Lecture 07:
Detecting selection with marker data.
I: Overview

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Detecting selection

- Bottom line: looking for loci showing departures from the equilibrium neutral model
- What kinds of selection are of interest?
- Time scales and questions
- KEY POINTS
  - False positives very common
  - MOST selective events will not be detected
  - Those that are likely represent a rather biased sample

Negative selection is common

- Negative (or purifying) selection is the removal of deleterious mutations by selection
- Leaves a strong signal throughout the genome
  - Faster substitution rates for silent vs. replacement codons
  - Comparative genomics equates strong sequence conservation (i.e., high negative selection) with strong functional constraints
  - The search for selection implies selection OTHER than negative

Positive selection

- An allele increasing in frequency due to selection
  - Can either be a new mutation or a previously neutral/slightly deleterious allele whose fitness has changed due to a change in the environment.
  - Adaptation
- Balancing selection is when alternative alleles are favored by selection when rare
  - MHC, sickle-cell
- The "search for selection" is the search for signatures of positive, or balancing, selection

Time scales of interest

- Ecological
  - An allele either currently undergoing selection or has VERY recently undergone selection
  - Detect using the nature of genetic variation within a population sample
  - Key: A SINGLE event can leave a signature
- Evolutionary
  - A gene or codon experiences REPEATED adaptive events over very long periods of time
  - Typically requires between-species divergence data
  - Key: Only informs us as to the long-term PATTERN of selection over a gene

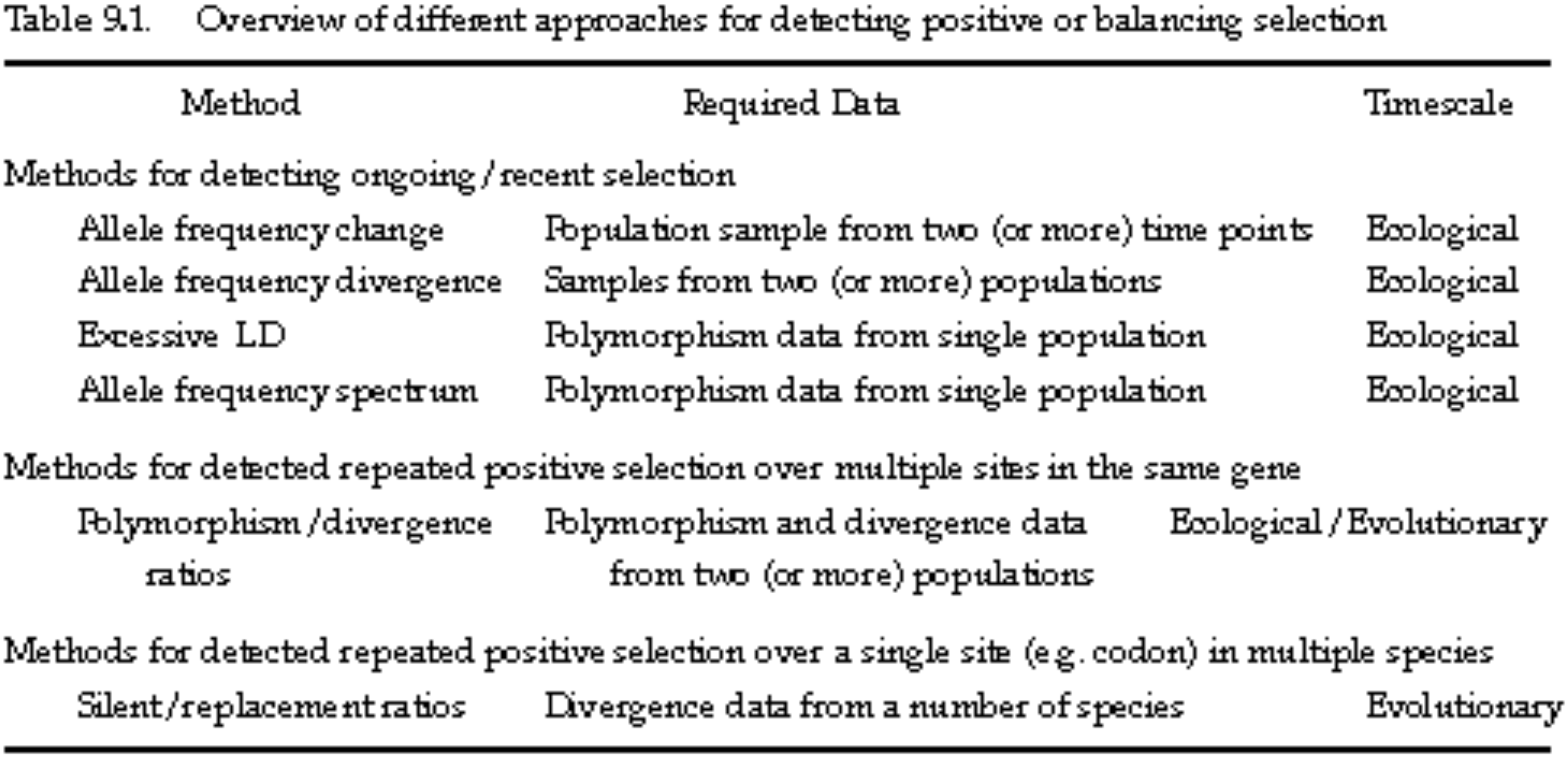


Table 9.1. Overview of different approaches for detecting positive or balancing selection

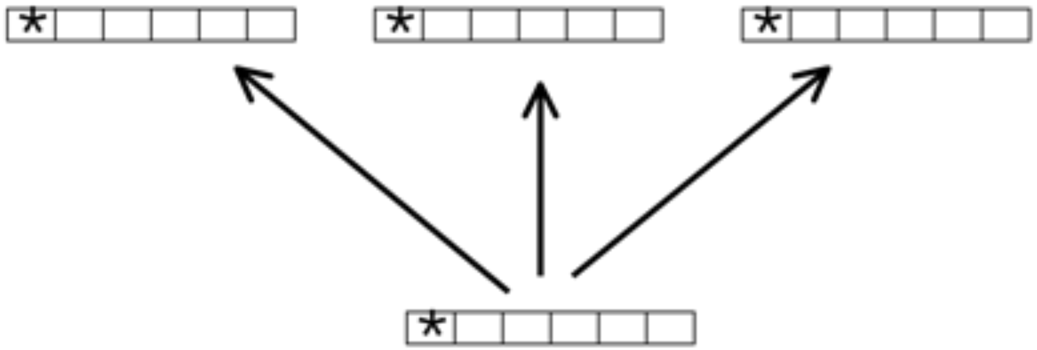
Method	Required Data	Timescale
Methods for detecting ongoing / recent selection		
Allele frequency change	Population sample from two (or more) time points	Ecological
Allele frequency divergence	Samples from two (or more) populations	Ecological
Excessive LD	Polymorphism data from single population	Ecological
Allele frequency spectrum	Polymorphism data from single population	Ecological
Methods for detected repeated positive selection over multiple sites in the same gene		
Polymorphism /divergence ratios	Polymorphism and divergence data from two (or more) populations	Ecological / Evolutionary
Methods for detected repeated positive selection over a single site (e.g. codon) in multiple species		
Silent/replacement ratios	Divergence data from a number of species	Evolutionary

Biased scan for selection

- Current/very recent selection at a single site requires rather strong selection to leave a signature.
  - Small shifts in allele frequencies at multiple sites unlikely to leave signatures
  - Very small time window (~0.1 Ne generations) to detect such an event once it has occurred.
- Recurrent selection
  - Phylogenic comparisons: Multiple substitution events at the same CODON required for a signal
  - OK for "arms-race" genes, likely not typical

- Recurrent selection at sites OVER a gene
  - Comparing fixed differences between two species with the observed levels of polymorphism
  - Requires multiple substitutions at different codons (i.e., throughout the gene) for any signal
  - Hence, a few CRITICAL adaptive substitutions can occur in a gene and not leave a strong enough signal to detect
  - Power depends on the number of adaptive substitutions over the background level of neutral substitutions

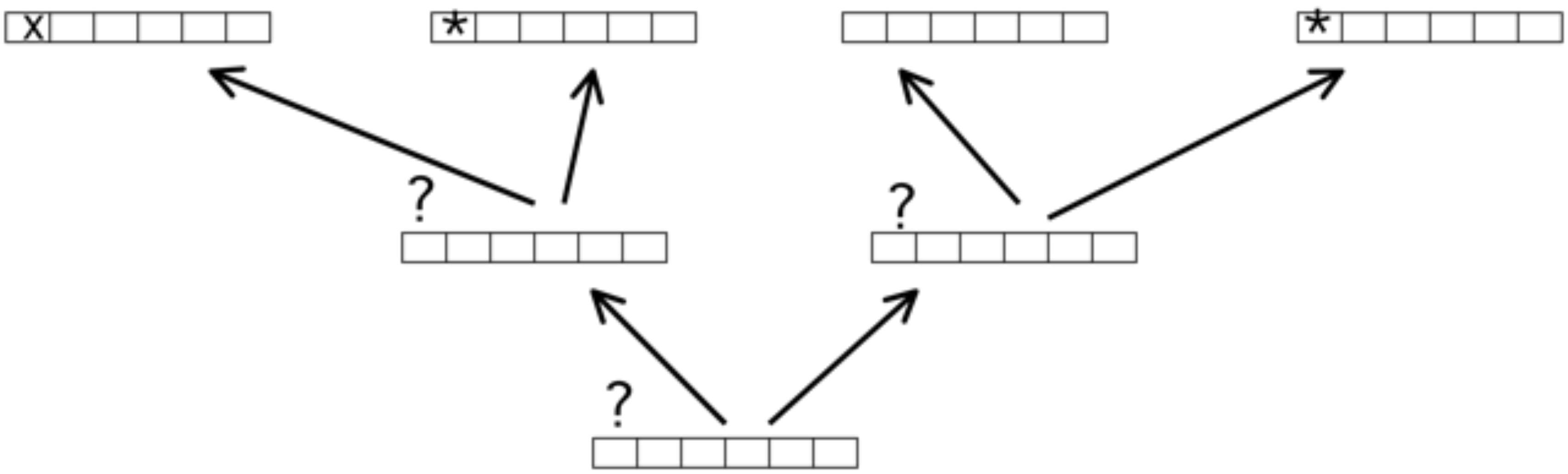
Sample of a gene from several individuals in the same population



Ongoing, or recent, selectio

Detecting ongoing selection within a population. Requires a population sample, in which we look for inconsistencies of the pattern of variation from the equilibrium neutral model. Can detect on-going selection in a single region, influencing the pattern of variation at linked neutral loci.

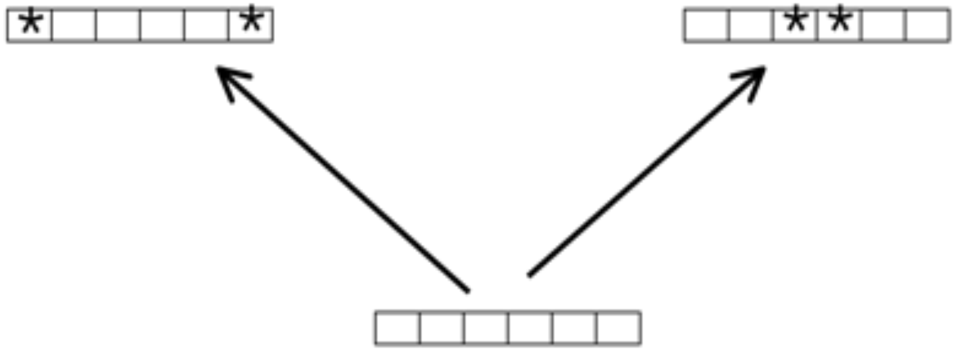
Sample of a gene over several species



Divergence data on a phylogeny.
Repeated positive selection at the same site

A phylogenic comparison of a sequence over a group of species is done on a codon-by-codon basis, looking for those with a higher replacement than silent rate. Requires MULTIPLE substitutions at the same codon over the tree

Fixed differences between two species



Positive selection occurring ov
multiple sites within the gene

Comparison of divergence data for a pair of species. Requires a background estimate of the expected divergence from fixation of neutral sites, which is provided from the polymorphism data (I’ll cover this shortly).

Key points

- Methods for detecting selection
  - Are prone to false-positives
    - The rejection of the null (equilibrium neutral model) can occur for reasons other the positive/balancing selection, such as changes in the population size
  - Are under-powered
    - Most selection events likely missed
  - Detect only specific types of selection events
    - Ongoing moderate to strong events
    - Repeated adaptive substitutions in a few codons over a phylogeny
    - Repeated adaptive substitutions over all sites in a gene

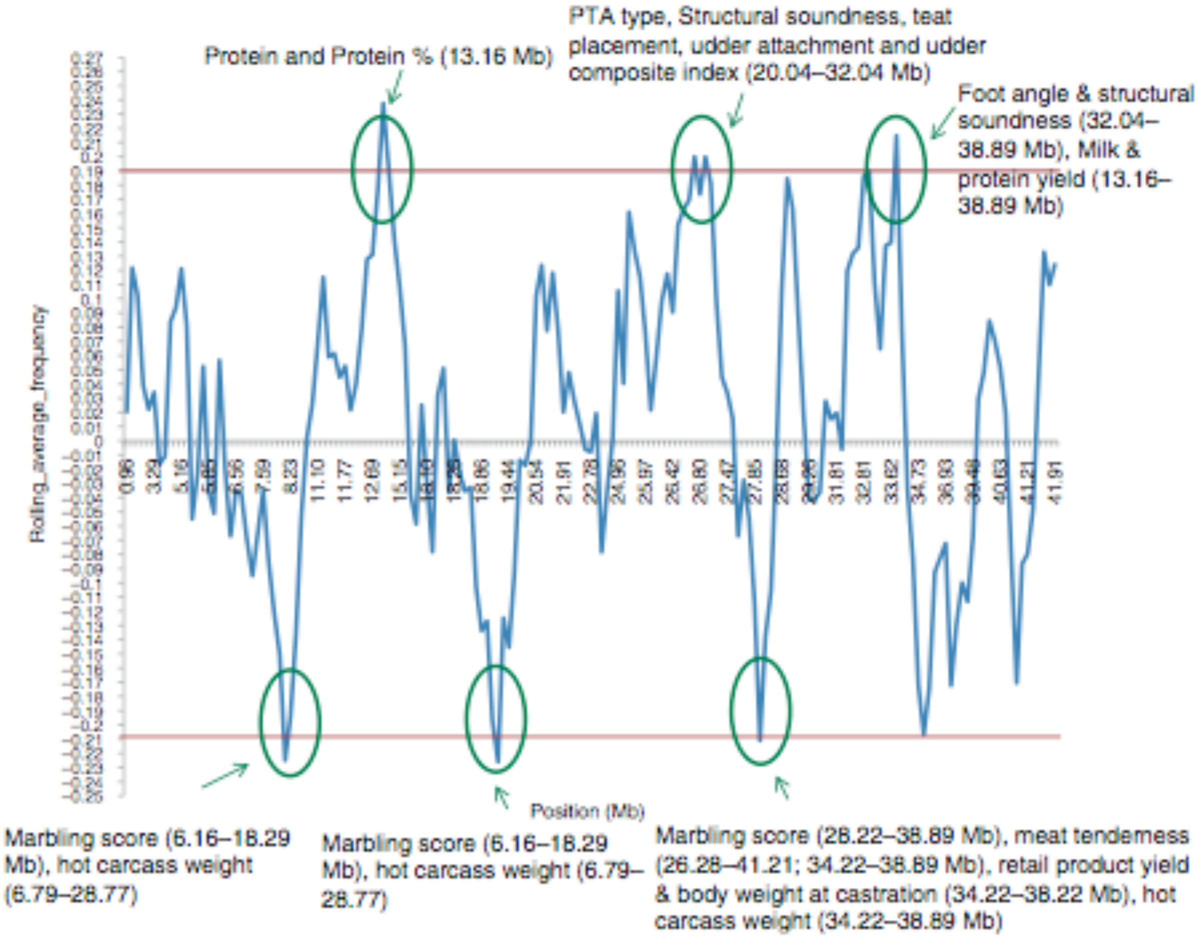
Detecting on-going selection

- Excessive allele frequency change/ divergence
- Selective Sweeps
  - Reduction in polymorphism around a selected site
- Shifts in the allele frequency spectrum
  - i.e., too many rare alleles
- Allelic age inconsistencies
  - Allele too common relative to its age
  - Excessive LD in a common allele

Excess allele frequency change

- Logically, most straightforward
- Need estimates of N_e, time
- Need two (or more) time points
- Generally weak power unless selection strong or time between sampling long
- Example: Divergence between breeds selected for different goals

Example 9.1. Angus and Holstein represent breeds of *Bos taurus* that have been selected, respectively, for beef and milk production. As such, might would expect allele frequency differences between the breeds, some of which represent differential selection on milk and beef traits. Prasad et al. (2008) uses 355 SNP markers on chromosome 19 (BT19) and another 175 SNPs on chromosome 29 (BT29) to search for significant allele frequency differences between these breeds. They used a five marker sliding window, computing the difference between the mean allele frequency in Holsteins and the mean frequency in Angus. Significantly positive values indicate potential alleles selected for milk production, while significant negatives values suggests alleles potentially selected for beef production. Figure 9.1 shows the result for chromosome 19. The authors used a permutation test to access the significance, with the species label for any given marker randomly assigned, and the difference for each five-marker window scored, generating an empirical distribution under the null hypothesis of breed-effects. Deviations above the upper significance line show alleles at a significantly higher frequency in Holsteins and deviations below the lower significance line indicates alleles that are significantly more frequency in Angus. The authors were able to relate these locations to locations of QTLs for various milk and beef production traits. Example 9.8 discusses Hayes et al. (2008), who also examine allele frequency differences between these two breeds.



Protein and Protein % (13.16 Mb)

PTA type, Structural soundness, teat placement, udder attachment and udder composite index (20.04–32.04 Mb)

Foot angle & structural soundness (32.04–38.89 Mb), Milk & protein yield (13.16–38.89 Mb)

Marbling score (6.16–18.29 Mb), hot carcass weight (6.79–28.77)

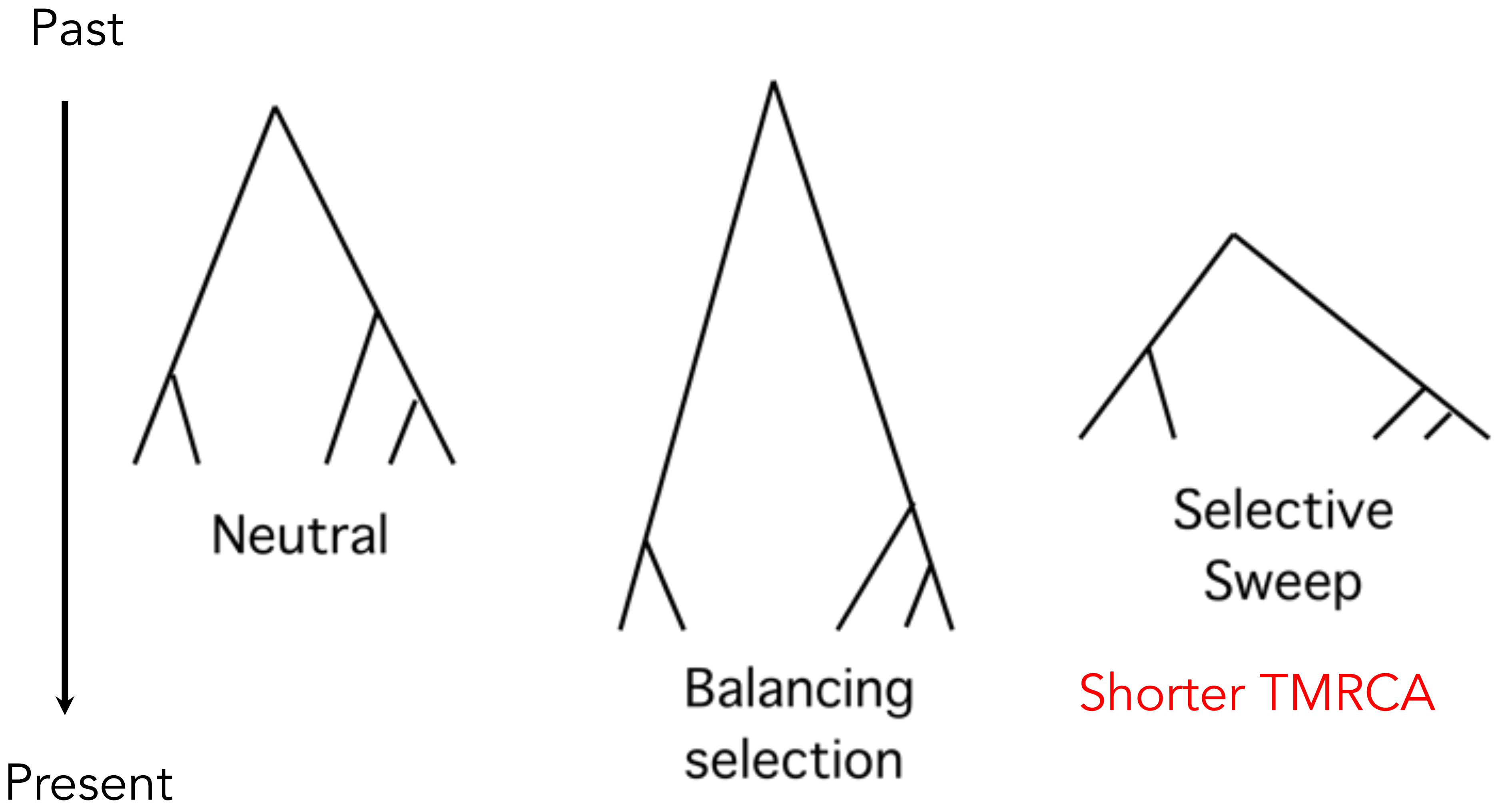
Marbling score (6.16–18.29 Mb), hot carcass weight (6.79–28.77)

Marbling score (28.22–38.89 Mb), meat tenderness (26.28–41.21; 34.22–38.89 Mb), retail product yield & body weight at castration (34.22–38.22 Mb), hot carcass weight (34.22–38.89 Mb)

Five-marker window scans of difference between Holstein & Angus breeds (dairy vs. beef selection)

Selective sweeps

- Classic visual tool to look for potential sites under selection
  - Common approach in the search for domestication genes
- Positive selection reduces N_e for linked sites
  - Reduces TMRCA and hence variation
- Balancing selection increases N_e for linked sites
  - Increases TMRCA and hence increase variation

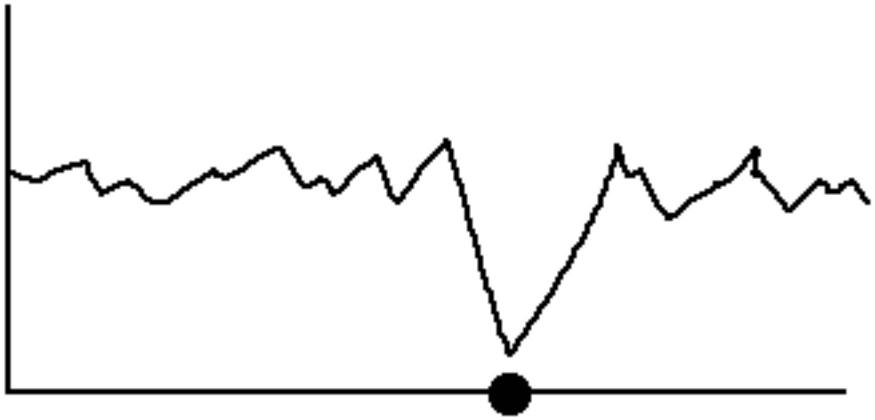


Shorter TMRCA

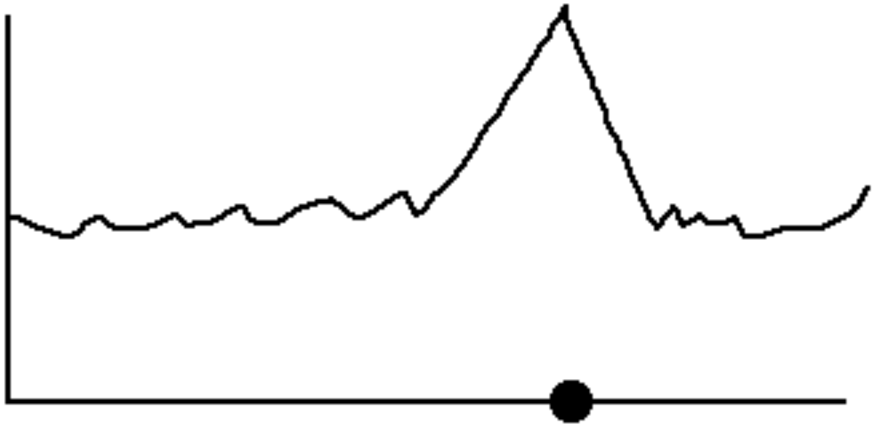
Longer TMRCA

Scanning for Sweeps

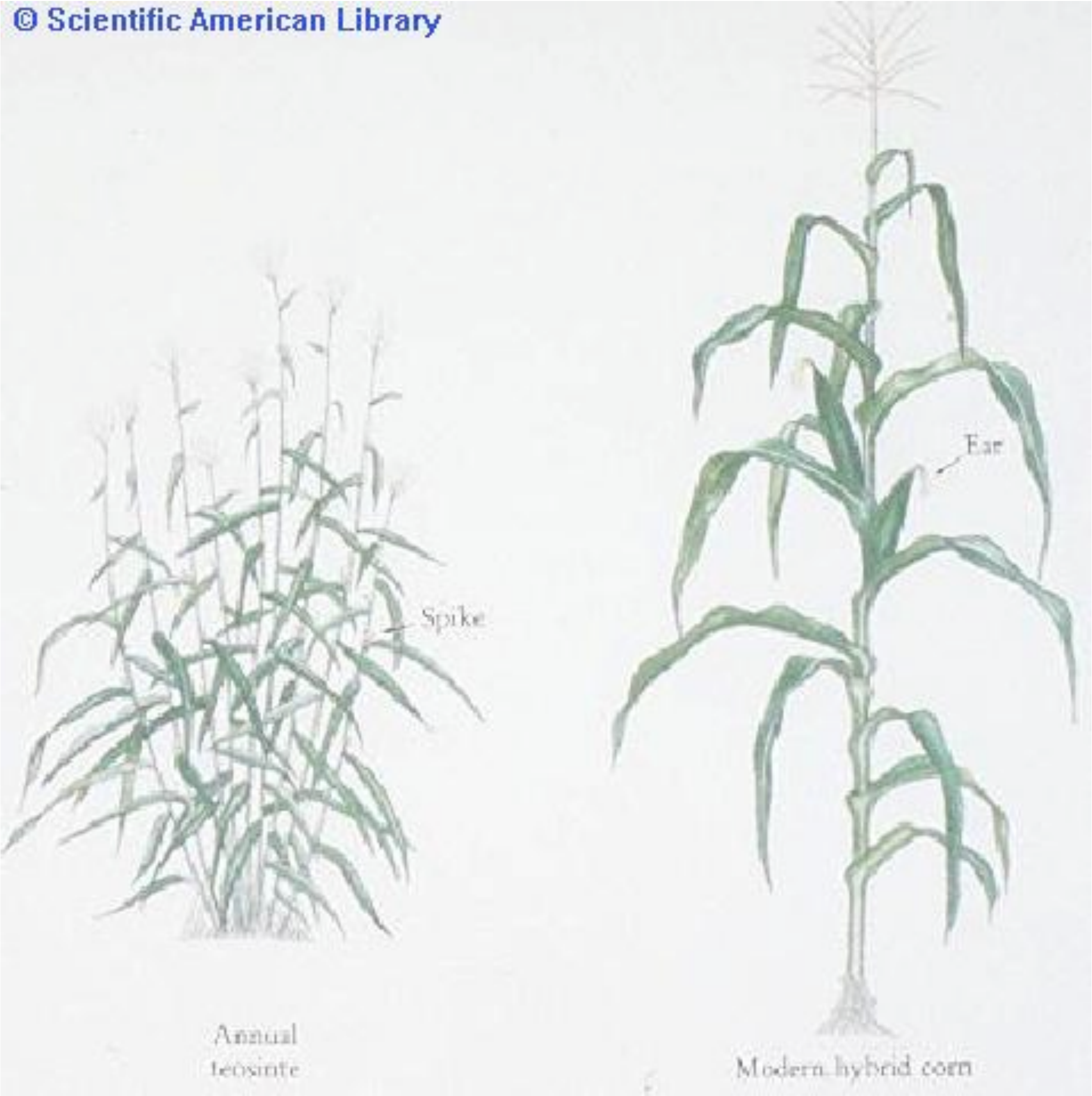
- Use a sliding window to look at variation along a chromosome (or around a candidate gene)
- Decrease (with respect to some standard) consistent with linked site under recent/ongoing positive selection
- Increase consistent with balancing selection



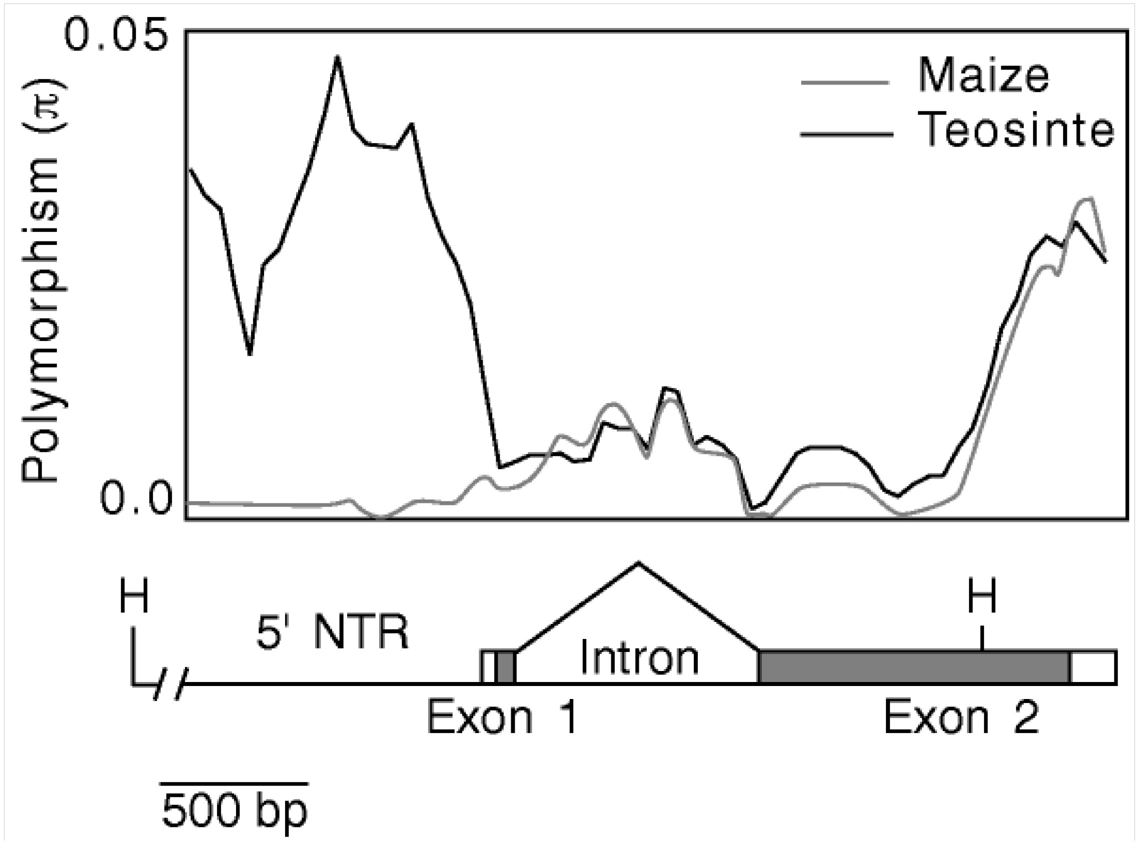
Signal of positive selection, OR reduction in mutation rate



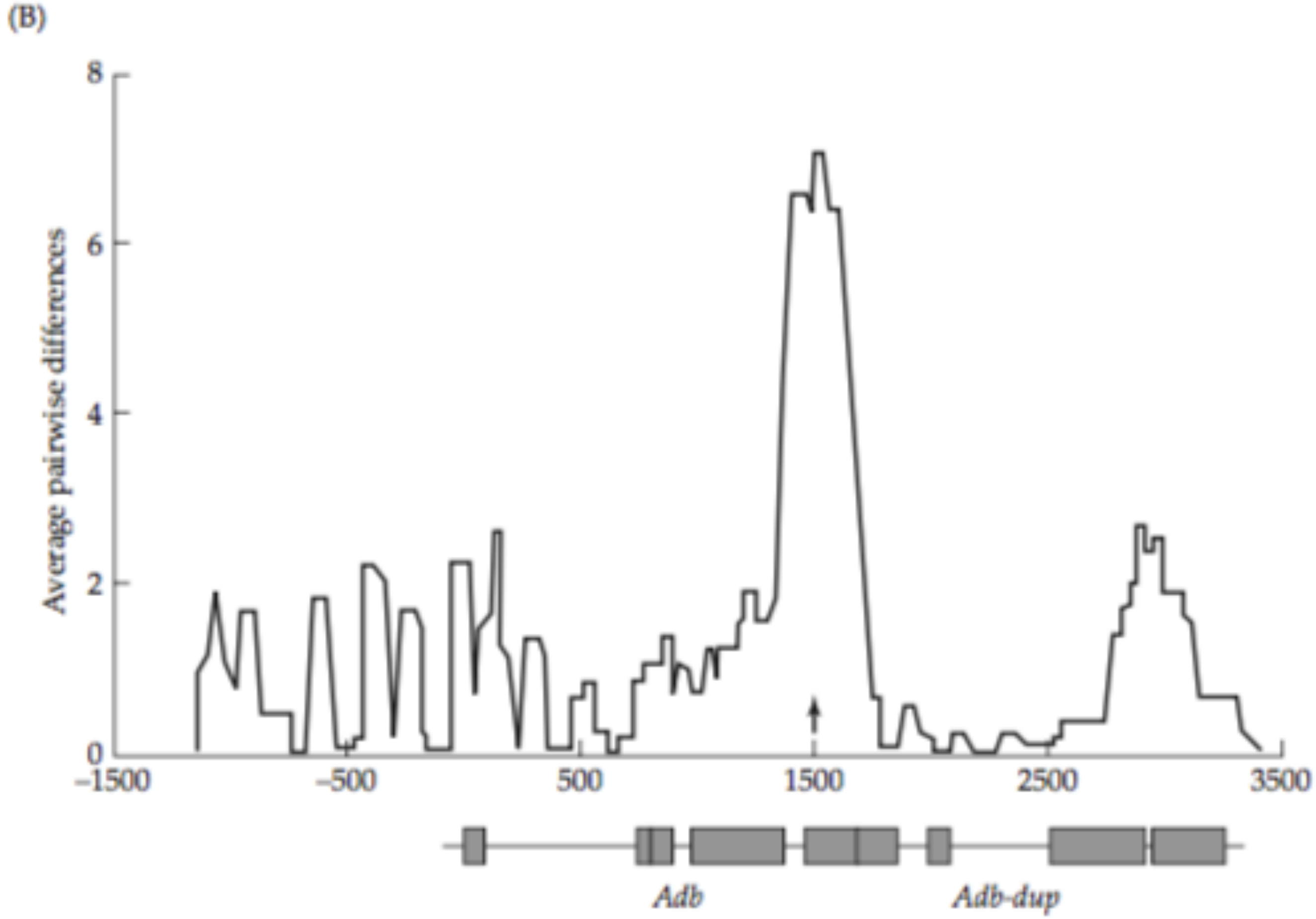
Signal of balancing selection, OR increase in mutation rate


© Scientific American Library

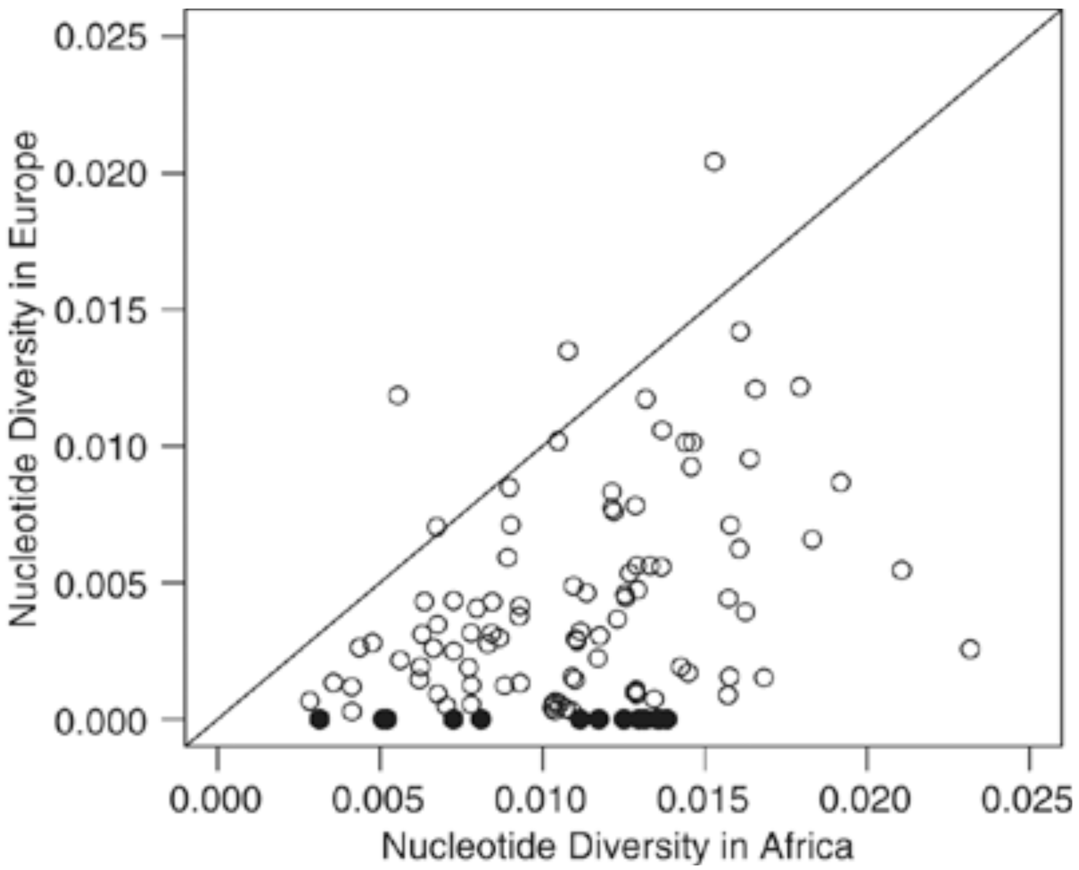
Domestication: Maize vs. teosinte



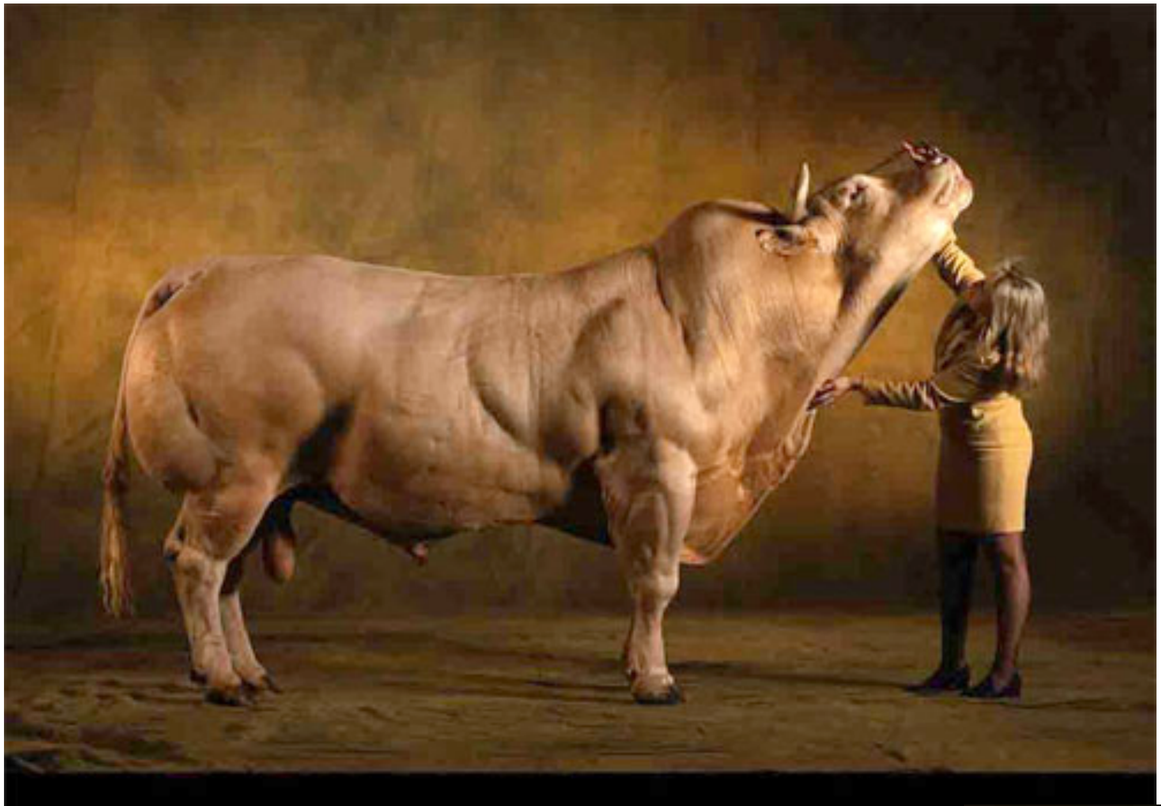
tb1 in maize. Used teosinte as a control for expected background levels of variation



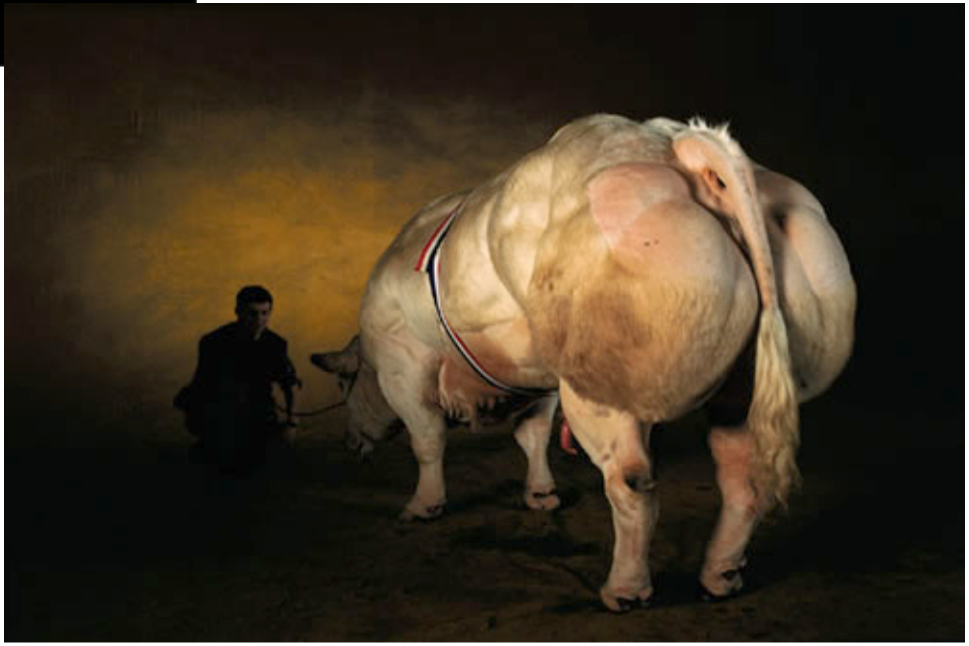
ADH in *Drosophila*. Strong candidate for balancing selection of the Fast and Slow alleles, due to a single aa replacement at the location marked by the arrow



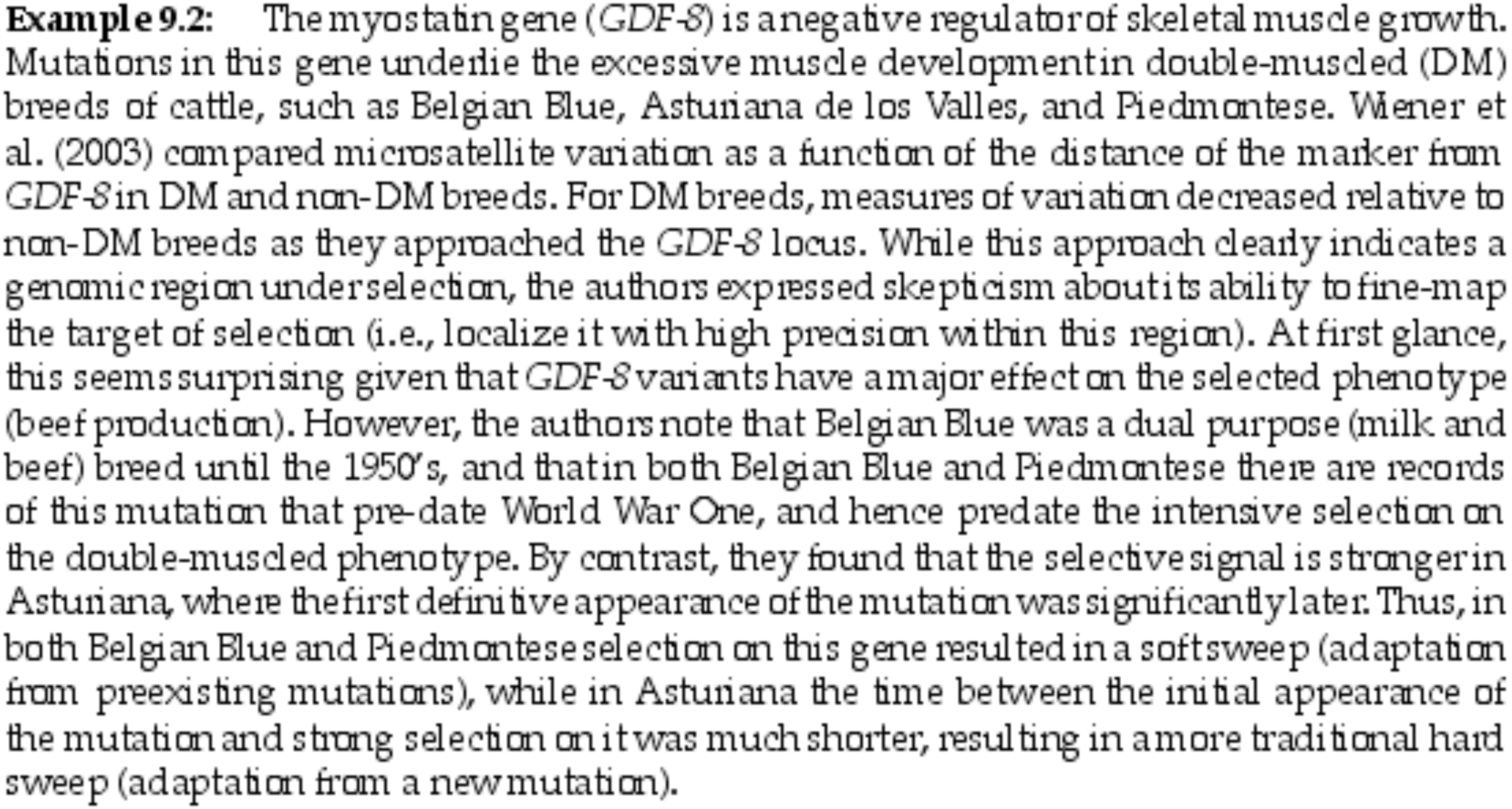
Scan of *Drosophila* genes in Africa (source population) and Europe (recently founded population). Less diversity in Europe, but some loci (filled circles) strong candidates for a sweep



Double-muscle cattle:
Belgian blue



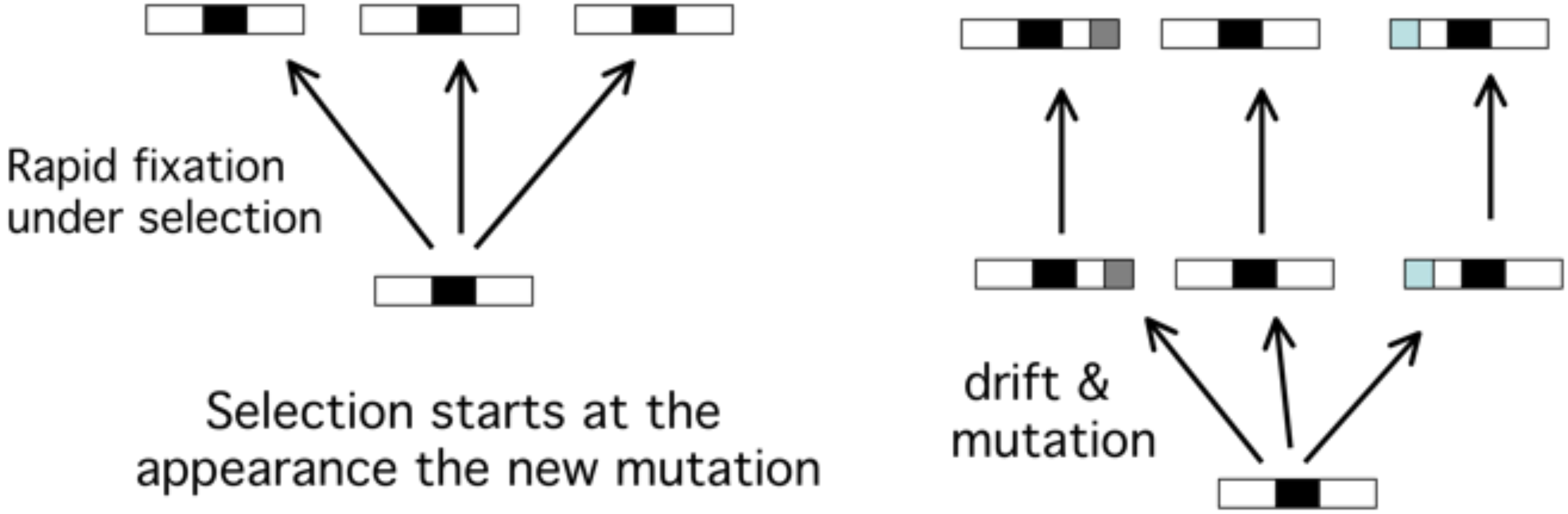
Reduction in microsatellite copy number variance often used



Example 9.2: The myostatin gene (*GDF-8*) is a negative regulator of skeletal muscle growth. Mutations in this gene underlie the excessive muscle development in double-muscled (DM) breeds of cattle, such as Belgian Blue, Asturiana de los Valles, and Piedmontese. Wiener et al. (2003) compared microsatellite variation as a function of the distance of the marker from *GDF-8* in DM and non-DM breeds. For DM breeds, measures of variation decreased relative to non-DM breeds as they approached the *GDF-8* locus. While this approach clearly indicates a genomic region under selection, the authors expressed skepticism about its ability to fine-map the target of selection (i.e., localize it with high precision within this region). At first glance, this seems surprising given that *GDF-8* variants have a major effect on the selected phenotype (beef production). However, the authors note that Belgian Blue was a dual purpose (milk and beef) breed until the 1950's, and that in both Belgian Blue and Piedmontese there are records of this mutation that pre-date World War One, and hence predate the intensive selection on the double-muscled phenotype. By contrast, they found that the selective signal is stronger in Asturiana, where the first definitive appearance of the mutation was significantly later. Thus, in both Belgian Blue and Piedmontese selection on this gene resulted in a soft sweep (adaptation from preexisting mutations), while in Asturiana the time between the initial appearance of the mutation and strong selection on it was much shorter, resulting in a more traditional hard sweep (adaptation from a new mutation).

Issues with sweeps

- Need sufficient background variation before selection for a strong signal
  - Strong domestication event (e.g. sorghum) can remove most variation over entire genome
  - Inbreeding greatly reduces variation
- The signal persists for only a short time
  - ~ 0.1 Ne generations
  - Distance for effects roughly 0.01 s/c
- Sweep region often asymmetric around target site
- Hard sweeps can be detected, soft sweeps leave (at best) a weak signal



A) Hard Sweep

Rapid fixation under selection

Selection starts at the appearance the new mutation

B) Soft Sweep

Rapid fixation under selection

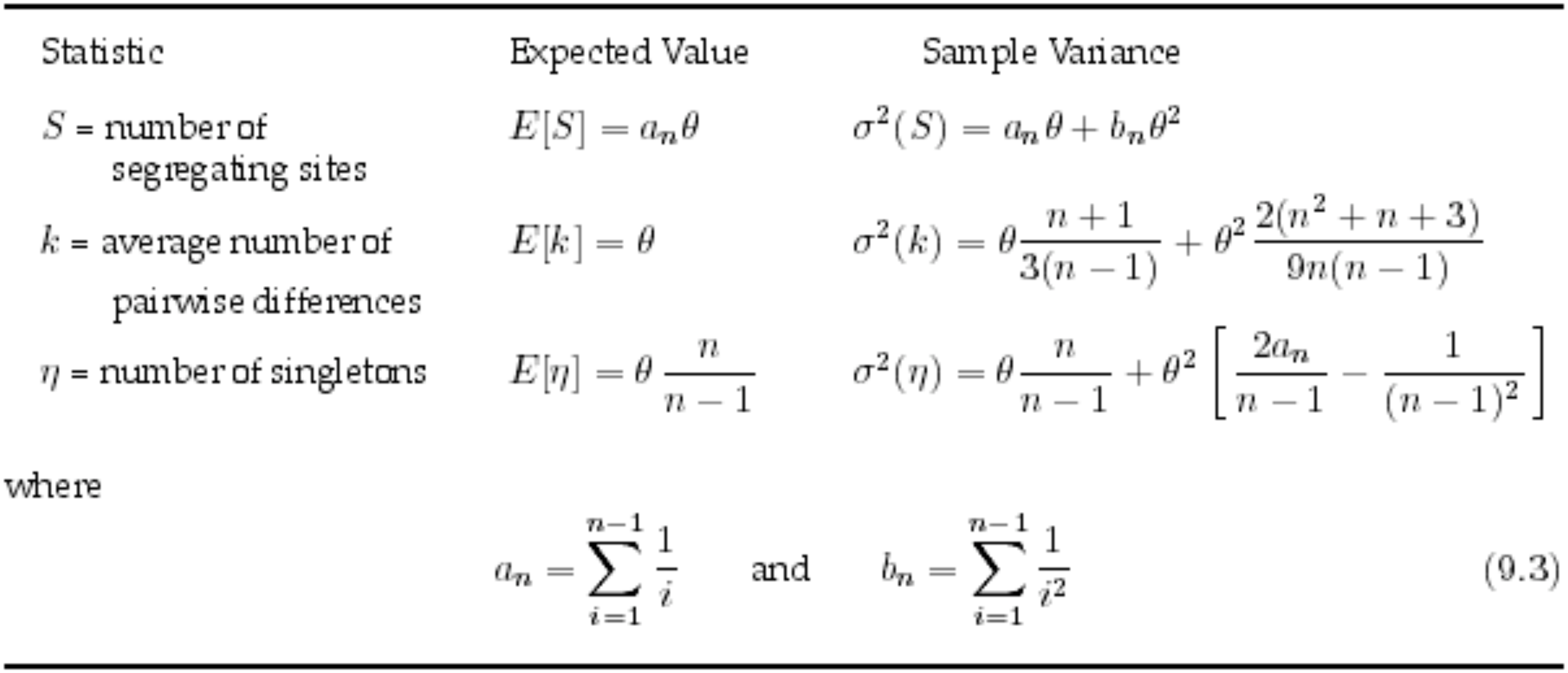
drift & mutation

Initially, new mutation neutral

Site frequency spectrum tests

- A large collection of tests based on comparing different measures of variation at a target site within a population sample
- Tajima's D is the classic
- Problem: significant result from either selection OR from changes in population size/structure (drift, mutation NOT at equilibrium)

Under the equilibrium neutral model, multiple ways to estimate $\theta = 4N_e u$ using different metrics of variation



Statistic	Expected Value	Sample Variance
S = number of segregating sites	$E[S] = a_n\theta$	$\sigma^2(S) = a_n\theta + b_n\theta^2$
k = average number of pairwise differences	$E[k] = \theta$	$\sigma^2(k) = \theta\dfrac{n+1}{3(n-1)} + \theta^2\dfrac{2(n^2+n+3)}{9n(n-1)}$
η = number of singletons	$E[\eta] = \theta\,\dfrac{n}{n-1}$	$\sigma^2(\eta) = \theta\dfrac{n}{n-1} + \theta^2\left[\dfrac{2a_n}{n-1} - \dfrac{1}{(n-1)^2}\right]$

where

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i} \qquad \text{and} \qquad b_n = \sum_{i=1}^{n-1}\frac{1}{i^2} \tag{9.3}$$

$$\widehat{\theta}_S = \frac{S}{a_n}, \qquad \widehat{\theta}_k = k, \qquad \widehat{\theta}_\eta = \frac{n-1}{n}\,\eta$$

All should be consistent if model holds.

Tajima's D

$$D = \frac{\widehat{\theta}_k - \widehat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}}$$

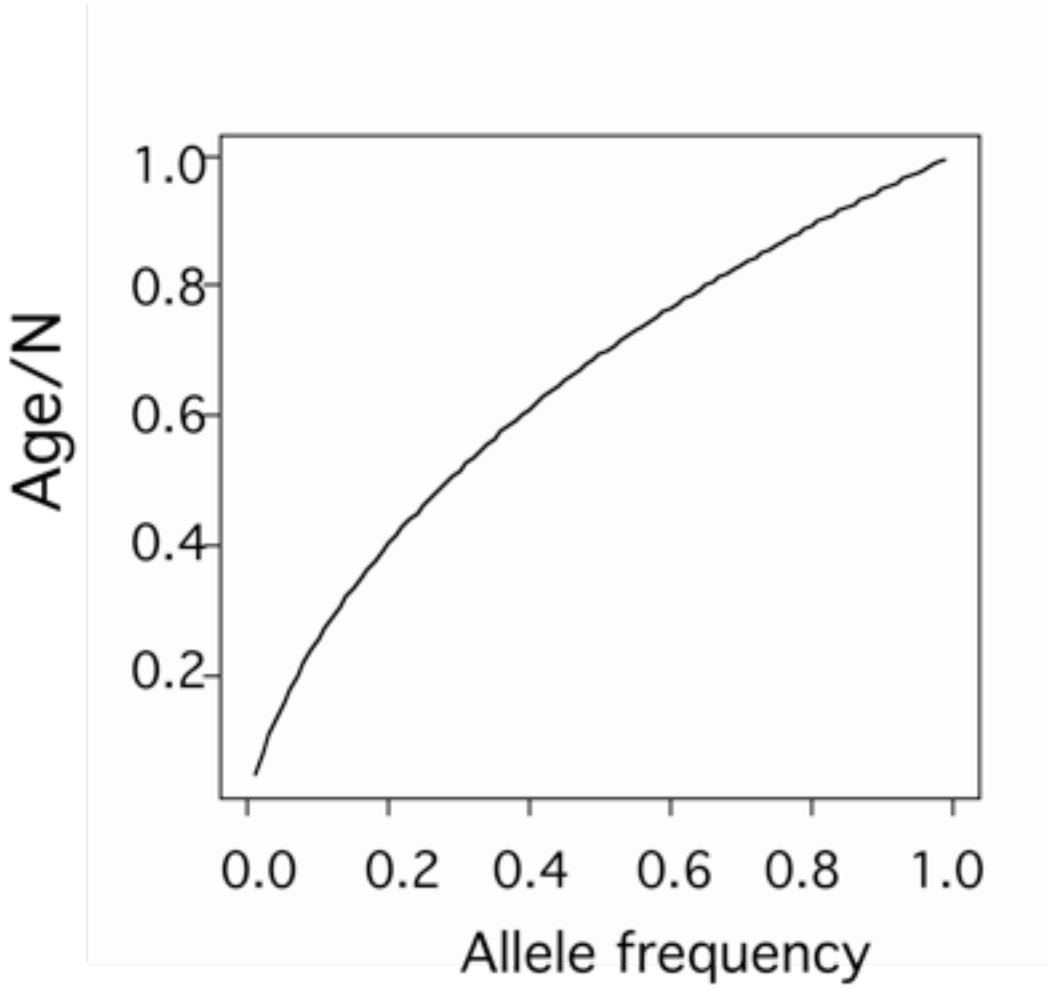
$$\alpha_D = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) - \beta_D$$

$$\beta_D = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right)$$

Negative value: excess number of rare alleles consistent with either positive selection OR expanding population size

Positive value: excess number of common alleles consistent with either balancing selection OR Population subdivision

Consistency of allelic age



$$E(t) = -4N\frac{x}{1-x}\ln(x)$$

Under drift, a common allele is an old allele

Common alleles should not be young

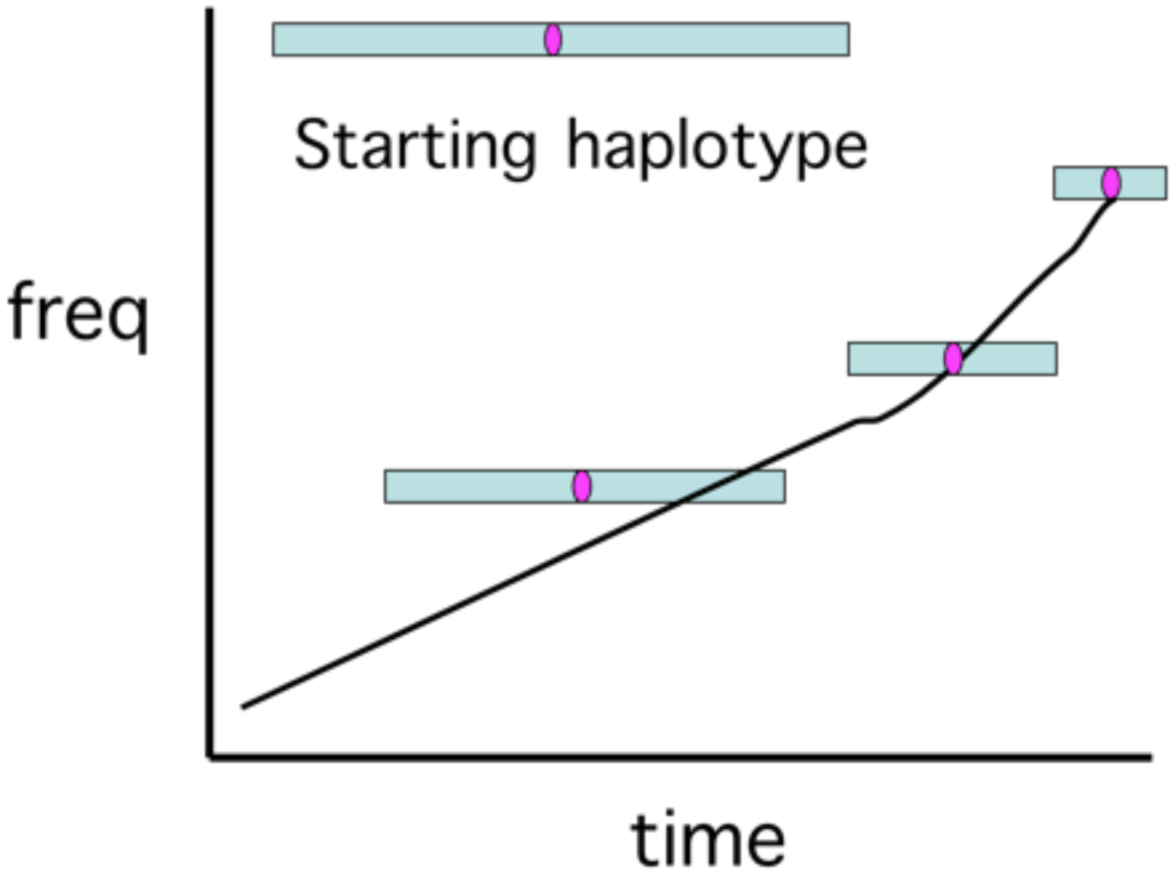
Example 9.4. The mutation *CCR5-δ32* destroys the *CCR5* receptor which is also used by the HIV virus, leading to significant resistance against HIV infection. This deletions occurs at frequencies up to 14% in Eurasia, but is absent in Africans, Native Americans and East Asians. Assuming a frequency of $x = 0.10$ and an effective population size $N_e = 5000$ for Caucasians, Stephens et al. (1998) used Equation 9.1 to estimate the age of this allele, based on its frequency, as

$$\widehat{t} = -4N_e \frac{x\ \log(x)}{1-x} = -4 \cdot 5000 \frac{0.1\ \log(0.1)}{0.9} = 5116 \text{ generations}$$

An independent estimate of age is offered by the variation in haplotypes among all sequences carrying this mutation. The δ mutation is in strong disequilibrium with allele *215* at the *AFMB* STR marker, to the extend that 84.8% (39 of 46) of the sampled δ mutations have the *δ32-215* haplotype. Clearly, the δ mutation at *CCR5* arose on a chromosome carrying the *215* allele. The recombination fraction between *CCR5* and *AFMB* was estimated by Stephens et al. (1998) to be $c = 0.006$. Using a calculation identical to that used in linkage disequilibrium mapping (LW Chapter 14), the probability p of a haplotype remaining intact after τ generations of recombination with fraction c is just $p = (1-c)^\tau$, or

$$\tau = -\log(p)/c = -\log(0.848)/0.006 = 27.5 \text{ generations}$$

Stephens et al. (1998) took these great disparities between age estimates as an indicator of strong selection on the δ mutation, generating much a higher frequency (under drift) for δ that expected from its age. Assuming it originated a single mutation, they estimated the selection coefficient to be between 20% and 40%, depending on assumptions about dominance.



Common alleles should have short haplotypes under drift -- longer time for recombination to act

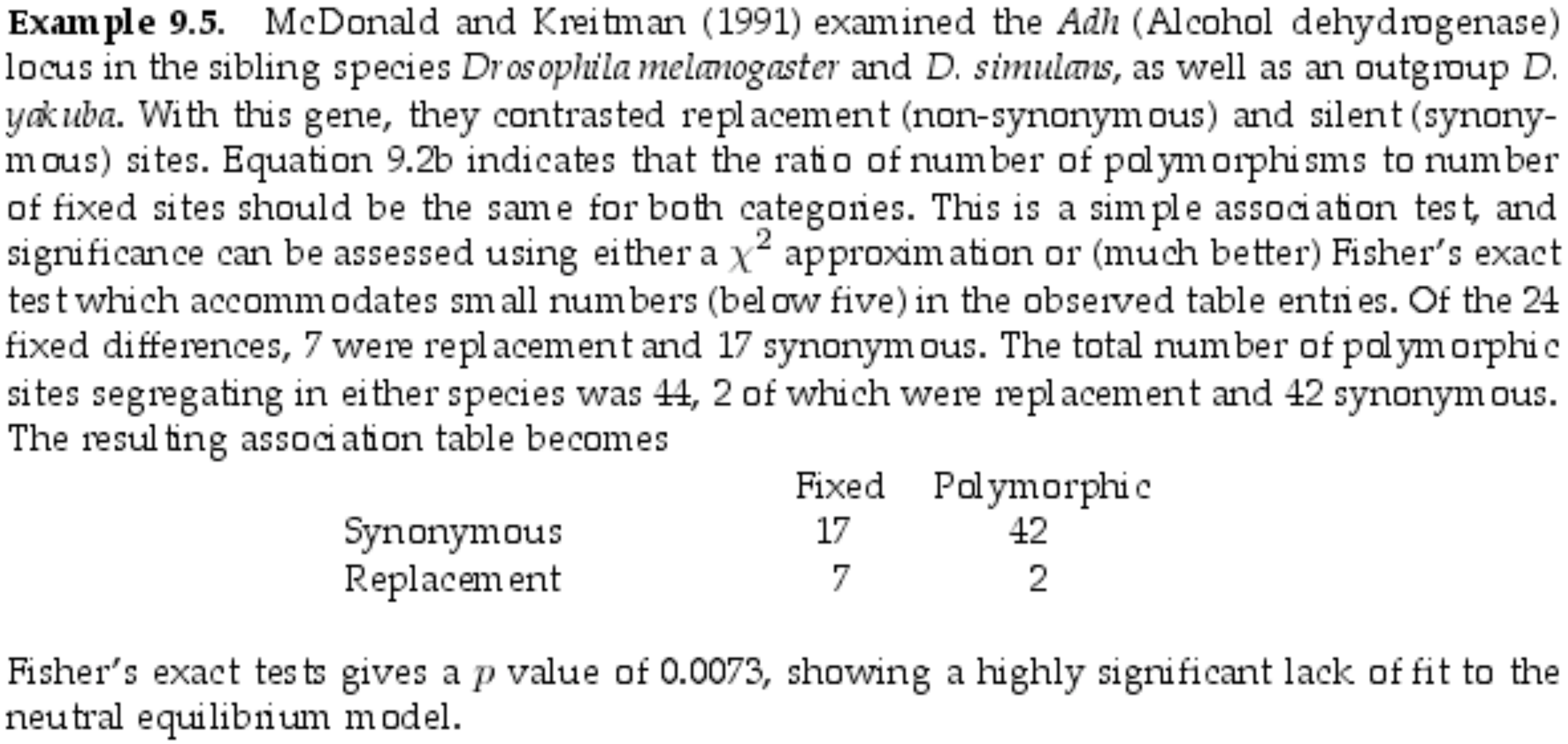
Common alleles with long haplotypes --- good signal for selection, rather robust to demography

Joint polymorphism-divergence tests

- HKA, McDonald-Kreitman (MK) tests
  - MK test is rather robust to demographic issues
- Require polymorphism data from one (or more) species, divergence data btw species
- Look at ratio of divergence to polymorphism

$$H_i = 4N_e\mu_i, \qquad d_i = 2t\mu_i$$

$$\frac{H_i}{d_i} = \frac{4N_e\mu_i}{2t\mu_i} = \frac{2N_e}{t}$$



Example 9.5. McDonald and Kreitman (1991) examined the *Adh* (Alcohol dehydrogenase) locus in the sibling species *Drosophila melanogaster* and *D. simulans*, as well as an outgroup *D. yakuba*. With this gene, they contrasted replacement (non-synonymous) and silent (synonymous) sites. Equation 9.2b indicates that the ratio of number of polymorphisms to number of fixed sites should be the same for both categories. This is a simple association test, and significance can be assessed using either a χ^2 approximation or (much better) Fisher's exact test which accommodates small numbers (below five) in the observed table entries. Of the 24 fixed differences, 7 were replacement and 17 synonymous. The total number of polymorphic sites segregating in either species was 44, 2 of which were replacement and 42 synonymous. The resulting association table becomes

	Fixed	Polymorphic
Synonymous	17	42
Replacement	7	2

Fisher's exact tests gives a p value of 0.0073, showing a highly significant lack of fit to the neutral equilibrium model.

Cool feature: can estimate # of adaptive substitutions

$$= 7 - 17(2/42) = 6$$

Robust to most demographic issues

However, replacement polymorphic sites can overestimate neutral rate due to deleterious alleles segregating

Strengths and weaknesses

- Only detects a pattern of adaptive substitutions at a gene.
  - Require multiple events to have any power
  - Can't tell which replacements were selectively-driven
- MK test robust to many demographic issues, but NOT fool-proof
  - Any change in the constraints between processes generating polymorphisms and processes generating divergence can be regarded as evidence for selection

Example 9.A6: An example in some of the potential difficulties in interpreting the results of a McDonald-Kreitman test is seen in Harding et al. (2000), who examined the human Melanocortin 1 receptor (*MC1R*), a key regulatory gene in pigmentation. Comparing the canonical *MC1R* haplotype in humans with a sequence from Chimp found 10 nonsynonymous (replacement) and 6 synonymous (silent) substitutions. An African population sample found zero nonsynonymous and 4 synonymous polymorphisms. The resulting DPRS table becomes

	Fixed (Human-Chimp)	Polymorphic (African)
Silent	6	4
Replacement	10	0

Fisher's exact test gives a p value of 0.087, close to significance. Taken on face value, one might assume that this data implies that the majority of the nonsynonymous substitutions between human and chimp were selectively-driven. However, the authors also had data from populations in Europe and East Asia, which showed ten nonsynonymous and three synonymous polymorphisms, giving the DPRS table as

	Fixed (Human-Chimp)	Polymorphic (Europe/East Asia)
Silent	6	3
Replacement	10	10

with a corresponding p value of 0.453. The authors suggest that the correct interpretation of these data is very stringent purifying selection due to increased functional constraints in African populations, with a release of constraints in Europe and East Asian. Asians in Papua New Guinea and India also showed very strong functional constraints, again consistent with a model of selection for protection against high levels of UV.

K_A/K_s tests

- THE classic test for selection, requiring gene sequences over a known phylogeny
  - K_A = replacement substitution rate
  - K_s = silent substitution rate
    - Neutral proxy
  - $\omega = K_A/K_s$
- $\omega > 1$: positive selection.
  - Problem: most codons have $K_s > K_A$, so that even with repeated adaptive substitutions throughout a gene, signal still swamped.

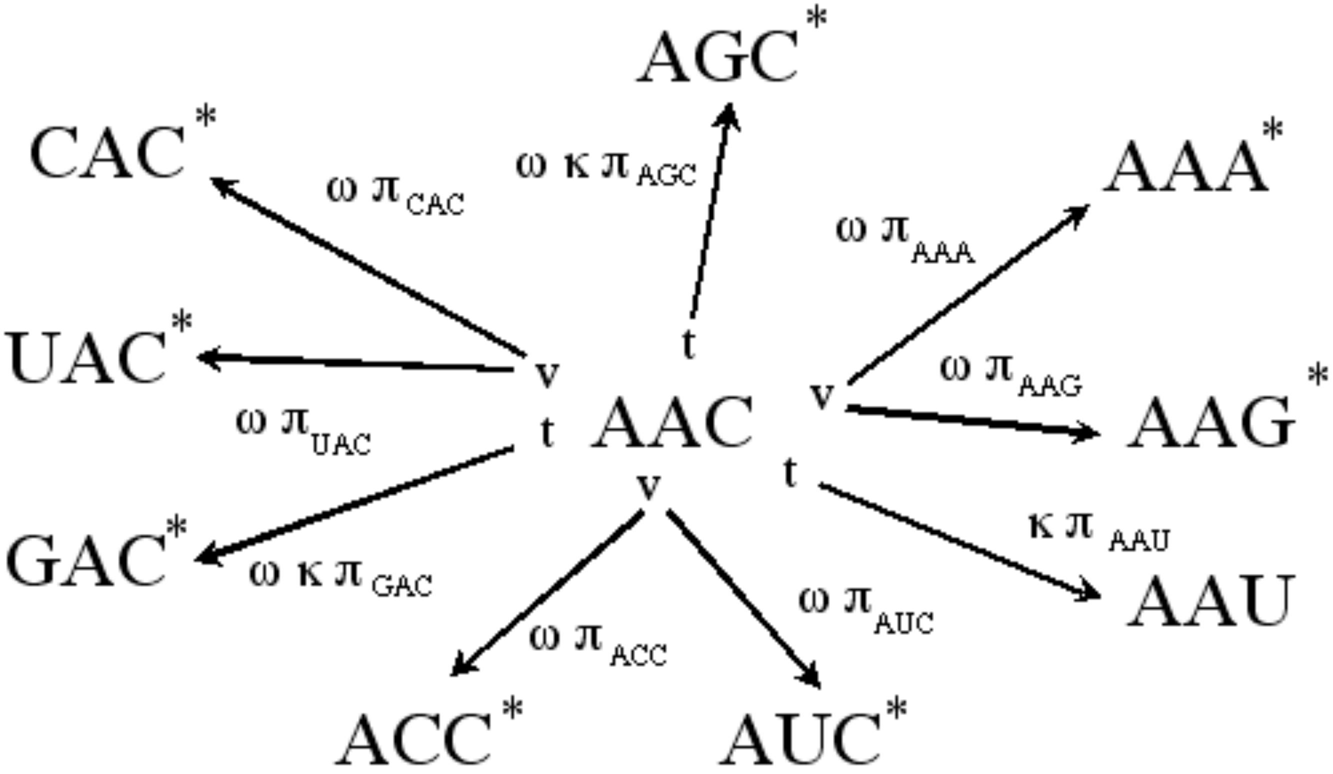
Example 9.6. One of the classic early examples of using sequence data to detect signatures of positive selection is the work of Hughes and Nei (1988, 1989) on mice and human major histocompatibility complex (MHC) Class I and Class II loci. These loci are highly polymorphic and are involved in antigen-recognition. Hughes and Nei compared the ratio of synonymous to nonsynonymous nucleotide substitution rates in the putative antigen-recognition sites versus the rest of these genes. For both classes of loci, they found a significant excess of nonsynonymous substitutions in the recognition sites and a significant deficiency of such substitutions elsewhere. If both types of substitutions are neutral, the rates per site are expected to be roughly equal. If negative selection is acting, the expectation is that the synonymous substitution rate would be significantly higher (reflecting removal of deleterious nonsynonymous mutations, as these change amino acids). However, if positive selection is common for many new mutations, then one would expect to see an excess of nonsynonymous substitutions. The observed patterns for both Class I and II loci were consistent with positive selection within that part of the gene coding for the antigen recognition site and purifying selection for the rest of the gene.

A large number of studies prior to Hughes and Nei found that an excess of nonsynonymous substitutions is by far the norm for almost all genes, implying that most nonsynonymous changes are selected against. Indeed, when one simply looks over an entire Class I (or II) MHC gene, this pattern is also seen. The insight of Hughes and Nei was to use data on protein structure to specifically focus on the putative antigen-binding site, and compare this region with the rest of the gene as an internal control. Further, there has to be a consistent pattern of new mutations being favored at the same few sites for such a signature to appear. A single favorable new mutation here and there through the evolution of a gene, when set against the background of most nonsynonymous mutants being deleterious, will still leave an overall signature of a vast excess of synonymous substitutions. Hughes and Nei concluded that a significant number of the new mutations that appear within the antigen-binding site are indeed favorable.

Codon-based models

- The way around this problem is to analyze a gene on a codon-by-codon basis
  - Such codon-based models assign all (nonstop) codons a value from 1 to 61
  - A model of transition probabilities between all one-nucleotide transitions is constructed
  - Maximum likelihood used to estimate parameters
  - Model with $\omega = 1$ over all codons contrasted with a model where $\omega > 1$ at some (unspecified) set of codons.

$$q_{ij} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega\pi_j & \text{for a nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases} \qquad \text{for } 1 \leq i, j \ \leq 61$$



Model easily expanded to allow for several classes of codons

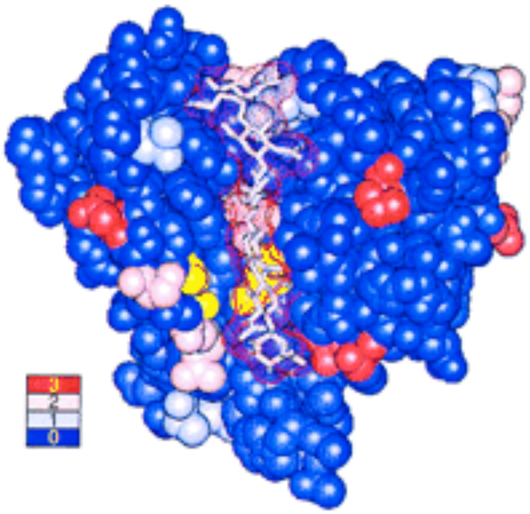
$$q_{ij}^{(k)} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega^{(k)}\pi_j & \text{for a nonsynonymous transversion} \\ \omega^{(k)}\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases}$$

$$\omega^{(k)} = \begin{cases} 0 & \text{deleterious class} \\ 1 & \text{neutral class} \\ \omega > 1 & \text{positively-selected class} \end{cases}$$

Can use Bayes' theorem to assign posterior probabilities that a given codon is in a given class (i.e., localize sites of repeated positive selection

$$\Pr(\text{class } i \,|\, D) = \frac{\Pr(D \,|\, \omega_i)\Pr(\text{class } i)}{\Pr(D)} = \frac{\Pr(D \,|\, \omega_i)\Pr(\text{class } i)}{\sum_{i=1}^{k}\Pr(D \,|\, \omega_i)\Pr(\text{class } i)}$$

Example 9.B. Bishop et al. (2000) examined the class I chitinase genes from 13 species of mainly North American *Arabis*, a crucifer closely related to *Arabidopsis*. Chitinase genes are thought to be involved in pathogen defense, as they destroy the chitin in cell walls of fungi. Many fungi have evolved resistance to certain chitinases, so these genes are excellent targets for repeated cycles of evolution. The authors found that phylogenies estimated by different methods all yielded similar results. Codon evolution models estimated that between 64 and 77% of replacement substitutions were deleterious, with 5-14% advantageous. These favored sites had an estimated value of $\omega = 6.8$. Using the criteria of a posterior probability of membership in the advantageous class in excess of 0.95 (i.e. $\Pr(\text{selective class} \mid D) > 0.95$), 15 putative sites were located. Seven of these sites involved only one alternative substitution, which evolved multiple times over the phylogeny. The authors had access to the three dimensional structure of chitinase, which shows a distinctive cleft, thought to be the active site. Mapping putative sites of positive selection onto this structure, the authors found a significant excess of sites cluster at the cleft, as opposed to the rest of the protein (28% of cleft sites versus 19% elsewhere). This example shows the power of combining this approach with solid biological data, and also care in checking the robustness of the methods by doing the analysis over slightly different phylogenies.



Class I Chitinase *(Arabis)*

Strengths and weaknesses

- Strengths
  - Can assign repeated selection to SPECIFIC codons
  - Requires only single sequences for each species
- Weaknesses:
  - Models can be rather delicate
  - Can only detect repeated selection at particular codons, NOT throughout a gene

The spandrels of San Marco (Gould and Lewontin 1979)

Very elaborate structure DOES not imply function nor adaptation

Structure vs. function

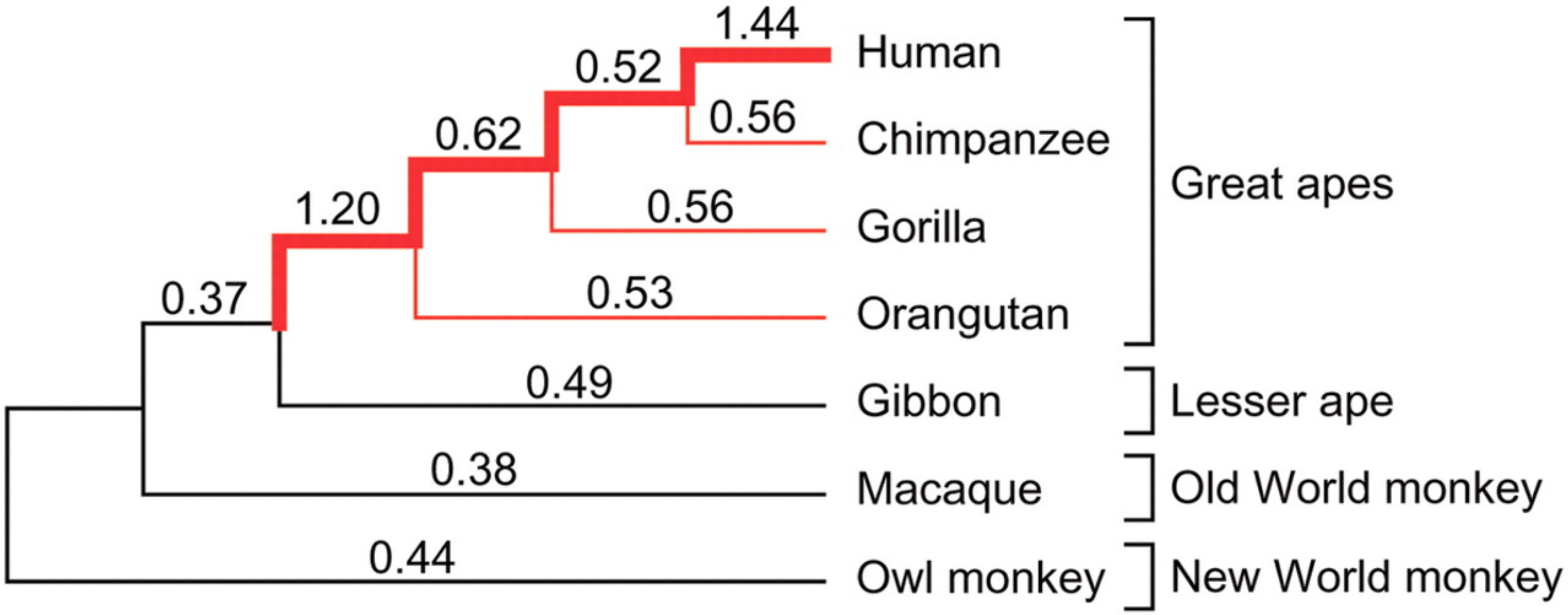
- Molecular biologists are largely conditioned to look for function through structure
- Problem: elaborate structures can serve little function
- Cannot simply assume an adaptive explanation because the structure is complex

Example 9.7. Humans show dramatic expansion of brain size with respect to most mammals, with this increase in (relative) size usually assumed to be corrected with increased cognitive abilities. Primary microcephaly is a condition in humans resulting in small heads, but othernormal features. Nonfunctional alleles at the genes *microcephalin* and *ASPM* (abnormal spindle-like microcephaly associated) both display the microcephaly phenotypes, with a typical individual having a brain size of around 400 cm^3 (versus the normal 1400 cm^3,) comparable to that in early hominids. Not surprising, several studies have looked for selection on these genes within the primate lineage. Zhang (2003) inferred a K_a/K_s ratio of 1.03 on the branch from the human-chimp common ancestor to humans, but a ratio of 0.66 on the branch from this ancestor to chimps. Values of 0.43 to 0.29 were found along other branches in mammals, suggested positive selection along the human lineage. Evans et al. (2004a) also examined *ASPM* over a larger phylogeny ranging from new world monkeys through humans. Accelerated ($K_a/K_s > 1$) rates of evolution were seen between gibbons and the ancestor the great apes, and a large acceleration ($K_a/K_s = 1.44$) was seen on the linkage from the human/chimp ancestor to humans. Evams et al. also performed a McDonald-Kreitman test (Example 9.5), comparing the polymorphisms within humans to the divergence since the human-chimp common ancestor, finding

	Fixed	Polymorphic
Synonymous	7	10
Replacement	19	6

Fisher's exact testgives a p value of 0.01, with an excess of around 15 replacement substitutions over what is expected from the replacement/synonymous ratio seen in the polymorphism data.

ω values shown on braches



ASPM

Building on these strong observations of selection leading to the human lineage, Mekel-Bobrov et al. (2005) and Evans et al. (2005) searched for *ongoing* selection in these two genes, and found strong signals in each. Evans et al (2005) found that the *microcephalin* gene had one haplotype (associated with a replacement substitution) at much higher frequencies than the others, with extended linkage disequilibrium and small intra-allelic variation. Indeed, using intra-allelic variation, the age of this haplotype was estimated at 37 thousand years (with a range of 14 to 60 thousand). Young alleles at high frequencies are hallmark indicators of positive selection (Example 9.4). Extensive coalescent simulations using a variety of population structures all gave high levels of significance to these results. The exact pattern, perhaps even more striking, was seen by Mekel-Bobrov et al. with *ASPM*: a common haplotype with long LD and a very recent estimated origin (5,800 years). Again, coalescent simulations of neutral drift under a variety of proposed models of human population growth and expansion showed these results to be highly significant. Together, these studies strongly suggested on-going selection in these two genes. They gathered a significant amount of attention, not the least of which was do to the finding that the putative adaptive haplotypes were in higher frequencies in Europe and Asia relative to Africa, and the connection that is often drawn between cognition and brain size.

Although Evans et al. (2005) cautioned that "it remains formally possible that an unrecognized function of *microcephalin* outside the brain is actually the substrate of selection", many interpreted the above data as an adaptive response in intelligence. After all, two functional genes that both influence brain size, a presumed correlate of intelligence, coupled with a history of past, and ongoing, selection does indeed suggest a case for selection on intelligence. This view, however, was quickly dispelled. Timpson et al (2007) and Mekel-Bobrov et al. (2007) showed in large sample sizes (900 and 2400, respectively) that there was no correlation between the putative adaptive halplotypes and increased intelligence. Any on-going selection on these genes does not appear to correlate with any selection for increased cognition. Currant et al. (2006) further noted that *spatial* models of growth were not considered, and here it is possible to see the above patterns for mutations that arise along the leading lead of a recent population expansion.