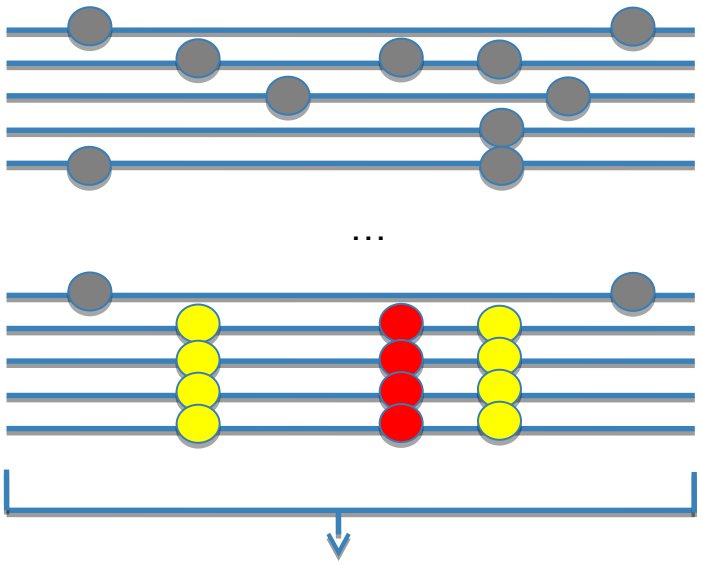
$$\theta_W = \frac{K}{a_n}$$

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

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

= 10, the number of comparisons you need to make

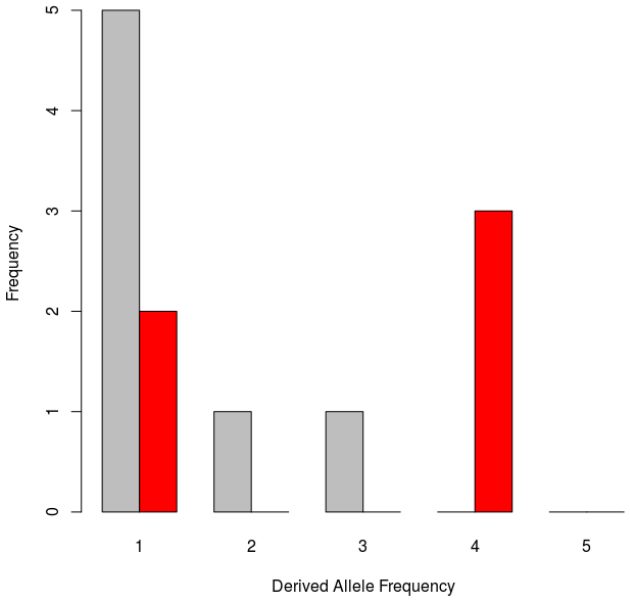
The importance of being…
The Site Frequency Spectrum



$t<T_{sel}$

…

$t>>T_{sel}$



K=5

a_n=1/1 + 1/2 + 1/3 + 1/4=(12+6+4+3)/12=25/12

Theta=5/(25/12)=12/5

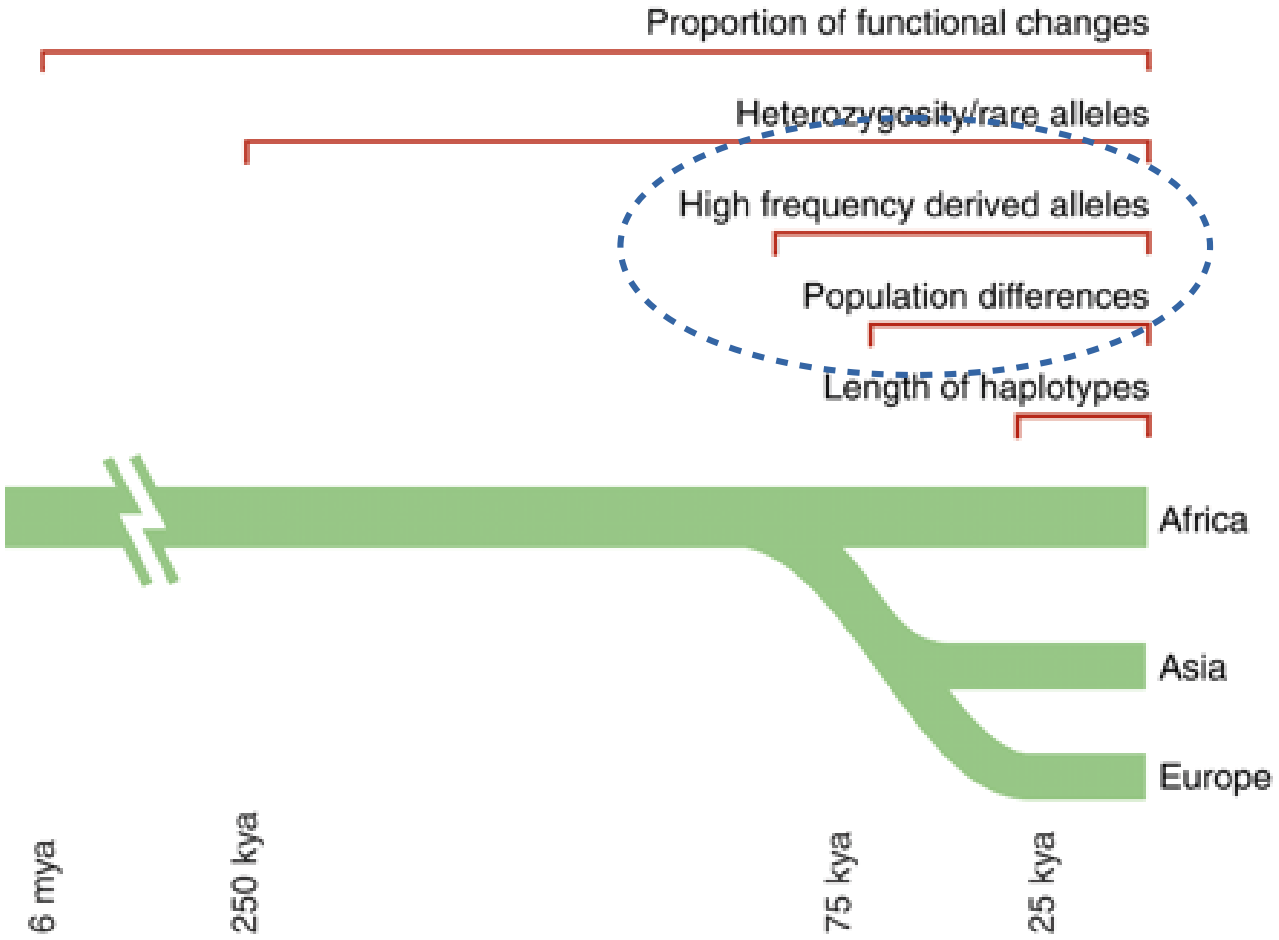
Pi=(5+5+5+5+0+0+0+0+0+0)/10=20/10=2

sd(D)=1/3

D=(2-12/5)/(1/3)=((10-12)/5)*3=-6/5=-1.2

D<0

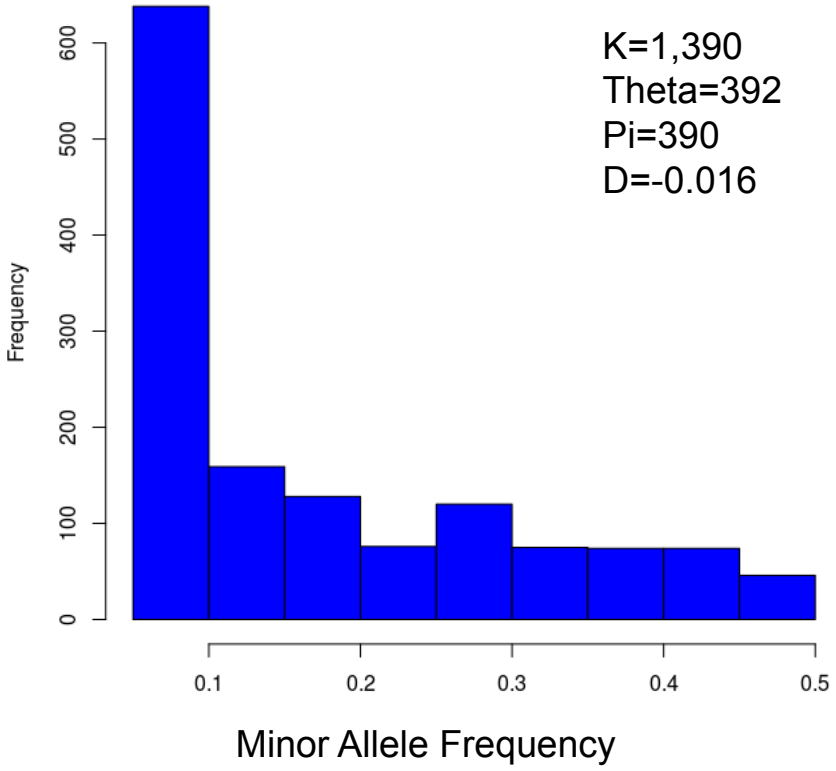
Inference of positive selection



How can we calculate these summary statistics from low-depth sequencing data?

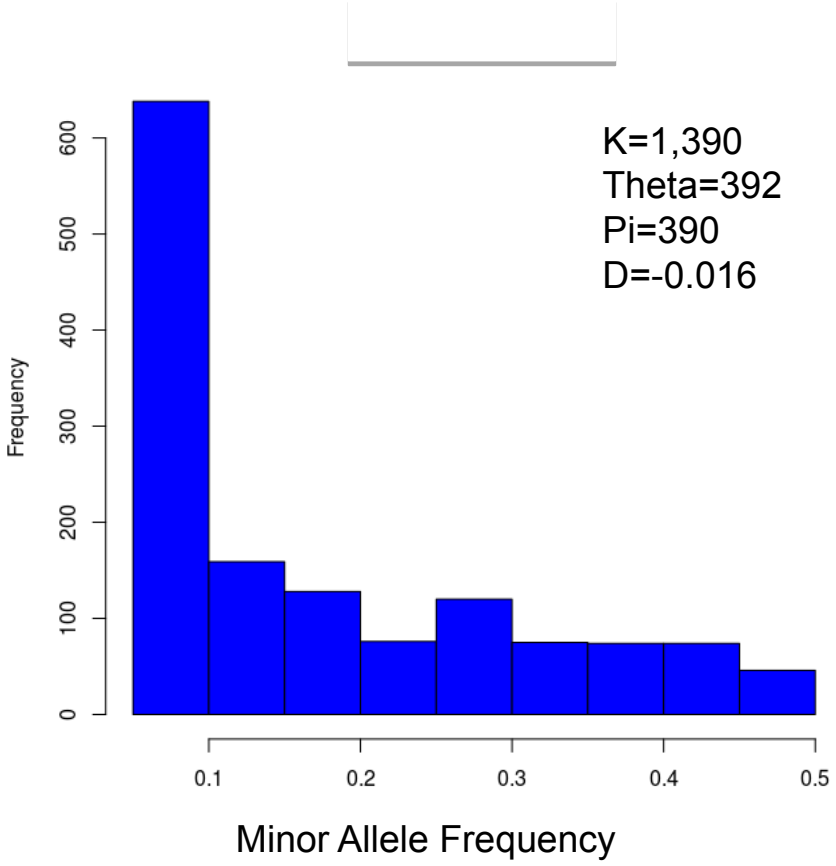
Confounding factor

n=20; L=500kbp; no selection

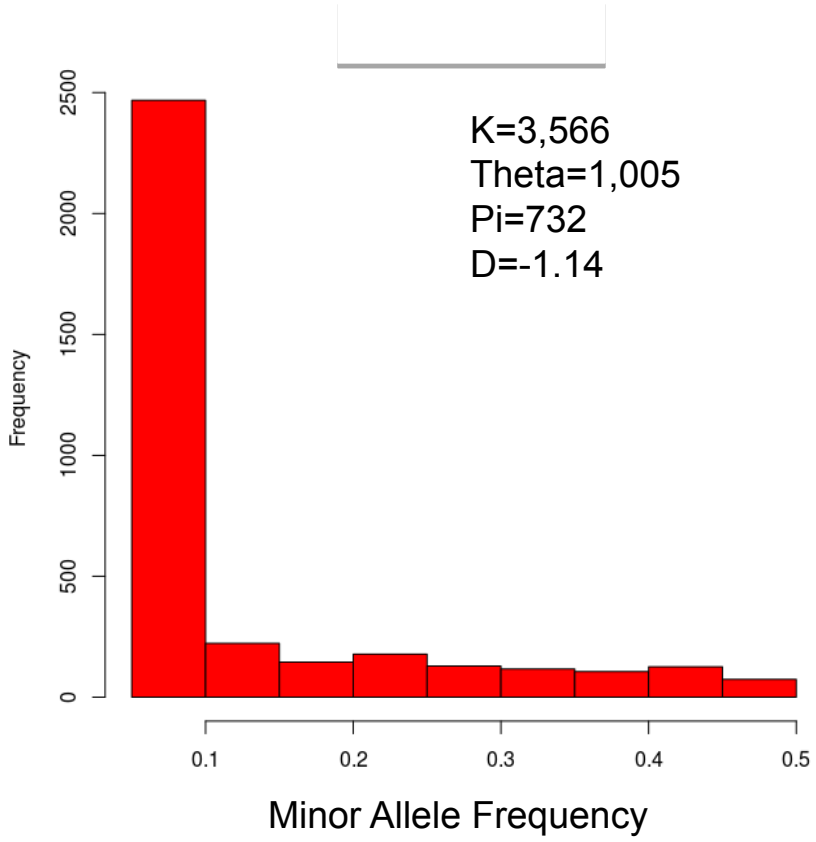


Confounding factor

n=20; L=500kbp; no selection

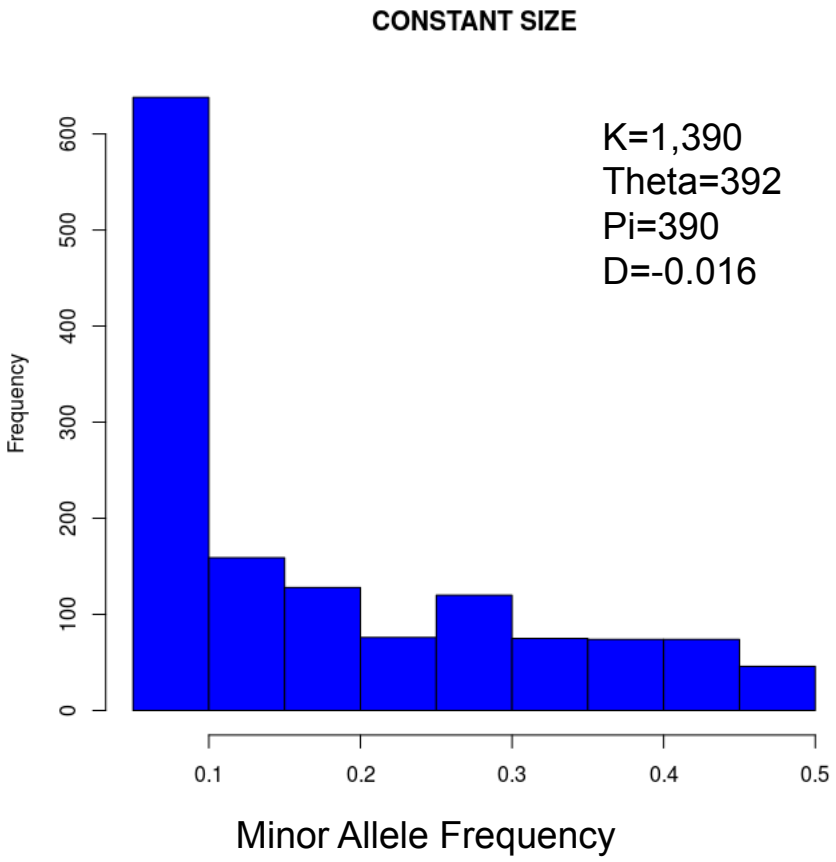


n=20; L=500kbp; no selection

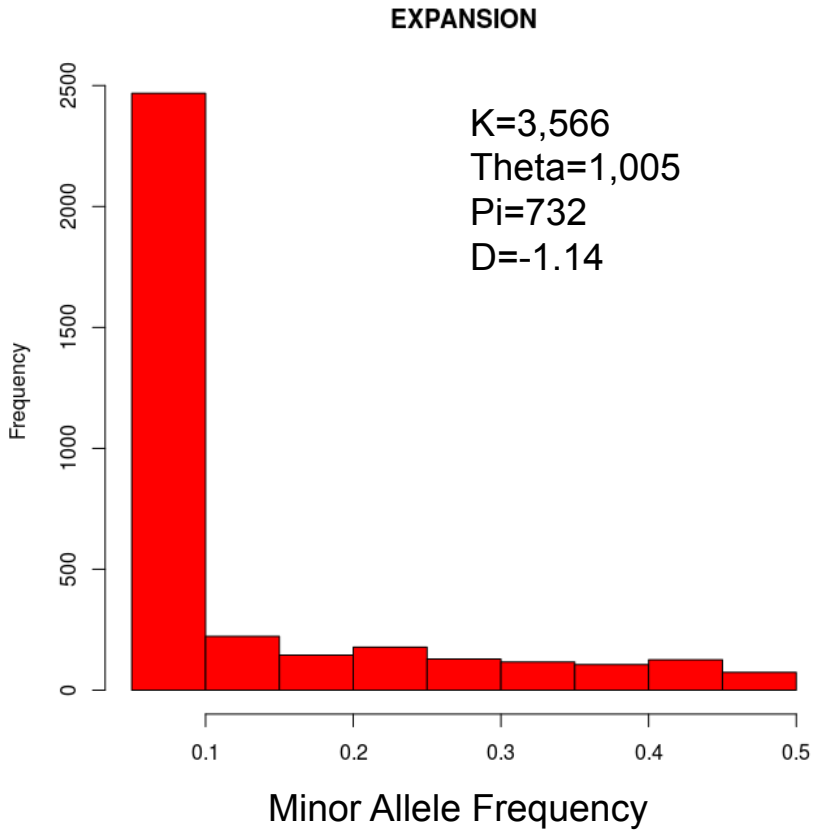


Demography matters!

n=20; L=500kbp; no selection



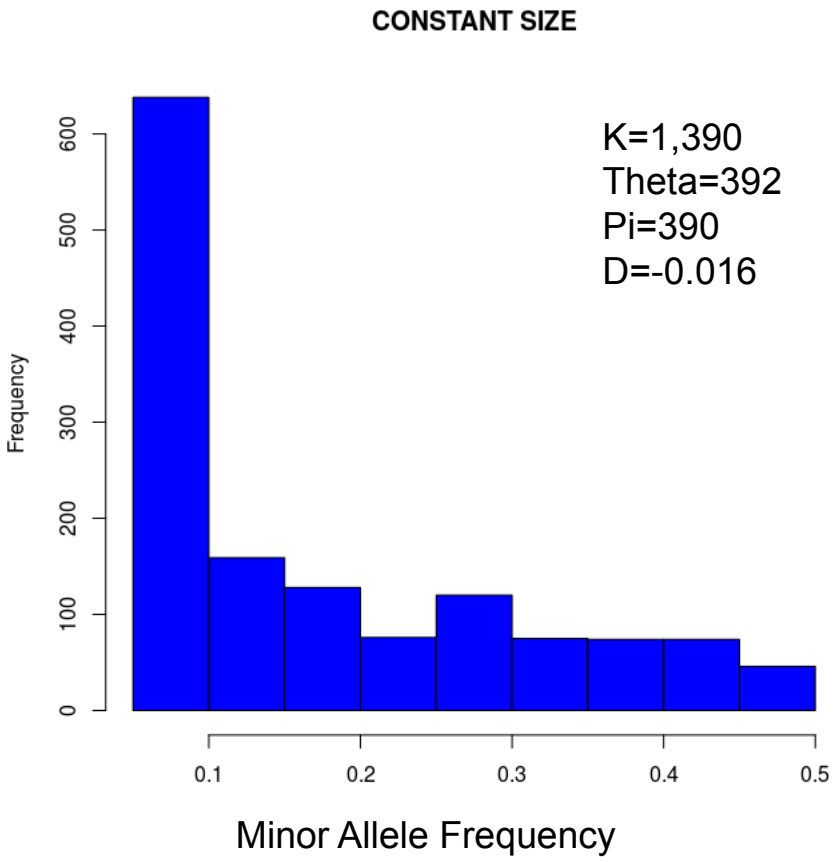
n=20; L=500kbp; no selection



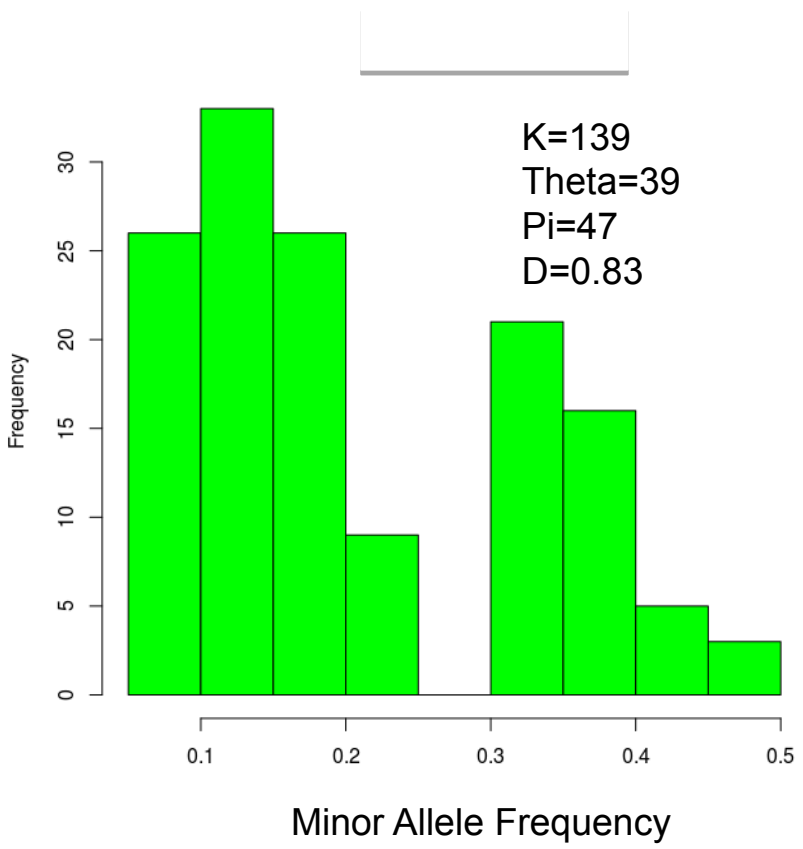
- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

Demography matters?

n=20; L=500kbp; no selection

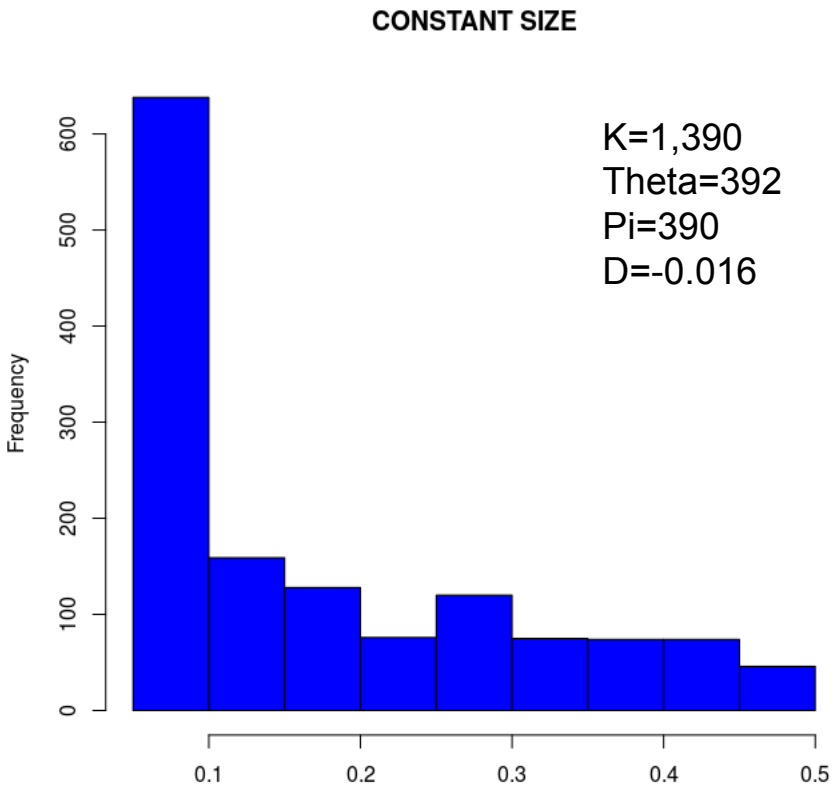


n=20; L=500kbp; no selection

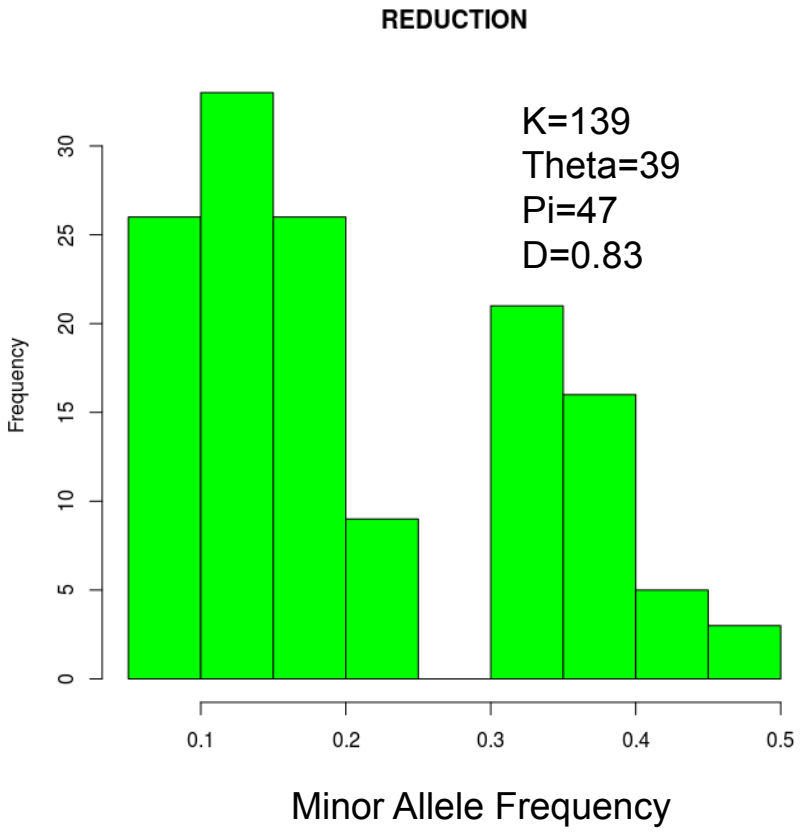


Demography matters!

n=20; L=500kbp; no selection



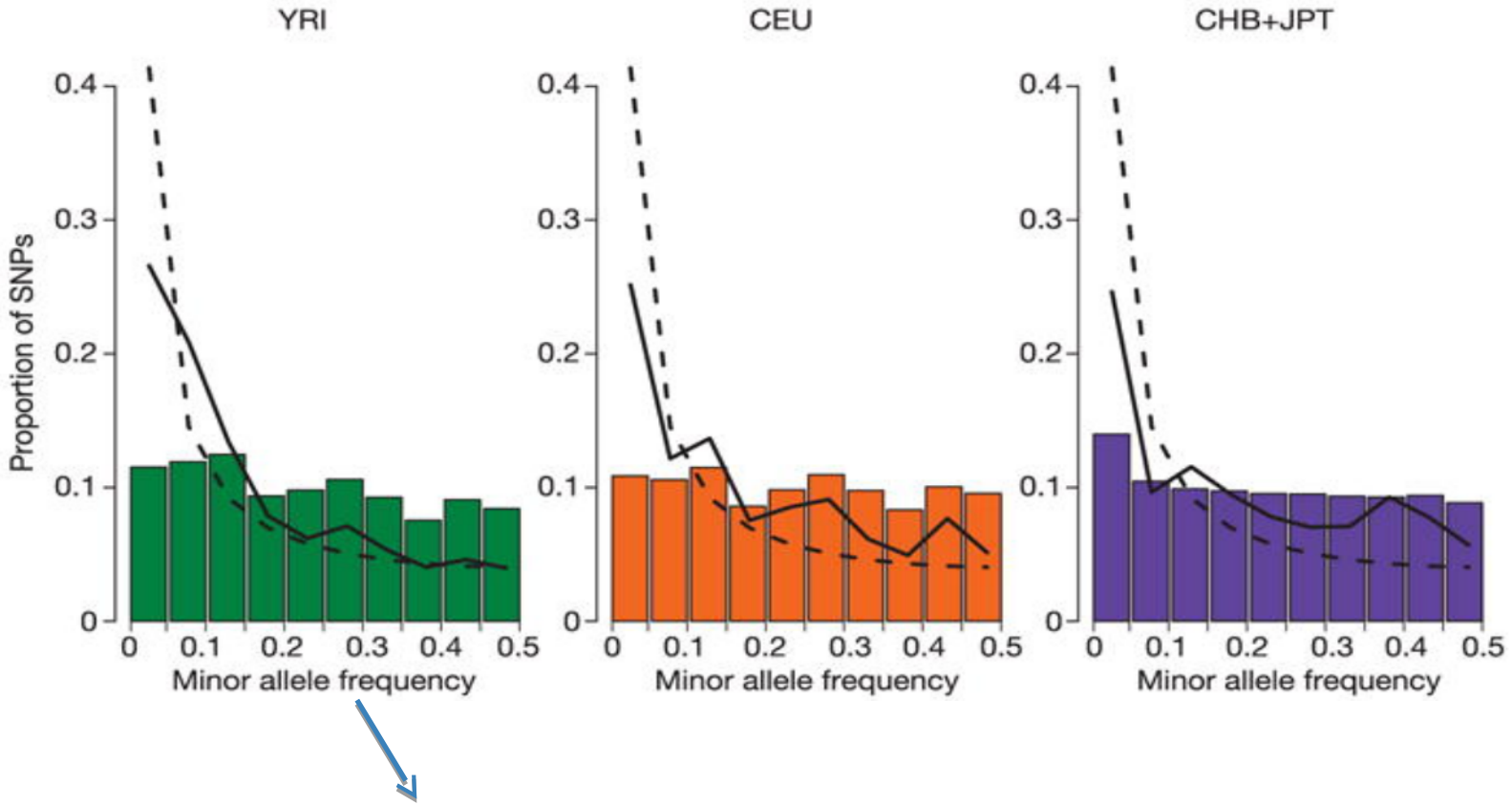
n=20; L=500kbp; no selection



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

Experimental design matters?

The effect of ascertainment bias

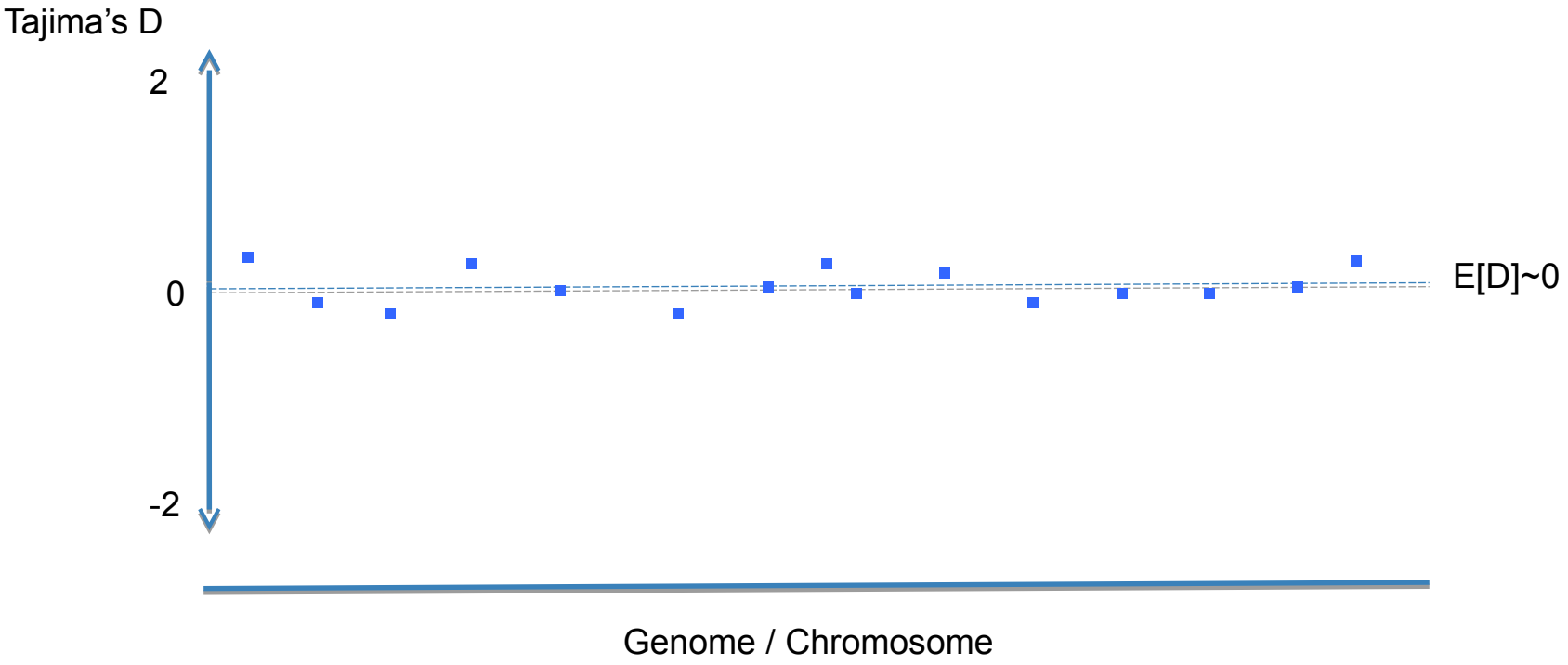


Deficiency of low-frequency variants

HapMap Consortium. Nature 2005

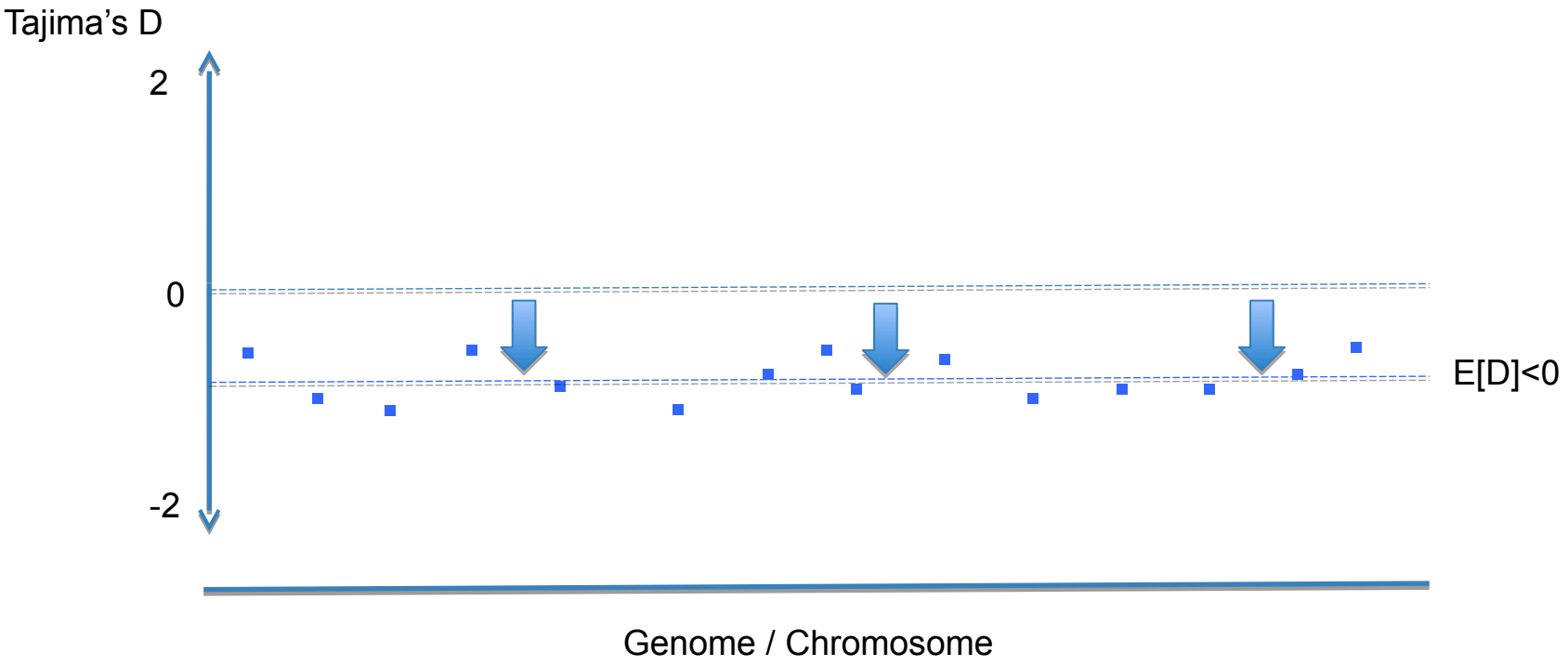
How to take neutral confounding factors into account?

Under constant population size:



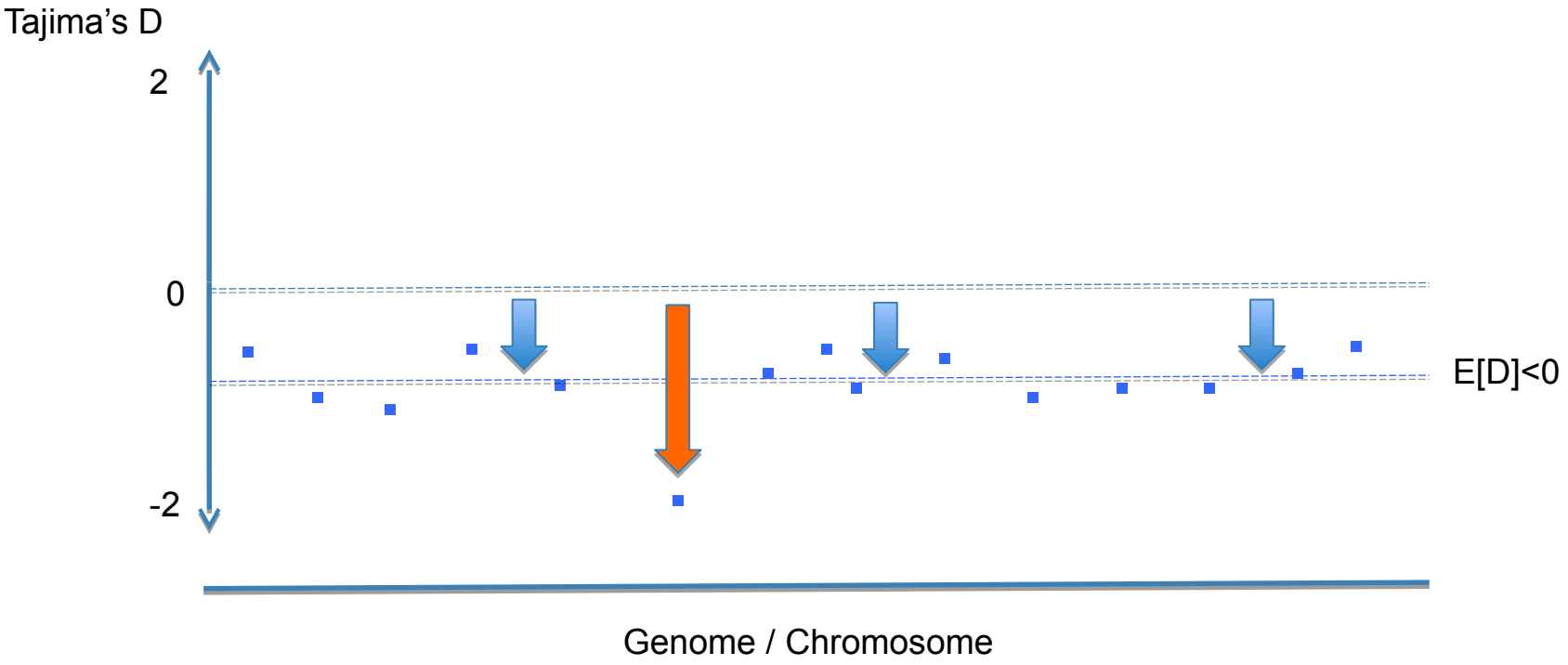
How to take neutral confounding factors into account?

Under expanding population size:



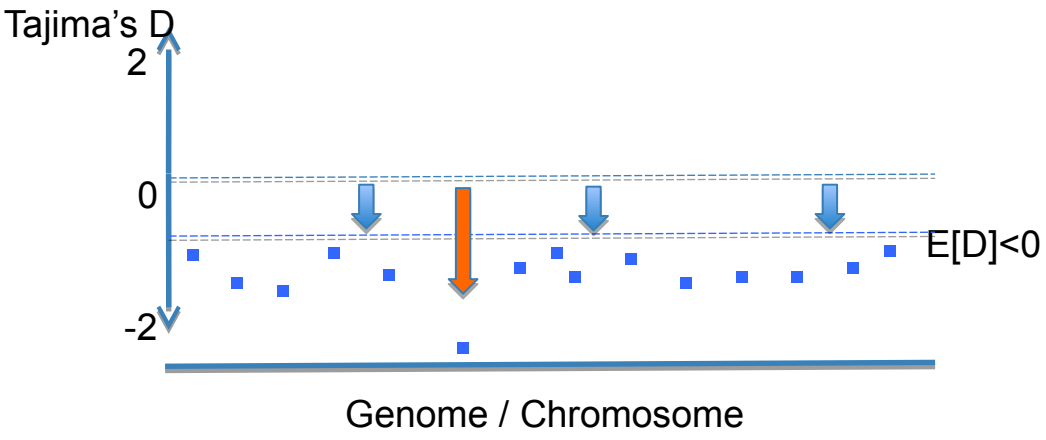
How to take neutral confounding factors into account?

Under expanding population size and positive selection:

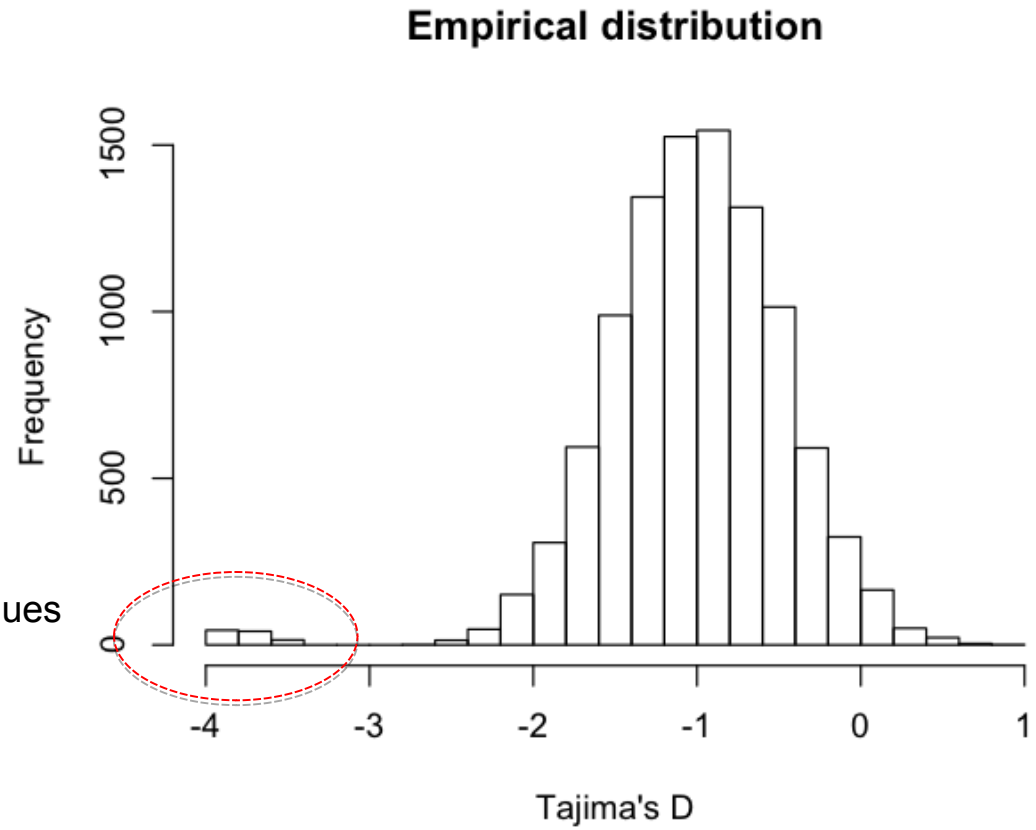


- Demography affects all loci equally, while selection changes local patterns

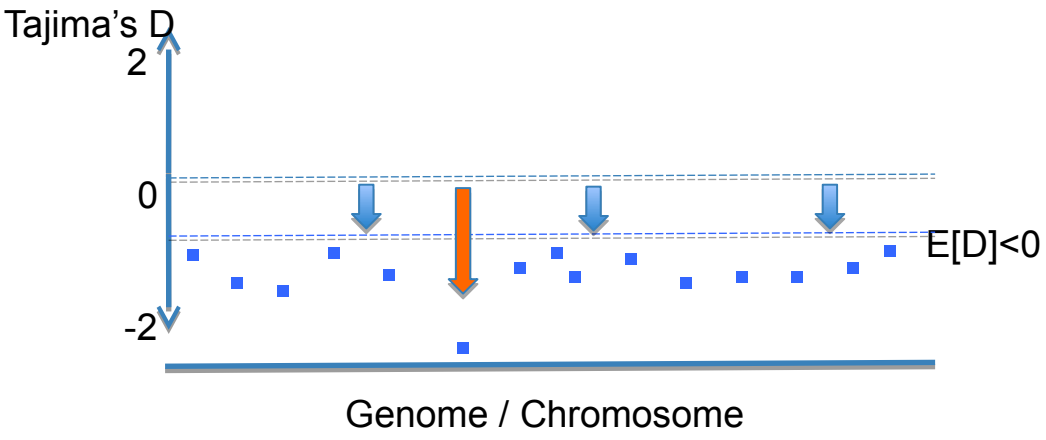
Outlier approach



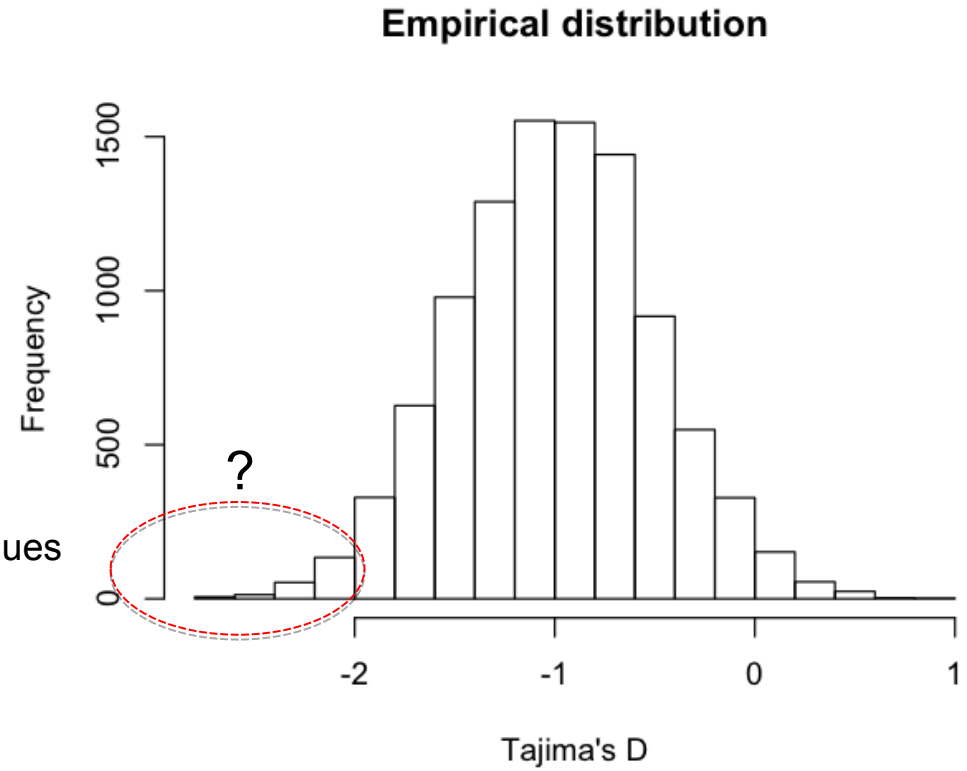
Assign empirical *p*-values
(ranked percentiles)



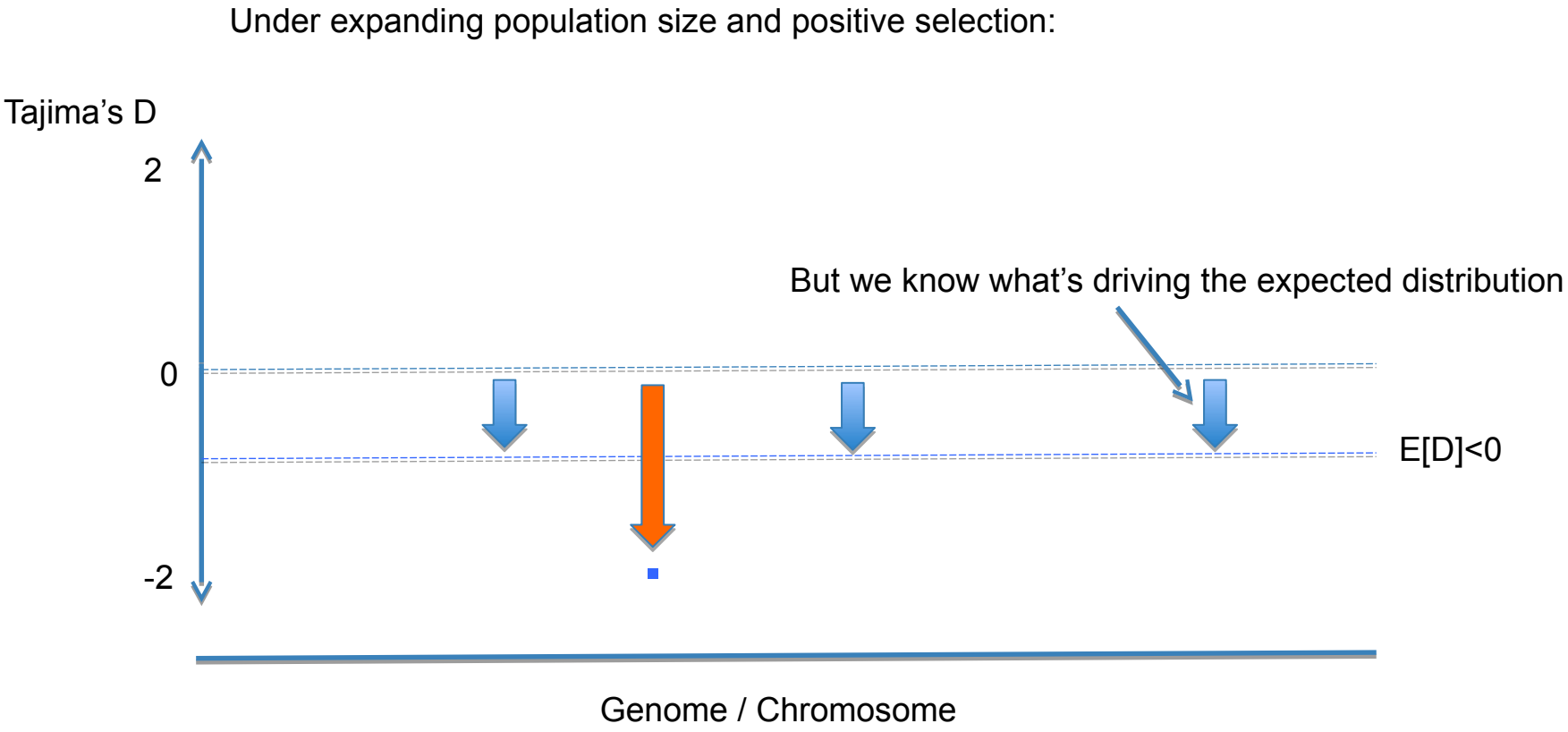
Outlier approach



Assign empirical *p*-values
(ranked percentiles)

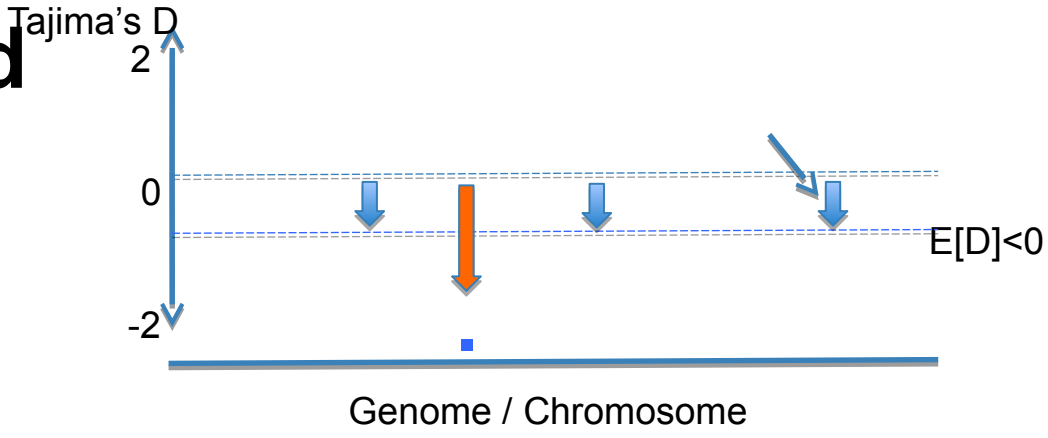


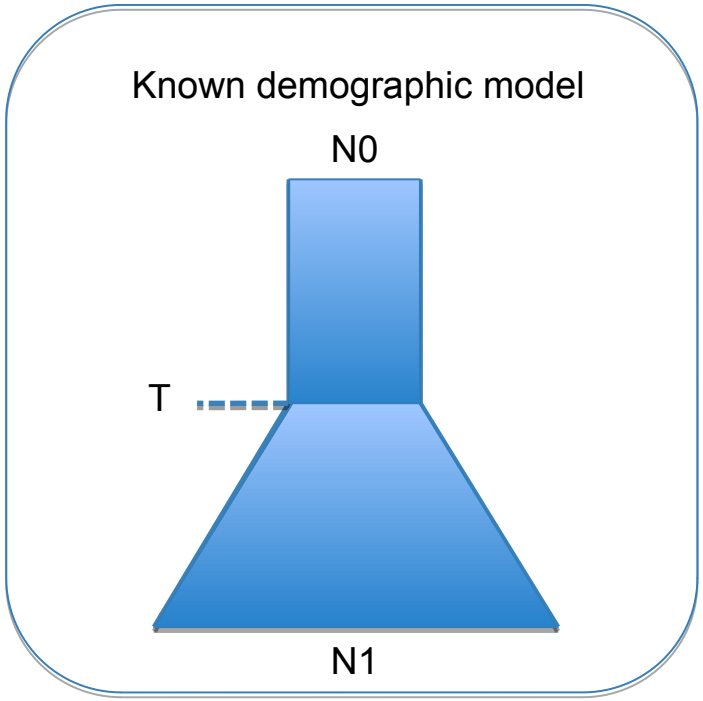
How to take neutral confounding factors into account?



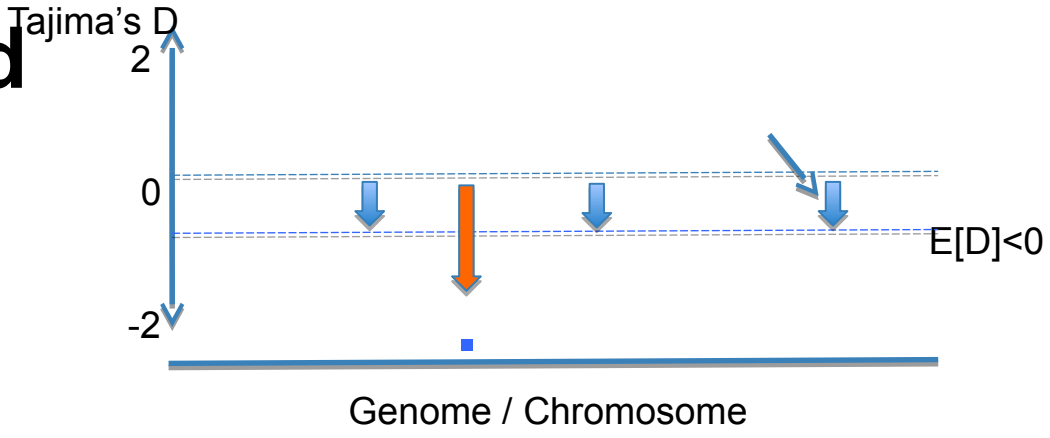
- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

Simulations-based approach

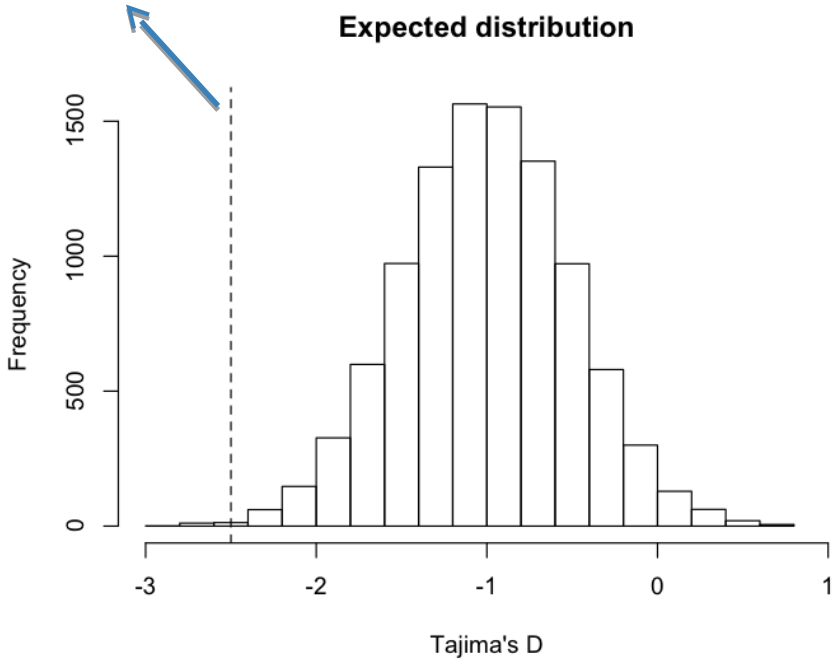


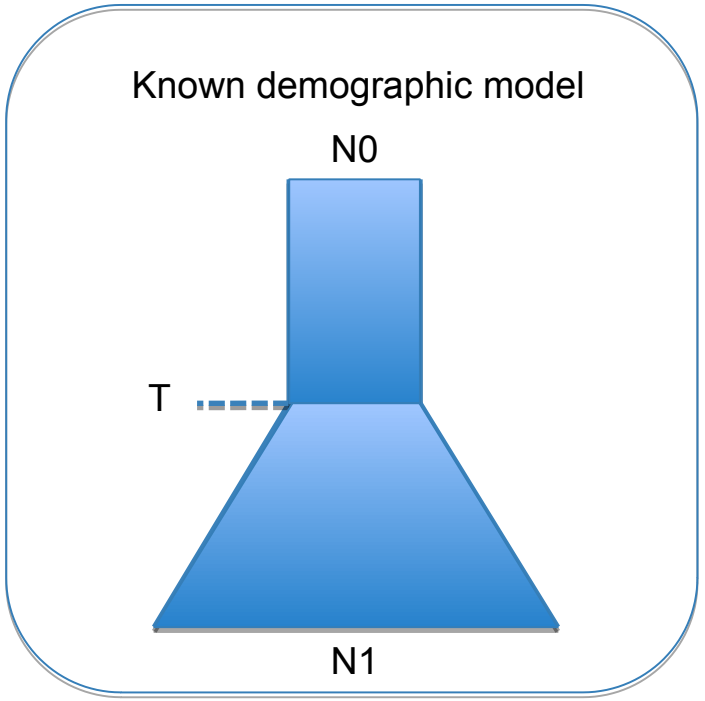


Simulations-based approach



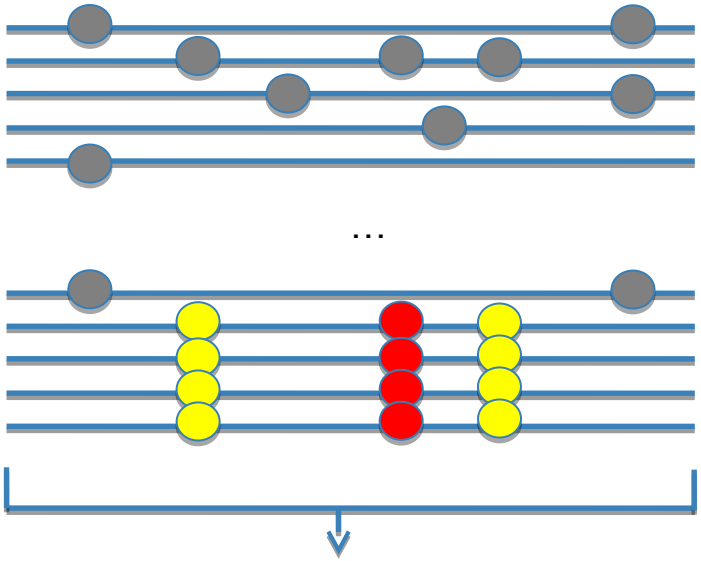
Assign *p*-values
(based on ranked percentile of observed value)





Let's assess statistical significance for our test of selection on EDAR.

Positive selection

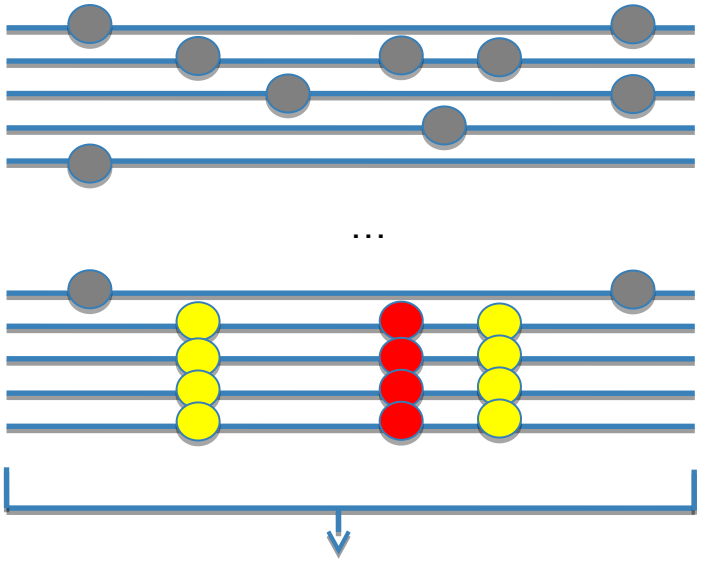


$t<T_{sel}$

$t>>T_{sel}$

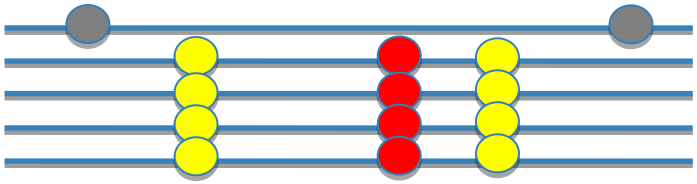
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- ?

Positive selection



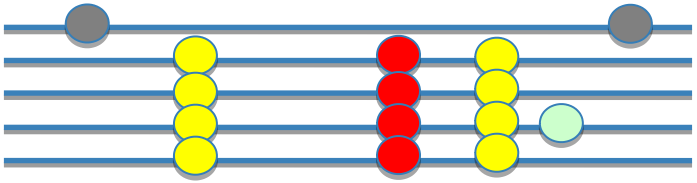
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- Extended haplotype homozygosity / Extended LD

Extended Haplotype Homozygosity



$t >> T_{sel}$

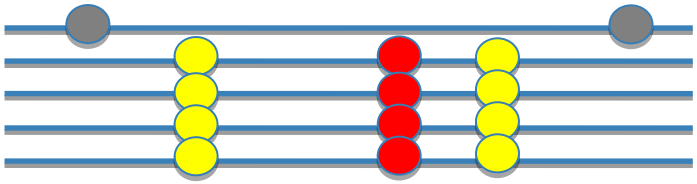
…



$t >>> T_{sel}$

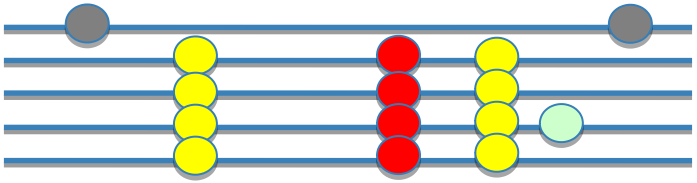
…

Extended Haplotype Homozygosity



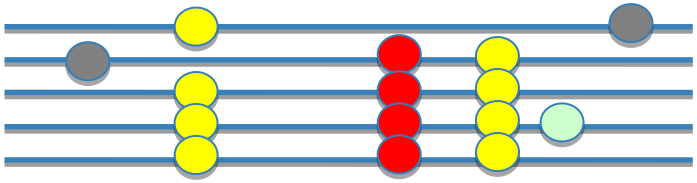
$t>>T_{sel}$

...



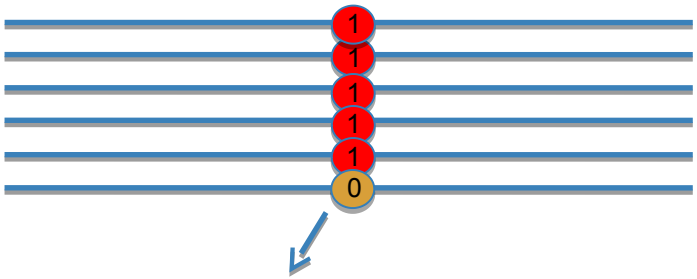
$t>>>T_{sel}$

...



$t>>>T_{sel}$

Extended Haplotype Homozygosity

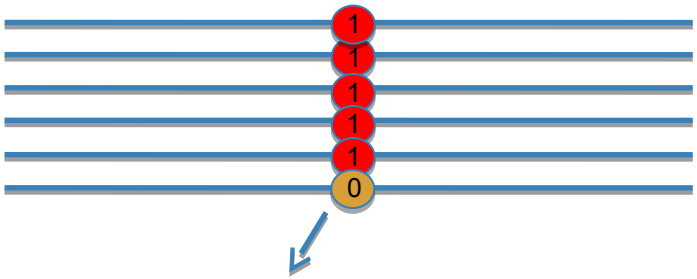


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Core SNP

Extended Haplotype Homozygosity

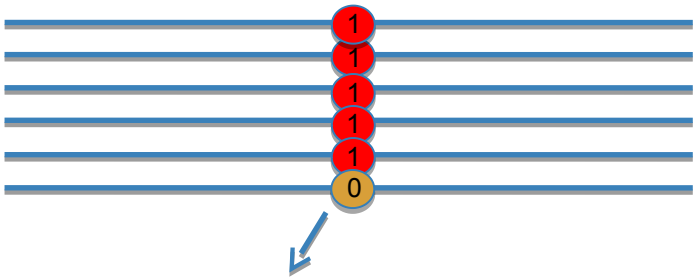


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

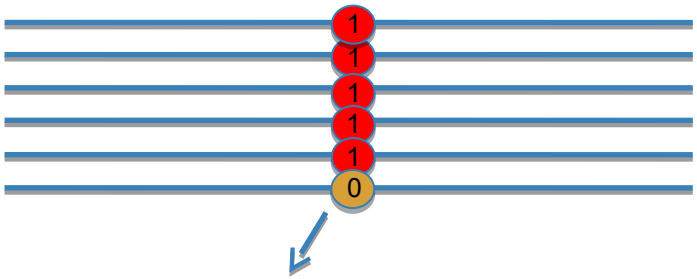


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Sum across all unique haplotypes carrying the core SNP

Extended Haplotype Homozygosity



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

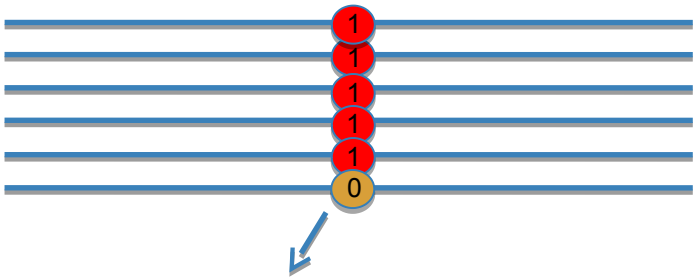
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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)

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

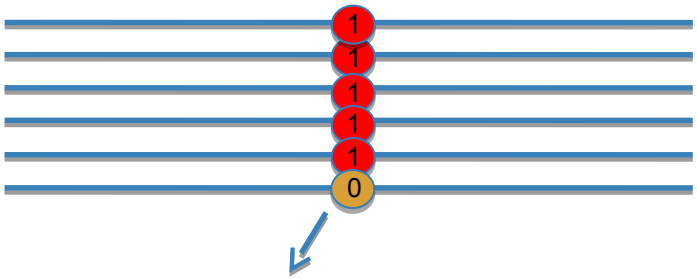
n_h is haplotype frequency of h

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)

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

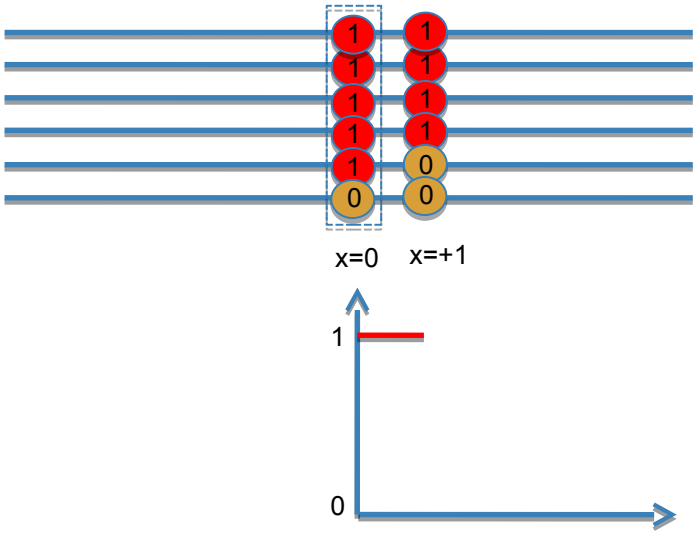
n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

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

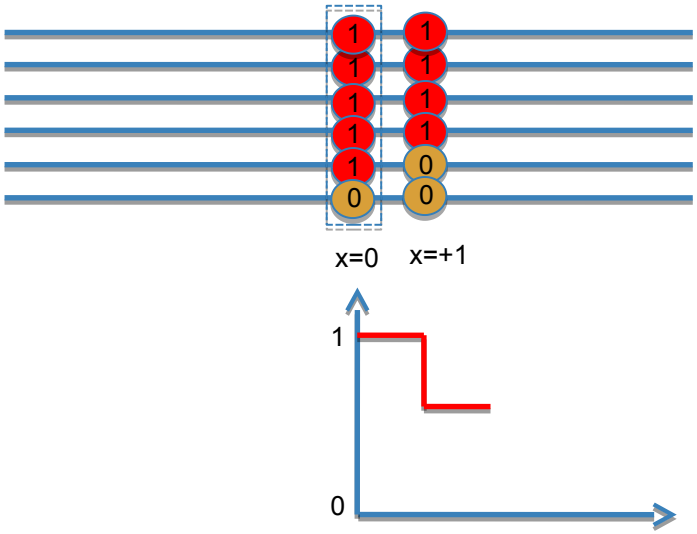
Extended Haplotype Homozygosity



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

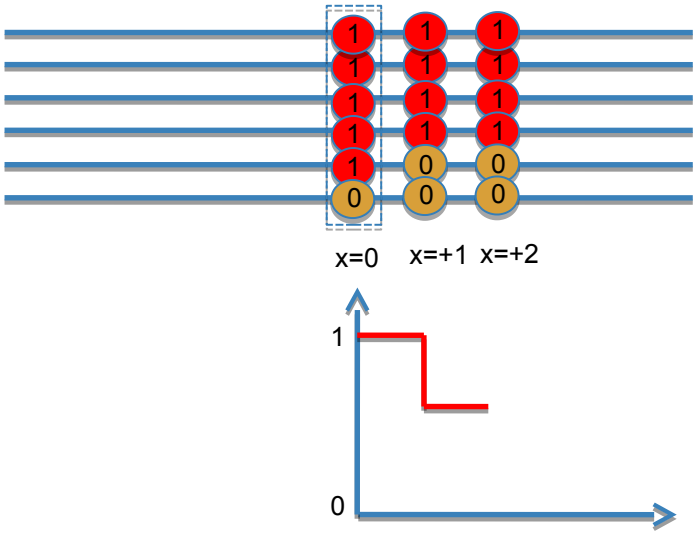
How many unique haplotypes carrying the core SNP?
What is their frequency?

Extended Haplotype Homozygosity



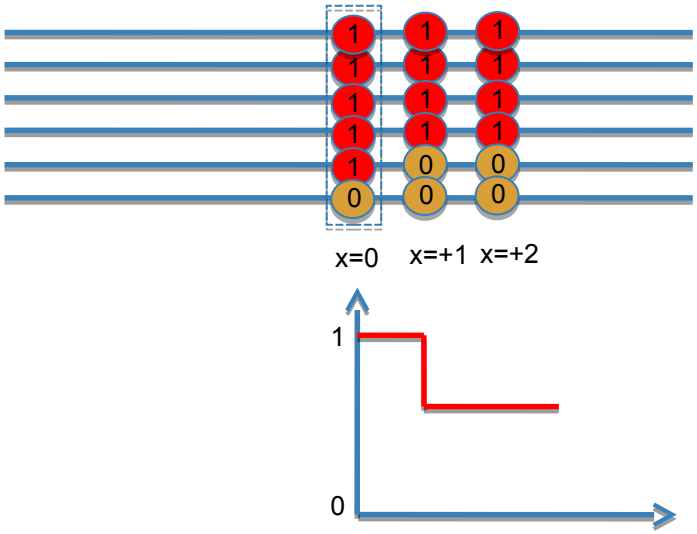
$$EHH_c(x_i = +1) = \frac{\binom{4}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{6+0}{10} = 0.60$$

Extended Haplotype Homozygosity



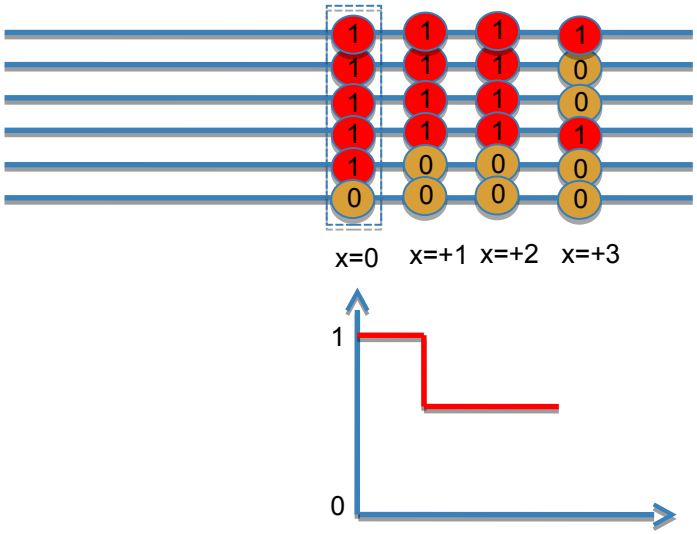
$$EHH_c(x_i = +2) = ?$$

Extended Haplotype Homozygosity



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

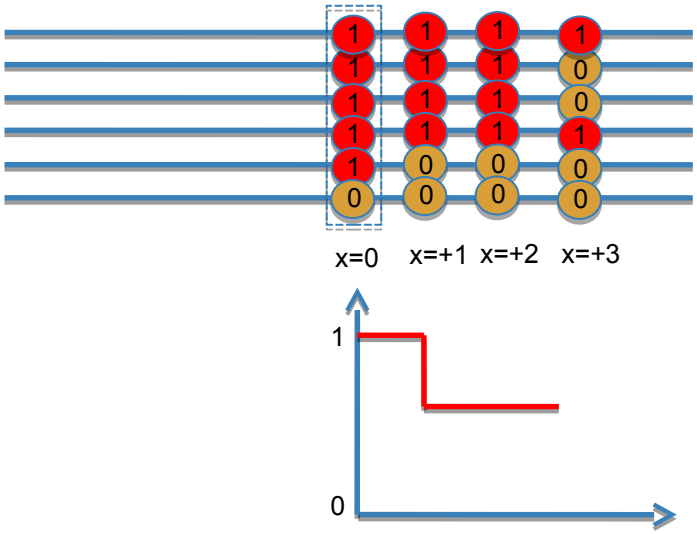
Extended Haplotype Homozygosity



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

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

Extended Haplotype Homozygosity

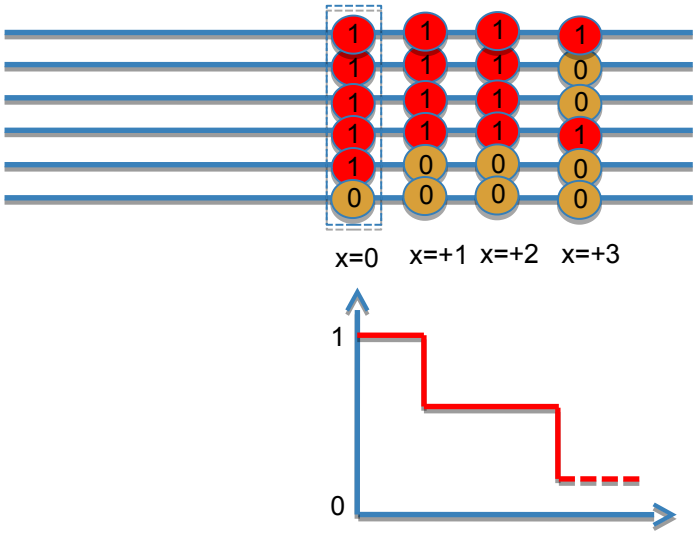


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

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

Extended Haplotype Homozygosity

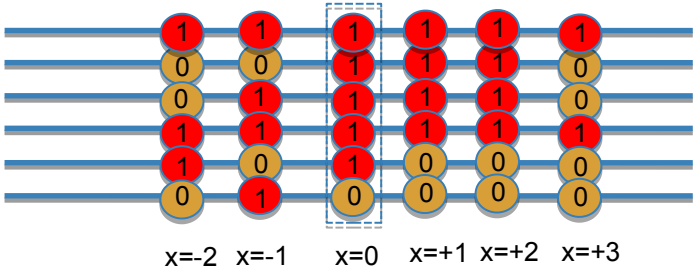


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

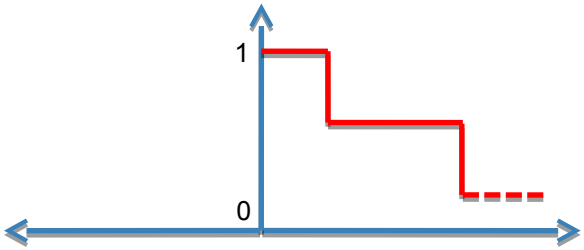
$$EHH_c(x_i = +3) = \frac{\binom{2}{2} + \binom{2}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+1+0}{10} = 0.20$$

Extended Haplotype Homozygosity



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

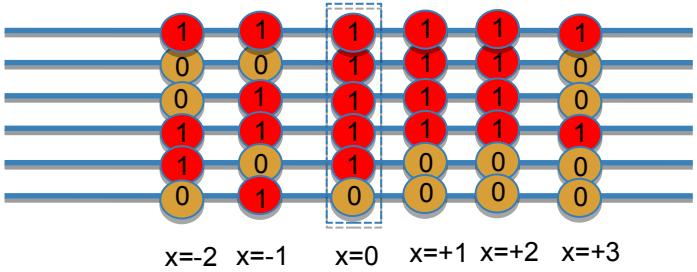
n	*n choose 2*
1	0
2	1
3	3
4	6
5	10
6	15



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

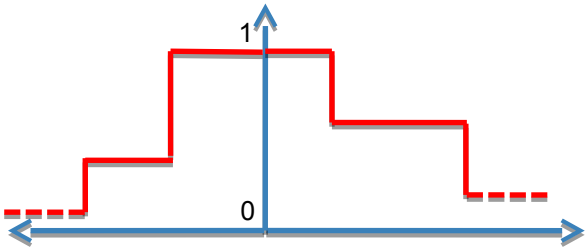
Comment on differences (if any) between *EHH(x=+2)* and *EHH(x=-2)*.

Extended Haplotype Homozygosity



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

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



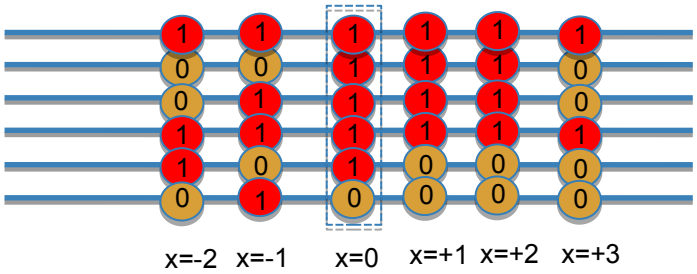
$$EHH_c(x_i = -1) = \frac{\binom{3}{2} + \binom{2}{2}}{\binom{5}{2}} = \frac{3+1}{10} = 0.4$$

+ (1 choose 2)

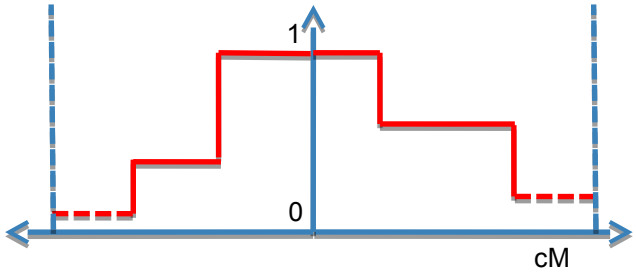
$$EHH_c(x_i = -2) = \frac{\binom{2}{2} + \binom{1}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+0+0}{10} = 0.1$$

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

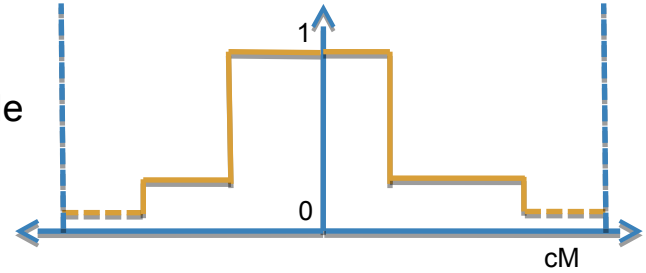
Integrated Haplotype Score



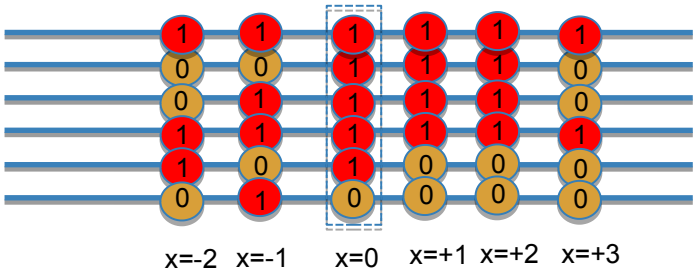
For the derived allele

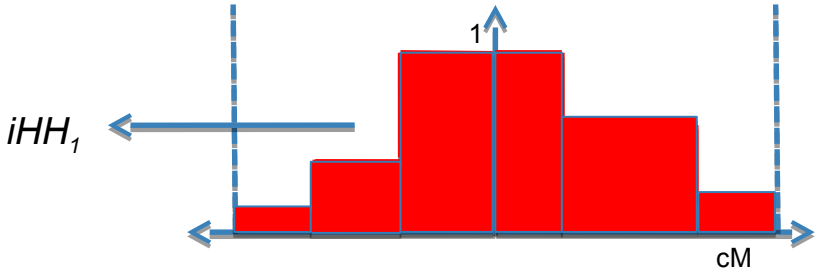


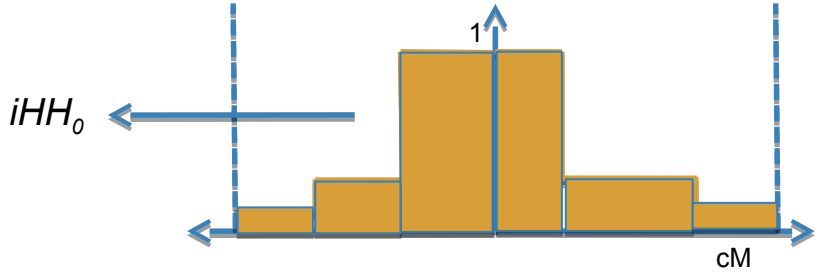
For the ancestral allele



Integrated Haplotype Score

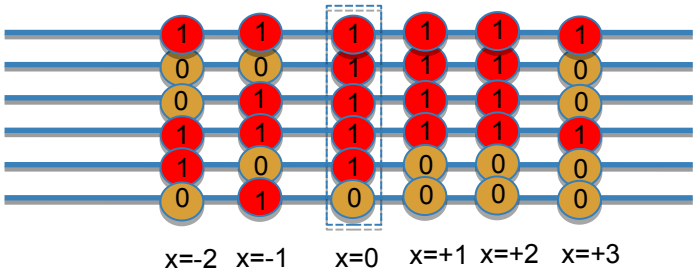


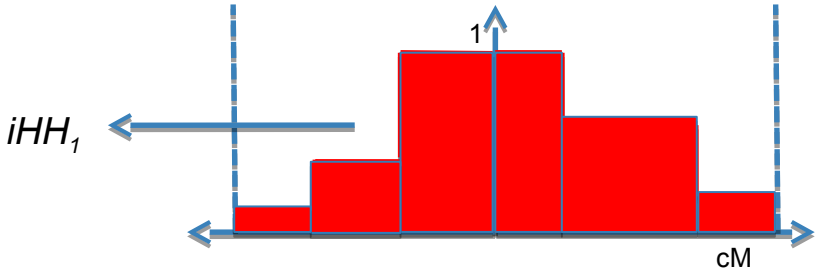




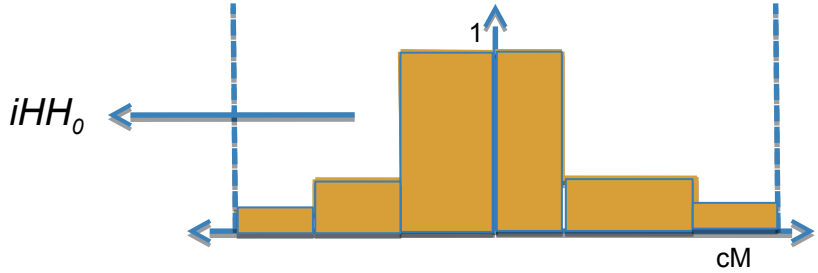
Integrated haplotype homozygosity (*iHH*)

Integrated Haplotype Score





Integrated haplotype homozygosity (*iHH*)

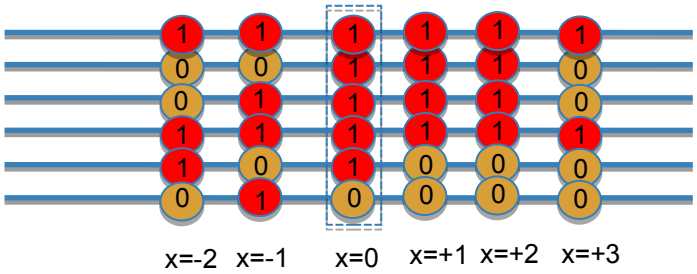


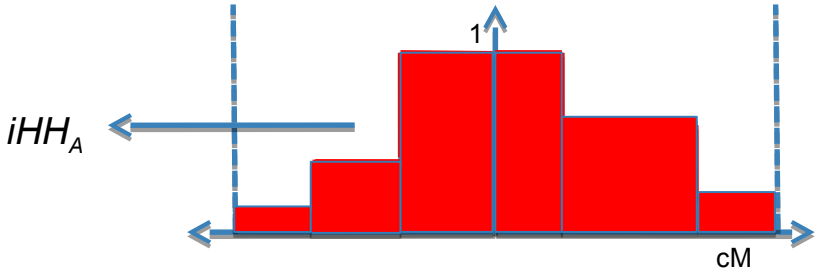
Integrated haplotype score:

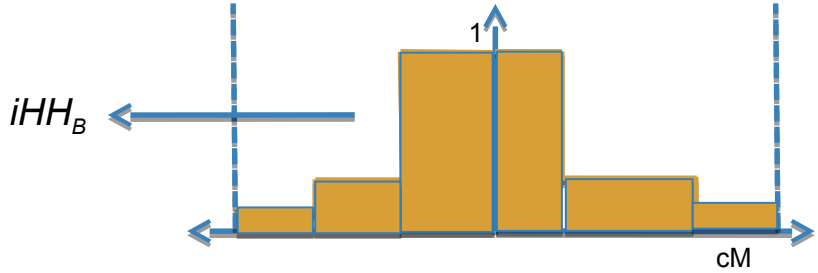
$iHs = ln(iHH_1/iHH_0)$

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

Cross-population Extended Haplotype Homozygosity





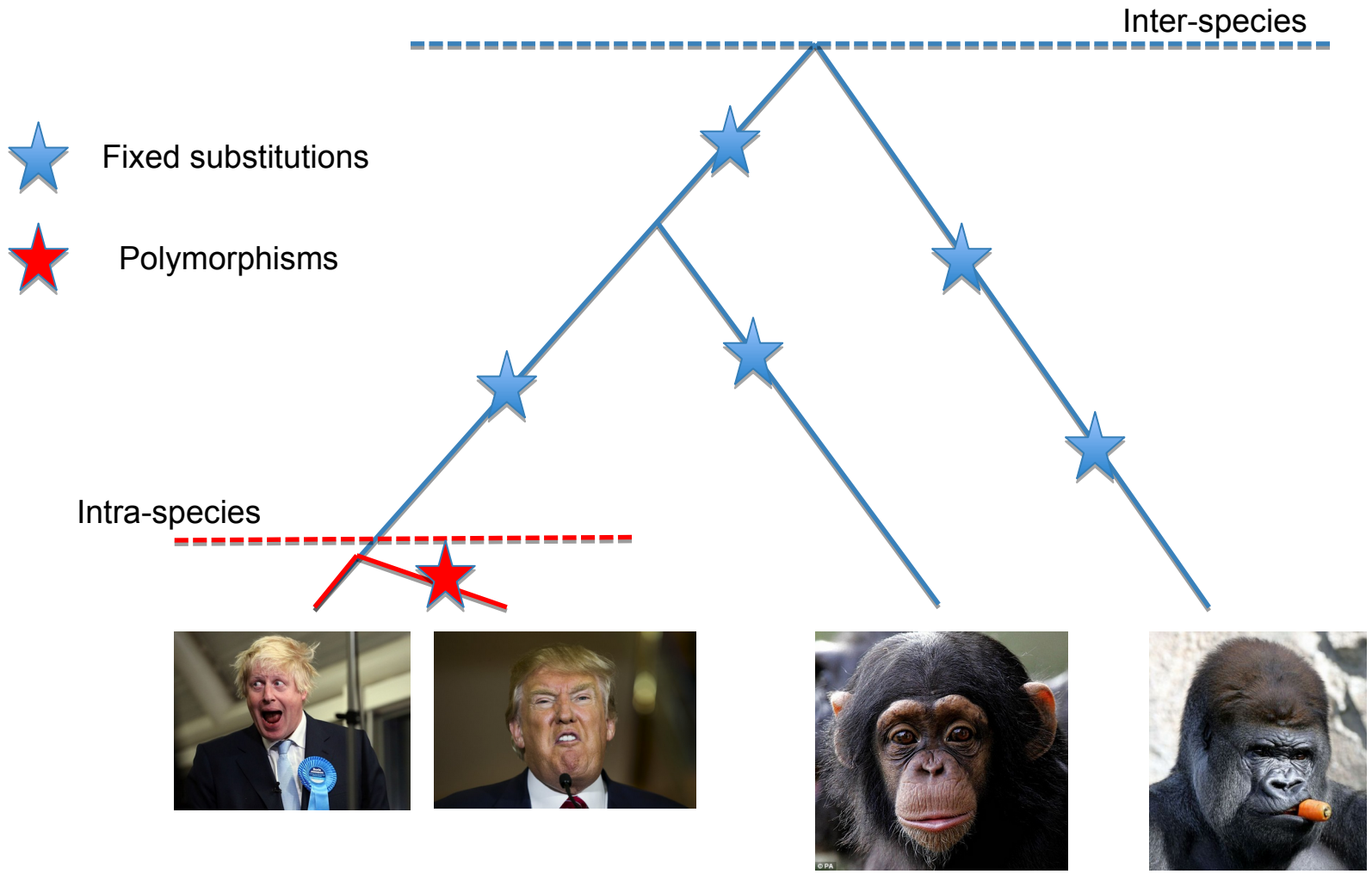


Integrated haplotype homozygosity (*iHH*) for **populations** A and B

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

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

Inferring inter-species selection



State-of-the-art methods to detect natural selection

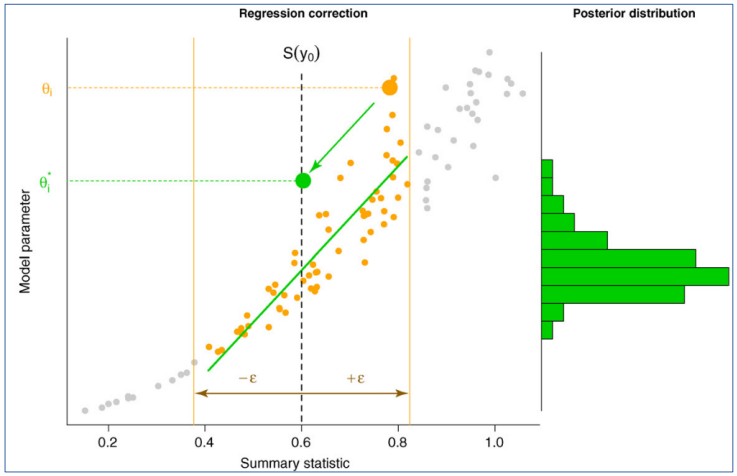
1. Composite scores (Grossman et al. 2013)

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

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

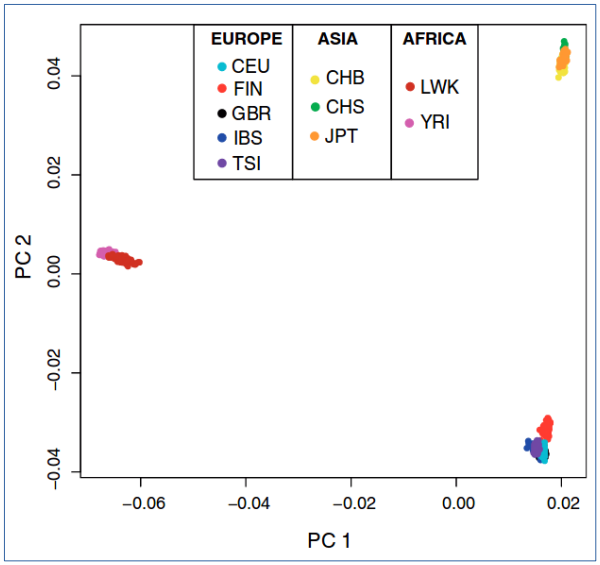
$$CMS_{GW} = \prod_{t \in tests} BF_t$$

2. Simulations-based (rejection, ABC)



3. Unsupervised machine learning

(PCA, Duforet-Frebourg et al. 2016)



4. Supervised machine learning

(SVM, Schrider & Kern 2018)