Lecture 20
Details on specific tests for selection

Bruce Walsh lecture notes
Synbreed course
version 8 July 2013
Key concepts

• There are a huge number of tests, using different information and hence testing to different features

• Most use the equilibrium neutral model
  - Alleles are effectively neutral, and mutation-selection balance has been reached (equilibrium assumption)

• Relatively recent changes in $N_e$ violate equilibrium assumption, can lead to false positives.

• Search for demography-independent tests largely unsuccessful.

• Tests can be based on within-sample variation or divergence data (allele frequency shifts over short time scales, fixations over longer time scales).
Detecting selection

• Bottom line: looking for loci showing departures from the equilibrium neutral model

• For breeders, provides a contrast to QTL mapping --- the search for genetic regions showing signatures of selection (unlike QTL mapping) does not require us to specific traits of interest.

• What kinds of selection are of interest?

• 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, including domestication and improvement genes as special cases
• 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

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

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.
Fixed differences between two species

Positive selection occurring over 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).
Sample of a gene over several species

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

A phylogenetic 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
Key points

• Methods for detecting selection
  - Are prone to false-positives
    • The rejection of the null (equilibrium neutral model) can occur for reasons other than 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 ongoing (or just finished) selection

Single site under either positive or balancing selection
Detecting on-going selection

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

These to compare different measures of variation (such as number of alleles vs. pair-wise distances among alleles)

Two sequence evolution frameworks are typically used: infinite alleles vs. infinite sites.

Both assume each new mutation generates a new (unique) sequence. (such is not the case for STRs)

How do these frameworks differ?
Consider the following five sequences (rows)

A A G A C C 1  Infinite alleles: Treat each
A A G G C C 2  different haplotype as a
different allele
A A G A C C 1
A A G G C C 2  Here, there are three alleles
A A G G C A 3

Infinitesites model: Treat each site (base
position = columns) separately. How many
polymorphic sites are there?

Here, 2 polymorphic sites
Site and allele frequency spectrum

- The **allele frequency spectrum** (AFS) is simply the distribution of the number of different types of allele at a locus in our sample
  - For example, in 20 sequences, 10 singletons, 2 present each as 3 copies, and one as four copies
  - $n_1 = 10$, $n_3 = 2$, $n_4 = 1$, all other $n_i = 0$, $\sum i*n_i = n$
  - Feature of a specific LOCUS

- The **site frequency spectrum** (SFS) is the number of sites in a sample of $n$ sequences with $k$ copies of the derived allele (unfolded SFS) or $k < n/2$ copies of the minor allele (folded SFS)
  - $n_j$ is the number of sites with $j/n$ derived/minor alleles.
  - Feature of a population of SITES
Detecting Hard Sweeps

• Features of a hard sweep
  - A single new mutation sweeps to fixation

• Pattern of variation
  - Changes in the site frequency spectrum

• Pattern of LD

• Sweep-specific tests
  - Bottleneck models
  - Spatial pattern of variation
  - Sweep-specific likelihood models
Table 10.1. Summary of various features associated with a selective sweep for a locus with fitnesses $1 : 1 + 2hs : 1 + 2s$ (for $h \neq 0$). Let $q$ denote the frequency of a neutral marker at the start of selection that is at distance $c$ from a strongly selected site ($4N_es >> 1$). Assume the frequency of the favorable allele is $p_0$ at the start of selection, and let $q_h$ and $H_h$ denote the final allele frequency for a marker allele associated with $A$ and the heterozygosity at the site immediately following the sweep.

Fraction $f$ of initial associations remaining after fixation:

$$f \simeq \begin{cases} 
(p_0)^{-c/(2hs)} \simeq 1 - \frac{c}{2hs} \log(p_0) & \text{for } p_0 >> 1/(2N) \\
(4N_es)^{-c/(2hs)} \simeq 1 - \frac{c}{2hs} \log(4N_es) & \text{for } p_0 = 1/(2N) 
\end{cases}$$

Total change $\Delta q$ in the frequency of a linked marker allele:

$$\Delta q \simeq (1 - q)f$$

Final frequency of a linked marker:

$$q_h = q + \Delta q = f + q(1 - f)$$

Reduction in heterozygosity immediately following the sweep:

$$\frac{H_h}{H_0} = 1 - f^2$$

Heterozygosity $t$ generations after a sweep:

$$\frac{H(t)}{H_0} = 1 - f^2 e^{-t/(2N_e)}$$

Can use length of reduction in variation to estimate $s$ given we know $c$. 20
Estimating strength of selection from size of sweep region

Kaplan, Hudson, and Langley (1989) showed that the distance $d$ at which a neutral site can be influenced by a sweep is a function of the strength of selection $s$ and the recombination fraction $c$, with $d = 0.01 \frac{s}{c}$.

Hence, $s = 100 \cdot d \cdot c$

For $tb1$, $s \rightarrow 0.05$.

With $s$ in hand, one can also estimate the expected time for selection to fix the allele, which Wang et al. estimated at 300 to 1000 years, indicating a fairly long period of domestication.
Example: *Waxy gene in Rice (Olsen et al. 2006)*

“Sticky” (glutinous) rice results from low amylose levels, and are typical of temperate japonica variety groups.

A number of groups showed this is due to a splice mutant in the Waxy gene. This is an example of an improvement (as opposed to domestication) gene

Olsen et al. observed a region 250kb in size around Waxy with a greatly reduced level of polymorphism compared to control populations.

Using the Kaplan et al. expression, this gives $s = 4.6!$
Variation declines around a sweep

Sweep signal around a candidate gene in Dachshunds for limb shortening (FGFR3)
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 suggest 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 frequent 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.
Five-marker window scans of difference between Holstein & Angus breeds (dairy vs. beef selection)
Distribution of allele frequencies under a sweep

Excess of low & high frequency sites
Site frequency spectrum (SFS)

- Under the equilibrium neutral model, the distribution of either the minor allele frequency (folded frequency spectrum) or the frequency of a derived allele (unfolded frequency spectrum) given by the Watterson distribution ($\theta/x$)

- A sweeps inflate the frequency of sites segregating rare alleles (folded spectrum)

- Sweeps inflate the frequency of derived alleles, increasing the number of high-frequency sites in the SFS

- Key idea: Can use different parts of the SFS to estimate theta
  - $\theta_s = S/a_n$, Watterson estimator, number of segregating sites $S$, an a constant based on sample size
  - $\theta_1 = S_1$, number of singletons in SFS
  - $\theta_\pi = \text{average pairwise difference } \pi \text{ btw two random sequences}$
Under the equilibrium neutral model, multiple ways to estimate \( \theta = 4Neu \) using different metrics of variation

\[
\begin{array}{lll}
\text{Statistic} & \text{Expected Value} & \text{Sample Variance} \\
S = \text{number of segregating sites} & E[S] = a_n \theta & \sigma^2(S) = a_n \theta + b_n \theta^2 \\
k = \text{average number of pairwise differences} & E[k] = \theta & \sigma^2(k) = \theta \frac{n + 1}{3(n - 1)} + \theta^2 \frac{2(n^2 + n + 3)}{9n(n - 1)} \\
\eta = \text{number of singletons} & E[\eta] = \theta \frac{n}{n - 1} & \sigma^2(\eta) = \theta \frac{n}{n - 1} + \theta^2 \left[ \frac{2a_n}{n - 1} - \frac{1}{(n - 1)^2} \right]
\end{array}
\]

where

\[
a_n = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and} \quad b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}
\]

\[
\hat{\theta}_S = \frac{S}{a_n}, \quad \hat{\theta}_k = k, \quad \hat{\theta}_\eta = \frac{n - 1}{n} \eta
\]

All should be consistent if model holds.
Tajima’s D

\[ D = \frac{\hat{\theta}_k - \hat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} \]

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

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

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
Changes in demography change \( D \)

**Figure 10.12** Distribution of Tajima’s \( D \) for 201 genes in African-American (A) and European-American (B) samples. The empirical distribution is given by the histogram, the solid line gives the simulated values under the equilibrium neutral model (allowing for recombination), and the dashed line gives the simulated distribution under the best-fitting demographic model. For (A) this is exponential growth starting 50,000 years ago, while for (B) this is a bottleneck starting 40,000 years ago. After Ronald and Akey (2005).

**Correction:** use the distribution for genome-wide tests to “adjust” distribution of \( D \) (akin to genomic control)
<table>
<thead>
<tr>
<th>Test</th>
<th>Contrast</th>
<th>Spectrum</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tajima's $D$</td>
<td>$\hat{\theta}<em>S$ vs. $\hat{\theta}</em>\Pi$</td>
<td>Folded</td>
<td>$&lt; 0$: Excess of rare alleles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sweep or population bottleneck</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$&gt; 0$: Excess of intermediate-frequency alleles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Balancing selection or population structure</td>
</tr>
<tr>
<td>Fu and Li's $D$</td>
<td>$\hat{\theta}_S$ vs. $\hat{\theta}_1$</td>
<td>Unfolded</td>
<td>Same as for Tajima's $D$</td>
</tr>
<tr>
<td>Fu and Li's $D^*$</td>
<td>$\hat{\theta}_S$ vs. $\hat{\theta}_1^*$</td>
<td>Folded</td>
<td>Same as for Tajima's $D$</td>
</tr>
<tr>
<td>Fu and Li's $F$</td>
<td>$\hat{\theta}_\Pi$ vs. $\hat{\theta}_1$</td>
<td>Unfolded</td>
<td>Same as for Tajima's $D$</td>
</tr>
<tr>
<td>Fu and Li's $F^*$</td>
<td>$\hat{\theta}_\Pi$ vs. $\hat{\theta}_1^*$</td>
<td>Folded</td>
<td>Same as for Tajima's $D$</td>
</tr>
<tr>
<td>Fay and Wu's $H$</td>
<td>$\hat{\theta}_\Pi$ vs. $\hat{\theta}_H$</td>
<td>Unfolded</td>
<td>$&lt; 0$: Excess of high-frequency derived alleles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sweep or allelic surfing</td>
</tr>
<tr>
<td>Zeng et al.'s $E$</td>
<td>$\hat{\theta}_\Pi$ vs. $\hat{\theta}_L$</td>
<td>Unfolded</td>
<td>$&lt; 0$: Excess of low- v. high-frequency derived alleles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signal of a recent past sweep</td>
</tr>
</tbody>
</table>
SFS tests

- Tajima’s D
  - Negative values (excess of rare alleles) following a sweep
  - Problem: changes in population size also generates this

- Fay and Wu’s H
  - Test for excess of high frequency derived alleles
  - This signal is unlikely due to demographic factors
SFS Power quickly dissipates

H loses power very quickly as high-frequency alleles following the sweep are fixed.
Distribution of LD around a hard sweep

At start, LD through sweep site. At fixation, strong LD on either side, but not through, a sweep site.
Summary: Signatures of a Hard Sweep

We can summarize the key findings of the population-genetics theory for a hard sweep as follows.

A recent or ongoing sweep leaves several potentially diagnostic signals:

1. An excess of rare alleles (in either the folded or unfolded frequency spectrum)
2. An excess of high frequency derived alleles
3. Depression of genetic variation around the site, often asymmetric with respect to site of selection

Signatures in the spatial pattern of LD differ during the sweep and after its completion. When a favorable allele is at moderate frequencies (a partial sweep), we see

4a. An excess in LD throughout the region surrounding the sweep

Following fixation of the favorable allele, the spatial pattern is rather different,

4b. An excess in LD on either side of the site, but a depression in LD around the site

Finally,

5. Signatures of a sweep are very fleeting, being on the order of $0.5N_e$ generations for signature (1), $0.4N_e$ generations for (2), $1.4N_e$ generations for (3) and $0.1N_e$ generations for (4B)
**Bottleneck tests**

**Example 10.10.** Wright et al (2005) used a multiple bottleneck approach in their search for genes under selection in maize. The authors used SNP data on 774 genes from 14 maize and 16 teosinte inbred lines. The sampled maize lines contained roughly 60% of the variation as the teosinte lines, showing a strong bottleneck signal across the entire maize genome. The authors quantified the strength of the bottleneck by \( b = N_b / d \), the ratio of the size of the population in the bottleneck divided by its duration. Using just the teosinte data, \( \theta_i \) was obtained using Watterson’s estimator (Equation 10.20c), and these were used in a likelihood function \( L(b|S) \) to estimate \( b \) based on the observed number of segregating sites \( S \) for each locus in maize. The likelihood \( L(b|S) \) just corresponds to the probability of the data \( S \) given \( b \), \( p(S|b) \), see LW Appendix 4. For the base model with a single \( b \) over all loci, the likelihood is just the product of \( L(b|S_i) \) over the sampled loci. The resulting maximum likelihood estimate (MLE) was \( \hat{b} = 2.45 \). The authors then assumed that two classes of loci exist in the populations: a fraction \( (1 - q) \) experiencing a bottleneck of strength \( b_1 \) and a fraction \( q \) experiencing a much stronger bottleneck of strength \( b_2 < b_1 \), giving the resulting likelihood for locus \( i \) as

\[
L(f, b_1, b_2|S_i) = (1 - q)L(b_1|S_i) + qL(b_2|S_i)
\]  

\[(10.32a)\]
Again, the full likelihood is the product over all loci. This is a mixture model (LW Chapter 13), with parameters $q$, $b_1$ and $b_2$. The resulting MLEs were $\hat{b}_1 = 2.45$, $\hat{b}_2 = 0.15$, $\hat{q} = 0.02$. However, many of the loci had low variation even in teosinte, and offer little information. Using a set of 275 genes with high variation (10 or more segregating sites in teosinte), $\hat{b}_1 = 2.45$, $\hat{b}_2 = 0.01$, $\hat{q} = 0.036$. In this sample of genes, almost 4% experienced a much greater bottleneck (smaller $b$) than the rest of the genome, and hence are strong candidate for sites that were influenced by a sweep. With these estimates in hand, one can use Bayes’ theory (Equation A2.2) to obtain the posterior probability of a locus being in this selected class, and hence can localize those genes potentially under past selection. Recall that Bayes’ theorem allows one to “flip” the condition, as we can easily compute $P(S_i \mid b_j)$ – indeed, this is the likelihood – but are much more interested in $P(b_j \mid S_i)$. Bayes’ theorem connects these as

$$\Pr(b_2 \mid S_i) = \frac{q P(S_i \mid b_2)}{(1 - q) P(S_i \mid b_1) + q P(S_i \mid b_2)} = \frac{q L(b_2 \mid S_i)}{(1 - q) L(b_1 \mid S_i) + q L(b_2 \mid S_i)} \quad (10.32b)$$

This gives a posterior probability for a particular locus (here $i$) being in the strong bottleneck class ($b_2$). This same approach for posterior prediction reappears in Example 11.16, where we are looking for actual sites in a protein under selection given divergence data.
Using spatial information (pattern of diversity along a chromosome) to detect sweeps

Likelihood of seeing $k_i/n$ derived alleles at a site $(i)$

\[
Pr(k_i \mid n, \Theta) = \binom{n}{k_i} \int_{1/(2N)}^{1-1/(2N)} x^{k_i} (1-x)^{n-k_i} \phi_i(x \mid \Theta) \, dx
\]

\[
\phi(x) = \begin{cases} 
\theta \left( \frac{1}{x} - \frac{1}{f} \right), & \frac{1}{2N} \leq x < f \\
0, & f \leq x \leq 1 - f \\
\frac{\theta}{f}, & 1 - f < x \leq 1 - \frac{1}{2N}
\end{cases}
\]

\[
f_i = (4N_e s)^{-c_i/(2hs)} = e^{-c_i \zeta}
\]

Key: varies in a defined way (i.e., with $c$) around the sweep
Several tests use this general approach

- **CLRT test**
  - Uses the Watterson distribution as the base SFS that is perturbed by the sweep

- **Sweepfinder**
  - Replaces the Watterson distribution for the SFS with an empirical SFS distribution given from the data

- **XP-CLRT**
  - Compares the $F_{st}$ values across two populations
Hard, partial, soft, & polygenic sweeps

- **Partial sweep**: favored allele not yet fixed, shift in SFS, generation of LD
- **Soft sweep**: multiple independent copies of the new mutation contribute to the sweep
  - Little signal in SFS, signal in certain LD features
- **Polygenic sweep**: small changes at a large number of loci
  - Almost no signal. Weak selection and small allele frequency change
LD-based sweep tests

**Kim and Nielson’s ω statistic:** LW with vs. across a test region

\[ \omega = C_{S,\ell} \frac{\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in L} r_{ij}^2}{\sum_{i \in L, j \in R} r_{ij}^2}, \quad C_{S,\ell} = \frac{1/(\ell(S-\ell))}{\binom{\ell}{2} + \binom{S-\ell}{2}} \]

Strong test for hard sweeps, little power for soft sweeps

**Kelly’s Z_{ns} Statistic:** Average pairwise LD through a region

\[ Z_{ns} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{ij}^2 \]

Strong test for soft sweeps, little power for hard sweeps
Haplotype-based tests

• A number of tests are based on haplotypes, rather than the distribution of individual sites
  - Based on infinite alleles, rather than infinite sites, model

• Can contrast number of haplotypes versus their frequency

• Can contrast estimates of age with haplotype frequency
  - Can look for long haplotypes at high frequency
  - Can look for recombination/STR age estimates different from frequency estimates
Consistency of allelic age

Under drift, a common allele is an old allele

Common alleles should not be young

\[ E(t) = -4N \frac{x}{1-x} \ln(x) \]
The mutation CCR5-δ32 destroys the CCR5 receptor, which is used by the HIV virus to enter the cell, leading to significant resistance against HIV infection. This deletion 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 2.12 to estimate the age of this allele as

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

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 extent 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 $q$ of a haplotype remaining intact after $\tau$ generations of recombination with fraction $c$ is $q = (1-c)^\tau$, giving

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

Stephens et al. took these great disparities between age estimates as an indication of strong selection on the δ mutation, generating much a higher frequency that expected from its age under a pure drift model. Assuming δ originated a single mutation, they estimated the selection coefficient of between 20% and 40%, depending on assumptions about dominance.
Ewen’s sampling distribution

For example, the number $k$ of different alleles in a sample of size $n$ is given by Ewens’ Sampling Formula (Evens 1972),

$$\Pr(k \mid \theta, n) = \frac{|S_n^k| \theta^k}{S_n(\theta)}$$

(10.19a)

where

$$S_n(\theta) = \theta(\theta + 1)(\theta + 2) \cdots (\theta + n - 1)$$

(10.19b)

and $S_n^k$ (a Stirling number of the first kind) is the coefficient on the $\theta^k$ term in the polynomial given by $S_n(\theta)$.

Further, the allele frequency spectrum, giving $k$, is independent of $\theta$:

$$\Pr(n_1, n_2, \ldots, n_k \mid n, k) = \frac{n!}{|S_n^k| 1^{n_1} 2^{n_2} \cdots n^{n_n} n_1! n_2! \cdots n_n!}.$$

This is used for several tests. These results depend on the equilibrium neutral assumptions.
Distribution of haplotypes

Several tests of selection are built around lack of fit with Ewens sampling distribution.

Observed vs. predicted allele frequency spectrum given $k$

Fit to Ewens distribution given independent estimate of $\theta$, for example using the Watterson estimator:

$$E[S] = a_n \theta,$$

where

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

This gives rise to the **Watterson estimator**

$$\hat{\theta}_w = \frac{S}{a_n}$$

which is widely used to estimate $\theta$ in a number of tests of selection.
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.
Long haplotypes

Expect new alleles to be RARE under drift. Look for new alleles at high frequency

One estimate of age offered by haplotype length: new alleles should have long haplotypes, older alleles shorter haplotypes

Haplotype homozygosity (HH): The probability that two randomly-chosen copy on an allele are identical (over some specified length)

Extended haplotype homozygosity (EHH): length of the region where the HH exceeds some threshold (say 0.05)
Relative EHH for allele $i$:

$$rEHH_i = \frac{EHH_i}{\text{ave}(EHH_j) \text{ for } j \neq i}$$
Test of selection: plot rEHH vs. allele frequency

Both genes (G6PD, CD) ligand showed no signature using SFS tests, likely because of partial/ongoing sweep.

Long haplotypes have power for detecting partial sweeps.
Summary of haplotype-based tests

Tests based on the allele-frequency spectrum (distribution of number of haplotype classes within a sample)

- Watterson’s Test: Observed allelic homozygosity vs. expected homozygosity under AFS(k)
- Slatkin’s exact Test: Observed AFS vs. expected AFS(k)
- Innan et al.’s HCT test: Observed AFS vs. expected AFS conditioned on observed S
- Hudson’s HP test: Frequency of most common haplotype given S
- Fu’s W: Test for deficiency of rare haplotypes given S
- Fu’s Fs: Test for excess of rare haplotypes given \( \hat{\theta}_n \) (average pair-wise difference estimator)
- Depaulis and Veuille’s K test: Observed number of haplotypes given S
- Depaulis and Veuille’s H test: Observed haplotype diversity given S

Tests based on averages of pairwise disequilibria

- Kelly’s \( Z_{m,n} \) test: Average of all pairwise disequilibria between all sites in a region
- Kim and Nielsen’s \( \omega_{max} \) test: Pairwise LD among sites within vs. between sides of a region

Tests based on frequency estimates of age vs. allelic diversity estimates of age

- Age estimated by decay of LD between allele and a linked marker
- Age estimated by number of segregating sites \( S \) within an allelic haplotype class
- Age estimated by copy number variance at tightly-linked STRs in the allelic class

Tests contrasting haplotype lengths of alternative alleles in the same population

- Sabeti’s \( rEHH \) test: ratio of the haplotype length \( (EHH) \) of two alternative alleles
- Voight’s \( iHS \) test: ratio of area under the \( EHH \) curve for ancestral vs. derived alleles
- Hanchard’s \( nHS \) test: standardized average haplotype diversity within a sliding window through a region

Tests contrasting haplotype lengths of the same allele in two populations

- Sabeti’s \( XP - EHH \), Tang’s \( \ln(Rsb) \) tests: ratio of area under the \( EHH \) curve in the different populations
- Kimura’s \( rHH \) vs. \( rMHH \) plots: ratios of overall \( HH, HH \) based on single most frequent haplotype
Detecting a history of repeated positive selection on a gene

Multiple substitutions have been fixed by positive selection within a gene
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

\[
H_i = 4N_e \mu_i, \quad d_i = 2t \mu_i
\]

\[
\frac{H_i}{d_i} = \frac{4N_e \mu_i}{2t \mu_i} = \frac{2N_e}{t}
\]
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 $\chi^2$ 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:

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>17</td>
<td>42</td>
</tr>
<tr>
<td>Replacement</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Fixed (Human-Chimp)</th>
<th>Polymorphic (African)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Replacement</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Fixed (Human-Chimp)</th>
<th>Polymorphic (Europe/East Asia)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Replacement</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

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 mouse 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.
Detecting a history of repeated positive selection at a site

Multiple substitutions have occurred at a site over a phylogeny of several species/very distant populations
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.
\[
q_{ij} = \begin{cases} 
0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\
\pi_j & \text{for a synonymous transversion} \\
\kappa \pi_j & \text{for a synonymous transition} \\
\omega \pi_j & \text{for a nonsynonymous transversion} \\
\omega \kappa \pi_j & \text{for a nonsynonymous transition} 
\end{cases} 
\text{ for } 1 \leq i, j \leq 61
\]
Model easily expanded to allow for several classes of codons

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

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

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)

\[
\Pr(\text{class } i \mid D) = \frac{\Pr(D \mid \omega_i) \Pr(\text{class } i)}{\Pr(D)} = \frac{\Pr(D \mid \omega_i) \Pr(\text{class } i)}{\sum_{i=1}^{k} \Pr(D \mid \omega_i) \Pr(\text{class } i)}
\]
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 mem-
bership 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 di-
mensional 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 signifi-
cant 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.
Strengths and weaknesses of Codon-based tests

• **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
Summary

• Lecture a **quick overview** of test of selection
  - Online notes (WL Chapters 9, 10) total over 140 pages, so MUCH detail left out

• **Keys:**
  - Most tests are NOT robust to demographic issues (e.g., change in population size) and must be suitably corrected (difficult)
  - Specific tests work well in specific time windows (e.g., during, right after, soon after a sweep).
  - NO ONE TEST WORKS BEST

• Potentially very useful to breeders
  - Genetic targets of past artificial selection
  - Can suggest candidate regions for response to natural selection (e.g., pathways for salt tolerance in a natural population under history of high salinity.)