

# *Inference of natural selection from sequencing data*

(at the intra-species level)

Matteo Fumagalli

## Intended Learning Outcomes

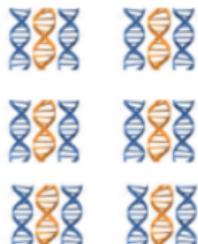
At the end of this session you will be able to:

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

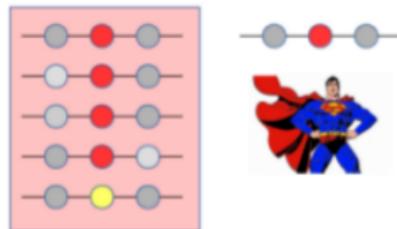
## Demographic history



## Whole genome sequencing



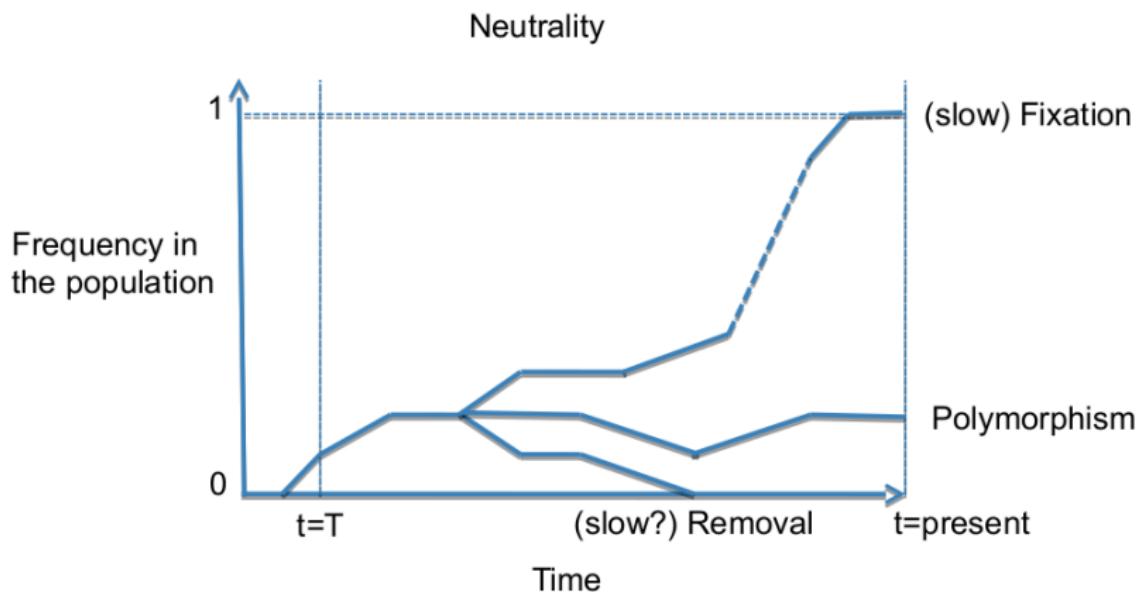
## Natural selection



# Natural selection

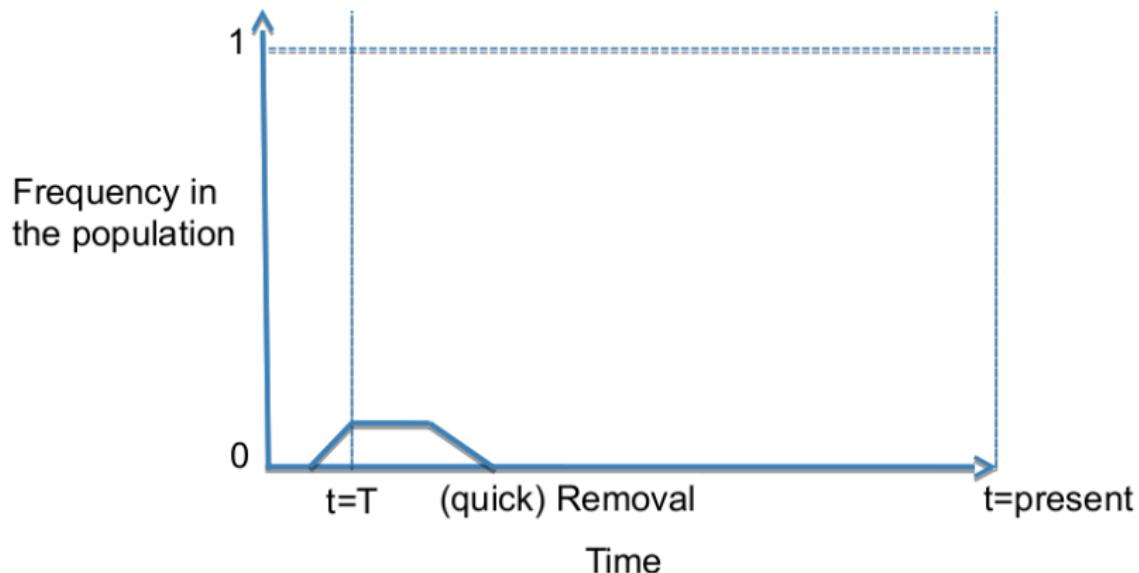
- Heritable traits that increase the fitness of the become more common.
- Sites targeted by natural selection are likely to harbour functionality.
- Mutations arise (almost) randomly and evolve according to their effect on the fitness of the carrier.

# Allele frequency trajectory

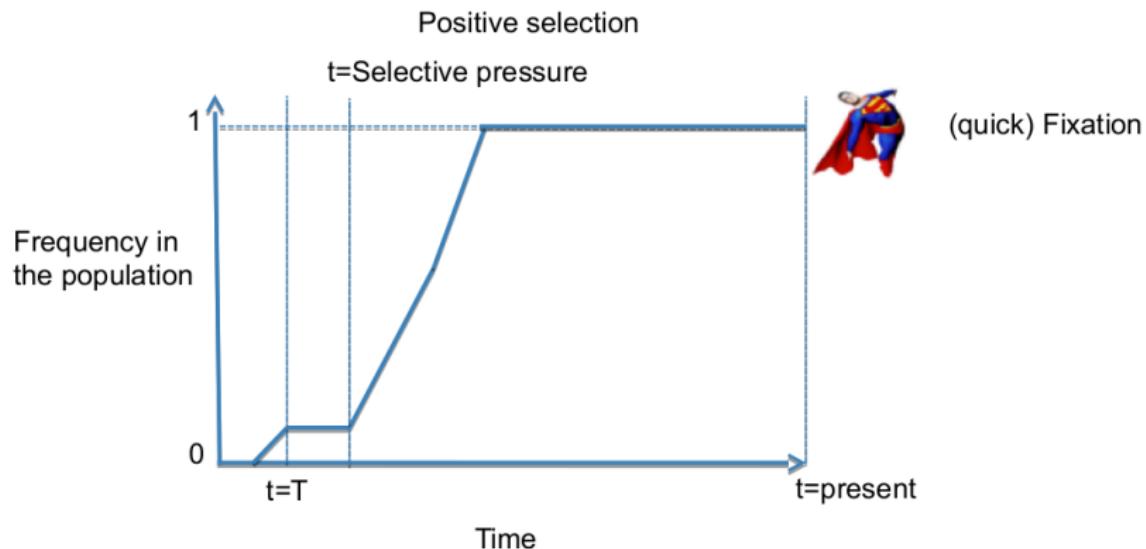


# Allele frequency trajectory

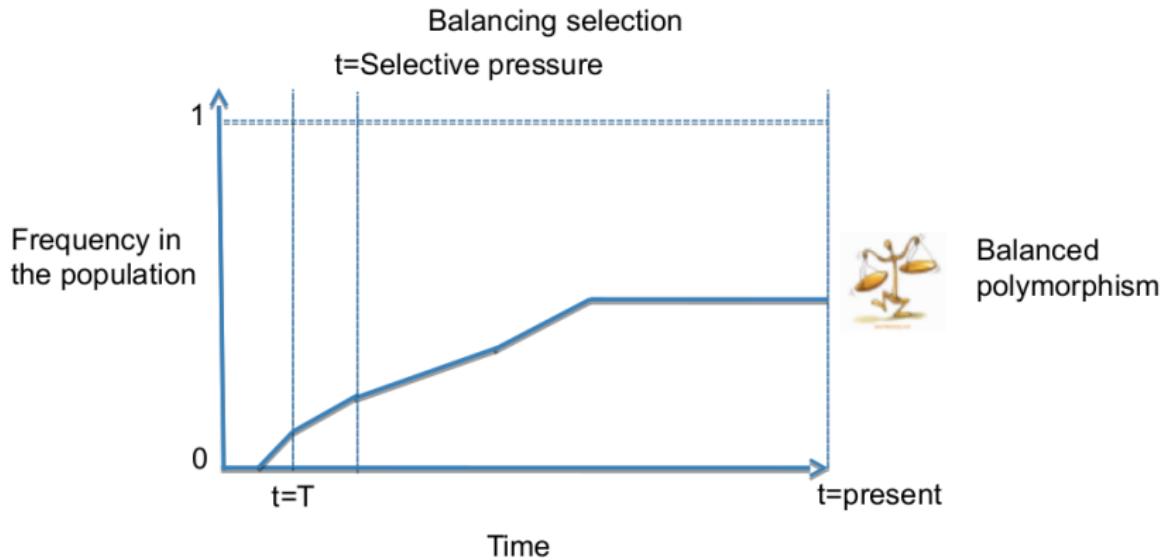
Negative selection



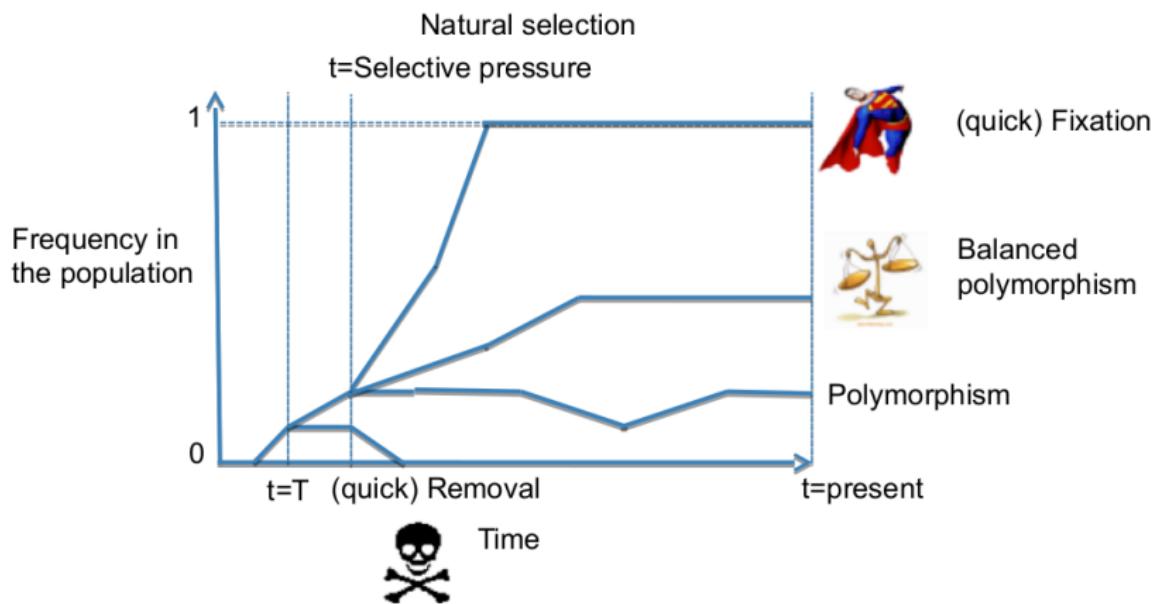
# Allele frequency trajectory



# Allele frequency trajectory



# Allele frequency trajectory



# Allele frequency trajectory - summary

Effect of selection on alleles:

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

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

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

(slide stolen from Anders)

## Testing for natural selection

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

## Testing for natural selection

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

- perform selection experiments
- use external information: candidate genes/biological knowledge, functional categories, association to phenotypes

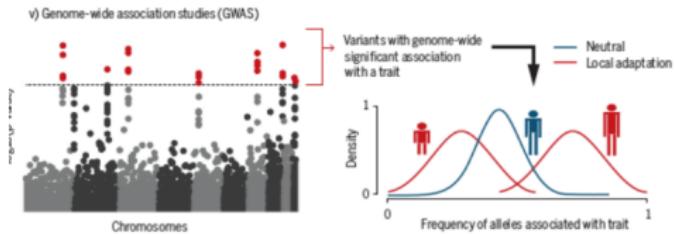
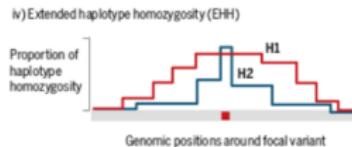
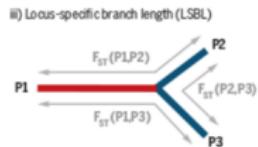
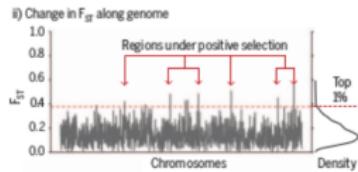
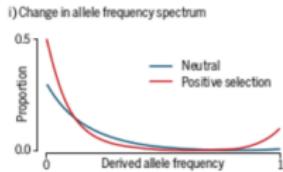
## Testing for natural selection

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

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

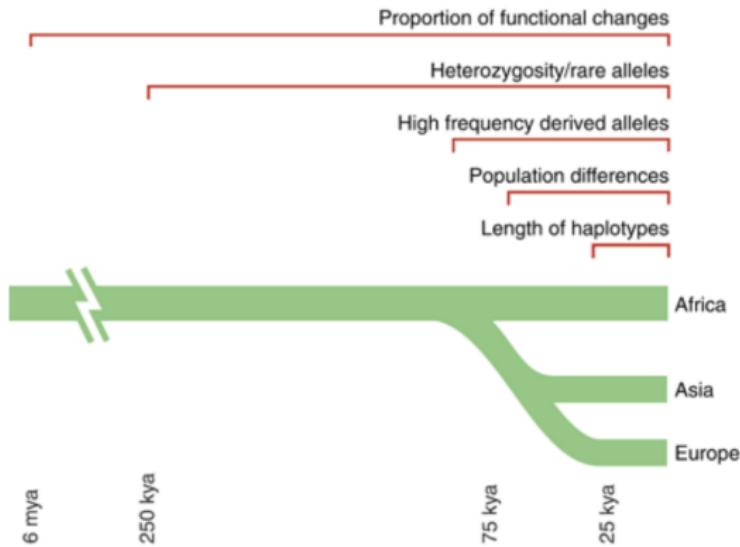
(slide stolen from Anders)

# Common methods to detect selection



(slide stolen from Anders)

# Detect recent selection within species / using shared variation

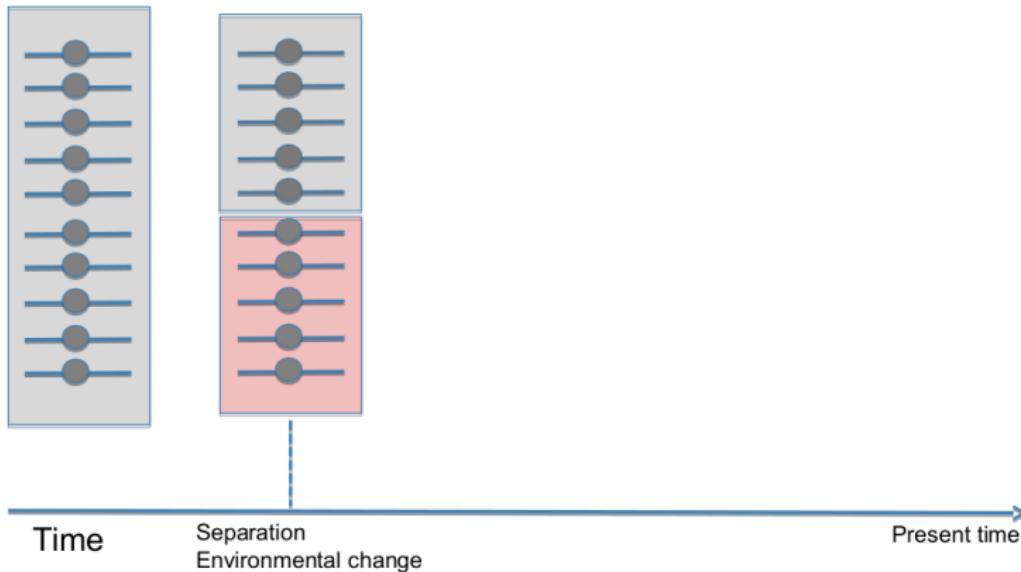


Sabeti et al. 2006 Science

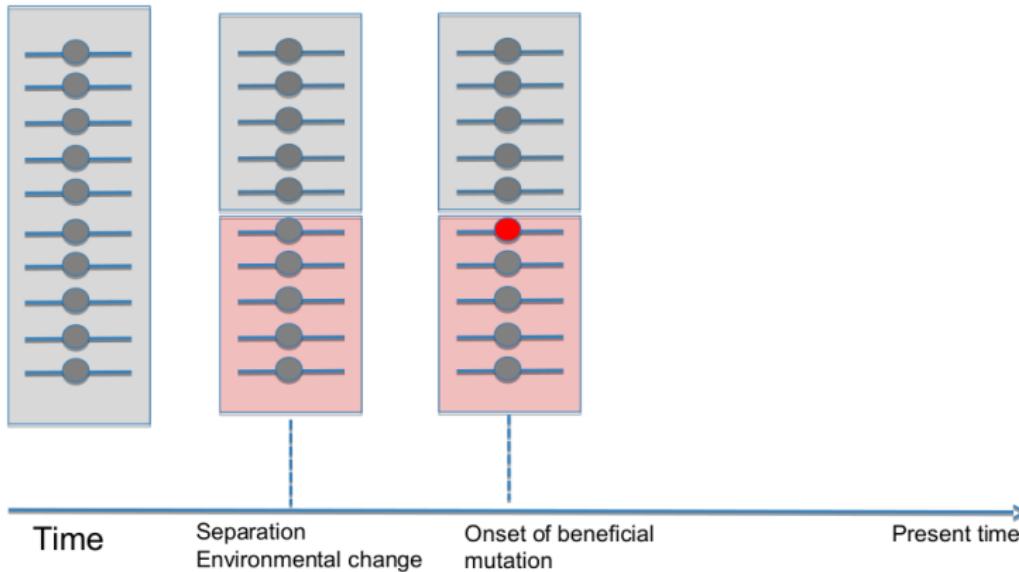
# Allele frequency differentiation



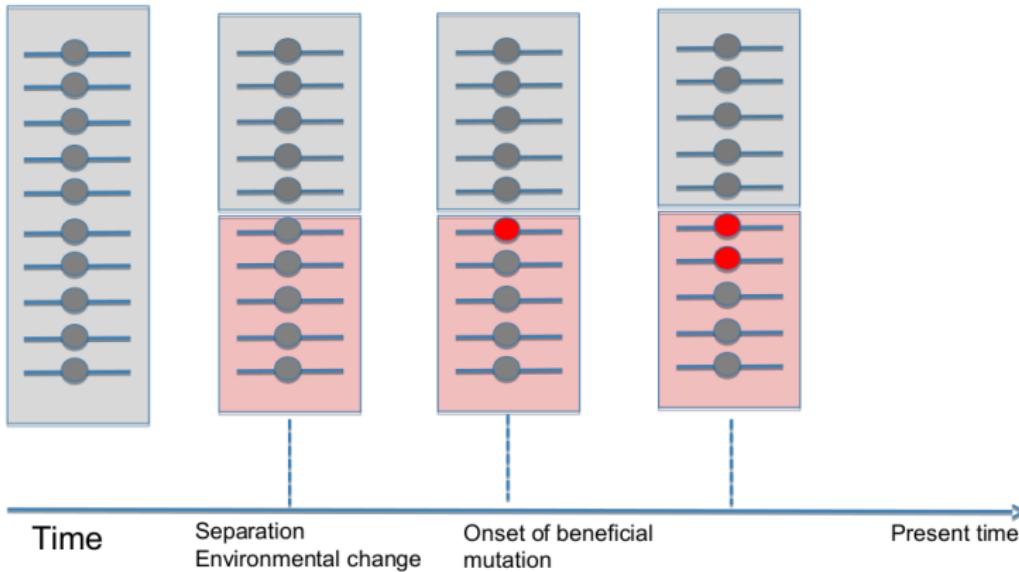
# Allele frequency differentiation



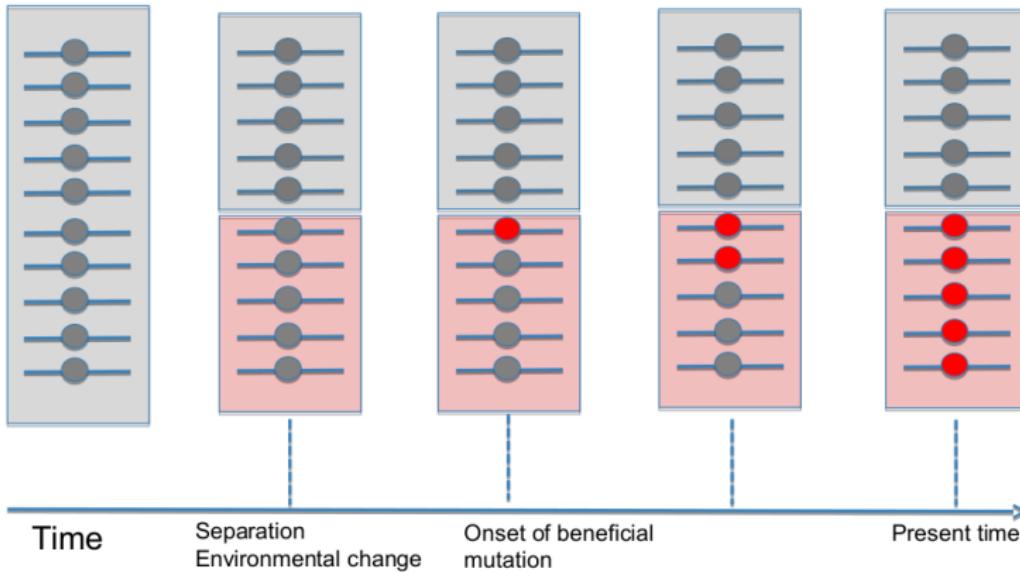
# Allele frequency differentiation



# Allele frequency differentiation



# Allele frequency differentiation



$$F_{ST}$$

Common measure for quantifying population subdivision.

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

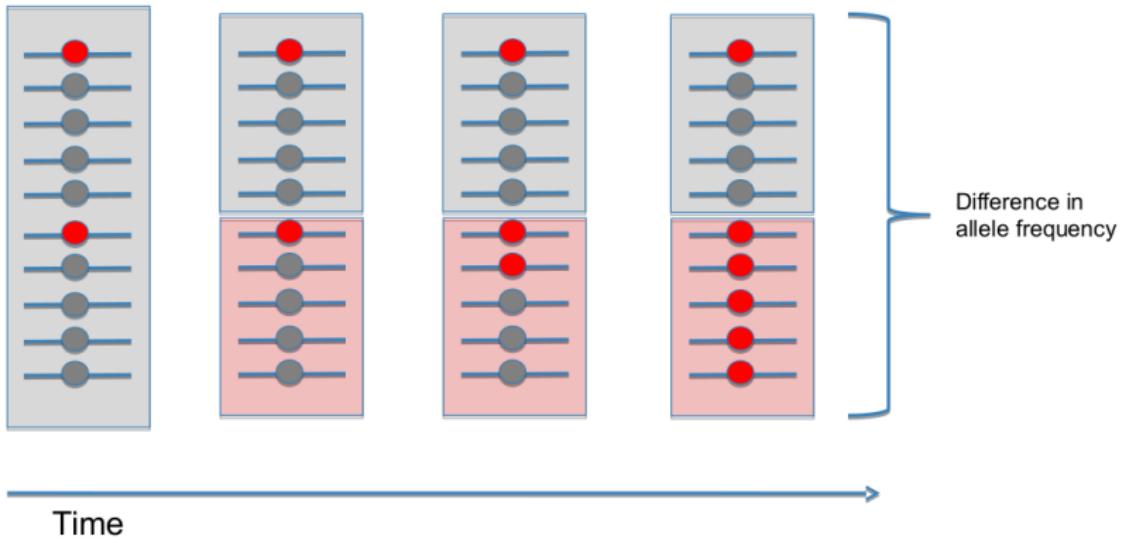
$H_B$ : between populations

$H_W$ : average within populations

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

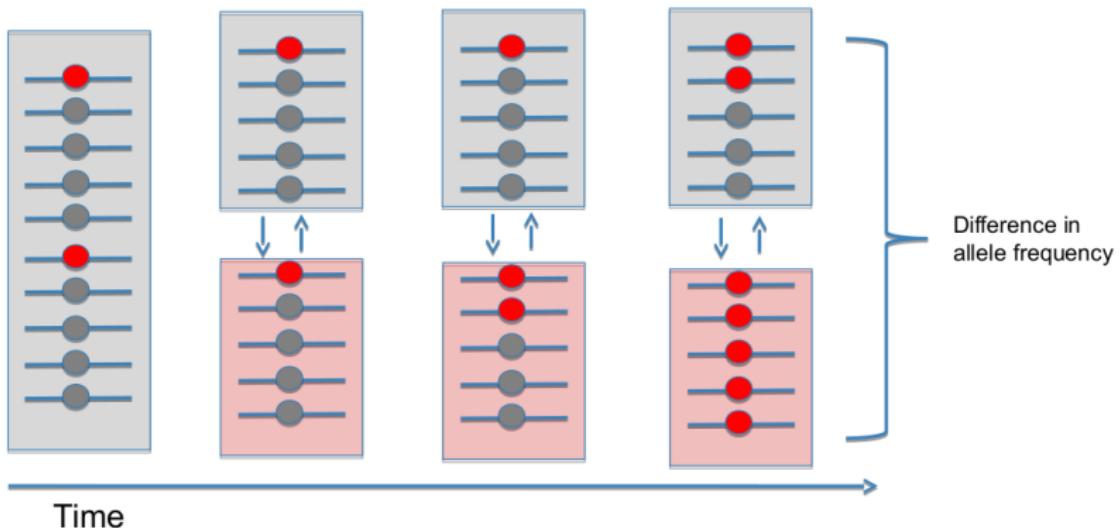
# Allele frequency differentiation

From standing variation

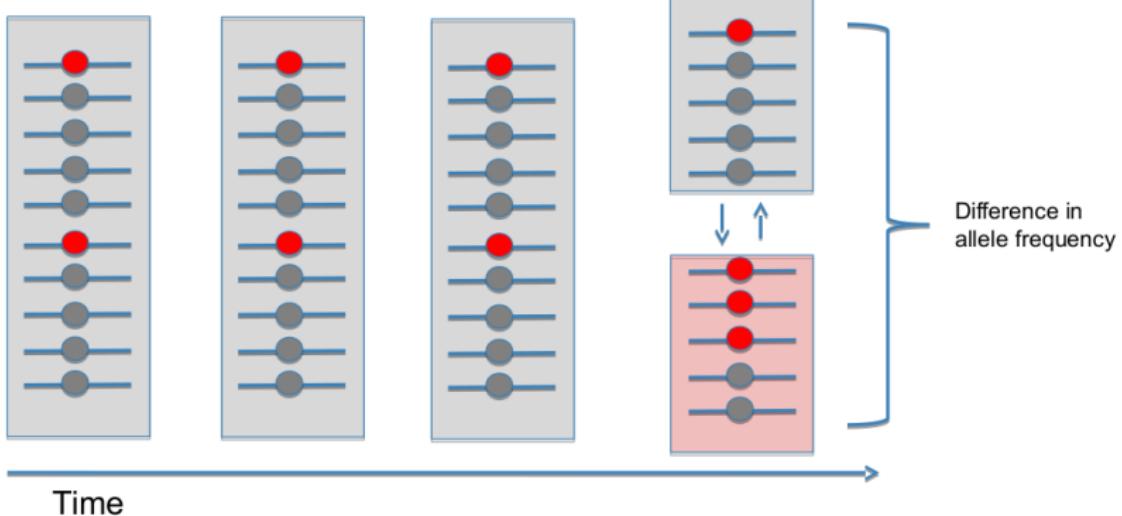


# Allele frequency differentiation

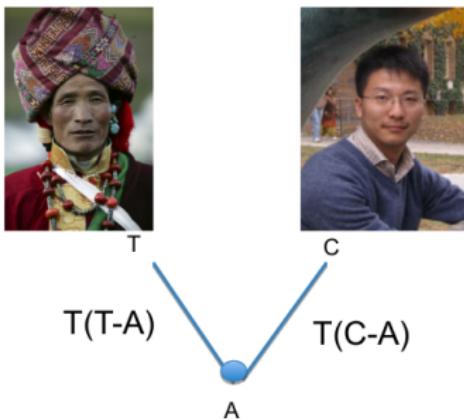
With migration



### With recent divergence



# Population genetic differentiation



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

# Population genetic differentiation

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



T

T(T-A)



C

?



T

T(T-A)



C

T(C-A)

A

T(C-A)

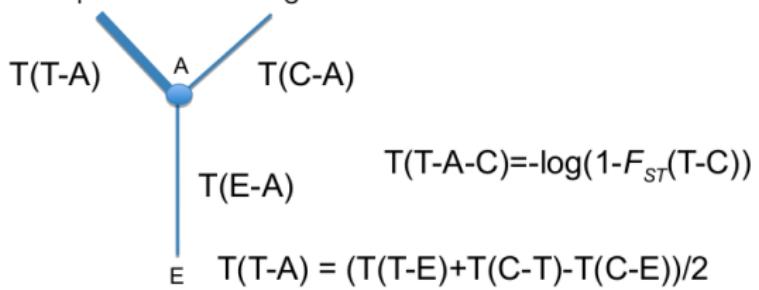
# Population genetic differentiation



T

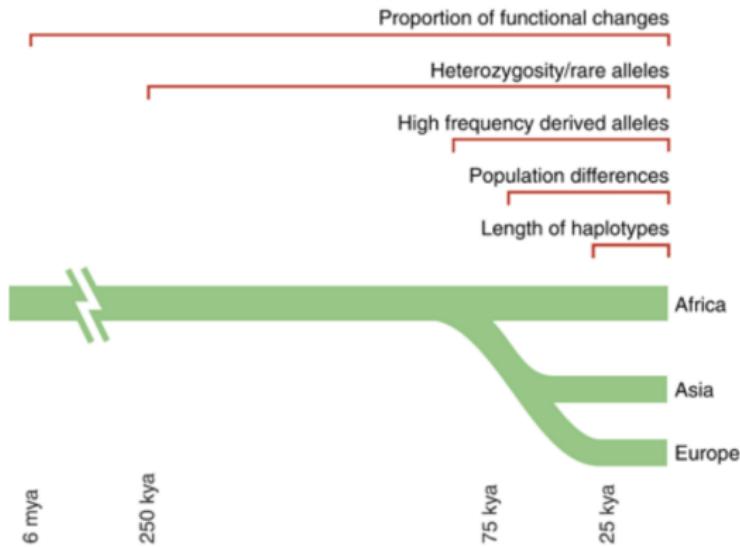


C



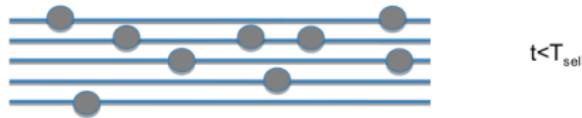
Population branch statistic (PBS) (see practical)

# Detect recent selection within species / using shared variation

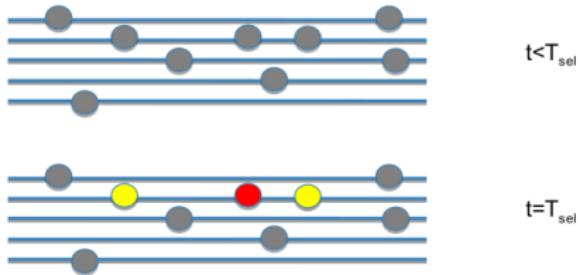


Sabeti et al. 2006 Science

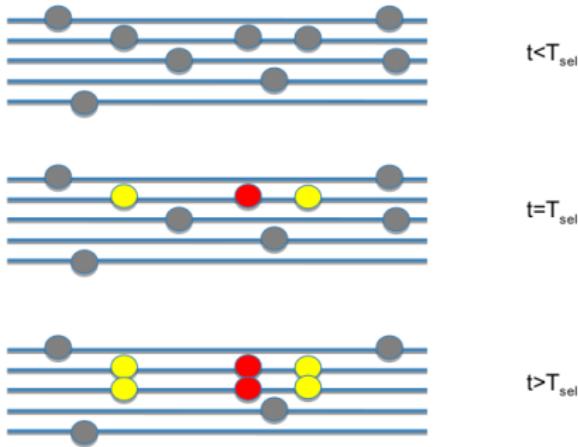
## Positive selection: effect on haplotypes



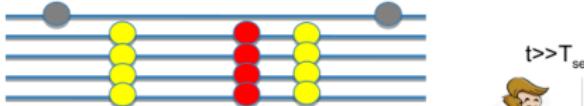
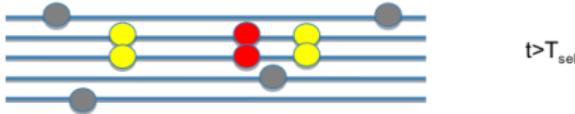
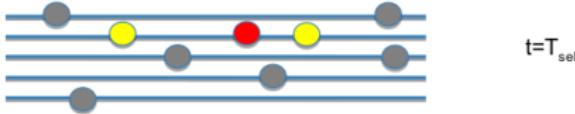
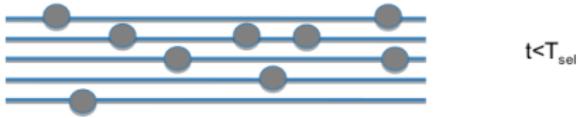
## Positive selection: effect on haplotypes



# Positive selection: effect on haplotypes



# Positive selection: effect on haplotypes



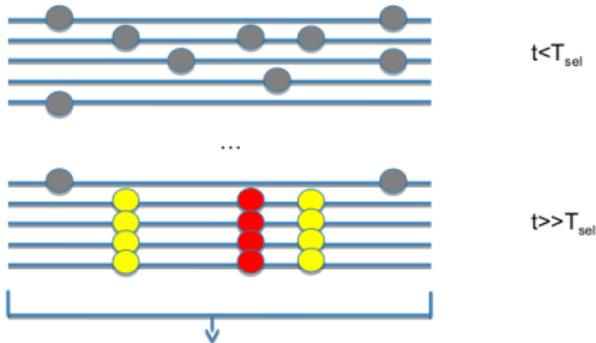
Selective sweep



Genetic hitch-hiking



# Positive selection



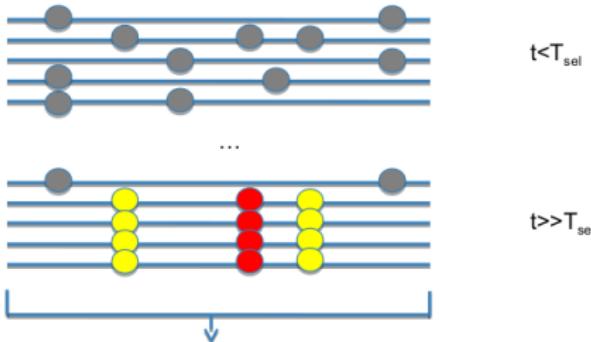
- Reduction of polymorphisms levels  
(e.g. from 7 to 5 SNPs)

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

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

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

# Positive selection

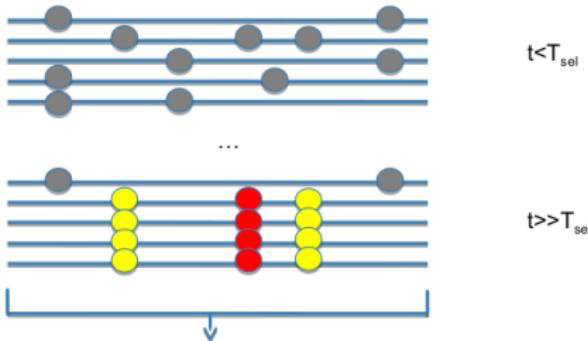


- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

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

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

# Positive selection



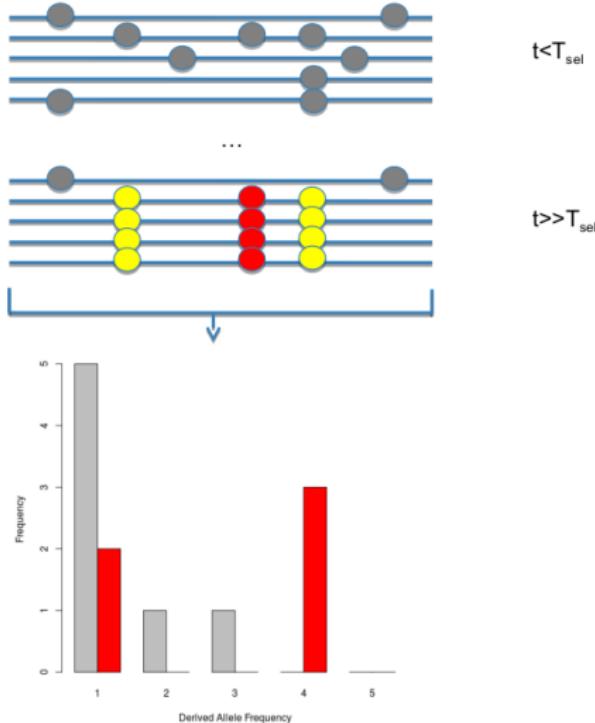
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

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

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

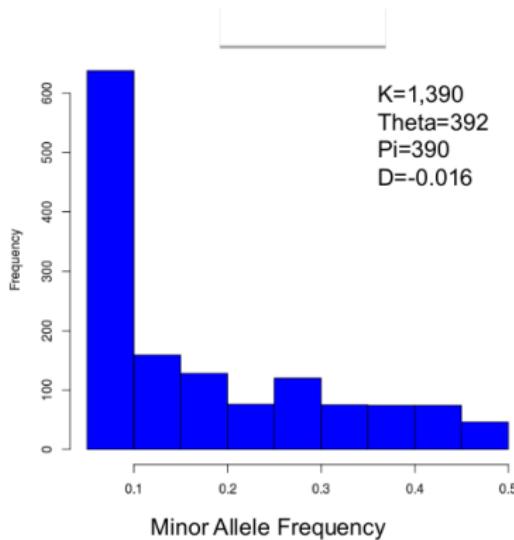
$D < 0$  is suggestive of an excess of low-frequency variants

# The Site Frequency Spectrum

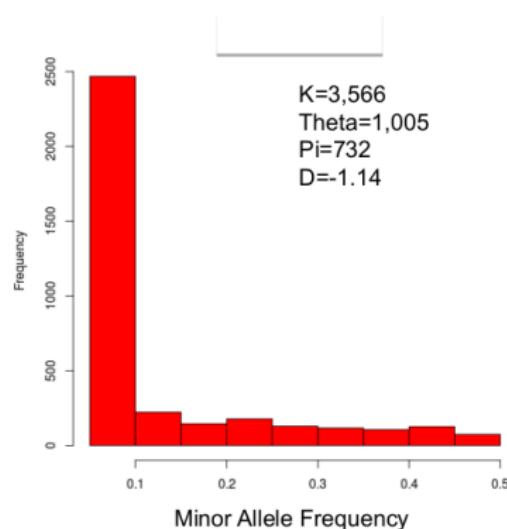


# Confounding factor

n=20; L=500kbp; no selection

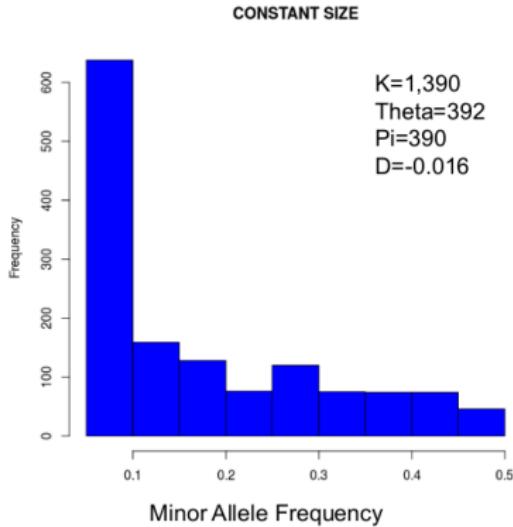


n=20; L=500kbp; no selection

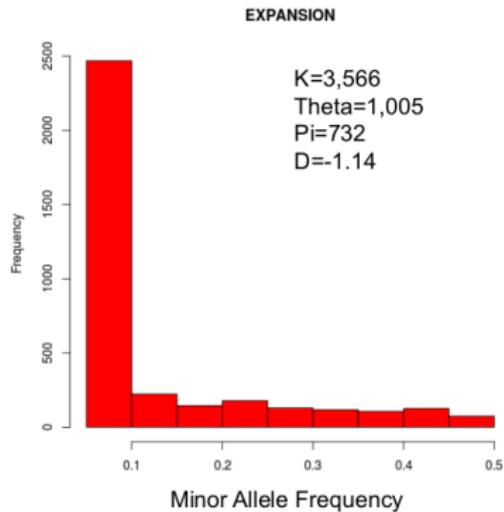


# Demography matters!

n=20; L=500kbp; no selection



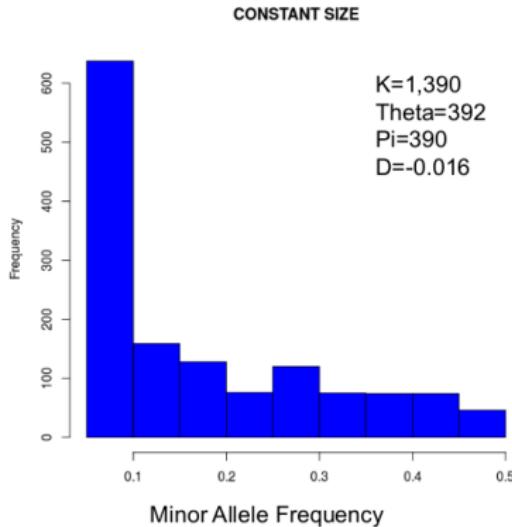
n=20; L=500kbp; no selection



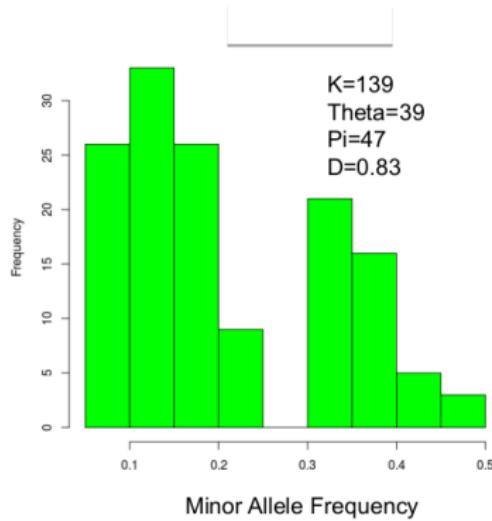
- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

# Demography matters?

n=20; L=500kbp; no selection

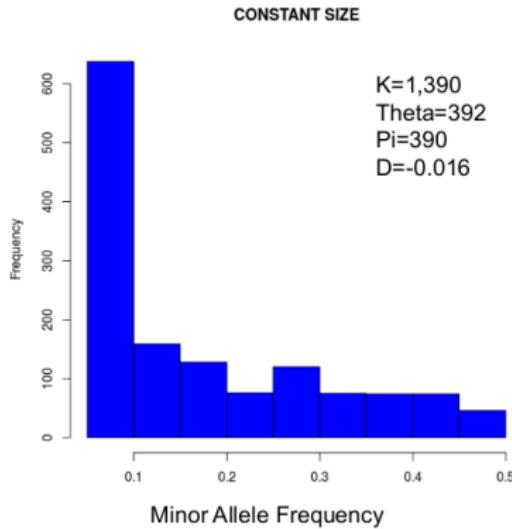


n=20; L=500kbp; no selection

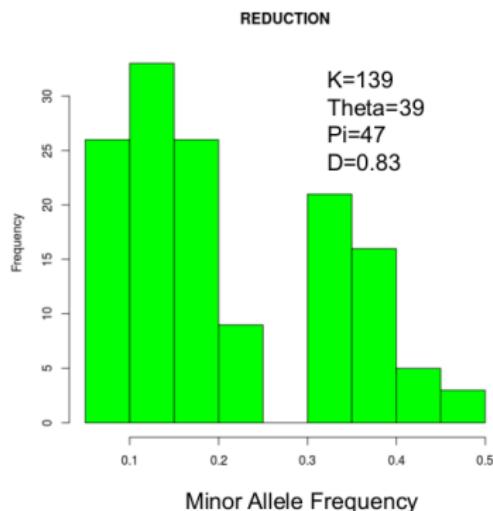


# Demography matters!

n=20; L=500kbp; no selection



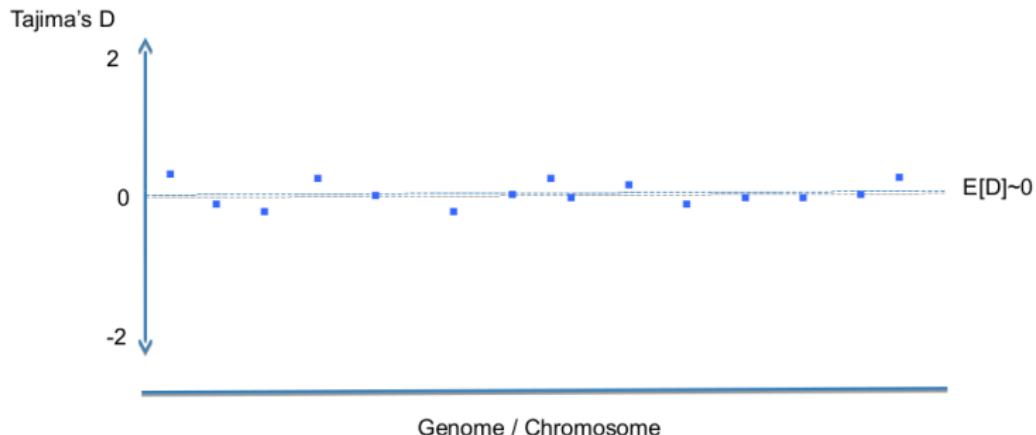
n=20; L=500kbp; no selection



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

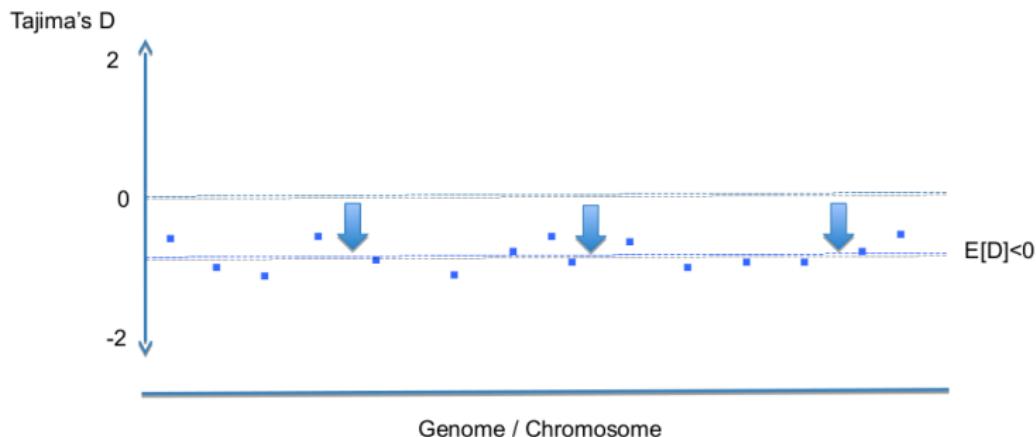
# How to take neutral confounding factors into account?

Under constant population size:



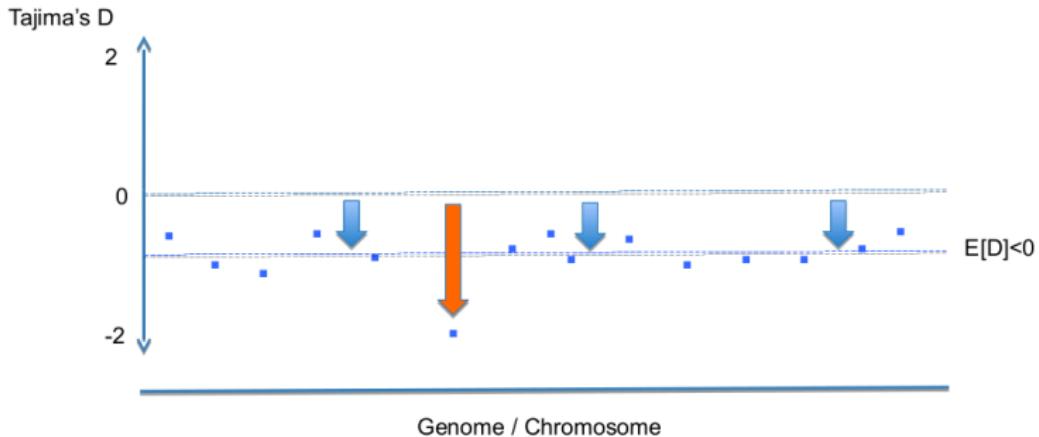
# How to take neutral confounding factors into account?

Under expanding population size:



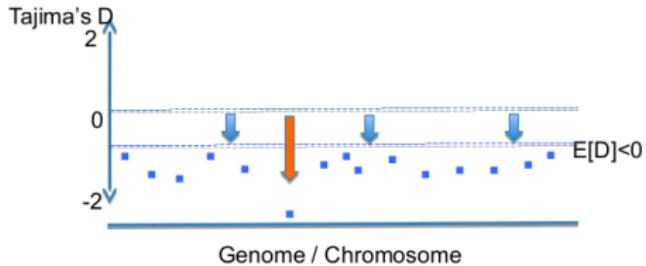
# How to take neutral confounding factors into account?

Under expanding population size and positive selection:

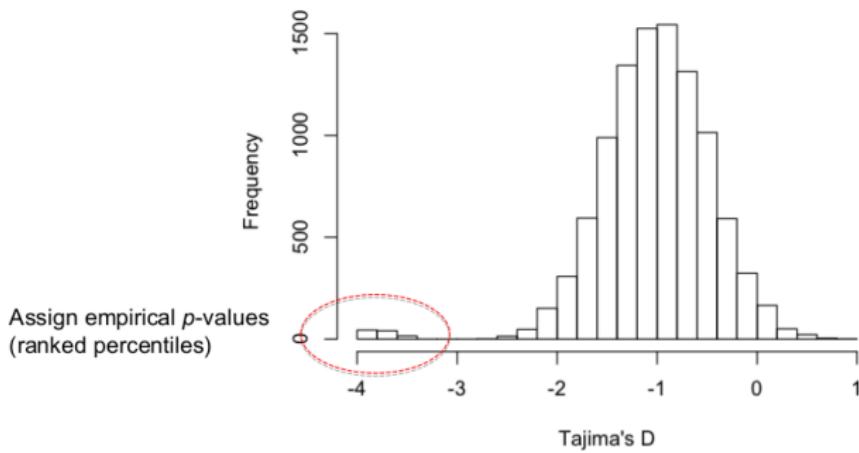


- Demography affects all loci equally, while selection changes local patterns

# Outlier approach

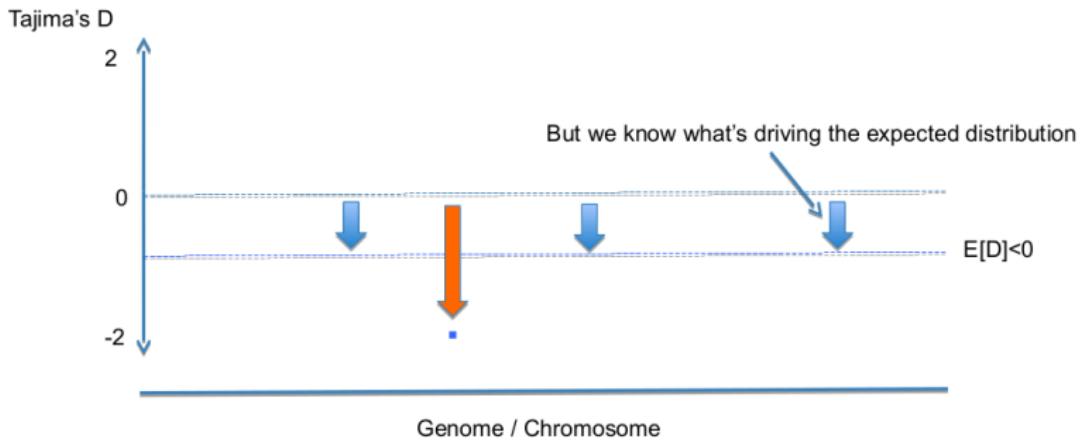


Empirical distribution



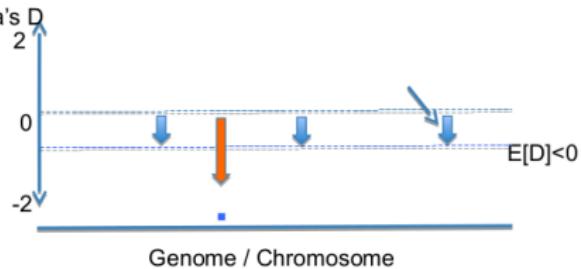
# How to take neutral confounding factors into account?

Under expanding population size and positive selection:

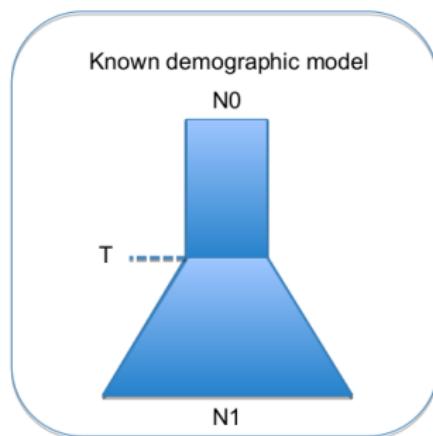
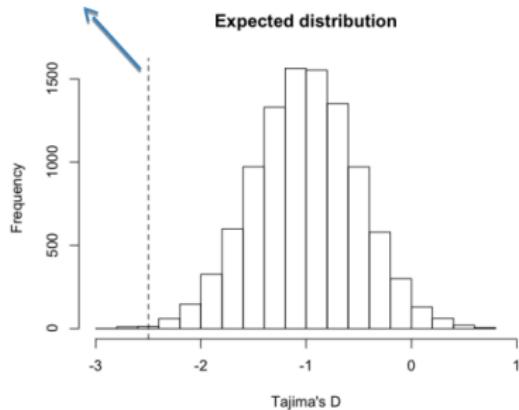


- Demography affects all loci equally, while selection changes local patterns  
What should we do if we don't have genome-wide data?

# Simulations-based approach

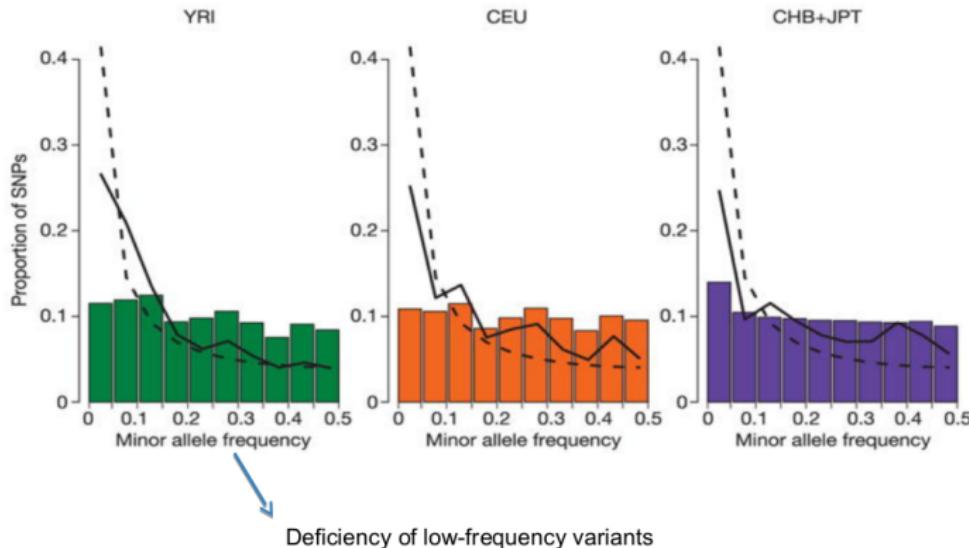


Assign  $p$ -values  
(based on ranked percentile of observed value)



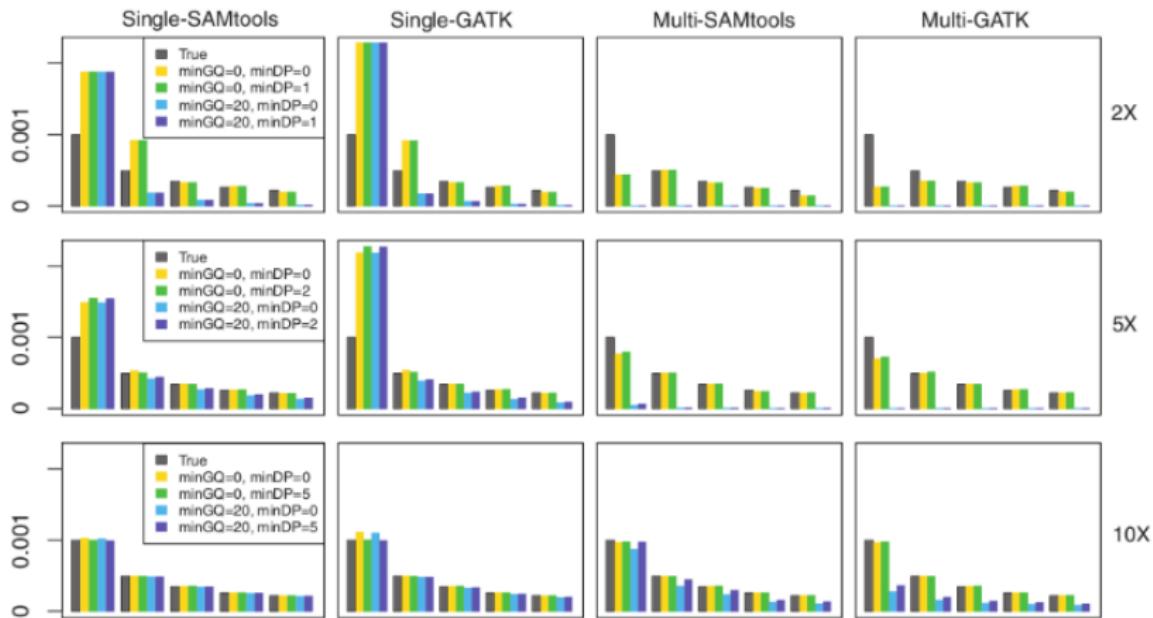
# Experimental design matters?

The effect of ascertainment bias



HapMap Consortium. Nature 2005

# Effect of low-depth sequencing



(slides stolen from Anders)

# Sample allele frequency posterior probabilities

With 6 chromosomes (3 diploids)

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

- SNP calling

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

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

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

# Sample allele frequency posterior probabilities

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

- SNP calling

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

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

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

# Nr of segregating sites

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

# Nr of segregating sites

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

# Nr of segregating sites

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

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

# Nucleotide diversity

Site 1

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

Site 2

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

Site 3

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

...

Site  $M$

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

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

$$E[D] =$$

# Nucleotide diversity

Site 1

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

Site 2

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

Site 3

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

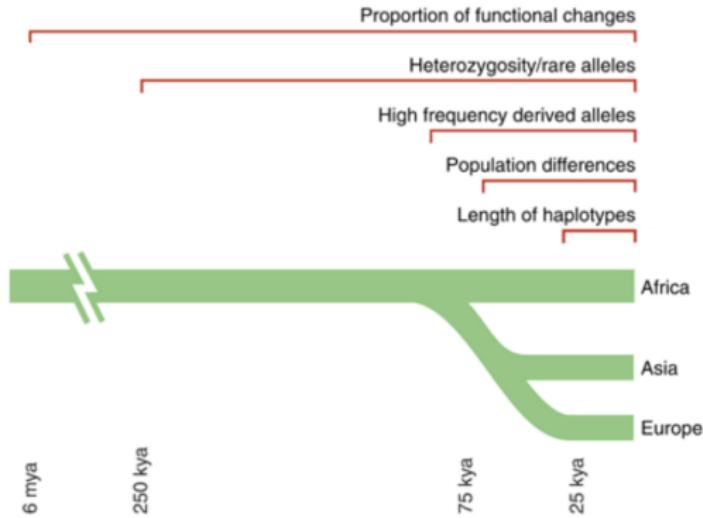
...

Site  $M$

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	...	$p(S_m=2k)$
------------	------------	------------	------------	-----	-------------

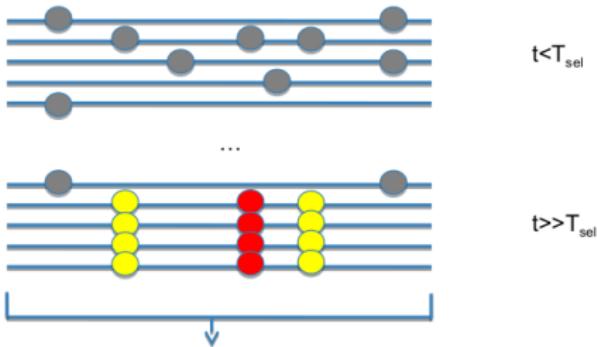
$$E[D] = \sum_{m=1}^M \sum_{j=0}^{2k} 2 \binom{i}{2k} \binom{2k-i}{2k} p(S_m = i)$$

# Detect recent selection within species / using shared variation



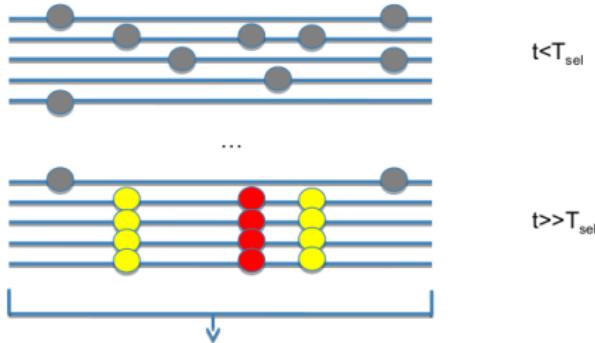
Sabeti et al. 2006 Science

# Positive selection



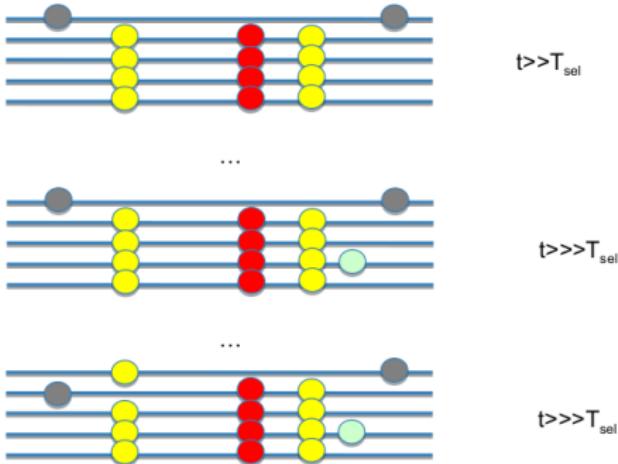
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- ?

# Positive selection



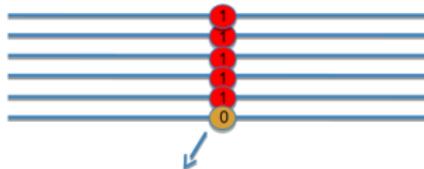
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- Extended haplotype homozygosity / Extended LD

## Extended Haplotype Homozygosity



Extended haplotype homozygosity (EHH): EHH at distance  $x$  from the core region is the probability that two randomly chosen chromosomes carry a tested core haplotype are homozygous at all SNPs for the entire interval from the core region to the distance  $x$ .

# Extended Haplotype Homozygosity

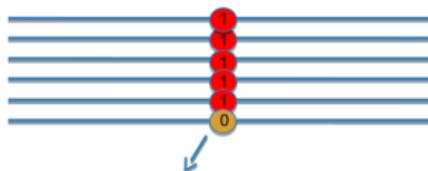


Core haplotype is 1  
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Core SNP

# Extended Haplotype Homozygosity

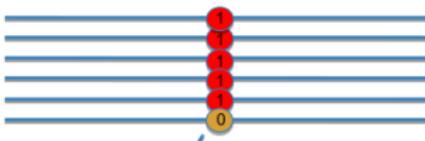


Core haplotype is 1  
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Until marker  $x_i$   
(starting from  $x_0$ )

# Extended Haplotype Homozygosity

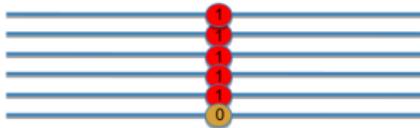


Core haplotype is 1  
(Biallelic: 0 is ancestral, 1 is derived allele)

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

Sum across all unique haplotypes  
carrying the core SNP

# Extended Haplotype Homozygosity



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

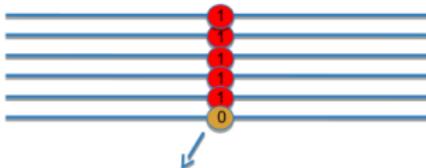
$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$n_h$  is haplotype frequency of  $h$

$n_c$  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)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \left[ \frac{n_h}{2} \right] \left[ \frac{n_h}{2} - 1 \right]$$

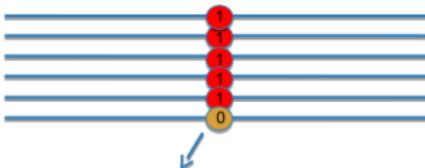
Sum across all unique haplotypes carrying the core SNP

$n_h$  is haplotype frequency of  $h$

$n_h$  is haplotype frequency of the core SNP

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

# Extended Haplotype Homozygosity



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

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

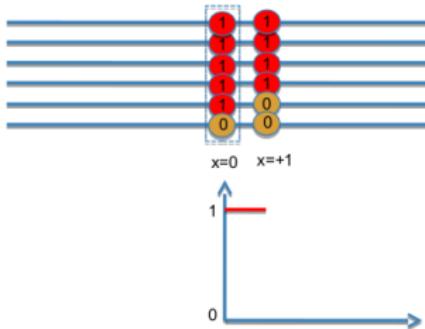
}  $n_h$  is haplotype frequency of  $h$

}  $n_c$  is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

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

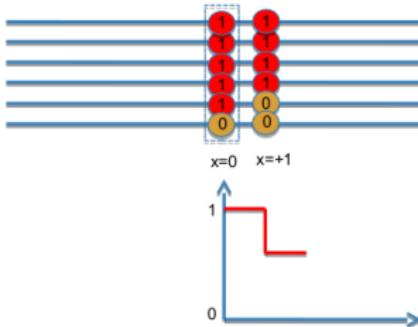
# Extended Haplotype Homozygosity



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

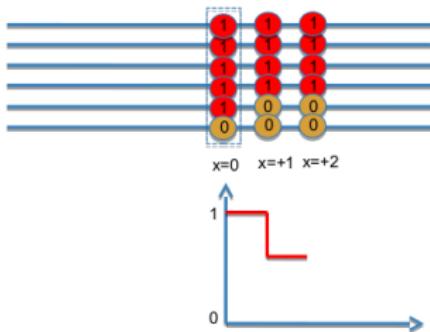
How many unique haplotypes carrying the core SNP?  
What is their frequency?

# Extended Haplotype Homozygosity



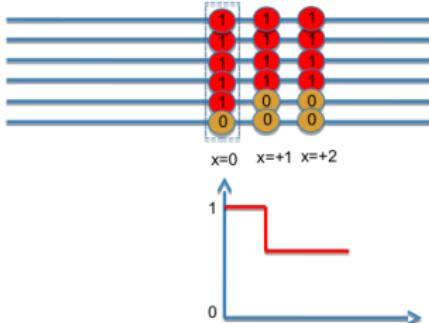
$$EHH_c(x_i = +1) = \frac{\binom{4}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{6 + 0}{10} = 0.60$$

# Extended Haplotype Homozygosity



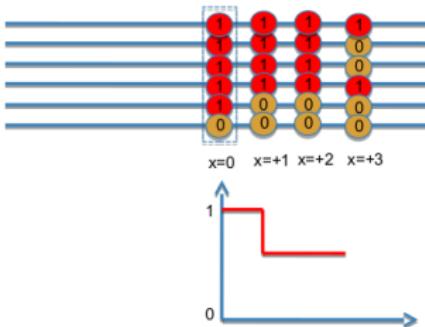
$$EHH_c(x_i = +2) = ?$$

# Extended Haplotype Homozygosity



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

# Extended Haplotype Homozygosity



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

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

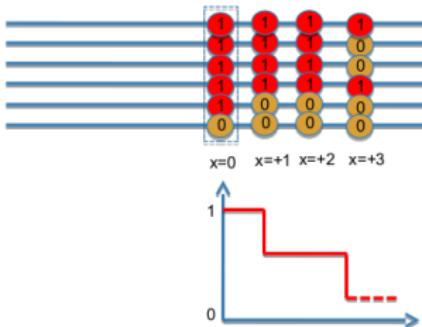
1111 with freq=2

1110 with freq=2

1000 with freq=1

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

# Extended Haplotype Homozygosity



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

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

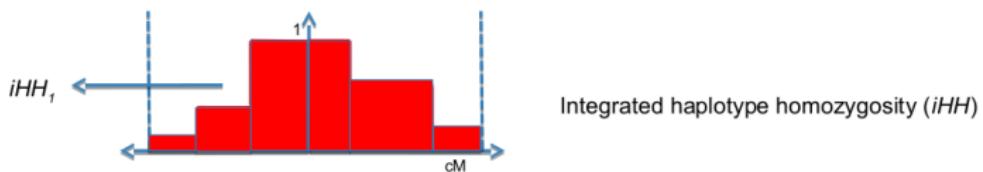
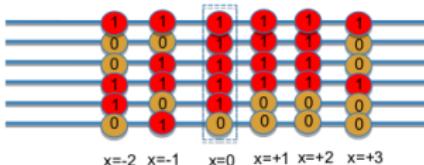
1111 with freq=2

1110 with freq=2

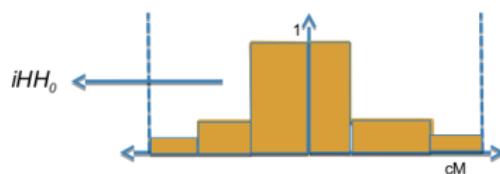
1000 with freq=1

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

# Integrated Haplotype Score



Integrated haplotype homozygosity ( $iHH$ )

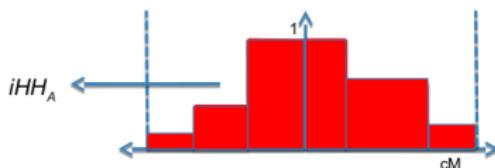
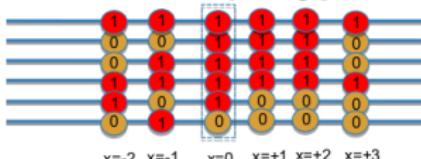


Integrated haplotype score:

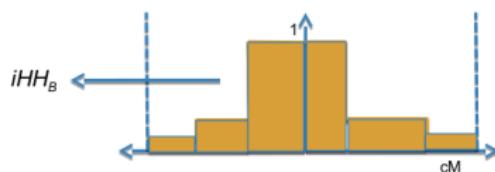
$$iHs = \ln(iHH_1 / iHH_0)$$

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

## Cross-population Extended Haplotype Homozygosity



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

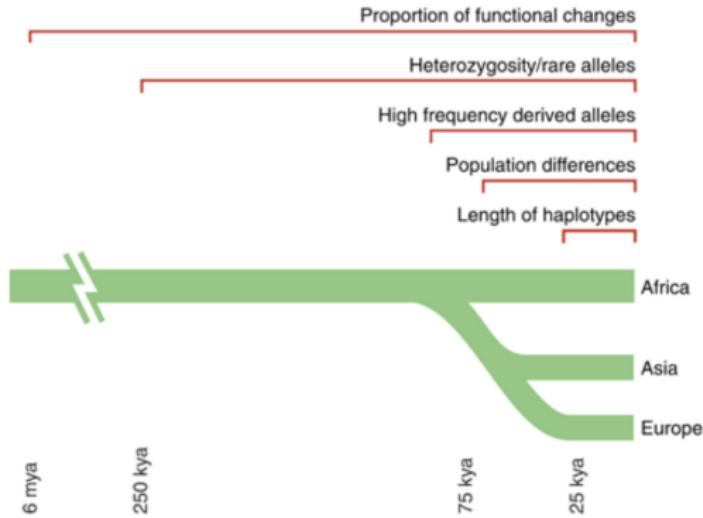


Integrated haplotype score:

$$XP-EHH = \ln(iHH_A/iHH_B)$$

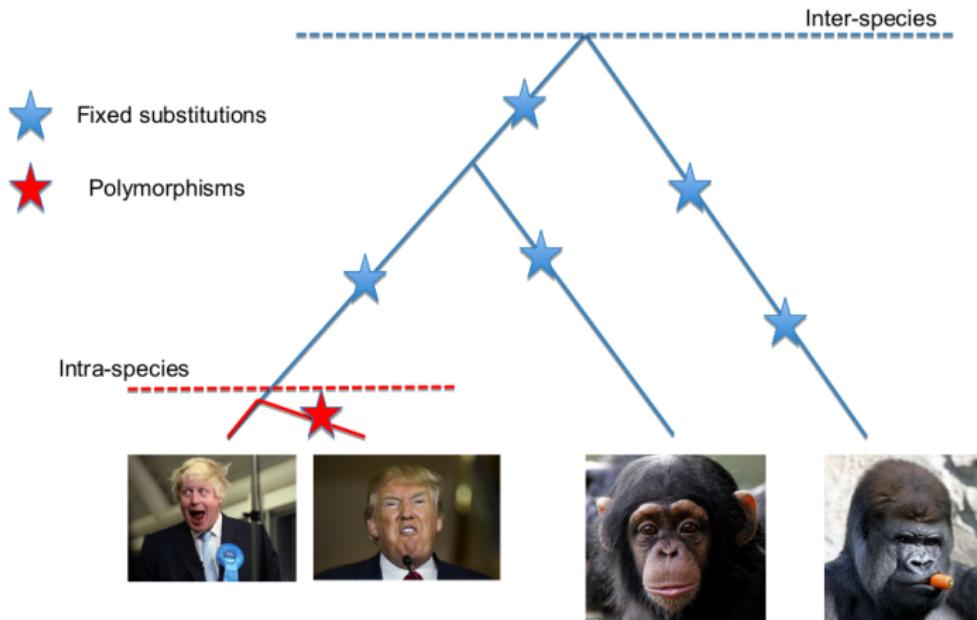
Genome-wide normalization in frequency bins  
(to mean=0 and sd=1)

# Detect recent selection within species / using shared variation



Sabeti et al. 2006 Science

# Infer inter-species selection



# Recent advances to detect selection

## 1. Composite scores (Grossman et al. 2013)

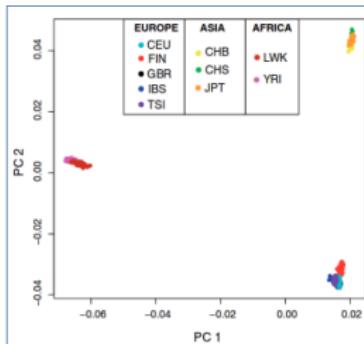
$$BF_t = \frac{P(v_t \in bin_{t,k} | selected)}{P(v_t \in bin_{t,k} | unselected)}$$

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

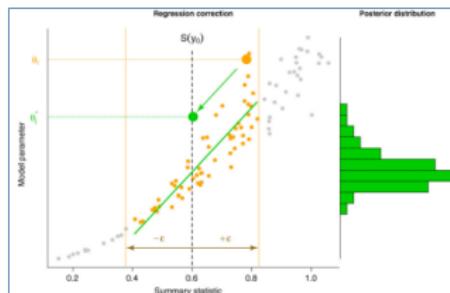
$$CMS_{GW} = \prod_{t \in \text{tests}} BF_t$$

## 3. Unsupervised machine learning

(PCA, Duforet-Frebbourg et al. 2016)

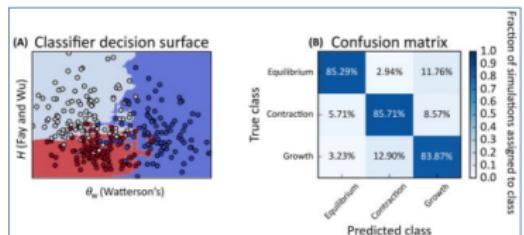


## 2. Simulations-based (rejection, ABC)



## 4. Supervised machine learning

(SVM, Schrider & Kern 2018)



## Intended Learning Outcomes

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

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