

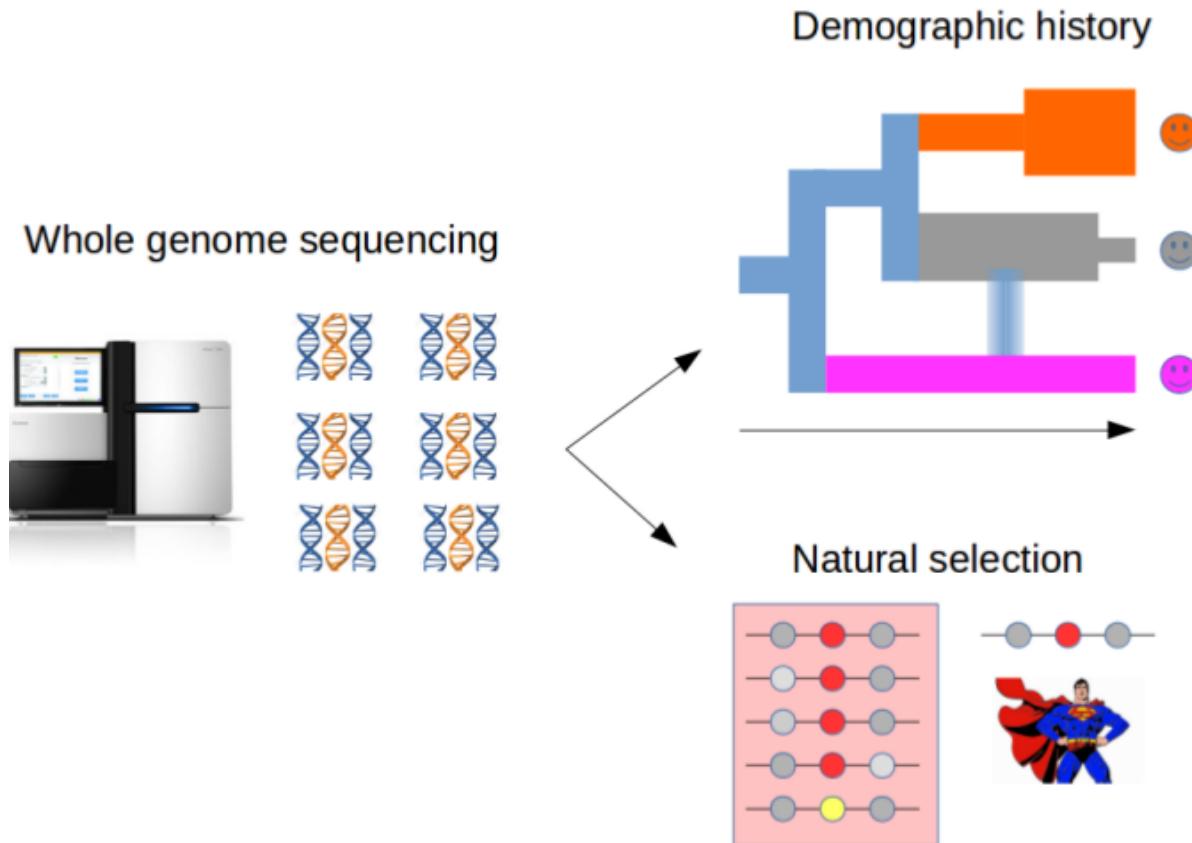
Detecting natural selection from genomic data

(part 1)

Matteo Fumagalli

August 25th 2017

Bioinformatics for Adaptation



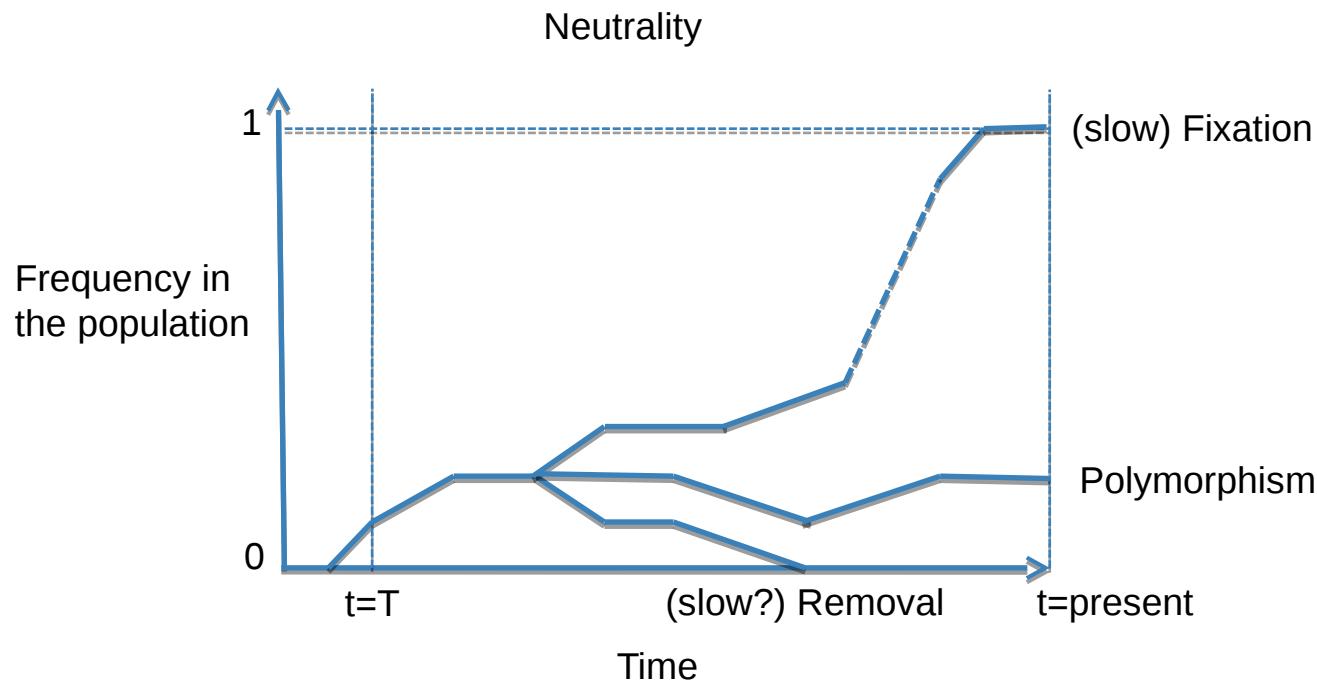
Outline

- Brief introduction to natural selection
- Inferring selection at the intra-species level
 - Genetic differentiation
 - Haplotype variation
 - Model-based approaches
 - Testing for significance
- Inferring selection at the inter-species level
- Detecting selection from low-depth sequencing data
- Experimental design

Natural selection

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

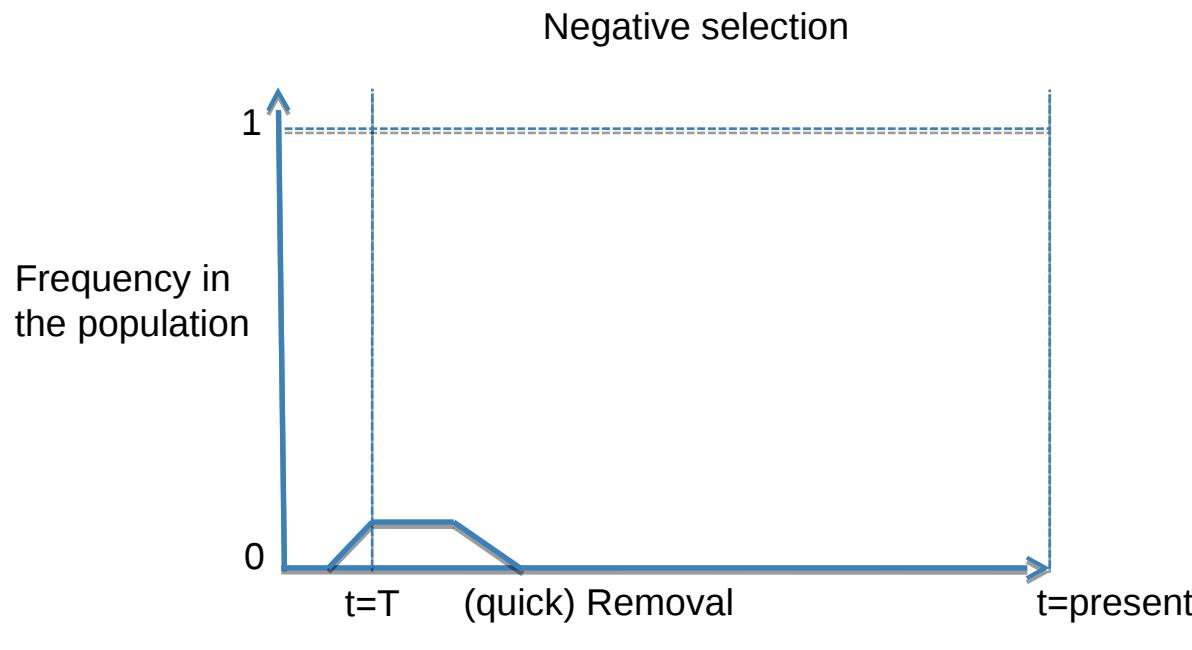
- 1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

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

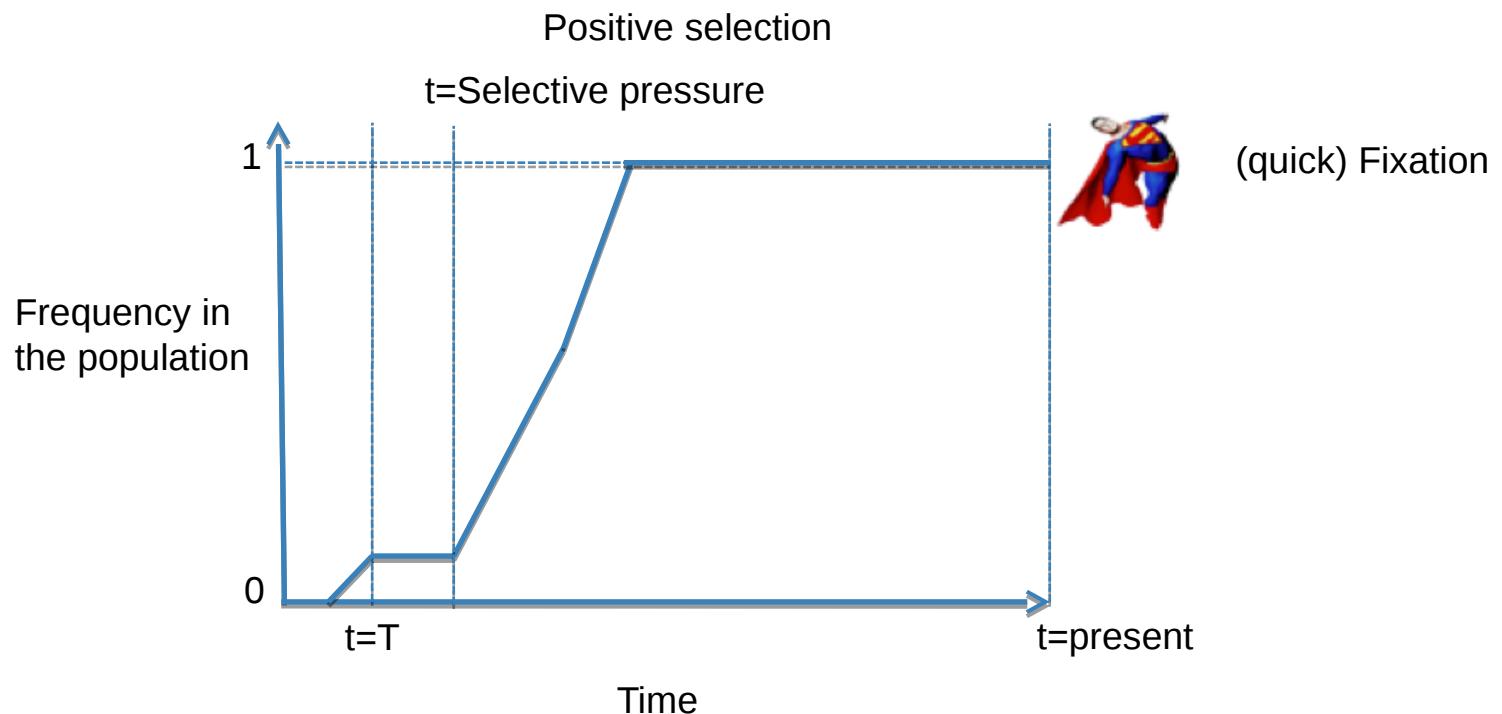
- 1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

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

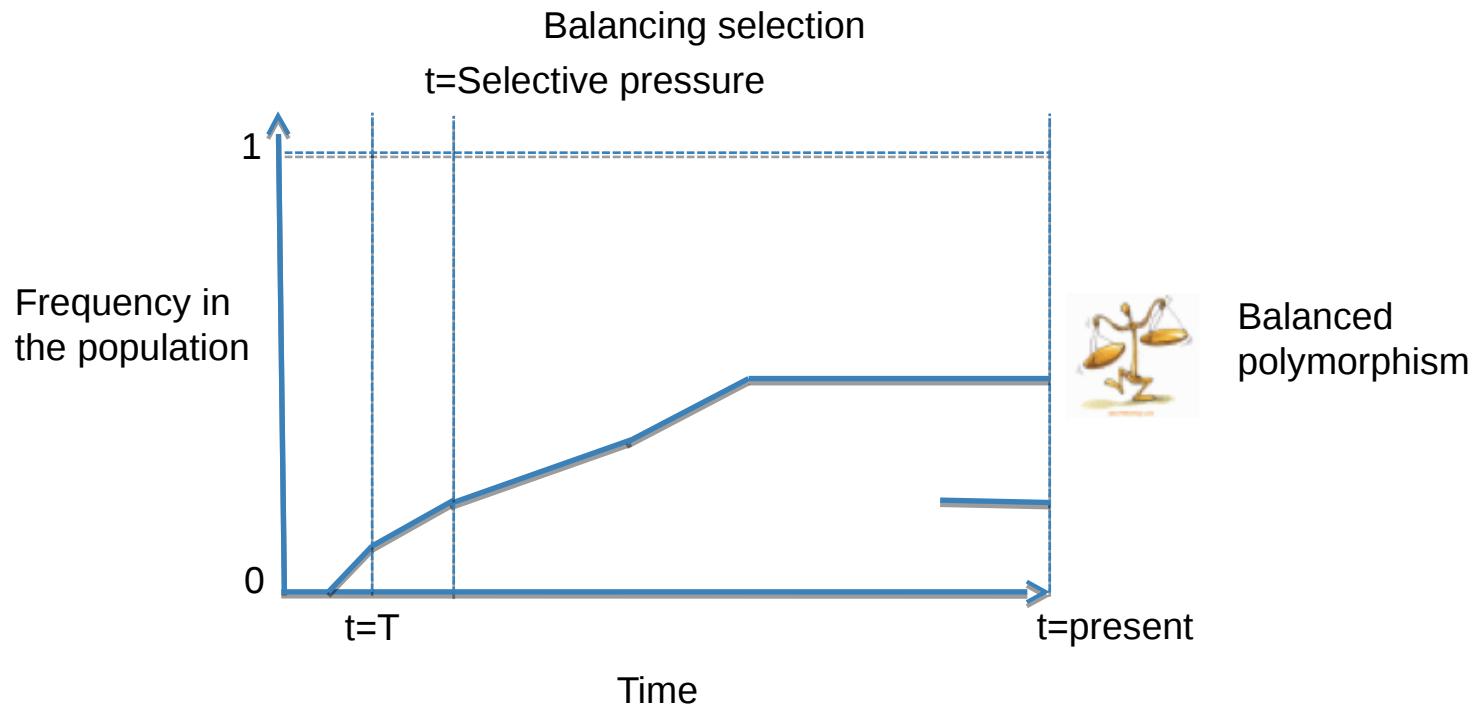
- 1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



Natural selection

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

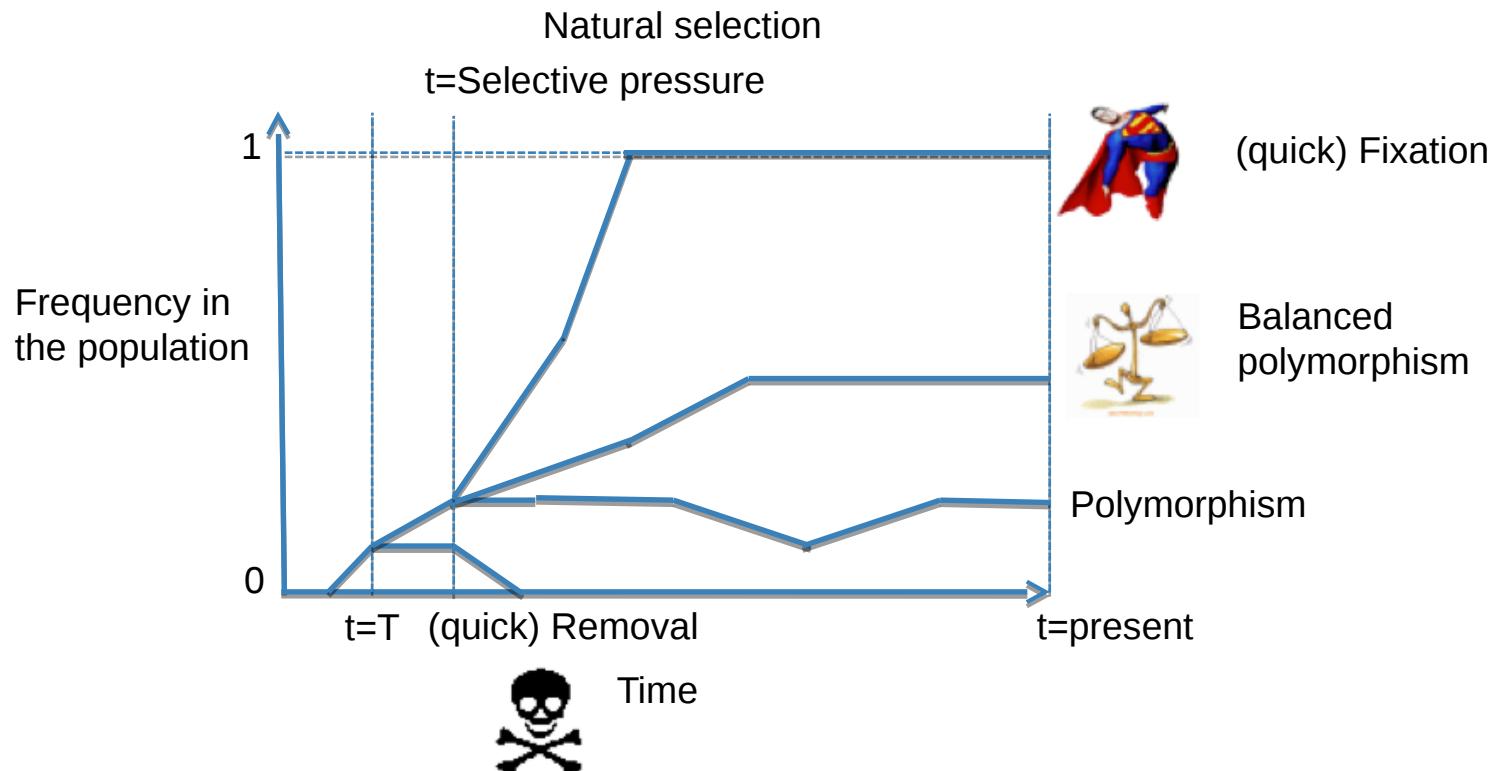
- 1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



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**

Outline

-
-
- Inferring selection at the intra-species level
 - Genetic differentiation
 - Haplotype variation
 - Model-based approaches
 - Testing for significance
-
-

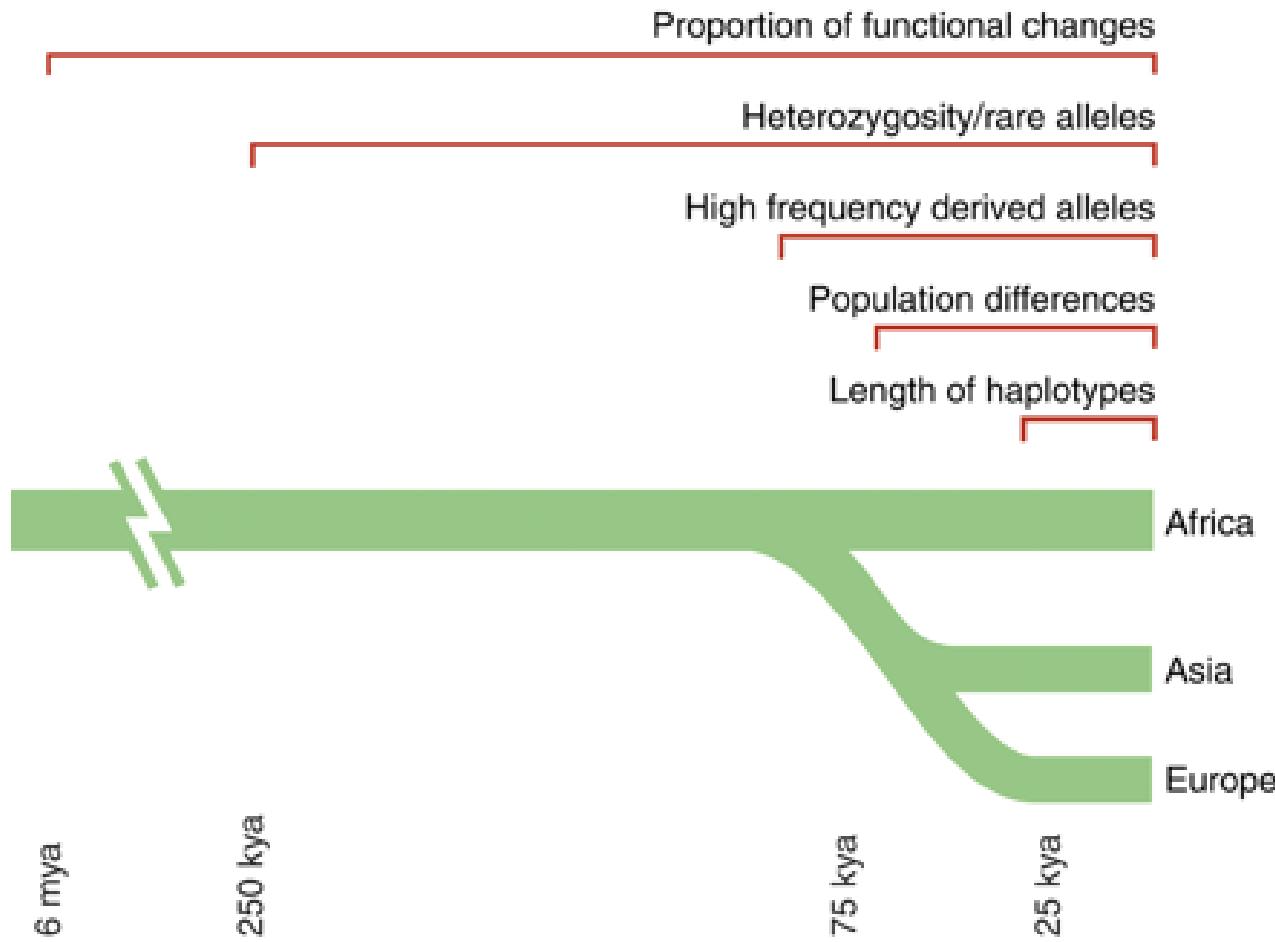
Methods to infer selection

- within-species:

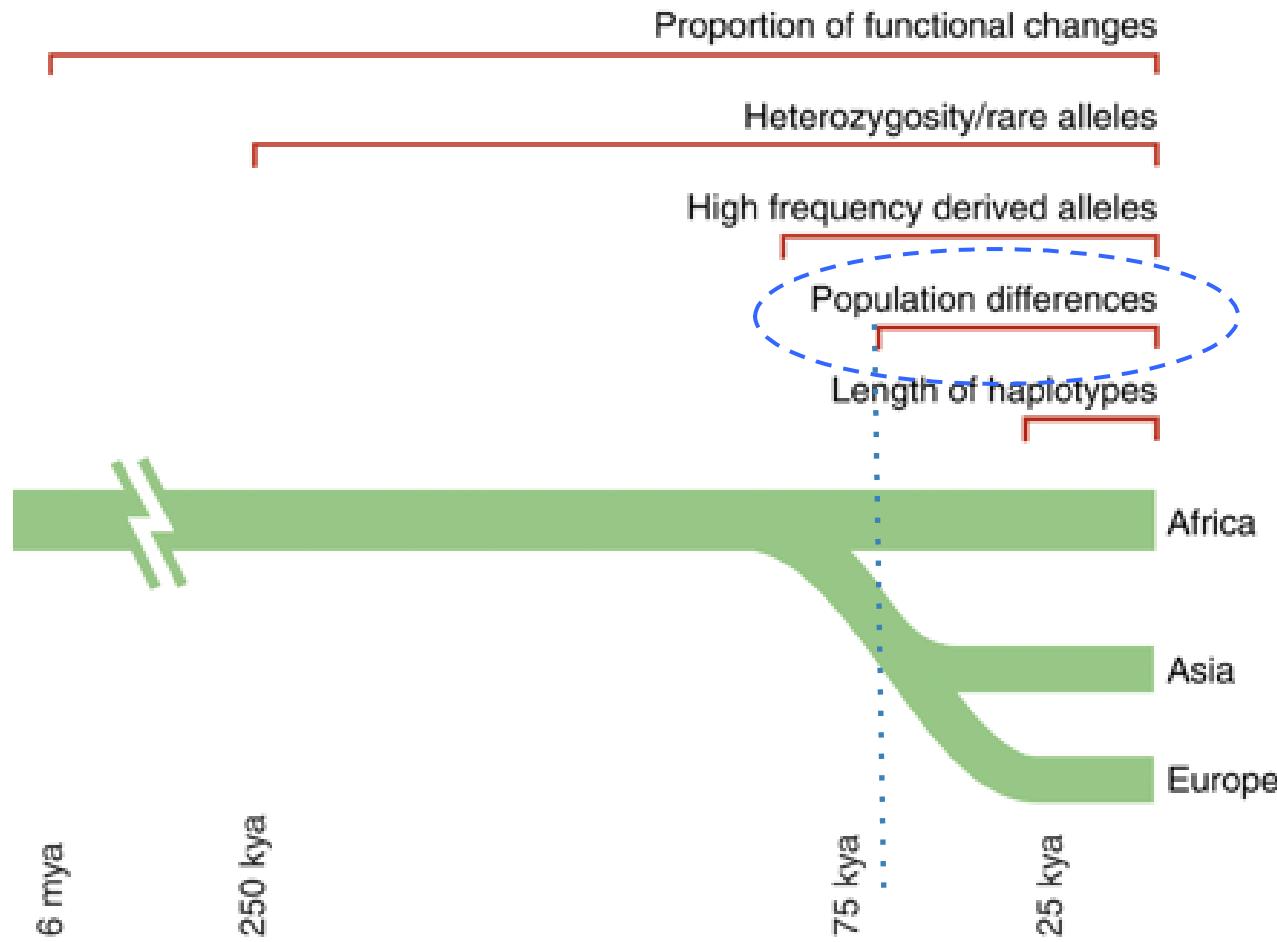
Micro-evolutionary events between populations, local adaptation



Methods to infer recent selection



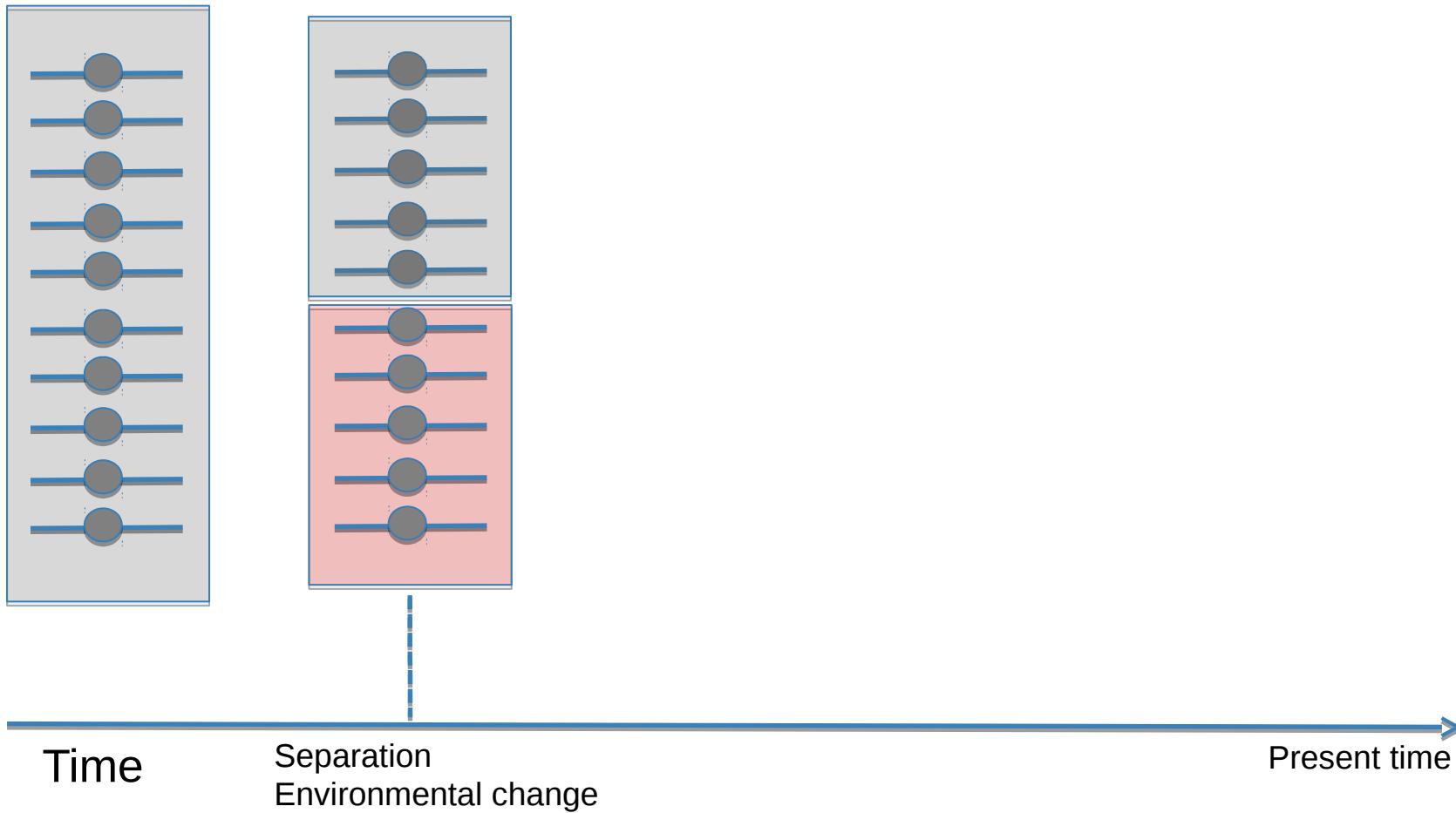
Methods to infer recent selection



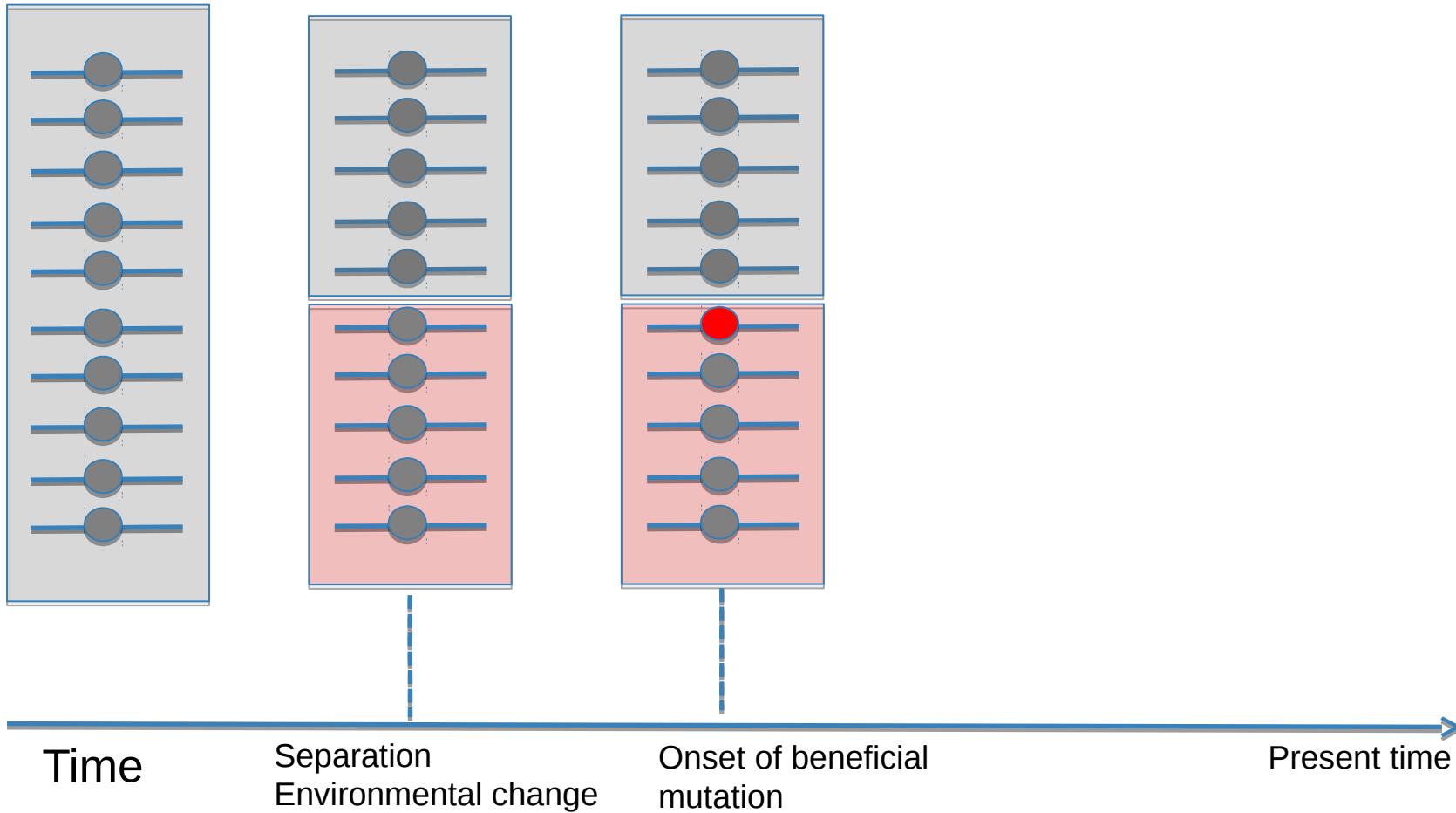
Allele frequency differentiation



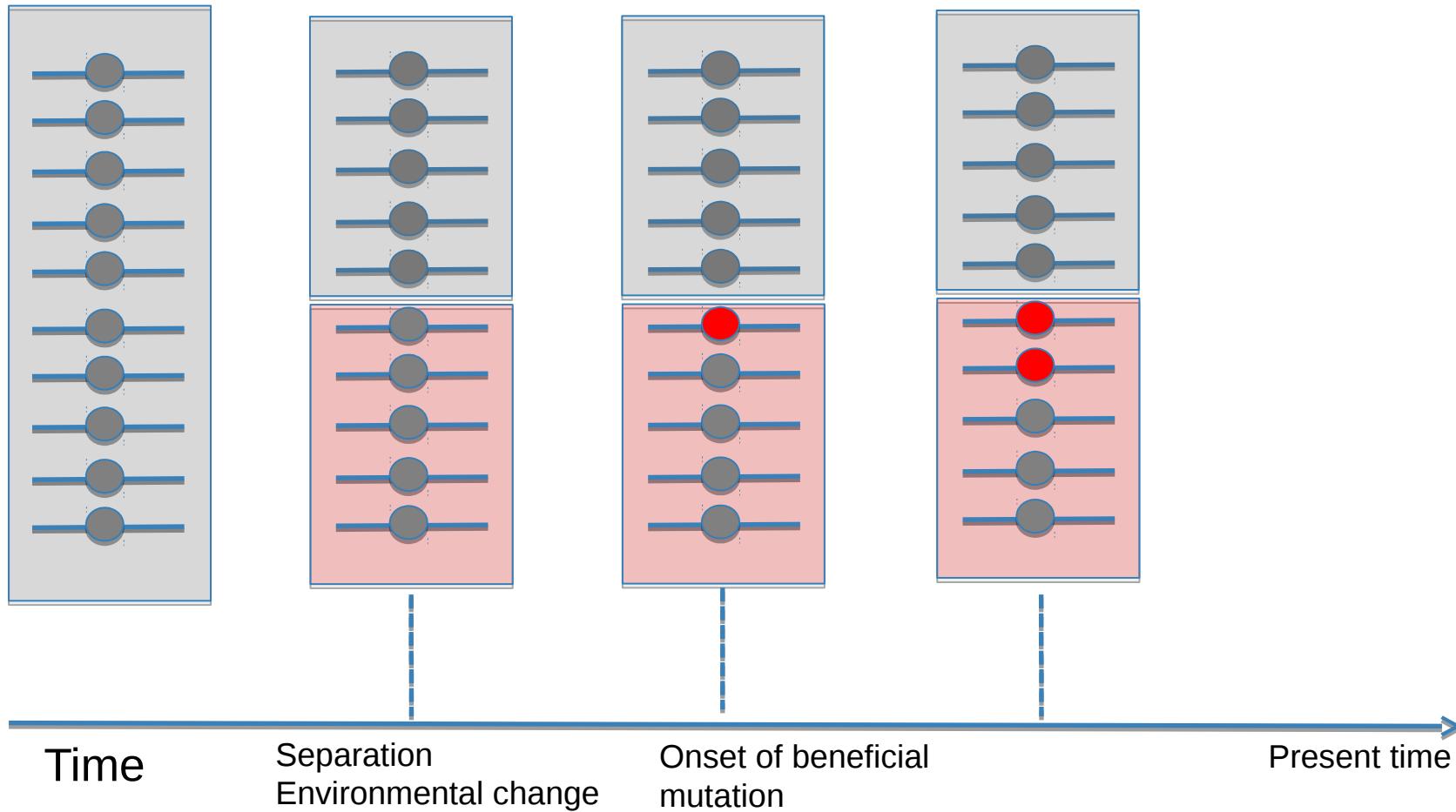
Allele frequency differentiation



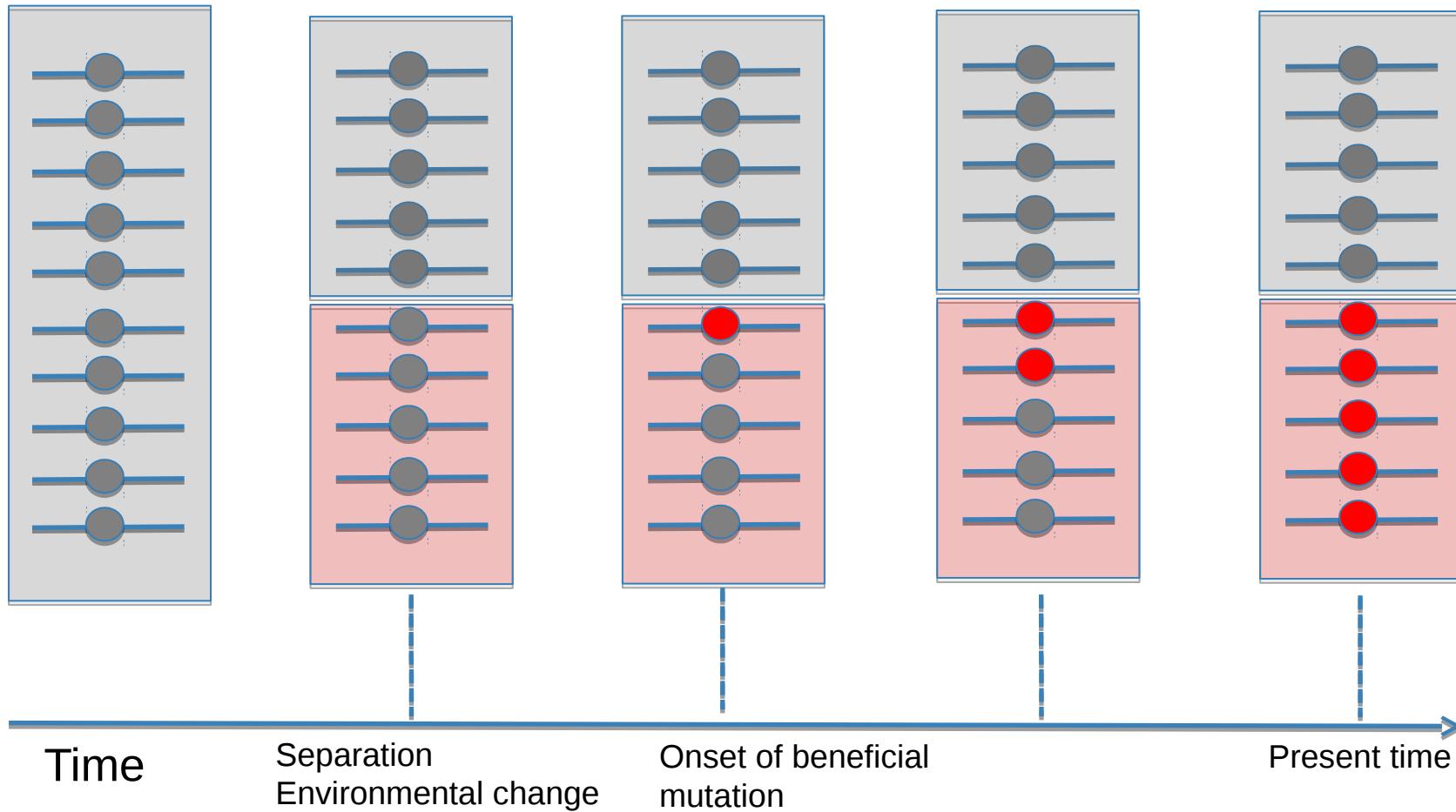
Allele frequency differentiation



Allele frequency differentiation



Allele frequency differentiation



$$F_{ST}$$

Common measure for quantifying population subdivision.

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

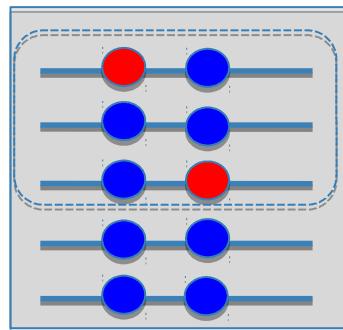
H_B : between populations

H_W : average within populations

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

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations

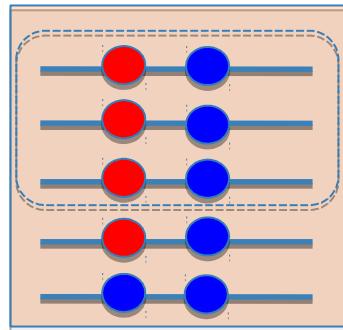


$$F_{ST} = 1 - (H_w / H_B)$$

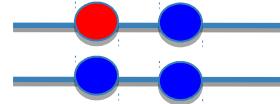
Within populations

Between populations

What is the variation within populations?



e.g. A vs B



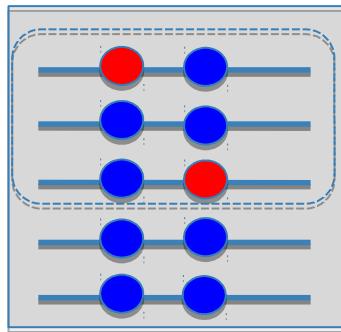
The differ by 1 site

Haplotype-based F_{ST}

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

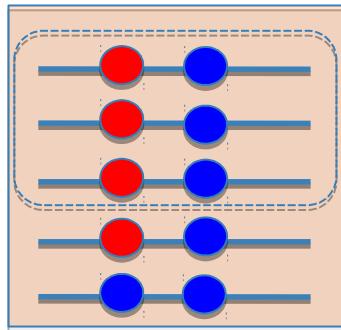
Within populations

Between populations



What is the variation within populations?

A	B	Mean=?
A	C	
B	C	
D	E	Mean=?
D	F	
E	F	



H_W is the average within-populations: ?

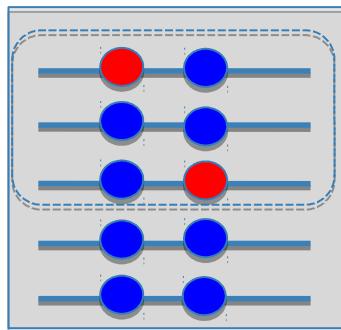
Haplotype-based F_{ST}

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

Within populations

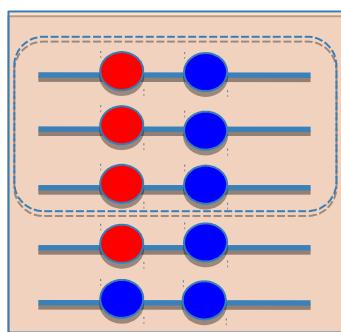
Between populations

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

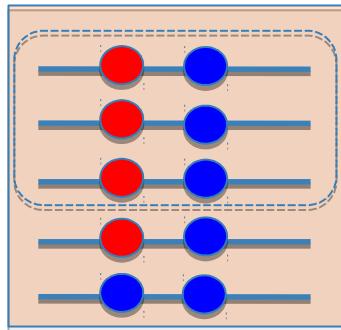
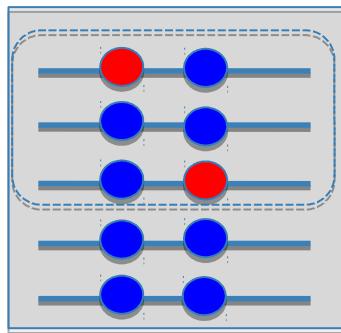
H_W is the average within-populations: $(4/3+0/3)/2=2/3$

Haplotype-based F_{ST}

$$F_{ST} = 1 - (\bar{H}_W / \bar{H}_B)$$

Within populations

Between populations



What is the variation between populations?

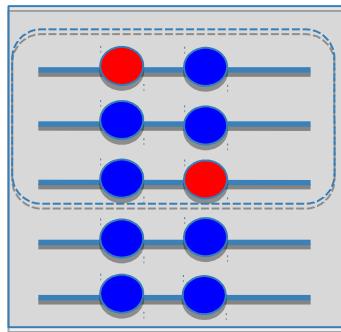
A	D	0
A	E	0
A	F	0
B	D	1
B	E	1
B	F	1
C	D	2
C	E	2
C	F	2

Mean=9/9

\bar{H}_B is the average between-populations: $9/9=1$

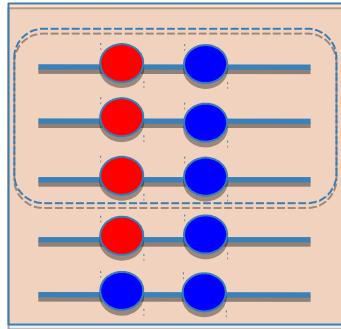
Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations



A
B
C

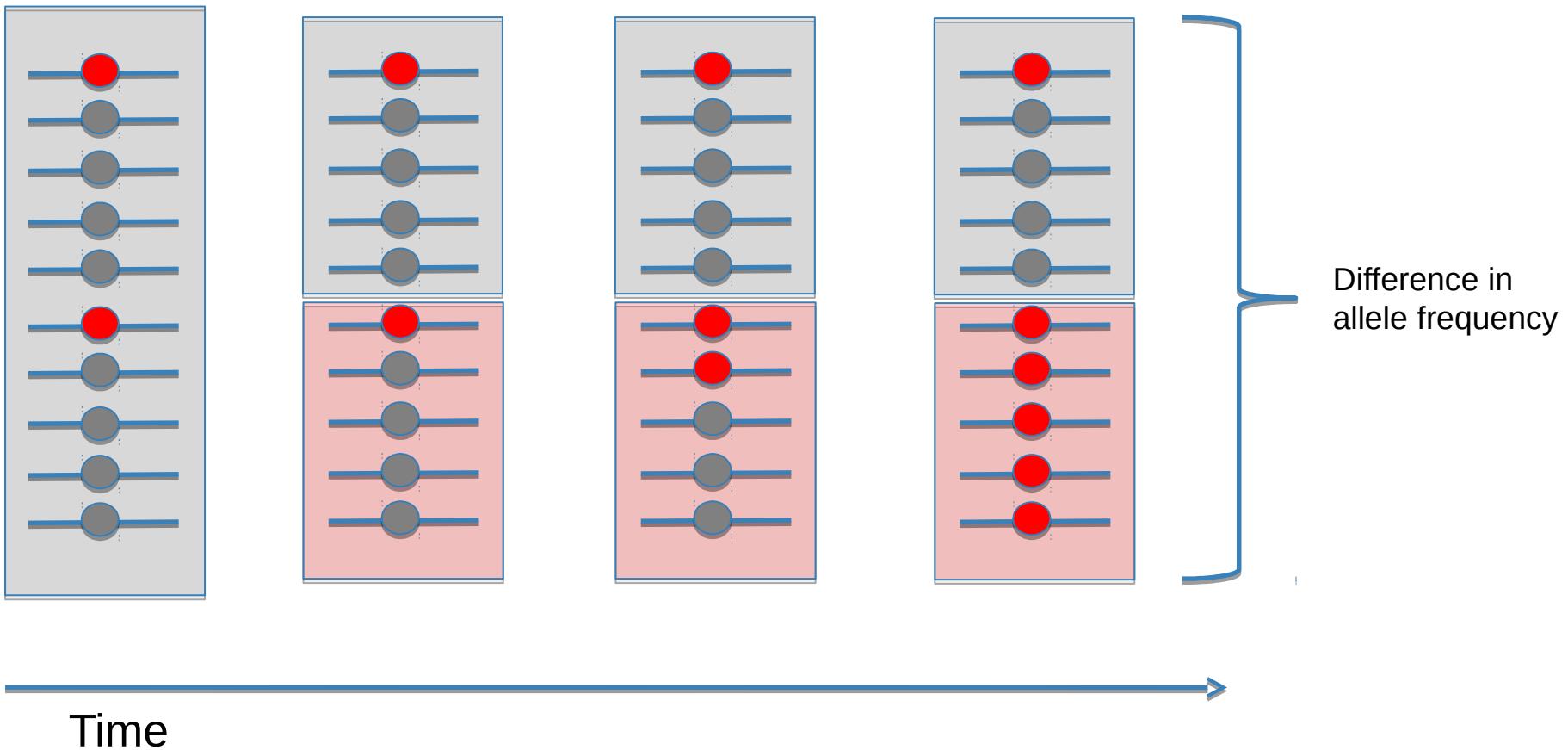
$$F_{ST} = 1 - (H_W / H_B) = 1 - ((2/3)/1) = 1/3 \sim 0.33$$



D
E
F

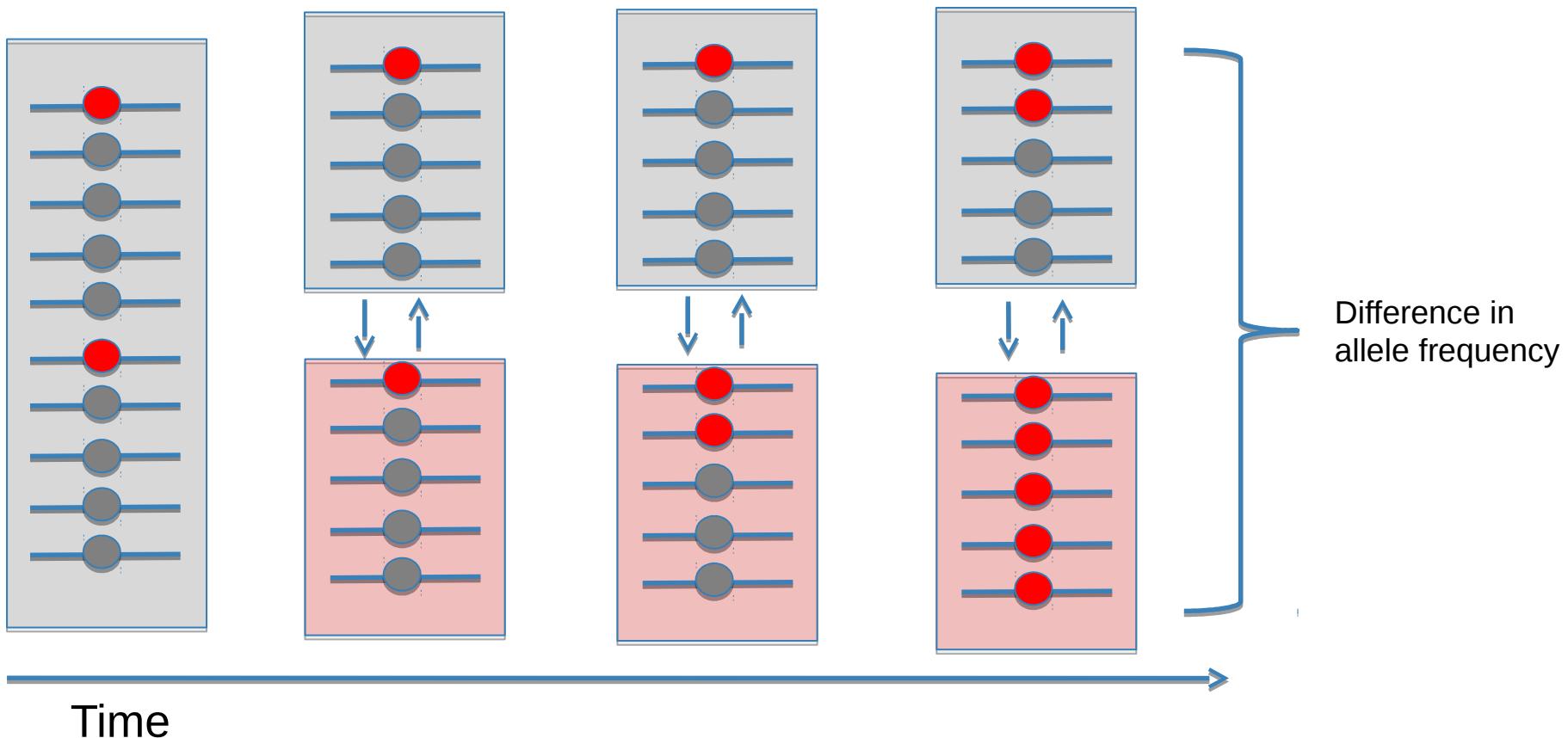
Allele frequency differentiation

From standing variation



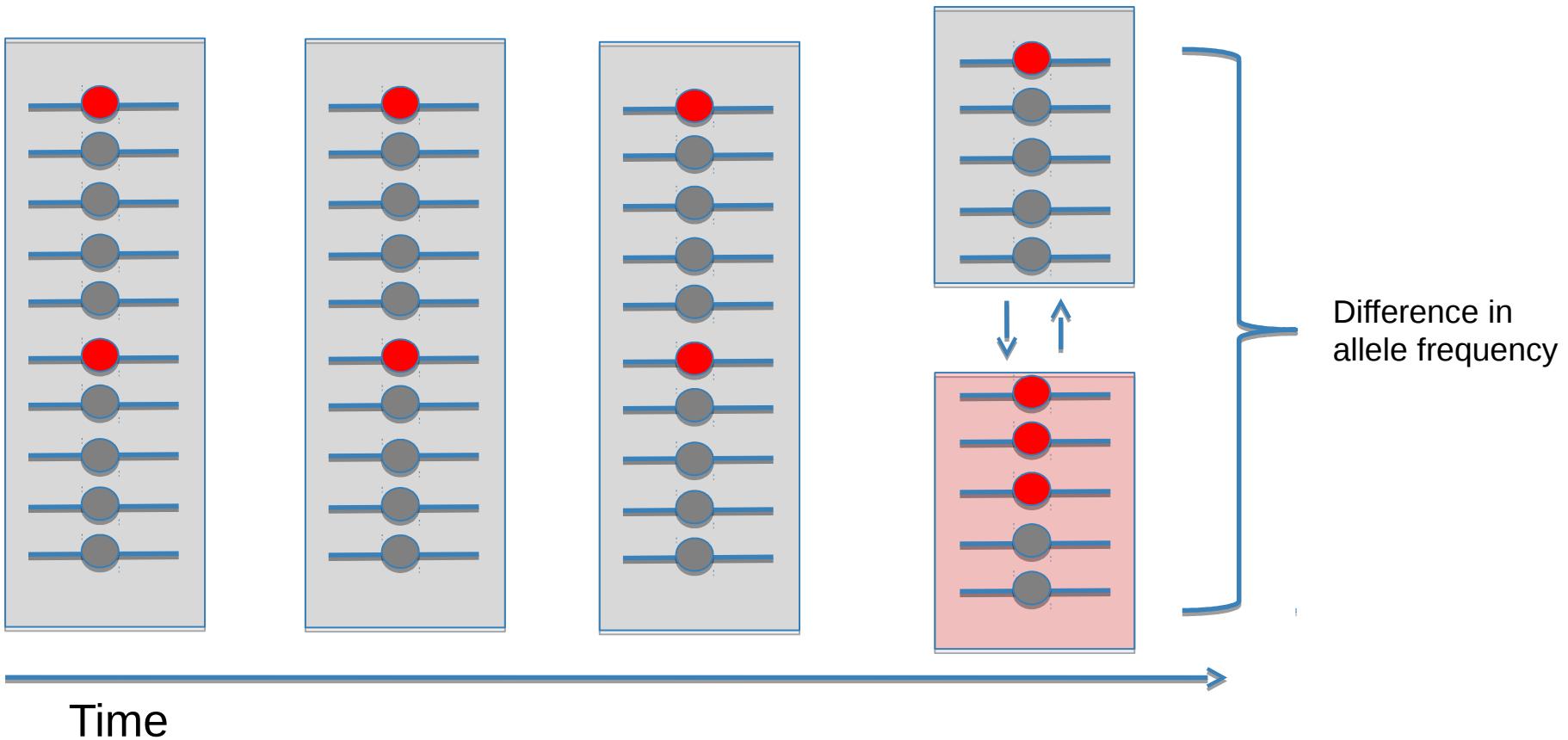
Allele frequency differentiation

With migration



Allele frequency differentiation

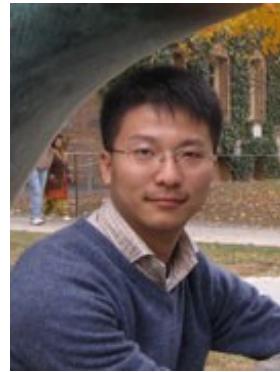
With recent divergence



Population genetic differentiation



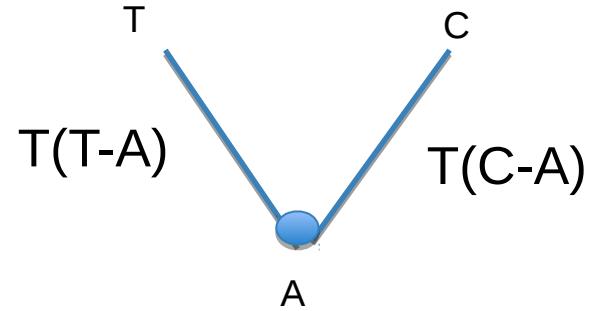
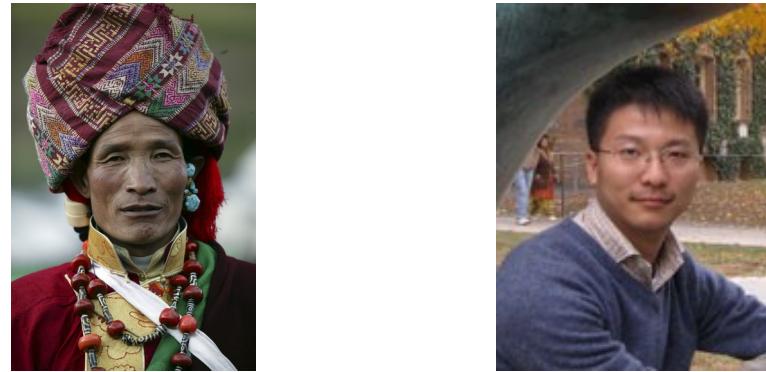
T



C

$$F_{ST}(T-C)$$

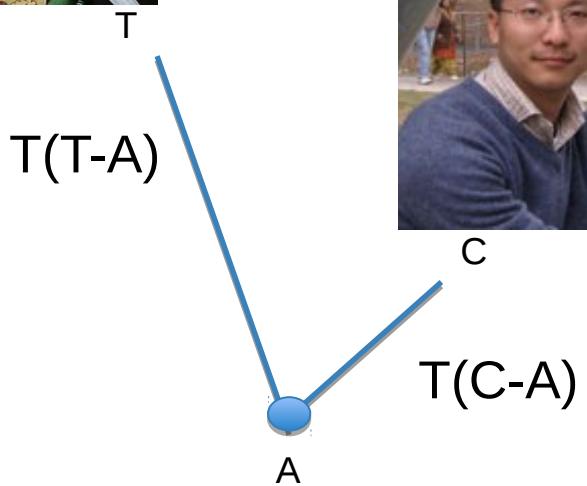
Population genetic differentiation



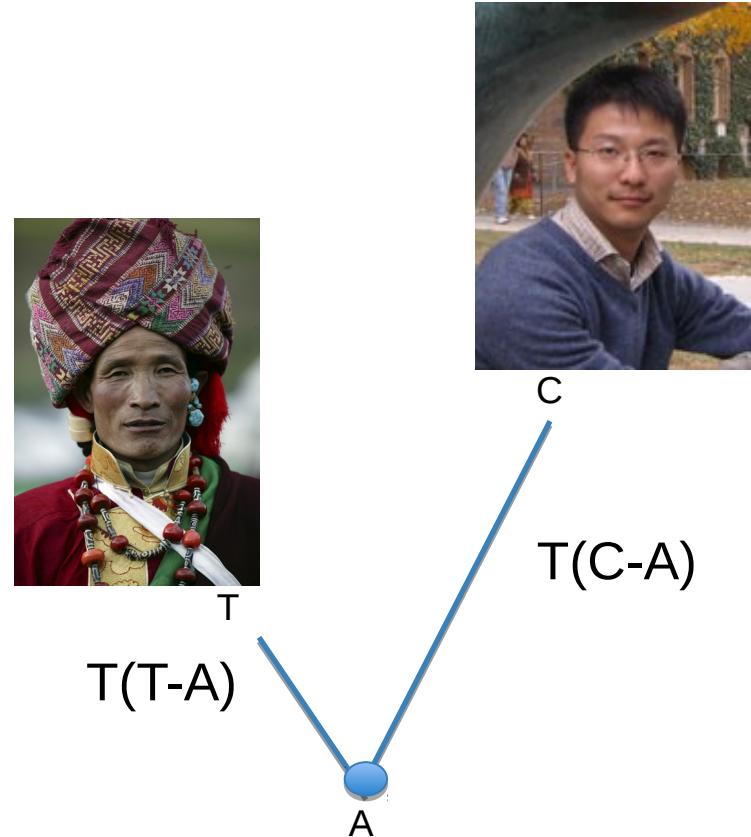
$$F_{ST}(T-C) \sim T(T-A-C)$$

Population genetic differentiation

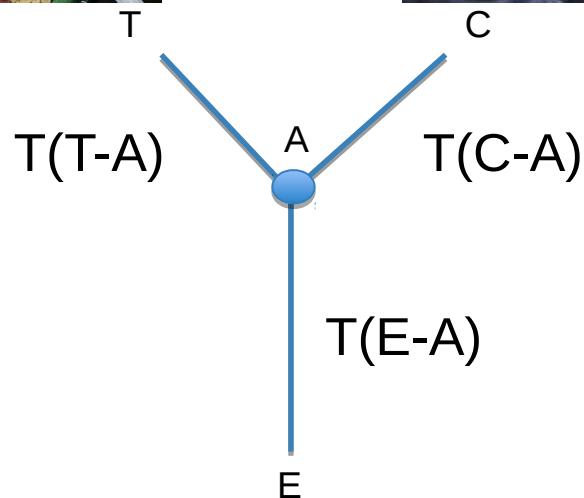
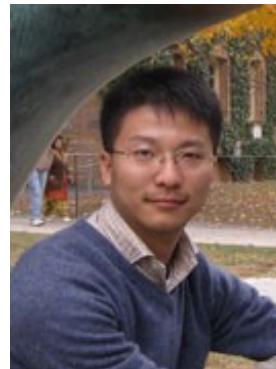
$$F_{ST}(T-C) \sim T(T-A-C)$$



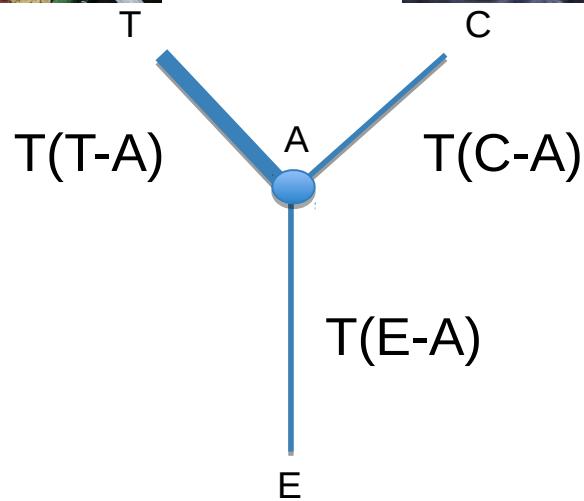
?



Population genetic differentiation



Population genetic differentiation

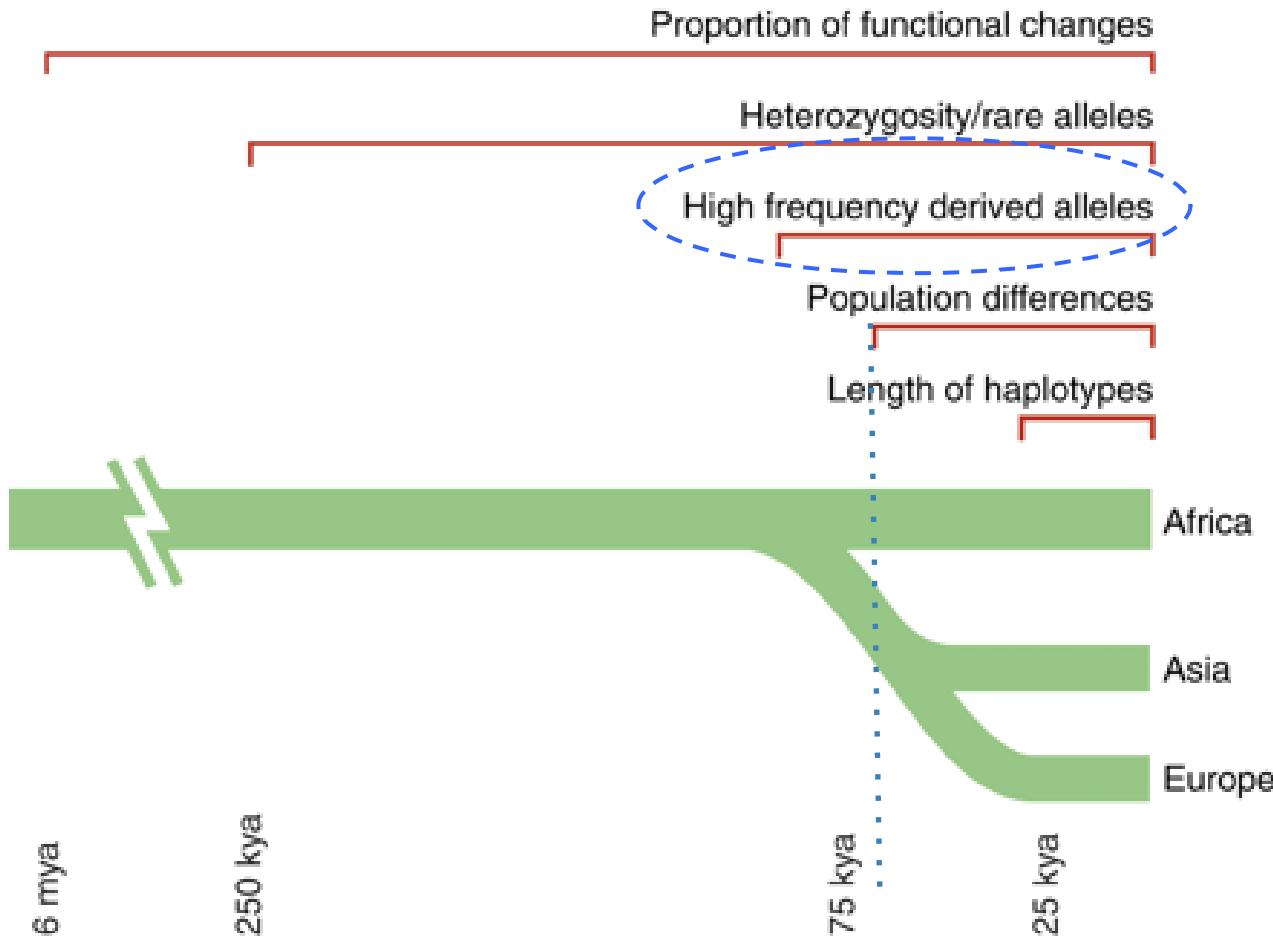


$$T(T-A-C) = -\log(1 - F_{ST}(T-C))$$

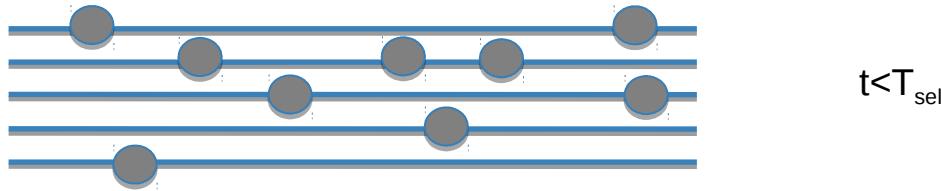
T(T-A)?



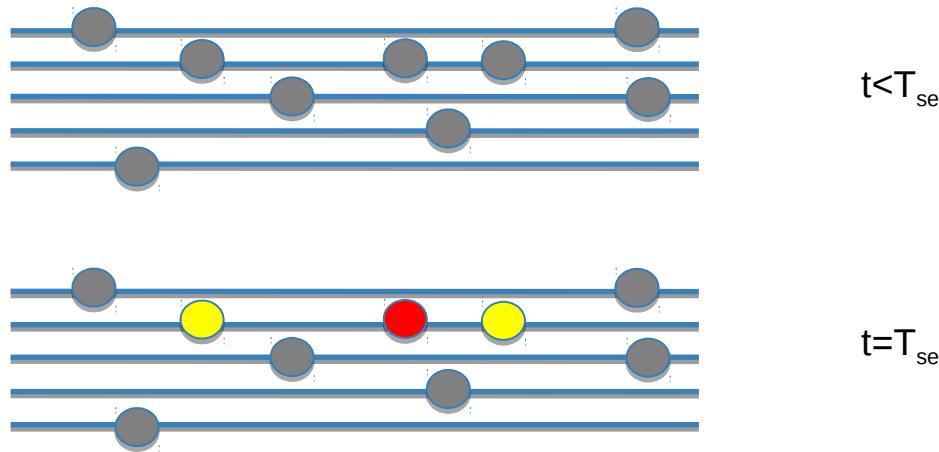
Methods to infer selection



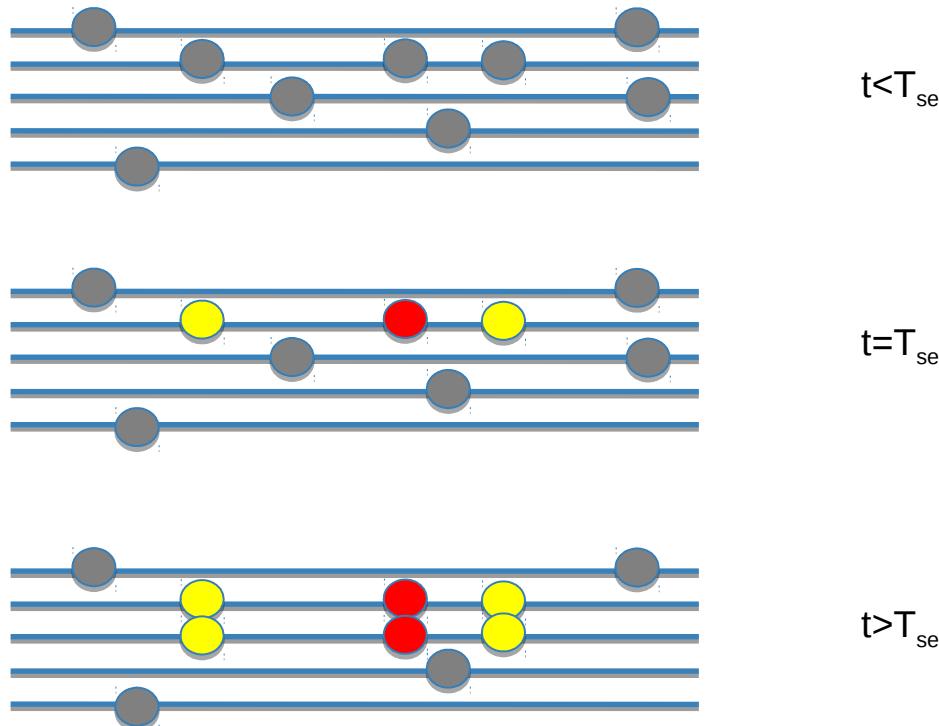
Positive selection: effect on haplotypes



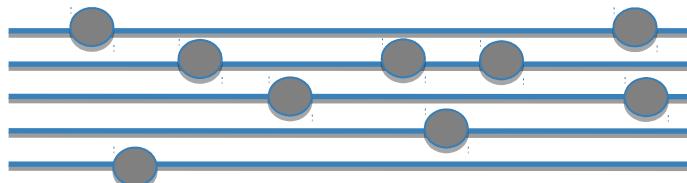
Positive selection: effect on haplotypes



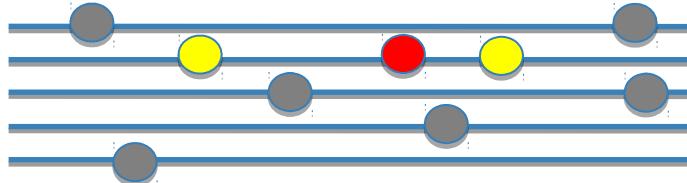
Positive selection: effect on haplotypes



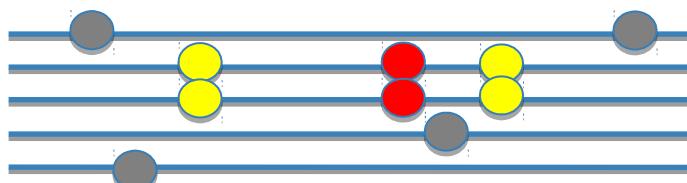
Positive selection: effect on haplotypes



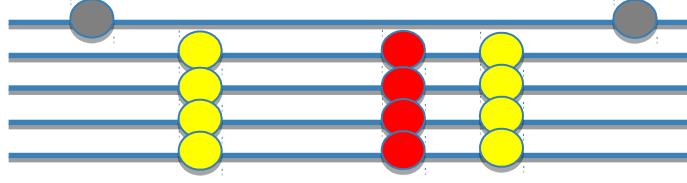
$t < T_{sel}$



$t = T_{sel}$



$t > T_{sel}$



$t \gg T_{sel}$

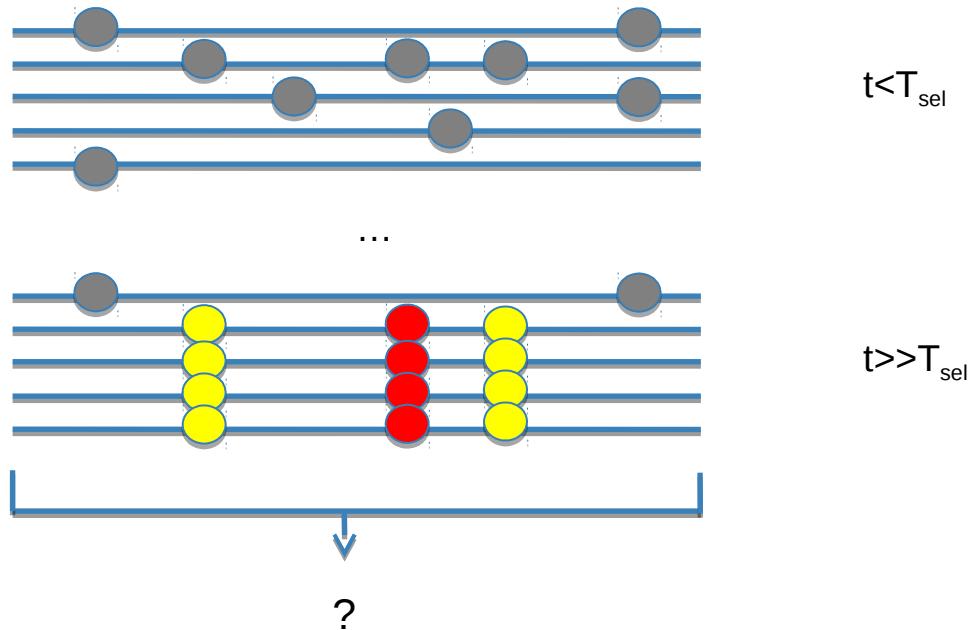
Selective sweep



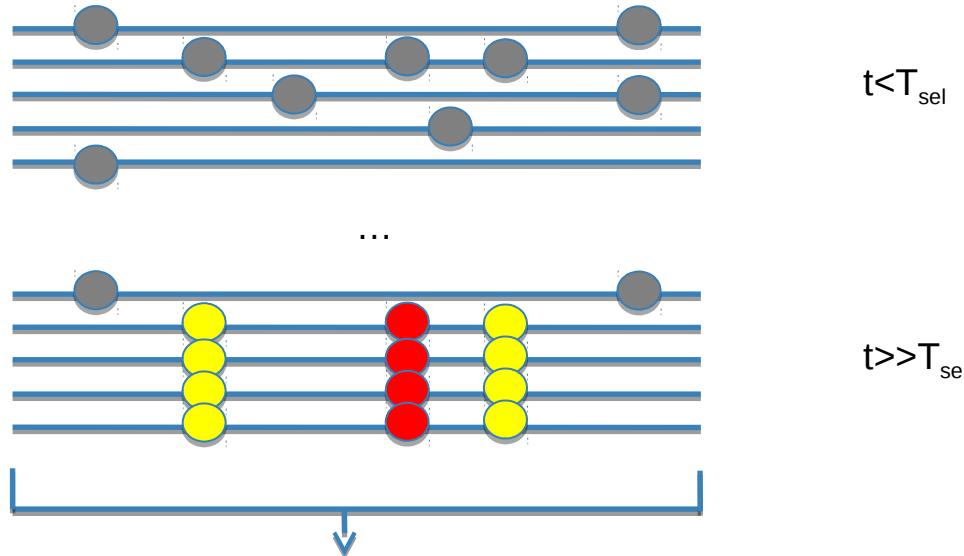
Genetic hitch-hiking



Positive selection

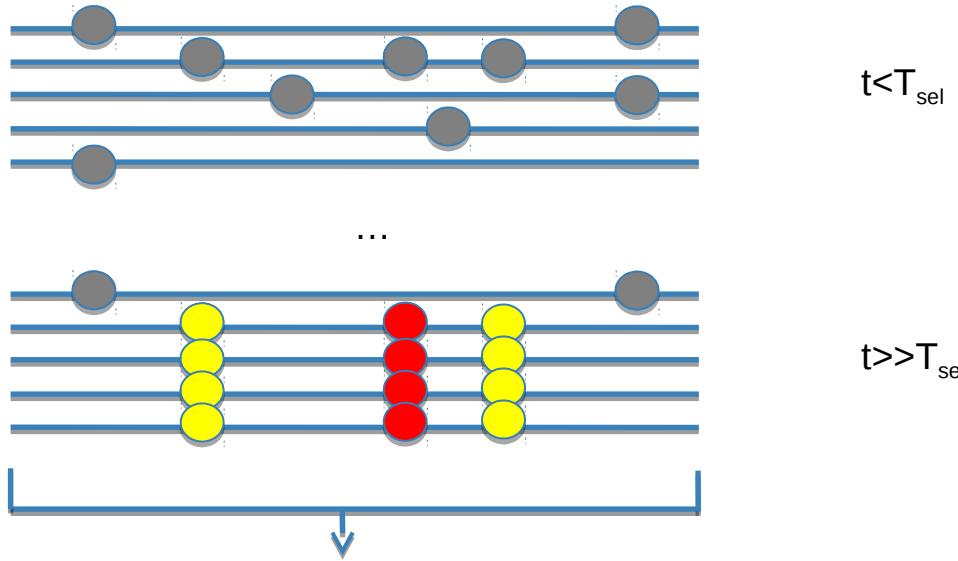


Positive selection



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

Positive selection



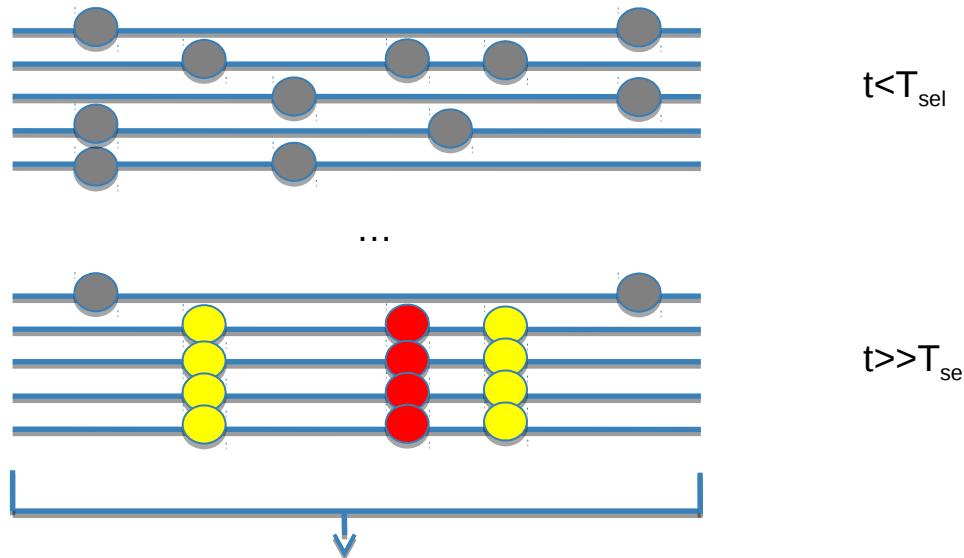
- Reduction of polymorphisms levels
(e.g. from 7 to 5 SNPs)

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

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

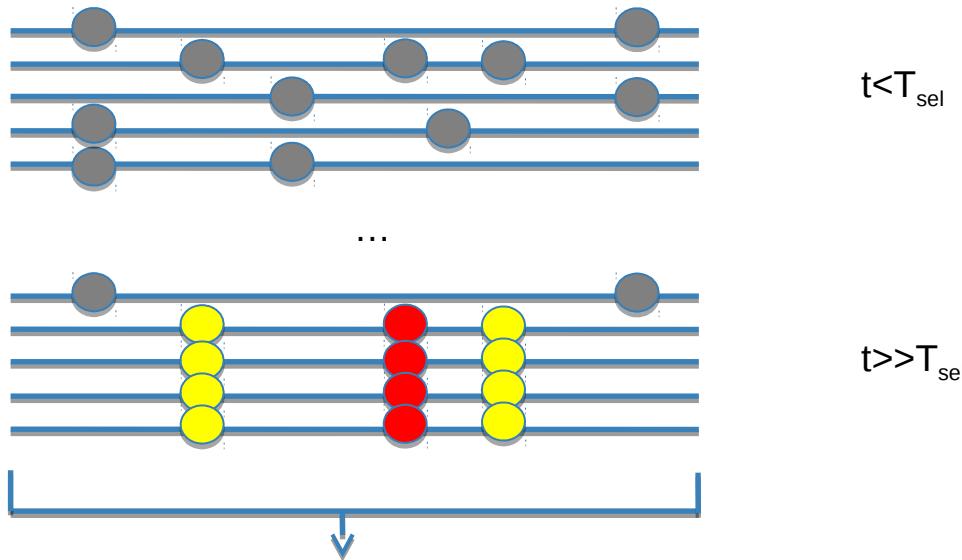
$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Positive selection



- Reduction of polymorphisms levels (Theta)
- ?

Positive selection

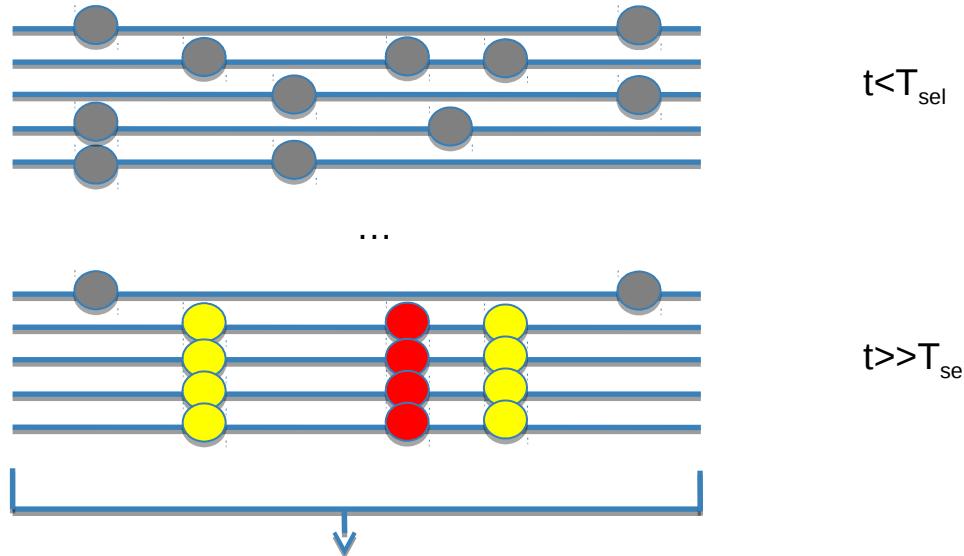


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

Nucleotide diversity index: average pairwise nucleotide differences (π) 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



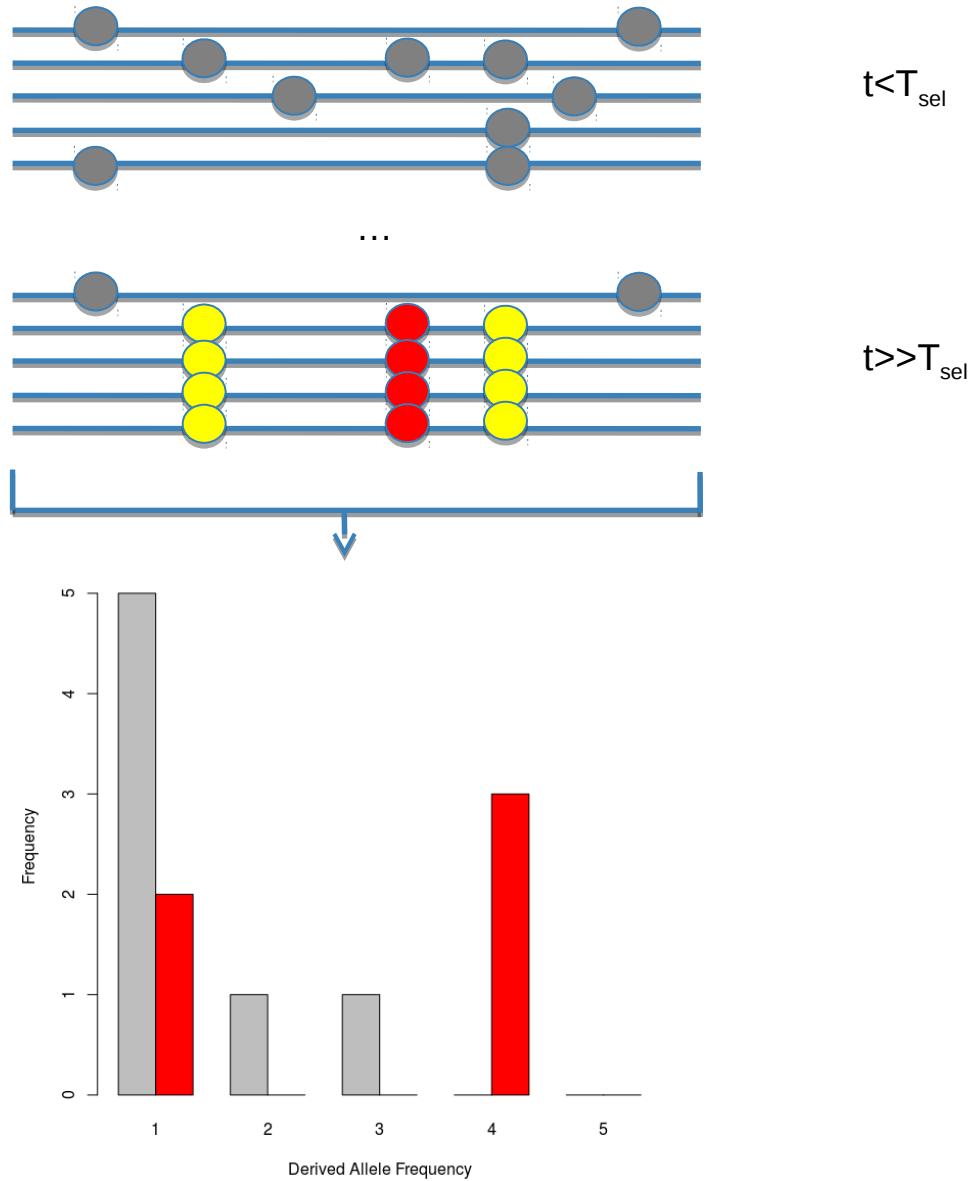
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

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

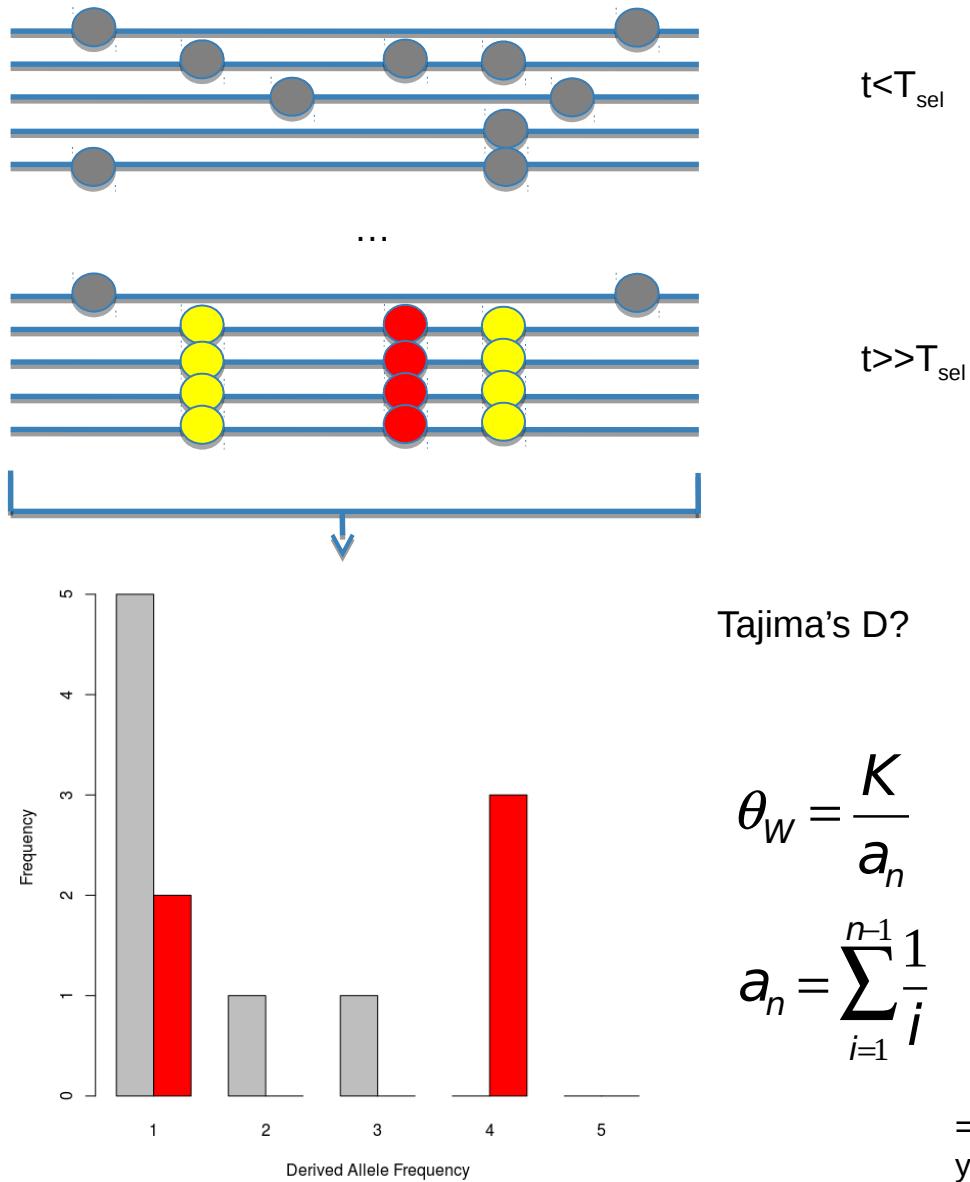
$$D = \frac{\pi - \theta_w}{\sqrt{\hat{V}(\pi - \theta_w)}}$$

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

The Site Frequency Spectrum



The Site Frequency Spectrum



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

$\sim 0.33 = 1/3$

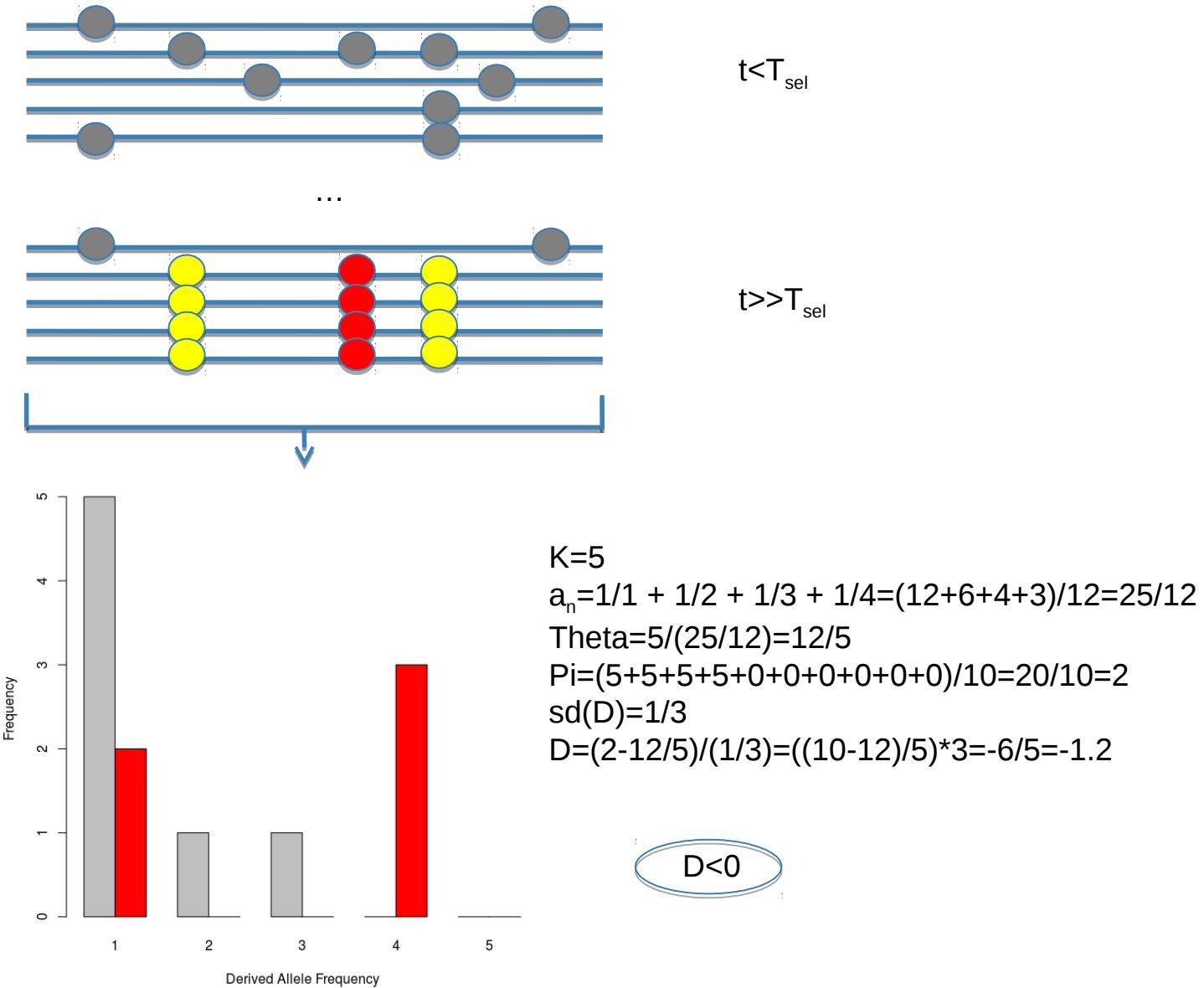
$$\theta_W = \frac{K}{a_n}$$

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

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

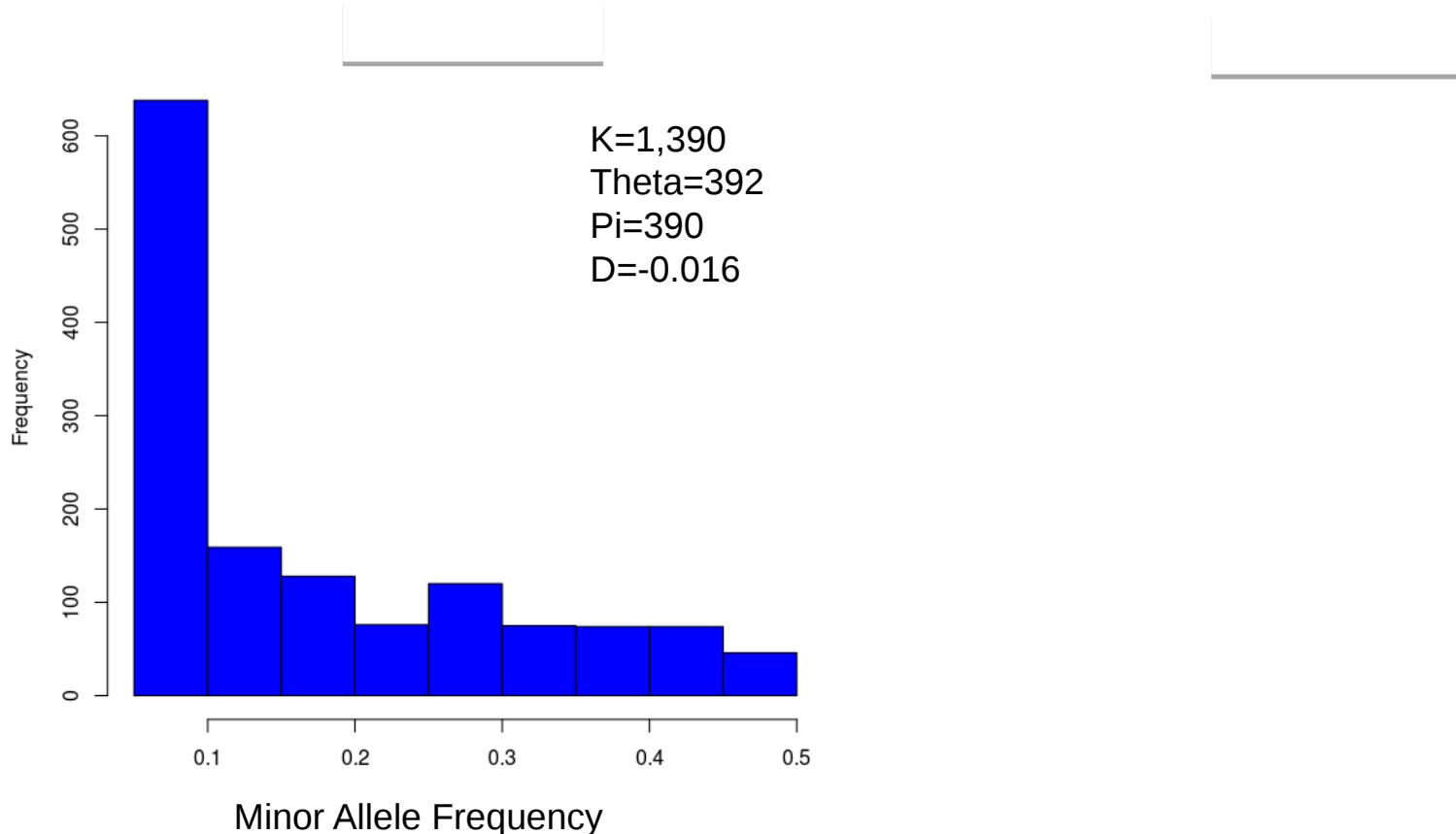
$= 10$, the number of comparisons you need to make

The importance of being... The Site Frequency Spectrum



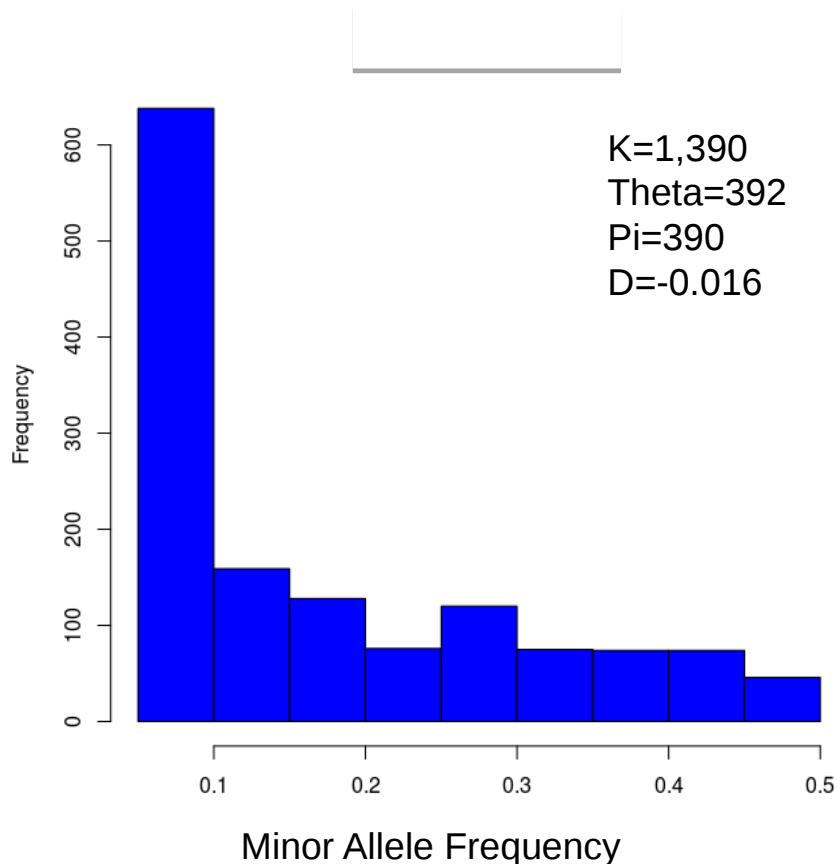
Demography matters?

n=20; L=500kbp; no selection

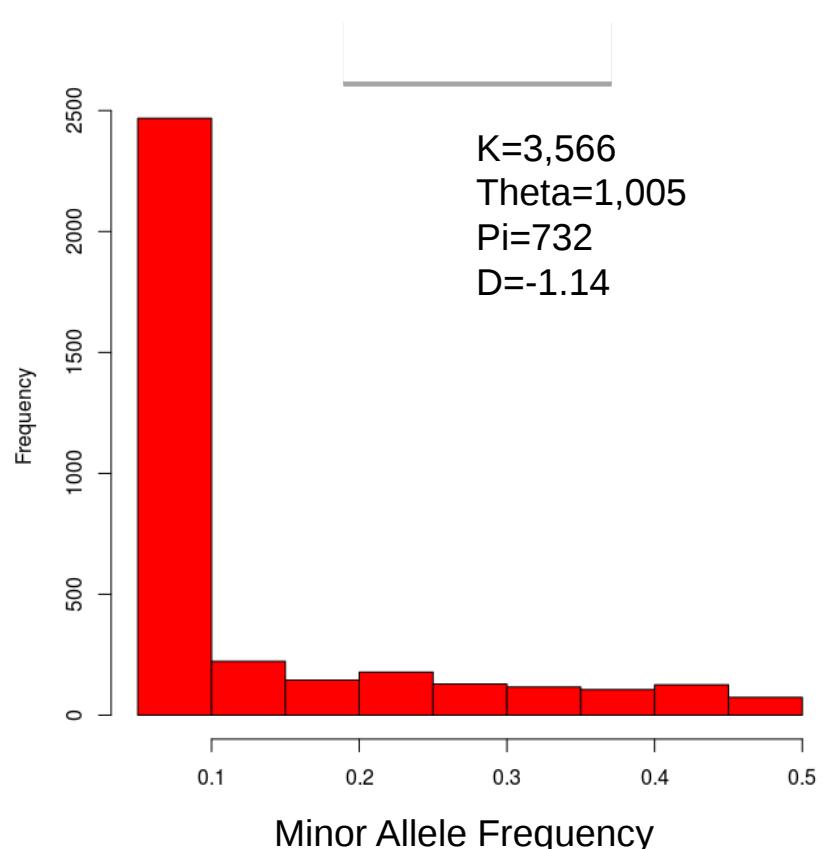


Demography matters?

n=20; L=500kbp; no selection

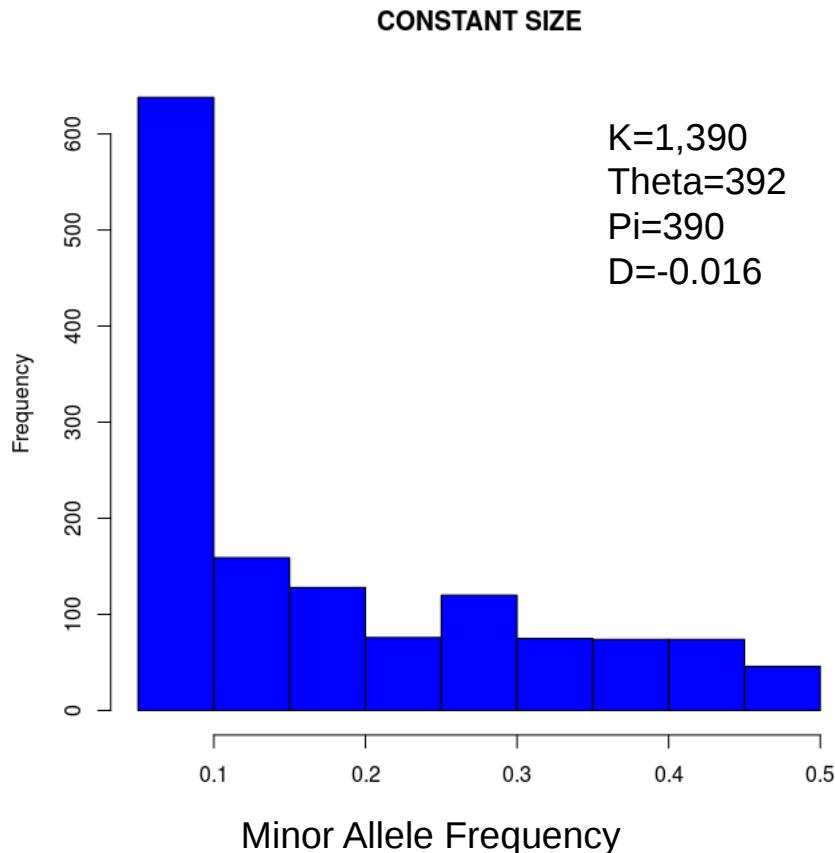


n=20; L=500kbp; no selection

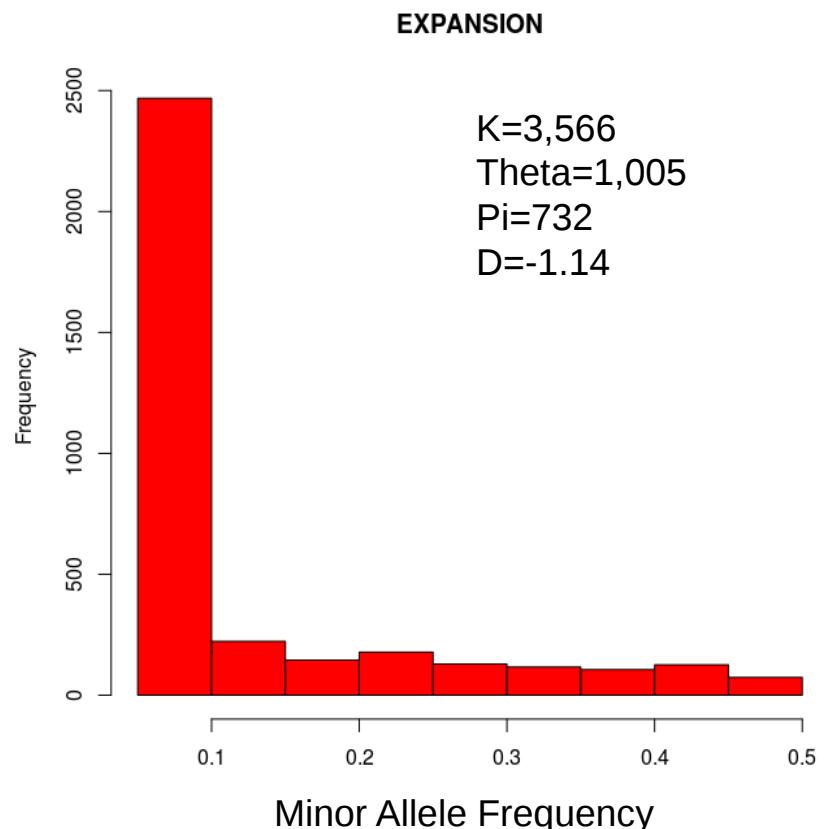


Demography matters!

n=20; L=500kbp; no selection

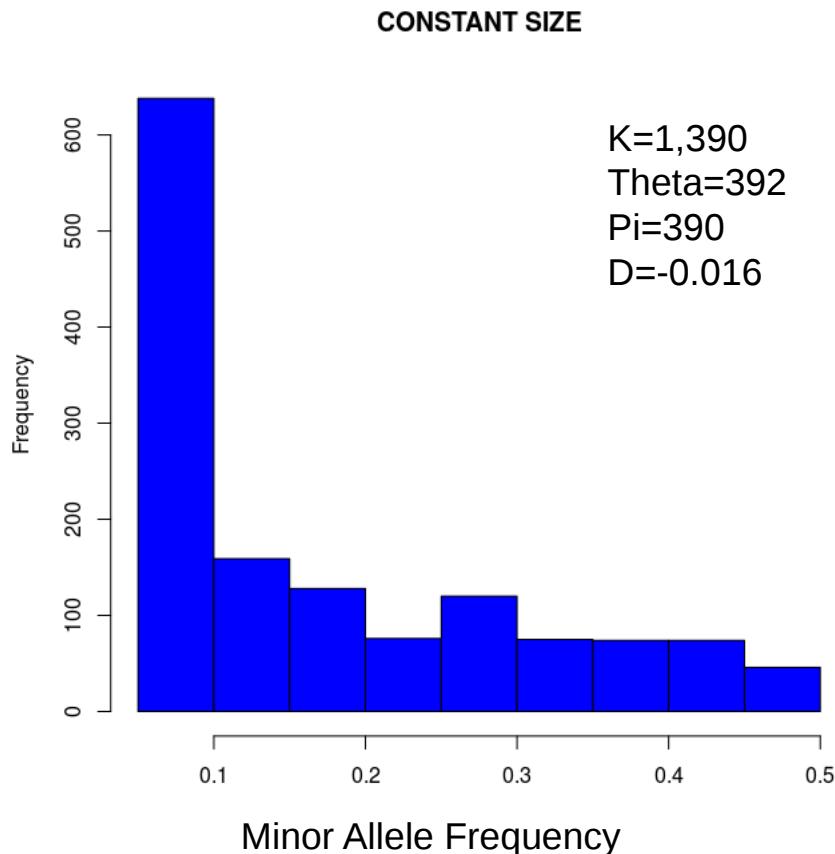


n=20; L=500kbp; no selection

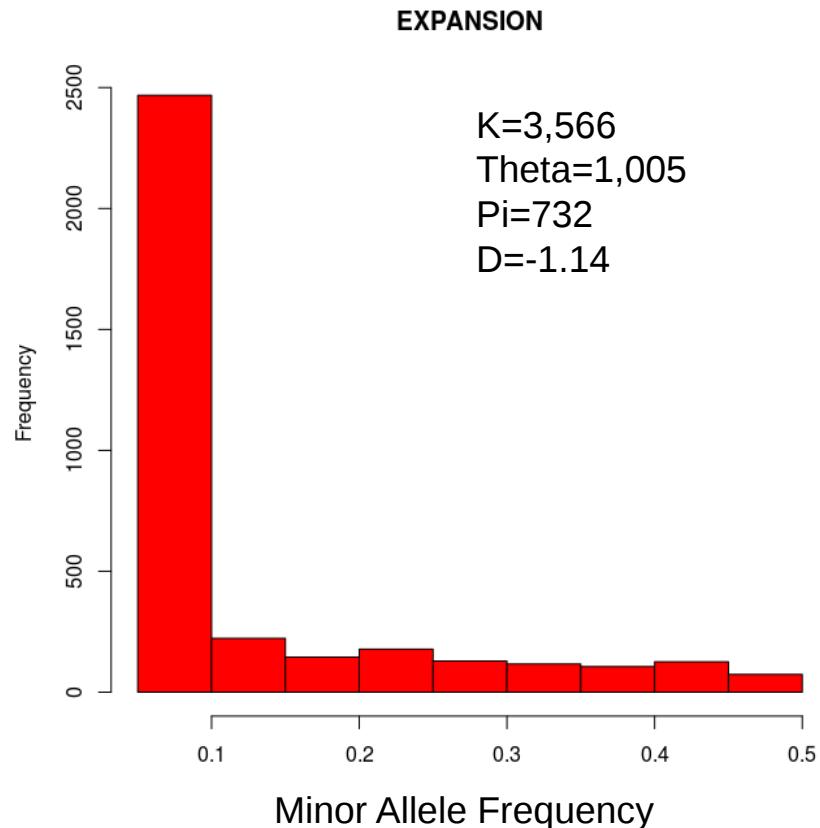


Demography matters!

n=20; L=500kbp; no selection



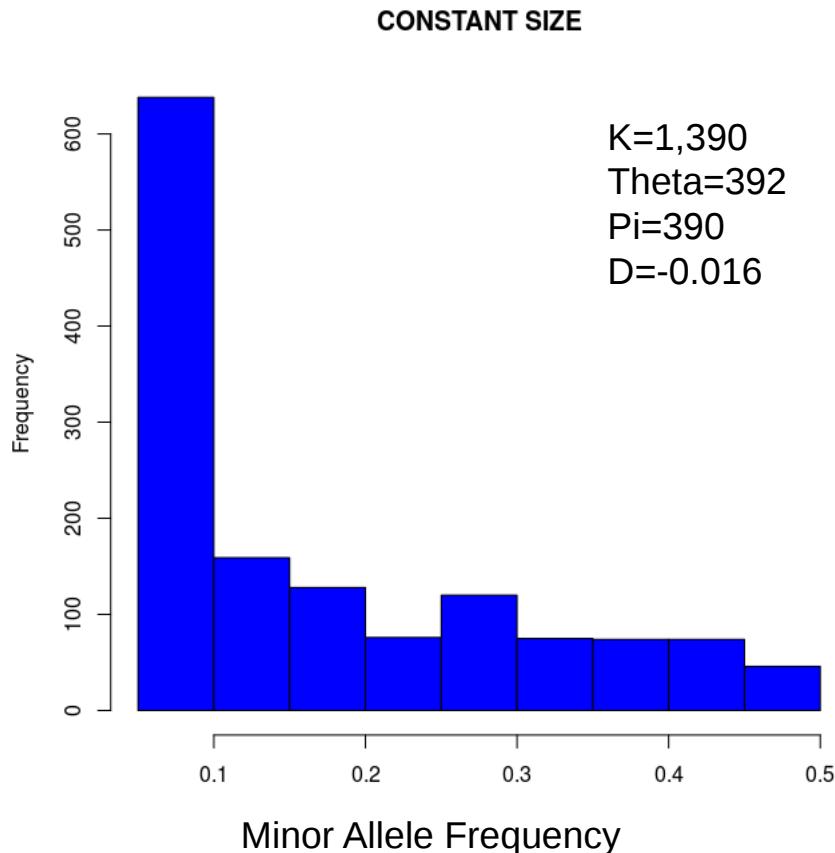
n=20; L=500kbp; no selection



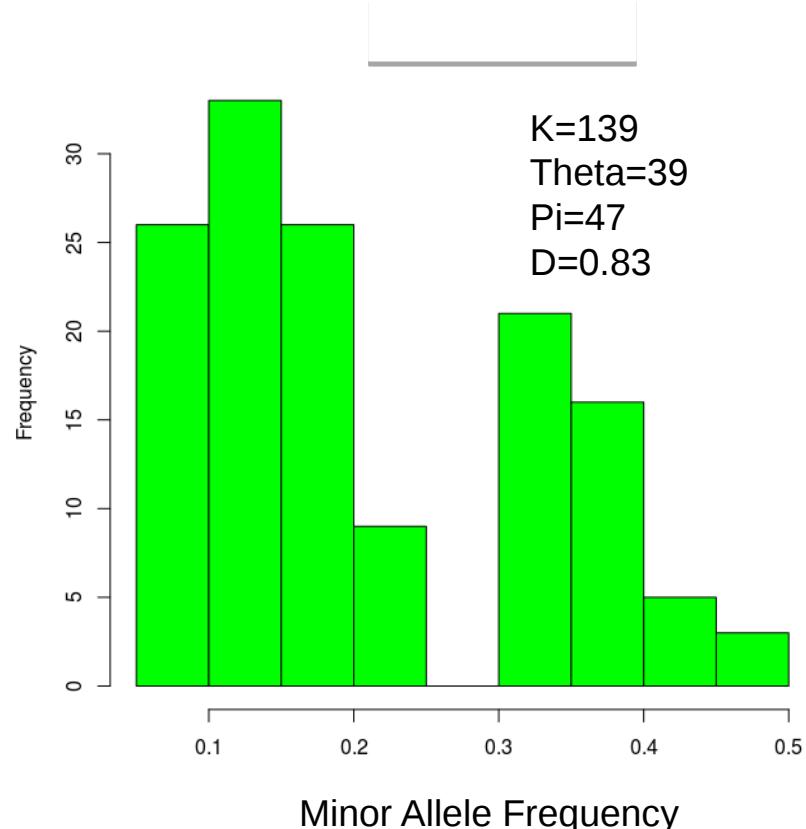
- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

Demography matters?

n=20; L=500kbp; no selection

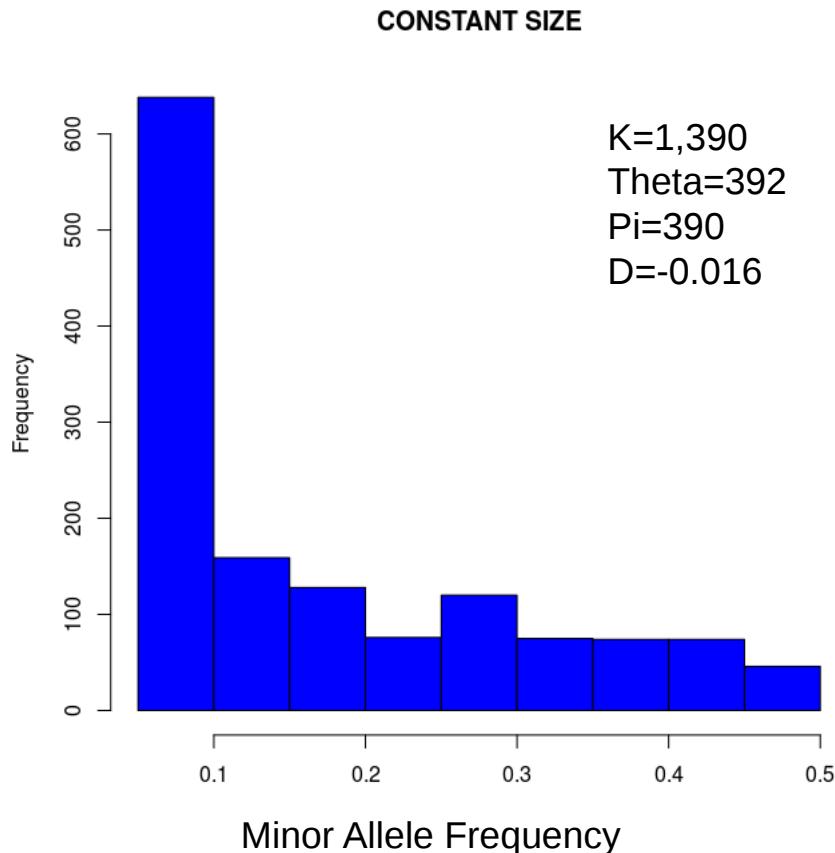


n=20; L=500kbp; no selection

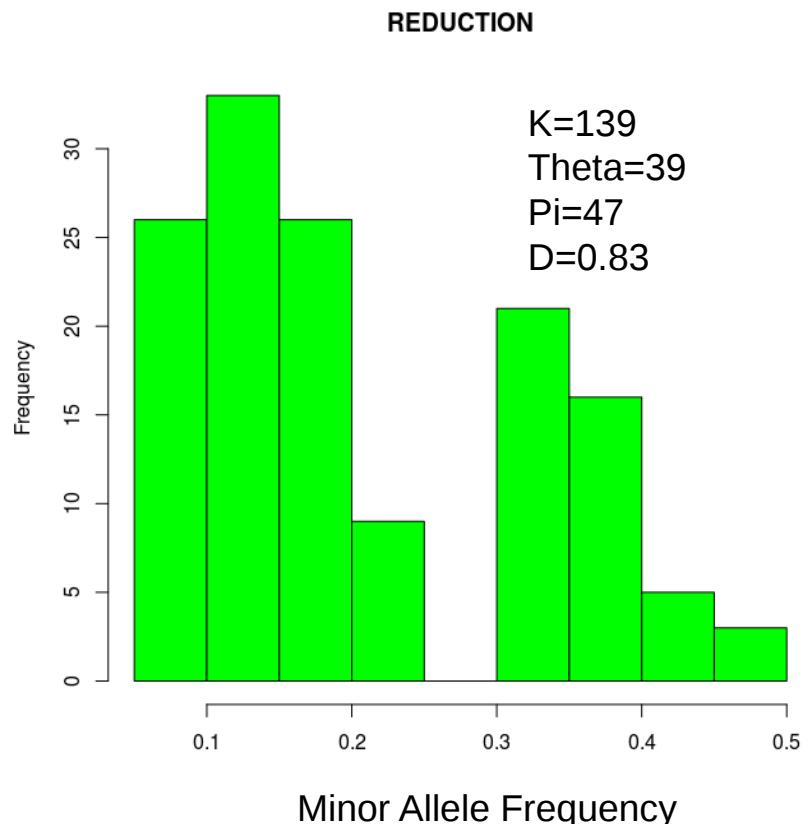


Demography matters!

n=20; L=500kbp; no selection

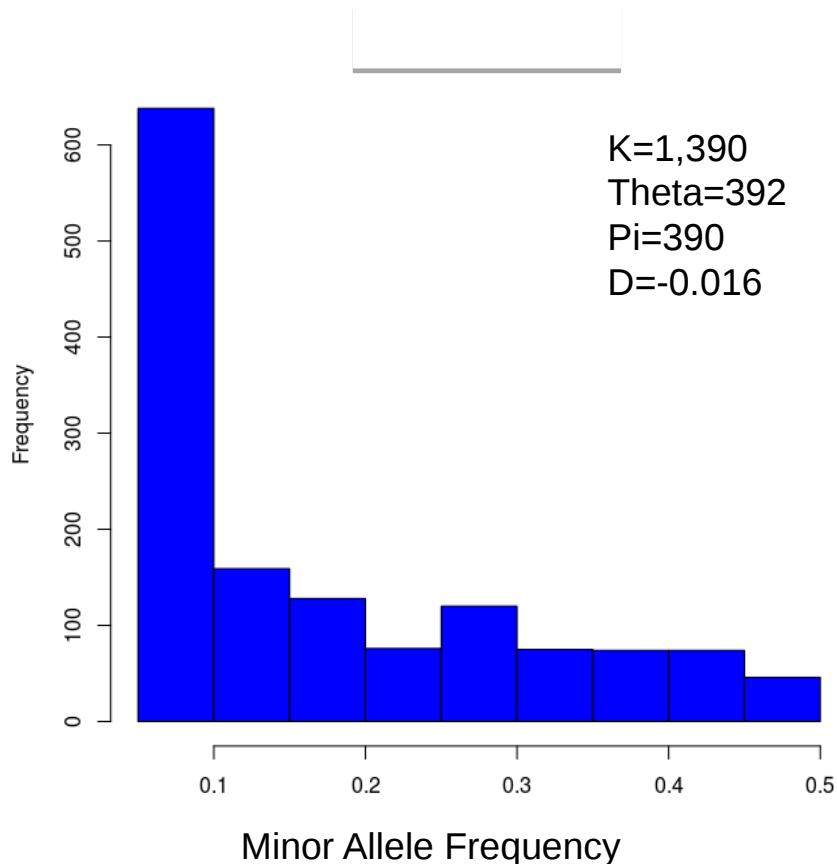


n=20; L=500kbp; no selection

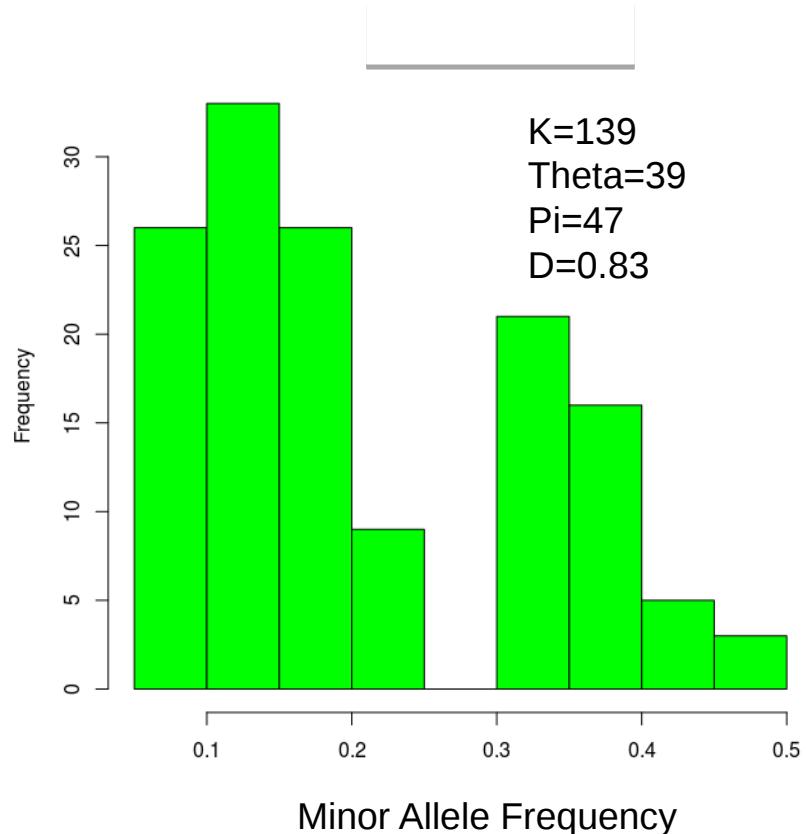


Demography matters!

n=20; L=500kbp; no selection



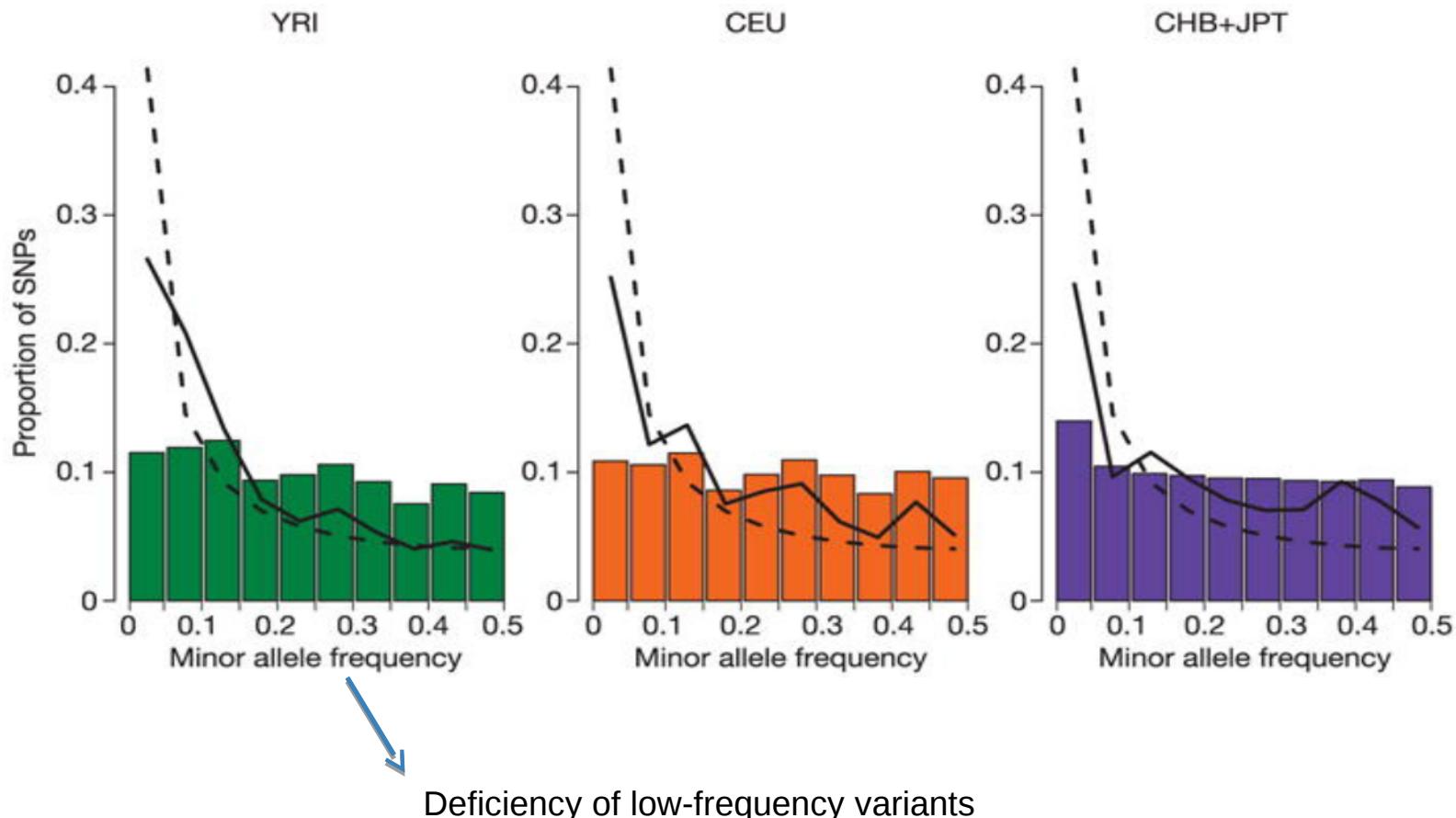
n=20; L=500kbp; no selection



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

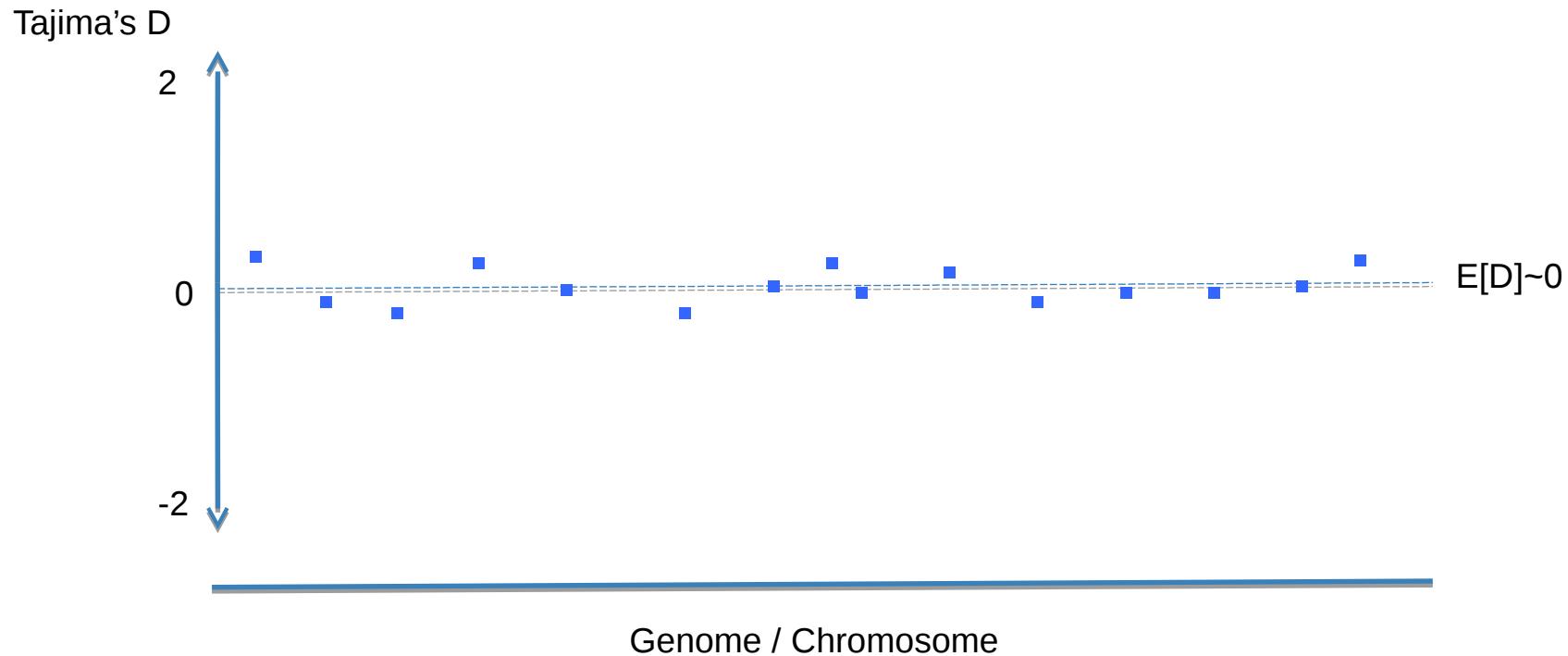
Experimental design matters?

The effect of ascertainment bias



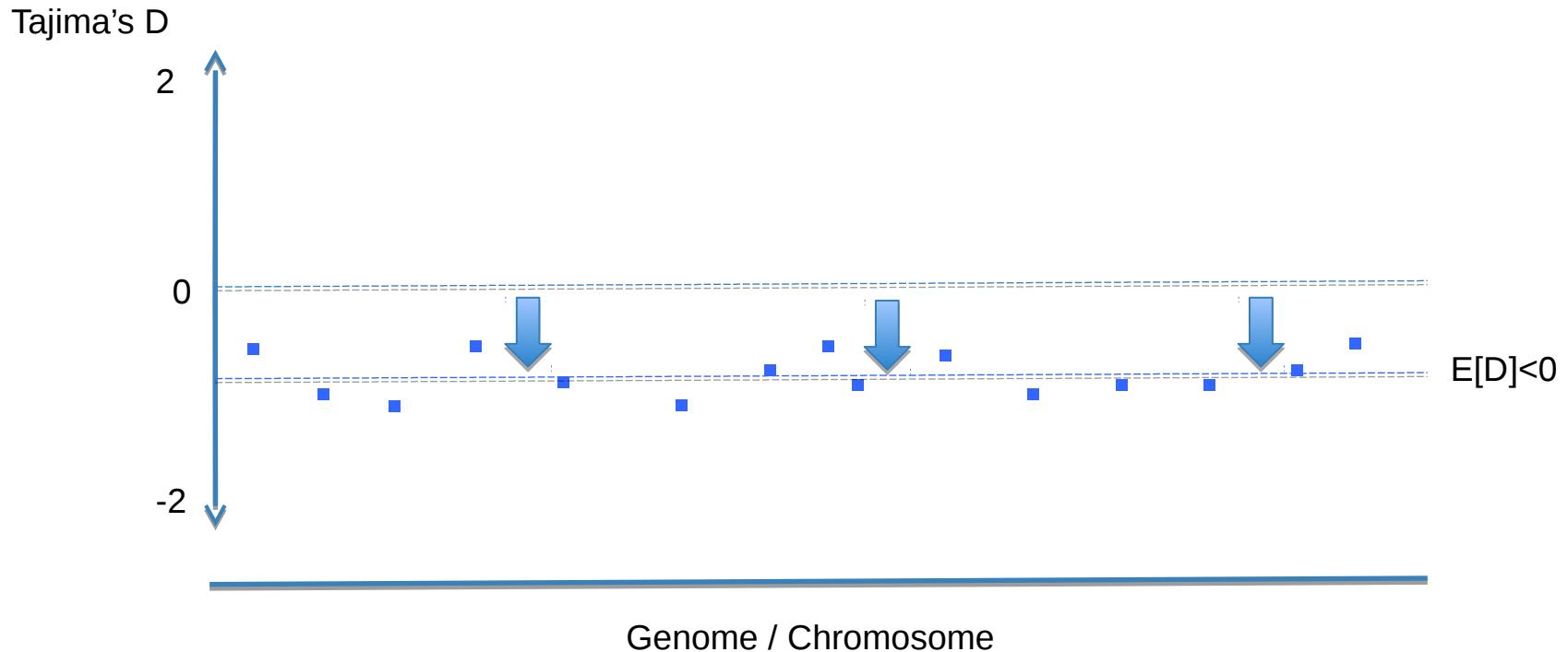
How to take neutral confounding factors into account?

Under constant population size:



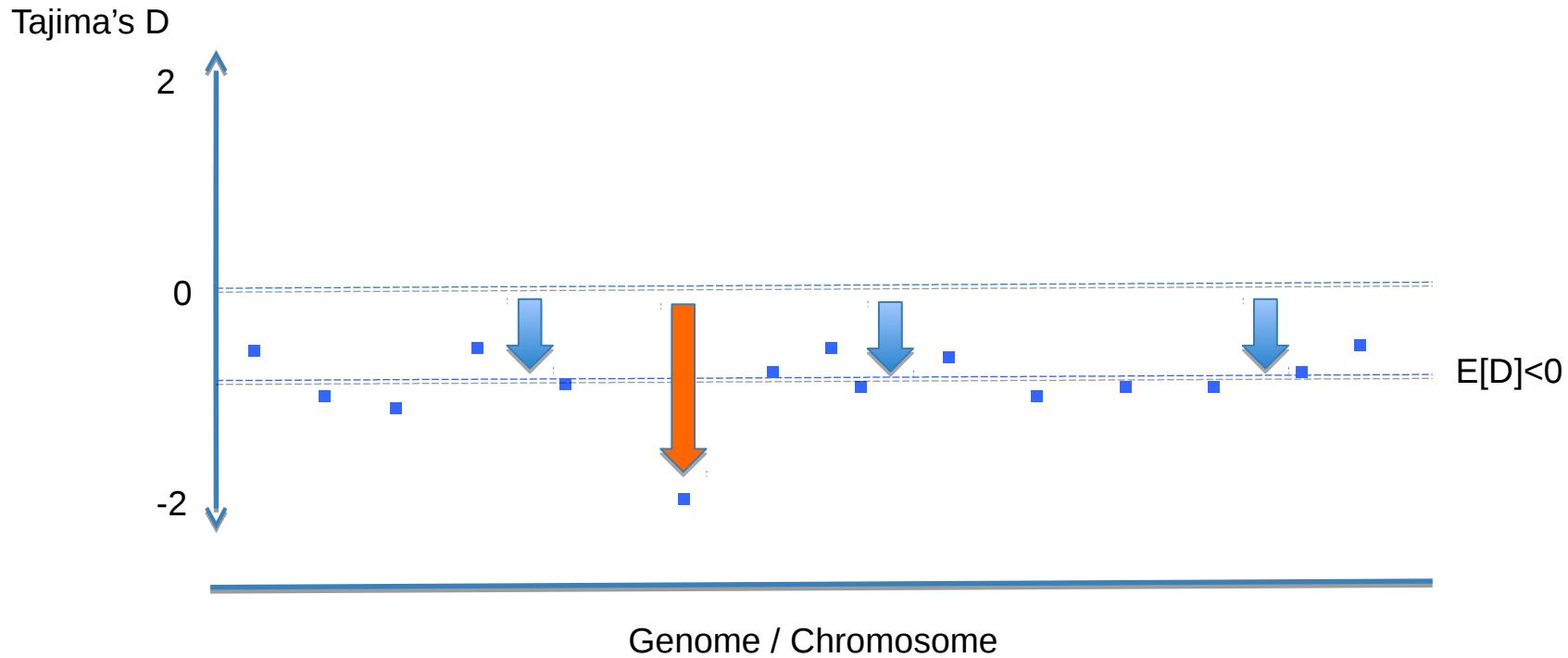
How to take neutral confounding factors into account?

Under expanding population size:



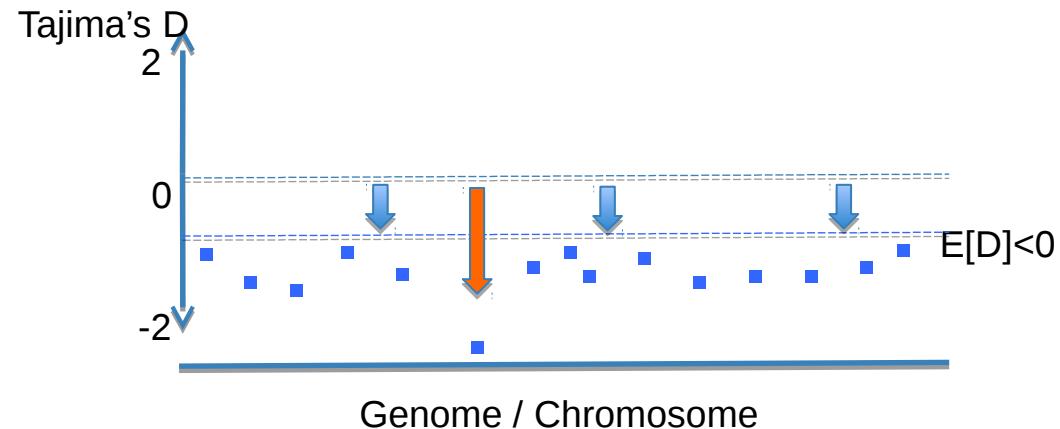
How to take neutral confounding factors into account?

Under expanding population size and positive selection:

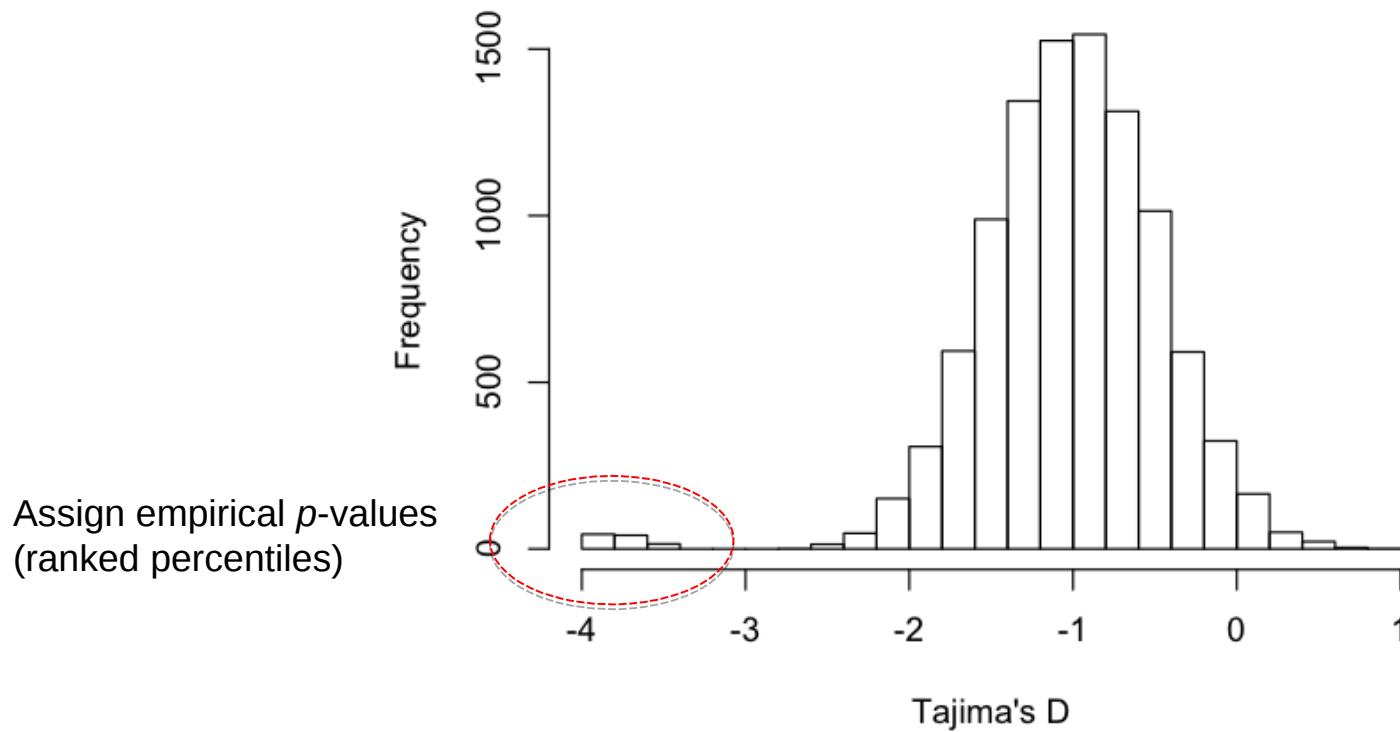


- Demography affects all loci equally, while selection changes local patterns

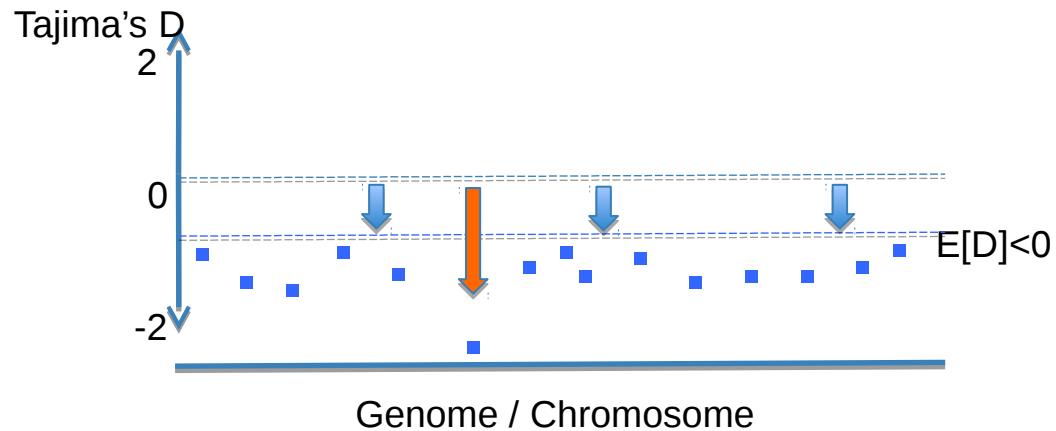
Outlier approach



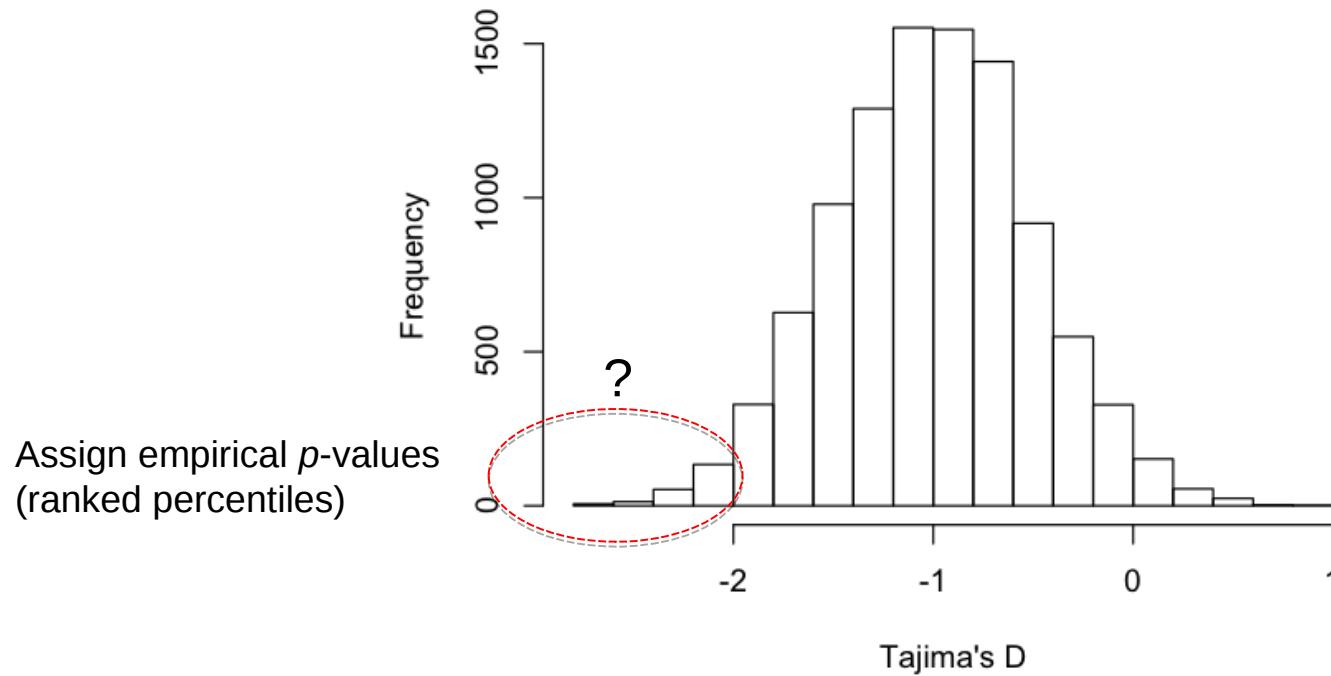
Empirical distribution



Outlier approach



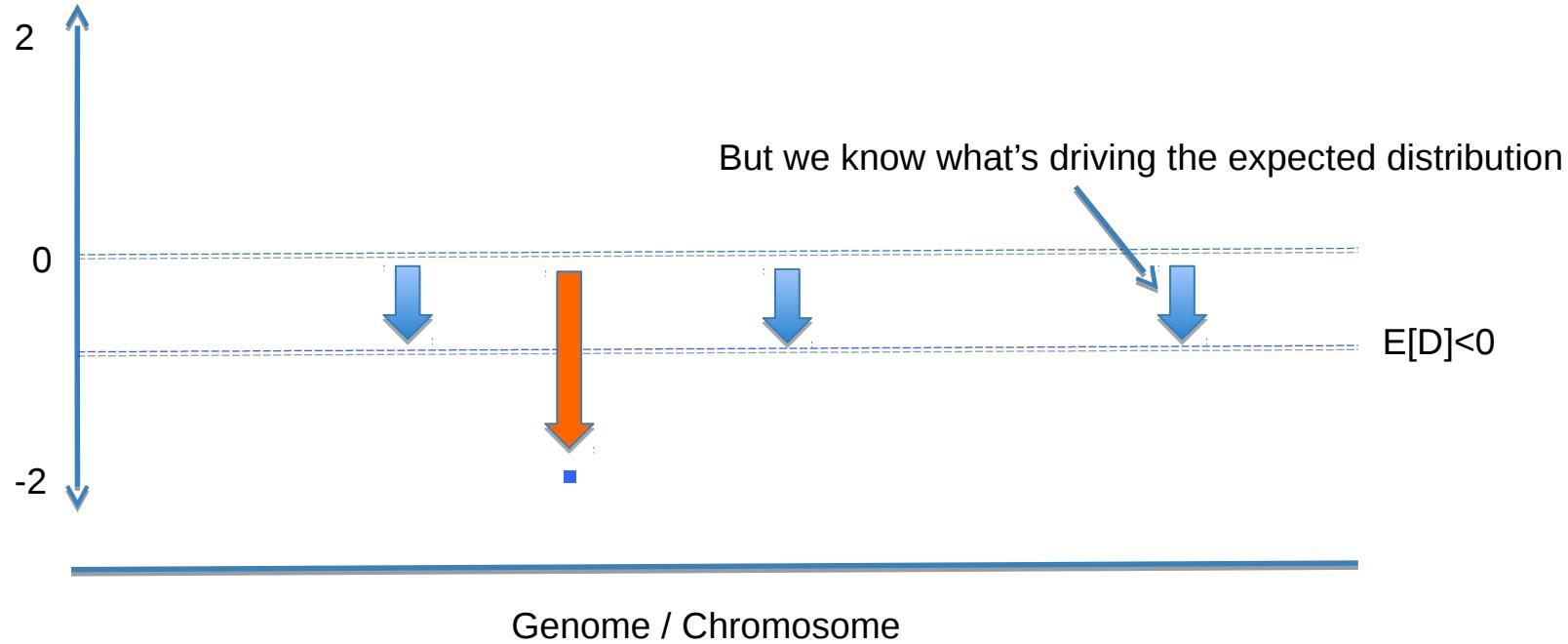
Empirical distribution



How to take neutral confounding factors into account?

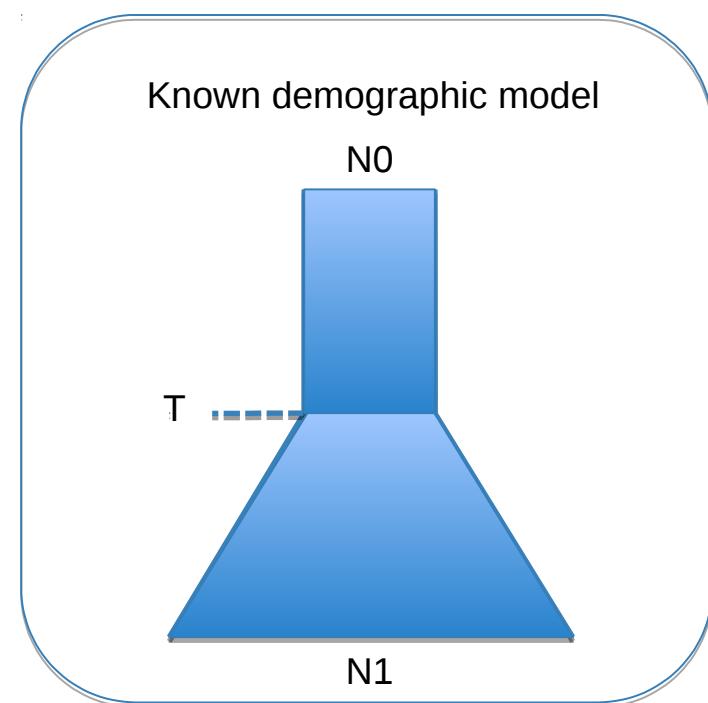
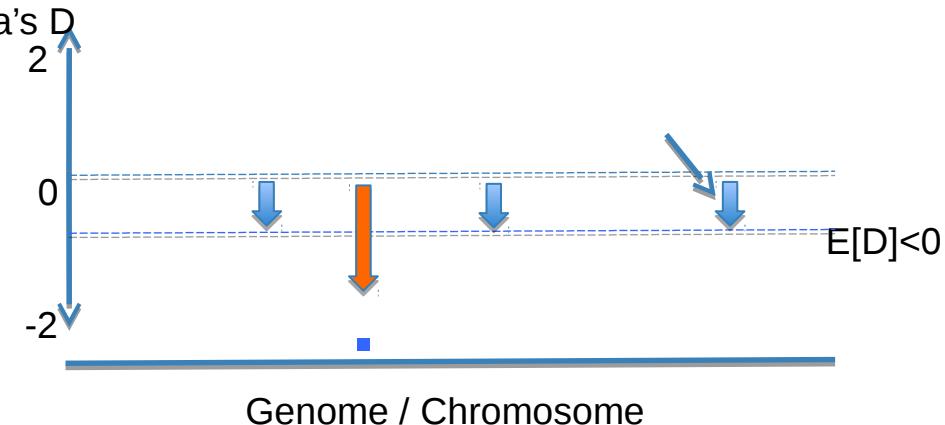
Under expanding population size and positive selection:

Tajima's D

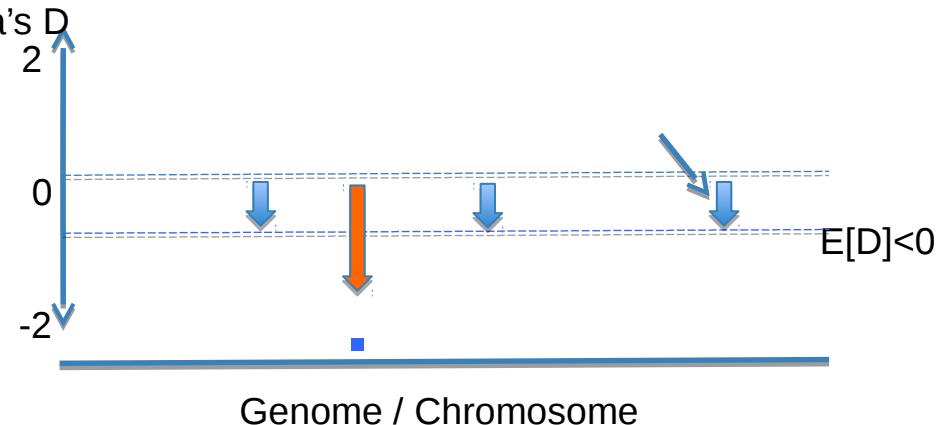


- Demography affects all loci equally, while selection changes local patterns
What should we do if we don't have genome-wide data?

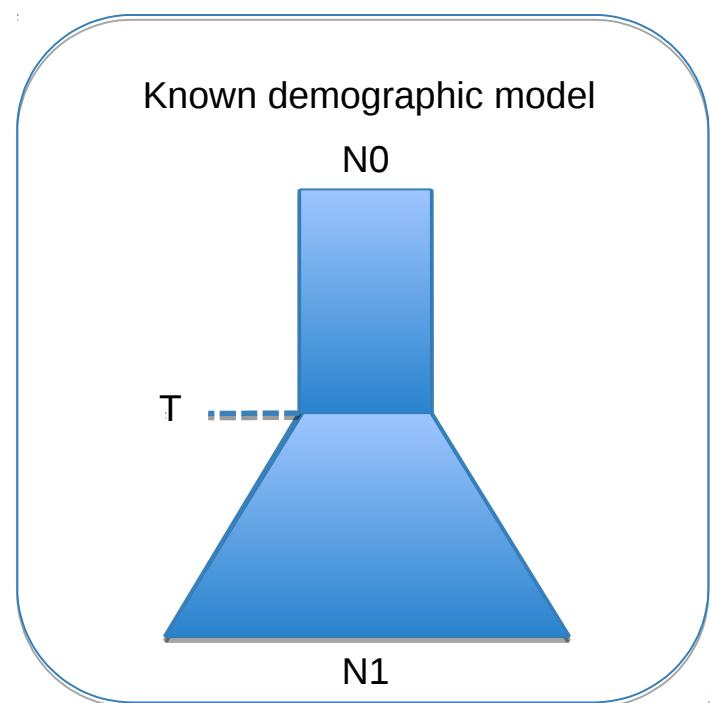
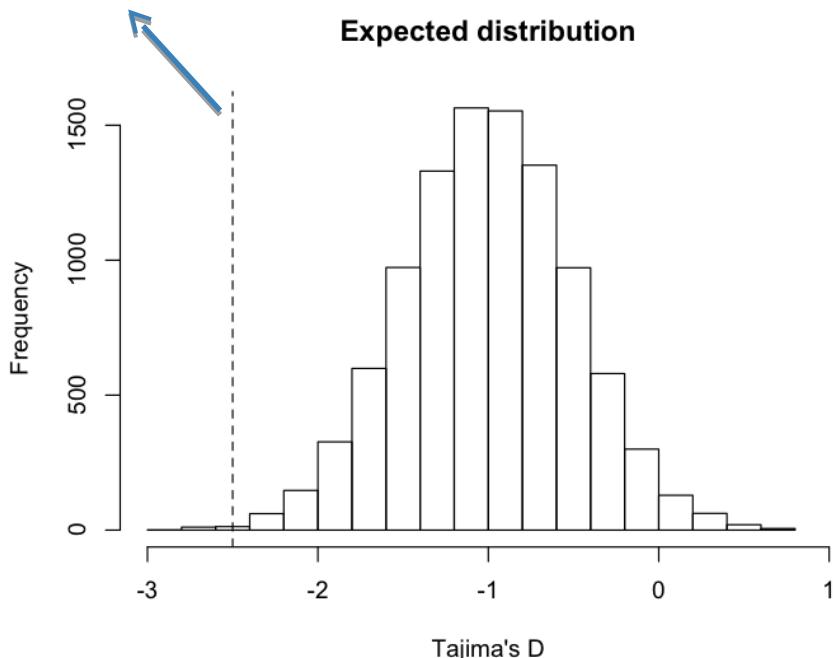
Simulations-based approach



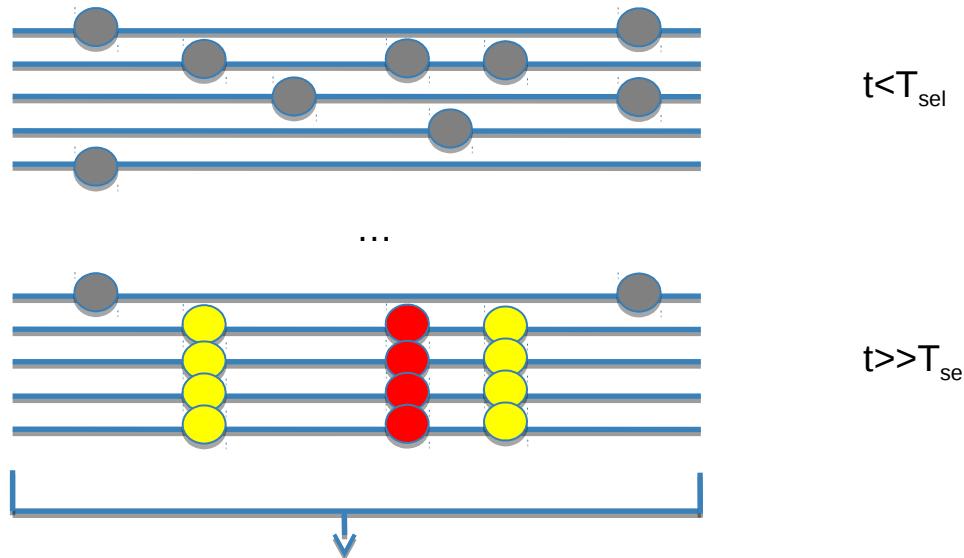
Simulations-based approach



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

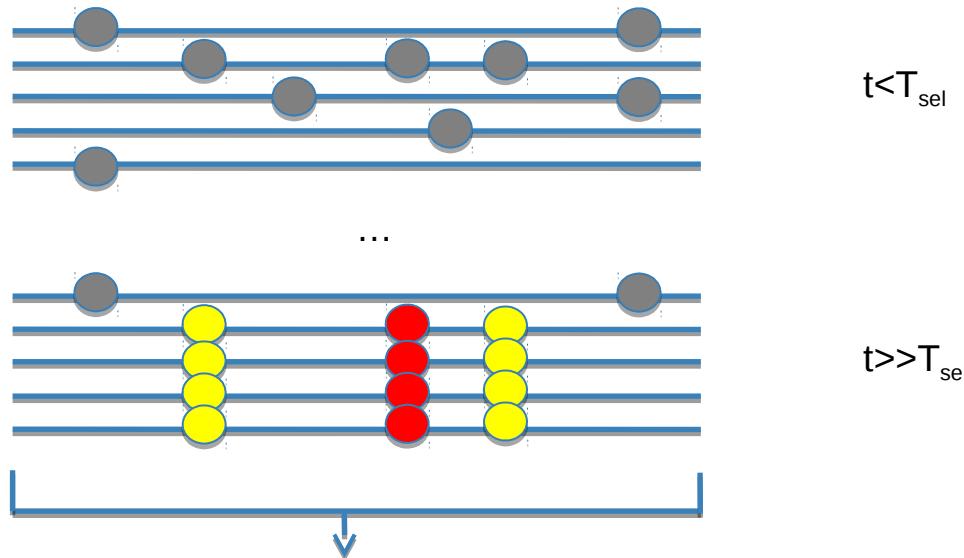


Positive selection



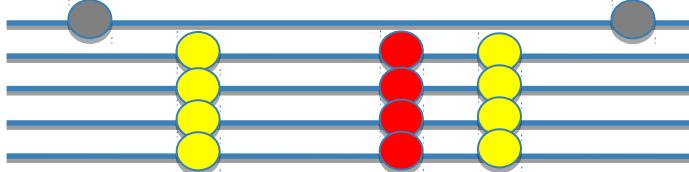
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- ?

Positive selection

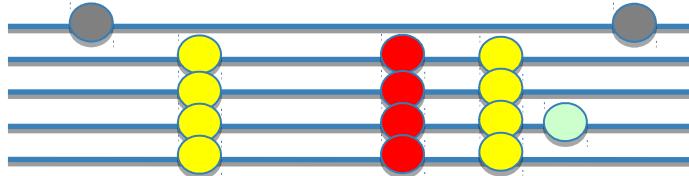


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- Extended haplotype homozygosity / Extended LD

Extended Haplotype Homozygosity

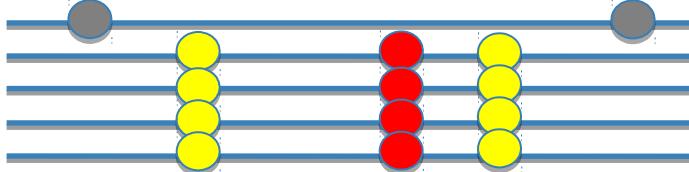


$t >> T_{sel}$

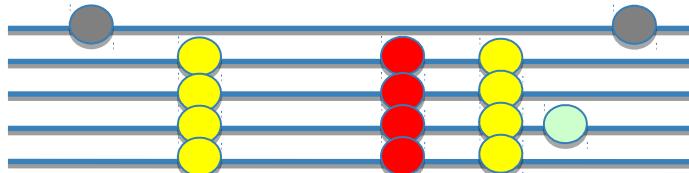


$t >>> T_{sel}$

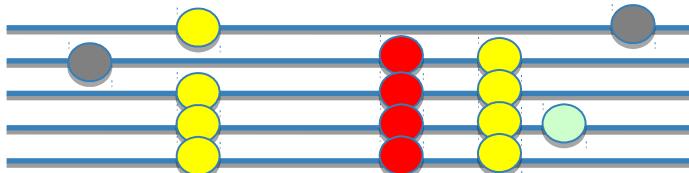
Extended Haplotype Homozygosity



$t >> T_{sel}$

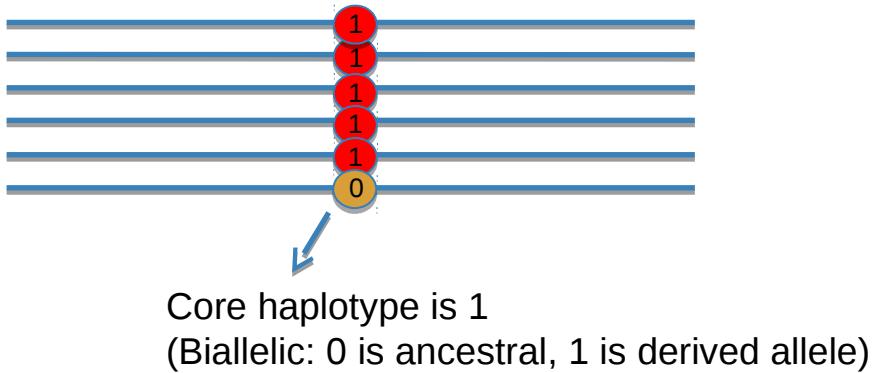


$t >>> T_{sel}$



$t >>> T_{sel}$

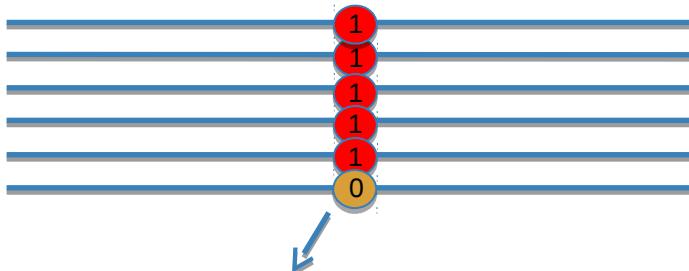
Extended Haplotype Homozygosity



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

Core SNP

Extended Haplotype Homozygosity

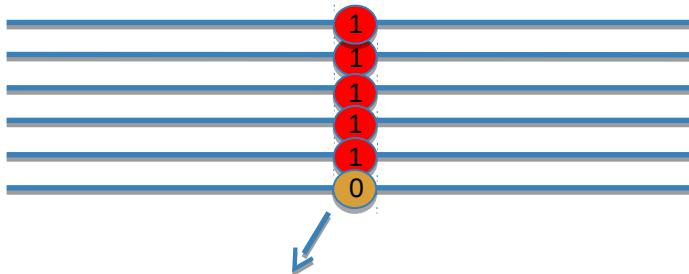


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

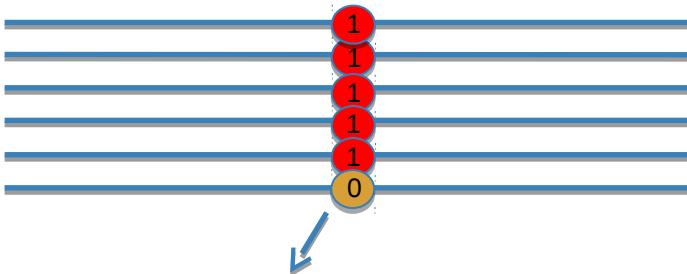


Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Sum across all unique haplotypes
carrying the core SNP

Extended Haplotype Homozygosity



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

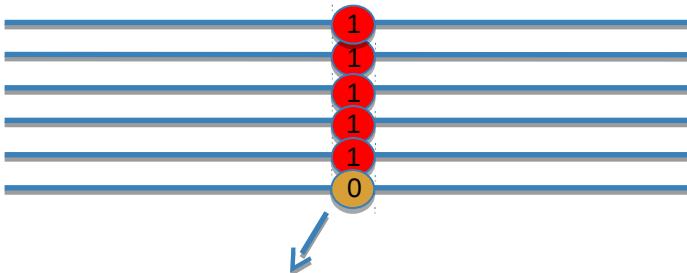
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \left(\frac{n_h}{2} \right)^2 / \left(\frac{n_c}{2} \right)^2$$

Sum across all unique haplotypes carrying the core SNP

n_h is haplotype frequency of h

n_h is haplotype frequency of the core SNP

Extended Haplotype Homozygosity



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

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

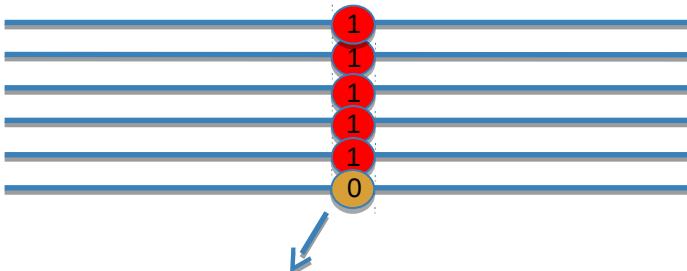
Sum across all unique haplotypes carrying the core SNP

n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

$$EHH_c(x_i = 0) = ?$$

Extended Haplotype Homozygosity



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

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

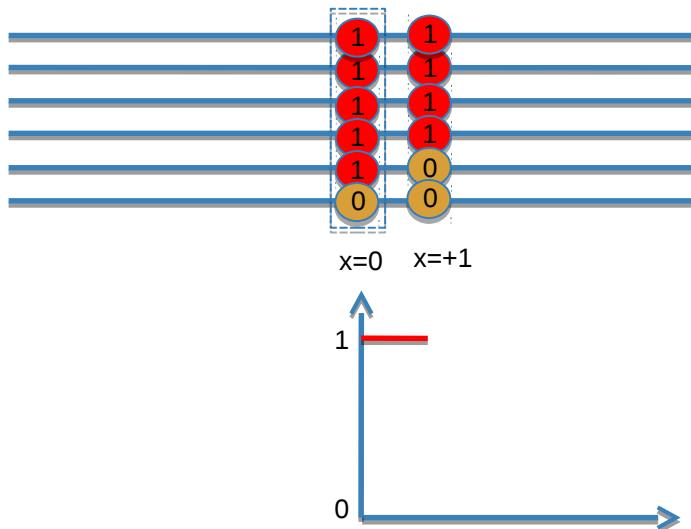
n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

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

Extended Haplotype Homozygosity

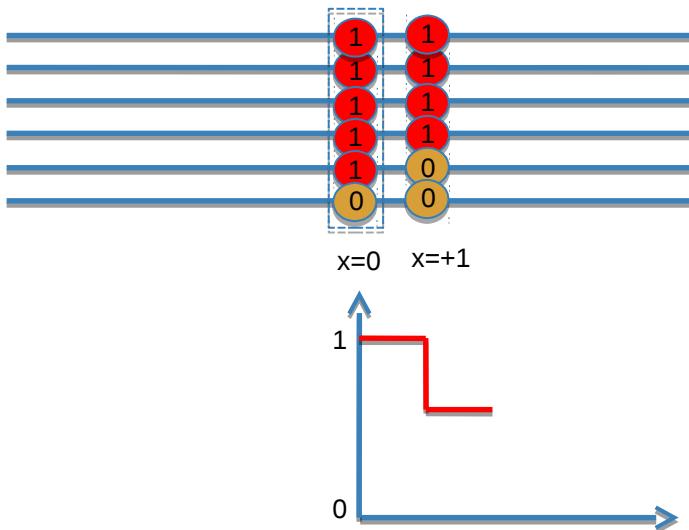


$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

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

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

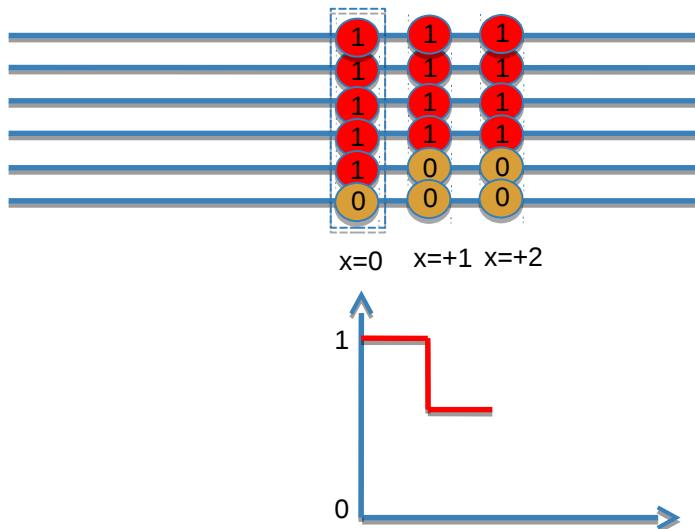
Extended Haplotype Homozygosity



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

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

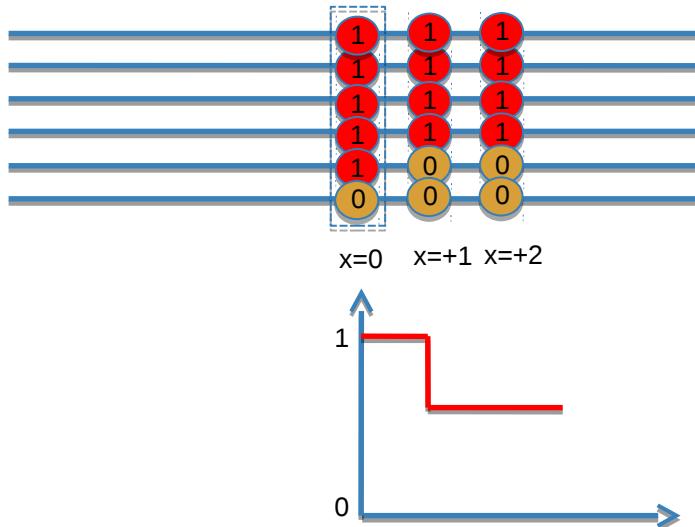
Extended Haplotype Homozygosity



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

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

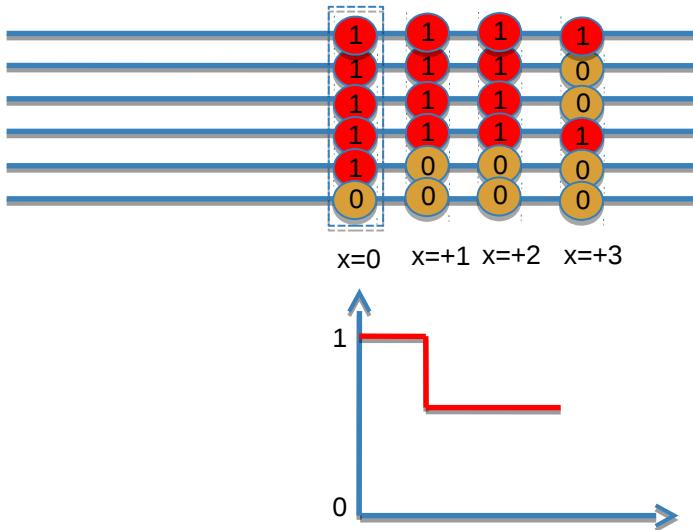
Extended Haplotype Homozygosity



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

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

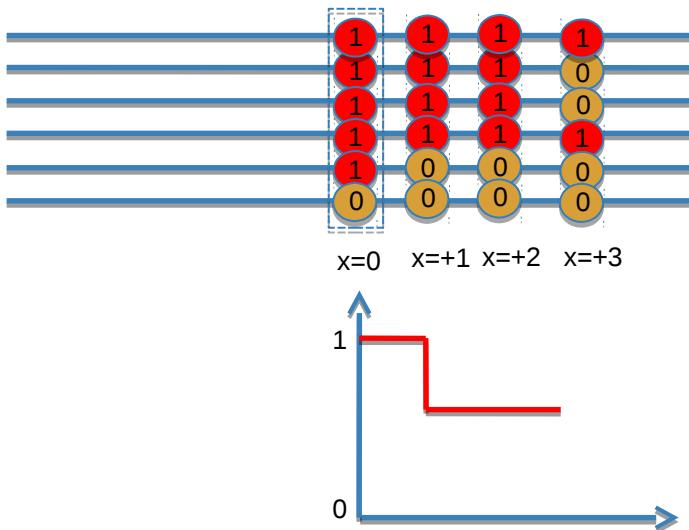
Extended Haplotype Homozygosity



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

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

Extended Haplotype Homozygosity



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

How many unique haplotypes carrying the core SNP?

What is their frequency?

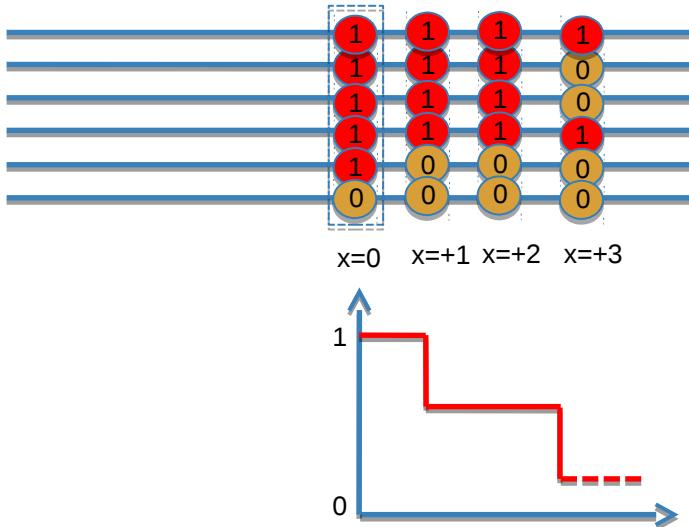
1111 with freq=2

1110 with freq=2

1000 with freq=1

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

Extended Haplotype Homozygosity



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

How many unique haplotypes carrying the core SNP?

What is their frequency?

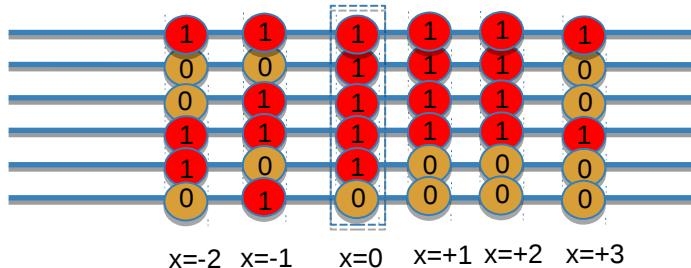
1111 with freq=2

1110 with freq=2

1000 with freq=1

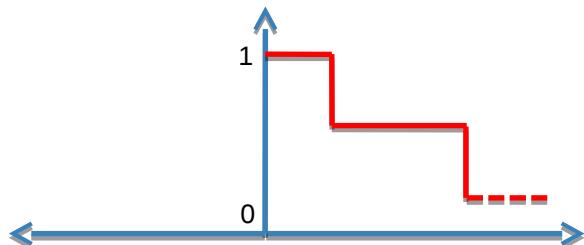
$$EHH_c(x_i = +3) = \frac{\binom{2}{2} + \binom{2}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+1+0}{10} = 0.20$$

Extended Haplotype Homozygosity



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

n	$n \text{ choose } 2$
1	0
2	1
3	3
4	6
5	10
6	15

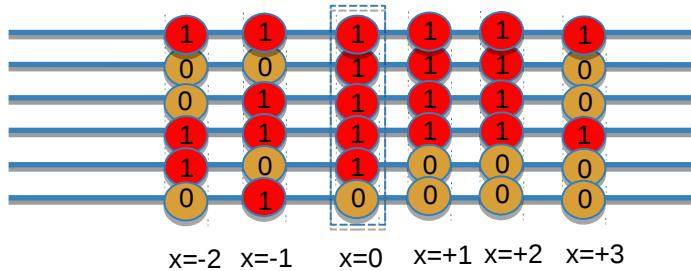


$$EHH_c(x_i = -1) = ?$$

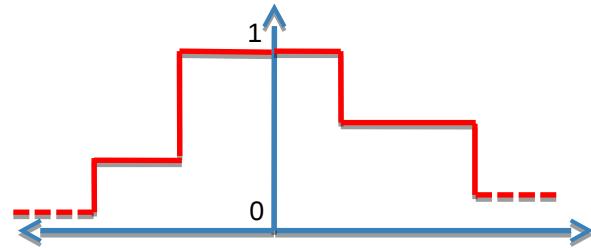
$$EHH_c(x_i = -2) = ?$$

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

Extended Haplotype Homozygosity



n	$n \text{ choose } 2$
1	0
2	1
3	3
4	6
5	10
6	15



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

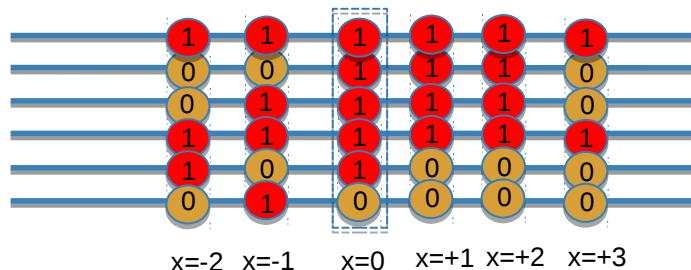
$$EHH_c(x_i = -1) = \frac{\binom{3}{2} + \binom{2}{2}}{\binom{5}{2}} = \frac{3+1}{10} = 0.4$$

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

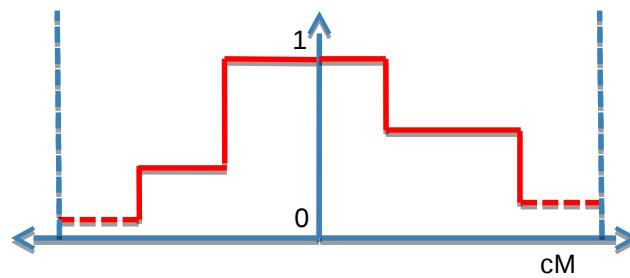
+ (1 choose 2)

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

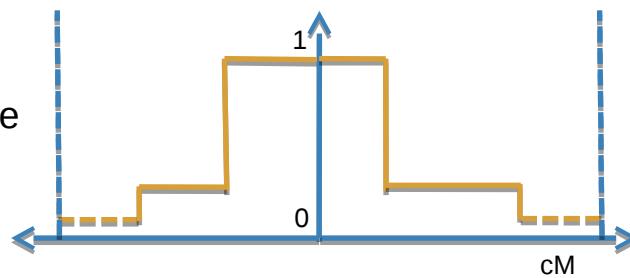
Integrated Haplotype Score



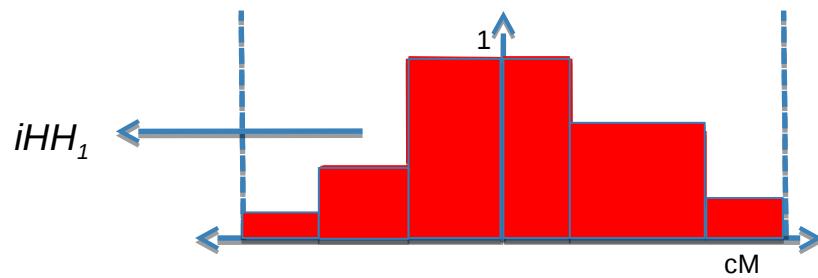
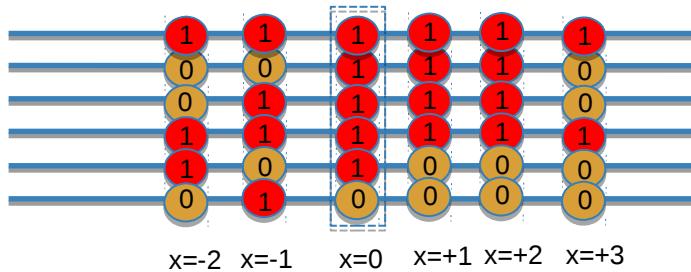
For the derived allele



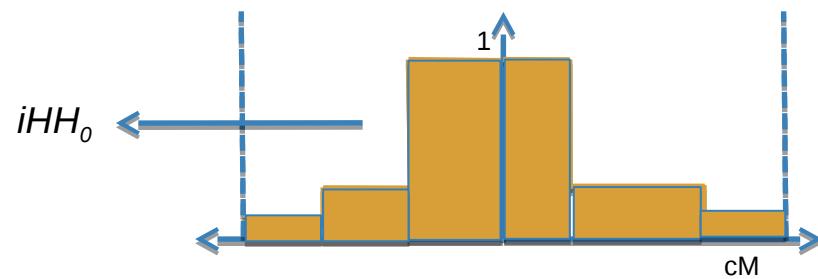
For the ancestral allele



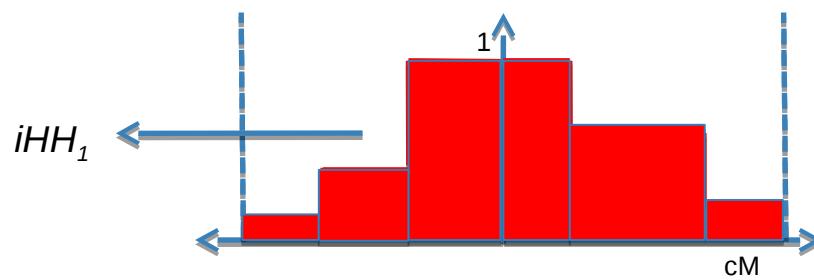
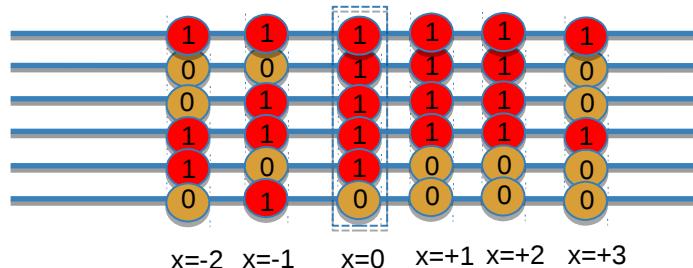
Integrated Haplotype Score



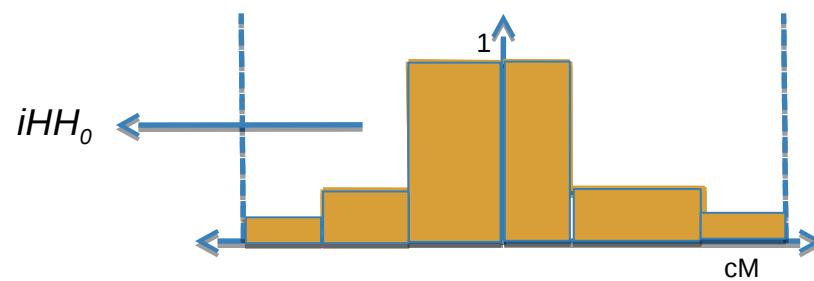
Integrated haplotype homozygosity (iHH)



Integrated Haplotype Score



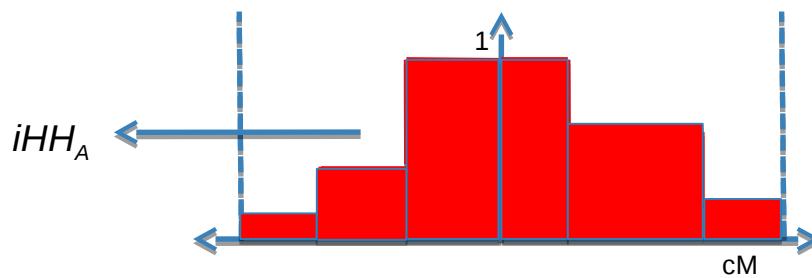
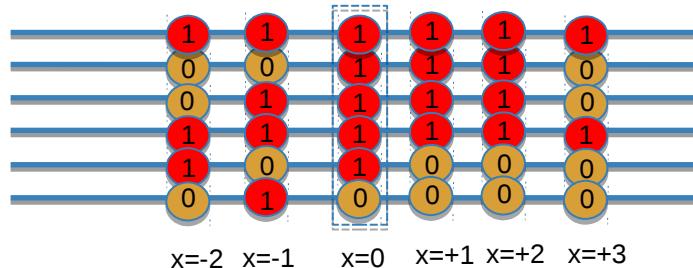
Integrated haplotype homozygosity (iHH)



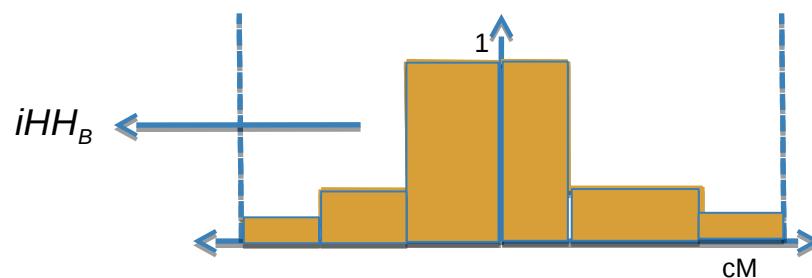
Integrated haplotype score:
 $iHs = \ln(iHH_1/iHH_0)$

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**



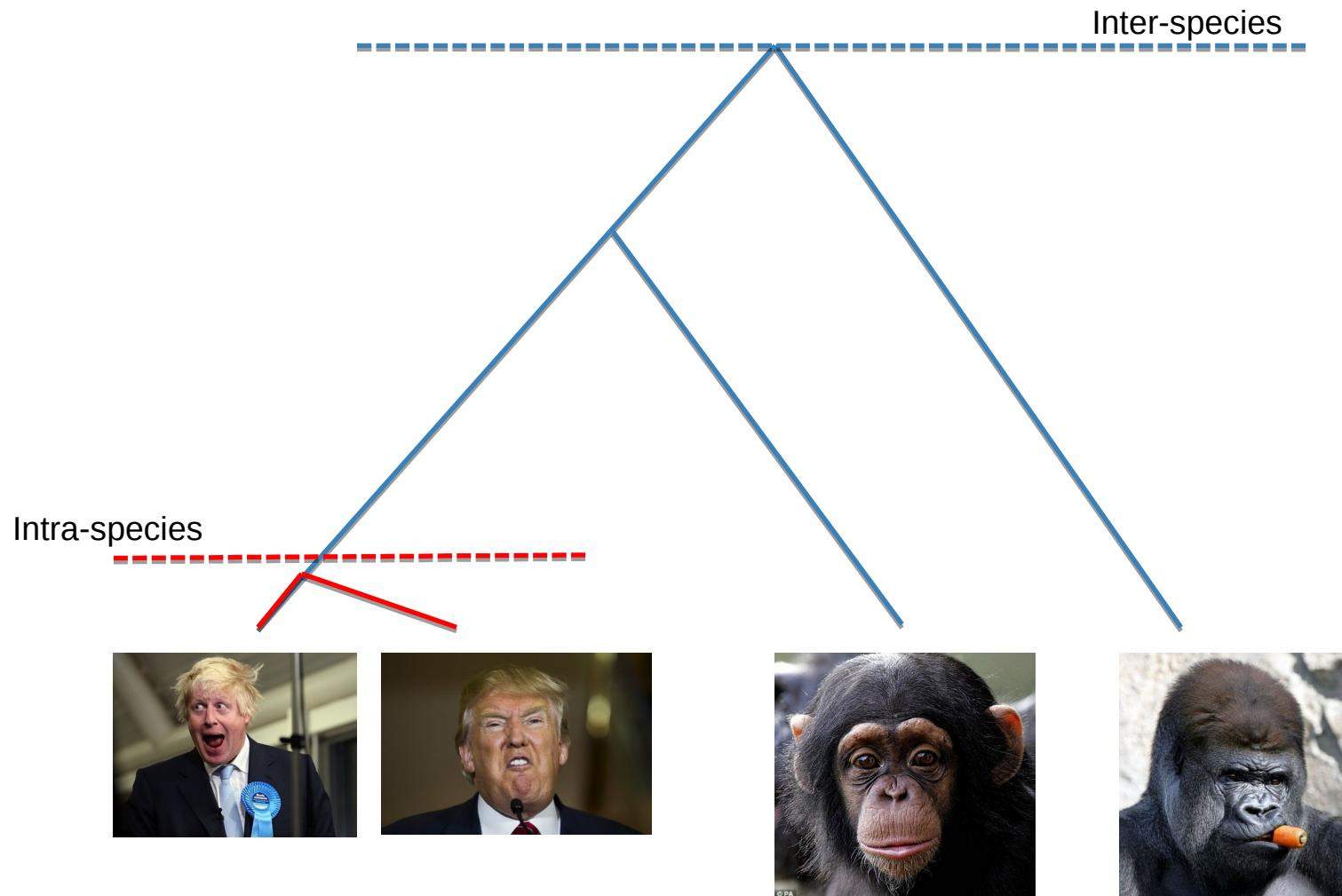
Integrated haplotype score:
 $XP-EHH = \ln(iHH_A/iHH_B)$

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

Outline

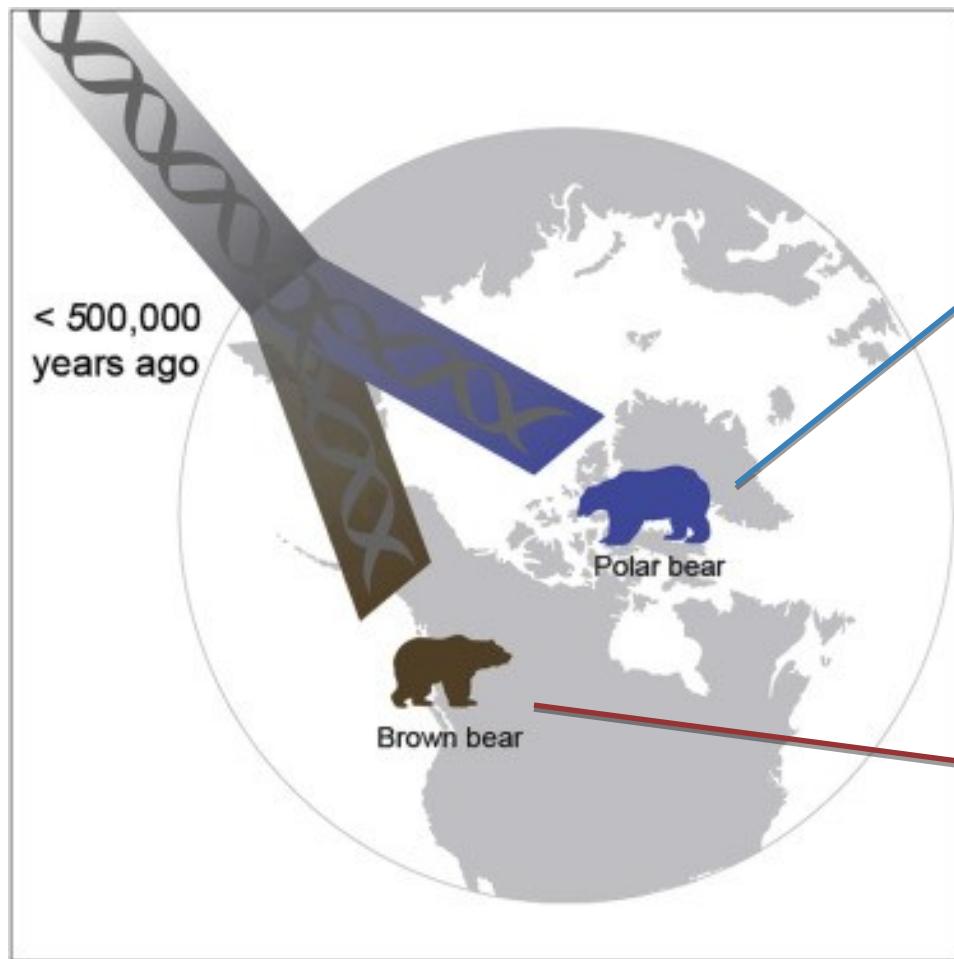
-
-
- -
 -
 -
 -
- Inferring selection at the inter-species level
-

Inter-species variation



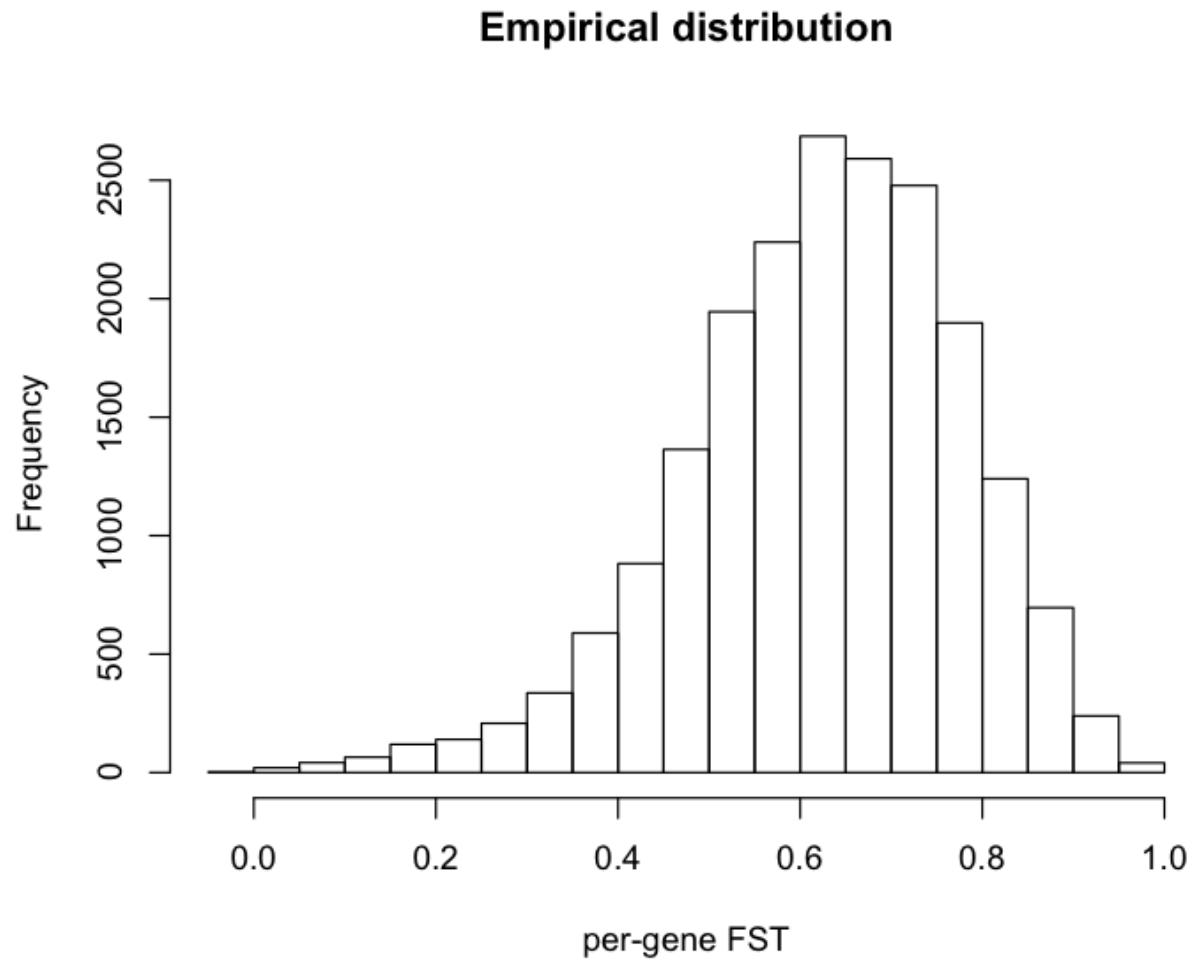
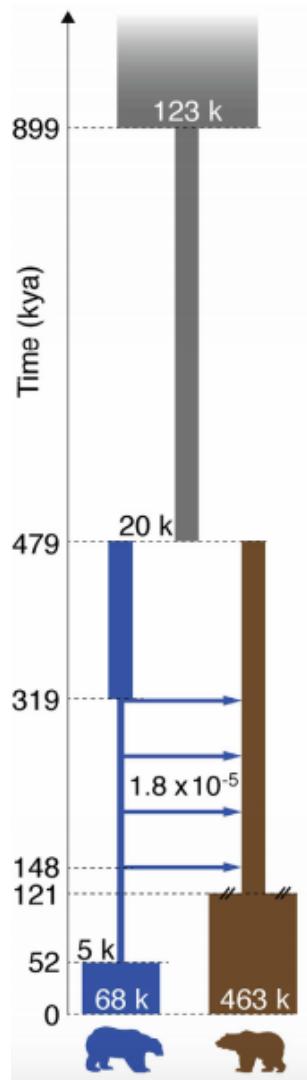
events in the deep past, macro-evolutionary trends, selection between species

Polar bears vs. Brown bears

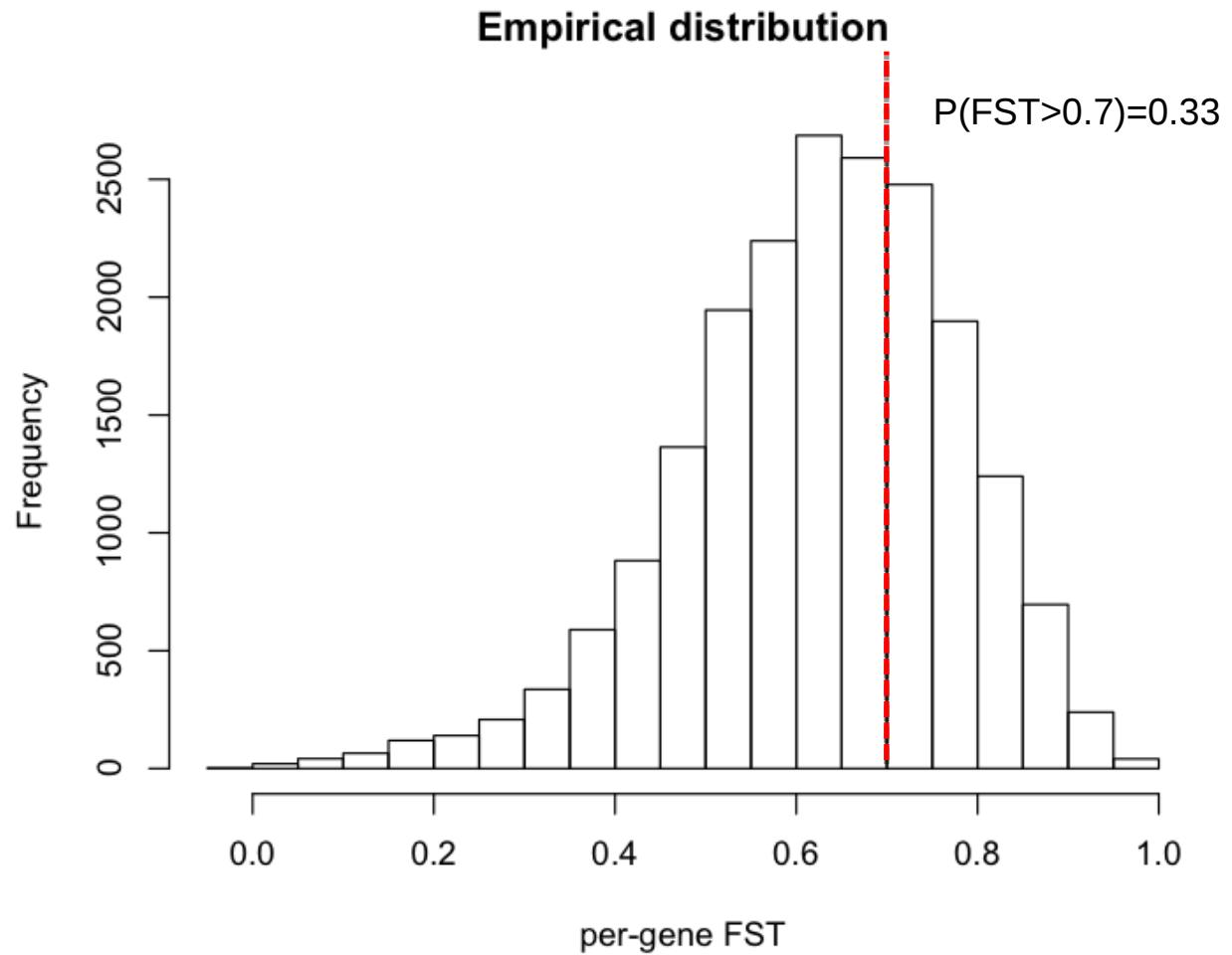
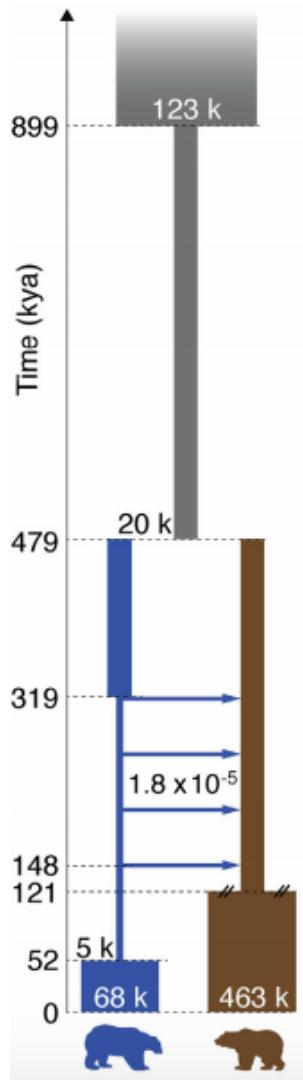


Question: what are the genetic signatures of polar bears' adaptation to the Arctic environment?

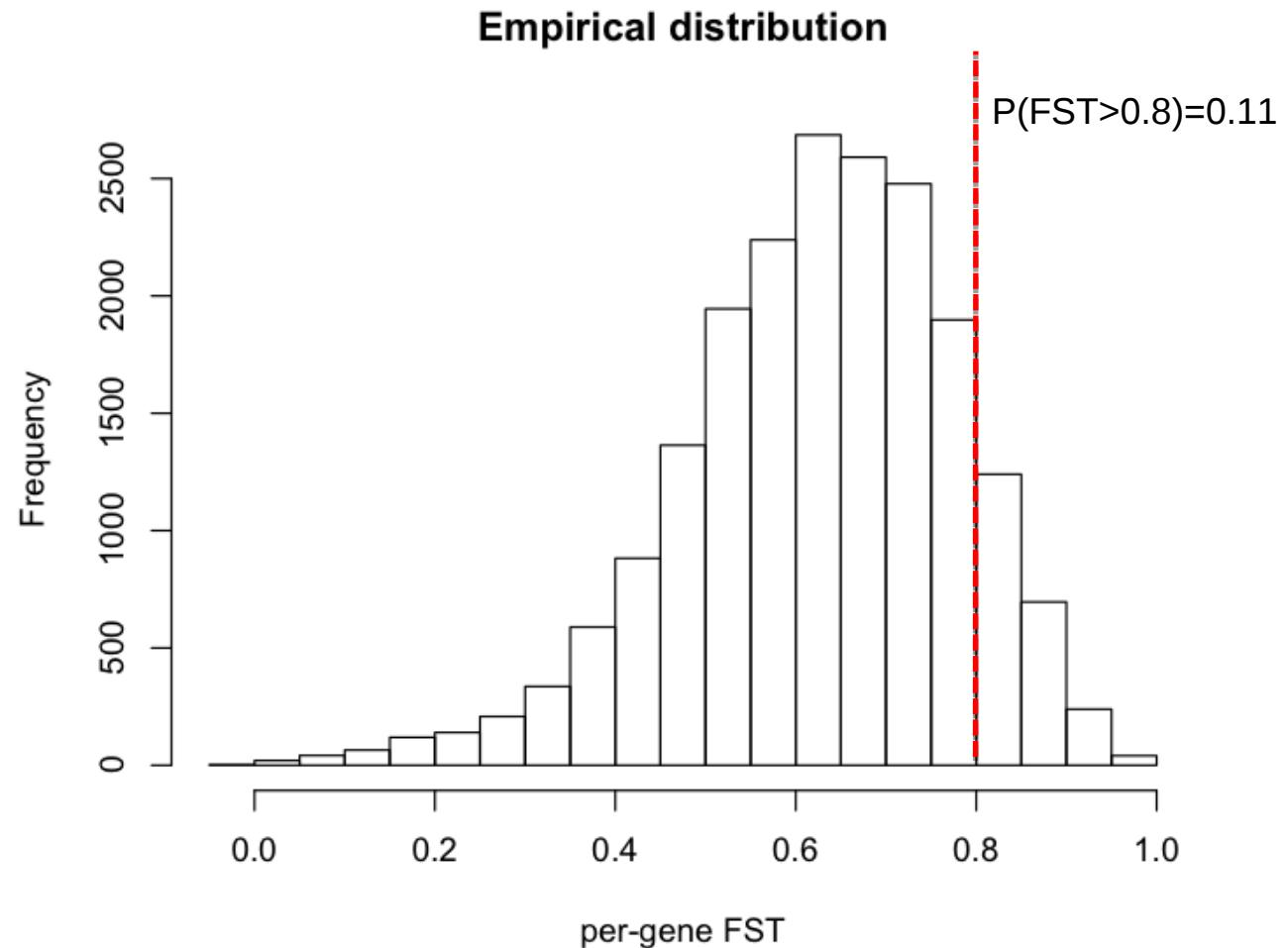
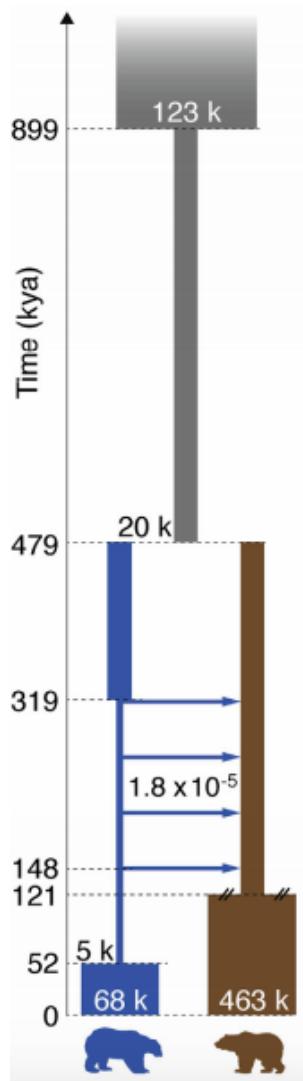
Expected genetic differentiation



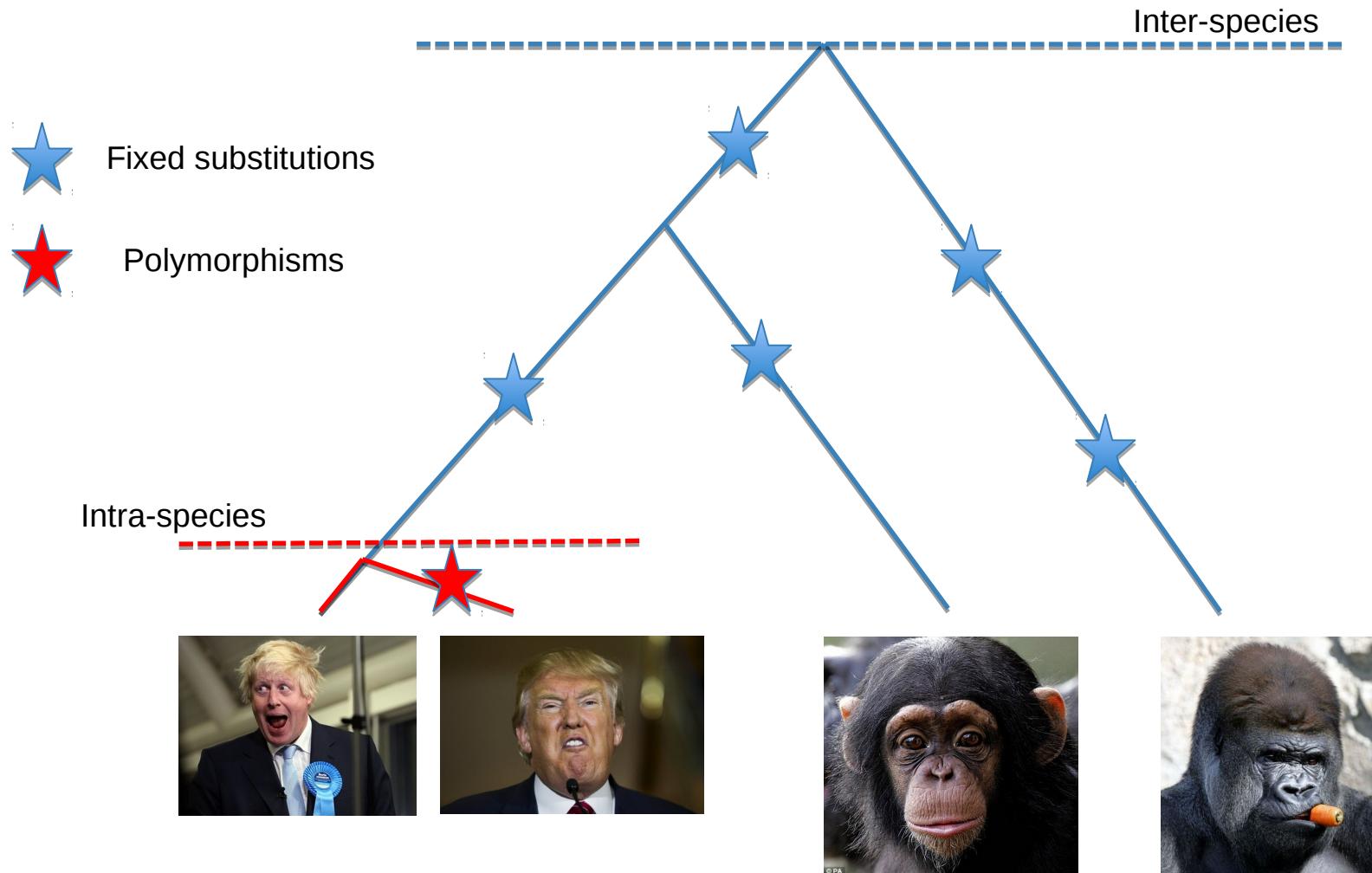
Expected genetic differentiation



Expected genetic differentiation

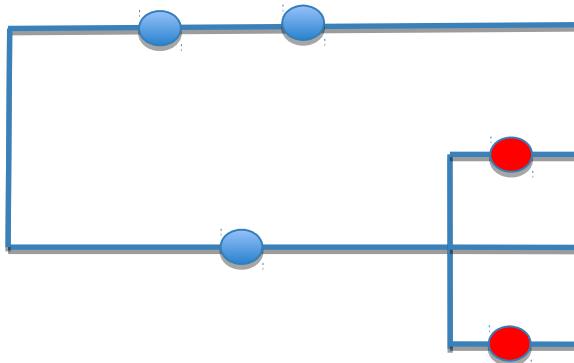


Inferring inter-species selection

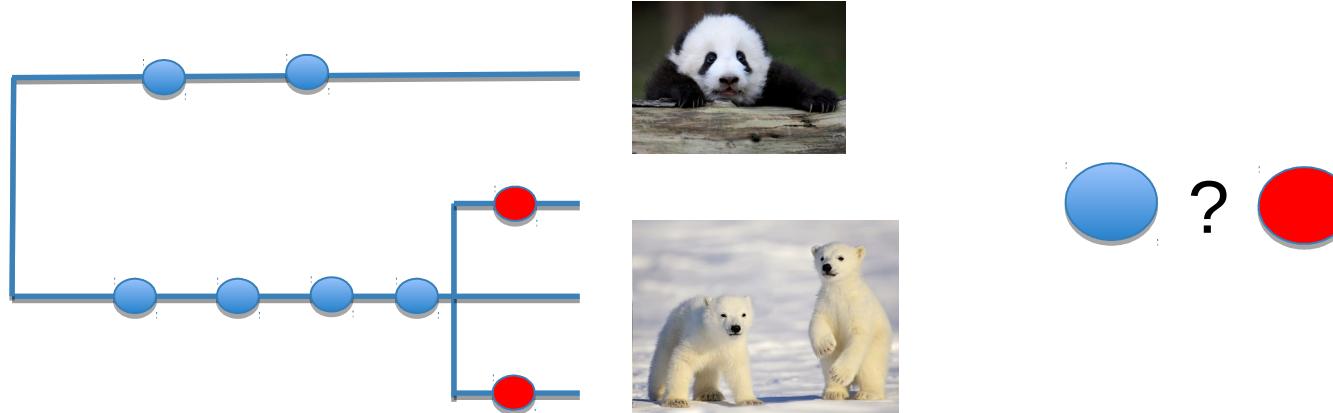
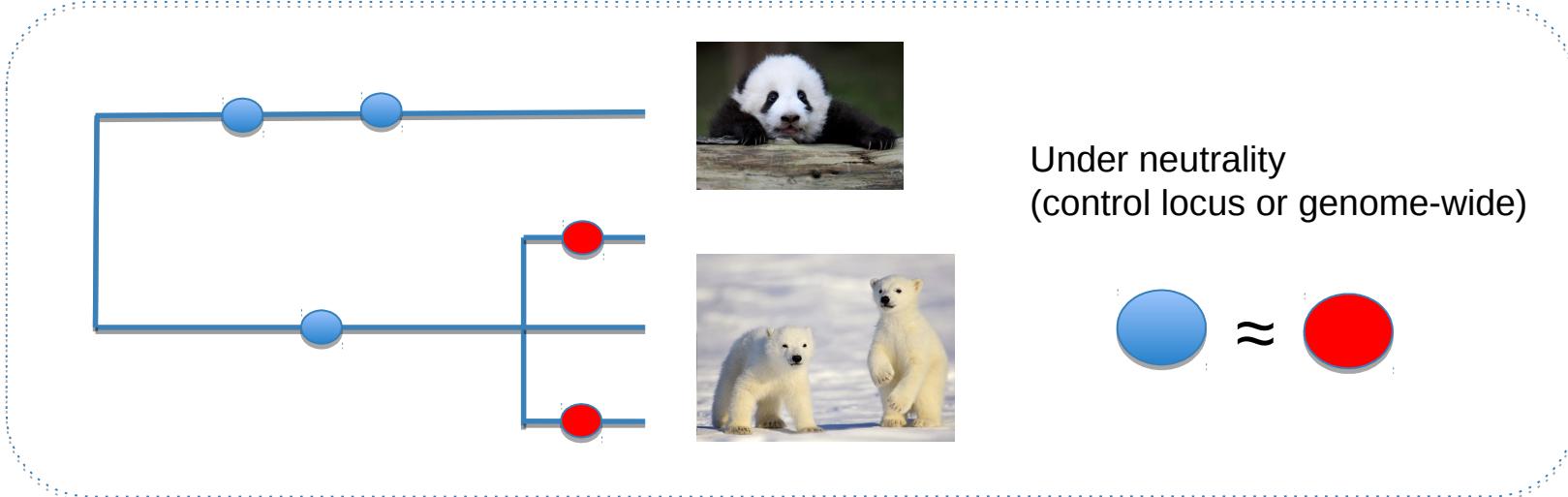


Polymorphisms and divergence

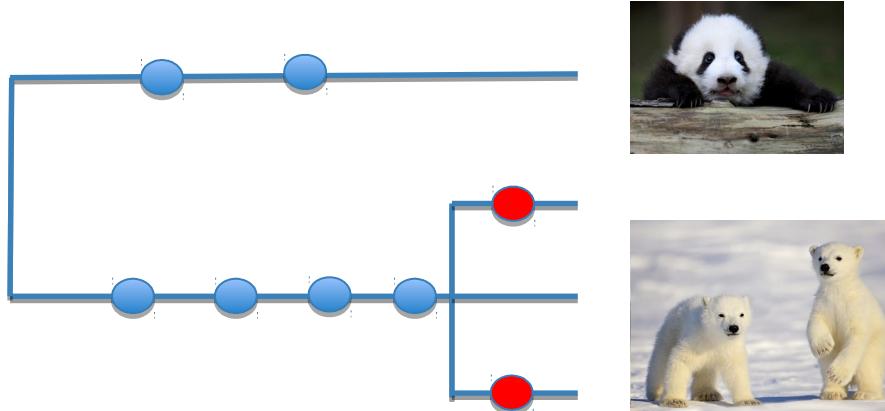
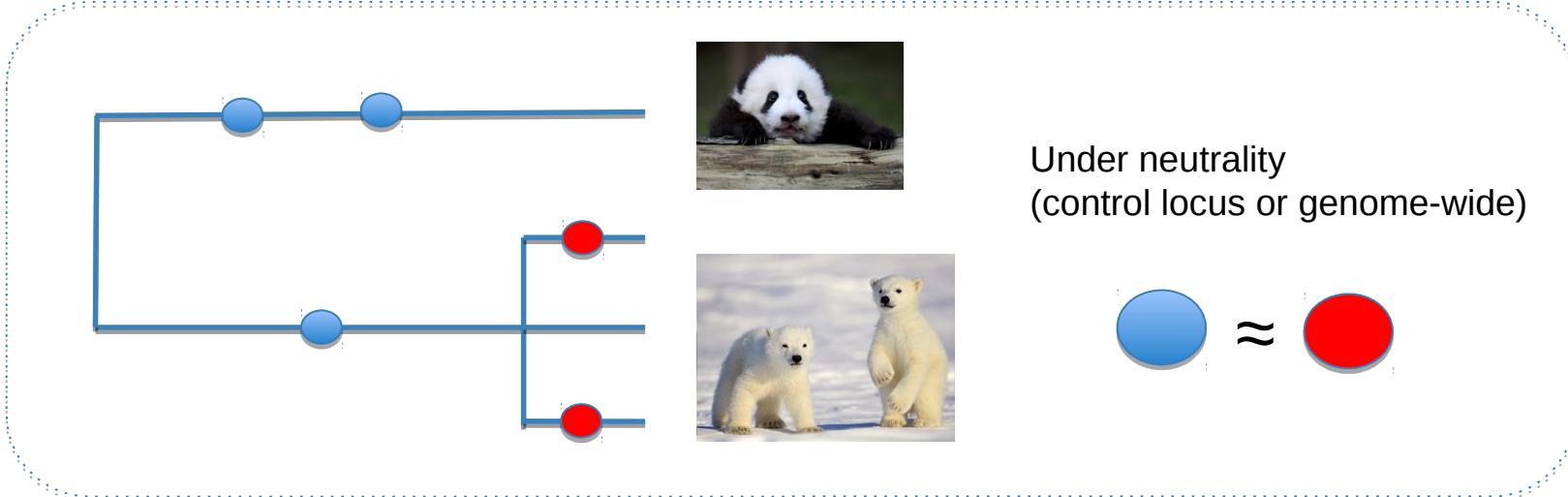
As both depend on mutation rates,  and  levels are expected to be proportional.



Polymorphisms and divergence



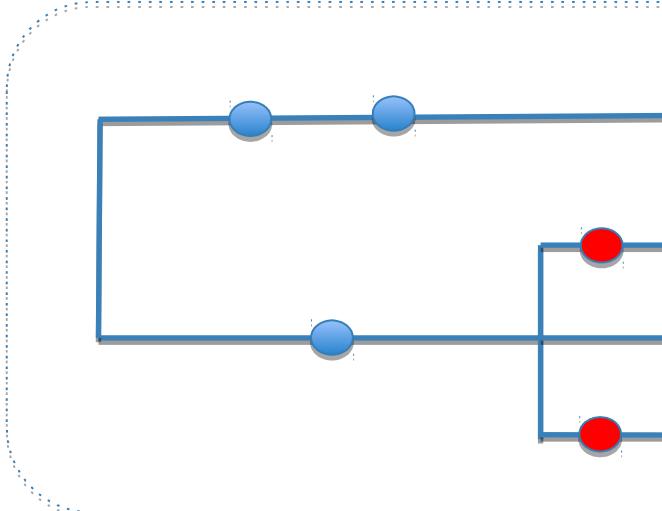
Polymorphisms and divergence



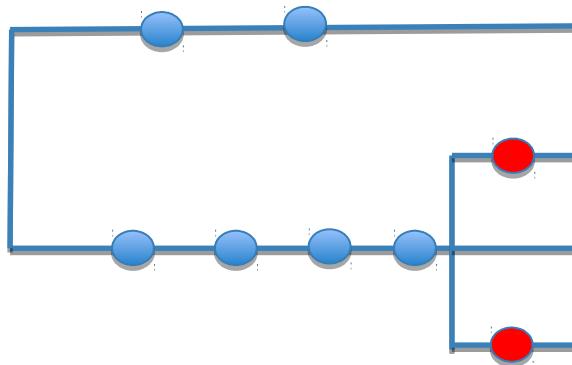
??? selection



Polymorphisms and divergence



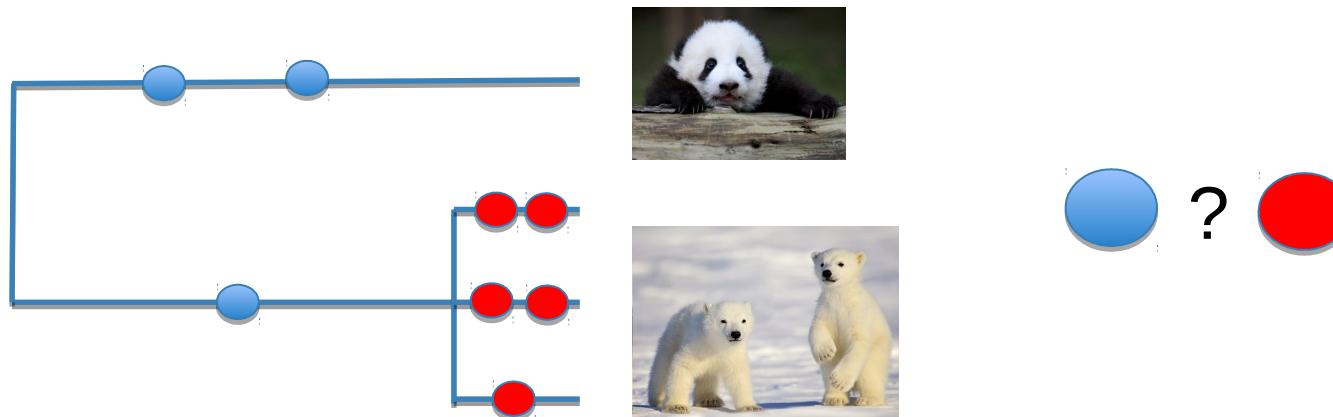
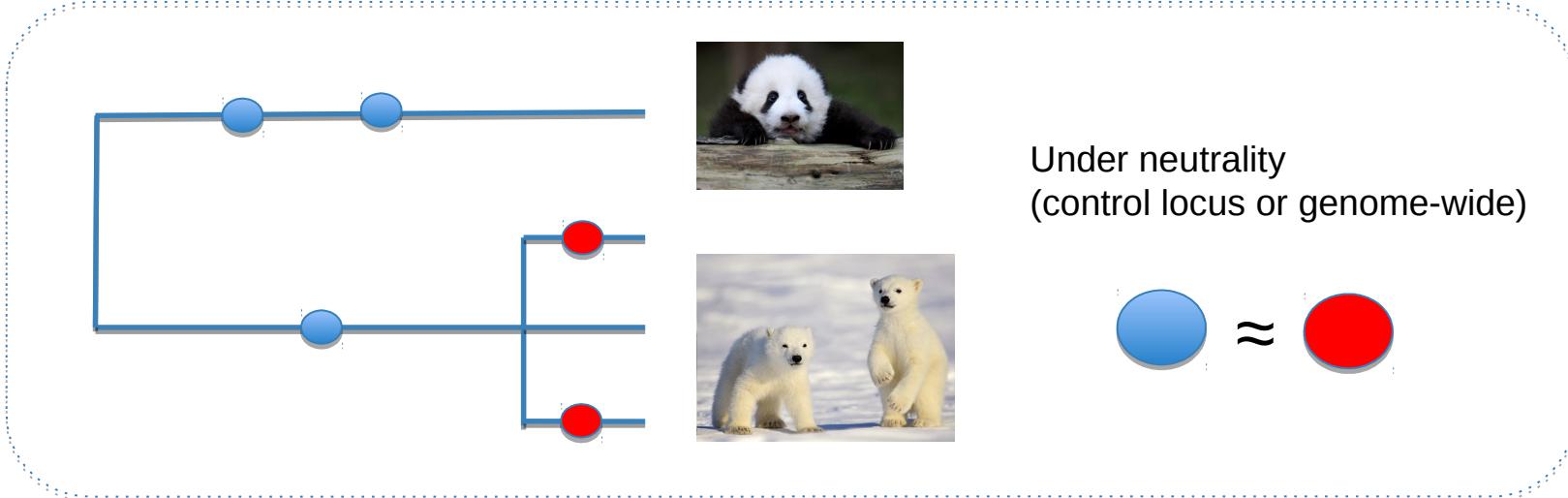
Under neutrality
(control locus or genome-wide)



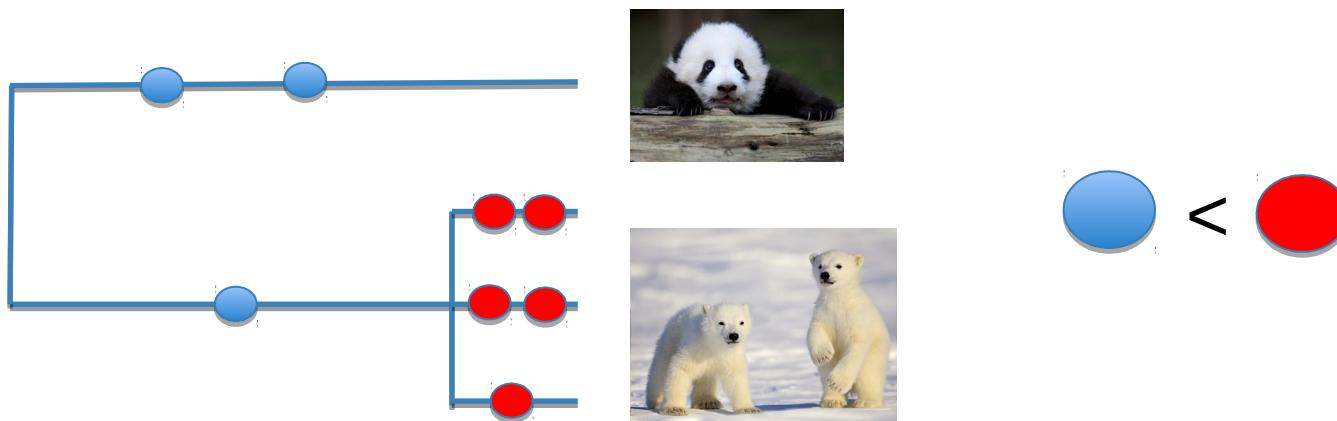
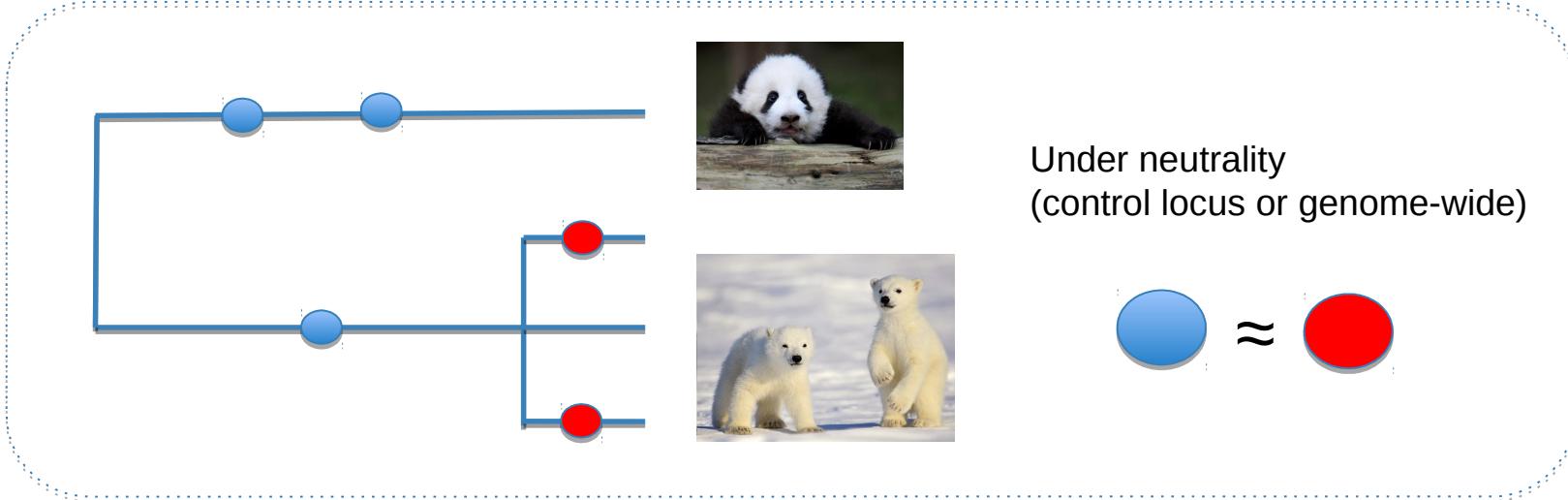
Positive selection
(or negative selection)



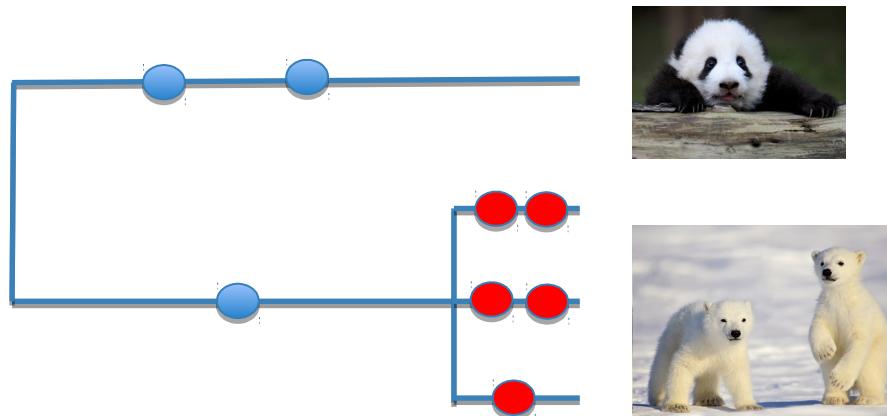
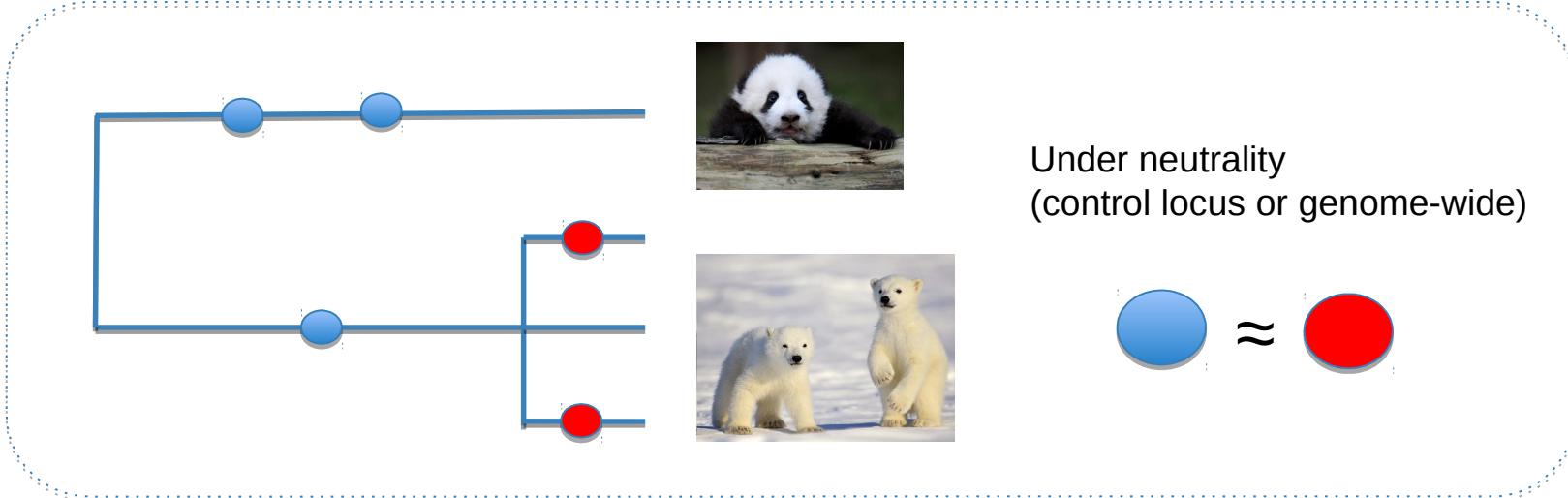
Polymorphisms and divergence



Polymorphisms and divergence



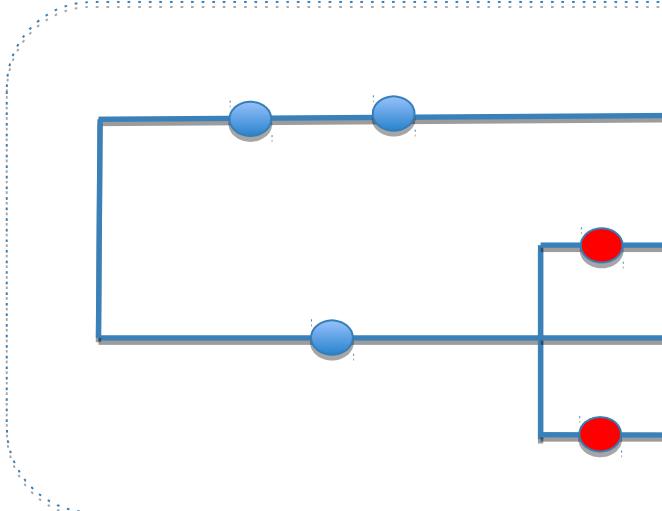
Polymorphisms and divergence



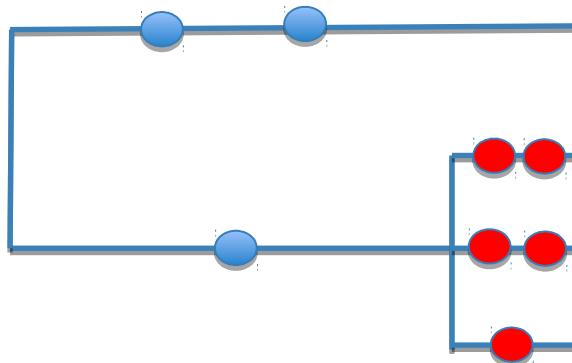
??? selection



Polymorphisms and divergence



Under neutrality
(control locus or genome-wide)



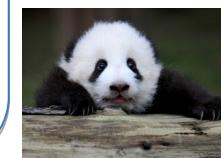
Balancing selection
(or relaxation of functional constraints)



HKA test

Hudson-Kreitman-Aguadè (HKA, Hudson et al. 1987) test

		Under investigation	Control / Neutrally evolving
		Gene1 G1-poly	Gene2 G2-poly
Polymorphic			
	Fixed	G1-fixed	G2-fixed

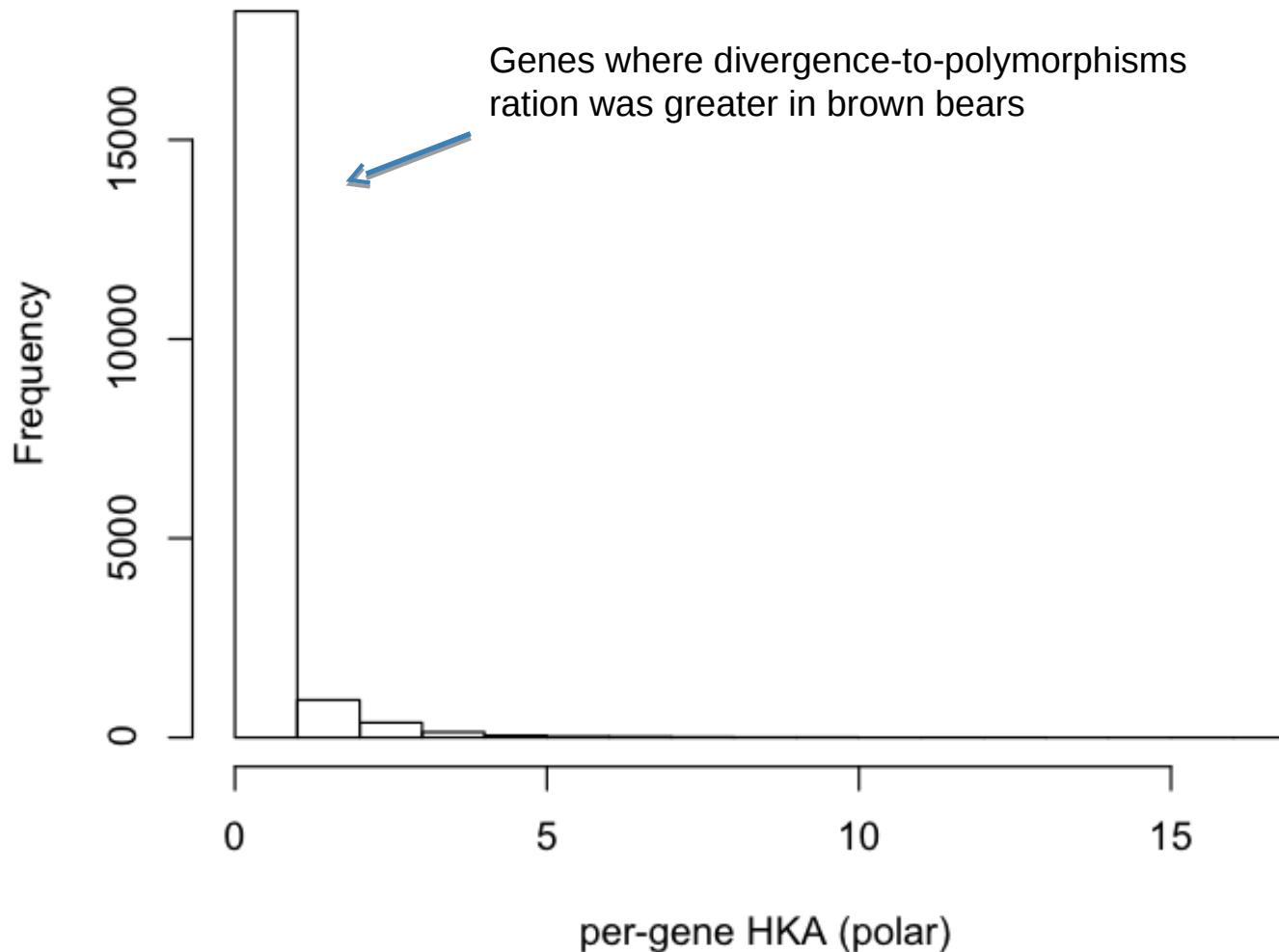
 

Contingency table 2x2: chi-square test

HKA test



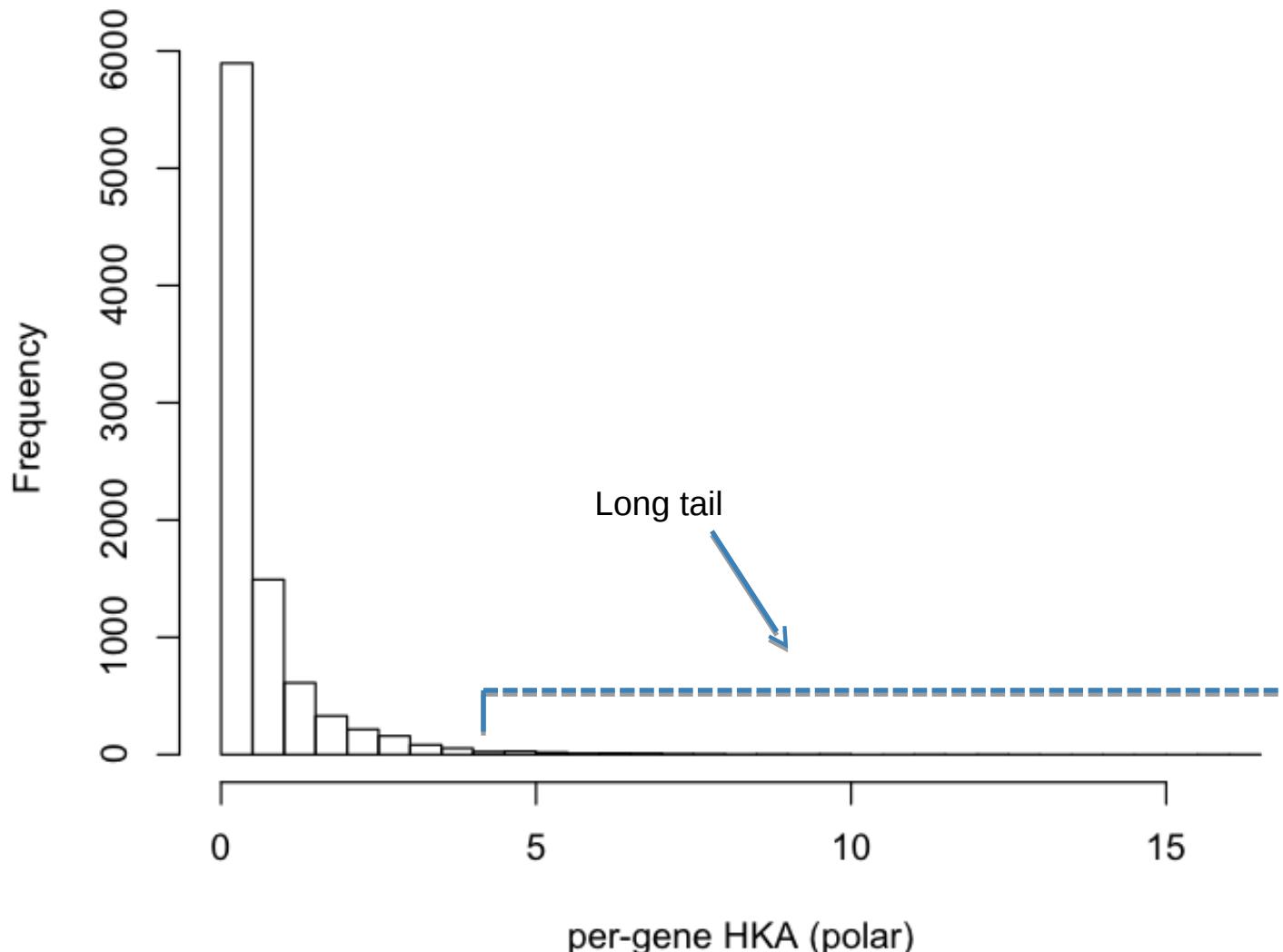
Empirical distribution (-log10)



HKA test



Empirical distribution (-log10)



Coding polymorphisms

Substitutions in the coding regions may alter (non-synonymous) or not (synonymous) the protein sequence:



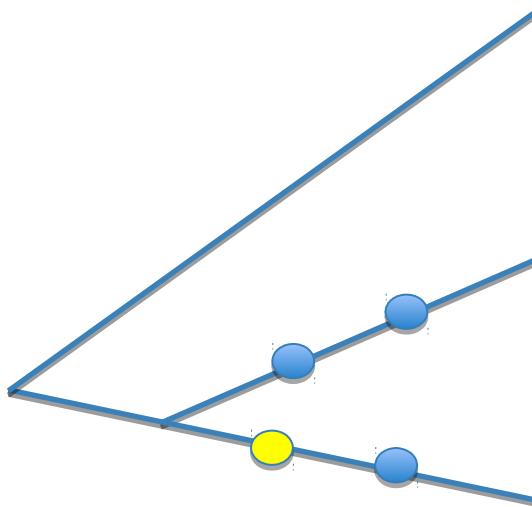
Non-synonymous



Synonymous

dN/ds

Comparison between rate of **nonsynonymous** substitutions and rate of **synonymous** substitutions (also called K_a/K_s or ω)



Non-synonymous



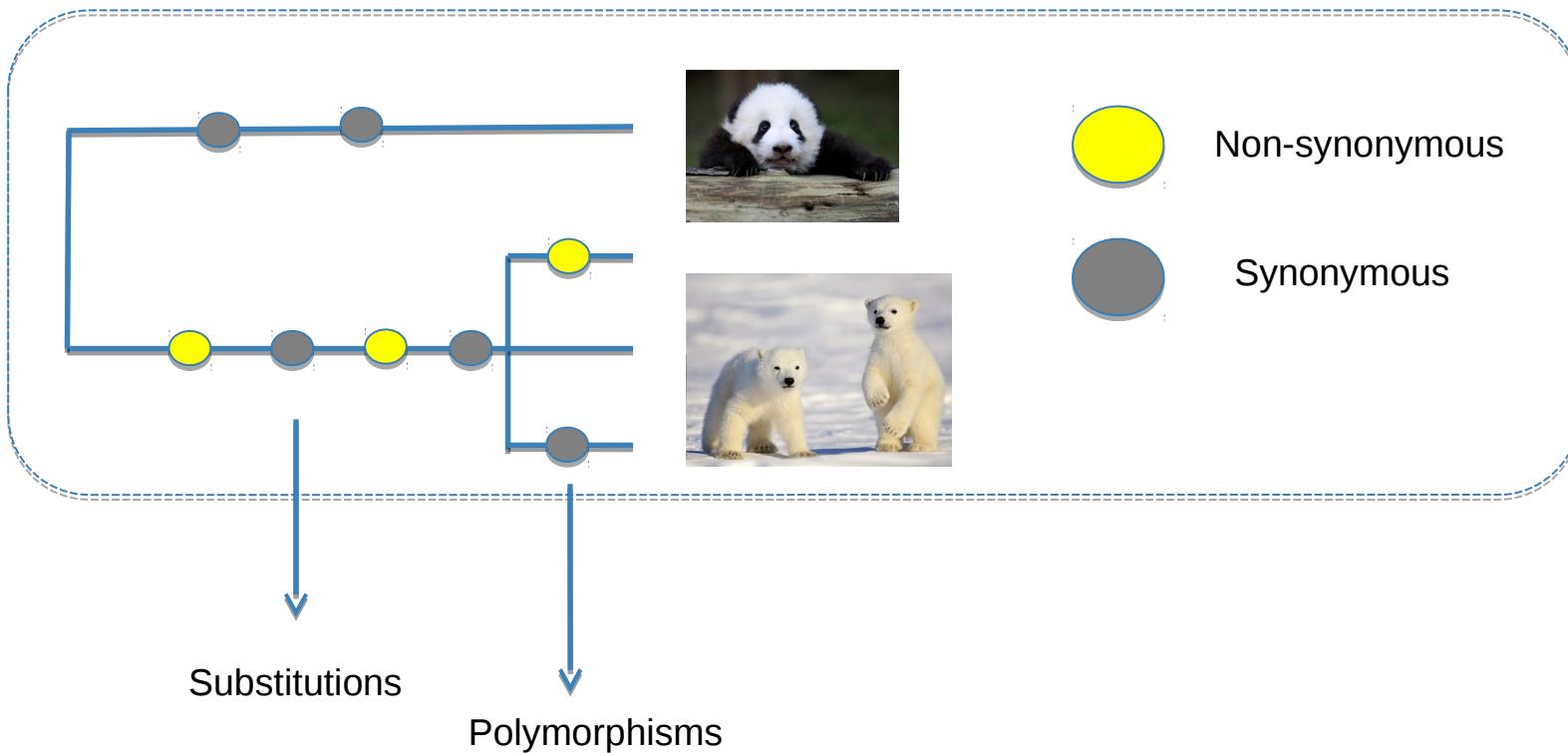
Synonymous

Positive selection
(novel proteins)
 $K_a/K > 1$

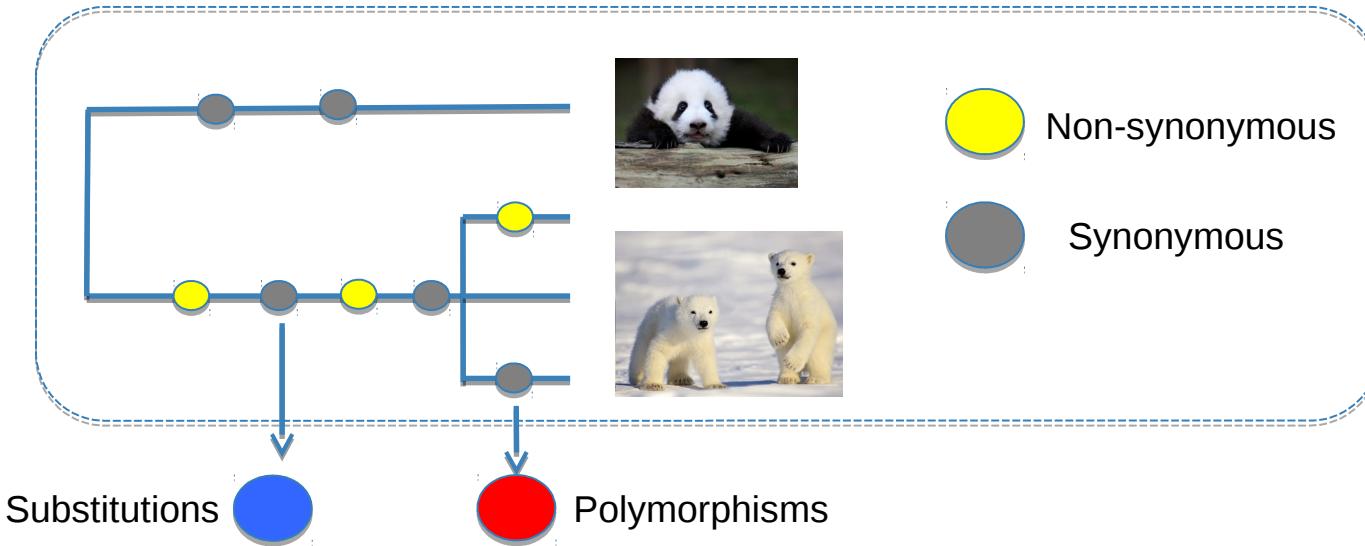
Negative selection
(against deleterious mutations)
 $K_a/K < 1$

Coding polymorphisms and divergence

Compare the amount of variation within a species to the divergence between species at non-synonymous and synonymous variants.

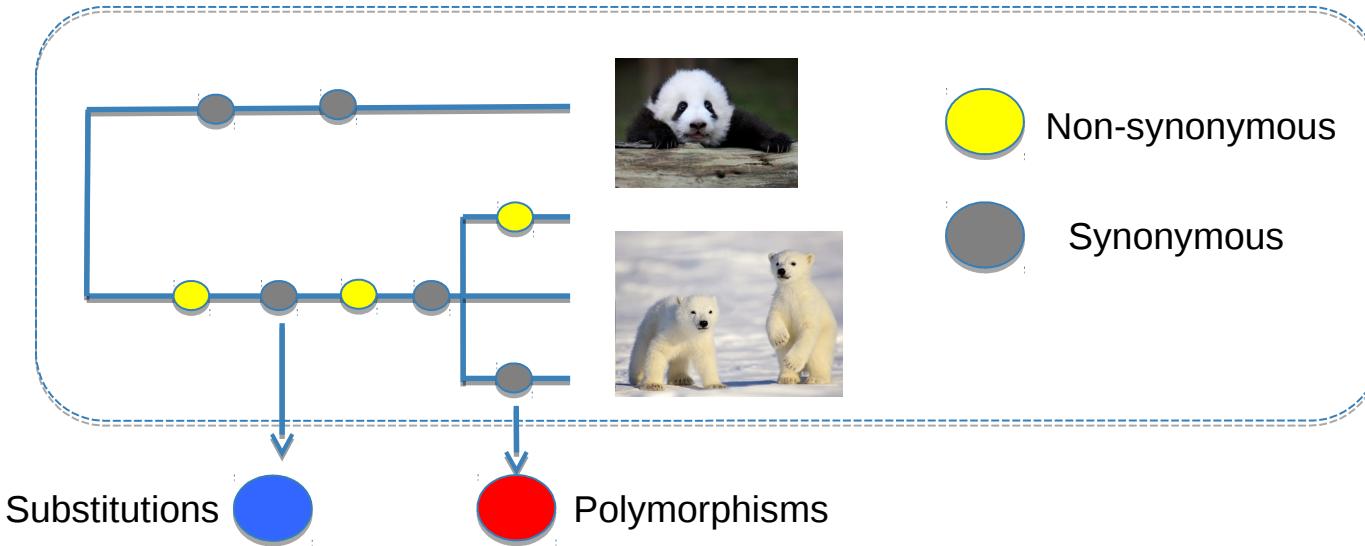


MK (McDonald–Kreitman) test



	Non-synonymous	Synonymous
Divergence		
Polymorphisms		

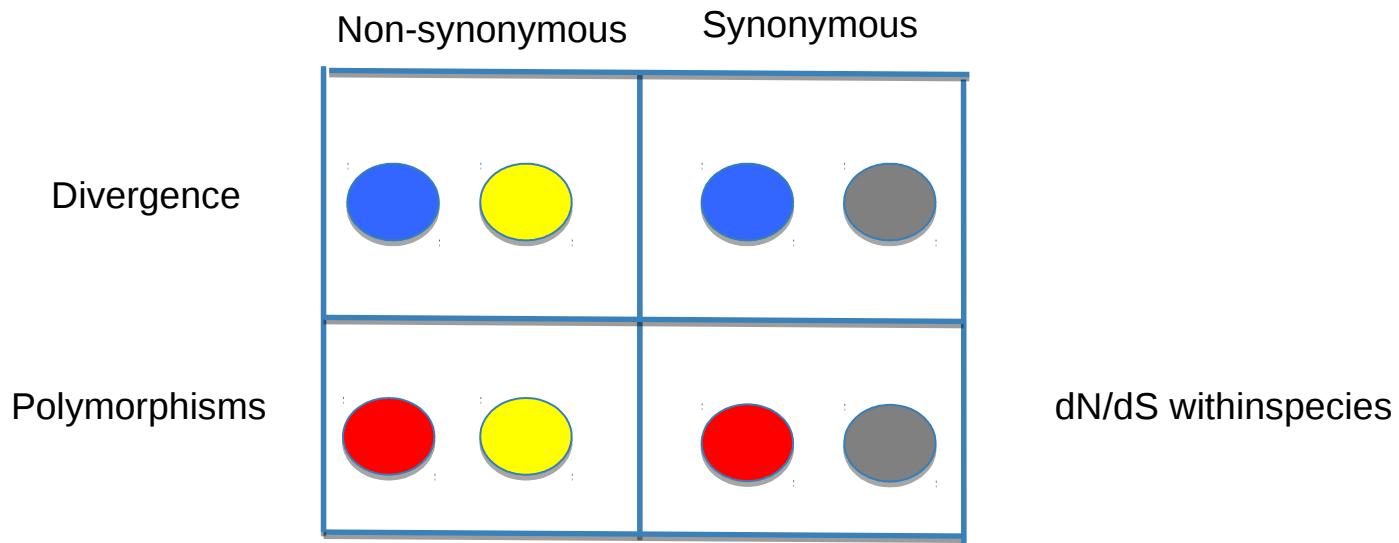
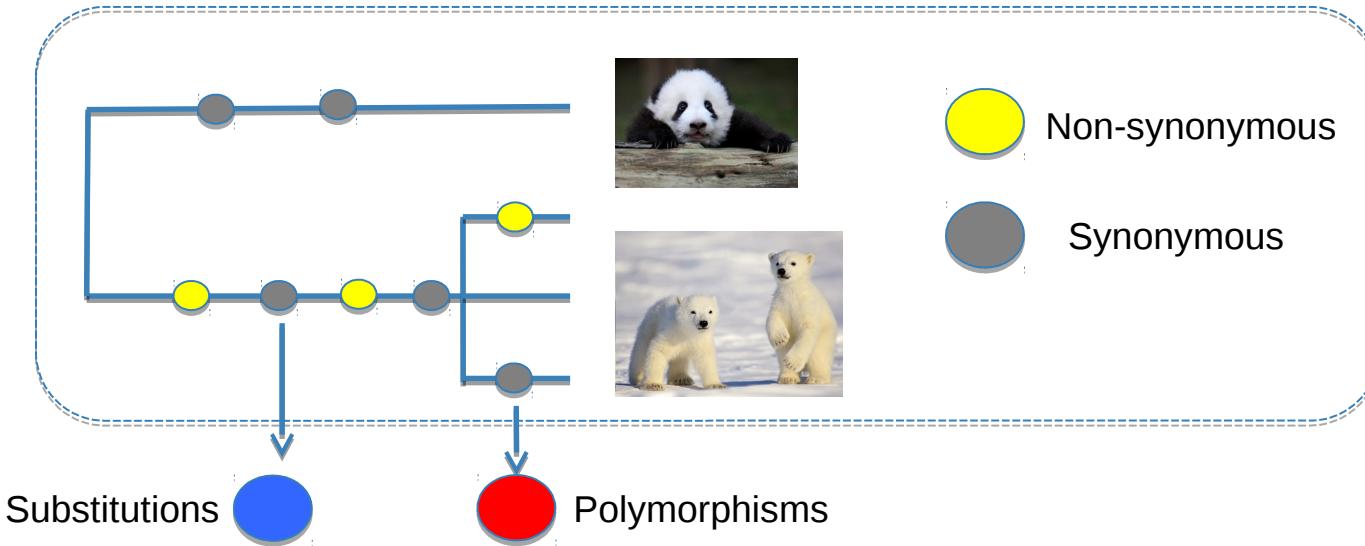
MK (McDonald–Kreitman) test



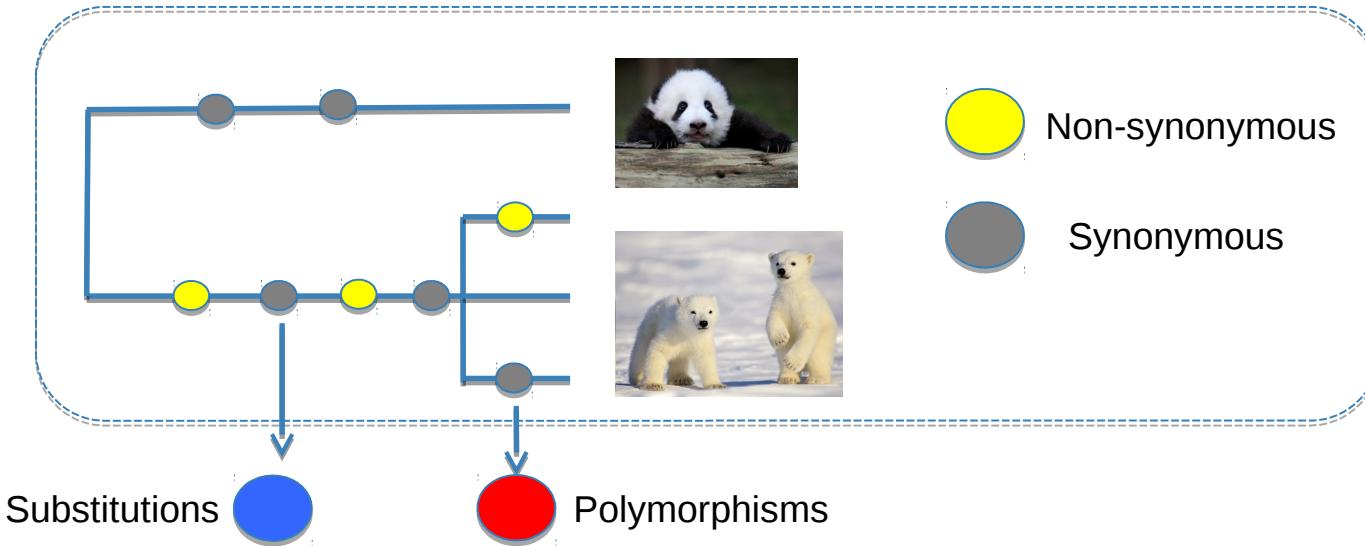
	Non-synonymous	Synonymous
Divergence	Substitutions	Polymorphisms
Polymorphisms		

dN/dS between species

MK (McDonald–Kreitman) test



MK (McDonald–Kreitman) test



	Non-synonymous	Synonymous
Divergence	Blue circle (Substitution)	Yellow circle (Substitution)
Polymorphisms	Red circle (Polymorphism)	Red circle (Polymorphism)

2x2 contingency table

Summary

- Methods to detect signatures of selection are grouped based on:
 - time of selection
 - summary statistics used
- Assessing statistical significance through empirical or expected distributions
- Investigating complex models of selection and adaptation