Inference of natural selection from NGS data using ABC

Matteo Fumagalli

September, 14th 2017

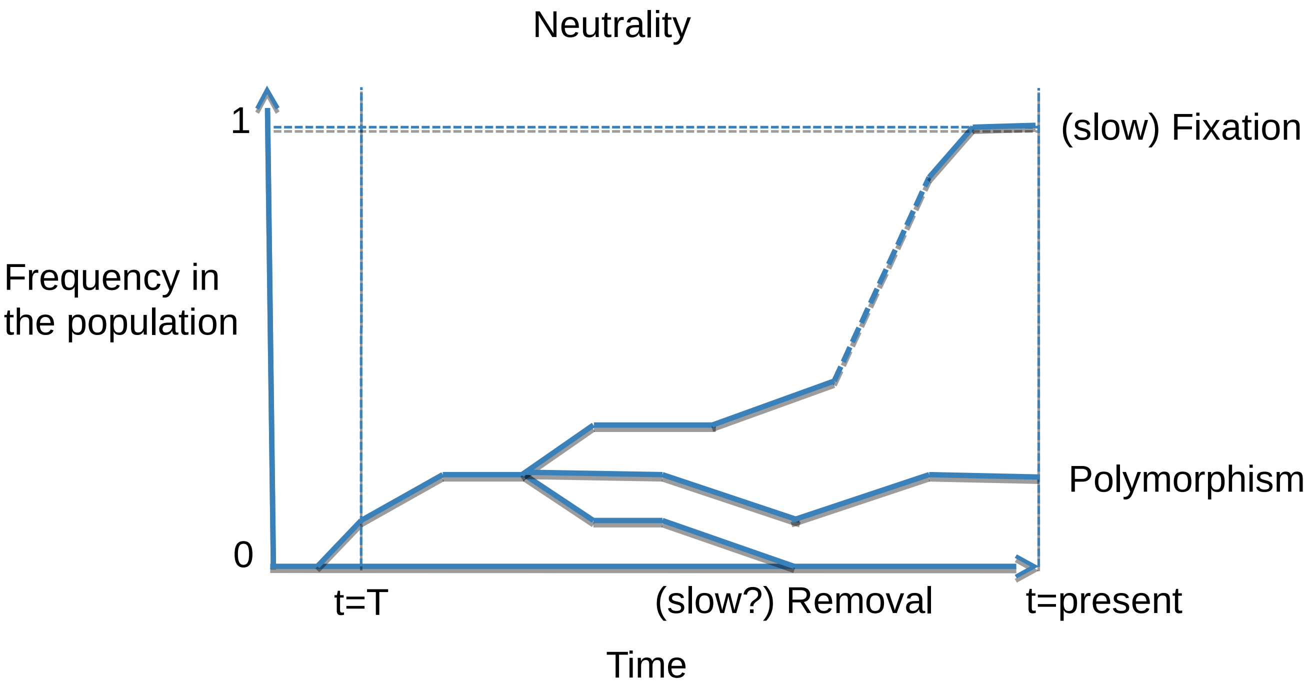
Outline

- Brief introduction to natural selection
- Inferring selection at the intra-species level using summary statistics
- PRACTICAL: detecting selection from low-depth NGS data
- The effect of demography on selection scans
- PRACTICAL: quantifying selection using ABC
- (Experimental design)

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

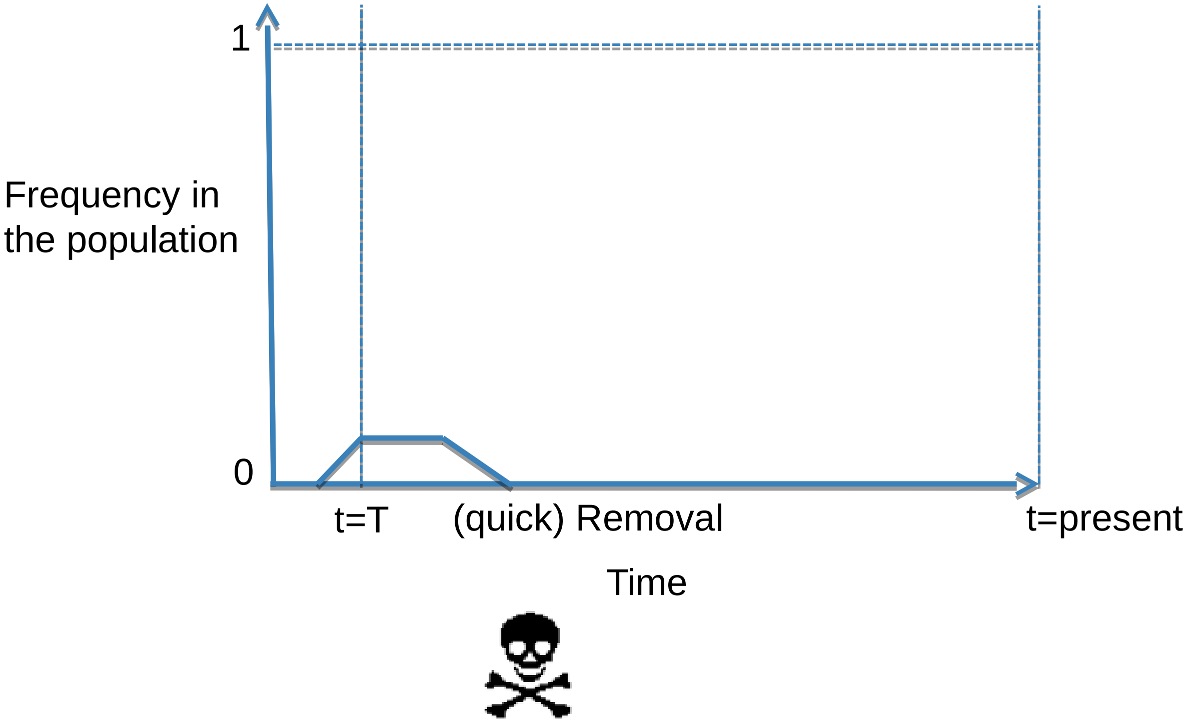


Natural selection

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

Negative selection

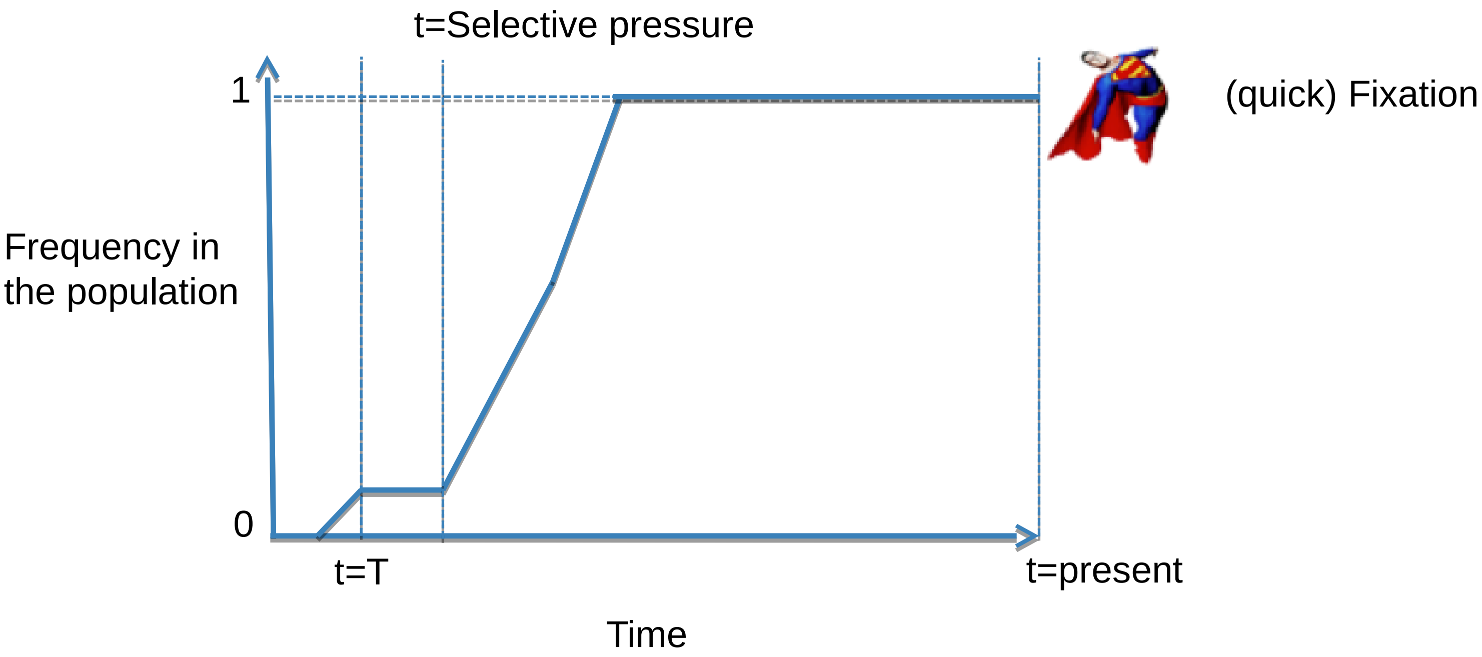


Natural selection

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

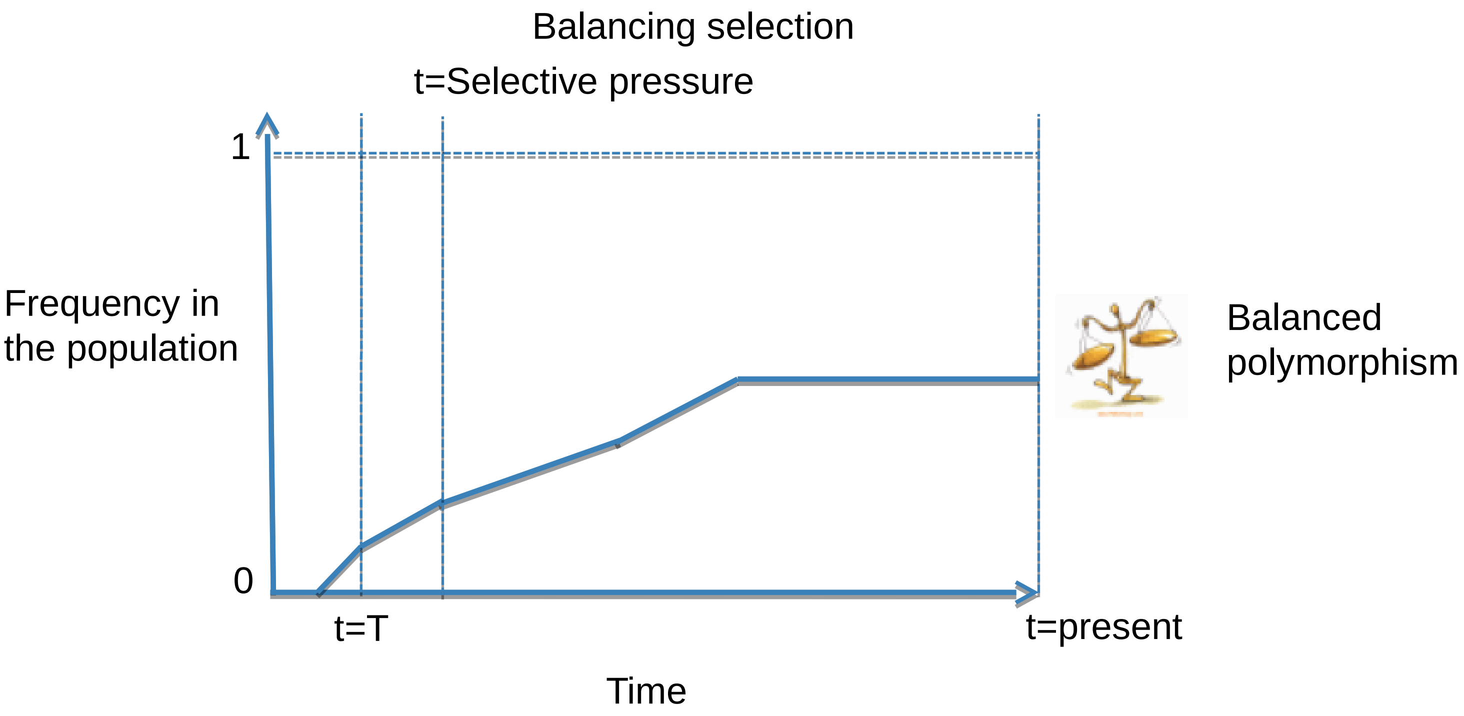
Positive selection



Natural selection

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

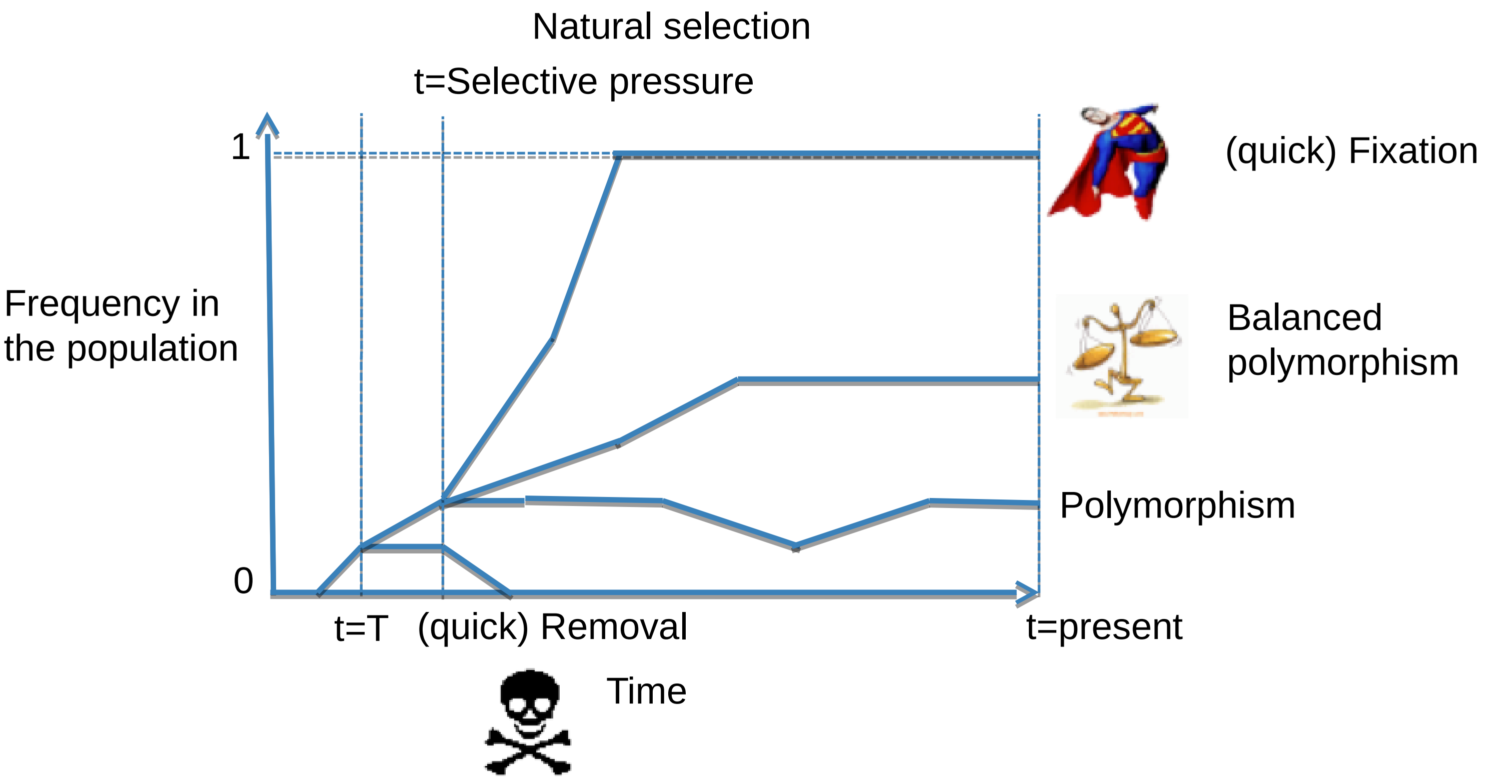
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



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



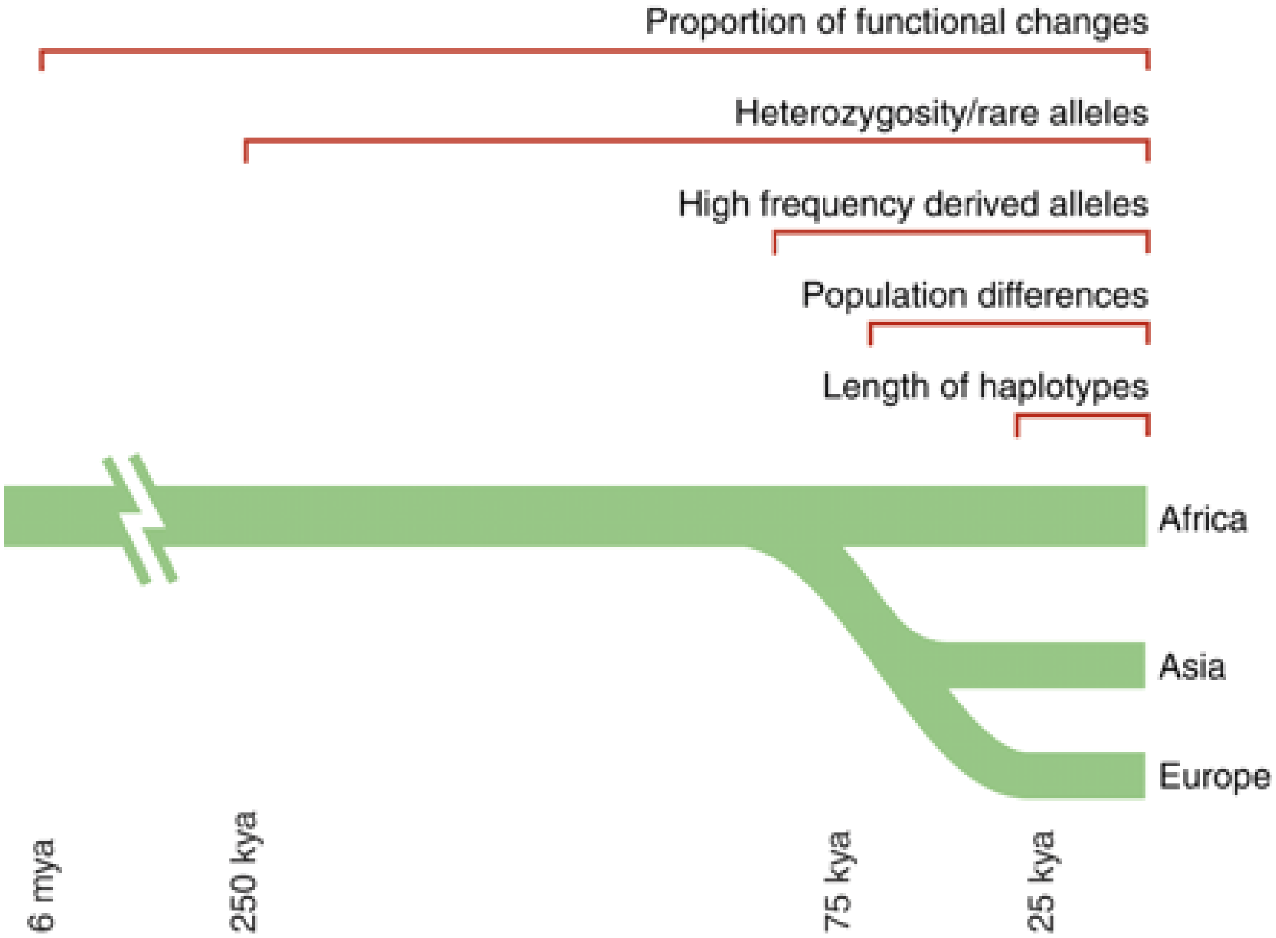
2) Sites targeted by natural selection are likely to harbour **functionality**

Methods to infer selection

- within-species:

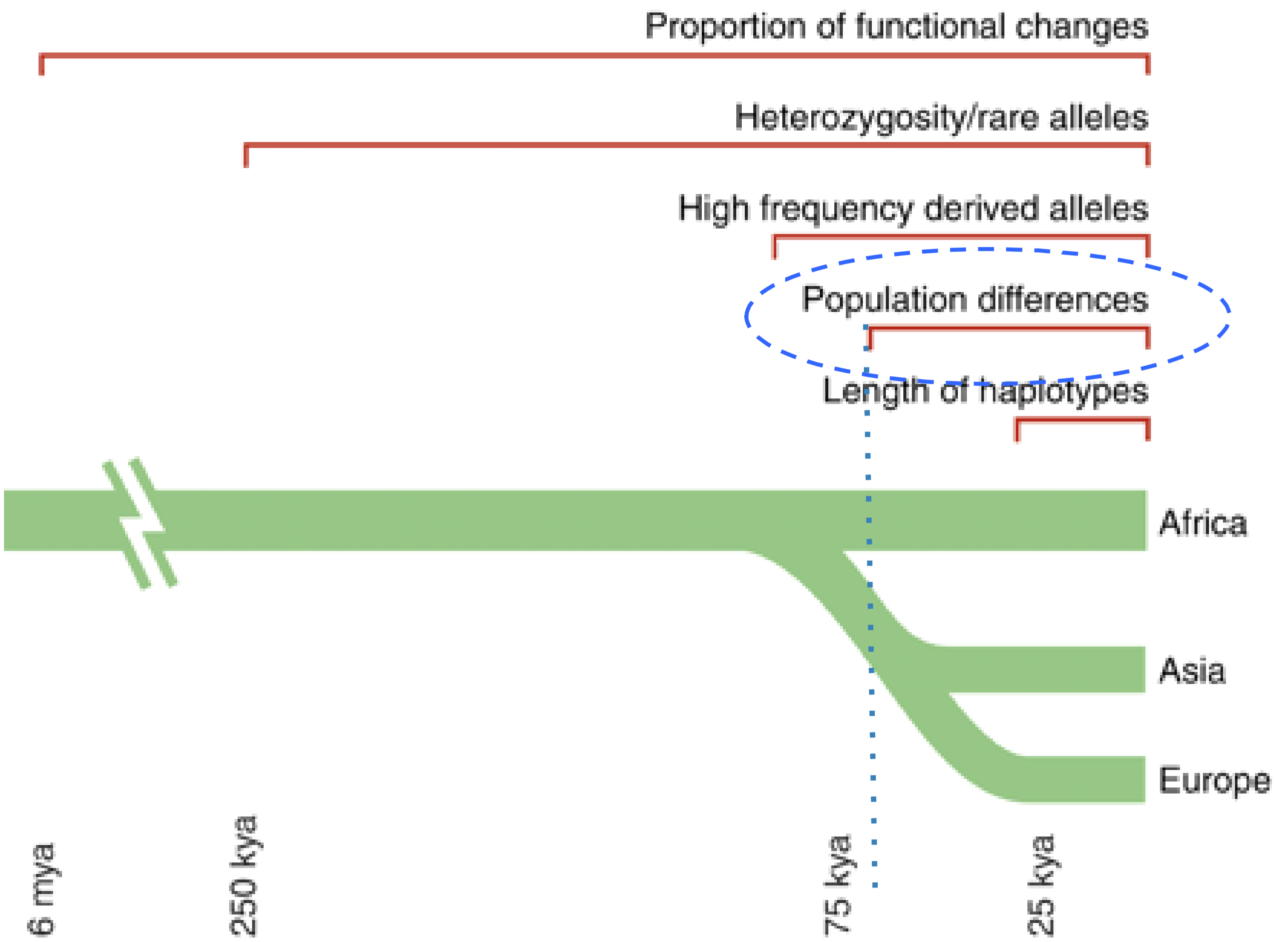
  Micro-evolutionary events between populations, local adaptation

Methods to infer recent selection


Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

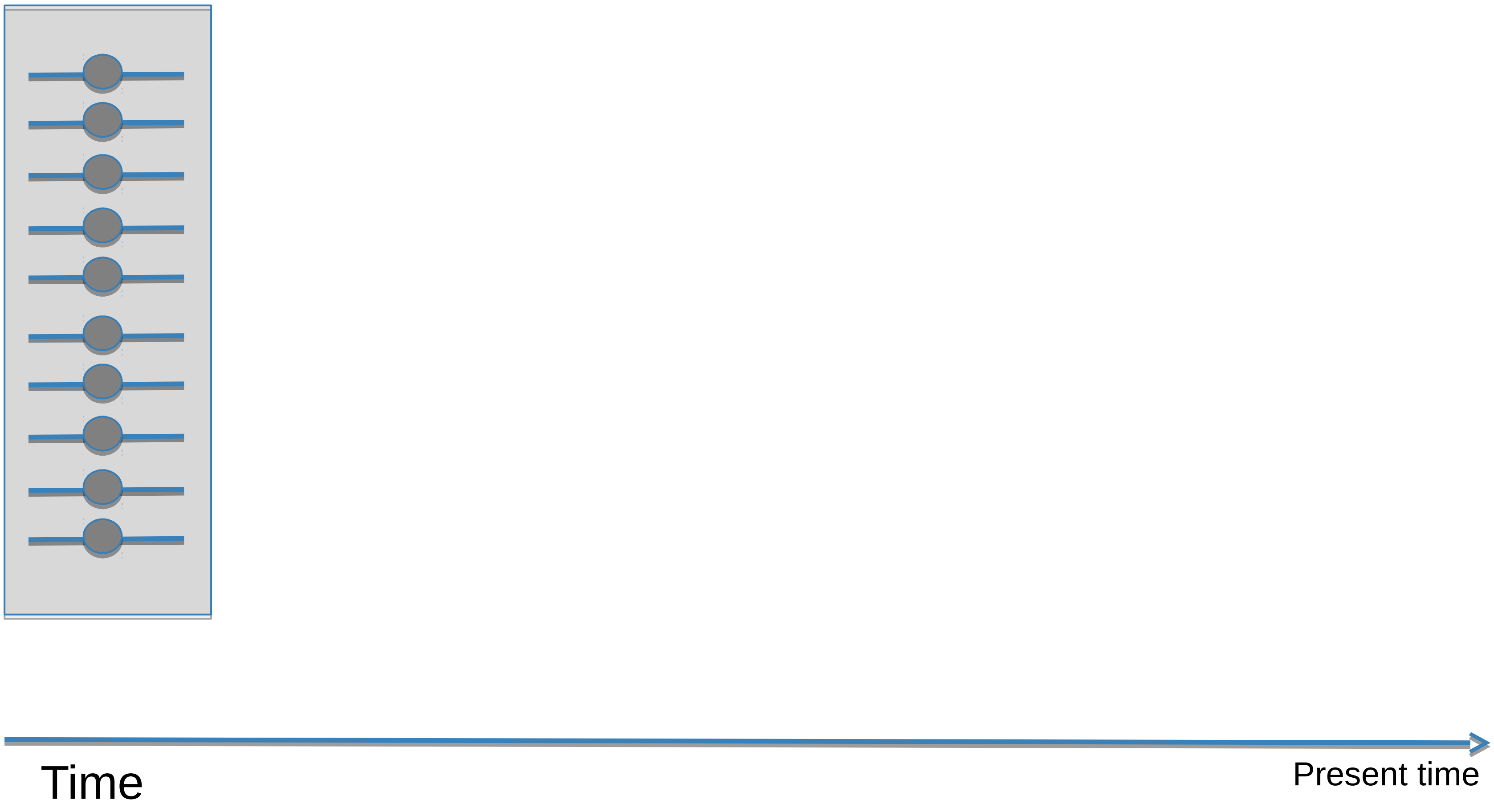
Sabeti et al. 2006 Science

Methods to infer recent selection

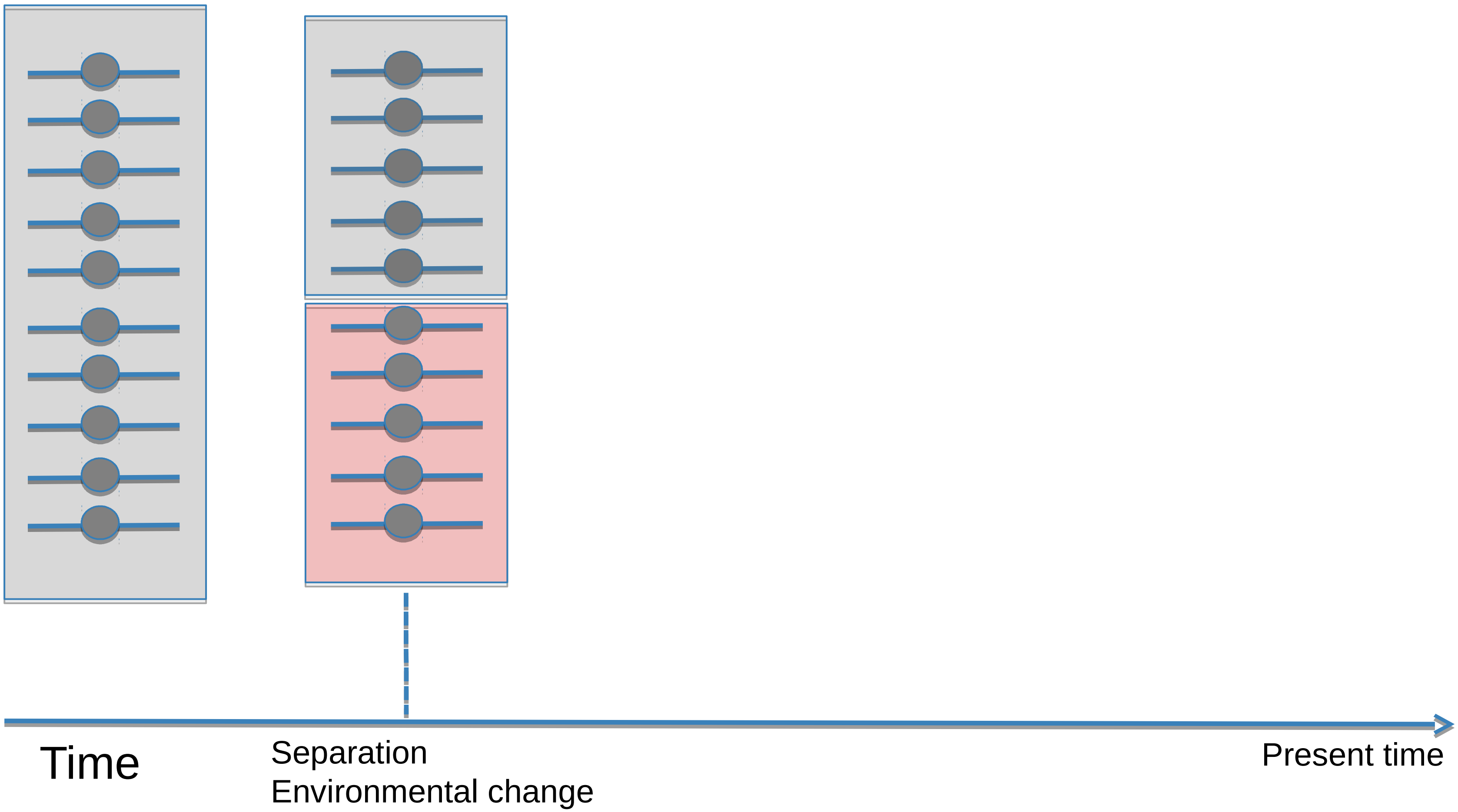

Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Sabeti et al. 2006 Science

Allele frequency differentiation



Allele frequency differentiation

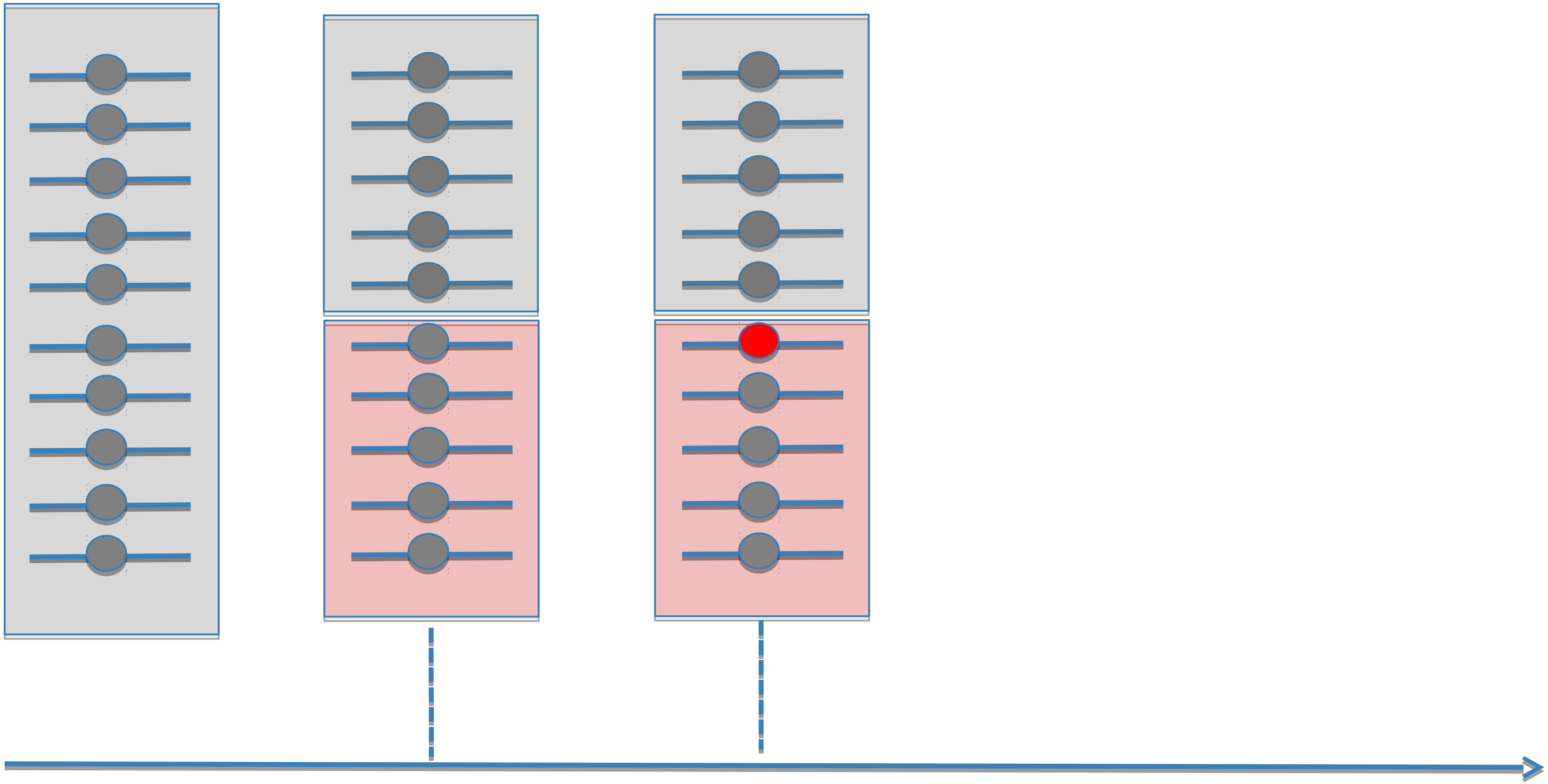


Time

Separation
Environmental change

Present time

Allele frequency differentiation



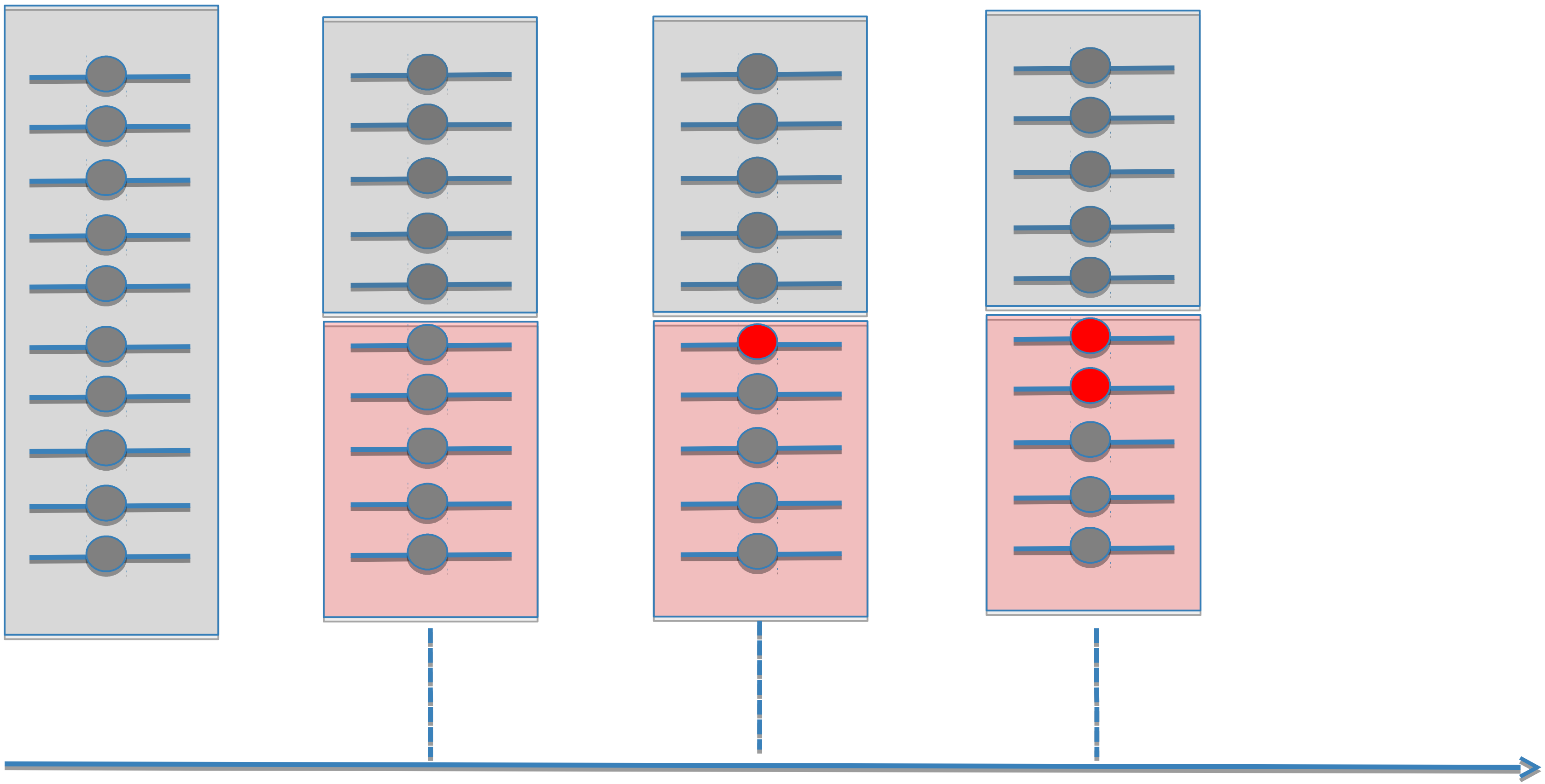
Time

Separation
Environmental change

Onset of beneficial
mutation

Present time

Allele frequency differentiation



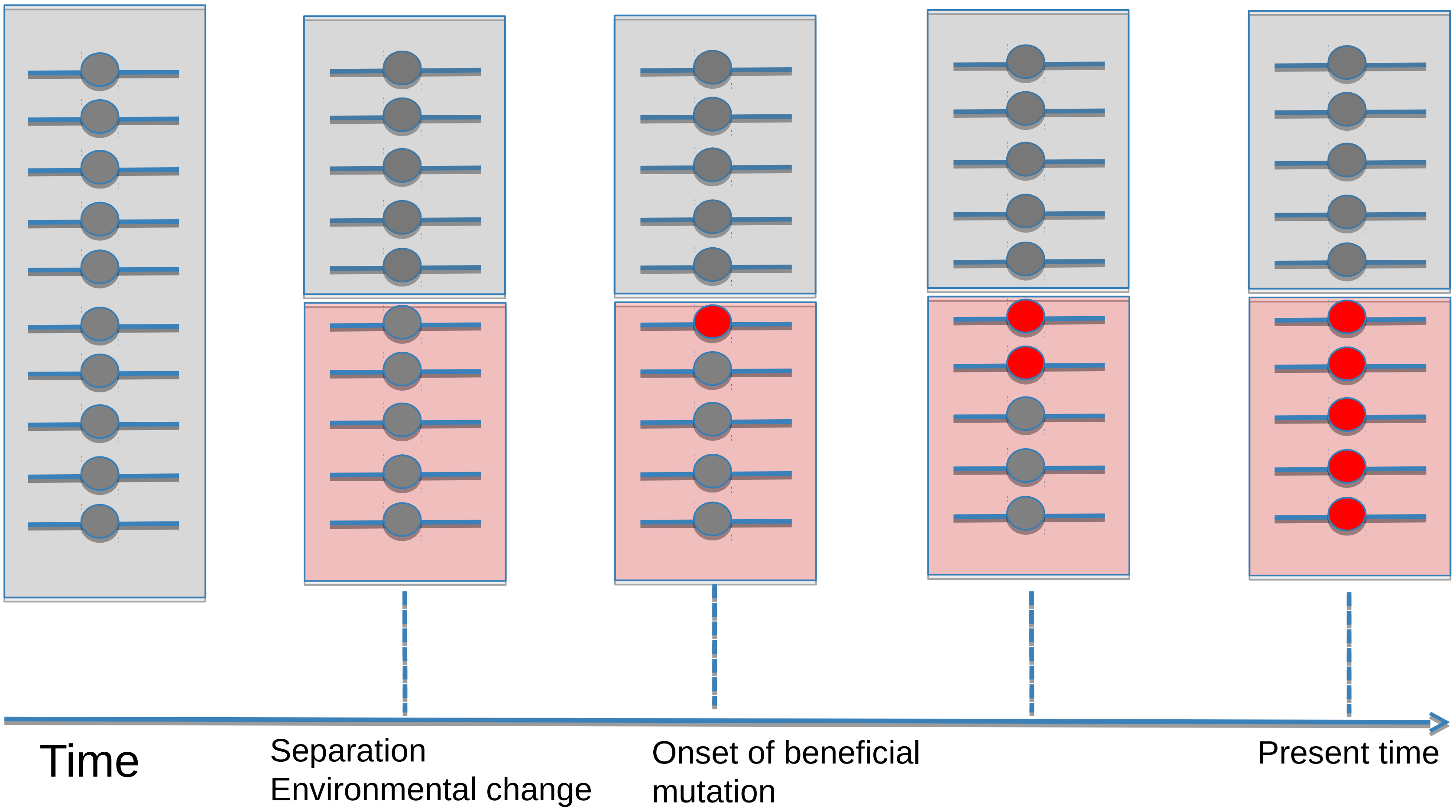
Time

Separation
Environmental change

Onset of beneficial
mutation

Present time

Allele frequency differentiation



F_{ST}

Common measure for quantifying population subdivision.

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

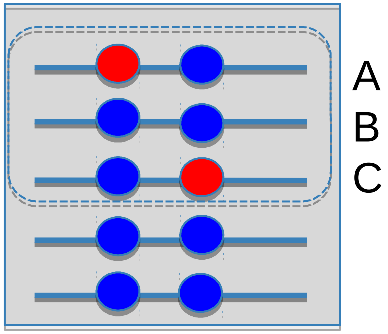
H_B: between populations

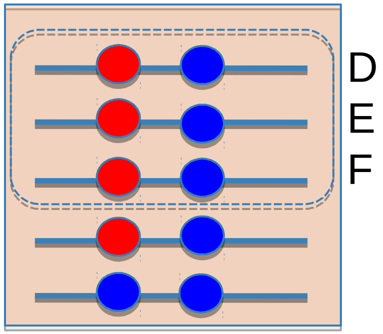
H_W: average within populations

- if $H_B >> H_W$ 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





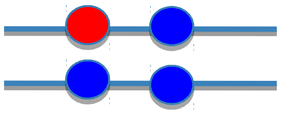
$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations

What is the variation within populations?

e.g. A vs B

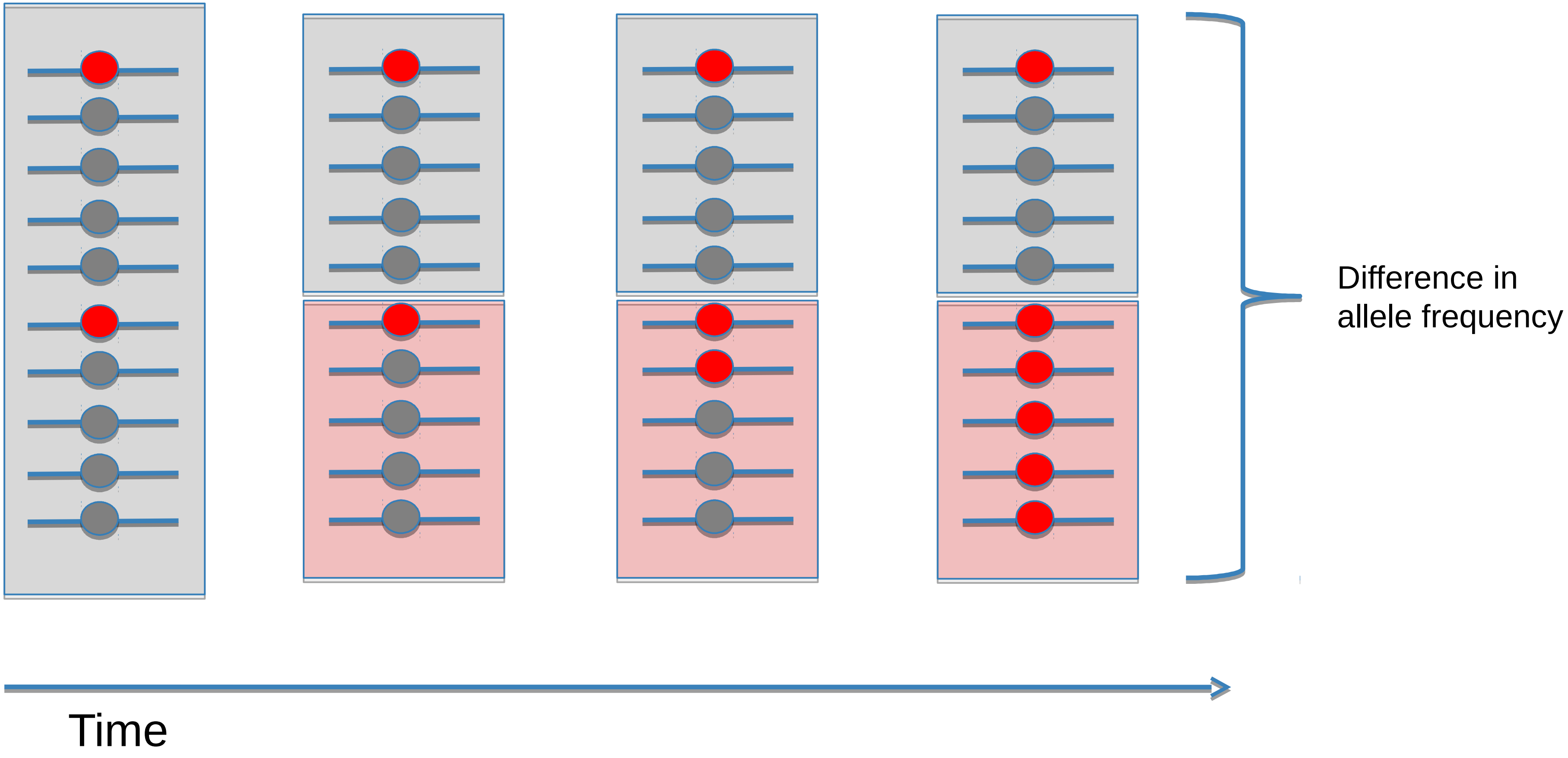


The differ by 1 site

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

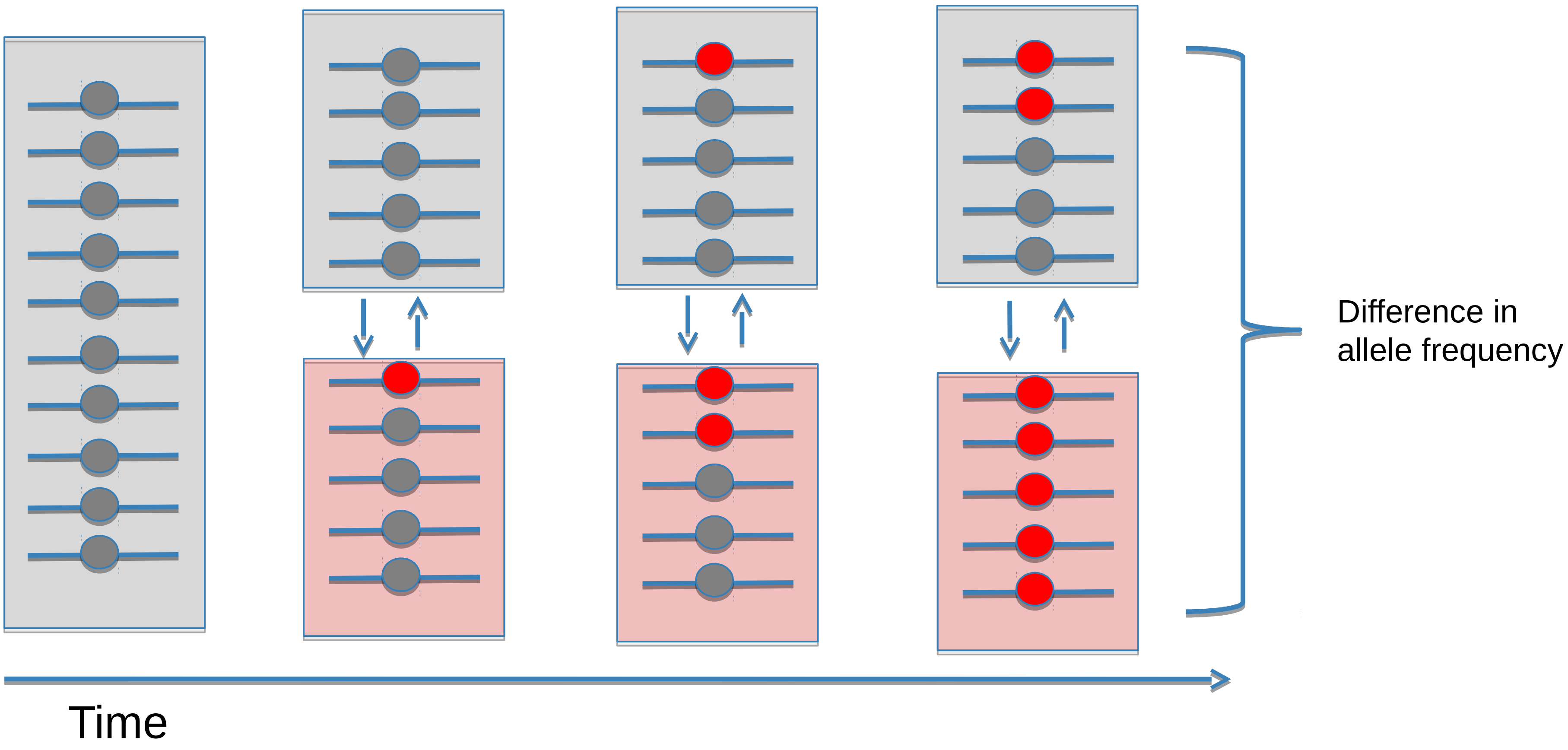
Allele frequency differentiation

From standing variation



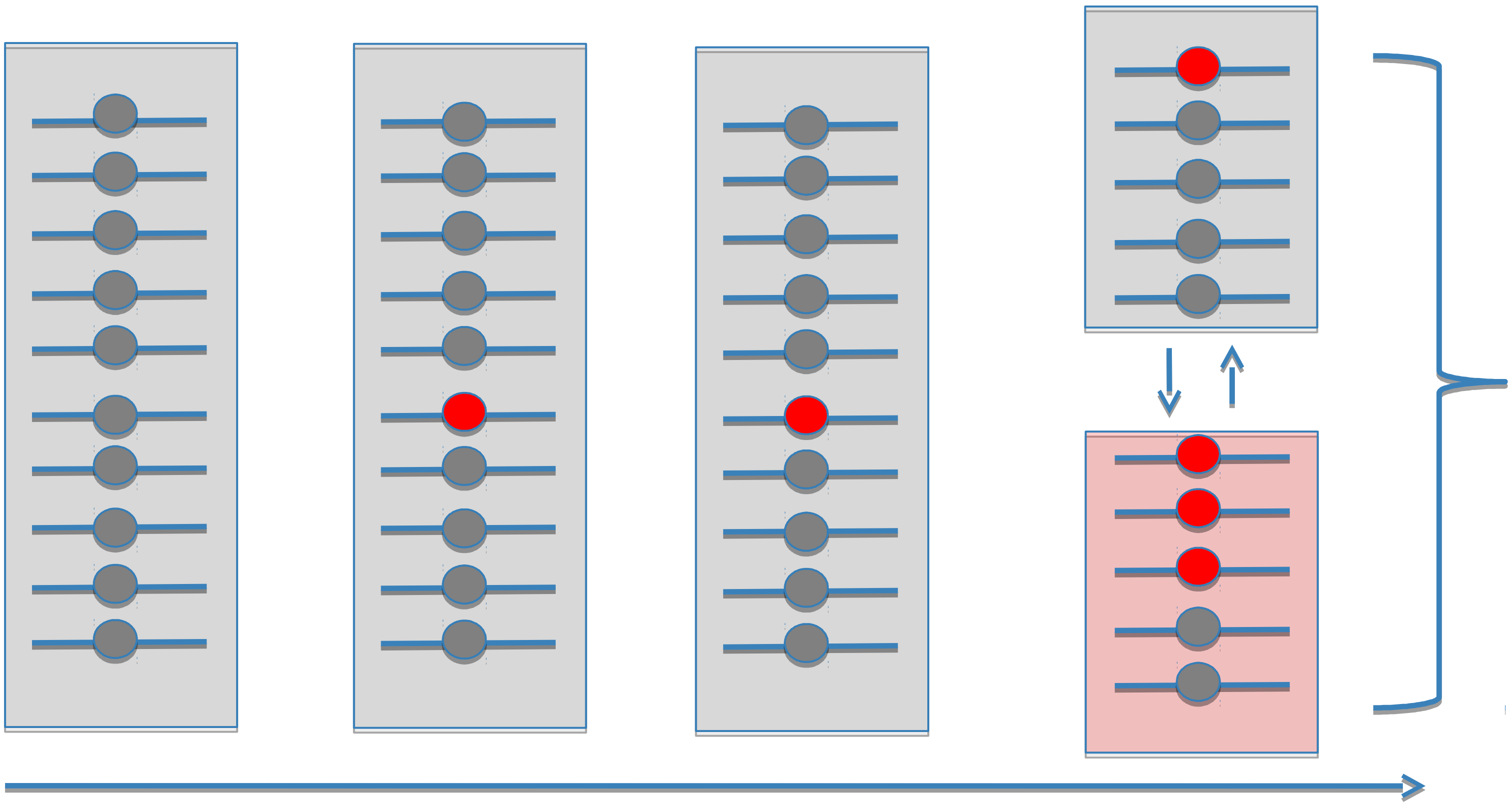
Allele frequency differentiation

With migration



Allele frequency differentiation

With recent divergence

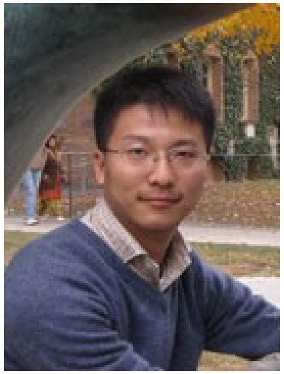


Difference in allele frequency

Time

Population genetic differentiation

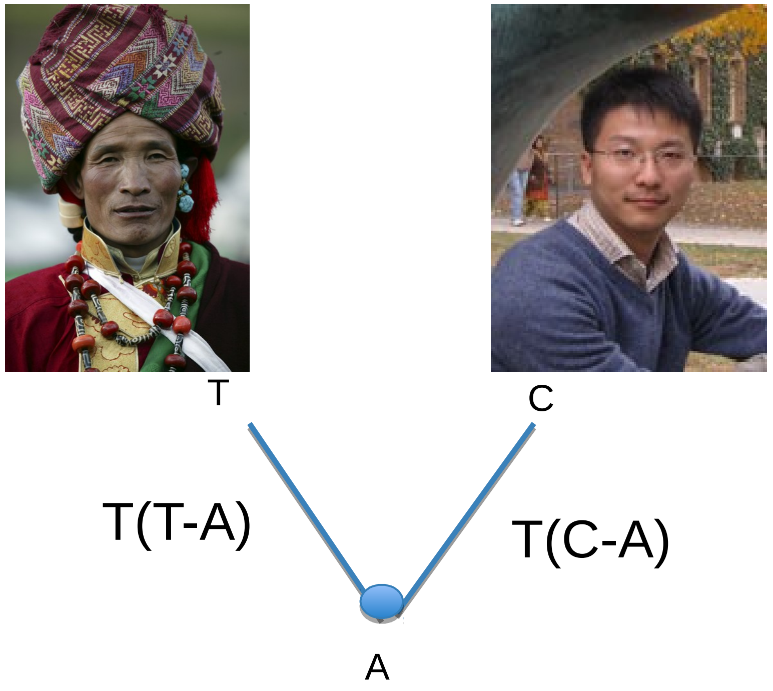
T



C

F_{ST}(T-C)

Population genetic differentiation



T

C

T(T-A)

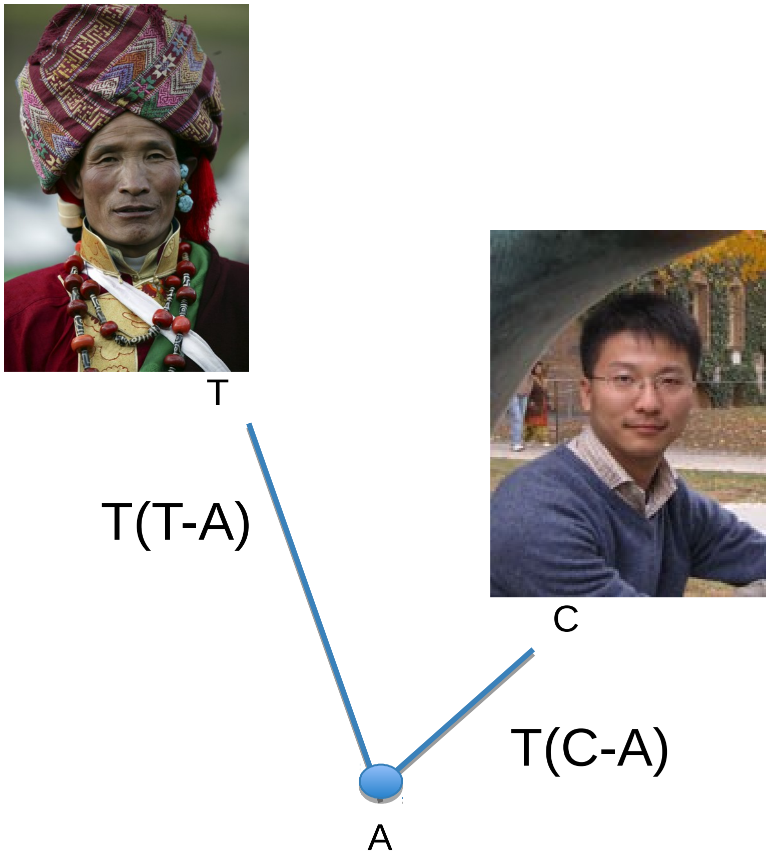
T(C-A)

A

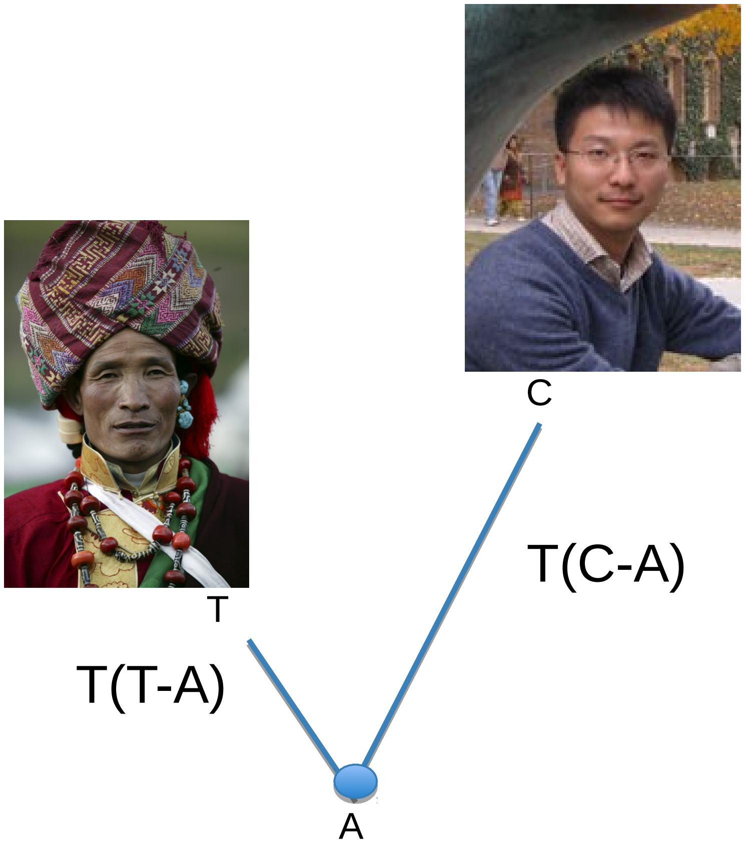
F_{ST}(T-C)~T(T-A-C)

Population genetic differentiation

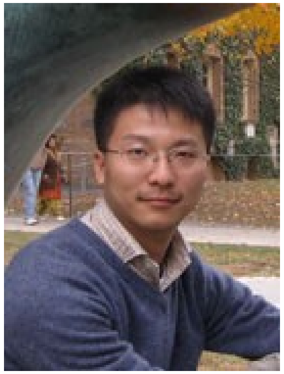
F_{ST}(T-C)~T(T-A-C)

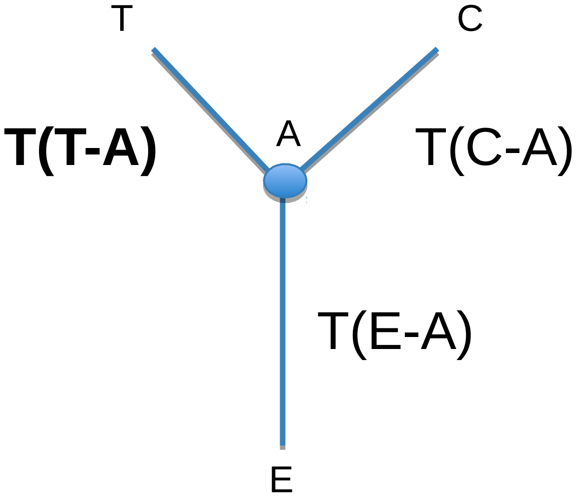


?



Population genetic differentiation





T

C

T(T-A)

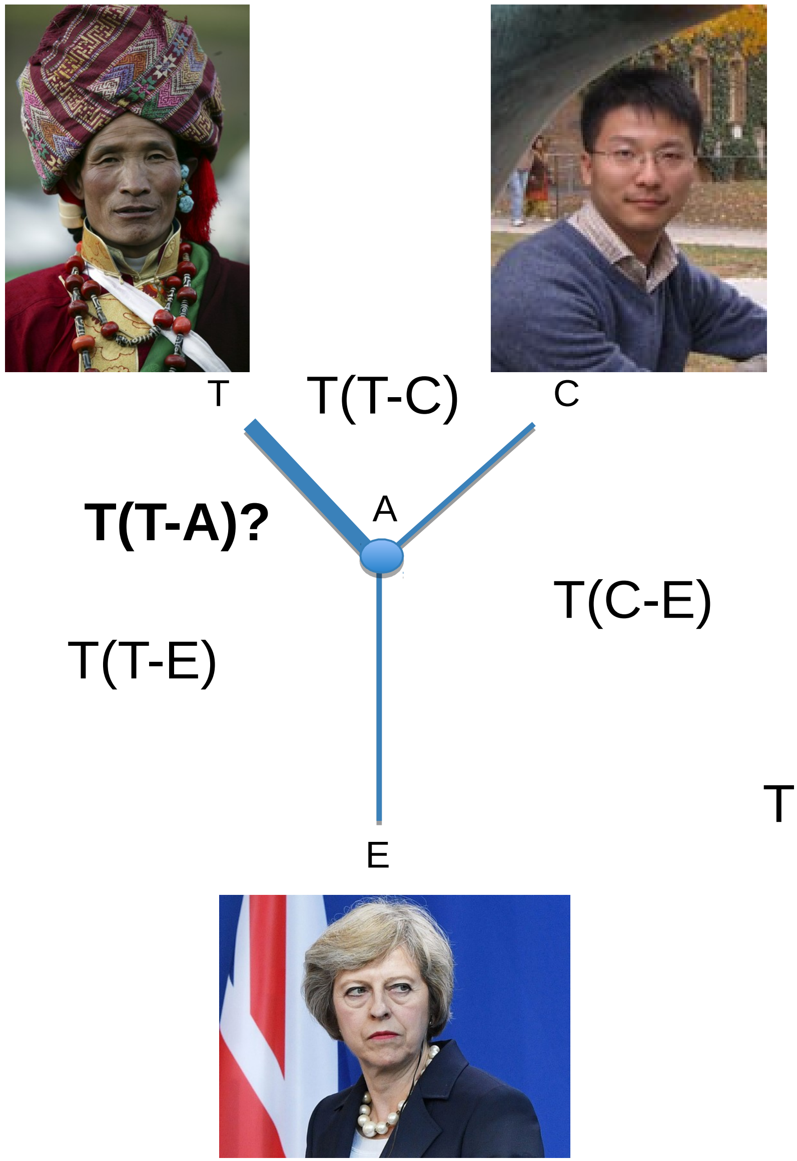
A

T(C-A)

T(E-A)

E

Population genetic differentiation



T

T(T-C)

C

T(T-A)?

A

T(C-E)

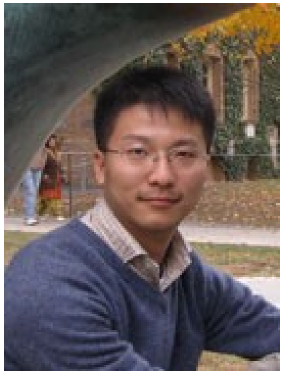
T(T-E)

E

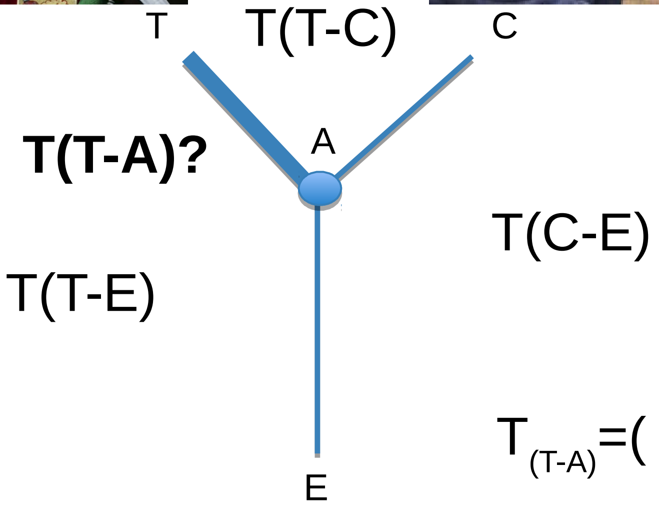
T(T-A-C)=-log(1-F_{ST}(T-C))

$T_{(T\text{-}A)}$= ...

Population genetic differentiation

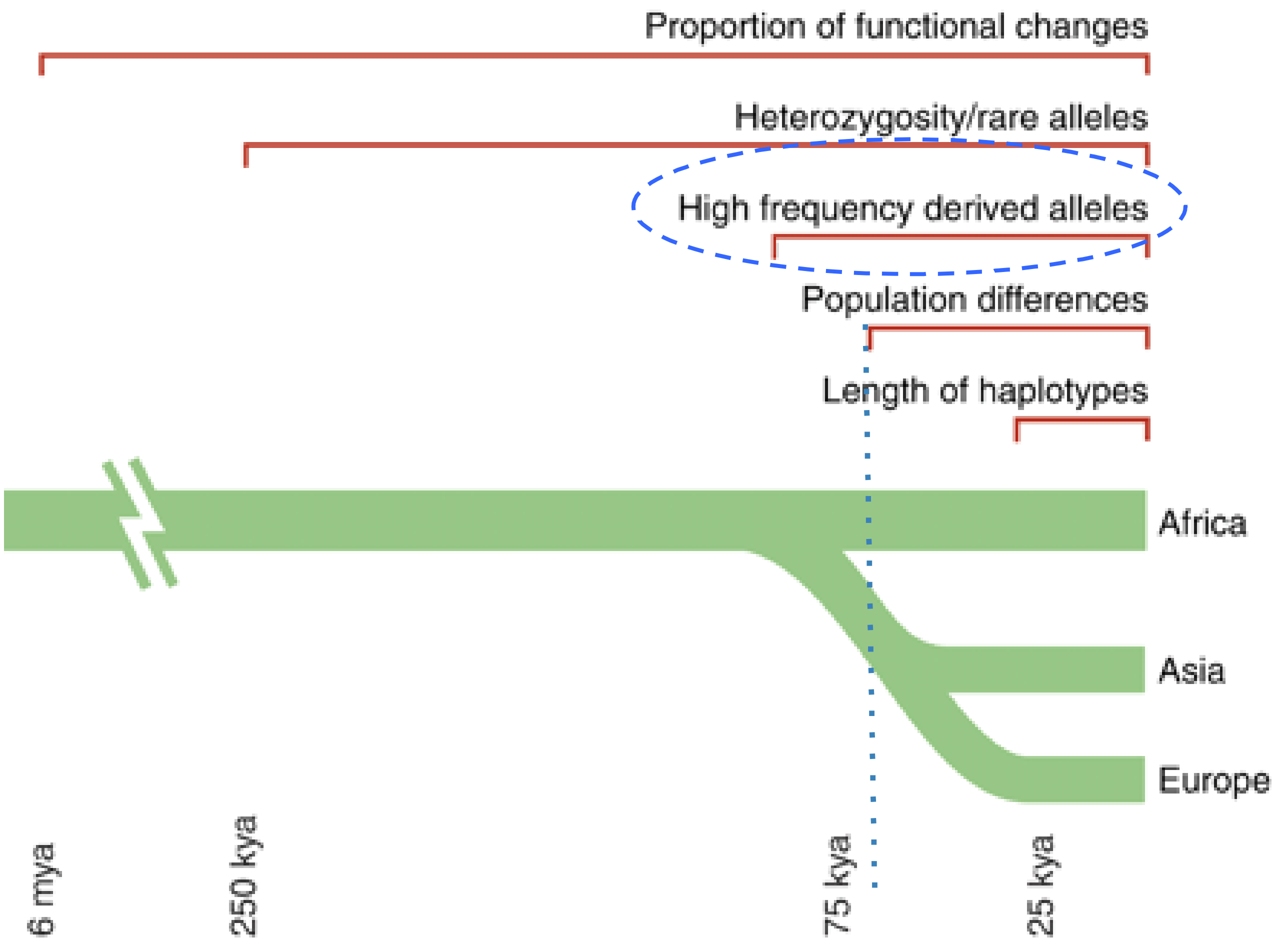


$T(T\text{-}A\text{-}C) = -\log(1 - F_{ST}(T\text{-}C))$



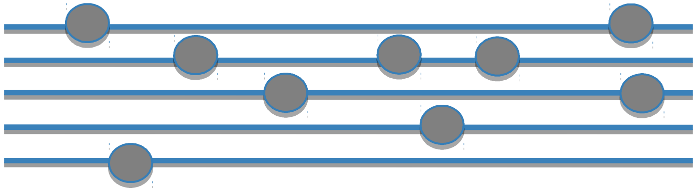
$$T_{(T\text{-}A)} = (\, T_{(T\text{-}E)} + T_{(T\text{-}C)} - T_{(C\text{-}E)} \,)/2$$

Methods to infer selection


Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

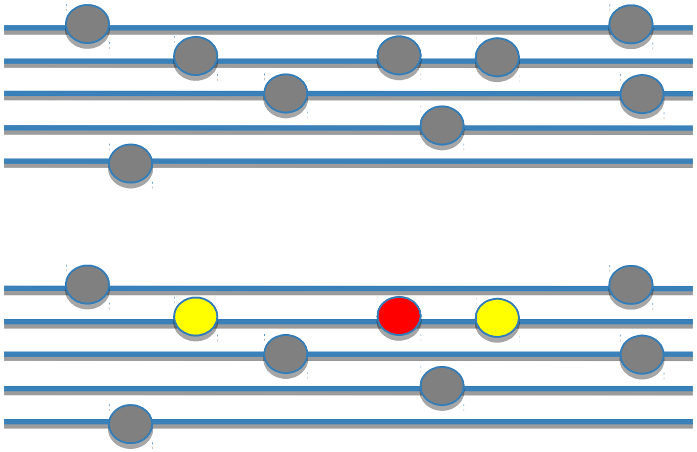
Sabeti et al. 2006 Science

Positive selection: effect on haplotypes



$t<T_{sel}$

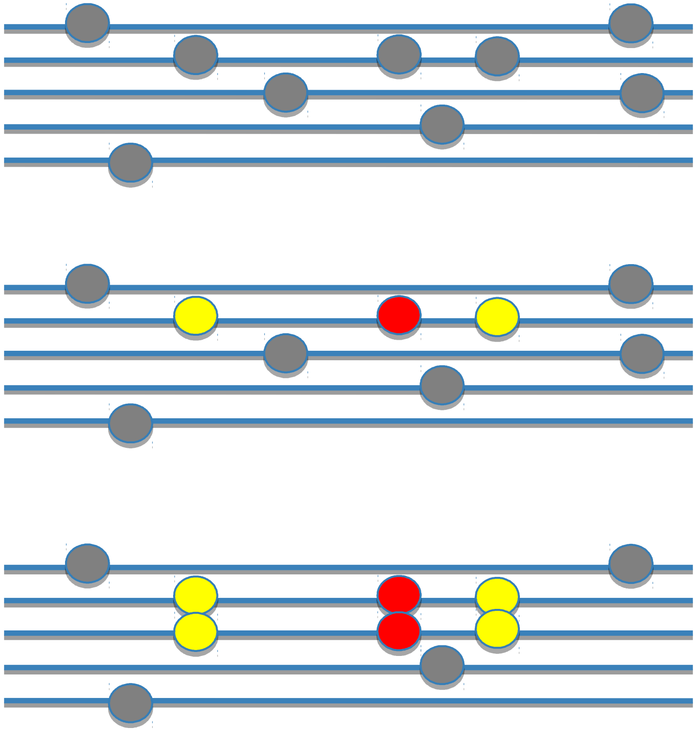
Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

Positive selection: effect on haplotypes

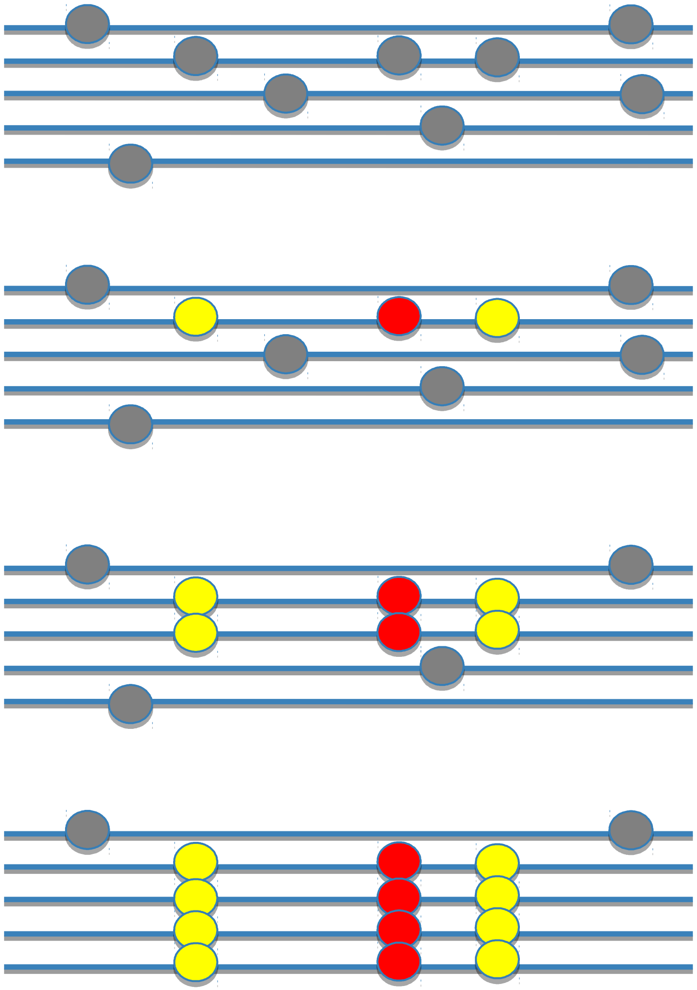


$t<T_{sel}$

$t=T_{sel}$

$t>T_{sel}$

Positive selection: effect on haplotypes



$t<T_{sel}$

$t=T_{sel}$

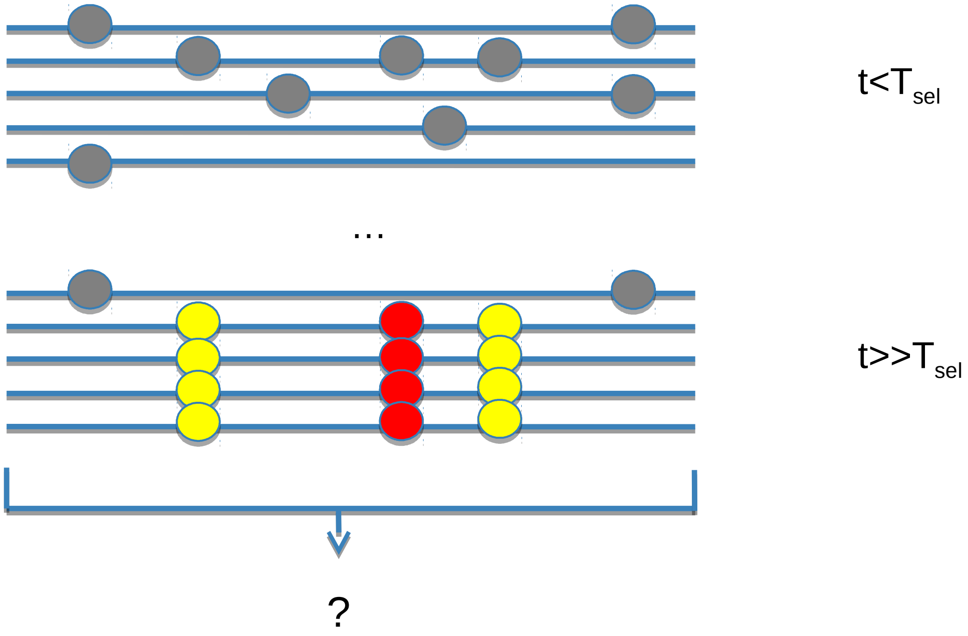
$t>T_{sel}$

$t>>T_{sel}$

Selective sweep

Genetic hitch-hiking

Positive selection



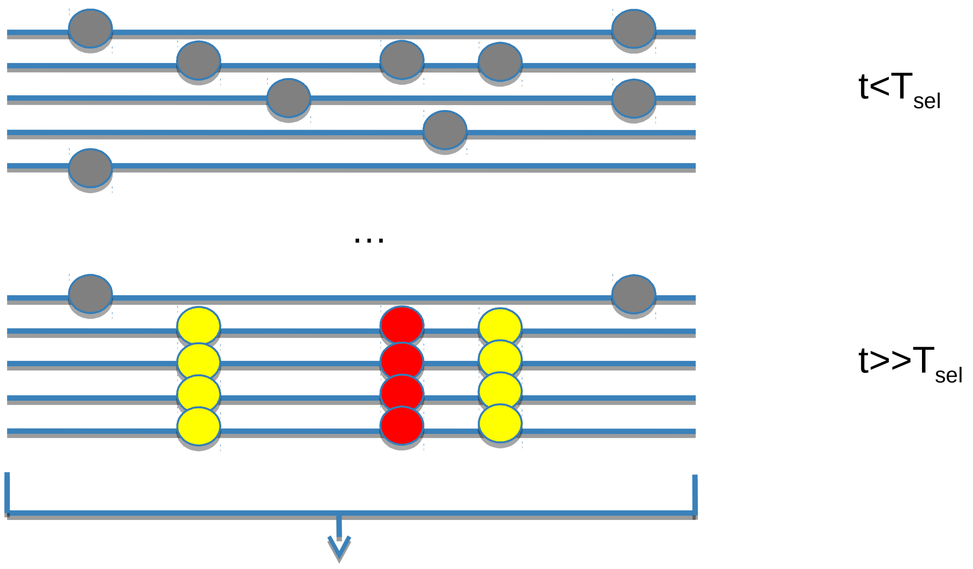
$t<T_{sel}$

…

$t>>T_{sel}$

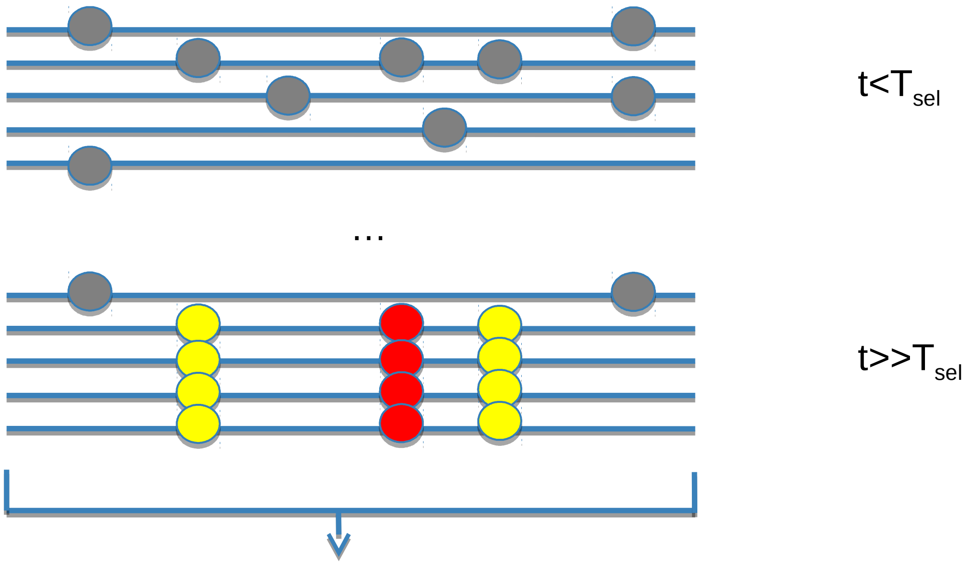
?

Positive selection



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

Positive selection



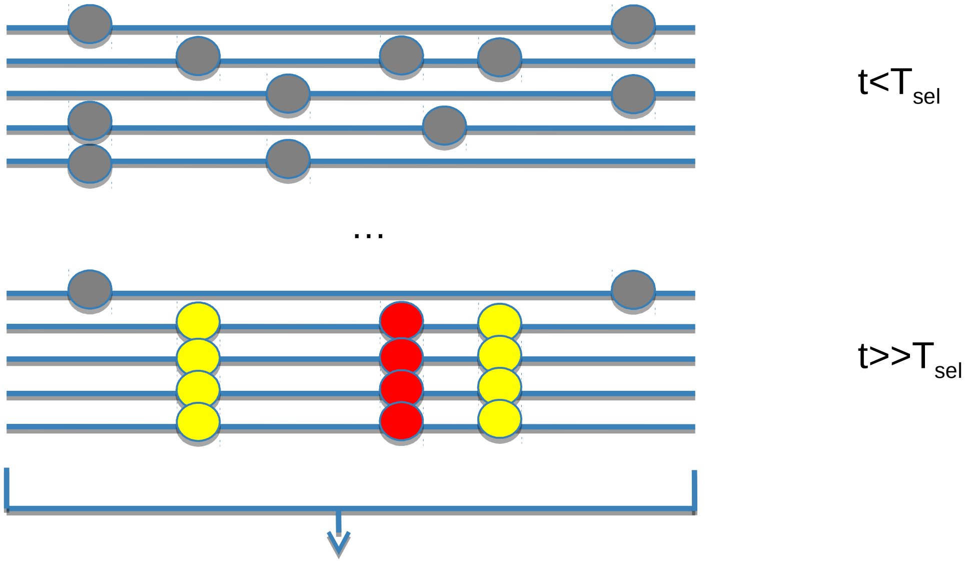
- Reduction of polymorphisms levels
(e.g. from 7 to 5 SNPs)

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

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

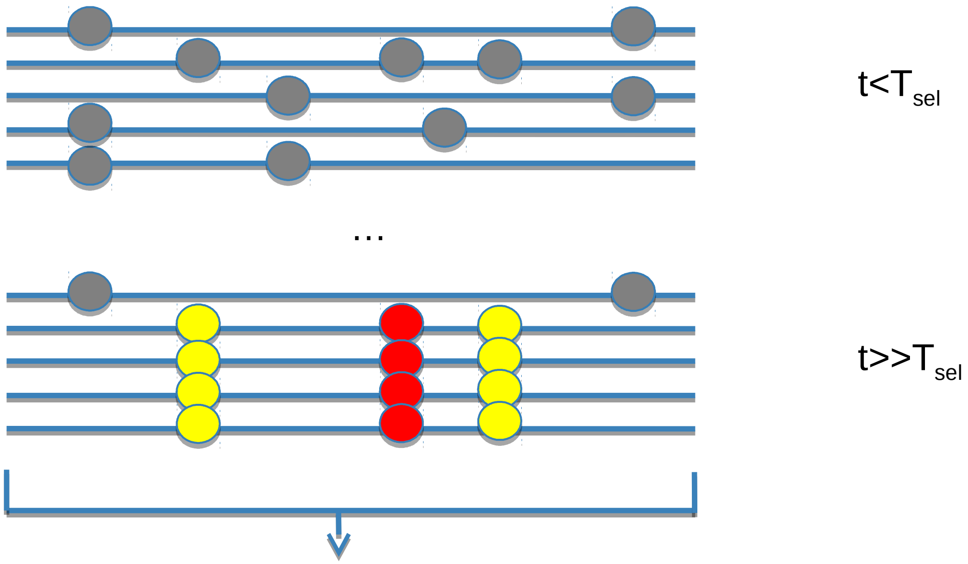
$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Positive selection



- Reduction of polymorphisms levels (Theta)
- ?

Positive selection

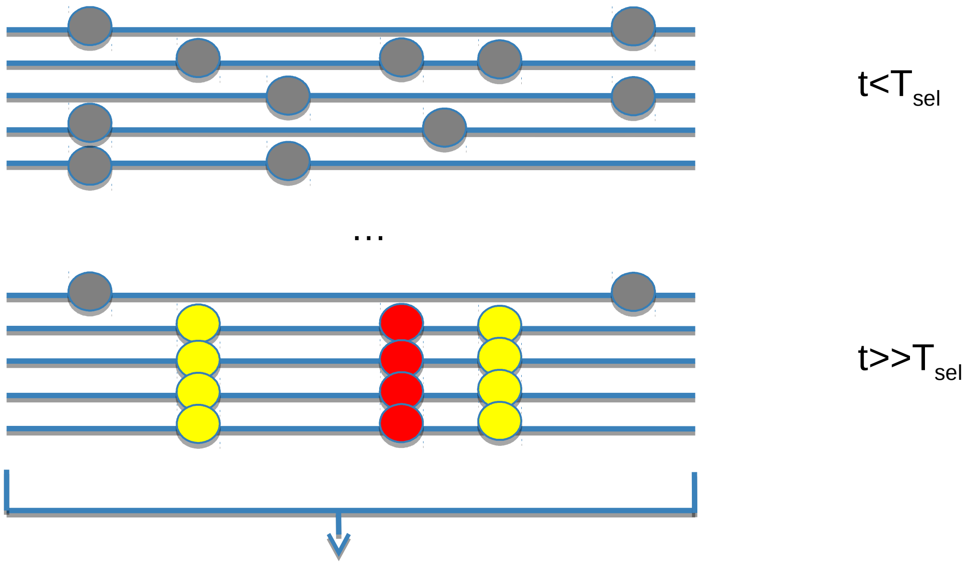


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

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

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

Positive selection



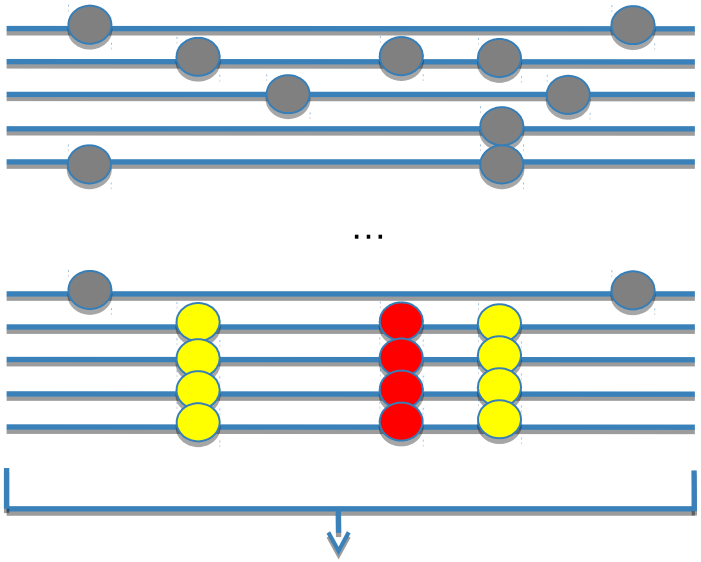
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

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

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

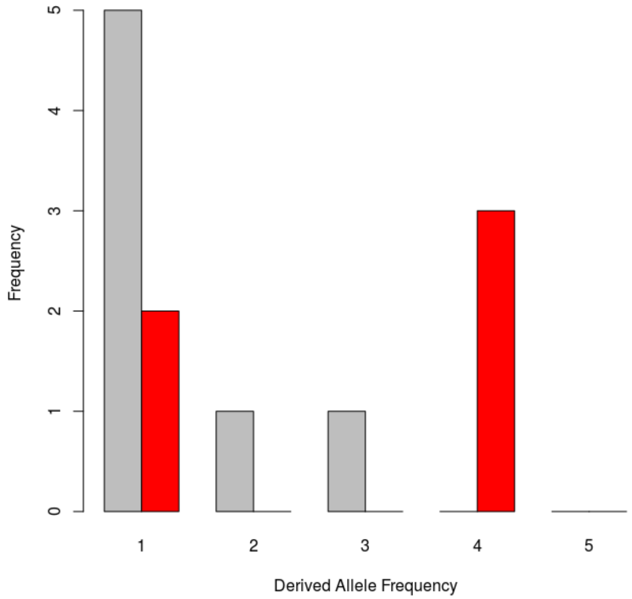
D<0 is suggestive of an excess of low-frequency variants

The Site Frequency Spectrum

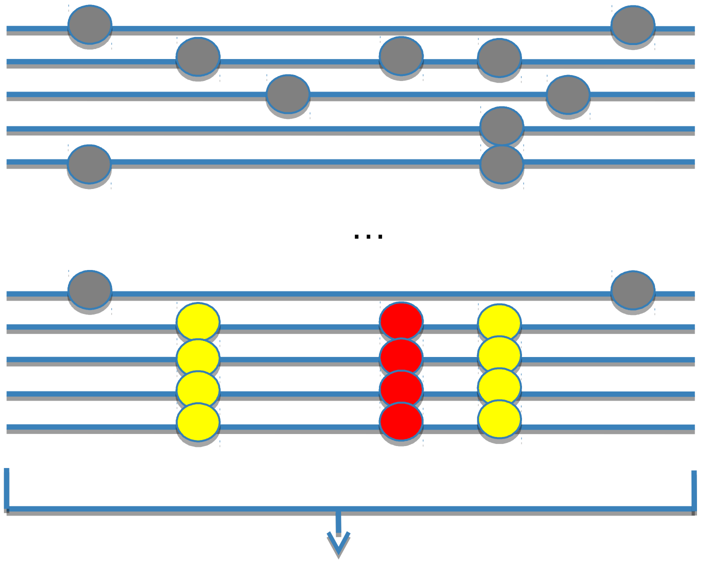


$t<T_{sel}$

$t>>T_{sel}$

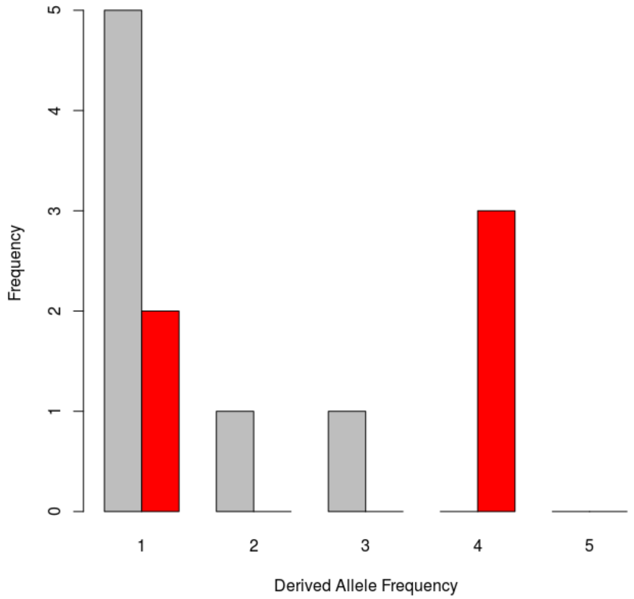


The Site Frequency Spectrum



$t<T_{sel}$

$t>>T_{sel}$



Tajima's D?

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

~0.33=1/3

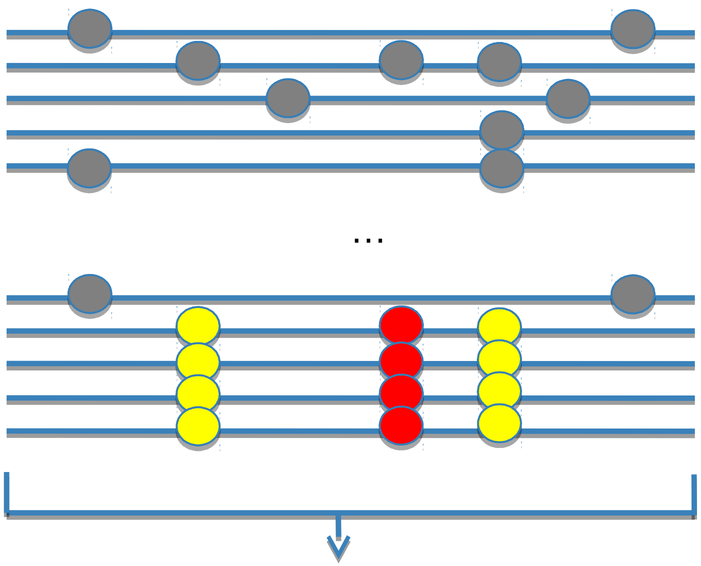
$$\theta_W = \frac{K}{a_n}$$

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

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

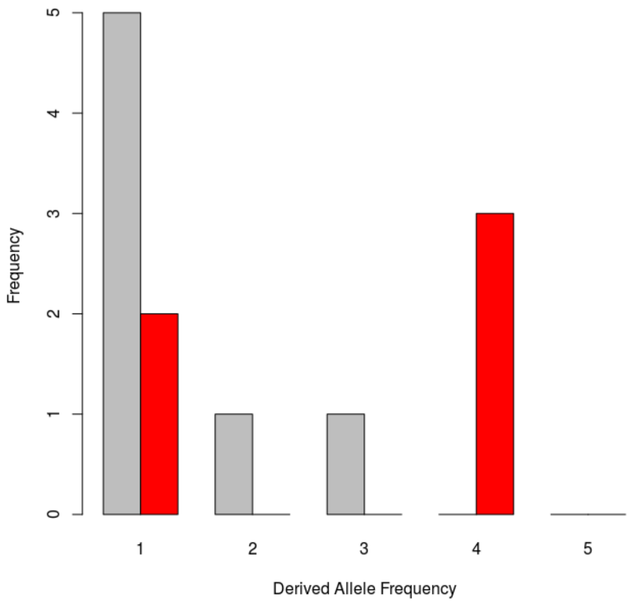
= 10, the number of comparisons you need to make

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

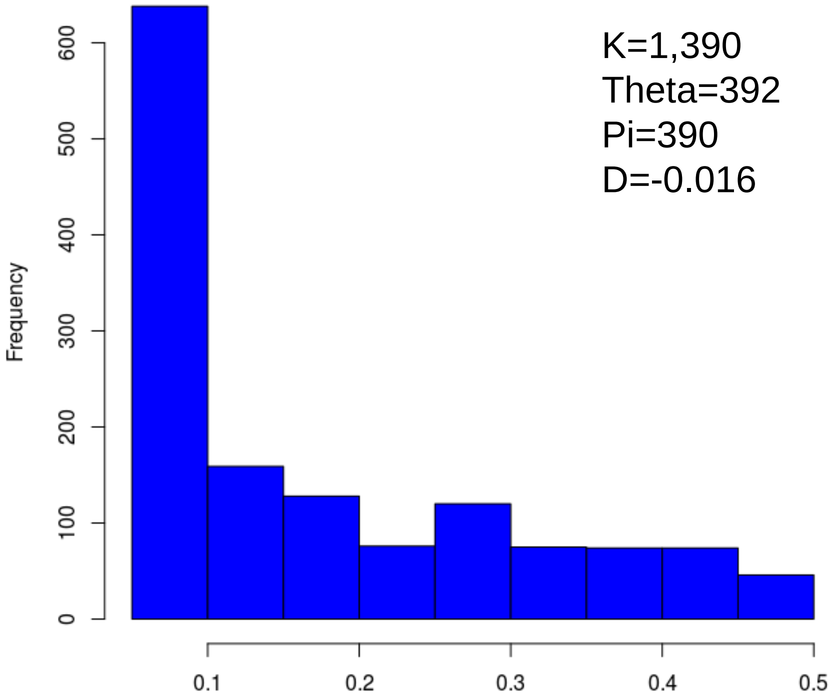
sd(D)=1/3

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

D<0

Demography matters?

n=20; L=500kbp; no selection

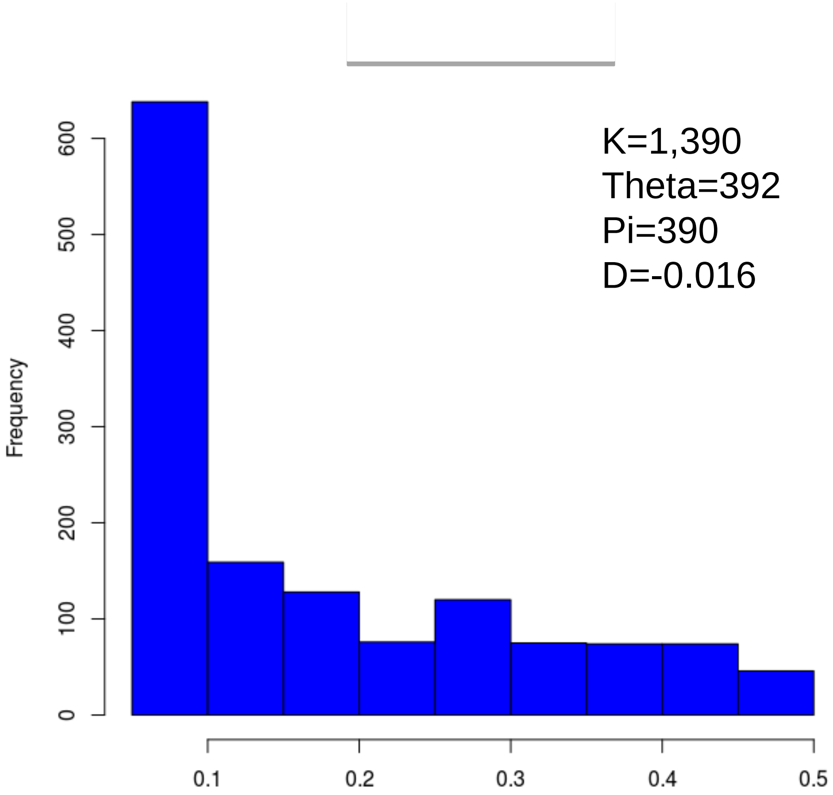


K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

Demography matters?

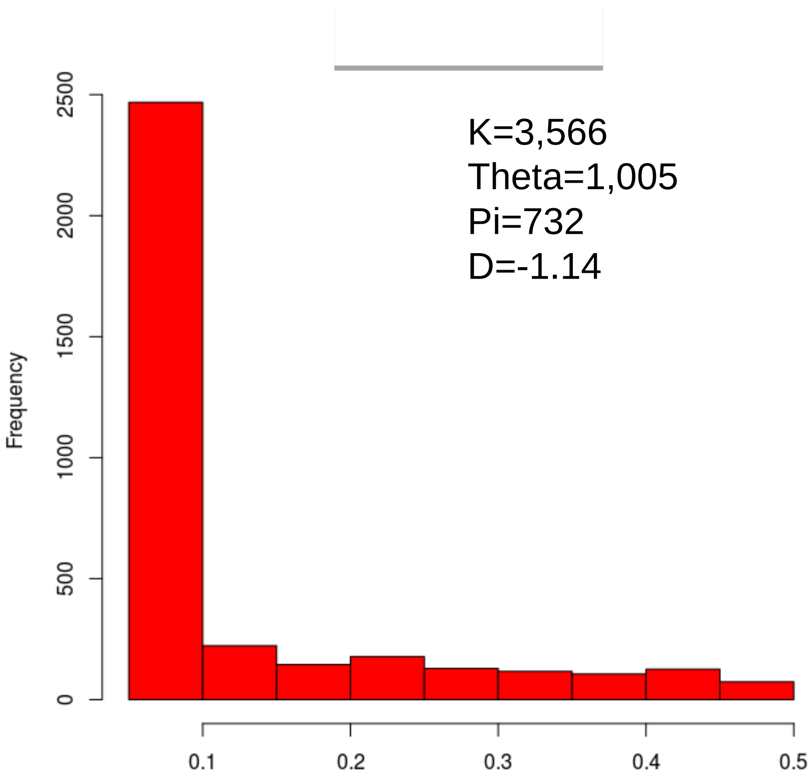
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection

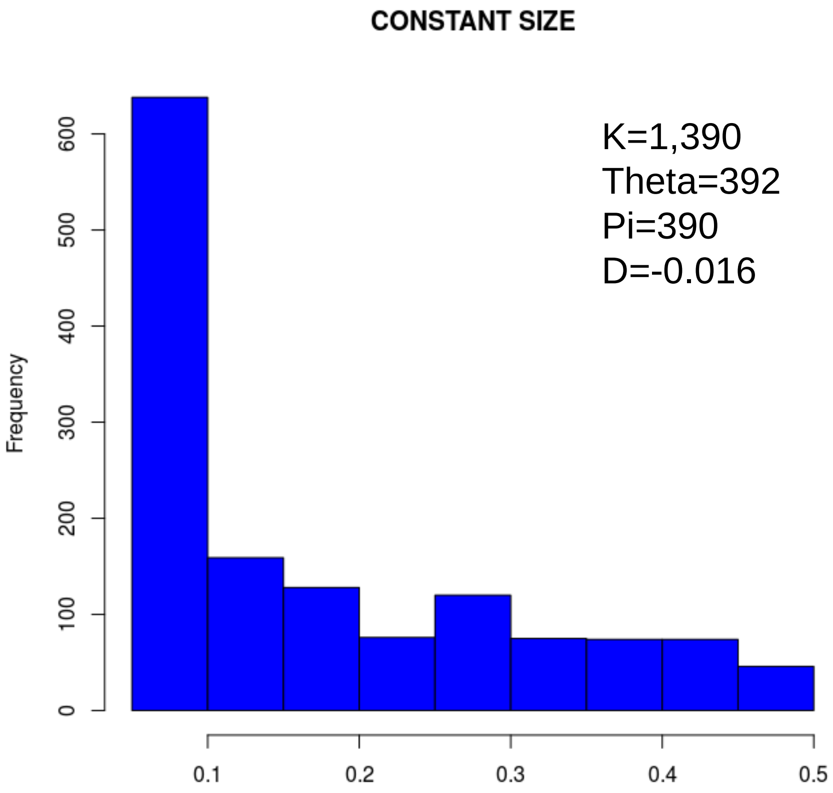


K=3,566
Theta=1,005
Pi=732
D=-1.14

Minor Allele Frequency

Demography matters!

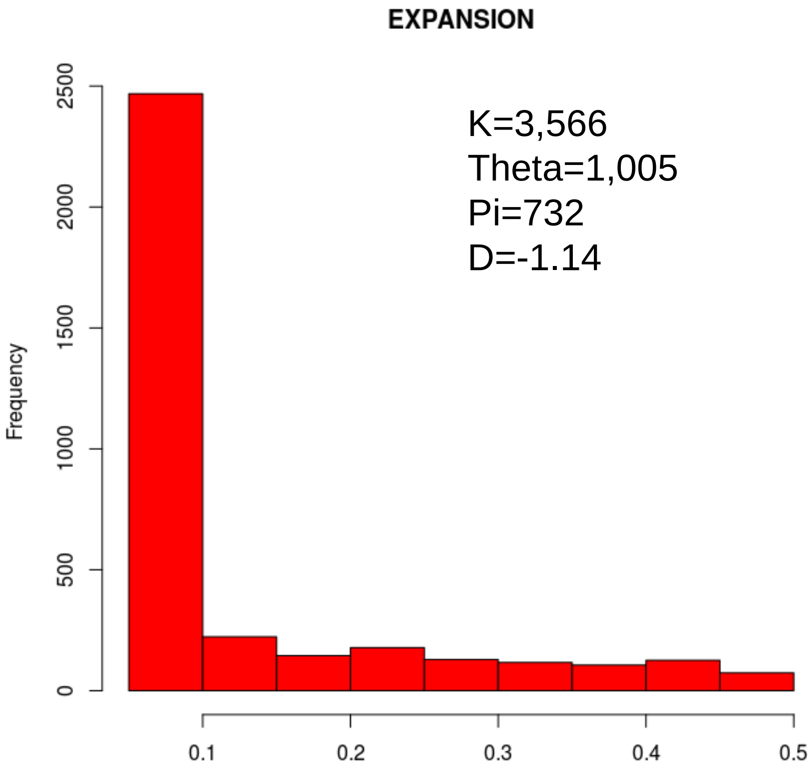
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection

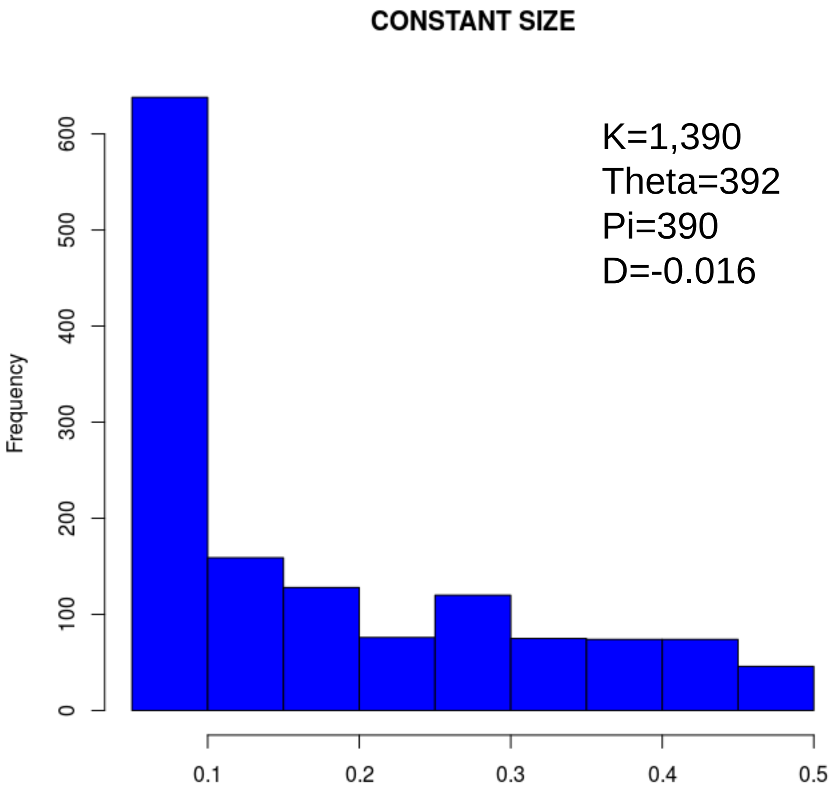


K=3,566
Theta=1,005
Pi=732
D=-1.14

Minor Allele Frequency

Demography matters!

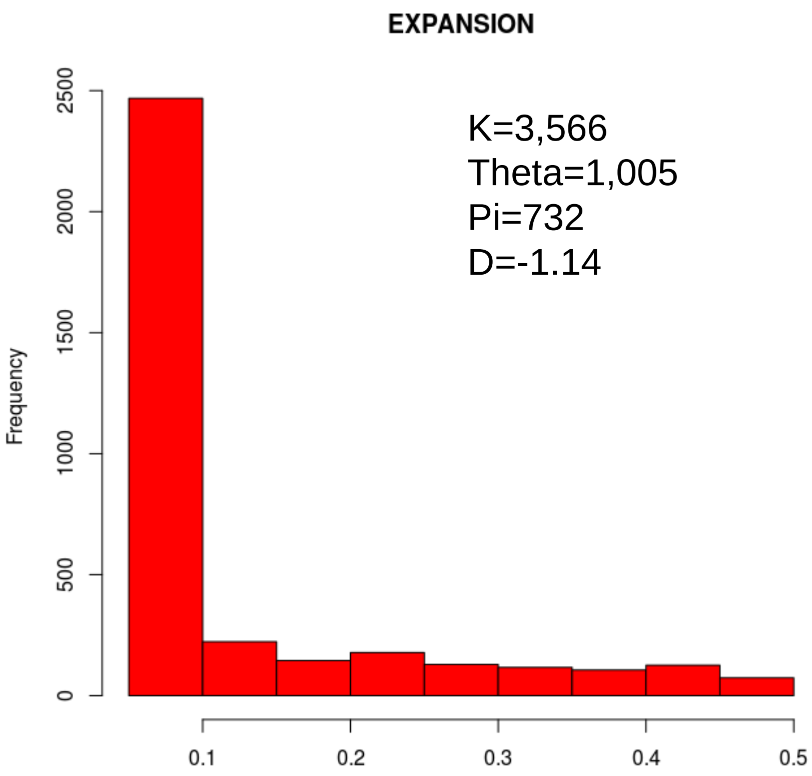
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection



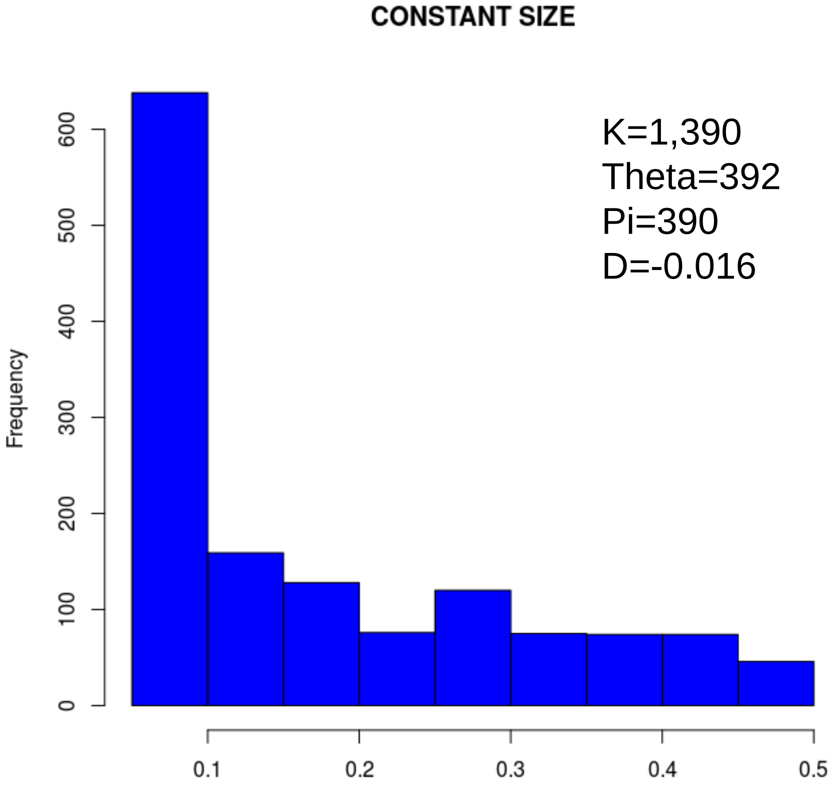
K=3,566
Theta=1,005
Pi=732
D=-1.14

Minor Allele Frequency

- 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?

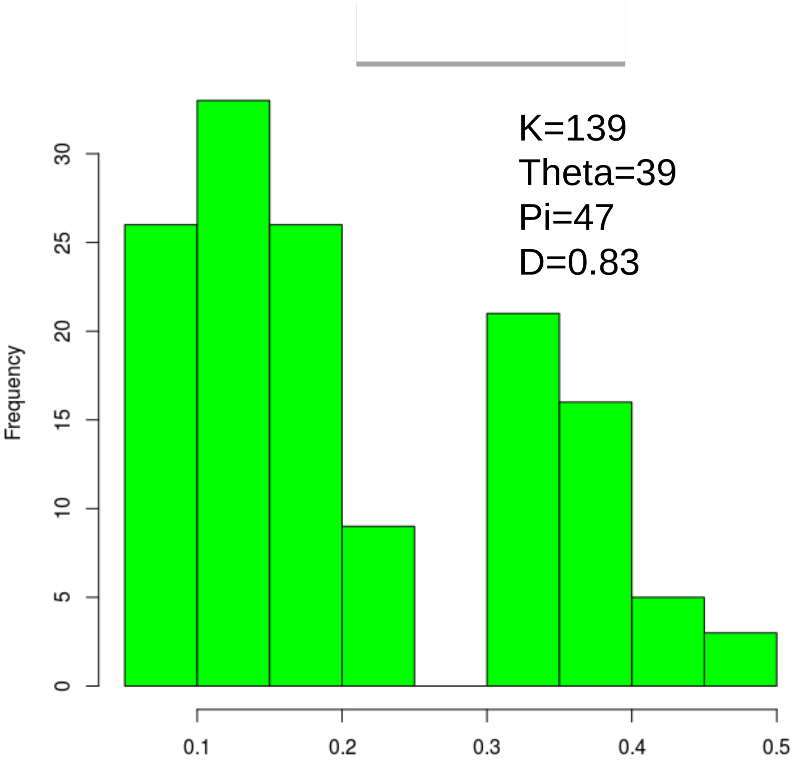
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection

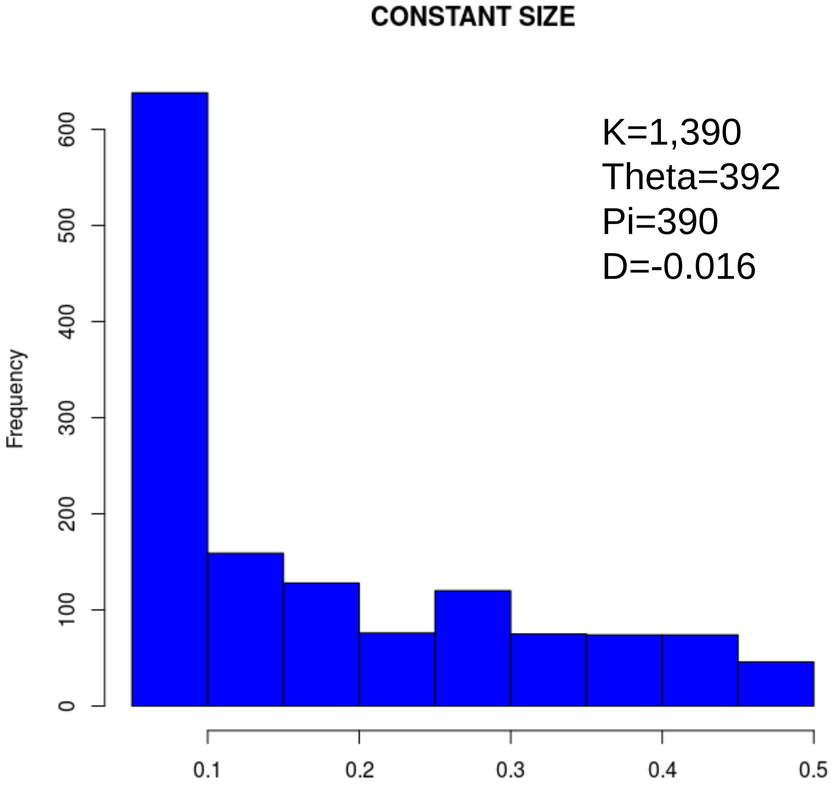


K=139
Theta=39
Pi=47
D=0.83

Minor Allele Frequency

Demography matters!

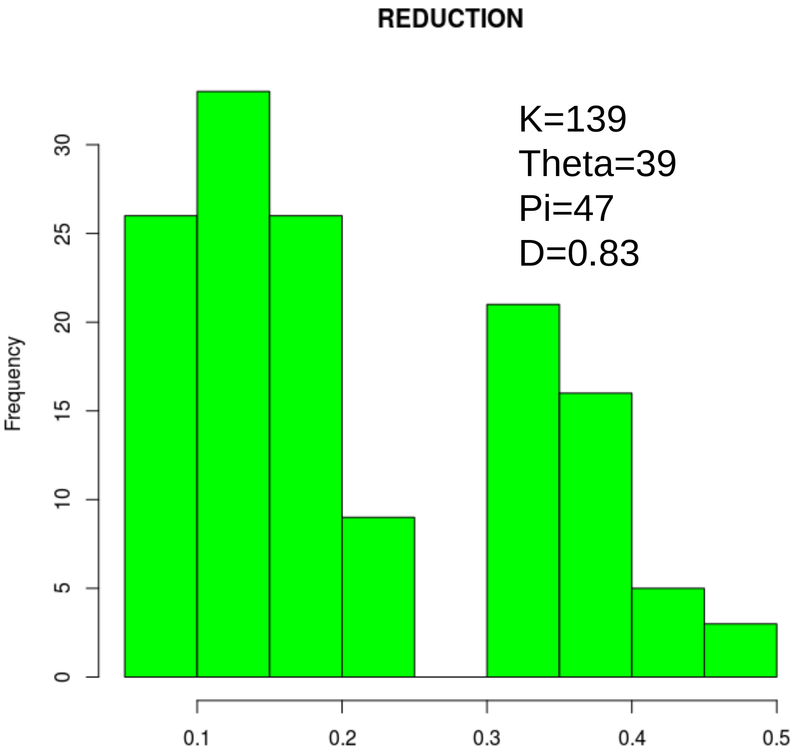
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection

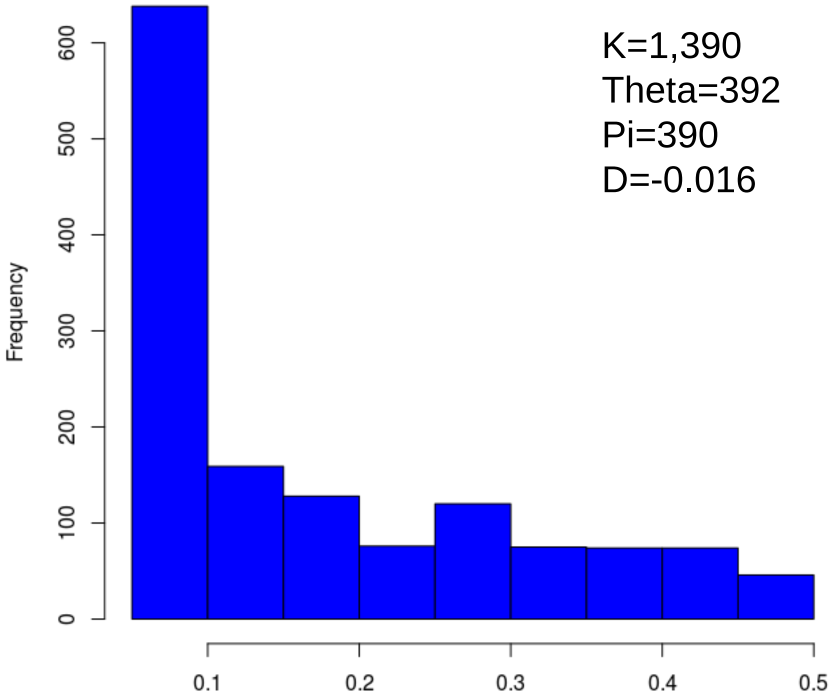


K=139
Theta=39
Pi=47
D=0.83

Minor Allele Frequency

Demography matters!

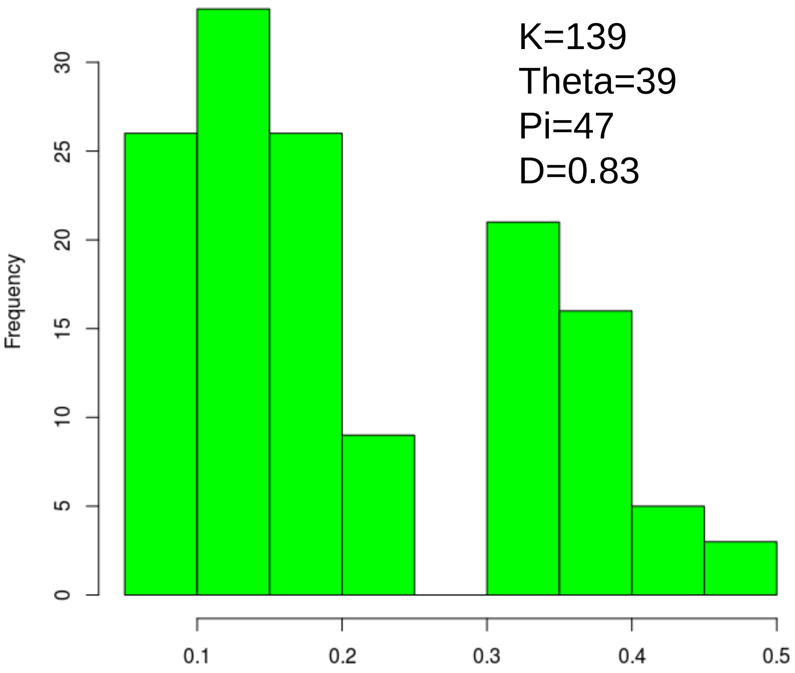
n=20; L=500kbp; no selection



K=1,390
Theta=392
Pi=390
D=-0.016

Minor Allele Frequency

n=20; L=500kbp; no selection



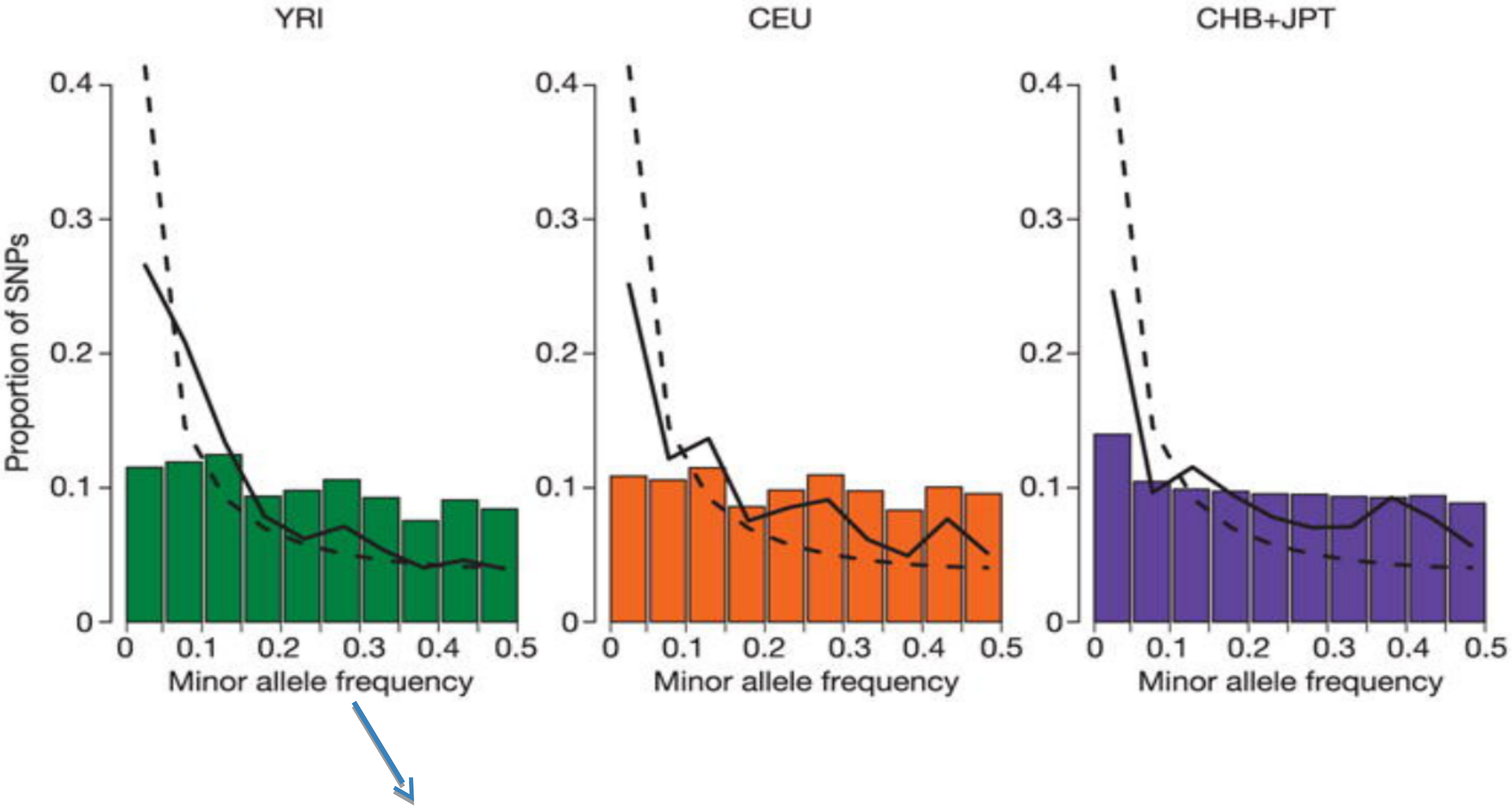
K=139
Theta=39
Pi=47
D=0.83

Minor Allele Frequency

- 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

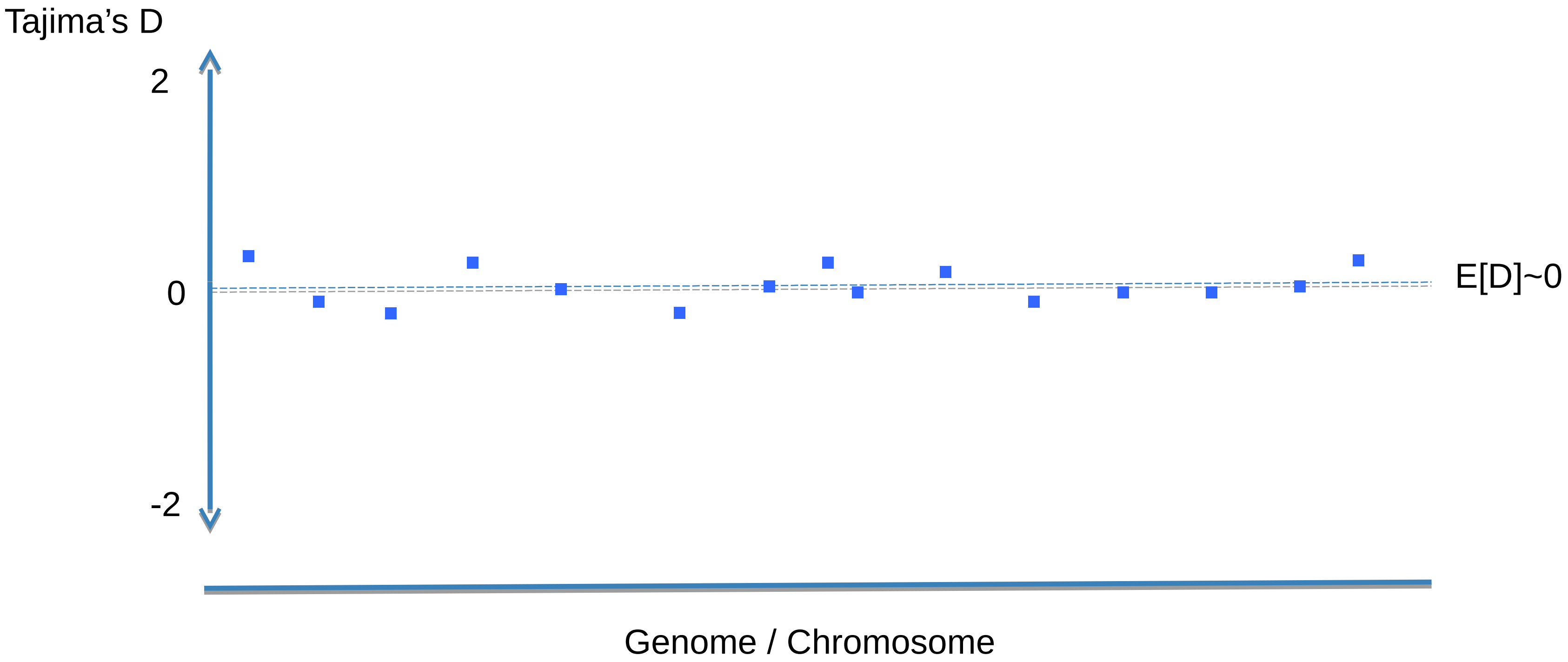


Deficiency of low-frequency variants

HapMap Consortium. Nature 2005

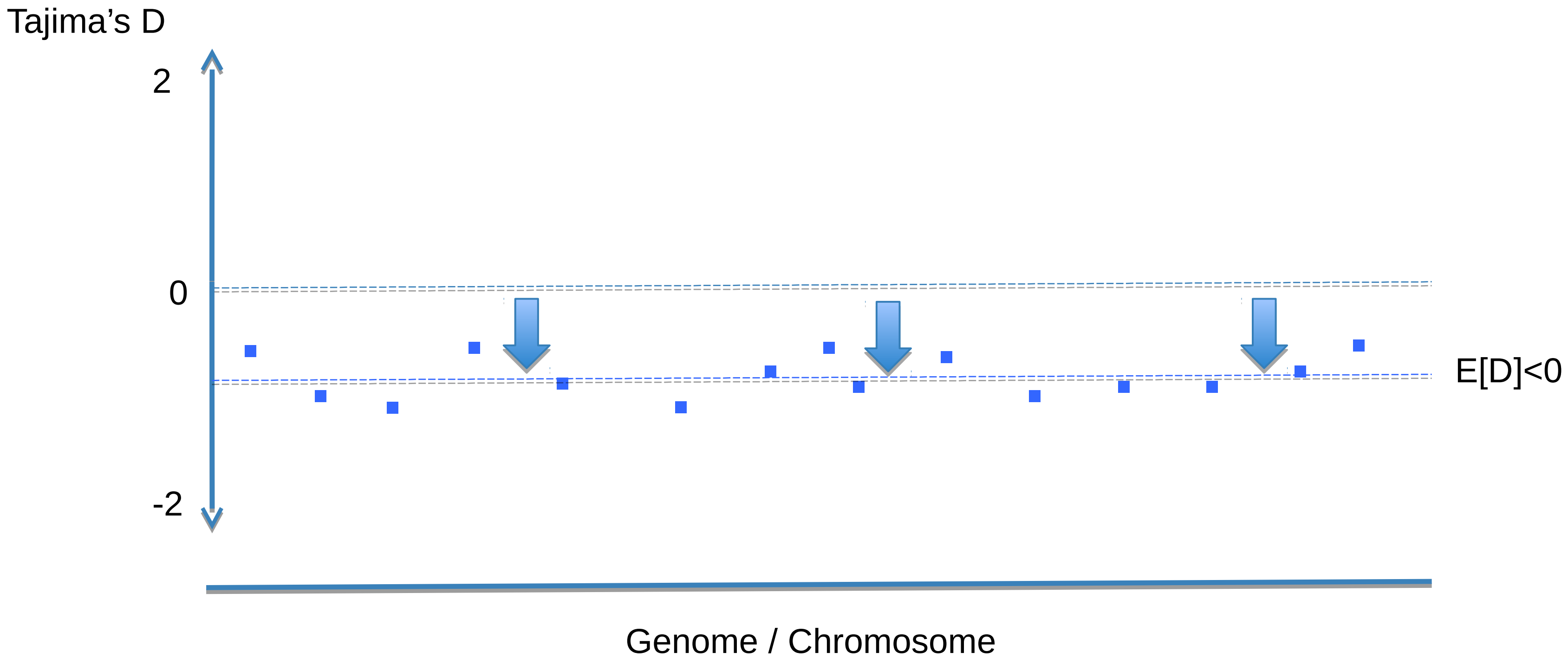
How to take neutral confounding factors into account?

Under constant population size:



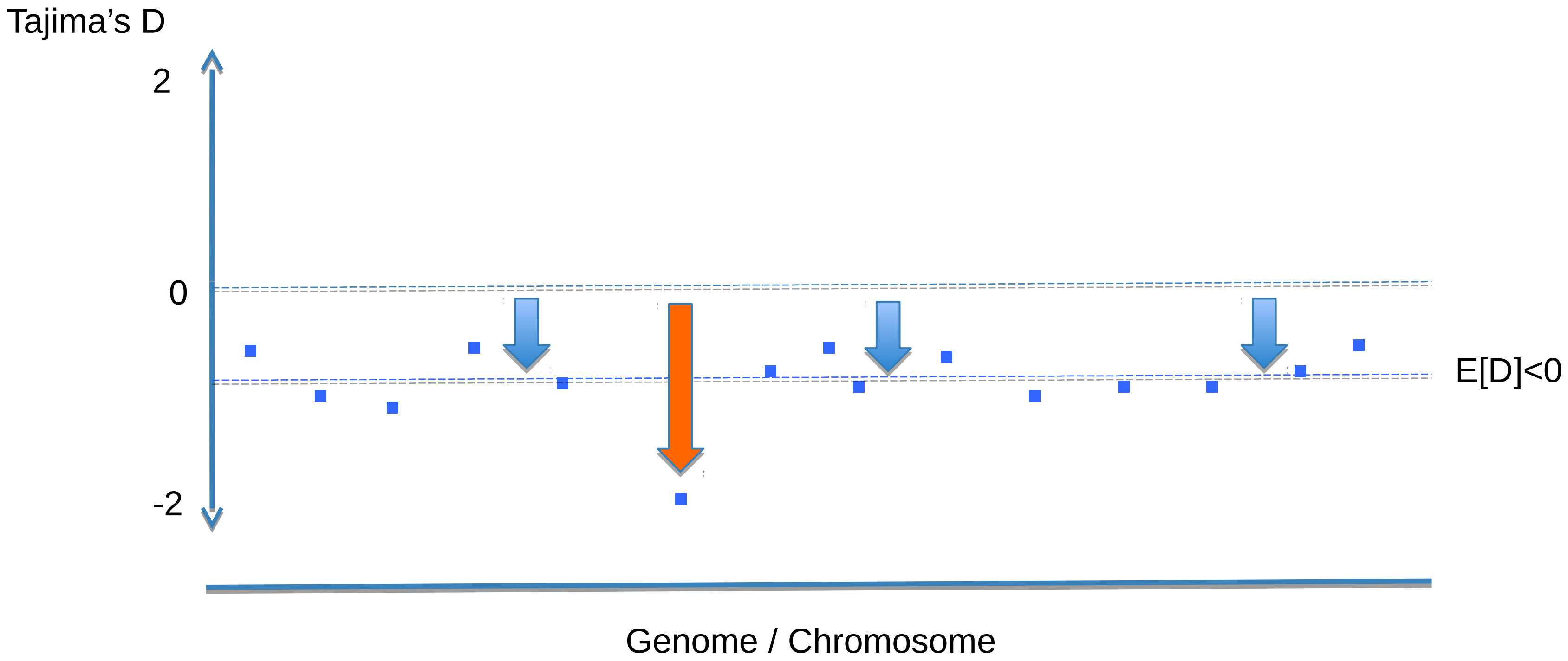
How to take neutral confounding factors into account?

Under expanding population size:



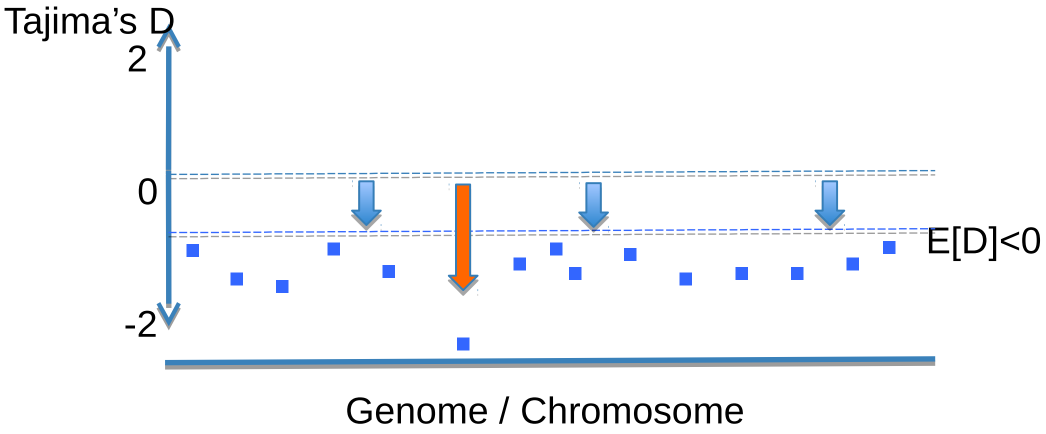
How to take neutral confounding factors into account?

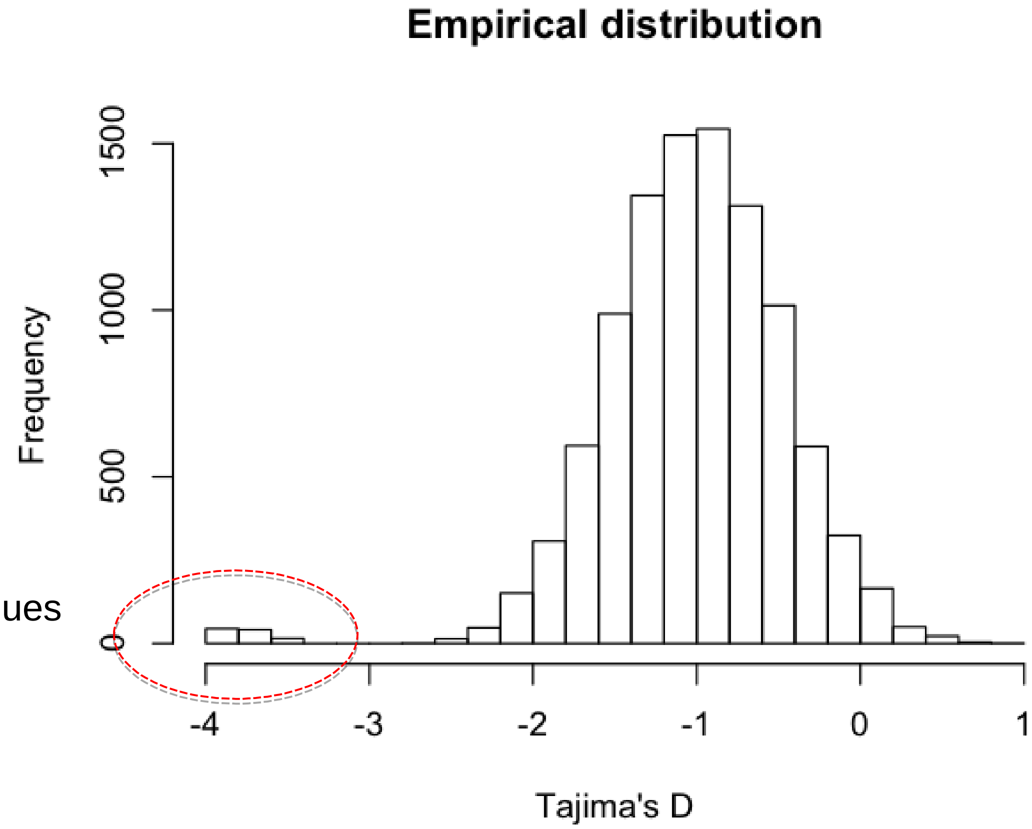
Under expanding population size and positive selection:



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

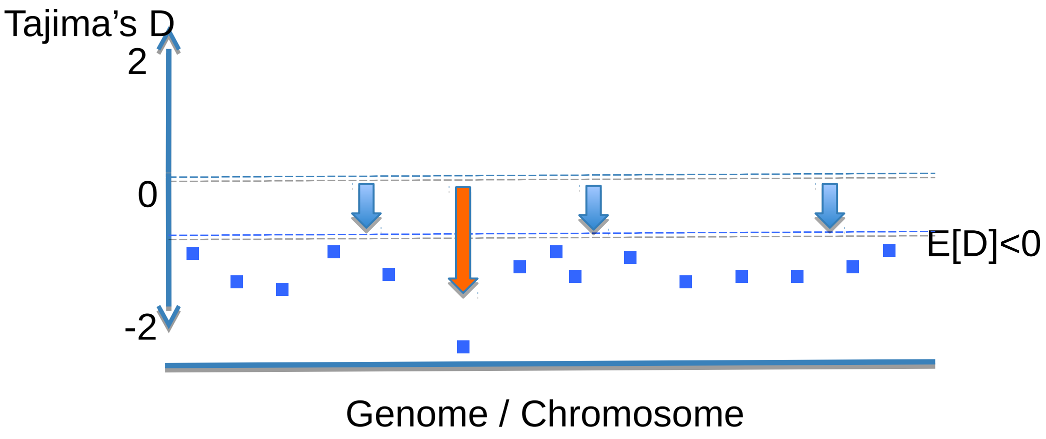
Outlier approach

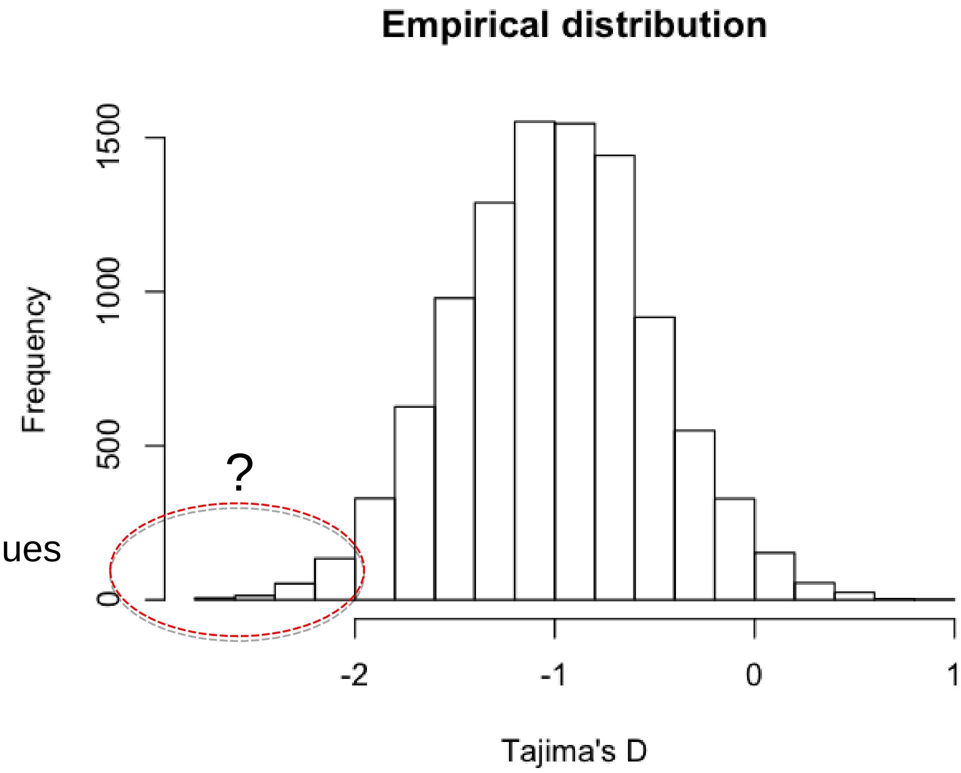




Assign empirical *p*-values
(ranked percentiles)

Outlier approach

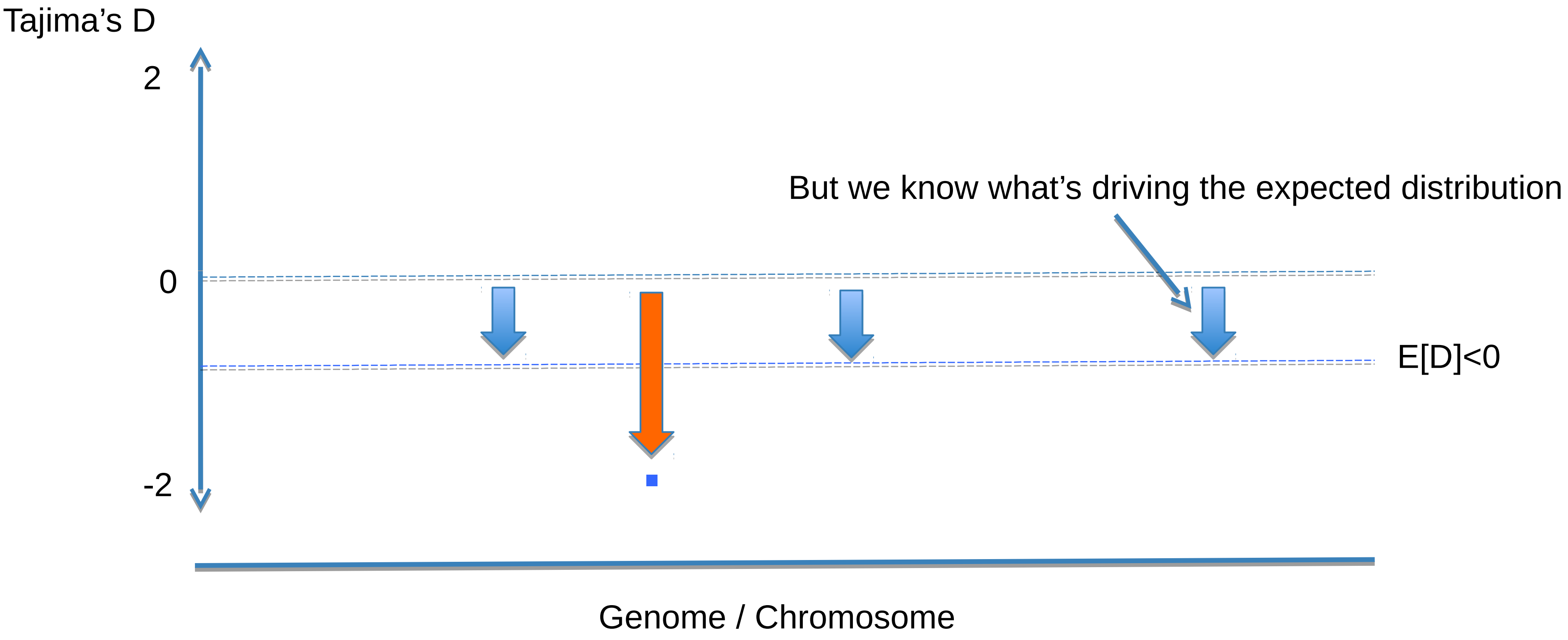




Assign empirical *p*-values
(ranked percentiles)

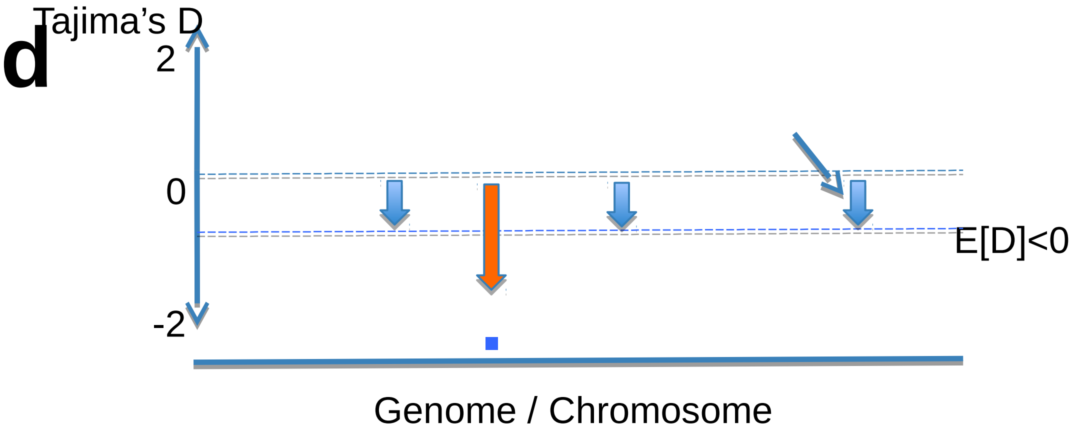
How to take neutral confounding factors into account?

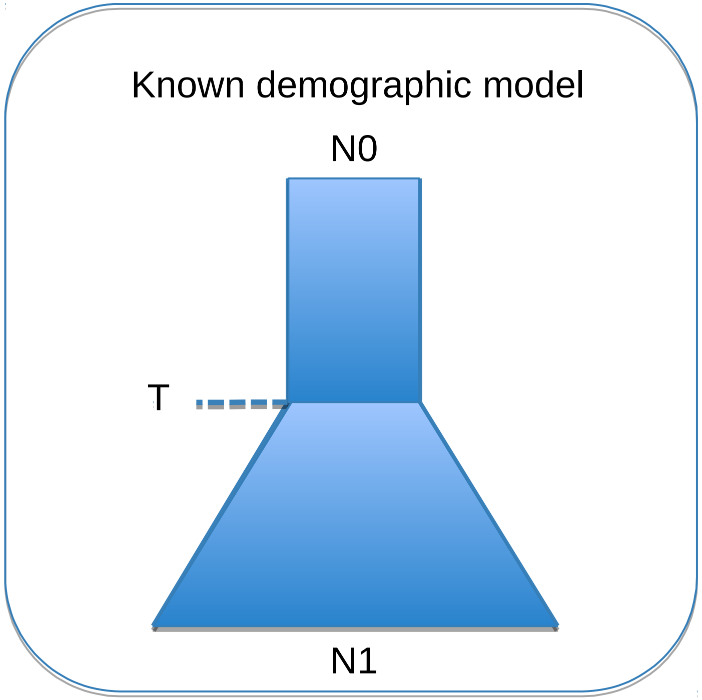
Under expanding population size and positive selection:



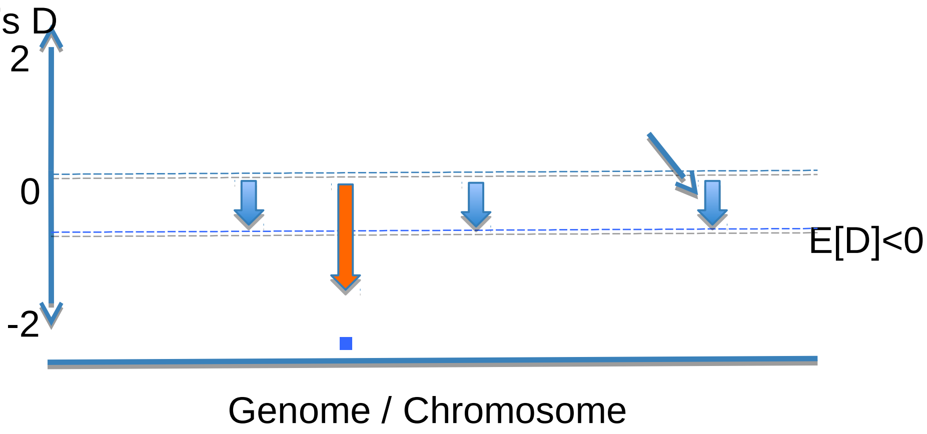
- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

Simulations-based approach

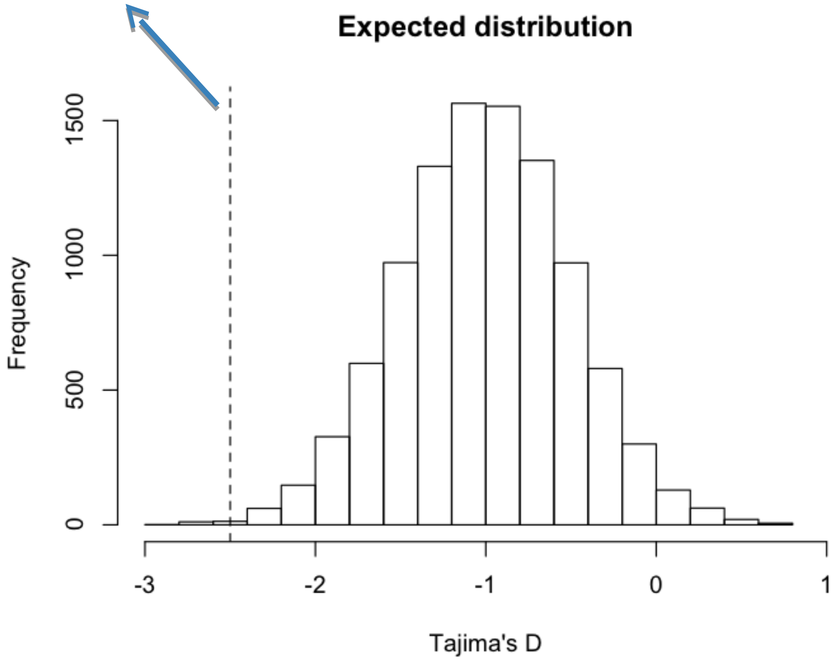


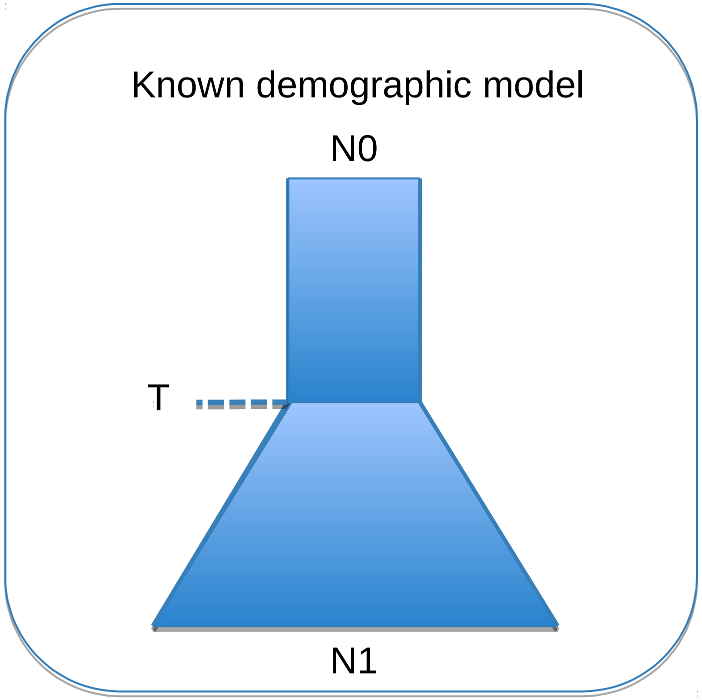


Simulations-based approach



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


Expected distribution
Frequency
Tajima's D


Known demographic model
N0
T
N1