Lecture 10
Tests for Molecular Signatures of Selection

Bruce Walsh. June 2008. Summer Institute on Statistical Genetics, Seattle

Changes in the level of neutral polymorphisms can provide signatures of recent selection.

There is a rich population genetics literature on tests for whether an observed pattern of polymorphism, or an observed between-species difference, or both, can be accounted for with a standard drift model. Rejection of this hypothesis offers the possibility that selection may play a role, but (as we will see) other forces (such as population history) can also cause strong deviations from the neutral model, especially when attempting to account for an observed pattern of polymorphism. While these tests have their roots in evolutionary biology, they have made their way into human genetics and animal/plant breeding.

Searching for Genes that have Experienced Recent Selection

Genomic scans for selection offer a complementary approach to testing candidate genes for association or using genome scans to find QTLs. These latter approaches assume a trait of interest and then look for associations between loci and this trait. In contrast, tests for selection make no assumptions about the trait(s) of interest, but rather simply look for loci under selection. To see the contrast between these approaches, suppose we are studying a rice strain that is adapted to growing under high water stress. Under a candidate gene/QTL approach, a plant physiologist might suggest a number of traits to examine, and then one looks for marker-trait associations. Under scans for selection, we simply look for loci under selection, and these may inform us as to other pathways for adaptation that were not apparent from physiological arguments. Thus, both approaches nicely complement each other. Further, if one finds a strong marker-trait association we can also look to see if this site shows indications of recent selection.

These scans for selection can also suggest the genes that were important in the domestication and improvement of crops and farm animals. One can distinguish between:

- **Domestication genes**: Alleles fixed in the course of the initial domestication. These should be shared by all domesticated lines.

- **Diversification/Improvement genes**: Alleles fixed in the course of improvement following domestication. These may be fixed in some lines and absent in others.

- **Adaptation genes**: Alleles in natural populations responding to natural selection on environmental conditions. Such alleles are candidates for transfer into elite germplasms.

Expected Patterns of Variation Under Genetic Drift

Tests of recent nature selection are all based on deviations from the patterns expected under mutation and genetic drift in the absence of selection (the *neutral theory*). Hence, a few basics on the interactions of drift and mutation are in order.

Basically, *drift removes variation*, while *mutation introduces variation*. Eventually (if the population size stays constant), these achieve forces balance each other out, with the removal of variation by drift balancing the new input from mutation. However, if the population size has recently changed (such as going through a bottleneck or an expansion), there may be an excess, or deficiency, of variation, as the population has yet to achieve the new mutation-drift equilibrium.
Figure 10.1 illustrates the dynamic nature of the interplay between drift and mutation. Notice that alleles randomly increase/decrease in frequency, with drift causing alleles to be lost, while new alleles are introduced by mutation. Drift may fix a lineage (the descendants of a single allele), but mutations can also occur along this lineage, so that there is still variation in the population even though all copies in the population descend from a single ancestor.

\[ H = \frac{4N_e \mu}{1 + 4N_e \mu} \]  

Thus, within-population variation is generated by the interactions of drift and mutation. The product \(4N_e \mu\) also appears in several other measures of variation, and this product is widely denoted by \(\theta\) in the population-genetics literature.

Likewise, drift can generate a between-population variation. If we split a starting population into two (initially) identical subpopulations, these diverge over time as drift randomly fixes alleles and (more importantly) mutation introduces new alleles. In particular, if two populations became isolated \(t\) generations ago, then the expected number of new mutations destined to become fixed between them is simply

\[ d = 2\mu t \]  

Thus, while the levels of within-population variation are a function of population size (and mutation rate), the amount of divergence between populations is strictly a function of the mutation rate.

**Drift and the Coalescence Process**

A very powerful way to think about drift processes is to think about time. If we start with a population of \(2N\) DNA sequences at a particular locus, eventually drift will result in all of the members in the population being direct descendants of just one common ancestral sequences. Thus, we can think about drift as fixing lineages, and also think of a collection of DNA sequences (for a particular locus) as a collection of ever-distant ancestors of a single progenitor sequence. Such a collection of nested lineages is called the coalescent process. Figure 10.2 illustrates this idea. If we sample 5 DNA sequences at a given locus, we can inquire about the time back until the five lineages coalesce into
four lineages. In our cases, samples 1 and 2 have the most recent common ancestor (or MRCA). At that point in time, we have gone from five lineages to four. We can continue this back in time until all of the sequence lineages coalesce to the common ancestral sequence from which all of the sequences in our sample descend.

![Diagram of coalescent process](image)

**Figure 10.2.** Drift as a coalescent process. We have sampled 5 DNA sequences for a given region. These are all related, having descended from a common ancestor at some point in the past. In our sample, 1 and 2 are the most closely related, while the 1,2 and 3,4,5 lineages met back at the common ancestor.

The key feature about the coalescent process is that we have said nothing about mutation. If the time back to the common ancestor is long, then there is a good chance that mutations will have occurred, and the sample will show variation. However, if this time is short, it is unlikely that mutations will have occurred, and the sample shows little sequence variation. Under an equilibrium drift model, the distribution of coalescent times is completely determined by the effective population size and sample size, following a geometric distribution with success parameter \( n(n-1)/4N_e \) for the first coalescent event (from \( n \) lineages to \( n-1 \) lineages), and \( j(j-1)/4N_e \) for the time between the coalescent event moving from \( j \) lineages to \( j-1 \) lineages.

We can make the connection between Equation 10.1 (for standing levels of variation) and the coalescent process as follows. For a population of size \( N \), the expected coalescence time for two random sequences if just \( 2N \) generations. Thus, \( 2n\mu \) mutations occur from one copy back to the ancestor, while \( 2N\mu \) mutations occur on the lineage from the other copy back to the ancestor, for a total of \( 4n\mu \) mutations separating the two present-day sequences.

Coalescence theory tells us to consider time, and this time is set by the population size – large \( N \) implies a long time back to the MRCA. Small \( N \) (and hence stronger drift) means a shorter time back to the MRCA.

**Hitch-hiking, Linkage Drag, and Selective Sweeps**

Against this neutral background, what is the nature of a signal that selection would leave? When selection rapidly increases the frequency of an allele, linked sites also hitch-hike along for the ride. Plant breeders are aware of this phenomenon, namely **linkage drag**, wherein a favorable introgressed region may drag along unfavorable linked genes. If the introgression is sufficiently rapid, the amount of linkage drag can be considerable. Likewise, when selection (natural or artificial) favors a particular allele, sites linked to that allele are also dragged along to fixation, resulting in a region around the selected site showing reduced genetic variation relative to the rest of the genome. Such a **selective sweep** occurs because the effect of selection is to reduce the effective population size at linked regions. This results in decreased times to the MRCA, and hence less polymorphism, due to a shortening of the coalescence times relative to pure drift (Figure 10.3). In the extreme case where a favorable allele starts as a single copy that is rapidly fixed, sites tightly linked to that region will also descend from this initial haplotype containing the favorable allele. The more rapid the fixation,
the more reduced the level of variation around the favored site and the larger the size of the region
influenced by the sweep. It is important to note, however, that while linkage may reduce the levels
of standing variation through their reductions in $N_e$ (and hence $\theta = 4N_e\mu$), linkage has essentially
no effect on the average substitution rate at linked neutral sites. This is because (as discussed above),
the per generation rate of divergence between neutral sites is simply the mutation rate, independent
of population size.

Figure 10.3. Coalescent times under pure drift and two types of selection. Under balancing selection
(over-dominance), the time back to the most recent common ancestor (MRCA) is longer than under pure drift. Under
directional selection (often called a selective sweep), an allele sweeps through a population far quicker than
under drift and hence has a more recent MRCA.

Thus, one signal for selection is a reduced level of polymorphism relative to the rest of the genome,
something that could be detected by either scoring a number of markers around a candidate gene or
using a dense marker screen of the entire genome, the equivalents of candidate gene and genomic
scan approaches (respectively) that one uses in QTL mapping. However, a significant reduction in
the level of polymorphism (by itself) is by no means sufficient to indicate selection, as this could
simply reflect a reduction in the neutral mutation rate in that region. Further, even under a neutral
model, a plot of the polymorphism level over a chromosome can often show significant dips, simply
due to sampling effects. Thus, while regions of significantly decreased polymorphism are certainly
suggestive, more formal tests are required, as we will review shortly.

Figure 10.4. The impact of selection on variability at surrounding neutral sites. The vertical axis plots
heterozygosity, the horizontal genome location. The upper graph shows the effect of a selective sweep, which
results in a decrease in linked neutral polymorphisms around the selected locus (indicated by the filled circle).
The width is a function of recombination (smaller $c$ = larger width), selection (larger $s$ = larger width) and
time of the sweep (longer the time since the sweep, the smaller the width). Plots such as this are generated
by computing the variation in a window (of, say 100-1000 bases) that we slide along the genome (see Figure
10.5). The lower graph shows that balancing selection results in an increase in the level of linked neutral
polymorphisms.
As shown in Figure 10.4, directional selection results in a reduction in the level of polymorphism at sites linked to the region under selection. Conversely, a region under overdominant selection will show an increase in the amount of polymorphism at linked neutral sites. Both of these observations can be thought of in terms of time. Under a selective sweep, markers tightly linked to a selected locus have a more recent common ancestor than the rest of the genome, while under overdominant selection, linked sites have a deeper (older) common ancestor. Figure 10.3 illustrated this idea by contrasting the coalescent times under pure drift, directional, and overdominant selection.

There is a final selective force that can cause a decrease in the level of polymorphism, namely background selection. Here, selection against newly arising deleterious mutations also reduces the effective population size in a linked region around the selected site. Highly deleterious alleles have little impact, as such mutations are removed almost immediately. However, slightly deleterious mutations may drift up to low (but not rare) frequencies, and their removal has a larger impact. While the effect for any single removal may be minor, one expects significantly more deleterious mutations arising within a region than beneficial ones, and hence background selection can potentially have a rather significant effect. Further, it can be very difficult to distinguish between selective sweeps (selection for new alleles) and frequent background selection (selection against new alleles). Of special interest to us is that the effects of background selection can be quite significant in highly selfing plant populations, due to their low effective recombination rates.

Parameters Associated with Selective Sweeps

The nature and persistence of the signal left by a selective sweep has been extensively investigated by population geneticists. As one might except, the stronger the selection, the quicker an allele becomes fixed, and the larger the linked region influenced by the sweep. Likewise, as recombination decreases, the length of the sweep-influenced region increases. Specifically, Kaplan et al. (1989) showed that an approximation for 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 \( c \),

\[
d \sim 0.01 \frac{s}{c}
\]

Equation 10.3 thus allows for an estimate of the historical value of \( s \) given an observed size \( d \) and an estimate of the local recombination fraction,

\[
s \sim 100 \cdot d \cdot c
\]

Kim and Stephan (2002) showed that the sweep region can often be asymmetric around the selected site, so one should choose \( d \) as the average of both sides of the sweep. Hence, if \( \delta \) is the total width of the genomic window of reduced polymorphism, then \( d = \delta/2 \), but the midpoint of this window is not necessarily a good estimate of the position of the selected site.

The time for the sweep to occur, namely the time to fixation of a favorable allele, is approximately \( 2 \ln(2N)/s \) generations, assuming that the favorable allele starts out as a single new mutation. Assuming one starts with a single new favorable mutant, a crude approximate for the time \( t \) (in generations) for fixation can be expressed as

\[
t \sim \frac{2 \ln(2N)}{s} \sim \frac{2 \ln(2N)}{100 \cdot d \cdot c} = 0.02 \frac{\ln(2N)}{d \cdot c}
\]

Once the favored site has become fixed (and indeed even on its way to fixation), signal for the sweep starts to decay through recombination and mutation. Przeworski (2002) suggests that the signal of a sweep persists for roughly \( N_e \) generations, so that recent sweeps can leave a signature, but more ancient sweeps do not. A number of investigators have suggested estimators for the time
since a sweep (Perlitz and Stephan 1997, Enard et al. 2002, Jensen et al 2002, Przeworski 2003). but these are very sensitive to model assumptions. Finally, a critical assumption to much of the theoretical work on sweeps is that they are initiated by the appearance of a single new mutation favored by selection. During domestication, the situation might have been different, with alleles already segregating in the population becoming favored under domestication. Such a scenario has significant implications for the type of signature left by a selective sweep, making them much more difficult to detect.

Signatures of Selective Sweeps in Crops


Figure 10.5. Signature of a selective sweep around the teosinte branched 1 gene in maize. Left: Reduction in variation in the 5’ NTR region of tb1. Note that slight reduction in variation across this gene in maize vs. teosinte, reflecting the population bottleneck created during domestication. Also note that the signature of the sweep does not extend into the coding region, suggesting the site of selection is in the 5’ regulatory region (after Wang et al. 1999). Right: A more focused analysis of the 5’ NTR region shows the limit of the sweep (after Clark et al. 2004).

tb1 in Maize

Perhaps the most convincing example is the work of Doebley and colleagues on the maize gene teosinte branched 1, henceforth tb1, see Figure 10.5. This gene was originally detected in QTL mapping studies in crosses between maize and teosinte and subsequently shown to be the previously characterized locus tb1, which has major effects on plant architecture. Given its obvious role as a candidate domestication gene, Wang et al. (1999) compared the levels of polymorphism around this locus in maize with the corresponding region in teosinte. Throughout this region, maize was found to have reduced levels of polymorphism (about 75%) relative to teosinte, but this is consistent with a bottleneck during domestication influencing all loci in modern maize. More importantly, they observed a significant decrease in the amount of polymorphism in the 5’ NTR region of maize (but not teosinte) tb1, suggesting a selective sweep influenced this region. Surprisingly, the sweep did not influence the coding region, suggesting that the selected site was in the 5’ regulatory region, as opposed to selection on a change in the amino acid sequence of tb1. Clark et al. (2004) examined the 5’ tb1 region in more detail, finding evidence for a sweep influencing a region of 60 - 90 kb in the 5’NTR. Both Wang et al. and Clark et al. controlled for the possibility of this reduction in neutral polymorphism being caused by reduced mutation rates in this region by comparing polymorphism levels with a close relative (teosinte). Wang et al. applied Equation 10.3 to estimate the average strength of selection on this site, obtaining \( s \approx 0.05 \). This value of \( s \) implies an expected time for selection to fix the domestication allele of around 300 to 1000 years, indicating a fairly long period of domestication.
Waxy in Rice

While *tb1* is a potential (and very likely) gene involved in very early domestication, a possible example involving a gene selected after initial domestication (during the improvement phase) is the *Waxy* gene in rice (Olsen et al. 2006). “Sticky” (glutinous) rice results from low amylose levels, which are typical of temperate japonica variety groups, and this has been shown to be due to a splice mutant in the *Waxy* gene. Olsen et al. observed a region of 250kb around *Waxy* with greatly reduced levels of polymorphism compared to control populations (lines of nonsticky rice). The size of this region (using the local recombination rate) gives the estimated strength of selection acting on this site as $s \approx 10.6$, implying selection much stronger than on *tb1* during maize domestication. Further, while the sweep around *tb1* did not even influence the coding region of its gene, the *Waxy* sweep covers 39 rice genes. If a number of sweeps covering large regions occur during the course of domestication, this can have significant evolutionary effects on the entire genome.

The Hill-Robertson Effect: Accumulation of Deleterious mutations in domesticated rice genomes

Hill and Roberston noted that one important consequence of selection at a linked site is reduction in the effective population size around that region. We have seen how this results in a reduction in the levels of variation. However, it also has a second important consequence: selection becomes much less effective in the present of linkage. The reduction in $N_e$ at some site makes the effects of drift (relative to selection) more important at linked sites. This reduces the probability of fixation for favorable alleles and increases the probability of fixation of deleterious alleles.

Noting the larger regions for selective sweeps in rice, Lu et al (2006) suggested that selection for domestication in rice significantly reduced its effective population size, and resulted in an increase in the fixation for harmful mutations. These authors compared the genomes of *Oryza sativa* ssp. *indica* and *japonica* with their ancestral relative *O. rufipogon*. They found that $K_A/K_S$ ratios (the ratio of the substitution rate of non-synonymous to synonymous changes) were much higher for *indica* vs. *japonica* (0.498) than for domesticated vs. wild rice (*japonica* vs. *rufipogon*, 0.259). Lu et al suggest that roughly 25% of the amino acid differences between *indica* and *japonica* are deleterious. They suggest that excessive reductions in $N_e$ due to selective-sweeps covering much of the genome during selection for domestication greatly reduced the efficiency of natural selection in removing deleterious alleles.

Basic Logic of Sequence-Based Selection Tests

While reductions (or increases) in the levels of variation around a gene of interest are suggestive of selection, this is not a formal test, and such changes can simply result from local changes in the mutation rate. A (large) number of more formal tests have been proposed to detect selection, so we focus on some of the classic tests as well as the general logic behind these. In attempts to detect recent selection, one can look at polymorphisms (within-population variation) and/or divergence (between population variation). The basic idea behind most tests is to contrast two different measures of variation that are connected in a particular way under the equilibrium neutral model with the actual pattern observed.

Logic Behind Polymorphism-Based Tests

In nutshell, the logic is time. If a locus has been under positive selection, it will have a more recent common ancestor (MRCA) than a sequence under pure drift. Conversely, if a locus is experiencing balancing selection, two random sequences will, on average, have a MRCA more distantly relative to pure drift. This difference in time to MRCA has consequences on levels of standing polymorphism (shorter the MRCA, the less the polymorphism). The time back to the MRCA also influences the

Lecture 10, pg. 7
length of a region under linkage disequilibrium. The longer the time, the shorter the expected block of disequilibrium around a gene. Hence, reduced level of polymorphism and/or longer blocks of disequilibrium relative to a neutral model are both potential signals of directional selection. Finally, selection shifts the frequency spectrum of alleles (the number of alleles in each frequency category) either producing too many rare alleles (alleles older than expected) or too many alleles at intermediate frequencies relative to pure drift (alleles younger than expected).

Logic Behind Divergence Tests

The simple logic behind many divergence-based (or phylogenetic comparison) tests (comparisons of a single sequence of the target gene from a collection of different species) is to examine $K_A/K_s$ ratios. $K_A$ refers to the substitution rate for replacement changes, while $K_s$ refers to the substitution rate for synonymous changes. Under the strict neutral theory, $K_A/K_s$ ratios should be one (after suitable adjustment accounting for the fact that a random change is more likely to give a replacement mutation than a synonymous one). In most settings, when averaging over an entire protein, one typically sees a $K_A/K_s$ ratio significantly less than one. This is evidence of selection, but this is expected under Kimura’s neutral theory, being a reflection of the selective constraints on the sequence (i.e., purifying selection). Many to most replacement mutations are deleterious. Conversely, there are cases for some genes where particular regions show a $K_A/K_s$ ratio above one. This suggests that these substitutions might be favored by selection. We can think about this by considering the expected substitution rate ratio. For a neutral site, $2\mu_{neu}$ is the expected number of mutations that arise per generation, each of which has probability $1/(2N)$ of being fixed, giving the neutral substitution rate as

$$d_{neu}/t = \mu_{neu}$$

Conversely, with favorable mutations, the expected number of such new mutations each generation is $2N\mu_{fav}$, however, their chance of fixation is now $2s$, twice their selective advantage, giving

$$d_{fav}/t = 2N\mu_{fav} \cdot 2s = 4Ns\mu_{fav}$$

The resulting ratio of favorable/neutral rates becomes

$$\frac{d_{fav}}{d_{neu}} = \frac{4Ns\mu_{fav}}{\mu_{neu}} = 4Ns\frac{\mu_{fav}}{\mu_{neu}}$$

Hence, even though favorable mutations are expected to be far less frequent, provided $4Ns\mu_{fav} > \mu_{neu}$, $K_A/K_s > 1$ is expected. While $K_A/K_s$ ratios greater than one for entire sequences are rare, such ratios can sometimes be found embedded within a sequence, for example at the critical residues a protein that may interact with some new target.

Logic Behind Joint Polymorphism and Divergence Tests

Under the neutral theory, the heterozygosity is a function of $\theta = 4N_e\mu$, while divergence is a function of $\mu t$. Tests jointly using information on within-species polymorphism and between-species divergence make use of these two different measures to test for concordance with neutral expectations. Under pure drift the amount of within-population heterozygosity and between-population (or species) divergence is positively correlation, as both are function of $\mu$. Recall that the standing heterozygosity and between-population divergence for the $i$th locus under drift are

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

Hence,

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

Lecture 10, pg. 8
Thus if we compare several loci both within the same population and between the same populations, the $H/d$ ratio should be the same (subject to random sampling), as under pure drift this ratio has expected value $2N_e/t$, which should be similar (if not identical) across loci (under the pure drift model). Variations on this theme have been proposed, as we detail below.

**Tests Based Strictly on Within-Population Variation**

There are a large number of tests that compare different features of standing variation (such as number of alleles versus average pairwise distance between alleles). Two sequence evolution frameworks are generally used as the basis for such comparison: the **infinite alleles** and **infinite sites** models. The key assumption of both models is that each mutation generates a new sequence, and hence leaves a unique signature. Such is *not* the case when using microsatellite (or STR) markers, as these follow a step-wise mutation model. When analyzing such markers, this very different mutation process must be explicitly modeled into the analysis.

So how are the two basic models different? Given a DNA sequence, an infinite alleles framework would treat each **haplotype** as a different allele (under the assumption of no intra-genic recombination), while the infinite sites framework looks at each position in the sequence separately. Figure 10.6 shows the difference. In this sample of five sequences (rows), there are three haplotypes (and hence three alleles in the infinite alleles framework). However, in an infinite sites framework, looking over the six sites (columns), we find that only two of these sites are segregating.

```
A A G A C C
A A G G C C
A A G A C C
A A G G C C
A A G G C A
```

*Figure 10.6.* A sample of five sequences (each row is a sequence), showing three haplotypes (rows 1,3; 2,4; and 5) but only two segregating sites (positions = columns 4 and 5). Thus, the number of alleles and the number of segregating sites are two different measures of standing variation.

Polymorphism-based tests compare the frequency of alleles with their expectations under the equilibrium neutral model. Two typical departures are seen: (i) an excess of common alleles and a deficiency of rare alleles (alleles younger than expected) and (ii) a deficiency of common alleles and an excess of rare alleles (alleles older than expected). Pattern (i) would be expected under directional selection, when the coalescent times have been shrunk by a selective sweep. Pattern (ii) would be expected under stabilizing selection, where the coalescent times are longer than expected under drift. The problem is that these patterns can also be generated by demographic events as well. A population bottleneck and/or recent population expansion can generation pattern (i), while population subdivision can generate pattern (ii). Thus polymorphism-based tests contrast the null (strict neutral model with constant population size) against a composite alternative hypothesis: selection and/or departures from a single random-mating population of constant size.

Obviously this is a serious limitation. However, demographic effects should leave a constant signature throughout the genome, while selection events leave a unique signature against this background. Hence, recent whole-genome scans of selection have performed polymorphism-based tests scanning a large, dense set of markers spanning the genome, and use this information to generate a null distribution of the test statistic given the population history. Selection is suggested by looking at the extreme outliers against this null distribution.
Polymorphism-based tests contrast independent estimates of $\theta = 4N_e \mu$ based on two (or more) different metrics of population variation (such as heterozygosity, number of alleles, number of singletons [alleles present as only a single copy in our sample]). These estimates are all based on the equilibrium neutral model, and if different metrics of variation give statistically significantly different estimates of $\theta$, this is evidence of departure from this strict neutral model, and hence possibly a selected site.

**The Infinite Alleles Model: Ewen’s Sampling Formula**

The foundation for several tests under the infinite alleles framework is Ewen’s Sampling Formula (Evens 1972): under the infinite alleles model, the probability that we see $K$ alleles (haplotypes) in a sample of size $n$ is

$$
Pr(K = k) = \frac{|S_n^k \theta^k|}{S_n(\theta)} \quad (10.7a)
$$

where

$$
S_n(\theta) = \theta(\theta + 1)(\theta + 2) \cdots (\theta + n - 1) \quad (10.7b)
$$

and $S_n^k$ is the coefficient on the $\theta^k$ term in the polynomial given by $S_n(\theta)$. $(S_n^k)$ is called a Stirling number of the first kind). For example, the probability that only a single allele is seen in our sample is

$$
Pr(K = 1) = \frac{(n - 1)!}{(\theta + 1)(\theta + 2) \cdots (\theta + n - 1)} \quad (10.7c)
$$

From Equation 10.7a, the mean and variance for the number of alleles can be found to be

$$
E(K) = 1 + \theta \sum_{j=2}^{n} \frac{1}{\theta + j - 1} \quad (10.8a)
$$

and

$$
\sigma^2(K) = \theta \sum_{j=1}^{n-1} \frac{j}{(\theta + j)^2} \quad (10.8b)
$$

One key feature to note is that all of these statistics are simple functions of $\theta$ and the sample size $n$.

**The Infinite Sites Model**

The building blocks for many of the early tests of neutrality are based on summary statistics from the infinite sites model. This model is the logical extension of the infinite alleles model to a DNA sequence, essentially treating each nucleotide as a new locus (or site). The infinite sites model assumes that each new mutation introduces a new site (i.e., only one mutation per site). This is not an unreasonable model unless we are scoring STR loci, which have high mutation rates and the very real possibility of back mutations.

The typical setting is a sample of $n$ sequences taken from a population, with the goal of estimating $\theta = 4N_e \mu$. Three common summary statistics are used for this purpose. The first is $S$, the number of segregating sites in sample. The second is $k$, the average pairwise difference between any two random sequences. The final is $\eta$, the number of singletons (variable sites present just once in our sample). The expected values and sample variances for these summary statistics are as follows:
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 | $E[k] = \theta$ | $\sigma^2(k) = \theta\frac{n+1}{3(n-1)} + \theta^2\frac{2(n^2 + n + 3)}{9n(n-1)}$

$\eta$ = number of singletons | $E[\eta] = \theta\frac{n}{n-1}$ | $\sigma^2(\eta) = \theta\frac{n}{n-1} + \theta^2\left[\frac{2a_n}{n-1} - \frac{1}{(n-1)^2}\right]$

where

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

(10.9)

Under pure drift, the following three are all estimators of $\theta$,

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

(10.10)

$\hat{\theta}_S$ is called with Waterson estimator (Waterson, 1975). Proposed tests for neutrality contrasts pairs of these estimates, with Tajima’s (1989) test comparing estimates based on $S$ and $k$, while two tests proposed by Fu and Li (1993) contrasts estimates based on $S$ and $k$ with those based on $\eta$.

**Example 10.1.** Suppose we sample 10 alleles from a population and observe $S = 12$, $k = 4$, and $\eta = 3$. What are the estimates of $\theta$ given these three summary statistics?

\[
\hat{\theta}_S = \frac{S}{a_{10}}, \quad a_{10} = \sum_{i=1}^{9} \frac{1}{i} = 2.83, \quad \text{giving} \quad \hat{\theta}_S = \frac{12}{2.83} = 4.24
\]

\[
\hat{\theta}_k = k = 4
\]

\[
\hat{\theta}_\eta = \frac{n}{n-1} \eta = \frac{10}{9} \cdot 3 = 3.33
\]

**Tajima’s $D$ Test**

One of the first, and most popular, polymorphism-based test is Tajima’s (1989) $D$ test, which contrasts $\theta$ estimates based on segregating sites ($S$) versus average pairwise difference ($k$),

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

(10.11a)

where

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

(10.11b, 10.11c)

Tajima’s motivation for this test was his intuition that there is an important difference between the number of segregating sites $S$ and the average number $k$ of nucleotide differences. For the former we simply count polymorphic sites (independent of their frequencies), while the later is a frequency-weighted measure. Hence, $S$ is much more sensitive to changes in the frequency of
rare alleles, while \( k \) is much more sensitive to changes in the frequency of intermediate alleles. A negative value of \( D \) indicates too many low frequency alleles, while a positive \( D \) indicates too many intermediate-frequency alleles. Expressed another way, \( D \) is a test for whether the amount of heterozygosity is consistent with the number of polymorphisms. Under selective sweeps (and background selection and population expansion), heterozygosity should be significantly less than predicted from the number of polymorphisms.

Example 10.2. Two interesting examples were offered by Tajima (1989). First, Aquadro and Greenberg looked at 900 base pairs in the mitochondrial DNA of seven humans, finding 45 segregating sites and an average number of nucleotide differences between all pairs of 15.38. Here

\[
\alpha_D = \frac{1}{2.45} \binom{7 + 1}{3(7 - 1)} - \frac{1}{2.45} \cdot 0.0417 = -0.0269
\]

\[
\beta_D = \frac{1}{2.45^2 + 1.49} \left( \frac{2(7^2 + 7 + 3)}{9 \cdot 7(7 - 1)} - \frac{7 + 2}{7 \cdot 2.45} + \frac{1.49}{2.45^2} \right) = 0.0417
\]

\[
D = \frac{\hat{\theta}_k - \hat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} = \frac{15.38 - 18.38}{\sqrt{-0.0269 \cdot 45 + 0.0417 \cdot 45^2}} = -0.3288
\]

Table 2 of Tajima (1989) gives the 95% confidence interval on \( D \) under strict neutrality for \( n = 7 \) as -1.608 to 1.932, so this value is not significantly different from its neutral expectations.

Second, Miyashita and Langley examined 64 samples of a 45-kb region of the white locus in \( D. melanogaster \). Taking large insertions/deletions as the polymorphic sites, they found \( S = 454 \) and \( k = 0.94 \), which gives \( D = -2.0709 \). Given that the 95% confidence interval under neutrality is -1.795 to 2.055, this locus shows evidence of either directional selection or a population bottleneck (or expansion).

Genome-Wide Polymorphism Tests

As mentioned several times, polymorphism-based tests suffer in that we reject the null hypothesis (the neutral equilibrium model), we are left with a composite alternative hypothesis that not only includes selection but also include departures from the standard demographic assumptions (a single random mating constant size population). Given this, much thought has gone into trying to estimate the coalescent process under neutrality, but allowing for the population structure inherent in the data. One approach is to make some assumptions about the demography, and then use these to generate a neutral coalescent under this structure, from which we can obtain a null distribution for comparison. More recently, a number of workers have used Cavalli-Sforza’s (1966) idea that all of the genome experiences the same demography (focusing here on the autosomal chromosomes). Hence, markers across the genome provide useful information on the null distribution. Using this approach, one could scan a huge number of loci, under the assumption that the vast bulk are essentially neutral (i.e., not under strong directional selection), and hence these can be used to generate the null distribution. Outliers in this null indicate potential loci under selection.

The Ghost of Lewontin-Krakauer: Genome Wide \( F_{ST} \)-based Scans

Lecture 10, pg. 12
One of the very first tests for selection with molecular data was proposed by Lewontin and Krakauer (1973), who looked at allele frequencies values in different populations by computing Wright’s $F_{st}$ statistic. $F_{st}$ is basically the fraction of between-group variation, the between-group variance divided by the total variance. The $F_{st}$ value for the data was compared with the expected neutral value. Lewontin and Krakauer reasoned (correctly) that if differential (directional) selection was occurring in the different populations, this would generate a larger than expected $F_{st}$ value. Likewise, if overdominant selection was operating, the between-population divergence would be less than expected. While their logic was sound, their test was heavily criticized, as the null distribution under neutrality depends very heavily on details of the (unknown) population structure. As a result, this test died a quick death. However, we are now starting to see its ghost reappear in the literature.

Several scans for selected loci in the human population have looked at the $F_{st}$ (or a related measure $R_{st}$ for STR loci) values over a very large number of sites, taking outliers from this distribution as indicators of potential loci under selection. For example, Akey et al. (2002) used 26,530 SNPs (single nucleotide polymorphisms) in three human populations, computing $F_{st}$ values for each, generating 174 candidate loci. Kayser et al. (2003) looked at 322 STR (short tandem repeats = microsatellite) loci in both Africans and Europeans. Of these, 11 showed usually high values. As a check, they then sequenced a nearby STR (for each of the candidates), finding that these new (and tightly linked) loci also have $R_{st}$ values larger than average. Storz et al. (2004) looked at 624 autosomal markers in multiple human populations, finding 13 that appeared to be outliers.

Tests Based on Long Haplotypes

As mentioned, one feature of selective sweeps is that they have an excess of newly-derived alleles at high frequency. A second class of tests is offered by the following observation: under a selective sweep, since some alleles are at much higher frequencies than their age would suggest under a neutral model, these alleles should also have longer regions of linkage disequilibrium. Again, the key here is time. The more time, the smaller the window of disequilibrium. If a sweep moves an allele quickly to high frequency, the amount of disequilibrium, given its frequency, should be excessive relative to a neutral model.

A number of tests have recently been proposed based on this idea of long haplotypes. Specifically, under the neutral theory, an allele at moderate to high frequency is an old allele, as it takes time for drift to move an allele to modest frequencies. As an old allele, it has also experienced a number of recombination events, and hence the region of linkage disequilibrium (the length of the haplotype) around this allele should be small. If we observe alleles at moderate to high frequencies with long haplotypes, this is strong evidence of selection. Population demography cannot generate such a feature, only selection to rapidly increase an allele frequency will generate such a pattern. Hence, this approach offers tests that circumvent much of the demographic concerns. The limitation is that we must be able to accurate estimate the length of haplotypes, which requires a very dense SNP map, especially if we are scanning a genome. If we are focusing on a candidate gene, we simply need a dense collection of SNPs spanning a region around that gene.

Ascertainment Bias

All genome-based test have an important caveat. The large number of markers used are typically generated by looking for polymorphisms first in a very small, and often not very ethnically-diverse, sample. As a consequence, there is a strong ascertainment bias inherent with these markers (for example, an excess of intermediate-frequency markers). If such biases are not accounted for, they can skew genome-wide tests (Nielsen 2005).
**Joint Polymorphism and Divergence Tests**

**McDonald-Kreitman Test**

One of the most straightforward tests of selection when one has both polymorphism and divergence data was offered by McDonald and Kreitman (1991). Their basic logic was very similar to that leading to Equation 10.6. Consider a single locus, where we contrast the polymorphism and divergence rate at synonymous versus replacement sites. The ratio of expected divergence between synonymous vs. replacement sites is

\[ \frac{d_{\text{syn}}}{d_{\text{rep}}} = \frac{2t\mu_{\text{syn}}}{2t\mu_{\text{rep}}} = \frac{\mu_{\text{syn}}}{\mu_{\text{rep}}} \]  

(10.12a)

Likewise, the ratio of heterozygosity at the two locations is

\[ \frac{H_{\text{syn}}}{H_{\text{rep}}} = \frac{4N_e\mu_{\text{syn}}}{4N_e\mu_{\text{rep}}} = \frac{\mu_{\text{syn}}}{\mu_{\text{rep}}} \]  

(10.12b)

Hence, these two ratio have the same expected value. We note that McDonald and Kreitman provide a more general derivation of 10.12a, replacing \( 4N_e \) (the equilibrium value) by \( T_{\text{tot}} \), the total time on all of the within-species coalescent branches, so that any effects of demography cancel. Hence, the McDonald-Kreitman is **not affected by population demography** (Nielsen 2001). Given the constancy of these ratios under neutrality, the McDonald-Kreitman test is performed by contrasting polymorphism vs. divergence data at synonymous versus replacement sites in the gene in question through a simple contingency table.

**Example 10.3.** McDonald and Kreitman (1991) examined the \( Adh \) (Alcohol dehydrogenase) locus in \( Drosophila \), specifically the sibling species \( D. melanogaster \), \( D. simulans \), and the outgroup \( D. yakuba \). Between the sibling species, a total of 24 fixed differences were found, 7 of which were replacement, 17 synonymous. Turning to polymorphisms, 44 polymorphic sites were found, 2 of which were replacement and 42 synonymous, giving

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

Fisher’s exact tests gives a \( p \) value of 0.0073. (In R, this is obtained using \( \text{x<-matrix(c(7,17,2,42),nrow=2)} \) to enter the data and \( \text{fisher.test(x)} \) to run the test.)

**Hudson-Kreitman-Aguade (HKA) Test**

Hudson, Kreitman, and Aguade (1987) proposed the first test to jointly use information from the standing levels of polymorphisms within a species and the amount of divergence between species. The result was the **HKA test**.

Consider two species (or distant populations) \( A \) and \( B \) that are at mutation-drift equilibrium with population sizes \( N_A = N \) and \( N_B = \alpha N \), respectively. Further assume they separated \( T = \tau/(2N) \) generations ago from a common population of size \( N^* = (N_A + N_B)/2 = N(1+\alpha)/2 \), the average of the two current population sizes. Now suppose \( i = 1, \ldots, L \) unlinked loci are examined in both species. The amount of polymorphism for locus \( i \) in species \( A \) is a function of \( \theta_i = 4N_e\mu_i \), while for species \( B, \theta = 4N_B\mu_i = 4(\alpha N_e)\mu_i = \alpha \theta_i \). The resulting summary statistics used are \( L S^A_i \) values, for the number of segregating sites at locus \( i \) in species (population) \( A \), another \( L S^B_i \) for the same \( i \) loci in species \( B \), and \( L D_i \) values, for the amounts of divergence (measured by the average number of differences between a random gamete from species \( A \) and a random gamete from species \( B \)).
This can be solved numerically, generating the estimated values for the $X$.

The divergence from partitioning of the polymorphism where the first term is the between-population divergence due to new mutations and the second term of fit $X$.

Equations 10.14a and 10.14b follow from our above results for the infinite sites model. Equation 10.14c follows by re-writing

$$
\theta_i \left( T + \frac{1 + \alpha}{2} \right) = 4N\mu_i \left( \frac{\tau}{2N} + \frac{1 + \alpha}{2} \right) = 2\mu_i T + 4\mu_i \frac{N(1 + \alpha)}{2}
$$

where the first term is the between-population divergence due to new mutations and the second term the divergence from partitioning of the polymorphism $4N\mu_i$ in the ancestral population. Thus the HKA test has $L+2$ parameters to estimate, the $L\theta_i^2$ values and two demographic parameters, $T$ and $\alpha$. The HKA test estimates these parameters and then (using Equation 10.14) computes the goodness of fit $X^2$ statistic (Equation 10.13), which is approximated $\chi^2$ distributed with $3L - (L+2) = 2L - 2$ degrees of freedom. Hudson et al. suggest the following system of equations for the estimating the $2L + 2$ unknowns,

$$
\sum_{i=1}^{L} S_i^A = a_{n_A} \sum_{i=1}^{L} \hat{\theta}_i \\
\sum_{i=1}^{L} S_i^B = \hat{\alpha} a_{n_B} \sum_{i=1}^{L} \hat{\theta}_i
$$

$$
\sum_{i=1}^{L} D_i = \left( \hat{T} + \frac{1 + \hat{\alpha}}{2} \right) \sum_{i=1}^{L} \hat{\theta}_i \\
S_i^A + S_i^B + D_i = \hat{\theta}_i \left( \hat{T} + \frac{1 + \hat{\alpha}}{2} \right) + a_{n_A} + \hat{\alpha} \cdot a_{n_B} \quad \text{for } 1 = 1, \ldots, L - 1
$$

This can be solved numerically, generating the estimated values for the $X^2$ statistic.

**Example 10.4.** Hudson et al. examined Adh locus silent variation as one locus and the 4-kb 5’ flanking regions of Adh as the second locus in D. melanogaster and its sibling species D. sechellia. A sample of 81 melanogaster alleles were sequenced, along with a single sechellia allele. Based on sequencing, the divergence was 210 differences in the 4052 bp flanking region and 18 differences in the 324 silent sites, for roughly equal levels of divergence between the two loci. Based on restriction enzyme data, within melanogaster, 9 of the 414 5’ flanking sites were variable, while 8 of 79 Adh sites were variable. Thus while the divergence was roughly equal, there was a four-fold difference in polymorphism. Hudson
et al. modify the HKA test to account for only polymorphism data from only a single species. Further, given the difference in number of sites between the polymorphism and divergence data, let $\theta$ be the per-nucleotide $\theta$ value, so that we have to weight the $\theta$ value for each term by the number of sites compared, giving Equation 10.15 as

$$S_A^1 + S_B^2 = 9 + 8 = a_{81} \left( 414 \cdot \hat{\theta}_1 + 79 \cdot \hat{\theta}_2 \right)$$

$$D_1 + D_2 = 210 + 18 = 4052 \cdot \hat{\theta}_1 + 324 \cdot \hat{\theta}_2 (\hat{T} + 1)$$

$$D_1 + S_A^1 = 210 + 9 = 4052 \cdot \hat{\theta}_1 (\hat{T} + 1) + a_{81} \cdot 141 \cdot \hat{\theta}_1$$

The solutions to this system were found to be

$$\hat{T} = 6.73, \quad \hat{\theta}_1 = 6.6 \times 10^{-3}, \quad \text{and} \quad \hat{\theta}_2 = 9.0 \times 10^{-3}$$

giving the resulting $X^2$ statistic as 6.09. Since $\Pr(\chi^2_1 > 6.09) = 0.014$, the test indicates a significant departure from neutrality.

### Closing Comments on the Search for Domestication Genes

As we have seen, there is no shortage of formal tests for selection. Unfortunately, most of these tests are also strongly influenced by demography, such as a recent passage through a bottleneck, something that is expected for most domesticated crops. Several approaches have been suggested to correct for demographic signals, such as using a large number of markers (most of which are likely not influenced by selection) to estimate features about the common demograph. The simplest approach is to use outliers as an enrichment procedure for candidates (as opposed to a formal test of selection). Another strategy used by several authors is to compute several different test statistics, with the idea that the appropriately-chosen tests use independent signatures of departures from the strictly neutral model, so that those showing significant results over a number of such tests provide a strong signature for selection. Again, this is an enrichment procedure, not a formal statistical test.

A more formal approach is to use the large amount of marker information in a genomic scan to estimate parameters for some assumed demographic model, which can then be used in coalescent simulations to generate null distributions for the test statistic. The concern with this method is that most coalescent simulations are based on simple demographic models (such as a single bottleneck in a randomly-mating population). However, the true demographic situation during domestication is likely to be much more complex, with migration between a number of subpopulations and (often) a considerable amount of selfing in addition to random outcrossing. Hence, the appropriate demographic model to underlie a coalescent simulation can be very difficult to ascertain. Further, even with a lack of demographic issues, most tests have been developed for fully outcrossing populations, and special modification may be required to correctly account for the presence of significant selfing in many crops. Such extensions have to be developed in order for plant breeders to fully exploit the power of selection scans. Demographics can also have important complications for association mapping. Population subdivision can introduced correlations between alleles that are unlinked, which can lead to markers being associated within an unlinked QTL and hence faultly marker-trait associations.

A final interesting unanswered question concerns the relative strength of selection on domestication versus improvement genes. With a series of diverse lines in hand, one can distinguish between these two different phases of selection, as domestication genes will leave a signal in all lines, while improvement genes may leave a line-specific (or collection of sublines) signal. The very few initial studies have shown stronger selection on improvement genes. For example, the maize domestication gene $tb1$ has a 90 kb sweep signature, leading to an estimated strength of selection of $s = 0.05$. However, the improvement gene $Y1$ has a 600 kb sweep, giving an estimated strength of...
selection of \( s = 1.2 \) (Palaisa et al 2004). Likewise, the strength of selection on the rice improvement gene *Waxy* is estimated at \( s = 4.5 \) (Olsen et al. 2006). Obviously, this is too small a sample upon which to draw conclusions, but it does suggest that the strength of selection during improvement may be considerably stronger than during domestication. As we have seen, too intense selection (especially when selfing can occur) can result in a considerable linkage drag allowing deleterious alleles to accumulate and potentially favorable alleles to become lost. Thus, wild species subjected to very strong selection may not have sufficient variation for subsequent improvement, so it may indeed be a good thing if selection during domestication was weak.

Finally, at the risk of stating the obvious, all of the approaches discussed here are suitable for searches for adaptation genes (such as to water stress or high salt) in the wild relatives of domesticated crops.