

Lecture 8

Tests for Molecular Signatures of Selection

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.

Notes from a short course taught June 2006 at University of Aarhus

The notes for this lecture were last corrected on 23 June 2006. Please email me any errors.

There is a rich population genetics theory 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.

Basic Logic of Sequence-Based Selection Tests

There are a huge (and every growing!) number of tests, so we will focus here on some of the classic tests as well as the general logic behind these. We can loosely group tests into two classes: These using population data and those using phylogenetic data. Tests using population data rely on that nature and amount of within-species polymorphisms and (in some cases) between-species divergence. These tests attempt to detect selection at a target region or gene. In contrast, the latest generation of phylogenetic tests go beyond trying to detect selection on a gene, attempting to further locate specific amino acids residues that have been under positive selection.

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 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, which is 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).

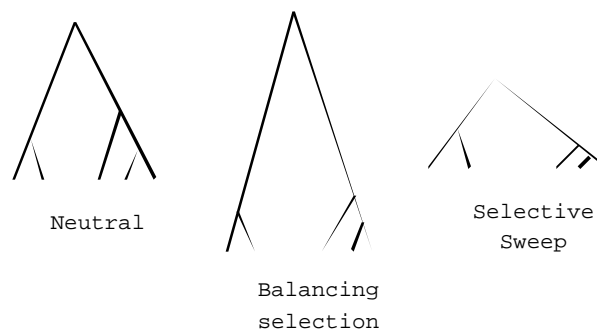


Figure 8.1. Coalescent times under pure drift and two types of selection. Under **balancing selection** (overdominance), the time back to the most recent common ancestor (MRCA) is longer than under pure drift. Under directional selection (often called a **selective sweep**), an allele sweeps through a population far quicker than under drift and hence has a more recent MRCA.

Recombination and Polymorphism

An important observation, first made in *Drosophila*, is that regions with reduced recombination (such as near centromeres and telomeres) tend to show reduced levels of polymorphism. If one plots the level of polymorphism within a sequence window versus the recombination rate in that window, a significant positive relationship is seen.

The initial explanation for this phenomena is that this is a signal of a **selective sweep**, with a linked site having a favorable mutation arising that sweeps through the population. In a region of reduced recombination, the size of the **hitchhiking** region (that part of the genome dragged along because of insufficient recombination with the selected locus) increases with decreasing recombination rate. Hence, regions of low recombination essentially have a higher chance of hitchhiking, as (everything else being equal), they contain more closely linked genes and hence more chances for favorable alleles to arise.

Put another way, the effect of frequent selective sweeps within a linked region is to lower the effective population size in that region. This results in decreased times to MRCA and hence less polymorphism. It is important to note, however, that while linkage may reduce the levels of standing variation through their reductions in N_e , linkage has essentially no effect on the average substitution rate at linked sites. This is because (Lecture 7), the rate of divergence between neutral sites is a function of the mutation rate, *independent* of population size.

A second explanation has been offered for the positive correlation between polymorphism and recombination rate, namely **background selection**. Here, selection *against* 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, there are a lot more deleterious mutations arising within a region than beneficial ones, and hence background selection can potentially have a very significant effect. Further, it is very difficult to distinguish between selective sweeps (selection for a new alleles) and frequent background selection (selection against new alleles), although we discuss some approaches for this.

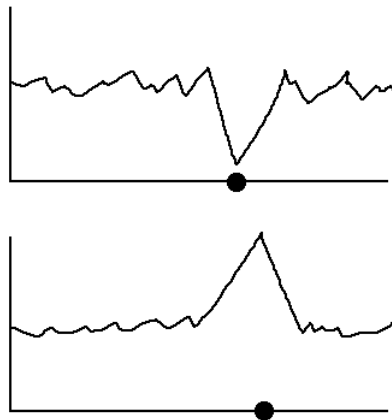


Figure 8.2. The impact of selection on variability at surrounding neutral sites. The vertical axis plots heterozygosity, the horizontal genome location. The upper graph shows the effect of a selective sweep, which results in a decrease in linked neutral polymorphisms around the selected locus (indicated by the filled circle). The width is a function of recombination (smaller c = larger width), selection (larger s = large width) and time of the sweep (longer the time since the sweep, the smaller the width). Plots such as this are generate by computing the variation in a window (of, say 100-1000 bases) that we slide along the genome. The lower graph shows that balancing selection results in an *increase* in the level of linked neutral polymorphisms.

Figure 8.2 shows another feature of recombination, with a decrease in variation at linked neutral sites around a locus under a selective sweep and an increase around a locus under balancing selection. While such a sliding window scan can be information, it is not definite proof of selection, as a local decrease (increase) in the mutation rate can also generate these patterns.

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, $2N\mu_{neu}$ is the expected number of mutations that arise per generation, each of which has probability $1/(2N)$ of being fixed, giving (Lecture 7) the neutral substitution rate as

$$d_{neu}/t = \mu_{neu} \quad (8.1a)$$

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

$$d_{fav}/t = 2N\mu_{fav} \cdot 2s = 4Ns\mu_{fav} \quad (8.1b)$$

The resulting ratio of favorable/neutral rates becomes

$$\frac{d_{fav}}{d_{nue}} = \frac{4Ns\mu_{fav}}{\mu_{nue}} = 4Ns \frac{\mu_{fav}}{\mu_{nue}} \quad (8.1c)$$

Hence, even though favorable mutations are expected to be far less frequent, provided $4Ns\mu_{fav} > \mu_{nue}$, $K_A/K_s > 1$ 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 μt . 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 μ . From Lecture 7, the standing heterozygosity and between-population divergence for the i th locus under drift are

$$H_i = 4N_e\mu_i, \quad d_i = 2t\mu_i \quad (8.2a)$$

Hence,

$$\frac{H_i}{d_i} = \frac{4N_e\mu_i}{2t\mu_i} = \frac{2N_e}{t} \quad (8.2b)$$

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 N_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 pair-wise 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 8.3 shows the difference. In this sample of five sequences, there are three haplotypes (and hence three alleles in the infinite alleles framework). However, in an infinite sites framework, looking over the six sites, we find that only two of these sites are segregating.

A	A	G	A	C
A	A	G	G	C
A	A	G	A	C
A	A	G	G	C
A	A	G	G	C

Figure 8.3. A sample of five sequences, showing three haplotypes but only two segregating sites.

Polymorphism-based tests compare the frequency of alleles with their expectations under the 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 generate 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.

The Infinite Alleles Model: Ewen's Sampling Formula

A classic result from population genetics 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

$$\Pr(K = k) = \frac{|S_n^k| \theta^k}{S_n(\theta)} \quad (8.3a)$$

where

$$S_n(\theta) = \theta(\theta + 1)(\theta + 2) \cdots (\theta + n - 1) \quad (8.3b)$$

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

$$\Pr(K = 1) = \frac{(n - 1)!}{(\theta + 1)(\theta + 2) \cdots (\theta + n - 1)} \quad (8.3c)$$

From Equation 8.3a, the mean and variance for the number of alleles can be found to be

$$E(K) = 1 + \theta \cdot \sum_{j=2}^n \frac{1}{\theta + j - 1} \quad (8.4a)$$

and

$$\sigma^2(K) = \theta \cdot \sum_{j=1}^{n-1} \frac{j}{(\theta + j)^2} \quad (8.4b)$$

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 (Lecture 3).

The typical setting is a sample of n sequences are 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 η , the **number of singletons**. 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	$E[k] = \theta$	$\sigma^2(k) = \theta \frac{n+1}{3(n-1)} + \theta^2 \frac{2(n^2+n+3)}{9n(n-1)}$
η = 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

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and} \quad b_n = \sum_{i=1}^{n-1} \frac{1}{i^2} \quad (8.5)$$

Under pure drift, the following three are all estimators of θ ,

$$\hat{\theta}_S = \frac{S}{a_n}, \quad \hat{\theta}_k = k, \quad \hat{\theta}_\eta = \frac{n-1}{n} \eta \quad (8.6)$$

$\hat{\theta}_S$ is called with **Waterson estimator** for θ (Watersons, 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 η .

Example 8.1. Suppose we sample 10 alleles from a population and observe $S = 12$, $k = 4$, and $\eta = 3$. What are the estimates of θ given these three summary statistics?

$$\begin{aligned} \hat{\theta}_S &= \frac{S}{a_{10}}, \quad a_{10} = \sum_{i=1}^9 \frac{1}{i} = 2.83, \quad \text{giving} \quad \hat{\theta}_S = \frac{12}{2.83} = 4.24 \\ \hat{\theta}_k &= k = 4 \\ \hat{\theta}_\eta &= \frac{n}{n-1} \eta = \frac{10}{9} \cdot 3 = 3.33 \end{aligned}$$

Tajima's D Test

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

$$D = \frac{\hat{\theta}_k - \hat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} \quad (8.7a)$$

where

$$\alpha_D = \frac{1}{a_n} \left(\frac{n+1}{3(n-1)} - \frac{1}{a_n} \right) - \beta_D \quad (8.7b)$$

$$\beta_D = \frac{1}{a_n^2 + b_n} \left(\frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2} \right) \quad (8.7c)$$

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 significant less than predicted from the number of polymorphisms.

Example 8.2. 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

$$a_7 = \sum_{i=1}^6 \frac{1}{i} = 2.45, \quad b_7 = \sum_{i=1}^6 \frac{1}{i^2} = 1.49$$

$$\hat{\theta}_S = \frac{S}{a_n} = \frac{45}{2.45} = 18.38, \quad \hat{\theta}_k = k = 15.38$$

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

$$\alpha_D = \frac{1}{2.45} \left(\frac{7+1}{3(7-1)} - \frac{1}{2.45} \right) - 0.0417 = -0.0269$$

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

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).

Fu and Li's D^* and F^* tests

Fu and Li (1993) introduced two tests, based on the two other contrasts of the three infinite-sites θ estimators (Equation 8.6). Their **D^* test** compares the segregating sites (S) versus singletons (η) estimator of θ ,

$$D^* = \frac{\hat{\theta}_S - \hat{\theta}_\eta}{\sqrt{\alpha_* S + \beta_* S^2}} \quad (8.8a)$$

$$\alpha_* = \frac{1}{a_n} \left(\frac{n+1}{n} - \frac{1}{a_n} \right) - \beta_* \quad (8.8b)$$

$$\beta_* = \frac{1}{a_n^2 + b_n} \left(\frac{b_n}{a_n^2} - \frac{2}{n} \left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n} \right) - \frac{1}{n^2} \right) \quad (8.8c)$$

While their **F^* test** compares the average pair-wise divergence (k) versus singletons (η) estimator of θ ,

$$F^* = \frac{\hat{\theta}_k - \hat{\theta}_\eta}{\sqrt{\alpha_F S + \beta_F S^2}} \quad (8.9a)$$

$$\alpha_F = \frac{1}{a_n} \left(\frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)} \right) - \beta_F \quad (8.9b)$$

$$\beta_F = \frac{1}{a_n^2 + b_n} \left(\frac{2n^4 + 110n^2 - 255n + 153}{9n^2(n-1)} + \frac{2(n-1)a_n}{n^2} - \frac{8b_n}{n} \right) \quad (8.9c)$$

These expressions are from Simonsen et al (1995), with Equation 8.9c correcting a typo in the original Fu and Li paper. Critical values are tabulated by Fu and Li (1993). While these tests are fairly widely used, Simonsen et al. (1995) found that they are not as powerful as Tajima's test for detecting a selective sweep or population structure departures (bottlenecks or population subdivision). However, Fu (1997) found that both tests have more power than Tajima's D for detecting signals of background selection.

Depaulis and Veuille's K and H tests

Depaulis and Veuille (1998) offered two interesting tests that, in effect, look at the sequence data from *both* the infinite alleles and infinite sites perspectives. Specifically, they looked at two infinite

allele measures, the number of haplotypes K (read alleles) and the diversity of haplotypes H (read heterozygosity of alleles), where

$$H = 1 - \sum_{i=1}^K p_i^2, \quad \text{where } p_i = \text{frequency of } i\text{th haplotype} \quad (8.10)$$

They compared the observed values of these two statistics conditioned on the number S of segregating sites (infinite sites model). Their **K test** and **H test** compare, respectively, the observed values of K and H with the neutral values expected given S . Hypothesis testing was accomplished by simulating neutral coalescents conditioned on the observed values of S and sample size n , and Depaulis and Veuille present tables for sample sizes up to $n = 60$. Also see Wall and Hudson (2001) for commentary on their simulation approach.

Fu's W and F_S Tests

Fu (1996, 1997) offered a number of tests of selection, with a goal of offering more refined test for specific settings, such as too few alleles or too many alleles. We consider on two of his proposed tests here.

Fu's **W test** (1996) is based on Ewen's sampling formula (Equation 8.3a). Suppose we have an estimate $\hat{\theta}$ of θ and we observe k alleles in our sample. The probability of seeing k (or fewer) alleles in our sample is just

$$W = \Pr(K \leq k) = \sum_{i=1}^k \Pr(K = i | \hat{\theta}) = \sum_{i=1}^k \frac{|S_n^i| \hat{\theta}^i}{S_n(\hat{\theta})} \quad (8.11)$$

where

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

the W test is essentially a test for an deficiency of rare alleles, and hence is a *one-sided* test. Fu's test uses the Watterson (1975) estimator $\hat{\theta} = S/a_n$, where S is the number of segregating sites. Fu (1996) showed that the W test is more powerful than Tajima's D and Fu and Li's D^* and F^* tests for detecting samples from a structured population (as also occurs with overdominant selection).

Fu's **F_S test** (1997) is the compliment of his W test, being a test for excess rare alleles. It starts by computing the probability of seeing k or more alleles in our sample,

$$S' = \Pr(K \geq k) = \sum_{i=k}^n \frac{|S_n^i| \hat{\theta}^i}{S_n(\hat{\theta})} \quad (8.12a)$$

but now using $\hat{\theta}_k$, the estimator of θ based on average number of pairwise differences. Fu notes that S' is not an optimal test statistic because its critical point are often too close to zero. Because of this, the test statistic S is the logistic of S' ,

$$F_S = \ln \left(\frac{S'}{1 - S'} \right) \quad (8.12b)$$

F_S is negative when there is an excess of rare alleles (as occurs with an excess of recent mutations as would occur with a selective sweep or population expansion), with a sufficiently large negative value being evidence for selection. Hence, F_S is a one-sided test. Fu (1997) showed that F_S is more powerful than Tajima's and Fu and Li's tests for detecting population growth/selective sweeps. Conversely, Fu and Li's tests are more power for detecting background selection.

Fay and Wu's H Test

Fay and Wu (2000) and Kim and Stephan (2000) note that a distinct signal is left by a selective sweep that is not left by background selection. Specifically, it is common to see alleles that have newly arise by mutation at high frequency following a sweep (as they hitched along for the ride). With background selection, this feature is not expected. This is the basis for Fay and Wu's **H test**, which disproportionate weights derived alleles at high frequency. Their test requires an **outgroup** so that one can access whether an allele occurs in the outgroup or has recently been derived by mutation. Such derived alleles are expected to be at lower frequency (as under neutrality, the frequency of an allele is a rough indicator of its age, with older alleles being more frequent). The test processed as follows. Let S_i denote the number of derived mutants found i times in our sample of size n . For example, if there are 5 unique (derived) alleles, 4 alleles each appearing twice, and one allele appearing 5 times in our sample of size 18, then $S_1 = 5$, $S_2 = 4$, $S_5 = 1$. The estimate of θ from the average pair-wise difference expressed in terms of the S_i is

$$\hat{\theta}_k = 2 \sum_{i=1}^{n-1} \frac{S_i i(n-i)}{n(n-1)} \quad (8.13a)$$

while an estimate of θ weighted by homozygosity is

$$\hat{\theta}_H = 2 \sum_{i=1}^{n-1} \frac{S_i i^2}{n(n-1)} \quad (8.13b)$$

Fay and Wu's **H test** is given (ala D , D^* , and F^*) by the scaled difference of $\hat{\theta}_H - \hat{\theta}_k$.

Given that Day and Wu's test weights derived allele at high frequency, a significant H and D test is consistent with a selective sweep, while a significant D test, but *not* a significant H test suggests background selection or demographic features more likely accounts for the departure from neutrality.

Genome-Wide Polymorphism Tests

As mentioned several times, polymorphism-based tests suffer in that we reject the null hypothesis (neutrality), 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

One of the very first tests for selection with sequence 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.

The Linkage Disequilibrium Decay (LDD) Test

As mentioned, one feature of selective sweeps is that they have an excess of newly-derived alleles at high frequency. We have already (Fay and Wu) seen one specific test for this. 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. This is the basis for the **Linkage Disequilibrium Decay, LDD Test** of Wang et al. (2006). They applied this idea on a massive human data set of 1.6 million SNPs, resulting in 1.6% of the markers showing some signatures of positive selection. Simulation studies by Wang et al. found that the LDD test effectively distinguishes selection from population bottlenecks and admixture (population structure).

All genome-based tests have an *important caveat*. The large number of markers used are typically generated by looking for polymorphisms 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 8.2. 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

$$\frac{d_{syn}}{d_{rep}} = \frac{2t\mu_{syn}}{2t\mu_{rep}} = \frac{\mu_{syn}}{\mu_{rep}} \quad (8.14a)$$

Likewise the ratios of heterozygosity at the two locations is

$$\frac{H_{syn}}{H_{rep}} = \frac{4N_e\mu_{syn}}{4N_e\mu_{rep}} = \frac{\mu_{syn}}{\mu_{rep}} \quad (8.14b)$$

Hence, these two ratios have the same expected value. We note that McDonald and Kreitman provide a more general derivation of 8.14a, replacing $4N_e$ (the equilibrium value) by T_{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 8.3. McDonald and Kreitman (1991) look at the *Adh* (Alcohol dehydrogenase) locus in *Drosophila*, specifically the sibling species *D. melanogaster*, *D. simulans*, and the outgroup *D. yakuba*. Looking at fixed differences, a total of 24 occur, 7 of which were replacement, 17 synonymous. Turning to polymorphisms, 44 polymorphic sites were found, 2 of which were replacement and 42 synonymous, giving

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

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

Hudson-Kreitman-Aguade (HKA) Test

Hudson, Kreitman, and Aguade (1987) proposed the first tests 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 $i = 1, \dots, L$ 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 $L S_i^A$ values, for the number of segregating sites at locus i in species (population) A , another $L S_i^B$ 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). Given these $3L$ summary statistics, the HKA test X^2 is given by

$$X^2 = \sum_{i=1}^L \frac{(S_i^A - \widehat{E}(S_i^A))^2}{\widehat{Var}(S_i^A)} + \sum_{i=1}^L \frac{(S_i^B - \widehat{E}(S_i^B))^2}{\widehat{Var}(S_i^B)} + \sum_{i=1}^L \frac{(D_i - \widehat{E}(D_i))^2}{\widehat{Var}(D_i)} \quad (8.15)$$

where, for n_A samples from species A and n_B samples from species B ,

$$\widehat{E}(S_i^A) = \widehat{\theta}_i a_{n_A}, \quad \widehat{E}(S_i^B) = \widehat{\alpha} \widehat{\theta}_i a_{n_B} \quad (8.16a)$$

$$\widehat{Var}(S_i^A) = \widehat{\theta}_i a_{n_A} + \widehat{\theta}_i^2 b_{n_A}, \quad \widehat{Var}(S_i^B) = \widehat{\alpha} \widehat{\theta}_i a_{n_A} + \widehat{\alpha}^2 \widehat{\theta}_i^2 b_{n_B} \quad (8.16b)$$

$$\widehat{D}_i = \widehat{\theta}_i \left(\widehat{T} + \frac{1 + \widehat{\alpha}}{2} \right) \quad (8.16c)$$

$$\widehat{Var}(D_i) = \widehat{\theta}_i \left(\widehat{T} + \frac{1 + \widehat{\alpha}}{2} \right) + \left(\frac{\widehat{\theta}_i (1 + \widehat{\alpha})}{2} \right)^2 \quad (8.16d)$$

Equations 8.16a and 8.16b follow from our above results for the infinite sites model. Equation 8.16c follows by re-writing

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

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 $L \theta_i^x$ values and two demographic parameters, T and α . The HKA test estimates these parameters and then (using Equation 8.16) computes the goodness of fit X^2 statistic (Equation 8.15), which is approximated χ^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,

$$\begin{aligned} \sum_{i=1}^L S_i^A &= a_{n_A} \sum_{i=1}^L \hat{\theta}_i \\ \sum_{i=1}^L S_i^B &= \hat{\alpha} a_{n_B} \sum_{i=1}^L \hat{\theta}_i \\ \sum_{i=1}^L D_i &= \left(\hat{T} + \frac{1 + \hat{\alpha}}{2} \right) \sum_{i=1}^L \hat{\theta}_i \\ S_i^A + S_i^B + D_i &= \hat{\theta}_i \left(\hat{T} + \frac{1 + \hat{\alpha}}{2} \right) + a_{n_A} + \hat{\alpha} \cdot a_{n_B} \quad \text{for } i = 1, \dots, L - 1 \end{aligned} \tag{8.17}$$

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

Example 8.4. 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 θ_i be the per-nucleotide θ value, so that we have to weight the θ value for each term by the number of sites compared, giving Equation 8.17 as

$$\begin{aligned} S_1^A + S_2^B &= 9 + 8 = a_{81}(414 \cdot \hat{\theta}_1 + 79 \cdot \hat{\theta}_2) \\ D_1 + D_2 &= 210 + 18 = 4052 \cdot \hat{\theta}_1 + 324\hat{\theta}_2(\hat{T} + 1) \\ D_1 + S_1^A &= 210 + 9 = 4052 \cdot \hat{\theta}_1(\hat{T} + 1) + a_{81} \cdot 141 \cdot \hat{\theta}_1 \end{aligned}$$

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 $\Pr(\chi_1^2 > 6.09) = 0.014$, the test indicates a significant departure from neutrality.

Tests Based on Between-Population (Species) Divergence

While tests based on between-species differences have the disadvantage of requiring sequences from the gene of interest from a number of closely-related species (or very divergent populations from the same species), they have the distinct advantage of not being confounded by issues of population demography. As mentioned above, this follows because while levels of polymorphism are functions of the effective population size, divergence between populations is simply a function of time and the mutation rate. Ford (2002) in a summary of published reports for evidence of selection noted that of the 119 (as of 2002) reports of positive selection, 25 were based on polymorphism and/or divergence tests (such as Tajima's and HKA), 9 had a significant McDonald-Kreitman test, and 92 were detected by tests based on K_A/K_S ratios.

Define $\omega = K_A/K_S$, the ratio of nonsynonymous to synonymous substitution rates. It has long been recognized (Kimura 1983) that evidence for positive (i.e., directional) selection is given when a protein shows $\omega > 1$. The problem is that while one or a few sites may be under strong directional selection ($\omega \gg 1$ at these residues), most sites in a protein are expected to be under some selective constraints ($\omega \ll 1$), so that the average over all sites gives that protein an $\omega < 1$. Indeed, a meta-analysis by Endo et al (1996) found that only 17 out of 3595 proteins showed $\omega > 1$. However, there were a few early success stories. For example, Hughes and Nei (1988) used the 3-D protein structure of genes in the major histocompatibility complex to suggest potential residues (those on the surface in critical positions) to test for, and find, $\omega > 1$ for these residues. However, for most proteins, we don't have this amount of biological detail to draw upon.

Given these constraints, two general approaches have been suggested to estimate ω given a set of sequences from a group of closely-related species. All need a phylogeny for these sequences, and issues such as the correct multiple sequence alignment as well as errors in the assumed tree potentially loom in the background, but are general not addressed (perhaps rightly so). **Parsimony-based** approaches reconstruct the sequence at each node in the tree, and then use these to count up the number of synonymous and nonsynonymous substitutions. **Likelihood approaches** are on a more firm statistical footing, but are computationally intense and can be rather model-specific. Both approaches allow for tests not just whether a protein is under selection, but more exciting what *sites* in that protein are under positive selection.

Parsimony-Based Ancestral Reconstruction Tests

Fitch et al. (1997) and Suzuki and Gojobori (1999) proposed similar parsimony-based approaches for detecting selection on single sites. Both methods start with a phylogeny and then use parsimony (choose the solution requiring the fewest number of changes) to reconstruct the ancestral sequences at all of the nodes in the tree. With these estimated sequences in hand, one can then simply count the number of synonymous and nonsynonymous substitutions on the tree. The approach of Fitch et al. assumes all sites are the same, while the approach of Suzuki and Gojobori allows sites to vary. The false-positive rate of these methods is generally small (Suzuki and Gojobori 1999, Suzuki and Nei 2002), but they suffer from low power (Wong et al 2004). Further, given that the ancestral states are likely estimated with error, the analysis has no formal procedure to take this uncertainty into account when computing a p value for any site. Bayesian posterior distributions can account for these errors, but this requires moving from a parsimony to a likelihood framework.

Maximum-Likelihood-Based Codon Tests

Maximum-likelihood methods make no assumptions as to the ancestral state at each of the nodes in a tree, but rather uses an explicit model of codon change. For example, Goldman and Yang (1994) considered the following model for transitions between codons i and j ($1 \leq i, j \leq 64$) [as the three

stop codons are ignored],

$$q_{ij} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega\pi_j & \text{for a nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases} \quad (8.18)$$

A similar model was suggested by Muse and Gaut (1994). Here, π_j is the equilibrium frequency of codon j (calculated from the nucleotide frequencies at the three codon positions), while κ and ω are estimated parameters to account for biases in codon changes. First, it is well known that **transitions** ($A \leftrightarrow G, C \leftrightarrow T$) can occur at different rates than **transversions** (e.g., $A \leftrightarrow T$, etc.) The parameter κ accounts for this, and can also account for codon usage bias in many cases. Of greater interest is ω , the substitution rate ratio. An estimated $\omega > 1$ is direct evidence for directional selection.

Tests for directional selection on a gene is accomplished by using this codon model superimposed on the phylogenetic tree, running likelihood calculation to find the ML solutions for \mathbf{Q} matrix parameters. This allows for a direct test that $\omega > 1$ using Likelihood Ratio tests (Lecture 1). The key to these likelihood calculations is that $\mathbf{P}(t)$, the codon transition matrix at time t , is related to the \mathbf{Q} matrix by

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \quad (8.19)$$

Here,

$$P_{ij}(t) = \Pr(\text{codon} = i \text{ at time } t \mid \text{codon is } j \text{ at time } t = 0) \quad (8.20)$$

Recalling that if we can diagonalize the matrix \mathbf{Q} , then $\mathbf{Q} = \mathbf{U}\mathbf{A}\mathbf{U}^T$ where \mathbf{A} is a diagonal matrix, with i th diagonal elements λ_i , the eigenvalues of \mathbf{Q} (Lecture 14). Then

$$\exp(\mathbf{Q}t) = \mathbf{U} \exp(t\mathbf{A}) \mathbf{U}^T$$

where

$$\exp(t\mathbf{A}) = \text{diag}(e^{t\lambda_1}, e^{t\lambda_2}, \dots, e^{t\lambda_n}) = \begin{pmatrix} e^{t\lambda_1} & 0 & \dots & 0 \\ 0 & e^{t\lambda_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & e^{t\lambda_n} \end{pmatrix} \quad (8.21)$$

A variety of likelihood models based on Equation 8.21 are tested (much in the same way that we test subset of complex segregation analysis models in Lecture 16), adding more factors (i.e., nonzero κ , etc.) if they improved model fit. Evidence for selection on a gene is indicated if the likelihood ratio test for $\omega > 1$ is significant.

The power of this approach has been examined by Anisimova et al. (2001). Power is a function of two different sample sizes: the number of codons in the sequence and the number of actual sequences. Typically, an investigator has very little control over the number of codons sequenced in a gene, but the more codons in a gene, the better, although 100 seems to give reasonable power (an amino acid chain 100 residues long). Power is more efficiently increased by adding more sequences, as opposed to looking at more codons. Five to six sequences offer little power, but 15-20 sequences can offer considerable power. Hence, for very short sequences, the method typically lacks power, but for moderately long sequences with a modest phylogeny (10-20 species), power can be quite reasonable. They also found that very similar, and very divergent, sequences both offer little power. Hence, sequences from divergent populations of the same species may lack power (viruses are a counterexample, due to their very high mutation rate). Likewise, a phylogeny based on rather distant sequences can also lack power.

Nielsen et al. (2005) uses this ML approach to compare 13,700 genes from humans with their chimpanzee orthologs, finding a number of genes with signatures of positive selection.

Bayesian Estimator of Sites Under Positive Selection

A limit with this type of likelihood approach is that we are still testing the entire gene, assuming a single ω ratio at all sites. As we mentioned above, this is a problem with gene-wide tests, as sites under strong directional can be obscured by the sea of sites that are conserved. A major innovation to resolve this problem was offered by Nielsen and Yang (1998) and Yang et al. (2000). Building on the likelihood approach, they further assumed that sites fell into several categories (for example, neutral, selected against, selected for). Further, within the selected sites, ω was allowed to vary, with values being drawn from a distribution (such as a Gamma), with the distirbutional parameters also fit by ML. The direct test for selection on the gene is whether adding the $\omega > 1$ distribution significantly improves model fit. If such selection is indicated, a very powerful feature with this approach is that, with the ML-estimated parameters in hand, one can use Bayes' theorem (Equation 1.4) to assign posterior probabilities that a particular site is in a specific category. Suppose there are two classes, neutral (n) and selected (s). From Bayes' theorem, the probability that a specific site is in the selected category is just

$$\Pr(s | A) = \frac{\Pr(s) \Pr(A | s)}{\Pr(s) \Pr(A | s) + \Pr(n) \Pr(A | n)} \quad (8.22)$$

where A is the pattern of codons for that site in the tree. Thus, Yang's approach allows us to directly assign probabilities of selection to any particular site.