Mapping Signatures of Positive Selection

Saber Qanbari

Department of Animal Sciences
Animal Breeding and Genetics Group
Georg-August-University Göttingen, Germany

Searching for the action of natural selection is a challenge…

…but there is also some promise!

ystack

le in a hay-
ublicity man,
nient pile of
ensive search
edle. Having
needle hunt.

copyright © 2008 Bill Frymire

Outline

- Overview of selection
- Background selection
- Balancing selection
- Positive selection
- Practical session (methods to detect positive selection)
  - Local variability
  - Allele frequency spectrum
  - Haplotype based approaches

Methods for detecting selection

- Difference between species
  - High proportion of function altering mutations
- Within-species variation
  - Differences between populations
  - Low diversity
  - Excess of derived alleles
  - Long unbroken haplotypes

Types of selection

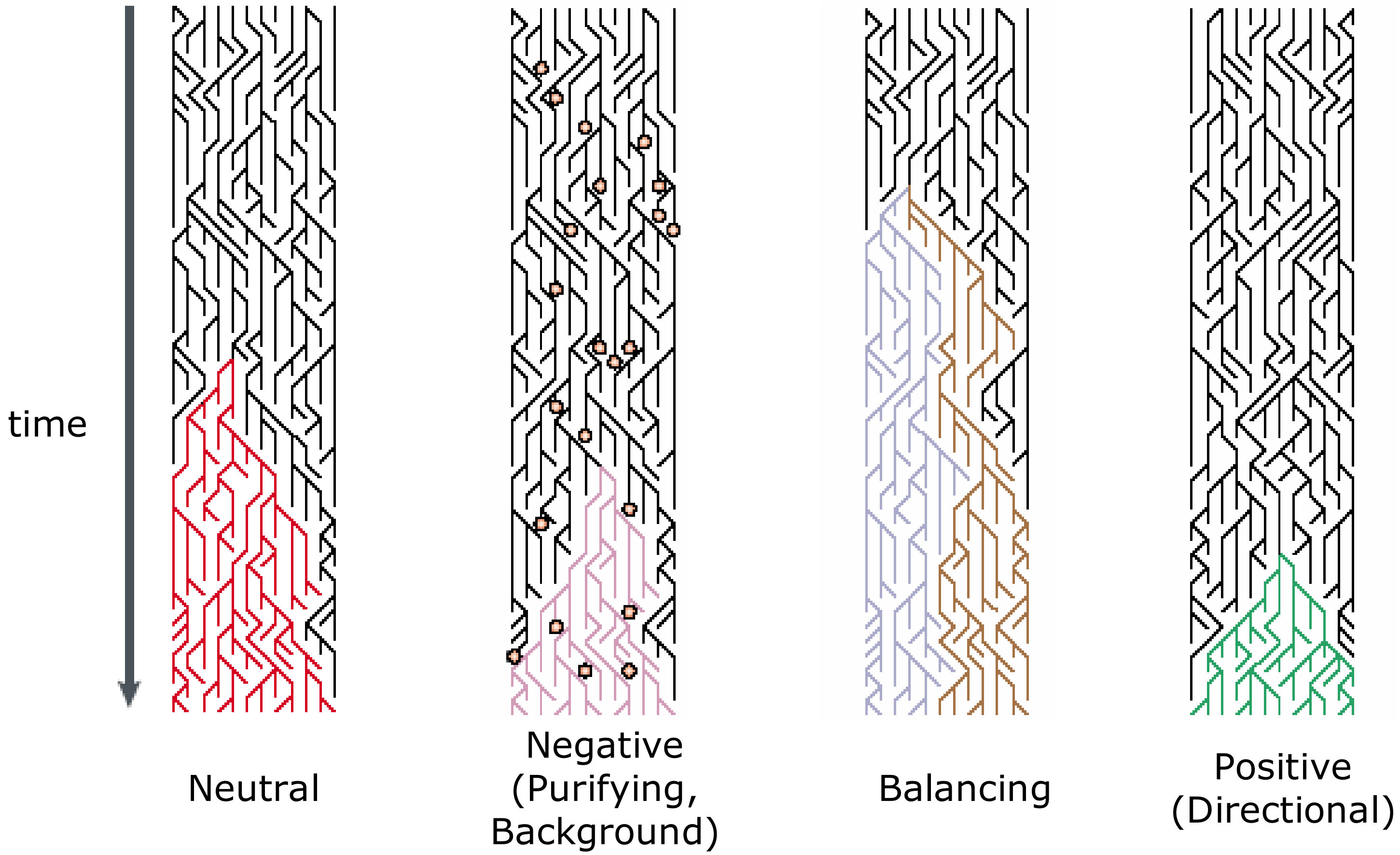
□ **Background selection** refers to the elimination of neutral polymorphism as a result of the negative selection of deleterious mutations (i.e. **purifying** or **negative selection**).

□ **Balancing selection** maintain variation in the population longer than expected

- Different functional mutations are favored
- Heterozygotes have a selective advantage

□ **Positive selection** favors for a adaptive (new/rare) mutation

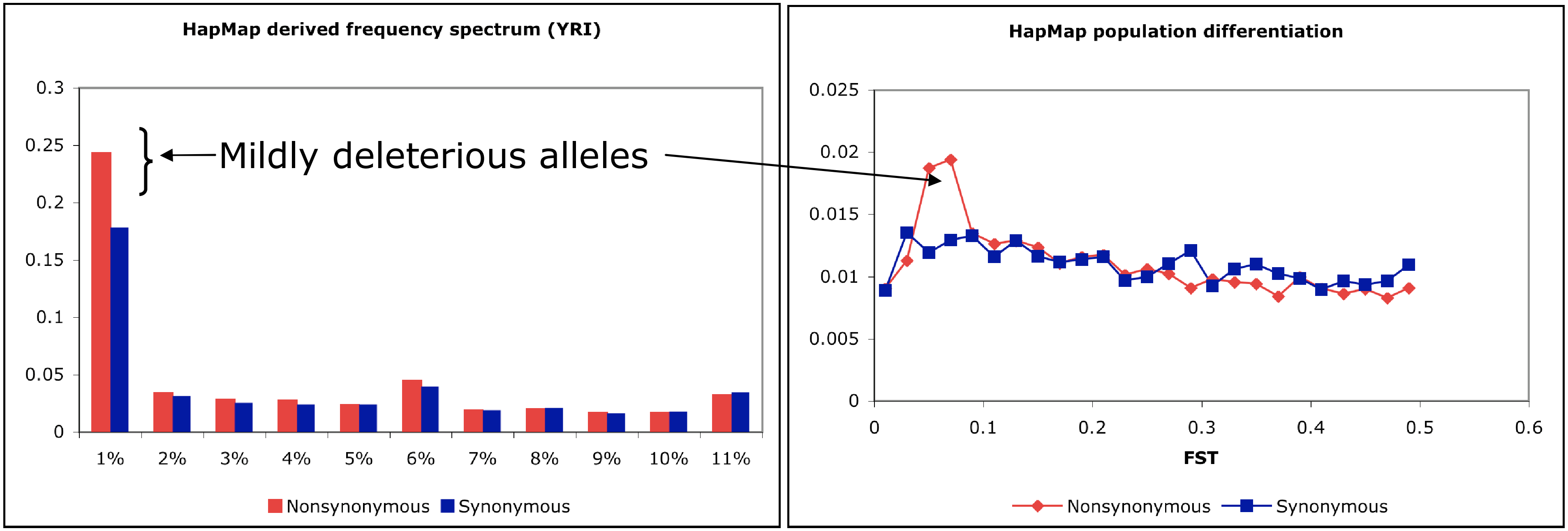
Types of selection



Bamshad & Wooding (2003) *Nature Rev. Genet.* **4**, 99-111

Background selection

Deleterious mutations stay at low frequency.
Nonsynonymous mutations are usually deleterious.



HapMap derived frequency spectrum (YRI)

HapMap population differentiation

Mildly deleterious alleles

Balancing Selection: selection for diversity

Balancing selection can lead to regions of unusually high genetic diversity

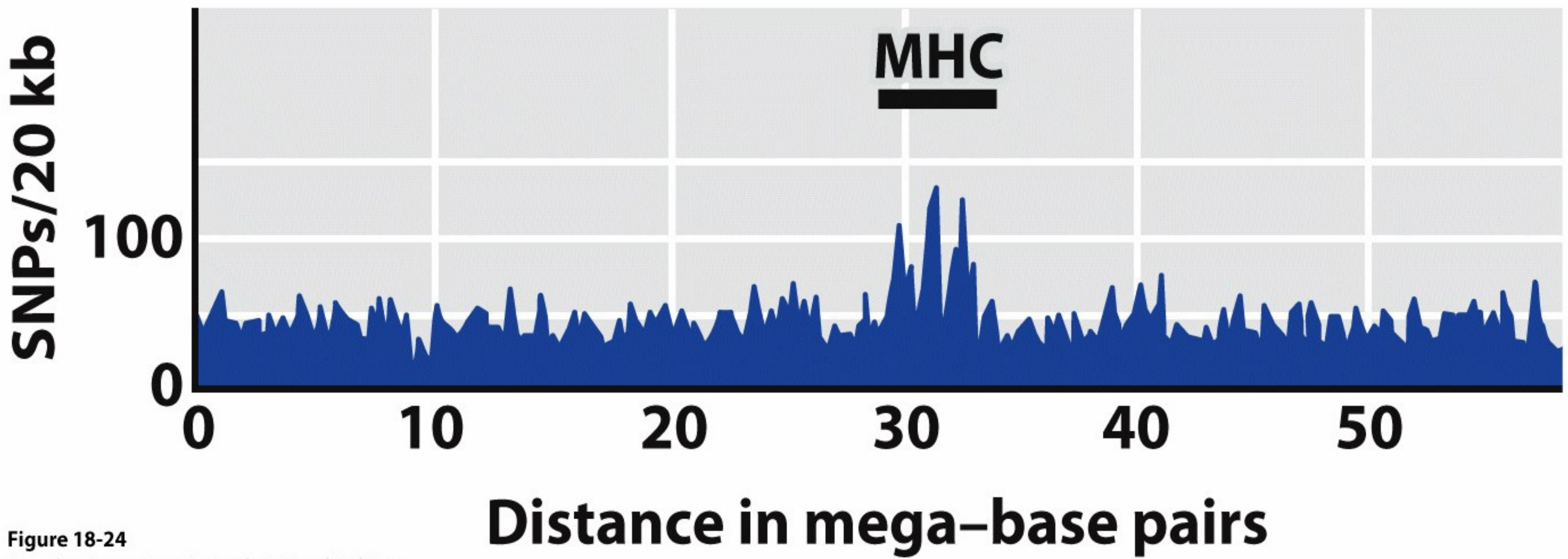
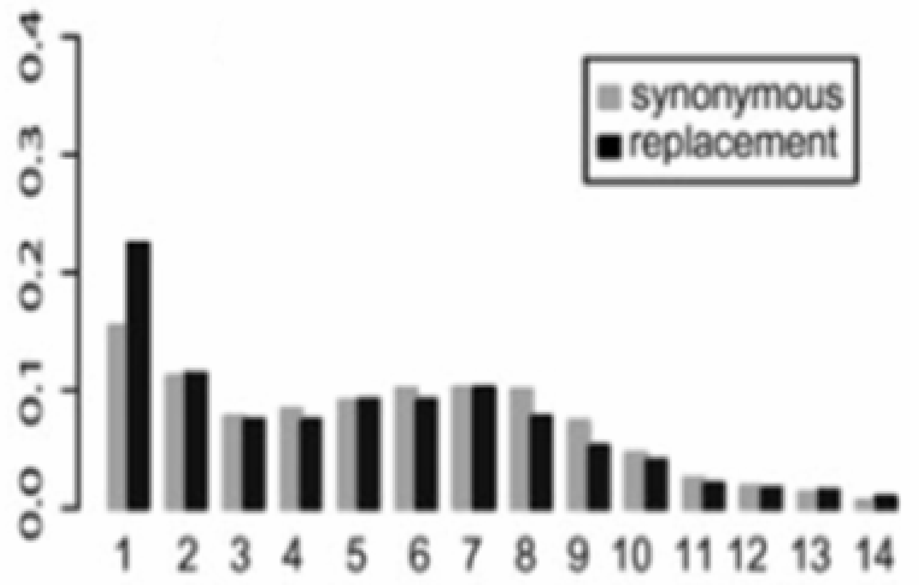


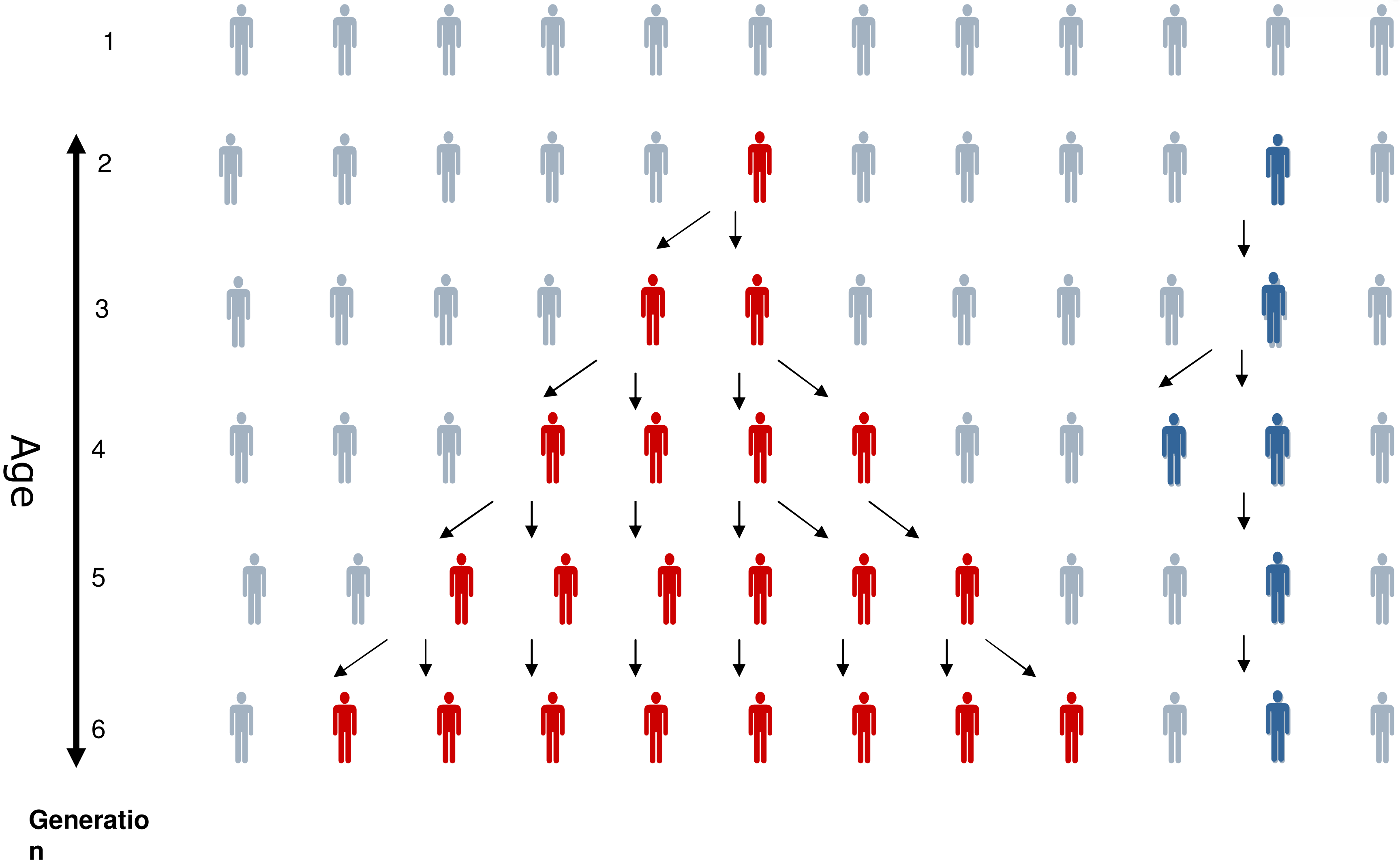
Figure 18-24
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

Detecting Balancing Selection

- Look for sites with excess polymorphism (Heterozygosity)
- Look for an excess of intermediate-frequency alleles at a site relative to rest of genome
- Compute site frequency spectra and perform Mann-Whitney U test
- CLR (in press ...)



Positive selection



Genetic variation: positive selection

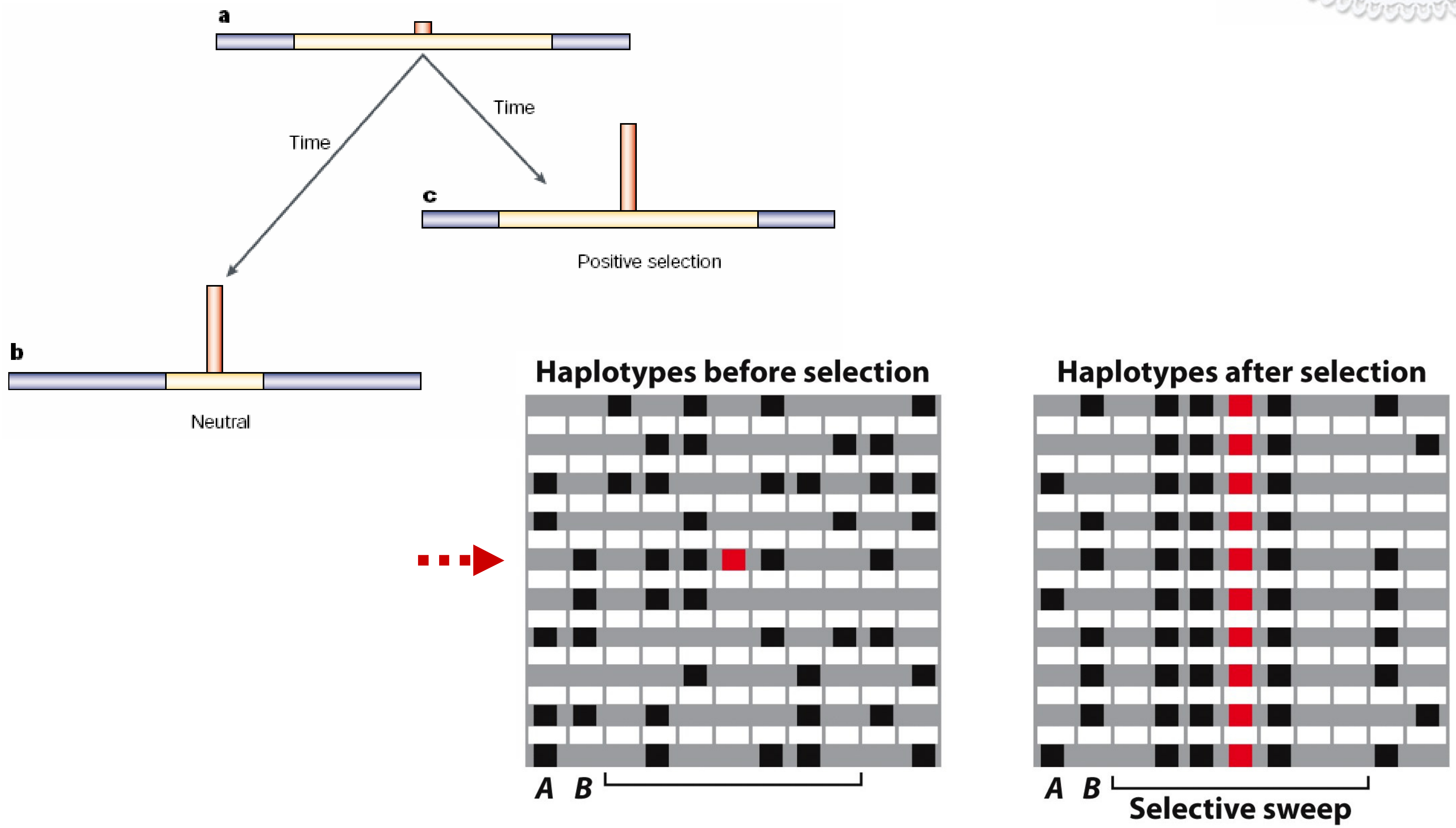
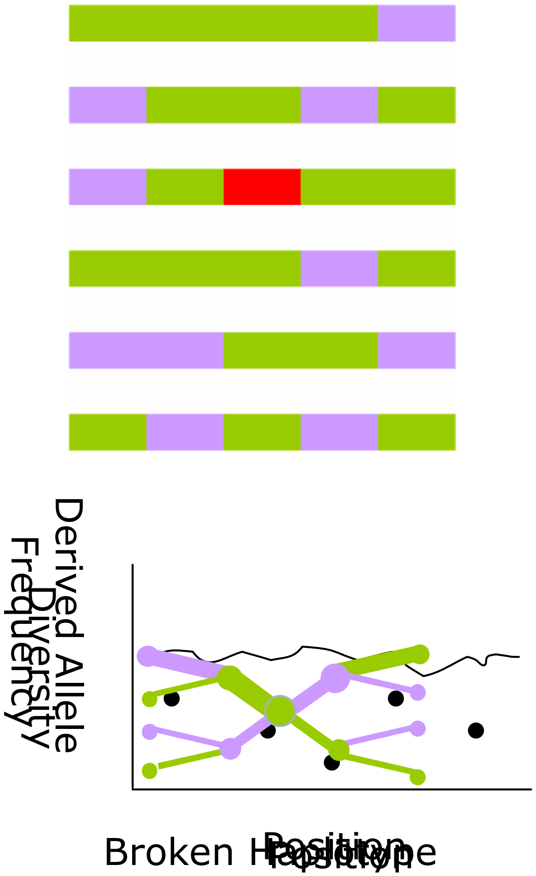


Figure 18-22
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

Signatures of a ‘selective sweeps’

before

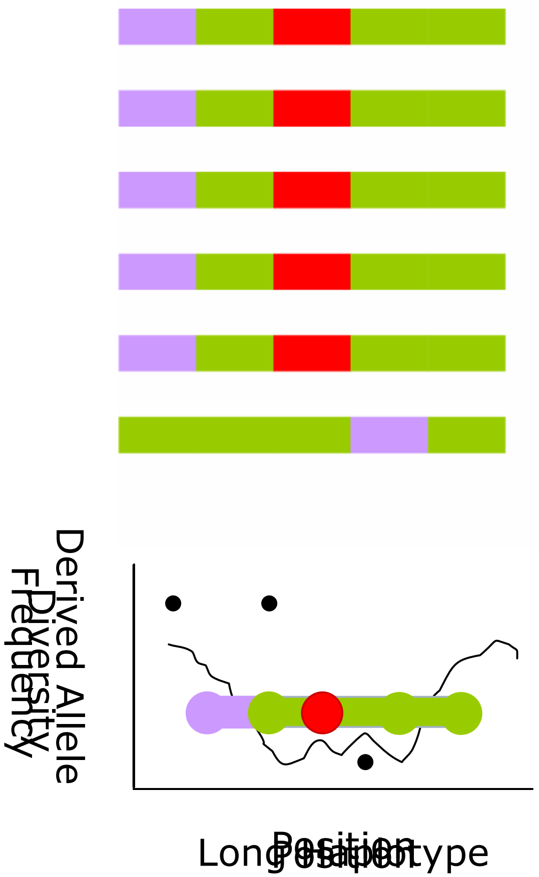


1) Low local variability
(many rare alleles)

2) Excess of frequent
and rare alleles

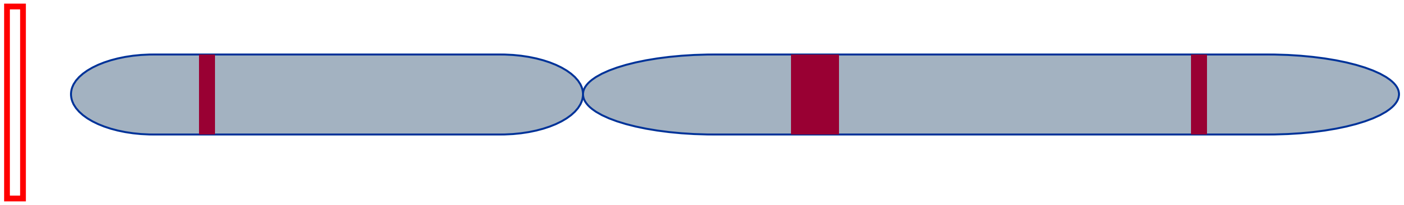
3) Long-range (unbroken)
haplotypes

after



Finding selective sweeps

- Pick a statistical test to detect sweeps
- Apply the statistic across the genome



- Validate the results
  - **Model-based**

    Compare genetic variation to ‘neutral’ model
  - **Purely empirical**

    Consider the ‘most extreme’ genomic regions
  - **Calibrated**

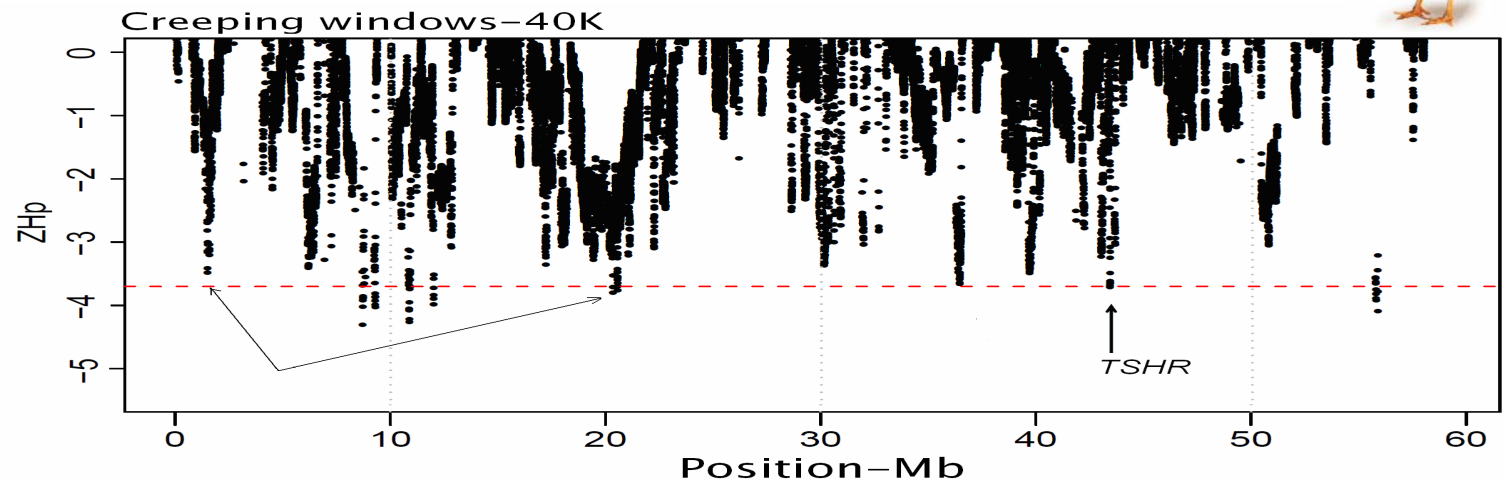
    Compare to examples of (very few) proven selective sweeps

Genome-wide searches for positive selection

- Low diversity
- Excess of frequent/rare haplotypes
- Long unbroken haplotypes

Low diversity:

- Simply look at diversity metrics (eg., proportion of polymorphic loci or heterozygosity, etc)



Locally reduced diversity region suggestive of a distinct selective sweep along with *TSHR* gene on GGA5 in Lohmann brown layers (Qanbari et al. 2012)

Genome-wide searches for positive selection

□ Low diversity

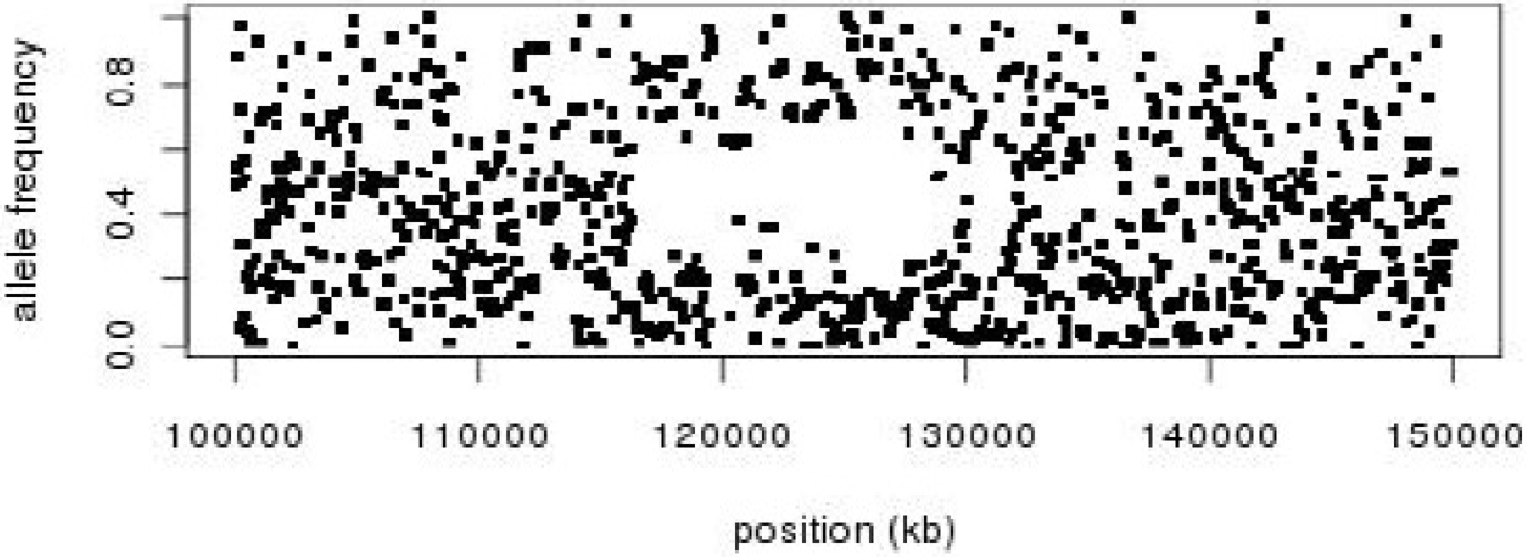
□ Excess of rare and frequent alleles

□ Long unbroken haplotypes

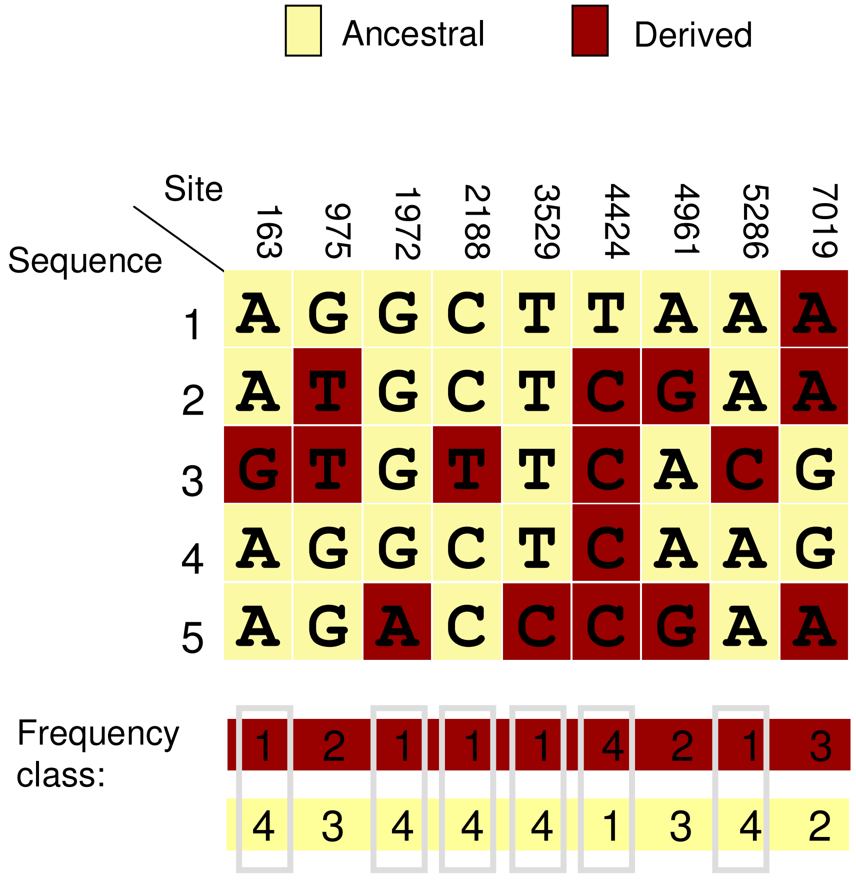
Site Frequency Spectrum (SFS):

Look for regions with deviated SFS

GDF8 gene in Texel sheep (Hapmap data)

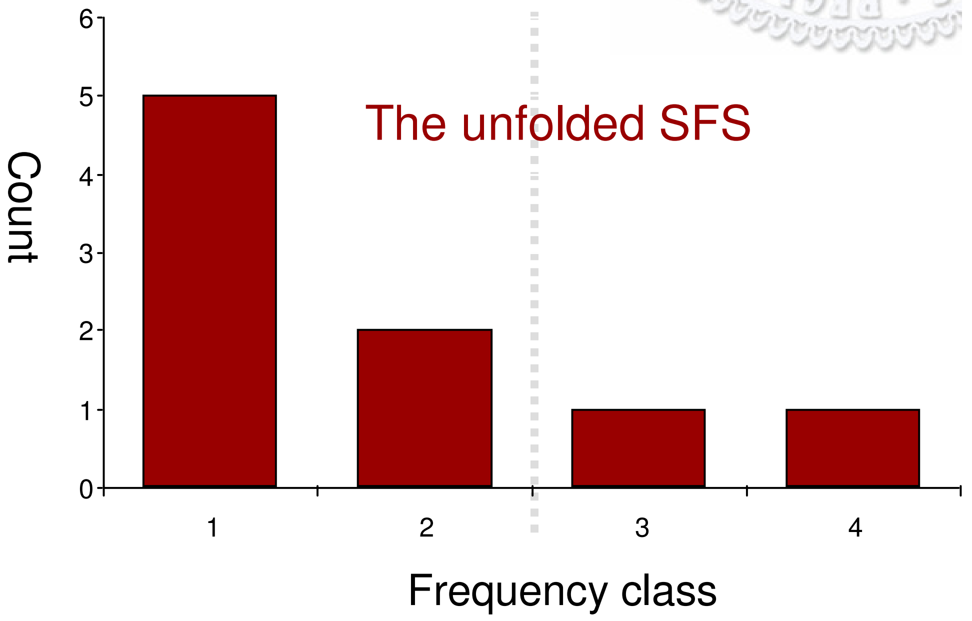


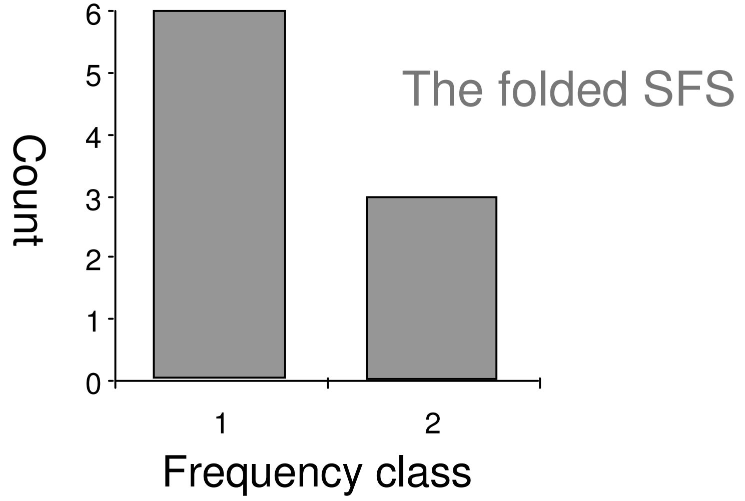
Folded vs., unfolded SFS



🟨 Ancestral 🟥 Derived

Sequence \ Site	163	975	1972	2188	3529	4424	4961	5286	7019
1	A	G	G	C	T	T	A	A	A
2	A	T	G	C	T	C	G	A	A
3	G	T	G	T	T	C	A	C	G
4	A	G	G	C	T	C	A	A	G
5	A	G	A	C	C	C	G	A	A
Frequency class:	1	2	1	1	1	4	2	1	3
	4	3	4	4	4	1	3	4	2



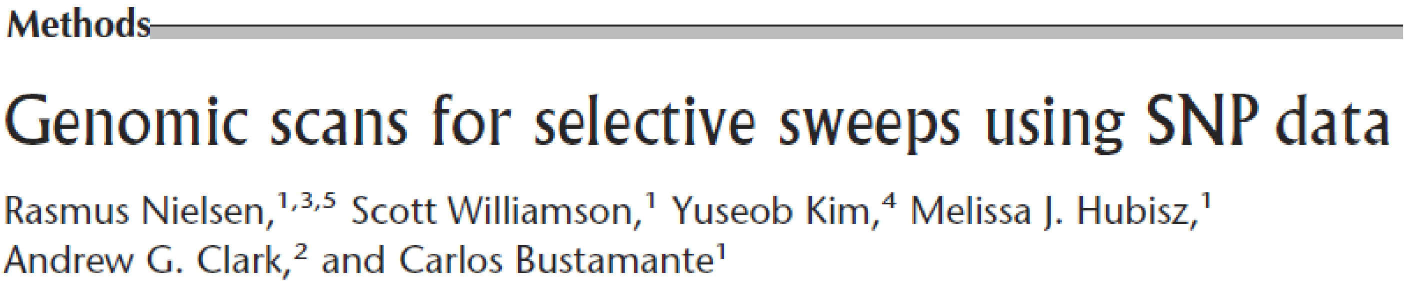


Finding dSFS regions …

- □ Nucleotide diversity
- □ Tajima D
- □ Fay & Wu H test
- □ Composite of Likelihood Ratio

... decides between the two hypothesis based on the value of the likelihood ratio.

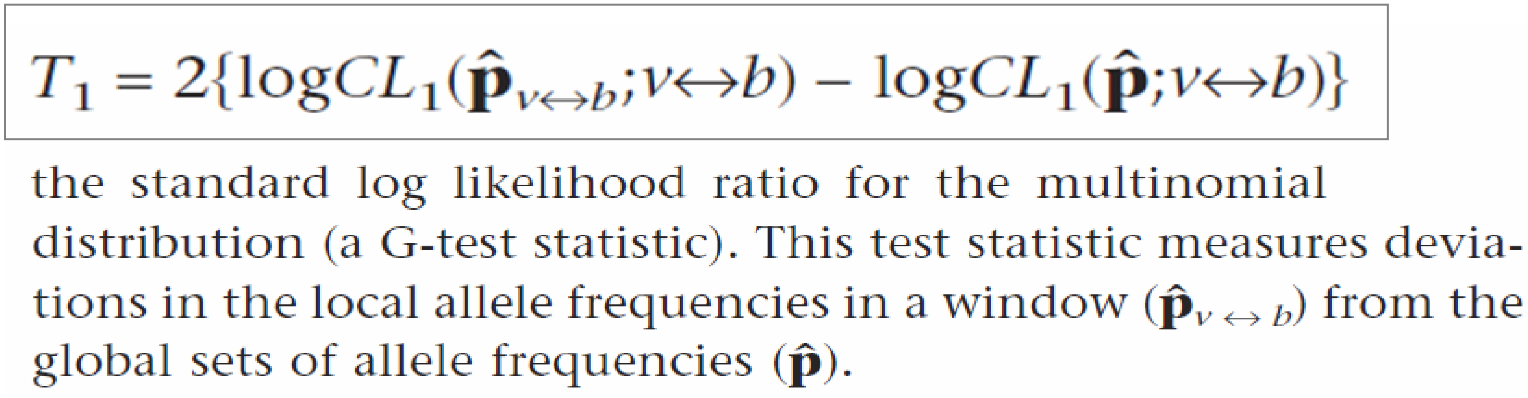
Finding dSFS regions …



Methods

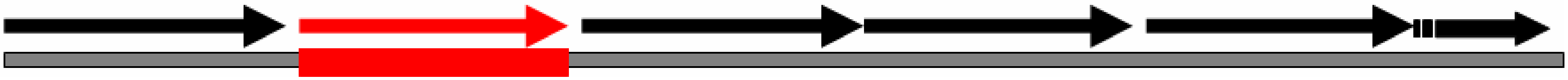
Genomic scans for selective sweeps using SNP data

Rasmus Nielsen,[1,3,5] Scott Williamson,[1] Yuseob Kim,[4] Melissa J. Hubisz,[1]
Andrew G. Clark,[2] and Carlos Bustamante[1]



$$T_1 = 2\{\log CL_1(\hat{\mathbf{p}}_{v\leftrightarrow b}; v\leftrightarrow b) - \log CL_1(\hat{\mathbf{p}}; v\leftrightarrow b)\}$$

the standard log likelihood ratio for the multinomial distribution (a G-test statistic). This test statistic measures deviations in the local allele frequencies in a window ($\hat{\mathbf{p}}_{v\leftrightarrow b}$) from the global sets of allele frequencies ($\hat{\mathbf{p}}$).

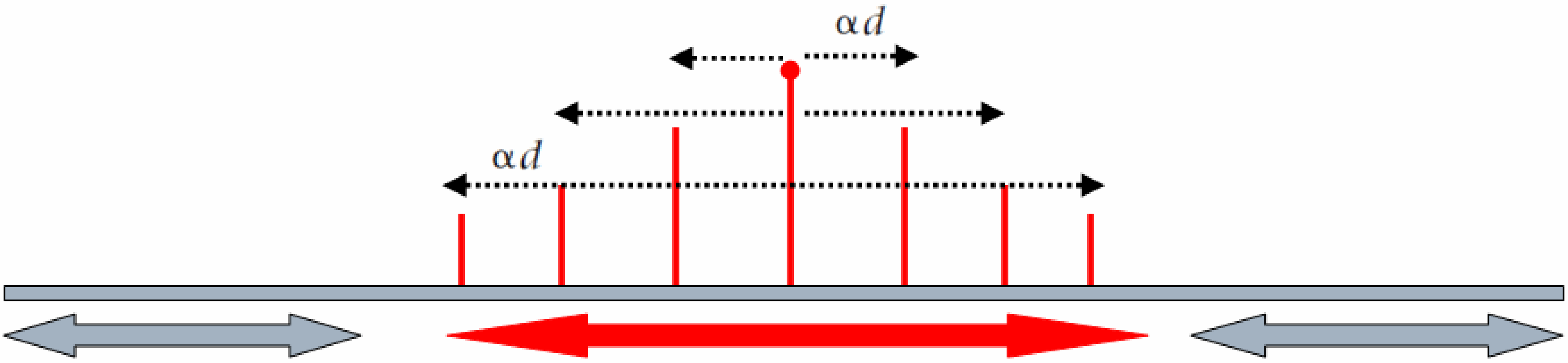


Finding dSFS regions …

$P_e = 1 - e^{-\alpha d}$, where d is the distance from the location of the sweep to the sampled SNP

$\alpha = r \ln(2N)/s$,

$$\boxed{p_B^{*} = P_e(n)p_B + \sum_{k=0}^{n-1} P_e(k)\left(p_{B+1-n+k,k+1}\frac{B+1-n+k}{k+1} + p_{B,k+1}\frac{k+1-B}{k+1}\right),}$$



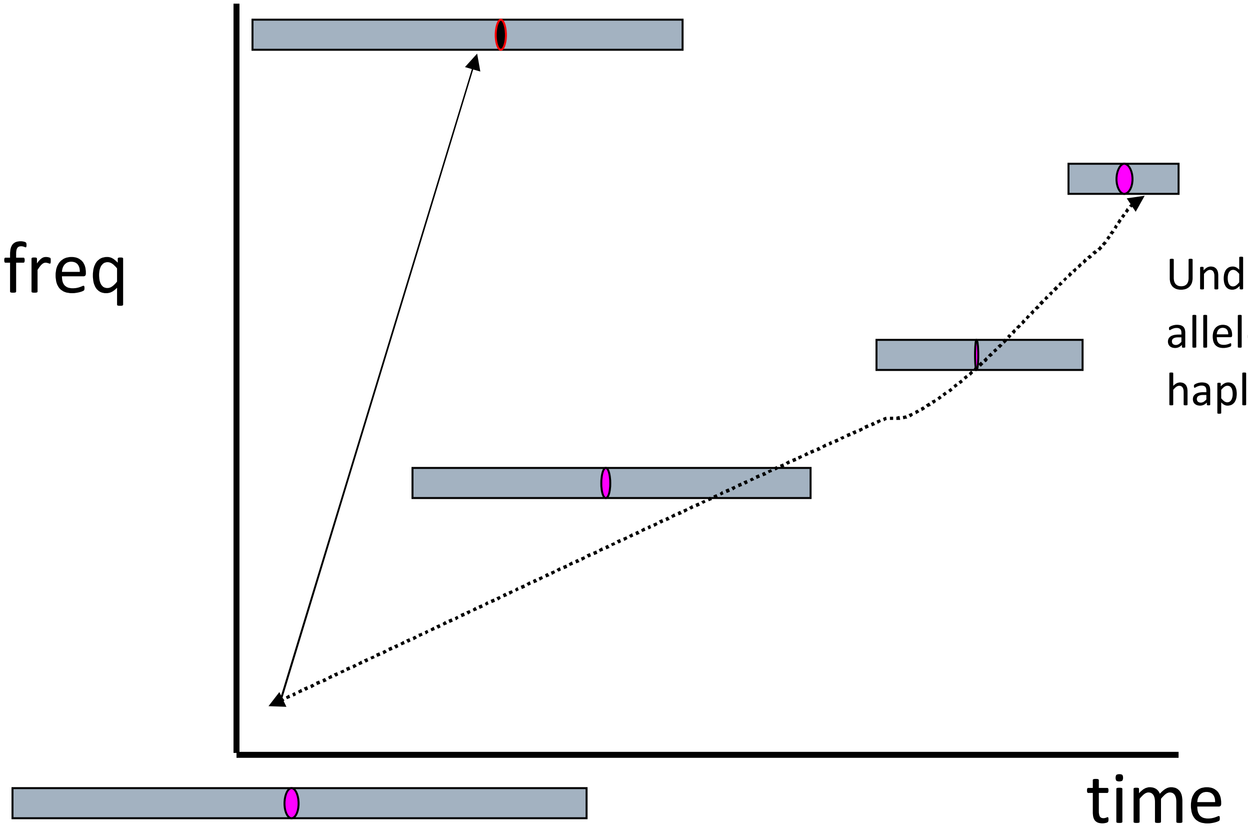
Finding deviated SFS (dSFS)

- Big CLR value indicates a sweep. How big is big?
- Do simulations to estimate significance.
- Repeat the CLR calculation for each simulation.
- Then for each region, find proportion of simulated CLRs that are bigger than its original CLR.
- That proportion is a p-value that tells if the region is a sweep.

Genome-wide searches for positive selection

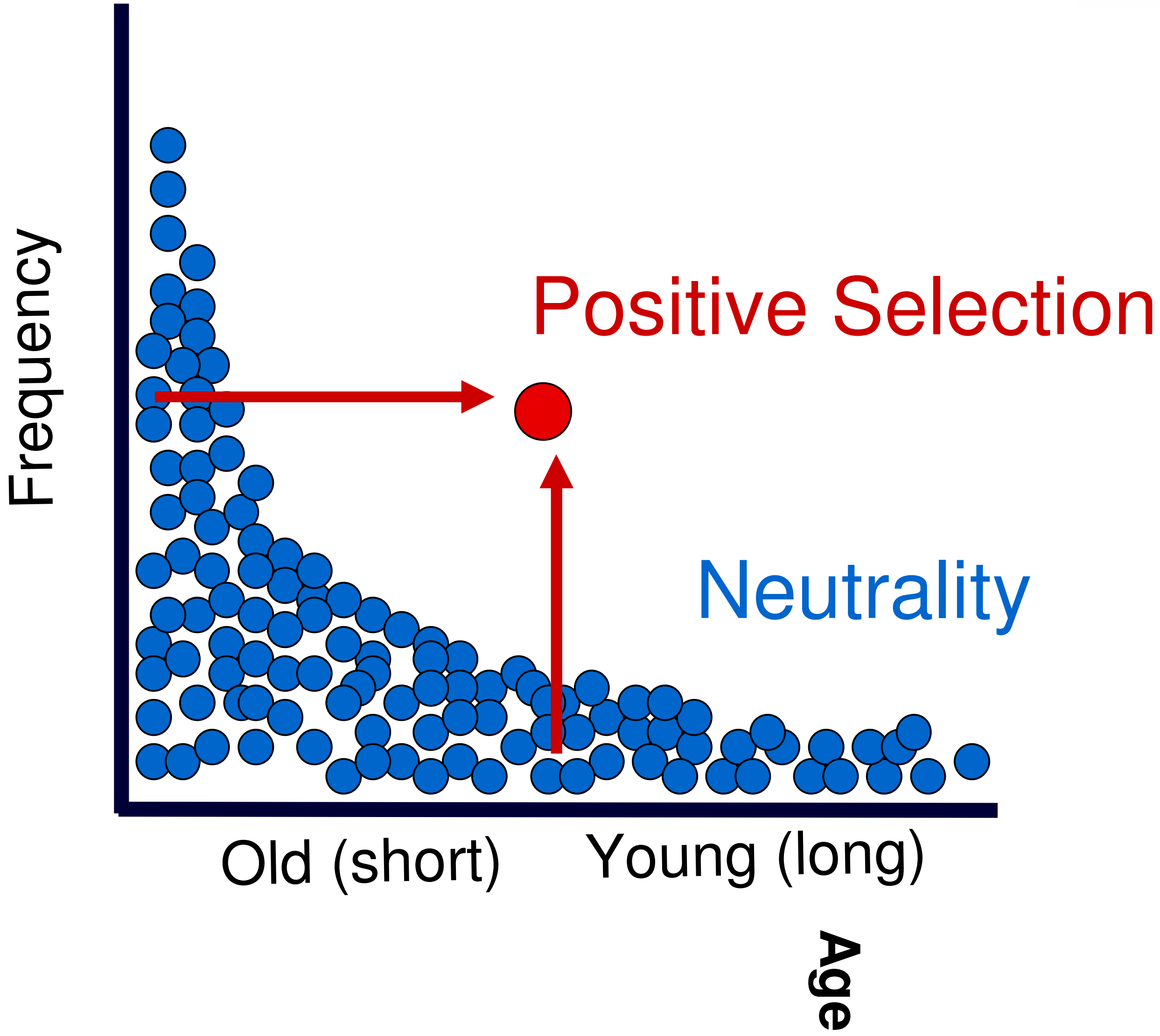
- Low diversity
- Excess of rare and frequent alleles
- Long unbroken haplotypes

Under directional selection, very fast change in allele frequency, and hence short time. Results in high-frequency alleles with long haplotypes

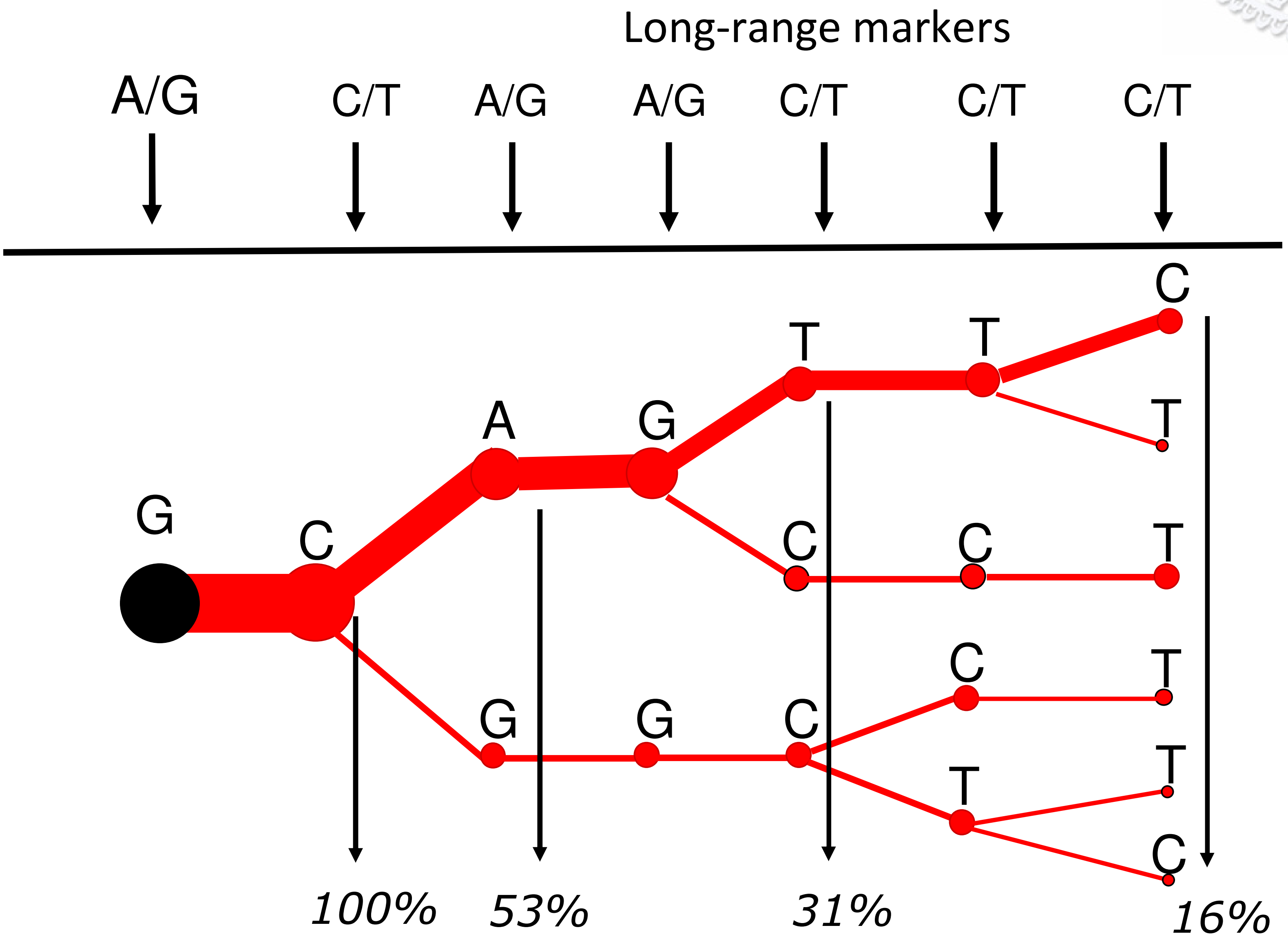


Under pure drift, high-freq alleles should have short haplotypes

Starting haplotype


Positive Selection
Neutrality
Frequency
Old (short)
Young (long)
Age

Measuring length of haplotype



Testing Long Range Haplotypes

- EHH (REHH); Sabeti et al. (Nature 2002)
  - Look for signal of “extended haplotype homozygosity”

- iHS; Voight et al. (PLoS Biology 2006)
  - Focus on potentially selected mutation
  - Compare selected/non-selected types

- iES; Rsb, XPEHH, nSL metrics use similar concept

Extended Haplotype Homozygosity

- Define "core regions" (eg with a higher LD) and estimate EHH

$$EHH_t = \frac{\sum_{i=1}^{s} \binom{e_{ti}}{2}}{\binom{c_t}{2}}$$

- REHH (relative EHH) – EHH normalized by EHH values of other haplotypes in the core region

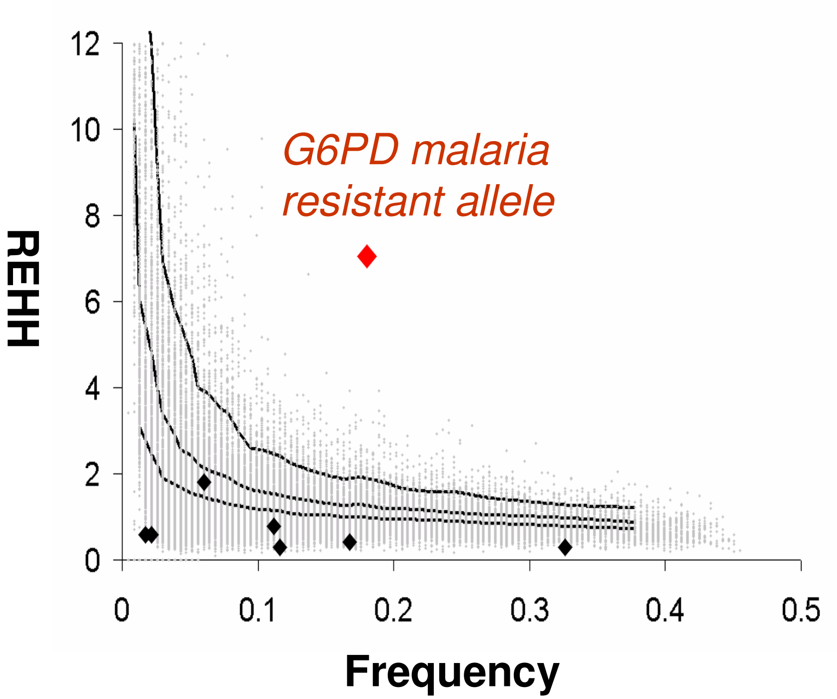
$$\mathrm{REHH} = EHH_t / \overline{\mathrm{EHH}}$$

- Bin SNPs by haplotype frequency
- Normalize ln(REHH) per bin
- Outlying values indicative of selection

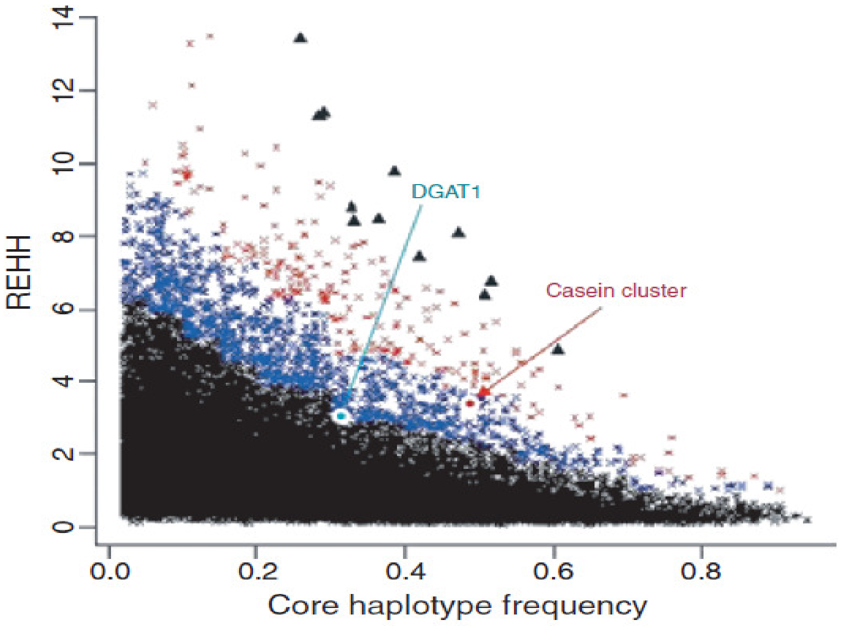
Extended Haplotype Homozygosity

Looking for a haplotype longer for its frequency (expected under neutrality)

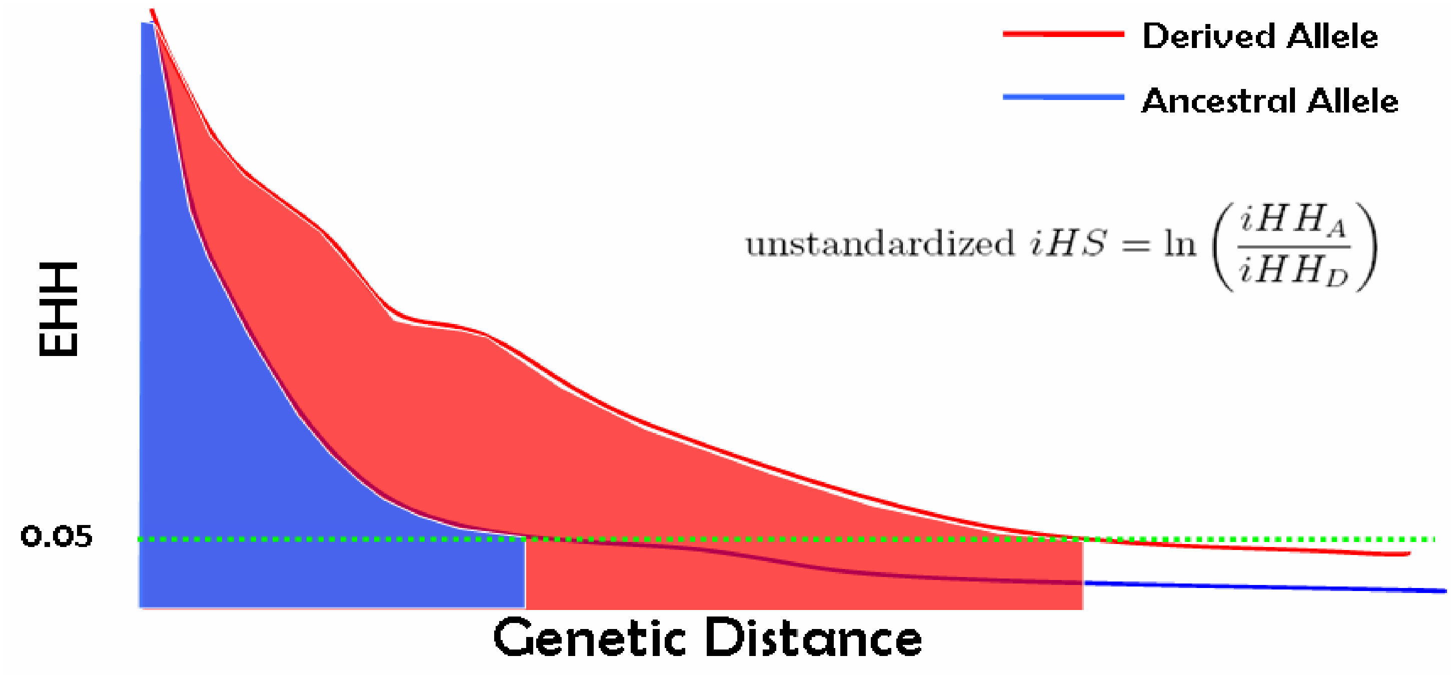
(Sabeti et al. 2002)



(Qanbari et al. 2010)



iHS: integrated Haplotype Homosygosity Score



$$\text{unstandardized } iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$

iHHD : iHH with respect to **D**erived core allele.
iHHA : iHH with respect to **A**ncestral core allele.

Integrated EHHS (iES)…

- Look at the marker at site *i* and calculate its expected (HW) homozygosity=$E(H_i)$.

- Then move to another site *j*, and look at the haplotypes that are defined by the variants between sites *i* and *j.*

- Next, calculate the expected (HW) homozygosity for these haplotypes = $E(H_{ij})$.

- The haplotype homozygosity between sites *i* and *j* normalized by the homozygosity at site *i* is:

$$EHHS_{i,j} = \frac{E(Ho_{i,j})}{E(Ho_i)}$$

Integrated EHHS (iES)…

As *j* increases, this ratio will decrease, and Tang et al. look at the 0.1 threshold. A measure of how fast homozygosity decays with increasing site distance until this threshold is reached is the area under the step function:

$$iES_i = \sum_{j=a+1}^{b} \frac{(EHHS_{i,j-1} + EHHS_{i,j})(Pos_j - Pos_{j-1})}{2}$$

Where *a* and *b* are the 5’ and 3’ positions from *i* at which the 0.1 threshold is reached, and Pos_j is the physical position of site *j* in the genome.

Thats it folks :)

Synbreed