9

Using Molecular Data to Detect Selection: Signatures From Recent Single Events

For the past 20 years, there has been a tendency on the part of journal editors and reviewers to assume that every case of alleged statistical evidence for positive selection is worthy of publication, even in the absence of a plausible biological mechanism underlying the alleged selection. — Hughes (2007)

Draft version 12 April 2014

While the ubiquity of **purifying** (or **negative**) **selection** (the removal of deleterious alleles) at the molecular level is well established, the frequency of **positive** (or **adaptive** or **Darwinian**) selection remains unclear. Because of this, the development of methods to detect the latter is a major growth industry in evolutionary genetics. There is a massive population-genetics literature on this subject, and a partial (but not exhaustive) list of recent reviews includes Kreitman (2000), Nielsen (2001), Ford (2002), Bamshad and Wooding (2003), Schlötterer (2003), Guinand et al. (2004), Storz (2005), Wright and Gaut (2005), Nielsen (2005), Sabeti et al. (2006), Biswas and Akey (2006), Thornton et al. (2007), Stinchcombe and Hoekstra (2008), Pavlidis et al. (2008), Akey (2009), Nei et al. (2010), Siol et al. (2010), Oleksyk et al. (2010), Stephan (2010a), and Fu and Akey (2013).

As detailed in Chapter 8, a single recent event of positive selection can leave a transient signal in the pattern of linked neutral variation, and the detection of such events is the subject of this chapter. Since most approaches for detecting recent/ongoing selection use the segregating variation in a sample from a contemporaneous population, we loosely refer to these as **polymorphism-based tests**. Because such signals are transient, these methods can only detect an ongoing, or very recent, single event over a very short time window. Hence, they work only over ecological time scales, detecting single events occurring over less than N_e generations ago. In contrast, a *history* of positive selection on a gene over evolutionary time can leave a *cumulative* signal in the pattern of substitutions. **Divergence-based tests** to detect these patterns, which require data on substitutions between species (or very distantly-related populations), are developed in Chapter 10 (which also covers tests that jointly use polymorphism and divergence data). These different approaches are complementary, as an adaptive substitution could leave a strong (but fleeting) signature over an ecological time scale, but essentially no signal in that gene over an evolutionary time scale (adding just one more substitution in a potential background of numerous neutral fixations). Likewise, the vast majority of adaptive events that have shaped a gene occurred in its distant past and hence leave no currently-detectable polymorphism pattern.

Since the search for sites under selection can be seductive, it is important to stress that the most important point about the methods developed in this chapter are their *limitations*. They *potentially* can be useful in detecting *some* events that involve the signatures from a single selective event at a single site. In principle, this allows for a crude localization of the site, and allows the prospect of studying *individual* (as opposed to cumulative) selective events. As discussed in Chapter 8, one source of these limitations is the nature of the signal left by the sweep itself. First, it is *very fleeting*, typically persisting for between $0.1N_e$ to N_e generations, depending on the features being examined. Second, *only a fraction of such*

events, even if ongoing or very recent, can be detected. A weak, but nontrivial, selection event may be too small to leave a meaningful signal against a noisy molecular background. Finally, even very strongly selected sites may not be detectable, especially if they involve soft or polygenic sweeps (Chapter 8). Even with all these concerns, the most critical problem with polymorphism-based tests are the ***confounding effects from demography/population structure***. The rapid expansion following a population bottleneck leaves sweep-like signal, while the presence of population structure can mimic balancing selection. One reason for the vast number of tests is that no single one is best in all settings, and the search to find strong signals *unique* to positive selection has, for the most part, as been unsuccessful.

Besides being fundamentally important to our understanding of evolution, tests of selection can also be a helpful tool to a breeder. Scans for ongoing and recently selected sites can provide a useful complement to QTL/association-mapping studies. In the latter, one specifies the traits of interest and then searches for marker-trait associations. While this is a powerful approach, it is also limited by the choice of traits. In a population undergoing selection for (say) water stress, one might miss pathways for adaptation that are not obvious and hence do not involve the traits chosen to be mapped. Conversely, one could perform a genomic scan on a population under relatively recent water stress to look for sites showing signatures of ongoing/recent selection. These in turn could suggest genome regions of interest, which in turn may suggest traits/pathways under selection, without every having to specify particular traits in the first place. This approach has been termed **natural selection mapping** (Kohn et al. 2000), **hitchhiking mapping** (Schlötterer 2003), and **reverse ecology** (Li et al. 2008), and is widely used in the search for **domestication genes** responsible for the transition from wild relatives into domesticated lines. We review several examples at the conclusion of the chapter.

AN OVERVIEW OF STRATEGIES BASED ON SEGREGATING VARIATION

For several reasons, there is no single omnibus test for selection. Different scenarios for positive selection (e.g., hard sweeps, soft sweeps, partial sweeps, balancing selection) leave different, and often conflicting, signatures (Chapter 8), so that tests for one type (e.g., hard sweep) may easily miss signatures from another (e.g., soft sweeps). Second, different tests are designed to detect signals from different time periods during (and following) a sweep. Third, different sampling schemes are possible. Most tests assume a single sample from a current population, but one might have contemporaneous samples from several related populations or a temporal series of samples from a single population. Finally, advances in genomics have enormously expanded the ability to score molecular variation and this is reflected in the historical development of tests. The first test followed changes in a single allele at one locus (Examples 9.1, 9.2), while later tests evolved to use data from **genomic scans**, where a very large number of sites are scored, and potentially phased (generating haplotype, as opposed to sequence, data). Approaches to detect recent selection can be classfied into five categories, which loosely follow the historical development of attempts to infer selection:

1) Excessive allele-frequency change. The first formal test of selection was proposed by Fisher and Ford (1947), who used the machinery developed in Chapter 2 for the divergence under drift to test for excessive change in a time series of allele frequencies from a single population. While perhaps the most unambiguous signature of selection, this approach requires long-term monitoring of a population and some reasonably independent estimate of N_e.

2) Excessive allele-frequency divergence. Lewontin and Krakauer (1973) proposed using the divergence between a series of contemporaneous-sampled populations (presumably from a common ancestor) to test for selection. The same machinery from Chapter 2 predicts the expected divergence, measured by Wright's F_{ST} statistic for population structure, under drift. Loci displaying excessive F_{ST} values relative to drift are selection candidates. Migration between populations following their divergence can compromise these tests.

The above two categories require samples from multiple populations (either temporally or spatially), which limits their widespread use. A less demanding design is a single population sample, as employed in the three remaining categories.

3) Chromosomal spatial patterns of variation. As detailed in Chapter 8, a sweep leaves a characteristic decrease in polymorphism around a selected site, and a number of formal likelihood-based tests are based on the expected pattern of the nucleotide diversity π as a function of the recombination distance c from the sweep (Equation 8.8a). Early versions of these tests assumed that the population was in mutation-drift equilibrium at the start of the sweep, while more recent versions attempt to remove this strong assumption.

The final categories divide tests by whether they assume an infinite-sites or an infinite-alleles framework, using the neutral-equilibrium results for these models developed in Chapter 2. Recall that the infinite-sites framework considers a sequence as a series of separate sites, while the infinite-alleles framework treats each different DNA sequence (haplotype) as a different allele (Figure 2.7). Both models assume that the region being considered is small enough that recombination within the sample can be ignored. Give the large (and diverse) number of tests in both of these categories, each section includes a summary table of the tests covered (Table 9.1 for infinite-sites tests, Table 9.3 for haplotype tests).

4) Changes in the site-frequency spectrum. At mutation-drift equilibrium, the frequency spectrum at neutral sites is given by the Watterson distribution (Equation 2.34). Starting with Tajima (1989), a number of tests have been proposed that look for shifts in this spectrum following a sweep, such an excess of sites with rare alleles or an excess of sites with high-frequency derived alleles. The major complication with this class of tests is that changes in population demography (such as a recent expansion or contraction) or the presence of population structure (migration between partly-isolated populations) mimic signatures of selection.

5) Tests based on haplotype information. Under the infinite-alleles model, the number of alleles in a sample at mutation-drift equilibrium is given by Ewens' sampling formula (Equation 2.30a) and their allele-frequency spectrum by Equation 2.33a. Starting with Ewens (1972) and Watterson (1977, 1978), a number of tests use these results to look for departures from the neutral equilibrium model. As with tests based on the site-frequency spectrum, significant departures can occur for neutral alleles if the population is not in equilibrium or if population structure is present.

Two other strategies also use haplotype information. The first searches for the distinct signatures in the pattern of pairwise linkage disequilibrium (LD) predicted around a hard or a soft sweep (Table 8.2). The second considers the frequency of a neutral allele as a function of its age (Equation 2.12). Under neutrality, a common allele is an old allele

with shorter blocks of LD, reflecting a longer history of recombination. The presence of high-frequency alleles with long haplotypes (large blocks of LD) offers a signature of selection. A key point is that haplotype (and LD) structure provides ***signals that can be missed by site-frequency and hard-sweep tests***, and thus offers more power in some settings. In particular, they can provide ***strong signals of ongoing/partial and soft sweeps***.

Attempts to Account for Departures from the Equilibrium Model

Most tests for selection are based on the null hypothesis of the neutral equilibrium model (Chapter 2), resulting in a composite alternative hypothesis. Rejection of this null can indeed imply a signature of selection, but can also simply imply that a neutral population is not in mutation-drift equilibrium. Cavalli-Sforza (1966) noted that demography and/or population structure should leave a common signal over all genes within a genome, and this observation has been used in attempts to correct for any non-equilibrium features in the data. The simplest approach is the **outlier method**, where values of the test statistic are computed for a large number of genes, with outliers suggesting potential targets of selection. This is an *enrichment method*, not a formal test. The second approach is to use data from presumably neutral markers unlinked to a region of interest to infer the population history (bottlenecks, expansions, population structure). These can then be used to simulate the coalescent structure (Chapter 2) for neutral alleles under this non-equilibrium model, which in turn can be used to generate the distribution of the test statistic under this more appropriate null. A final approach is to use presumably neutral sites to generate an empirical site-frequency spectrum to use in place of the equilibrium Watterson distribution.

These approaches are based on using information from a large number of loci obtained in a genomic scan, with the assumption that most sites are not under positive selection and hence provide information to better shape the null hypothesis. This critically relies on the validity of Cavalli-Sforza's assumption of a common demographic/population structure signal over all loci, upon which any additional signal from selection is placed. Unfortunately this not need be the case, especially in a population that is expanding over space. **Allelic surfing** can occur, wherein random alleles (and new mutations) on the leading edge of a wave of population expansion surf rather quickly to high frequencies in newly-founded parts of the population (Edmonds et al. 2004; Klopfstein et al. 2006; Hallatschek et al. 2007; Travis et al. 2007; Excoffier and Ray 2008; Hallatschek and Nelson 2008, 2009; Excoffier et al. 2009; Hallatschek 2011). Since neutral alleles on the leading wave of expansion are largely random, surfing *does not affect all genomic locations equally*, and as a result can mimic signatures of selection even after corrections for demography/structure based on others markers within the sample. This is especially troublesome as the model species most surveyed for recent selection — humans, cosmopolitan human commensal *Drosphila* (*melanogaster* and *simulans*), and *Arabidopsis* — are all known to have undergone relatively recent massive population expansions. Hofer et al. (2009) found that while a large fraction of the human SNPs, STRs, and indels show large (greater than 0.3) differences in frequency across world populations, this pattern is easily accounted by allelic surfing, suggesting that this can be a considerable problem in the search for sites under recent selection.

SNP Ascertainment Bias

Another concern common to many tests is **SNP ascertainment bias**, which arises when molecular variation is scored using SNPs. In a typical SNP discovery setting, one sequences a relative small pool of individuals (the **SNP discovery panel**) to "discover" SNPs — polymorphic nucleotides whose minor allele is above some critical frequency in the panel. These are then used to score a much larger sample of individuals, creating a severe bias in favor

of SNPs at intermediate frequency and against rare SNPs (sites with very low minor allele frequencies). Likewise, if the SNP discovery panel is from a different population than the screened sample, this also creates bias in that important SNPs in the population of interest can be missed (e.g., Ptak and Przeworski 2002). When the frequency of SNP minor alleles in the discovery panel are known, corrections for SNP ascertainment can be straightforward (Nielsen et al. 2004). However, often SNP discovery is a more complex process, creating biases that simple methods can reduce, but not remove (Clark et al. 2005). While direct sequencing resolves these issues, sequencing errors can inflate the number of rare alleles in the sample, which is an important signal for several tests.

SNP Polarity Assignment Errors

A final source of bias can appear in tests requiring the **polarity** status (i.e., ancestral or derived) of a SNP allele. Recall from Chapter 2 the distinction between **unfolded** and **folded** frequency spectrum. The former is based on the frequency of derived alleles (i.e., the Watterson distribution), whereas the latter is based on the frequency of the minor allele and thus immune to polarity assignment errors. Typically, polarity is accessed using an outgroup, with the outgroup allele assumed to be the ancestral stage. This is a parsimony assumption, assuming that no back- or parallel-mutations occur. Incorrect polarity assignments can result in miscalling a low-frequency derived allele as a high-frequency ancestral one, and even a few such miss-calls can significantly impact certain tests (Baudry and Depaulis 2003; Hernandez et al. 2007).

Structure of the Remainder of this Chapter

The rest of the chapter is structured into treatments based on our five categories of tests. These are largely constructed for convenience of presentation, and some tests draw upon ideas from several different categories. Given the amount of information in this chapter, we have tried to make our discussion of each category largely autonomous of the others, allowing the reader to skip directly into the section they feel most appropriate for their needs. We conclude with a brief review of scans for recent positive selection in humans and domesticated/commensal organisms, with a focus on the search for domestication genes.

ALLELE-FREQUENCY CHANGE IN A SINGLE POPULATION

There are several settings where tests based on allele-frequency change may be appropriate. One is a population monitored over some reasonable period of time, which was the basis for the first formal test by Fisher and Ford (1947) of whether a specific gene is under selection (Examples 9.1 and 9.2). The second is a population under artificial selection, which has also been proposed as an approach for QTL mapping of the trait of interest (Nuzhdin and Pasyukova 1991; Nuzhdin et al. 1993; Keightley and Bulfield 1993; Ollivier et al. 1997). While excessive allele-frequency change is perhaps the most unambiguous signal of selection, there are also power issues when the number of generations separating the first and last samples is modest (De Kovel 2006). Given that the time scale for significant allele-frequency change under selection is $\sim 1/s$, sampling based on a modest number of generations requires strong selection for a signal (to detect a significant change even in the absence of drift requires that the sample size n and number of generations t satisfy $tn \gg 1/s$). We start by developing formal tests based on two time points and then expand to a time series with k points. Because these tests are for single loci, we conclude with a discussion of genomic scans for frequency change.

Allele-frequency Change Over Two Sample Points: The Waples Adjusted Test

Chapter 4 considered the estimation of N_e from allele-frequency change, a setting where one typically averages over a number of loci to reduce the evolutionary sampling variance. Here our task is the complementary problem, *given* some estimate of N_e, is the observed change in allele frequency at a candidate locus excessive? If so, this presumably reflects directional selection acting at, or close to, this region. In theory, one could also test for too little divergence (reflecting balancing selection), although this is rarely done given the high sampling variance unless sample sizes are extremely large.

In the early literature, a number of workers incorrectly tested for excessive divergence with a simple χ^2 test of whether allele frequencies in two samples were significantly different. As noted by Gibson et al. (1979) and Waples (1989b), this is inappropriate, as it does not account for the *drift variance* in allele frequencies,

$$\sigma^2(p_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right] \simeq p_0(1-p_0)\frac{t}{2N_e} \quad \text{for } t \ll N_e$$

where p_0 is the initial allele frequency (Equation 2.14a). Consider a population sampled at two time points (0 and t), with samples sizes n_0 and n_t (respectively). The estimated divergence is

$$\widehat{\delta}_t = \widehat{p}_t - \widehat{p}_0, \tag{9.1a}$$

which has variance

$$\sigma^2\left(\widehat{\delta}_t\right) = \sigma^2\left(\widehat{p}_t - \widehat{p}_0\right) = \sigma^2\left(\widehat{p}_t\right) + \sigma^2\left(\widehat{p}_0\right) - 2\sigma\left(\widehat{p}_t, \widehat{p}_0\right) \tag{9.1b}$$

where $\widehat{p}_i = p_i + e_i$, the true value plus an error due to finite sampling. From binomial sampling, the sampling variance of the initial frequency is given by

$$\sigma^2\left(\widehat{p}_0\right) = \frac{p_0(1-p_0)}{2n_0}, \tag{9.2a}$$

while the final allele frequency is influenced by both the drift and sampling variances (Waples 1989a,b),

$$\begin{aligned}\sigma^2\left(\widehat{p}_t\right) &= p_0(1-p_0)\left[1-\left(1-\frac{1}{2n_t}\right)\left(1-\frac{1}{2N_e}\right)^t\right] \\ &\simeq p_0(1-p_0)\left[\frac{1}{2n_t}+\frac{t}{2N_e}\left(1-\frac{1}{2n_t}\right)\right]\end{aligned} \tag{9.2b}$$

If sampling is done (without replacement) before reproduction, then $\sigma\left(\widehat{p}_t, \widehat{p}_0\right) = 0$. Substitution of Equations 9.2a, b into Equation 9.1b gives

$$\sigma^2\left(\widehat{\delta}_t\right) \simeq p_0(1-p_0)\left[\frac{1}{2n_0}+\frac{1}{2n_t}+\frac{t}{2N_e}\left(1-\frac{1}{2n_t}\right)\right] \tag{9.2c}$$

If sampling is done either after reproduction, or with replacement, this generates a covariance between the sample estimators $\widehat{p}_t$ and $\widehat{p}_0$, see Nei and Tajima (1981) and especially Waples (1989a,b) for details. Assuming this is not the case, so that $\sigma(\widehat{p}_0, \widehat{p}_t) = 0$, Equation 9.2c then gives the correct variance for the null hypothesis of random genetic drift, giving the test statistic as

$$\frac{\widehat{\delta}_t^{\,2}}{\sigma^2\left(\widehat{\delta}_t\right)}, \tag{9.2d}$$

which is approximately χ^2_1-distributed. This is the **Waples adjusted test** (Waples 1989b).

Application of this test requires an accurate estimate of p_0. While the sample estimate $\widehat{p_0}$ can be used, this can be improved upon by noting that the expected allele frequency change is zero, so that $\widehat{p_t}$ also contributes information about p_0. A simple average of the two frequencies is not appropriate, as $\widehat{p_0}$ has a smaller drift variance, and the two estimates may differ in information value due to differences in sample size n_i. Given these concerns, Schaffer et al. (1977) and Waples (1989b) propose using a generalized (i.e., weighted) least-squares (GLS) estimator (LW Chapter 8). Let $\mathbf{p} = (\widehat{p_0}, \widehat{p_t})^T$ denote the two sample estimates and denote the sampling variance-covariance matrix for these by

$$\mathbf{V} = \begin{pmatrix} \sigma^2(\widehat{p_0}) & \sigma(\widehat{p_0}, \widehat{p_t}) \\ \sigma(\widehat{p_0}, \widehat{p_t}) & \sigma^2(\widehat{p_t}) \end{pmatrix} \tag{9.3}$$

Finally, let $\mathbf{u} = (1,1)^T$ be a vector of ones. The underlying statistical model is $p_i = p_0 + e_i$, which can be written in general linear model form as $\mathbf{p} = p_0\mathbf{X} + \mathbf{e}$, where $\mathbf{V}$ is the covariance matrix for the vector $\mathbf{e}$ of residuals and $\mathbf{X} = \mathbf{u}$. Recalling LW Equation 8.34, the resulting GLS estimate of p_0 is given by $\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{p}$, which reduces to

$$\text{GLS}(\widehat{p_0}) = \frac{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{p}}{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}} \tag{9.4}$$

since both quadratic products are scalars.

Example 9.1. One of the classic papers in evolutionary biology is Fisher and Ford's (1947) study of the *medionigra* gene in the Scarlet Tiger moth *Callimoprha* (*Panazia*) *dominula*, a colorful day-flying species with a single generation per year. A single diallelic locus has a major effect on the forewing pattern. Individuals homozygous for the *dominula* allele have multiple forewing spots, while individuals homozygous for the *medionigra* allele have a darkly-suffused forewing with typically two small spots. Heterozygotes show an intermediate pattern. In 1938, Ford began studying a small colony of this species in Cothill Fen, just southwest of Oxford, England. Starting in 1941, capture-recapture data was used to estimate the census population size, with the smallest estimated size between 1941 and 1947 being 1000. In 1939 (t=0) the frequency of the *medionigra* allele was estimated (from a sample size of $n_0 = 223$) as $\widehat{p_0} = 0.092$, while by 1947 ($t = 8$), its sample frequency decreased to $\widehat{p_8} = 0.037$ ($n_8 = 1341$). Taking $N_e = 1000$ (this being the smallest estimated census value over any of the generations, and hence most favorable to supporting drift), do these data show evidence of a departure from drift?

For simplicity, assume sampling without replacement, so that $\sigma(\widehat{p_0}, \widehat{p_t}) = 0$, while the variances are given by Equations 9.2a and 9.2b. The resulting covariance matrix $\mathbf{V}$ becomes

$$\frac{\mathbf{V}}{p_0(1-p_0)} = \begin{pmatrix} \dfrac{1}{2\cdot 223} & 0 \\ 0 & \dfrac{1}{2\cdot 1341} + \dfrac{8}{2000}\left(1 - \dfrac{1}{2\cdot 1341}\right) \end{pmatrix} = \begin{pmatrix} 0.0022 & 0 \\ 0 & 0.0044 \end{pmatrix}$$

Since $\mathbf{V}^{-1}$ appears in both the numerator and the denominator of Equation 9.4, the unknown constant $p_0(1-p_0)$ cancels, allowing us to just use the above right-hand matrix for $\mathbf{V}$, giving

$$\text{GLS}(\widehat{p_0}) = \frac{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{p}}{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}} = \frac{49.496}{674.762} = 0.0734$$

Equation 9.2c gives the sampling variance for the difference in allele frequencies as

$$\sigma^2\left(\widehat{\delta_t}\right) \simeq p_0(1-p_0)\left[\frac{1}{2n_0}+\frac{1}{2n_t}+\frac{t}{2N_e}\left(1-\frac{1}{2n_t}\right)\right]$$
$$= 0.0734 \cdot 0.9266\left[\frac{1}{446}+\frac{1}{2682}+\frac{8}{2000}\left(1-\frac{1}{2682}\right)\right] = 0.0004495$$

The resulting Waples test statistic for fit to pure drift becomes

$$\frac{(0.037-0.092)^2}{0.0004495} = 6.729.$$

The probability that a χ_1^2 random variable is this big or larger is 0.0095, giving a highly significant p value and strong support against drift. By using different values of N_e in the above calculation, we can find the largest effective population size that could still allow drift to account for these data. For $N_e = 500$, the test statistic becomes 4.19 ($p = 0.040$), while for $N_e = 250$, the statistic is 2.39 ($p = 0.12$). Hence, any effective population size slightly smaller than 500 would not reject the hypothesis of drift accounting for the observed allele-frequency change.

Except for a gap between 1979 and 1987, this population has been surveyed yearly since 1939, see Jones (1989) and Cook and Jones (1996) for reviews (with Jones providing a handy table of all of the data through 1988). O'Hare (2005) used a hierarchical Bayesian analysis to examine a 60-year time series of these data. He assigned genotypes fitness drawn from a lognormal prior, allowing then to vary yearly. While selection was found to significantly contribute to the change in allele frequency, most of the variance was attributable to drift.

While one might think that tests based solely on allele-frequency change are among the most convincing, this is not the case. As Example 9.1 shows, rejection of drift can easily result from an overestimation of the effective population size. Fisher and Ford took their results as evidence against Sewall Wright's notion of the importance of genetic drift. In his reply, Wright (1948) noted that values of N_e simply based on census numbers can easily be contested by the widespread observation that the effective population size is lower, often significantly so, than the observed number of individuals in the population (Chapter 3).

As mentioned, tests of allele-frequency change suffer from low power. If selection is modest relative to $1/N_e$ or $1/n$ (with n the sample size), the sample variance can obscure any selection signal. Waples (1989b) examines some of these design issues. While we have presented this test for a single locus with just two alleles, extensions to multiple alleles are straightforward (e.g., Waples 1989b; Goldringer and Bataillon 2004). A more subtle issue is the fit of χ^2 distribution to the test statistic given by Equation 9.2d, which can be poor with rare or excessive number of alleles or a large number of generations (Goldringer and Bataillon 2004). While more sophisticated modifications (e.g., Sandoval-Castellanos 2010) can avoid some of these issues, the use of simulations for the change in neutral alleles under drift incorporating as much of the specific biology of the species as is known (e.g., Mueller et al. 1985) is strongly preferred over parametric tests.

Allele-frequency Change Over a Times Series: The Fisher-Ford Test

The test given by Equation 9.2d assumes data from just two time points, but often one has data from a series of generations. In such cases, the strong temptation to simply test the two most extreme values should be avoided, as such sampling gives a highly biased test. Rather,

specific tests have been developed to jointly consider all of the data. Indeed, the original test of Fisher and Ford involved such a temporal sequence of data. While one can use frequencies directly, Fisher and Ford used the arcsin-squareroot transform to both stabilize the variance (making it independent of the initial frequencies) and also to improve the fit to normality, especially at extreme frequencies (note that the arcsin is measured in radians, rather than degrees). Such variance-stabilizing transformations are discussed in LW Chapter 11.

Let y_t denote the transformed frequency of the allele in generation t . For t small relative to N_e, we have (approximately) that

$$y_t = 2\sin^{-1}\left(\sqrt{p_t}\right) \sim N\left[y_0, t/(2N_e)\right] \tag{9.5a}$$

where $y_0 = 2\sin^{-1}\left(\sqrt{p_0}\right)$ is the transformed value of the initial frequency. Estimates of allele frequencies are made at k time points, with no requirements about the temporal spacing of sampling. Let $\mathbf{y}$ denote the vector of the transformed estimates of the k sample allele frequencies, and let $\mathbf{u}$ denote a vector of ones of the same length,

$$\mathbf{y} = 2\begin{pmatrix} \sin^{-1}\left(\sqrt{p_1}\right) \\ \vdots \\ \sin^{-1}\left(\sqrt{p_k}\right) \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{9.5b}$$

Here, the indices denote the sequence of samples, not the actual sampled generation itself (see Example 9.2). Finally, we need the covariance matrix $\mathbf{V}$ associated with these measures. Let t be the generation number associated with the ith sample. The diagonal terms of $\mathbf{V}$ are given from Equation 9.2c,

$$V_{ii} = \frac{1}{2n_t} + \frac{t}{2N_e}\left(1 - \frac{1}{2n_t}\right) \simeq \frac{1}{2n_t} + \frac{t}{2N_e} \tag{9.5c}$$

Note that the use of this transformation makes the elements in $\mathbf{V}$ independent of the allele frequency. Now consider the covariance between samples i and j, which correspond to times t and τ, respectively (where $i > j$ and $t > \tau$). The estimates for these two sample points have a shared history (from the base value p_0) of drift up through generation τ, giving

$$V_{i,j} = V_{j,i} = \frac{\tau}{2N_e}, \quad \text{where } \tau < t \tag{9.5d}$$

Note that the covariance against the base generation ($t = 0$) is always zero. The matrix $\mathbf{V}$ contains only those rows and columns corresponding to the k specific generations sampled, so that $\mathbf{V}$ is $k \times k$.

This is now a goodness-of-fit problem for a linear model. Using Equation 9.4, we obtain a generalized least-squares estimate of the initial frequency,

$$\widehat{y_0} = \frac{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{y}}{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}} \tag{9.6a}$$

Using this value, the vector of deviations is computed by

$$\boldsymbol{\delta}_\mathbf{y} = \mathbf{y} - \widehat{y_0}\,\mathbf{u} \tag{9.6b}$$

with the test statistic, the weighted sum of the square allele-frequency differences, given by

$$\boldsymbol{\delta}_\mathbf{y}^T\mathbf{V}^{-1}\boldsymbol{\delta}_\mathbf{y} \tag{9.6c}$$

which approximately follows a χ^2_{k-1} distribution due to the normality assumption on the y_i.

Example 9.2. Let's revisit Fisher and Ford (Example 9.1), and now consider a test based on the data from 1939, 1943, and 1947,

Year	t	$\widehat{p}$	$y = 2\sin^{-1}\left(\sqrt{p}\right)$	n
1939	0	0.092	0.616	223
1943	4	0.056	0.478	269
1947	8	0.037	0.387	1341

Again, assuming $N_e = 1000$, the resulting covariance matrix $\mathbf{V}$ becomes

$$\mathbf{V} = \begin{pmatrix} V_{0,0} & V_{0,4} & V_{0,8} \\ V_{4,0} & V_{4,4} & V_{4,8} \\ V_{8,0} & V_{8,4} & V_{8,8} \end{pmatrix} = \frac{1}{2000} \begin{pmatrix} \frac{2000}{2\cdot 223} + 0 & 0 & 0 \\ 0 & \frac{2000}{2\cdot 269} + 4 & 4 \\ 0 & 4 & \frac{2000}{2\cdot 1341} + 8 \end{pmatrix}$$
$$= \frac{1}{2000} \begin{pmatrix} 4.484 & 0 & 0 \\ 0 & 7.717 & 4 \\ 0 & 4 & 8.745 \end{pmatrix}$$

In addition,

$$\mathbf{y} = \begin{pmatrix} 0.616 \\ 0.478 \\ 0.387 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \text{giving} \quad \widehat{y_0} = \frac{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{y}}{\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}} = \frac{418.851}{774.701} = 0.541.$$

Using this estimate for y_0, the vector of deviations from the initial value becomes

$$\boldsymbol{\delta}_\mathbf{y} = \mathbf{y} - 0.541 \cdot \mathbf{u} = \begin{pmatrix} 0.0757 \\ -0.0628 \\ -0.1535 \end{pmatrix},$$

giving the test statistic as $\boldsymbol{\delta}_\mathbf{y}^T\mathbf{V}^{-1}\boldsymbol{\delta}_\mathbf{y} = 7.964$, which approximately follows a χ^2_2 distribution, for a p value of 0.0186. If we used $N_e = 500$, then $\boldsymbol{\delta}_\mathbf{y}^T\mathbf{V}^{-1}\boldsymbol{\delta}_\mathbf{y} = 5.398$, or a p value of 0.067. Hence, the hypothesis that drift alone accounts for this observed pattern of change could not be rejected under this smaller value of N_e.

A number of generalizations and increasingly sophisticated tests building on the basic elements the Fisher-Ford test have been proposed by Templeton (1974), Schaffer et al. (1977), Gibson et al. (1979), Wilson (1980), Watterson (1982), Waples (1989b), De Koeyer et al. (2001), Goldringer and Bataillon (2004), Wisser et al. (2008), and Sandoval-Castellanos (2010). While the above machinery can be applied to samples from natural populations, it is often used in artificial selection experiments to detect regions of interest. In the pre-genomics era, this approach was pioneered by Stuber and Moll (1972) and Stuber et al. (1980), who looked for shifts in the frequencies of isozyme markers in lines of maize selected for yield. More recent examples for maize include Labate et al. (1999) and Coque and Gallais (2006) for yield selection, and Wisser et al. (2008) for disease resistance, while De Koeyer et al. (2001) examined yield in oats.

Schaffer's Linear Trend Test

A variation of the Fisher-Ford test was suggested by Schaffer et al. (1977), who noted that power might be improved by not just looking for a lack of fit against the model $y_t = \mu + e$ (where μ is the transformed initial allele frequency), but rather asking if a *significant linear trend* is present. If such a trend is present, i.e., if the slope term β for the model

$$y_t = \mu + \beta\, t + e \tag{9.7a}$$

is significantly different from zero (the Fisher-Ford test assumes that $\beta = 0$). Such a linear trend is not expected under drift, but would be under directional selection, assuming the direction of selection is not changing. In general linear model form, Equation 9.7a becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_k \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{9.7b}$$

For the data in Example 9.2, the resulting $\mathbf{X}$ matrix and the GLS estimate $\widehat{\boldsymbol{\beta}}$ of the vector of parameters becomes

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.609 \\ -0.028 \end{pmatrix}$$

Applying LW Equation 8.35, the standard error on the slope is found to be 0.0086, showing that it is highly significant. Stuber et al. (1980) used this approach to suggest selection at sites linked to several allozyme markers in a series of selected maize lines.

Scans and Simulation-Based Approaches

As presented, the above tests for shifts in allele frequencies are performed one marker at a time, as they herald from the days of testing just one or a few unlinked candidate genes. With a few unlinked markers, Bonferroni corrections (or the slightly more powerful sequential methods, Appendix 4) can be applied to assign overall significance levels. However, in the genomic scan era, with the potential for thousands of linked markers on each chromosome, tests are no longer independent. Even if they were, their vast number makes Bonferroni-type corrections untenable for a test to have any power. How are these tests extended to the dense marker maps used in genomic scans?

For starters, analyzing markers one at a time is rather inefficient, in that one potentially loses shared information from linked markers. A better approach is to compute the average allele frequency change within a small sliding window. Significance can be assessed using simulation models specific to the recombination pattern for the marker map being used in the experiment to generate the distribution of changes across windows under the null hypothesis of drift. Genome-wide significant levels for tests can be computed from the simulated null distribution for any particular window (e.g., Johansson et al. 2010). Other less formal approaches can also be used to indicate regions of interest (as opposed to regions formally statistically supported). See Figure 9.1 and Example 9.3 (below) for examples of this in the context of divergence between two differentially-selected populations.

Birth Data Selection Mapping, BDSM

A very interesting approach for a genomic scan for sites under selection when one has extensive pedigree data (such as in cows) was proposed by Decker et al. (2012). With extensive pedigrees, one has information on the date of birth (DOB) of most individuals, whose value can be expressed in fractions of years since the start of the pedigree. Individuals with a later

date of birth have likely experienced more selection than earlier-born individuals. As such, one expects a positive relationship between a marker linked to a selected site and DOB, for example if A was initially rare, but favored, AA individuals are far more common in animals with a higher DOB. Decker et al. turned this relationship around, realizing that by treating DOB as a quantitative trait, one could use association mapping (LW Chapter 16), with sites showing an association with DOB being under selection. This approach is called **Birth Data Selection Mapping** (or **BDSM**), and because association mapping is done in a mixed-model framework (Chapter 19), it accounts for biases introduced by family structure. The authors applied this approach to US Angus cattle born over a 50-year period (roughly 10 generations). A standard random-effects mixed model of association mapping detected 11 loci significantly associated with DOB, while a Bayesian model found around two percent of the SNPs were strongly associated with DOB. The former model assumes an infinitesimal structure, while the latter allows for genes of larger effect embedded in a sea of smaller-effect genes. While BDSM requires large, deep pedigrees with DNA data, it is intriguing and powerful, and should be explored more in natural populations.

ALLELE-FREQUENCY DIVERGENCE BETWEEN POPULATIONS

The within-population change allele-frequency due to drift just considered directly translates into divergence in allele frequencies among replicate populations. One natural measure of this divergence at a biallelic locus compared between two population is the squared allele-frequency difference. As this statistic was also used in tests developed the previous section, these are easily modified to test for excessive divergence. When comparisons involve either more that two populations and/or markers with more than two alleles, a more natural measure of divergence is given by Wright's F_{ST} statistic of population structure (Wright 1951). Recall from Chapter 2 that this statistic measures the fraction of total variation of a set of populations that is due to among-population differences in allele frequencies, and is easily extended to multiple alleles and multiple populations. For a biallelic locus, $F_{ST} = \sigma_B^2(p)/[\,\overline{p}(1-\overline{p}\,)\,]$ where $\sigma_B^2(p)$ is the variance in allele frequency p over the populations and $\overline{p}$ its average value.

Our treatment first considers simple modifications of excess allele-frequency change tests from the last section for detecting excessive divergence at a candidate locus. We then consider scan-like procedures using a sliding window of linked markers to detect excessive divergence over a region. Finally, we examine formal F_{ST}-based tests and their potential limitations. It is worth emphasizing that tests considered in this section apply over ecological time scales — they assume the marker is generally expected to be segregating in both populations (if only drift is operating), which puts the time scale for these at no more than $\sim N_e$ generations. New mutations are also ignored. Chapter 10 examines divergence-based tests built on detecting excess *substitutions* at a site, examining the cumulative effect of multiple fixations over evolutionary time scales.

Modifications of Within-population Divergence Tests

Assume a biallelic locus compared over two populations. Whether an observed squared allele-frequency difference is too large, or small, relative to drift can be tested using a simple modification of Waples test (Equation 9.2d). The denominator in Equation 9.2d is replaced by $\sigma^2\left(\widehat{p}_{t,1}\right) + \sigma^2\left(\widehat{p}_{t,2}\right)$, the sampling variances for the two populations (Equation 9.2b). This requires estimates of the divergence time t as well as the average effective size for both populations, but this may not be unreasonable when comparing domesticated lines, or artifically-selected populations, with known histories.

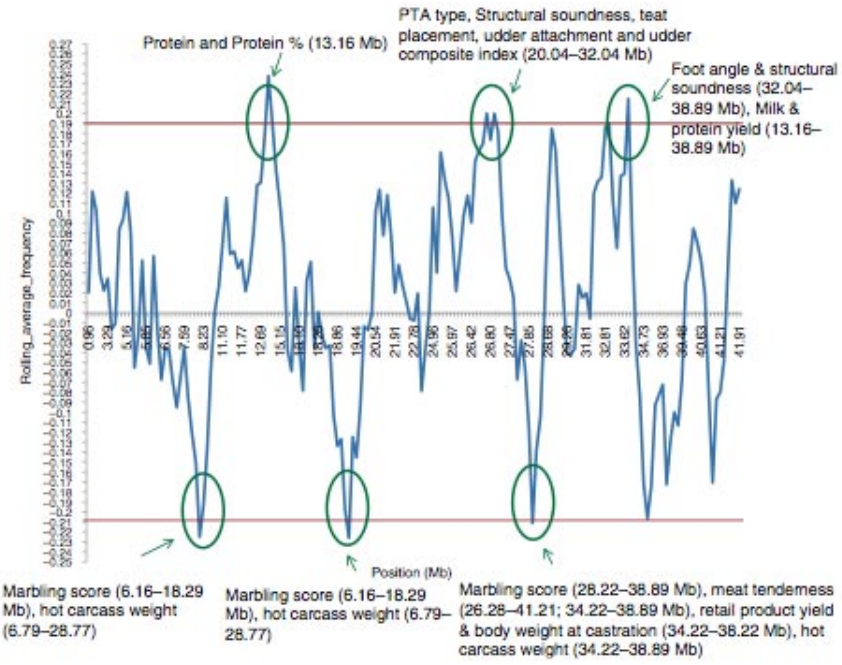


Figure 9.1. A scan of *Bos taurus* chromosome 19, contrasting differences in SNP allele frequencies between specialized dairy (Holstein) and meat (Angus) breeds. Differences were computed based on a sliding window of five adjacent markers, using a set of 175 SNPs on chromosome 19. The horizontal axis represents chromosomal position, the vertical axis is the average between-breed difference in SNP allele frequencies over the five-SNP window. The upper and low lines indicate the five percent significance levels as accessed via a permutation test. As shown in the figure, the authors were able to correlate significant peaks and valleys with known QTLs for dairy and beef traits. Since QTL intervals tend to be rather vague (averaging around 20 megabases, or roughly 20 cM, for these traits), the significance of these correlations with known QTLs, while suggestive, is unclear. After Prasad et al. (2008).

Modifications of genomic scans using a sliding window of makers has been applied to divergence data, in particular to contrast allele frequencies between meat and dairy cattle. Figure 9.1 shows a scan for excess SNP allele frequency differences in a sliding window analysis over *Bos taurus* chromosome 19. Regions of interest were accessed by a permutation test where breed labels were randomized over SNPs to generate the null. Threshold values indicate the 95% limit for the range of maximal window score (over the entire chromosome) in the randomized data sets. This approach is only approximate, as it ignores LD among neutral markers (and thus adjacent markers are correlated). However, it still serves to indicate sites likely enriched for differentially-selected genes. Example 9.3 discusses an independent analysis of the same question, where direct simulations were used to access significance.

Example 9.3. Hayes et al. (2008) compared roughly 400 Australia Holstein (dairy) and Angus (beef) cattle. Roughly 7600 SNPs spanning the whole genome were scored using a sliding window of ten adjacent SNPs, resulting in 7032 sliding 10-SNP windows (averages were not computed within ten SNPs from a chromosome end). To assess the significance of any

particular window, the authors performed a detailed simulation, starting with a common base population of $N_e = 1000$ simulated for 900 generations. This was then subdivided into two populations with $N_e = 125$ and simulated for an addition 100 generations, to reflect breed formation. These N_e values reflect the initial domestication followed by breeding for more specialized objectives. Throughout, roughly 300 markers per chromosome were simulated, with F_{ST} values computed for the final divergence between the two split populations. Only those simulations whose F_{ST} matched the actual data set (0.08) were kept. The uppermost p of excessive divergence (positive or negative) was computed by averaging over simulations. For $p = 0.001$, this gave 15 significant tests (windows), while for $p = 0.005$, 84 tests were declared significant. To assess what fraction of these might be false positives, the authors computed the false discovery rate (FDR), the fraction of those tests declared to be significant that are actually false positives (Appendix 4). FDR provides a measure of how enriched the data set is for true positives. At a $p = 0.001$ level of significance, one expects $7032 \cdot 0.001 = 7.03$ tests to be significant by chance alone, while for $p = 0.005$, this increases to 35 false positives among the 7032 tests for each window. The FD rates are 7/15 = 47% for tests of $p = 0.001$, and 35/84 = 42% for tests at the $p = 0.005$ significance level. Hence, the expectation is that roughly half of the significantly different windows are indeed true positives.

Grossman et al.'s ΔDAF Test

A simple test growing in popularity (especially in the animal domestication literature) is the ΔDAF statistic of Grossman et al. (2010). This test is a natural outgrowth of the type of comparisons see in Figure 9.1, focusing on the difference in the derived allele frequency (DAF) between a control and selected population. For a candidate SNP, let $\overline{D}_{NS}$ denote the average frequency of the derived allele in a control population(s) and D_S its frequency in the putatively-selected population, with $\Delta DAF = D_S - \overline{D}_{NS}$. This statistic ranges between plus and minus one, and standard outlier approaches are used to highlight SNPs with excessive values.

F_{ST}-based Divergence Tests: Basic Approaches

The underlying premise for most F_{ST}-based tests of selection is the suggestion by Cavalli-Sforza (1966) that all neutral loci have the same expected value of F_{ST}, reflecting the genome-wide impact of common demographic and population structure forces. Thus, one can (in theory) use a large number of marker loci to estimate this baseline F_{ST} value for the set of populations being compared, and then look for outliers. Loci with excessively high values indicate more divergence than expected under drift, and the possibility that the marker is linked to a site under directional selection. This approach is easily modified to look for specific loci being outliers in specific populations (e.g., Akey et al. 2010). Likewise, excessively low values indicate less divergence than expected under drift, and hence the potential for a site under balancing selection near the marker. Caution is in order, however, as Cavalli-Sforza's premise of equal effects of demography over all loci does not necessarily hold for populations undergoing range expansions. As discussed in the introduction, new alleles arising on the leading wave of expansion can 'surf' to high frequencies, generate excessive values over the expected background.

When migration and new mutation can be ignored, F_{ST} provides an estimate of the divergence time T (in N_e generation) between two populations. Rearranging Equation 2.43, taking logs of both sides, and recalling (for $|x| \ll 1$) that $\ln(1-x) \simeq -x$ yields

$$\ln(1 - F_{ST}) = t \ln\left(1 - \frac{1}{2N_e}\right) \simeq -t/2N_e$$

Hence $T = -\log(1 - F_{ST}) \simeq t/2N_e$. Thus, one can recast an excessive F_{ST} value into an

excessive divergence time required for drift to account for the observed divergence. These estimated times are called **branch lengths**, and (following the Cavalli-Sforza premise) should have the same expected value over all neutral genes. An excess branch length for a candidate gene relative to some reference set of genes suggests excessive change relative to drift (Vitalis et al. 2001; Rockman et al. 2003; Yi et al. 2010), see Figure 9.2.

Finally, an important caveat is that analysis of the F_{ST} statistic assumes the infinite alleles model, with new mutations sufficiently rare to be safely ignored. This is not the case for microsatellite (STR) markers, which require specific divergence metrics, such as R_{ST} (Slatkin 1995a; Goodman 1997). Excoffier et al. (2009) showed that using F_{ST} in place of R_{ST} for the analysis of STR data can significantly inflate the false-positive rate.

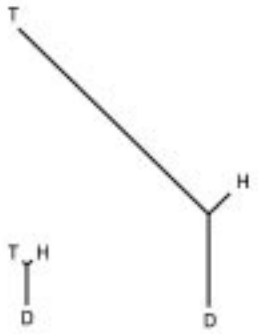


Figure 9.2. F_{ST}-based branch lengths for Tibetan (T), Han (H), and Danish (D) populations. On the left are lengths based on the average F_{ST} values for all sampled markers. The tree on the right is for the *EPAS1* gene, which shows a dramatic increase in divergence times in the Tibetan population for this gene, consistent with excessive allelic divergence due to selection or perhaps other features, such as allelic-surfing. After Yi et al. (2010).

Example 9.4. A very nice use of F_{ST} to provide evidence for selection was the *a priori* prediction by Taylor et al. (1995) of the behavior of alleles at a putative site of selection. The tobacco budworm (*Heliothis virescens*) is a noctuid moth and a major cotton pest in the US. Pyrethroid insecticides have been used in control efforts, and these act on voltage-gated sodium channels in the nervous system. Because of geographical differences in the historical application of pyrethroids, the authors predicted that F_{ST} values at the sodium channel *Hpy* gene should be significantly higher than for background loci, reflecting differential selection. Samples of adults from five widely spaced locations in the US (two in the SW, two in the SE) found an F_{ST} value of 0.041 ± 0.005 at the *Hpy* marker, in contrast to values of 0.002 ± 0.001 at 14 other loci.

The Lewontin-Krakauer Test and its Extensions

The above outlier methods (for either F_{ST} or branch lengths) in the absence of simulations to access significant are rather ad-hoc. They are best viewed as *enrichment* methods, distilling a reduced set of makers that is likely enriched for any selected sites. More formal statistical tests start with Lewontin and Krakauer (1973), who showed that the expected large-sample value of F_{ST} for a random locus sampled over n populations approximately follows a $\lambda\chi^2_{n-1}$ distribution, where the scaling factor $\lambda = E(F_{ST})/(n-1)$. Given Cavalli-Sforza's assump-

tion that, on average, population structure influences all neutral loci equally, Lewontin and Krakauer estimated $E(F_{ST})$ from the average $\overline{F}_{ST}$ over all scored loci, giving

$$\frac{1}{\lambda}F_{ST} = \frac{(n-1)F_{ST}}{\overline{F}_{ST}} \sim \chi^2_{n-1} \tag{9.8a}$$

This is a large-sample approximation, as the sampling error in estimating the true realization of the F_{ST} value for a given marker is ignored. Since the variance of a χ^2_n random variable is $2n$ (LW Equation A5.15b), the sampling variance is approximately

$$\sigma^2\left(F_{ST}\right) \simeq \lambda^2 2(n-1) = \left(\frac{E(F_{ST})}{n-1}\right)^2 2(n-1) \simeq 2\,\frac{\left(\overline{F}_{ST}\right)^2}{n-1} \tag{9.8b}$$

A potential problem with using $\overline{F}_{ST}$ to estimate λ is that it can be skewed by a few excessive values, inflating the control factor. A more robust estimate can obtained by ordering the F_{ST} estimates and equating their median value with the corresponding 50% value for a χ^2_{n-1} (Devlin and Roeder 1999). Suppose the median value for F_{ST} among our collection of loci sampled over five populations is 0.168. The medium value of a χ^2_4 is 3.357, giving $0.168 = \lambda \cdot 3.357$ or $\lambda = 0.05$.

Lewontin and Krakauer proposed that either Equation 9.8a or 9.8b could be used to detect excessively large (or excessively-small) F_{ST} values among a collection of loci, but this approach depends on the validity of the χ^2 approximation for the distribution of F_{ST} values. This approximation fails when too many alleles (> 5) are present at a locus, or there are rare minor alleles (minor allele frequency < 0.1), or if the divergence time is too large (Goldringer and Bataillon 2004). Further, when migration occurs among the sampled populations, correlations in gene frequencies across populations are generated by the migration structure, and this further impacts the χ^2 assumption (Nei and Maruyama 1975; Robertson 1975a,b; Tsakas and Krimbas 1976). As a result of these concerns (and others, see Nicholas and Robertson 1976), the Lewontin-Krakauer quickly languished. However, its basic simplicity, coupled with requiring only the type of data routinely gathered by ecological geneticists (estimates of locus-specific F_{ST} values), fueled the search for ways to correct these initial flaws.

One proposed correction is again an outlier approach: replace the parametric χ^2 assumption with an empirical distribution of F_{ST} values from the loci being scanned, with outliers representing a set enriched for loci under selection (e.g., Akey et al. 2002; Kayser et al. 2003; Akey et al. 2010). Beaumont and Nichols (1996) suggested that a plot of F_{ST} versus heterozygosity offered a more robust approach for the detection of outliers. Their logic is that the sampling variance in F_{ST} becomes more sensitive when allele frequencies are skewed over populations, so that sites with low-frequency alleles can generate an excess of extreme values simply by chance.

A second strategy to resuscitate the Lewontin-Krakauer test is to use knowledge of the population structure (when it exists) as the basis for simulations of the distribution of F_{ST} under the null of no selection. This was first proposed by Bowock et al. (1991), who had a rough idea of the structure of the five human populations they surveyed. Vigouroux et al. (2002) used coalescent simulations (Chapter 2) incorporating a founding bottleneck in a screen of 501 maize genes for signatures of selection to find which showed F_{ST} values that (conditioned on the number of segregating alleles) were significant. Similarly, Ross-Ibarra et al. (2008) estimated the parameters for a complex demographic model for the population structure of *Arabidopsis lyrata*, and then used simulations based on their estimated demography to detect outliers. The concern with any null distribution of F_{ST} values generated by simulations is robustness to assumptions about the population structure, as even the most careful simulations can be misleading (e.g., Curret et al. 2006; Excoffier et al. 2009). For

example, most analyses of robustness to different demographic models fail to consider the effects of spatial expansion, and hence miss concerns raised by allelic surfing.

Finally, Bayesian approaches that partly account for the correlation among allele frequencies have been proposed by Beaumont and Balding (2004), Riebler et al. (2008), Foll and Gaggiotti (2008), and Bazin et al. (2010). The bottom line for F_{ST} methods is that they continue to remain popular and have strong supporters (e.g., Beaumont 2005; Novembre and Di Rienzo 2009) as well detractors (Hermisson 2009). There is no question they can be a very useful tool for finding potential *regions of interest*, but great care should be exercised in making anything other than cautious statements about the *statistical significance* of such regions.

CHANGES IN THE CHROMSOMAL SPATIAL PATTERN OF NEUTRAL VARIATION

The classic signature of a recent hard sweep is a chromosomal region of depressed variation, while a site under long-term balancing selection displays enhanced variation (Chapter 8). This section develops methods based on the expected spatial pattern of variation around a selected site following a hard sweep. We start with a gentle introduction to simple graphical methods for suggesting interesting regions, and subsequently move onto Maximum Likelihood (LW Appendix 4) models. These become increasingly more complex as we attempt to capture explicitly the spatial pattern of variation through a chromosomal region about a putative sweep. For the casual reader, these ML-based spatial models can be rough going, but the take-home message is that using all of this spatial information can significant improve our ability to detect a recent hard sweep, as well as provide estimates of the required strength of selection.

Simple Visual Scans for Changes in Nucleotide and STR Diversity

The most basic approach is a simple plot of variation as a function of genomic location, looking for either peaks (long-term balancing selection) or valleys (a recent sweep), see Figures 8.1 and 8.2. With SNP data, variation is typically scored as average nucleotide diversity π (Chapter 4) within a sliding window to smooth out the inherent noisiness from individual sites. With STR (simple tandem repeat or microsatellite) data, several different metrics of variation are available, e.g., copy number variance, number of alleles, probability of heterozygosity. With their large number of alleles/marker and high mutation rates, STRs provide a more consistent signal, and are usually plotted on a per-marker basis (as opposed to a sliding-window analysis); see Figure 9.3 and Example 9.5. A point of caution is that mutation rates at STRs can be length dependent, with smaller arrays expected to show less variation.

Example 9.5. Domesticated breeds of dogs are ideal candidates for detection of regions under a selective sweep. Most breeds are rather recent (roughly in the last 100 years, or 200 generations) and often involve large phenotypic effects (and hence the potential of strong selection on just a few genes). Simulation studies by Pollinger et al. (2005) found that a few moderately linked highly polymorphic loci can give a strong sweep signal under realistic conditions for the formation of dog breeds. They tested their idea using a scan of microsatellites around candidate genes in two different breeds. The Large Munsterlanders are a recent breed (originating around 1910) and categorized by a black coat color, with the pigment gene *TYRP1* being suggested as a candidate for this trait. As Figure 9.3A shows, there is a roughly 50 Mb region of depressed microsatellite variation around *TYRP1* relative to control and brown-coated populations. Note that the region under the sweep is rather large, and if indeed *TYRP1*

was the actual target, this region is asymmetric, showing more reduced variation to the right of *TYRP1* than at the locus itself.

A more striking example is offered by Dachshunds (Figure 9.3B). These show no variation at three microsatellites surrounding the *FGFR3* candidate gene that is involved in achondroplasia (limb-shortening). Both of these breeds went through strong bottlenecks during their formation, so perhaps sampling noise overlaid on the general reduction in variation from the bottleneck generated these depressions. To test for this, the authors simulated the founding process. Assuming highly polymorphic loci (heterozygosity in excess of 30%), the authors' simulations showed a less than 5% chance of finding three adjacent loci with no variation under a simple model of a genome-wide population bottleneck during founding (conditioned on the average levels of heterozygosity seen throughout the genome).

These dog examples show the power (strong signal with only modestly-dense markers) but also the pitfalls (poor localization) when a strong sweep occurs. Since the sweep depresses variation throughout a region, additional sweep-based fine-mapping within this region would be futile. Ironically, localization is easier under a weak sweep than a strong one, as the region of depressed variation is smaller in the latter case. However, since the same traits often appear in independent dog breeds, if the same gene(s) are the targets of selection in each, there is the potential for improved resolution by looking for overlapping intervals in the detected sweep regions.

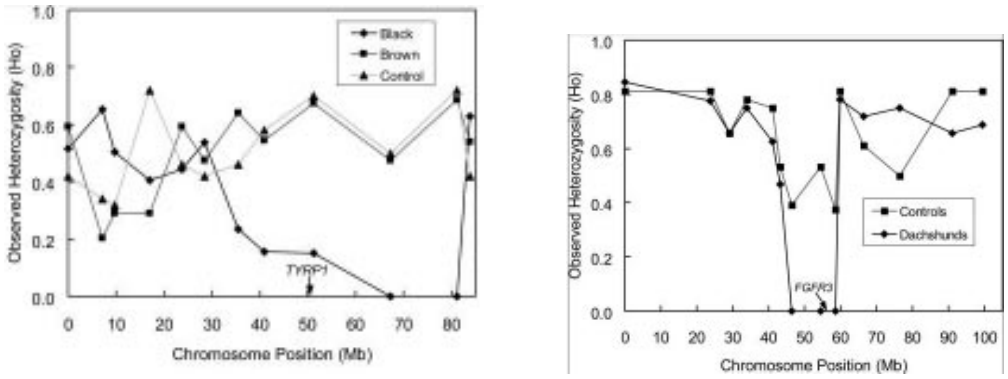


Figure 9.3. Using microsatellites in the search for dog domestication genes. **A** (Left)**:** Large Munsterlanders have a black coat, and hence the pigment gene *TYRP1* on chromosome 11 is a candidate. A plot of variation for this breed (Black) relative to control and brown individuals for the pool that founded Munsterlanders shows depressed variation spanning this gene. **B** (Right)**:** Dachshunds are characterized by shortened limbs, suggesting the *FGFR3* gene on chromosome 3 as a candidate. Dachshunds have an absence of variation at three microsatellites spanning this gene, while variation is present in controls. After Pollinger et al. (2005).

While such plots of the spatial structure of genome variation are visually appealing, they are not formal tests of selection and can indeed be rather misleading. A change in the background level of variation can arise for several reasons besides selection, such as variation in the local mutation rate. This is especially true for STR markers, whose mutation rates are expected to vary considerably, reflecting differences in the composition (size and sequence) of their repeating units. Second, the inherent stochasticity of recombination and drift can generate considerable variation in the coalescent process across a genome, so that even with a sliding window analysis for a set of markers with the same mutation rate, strong peaks

and valleys are routinely found in neutral simulations (Kim and Stephan 2002; Jensen at al. 2005).

Two different strategies can address these complications. The first is to use information from the same (and other) regions in several populations. If a region has a low (or high) mutation rate relative to the rest of the genome, presumably this will also be true in other populations or closely related species. One interesting application of this approach found a large reduction in STR variation a region around the *pfcrt* gene implicated in Chloroqunie resistance in the malaria parasite *Plasmodium falciparum* (Wootton et al. 2002). This reduction was seen in resistant lines from both South America and Asia/Africa, but was absent in sensitive lines from Asia/Africa.

However, comparing *just* the region of interest in two populations is uninformative, as differences in N_e (for example, one has just passed through a bottleneck, as might occur in the domestication process) can generate selection-like signals in one. Ideally, one would compare k different regions in $n \geq 2$ populations. As detailed below, this can be done with a simple ANOVA approach or through more formal maximum-likelihood (ML) based **bottleneck models**, which test whether the variation at some loci is consistent with a more extreme bottleneck than experienced by the rest of the genome. Second, in the absence of information from other populations, **explicit spatial models** can be applied. These use ML to compare the genomic spatial structure of variation as a function of recombination fraction over a region of interest, testing whether the fit is consistent with the pattern expected from a sweep (e.g., Equation 8.8a). This approach is both the most powerful (when modeling assumptions hold) and potentially the most fragile when they do not, and we will address some of these concerns and their potential corrections.

Tests Based on STR Variation Across Populations

Schlötterer and colleagues (Schlötterer et al. 1997, 2002; Harr et al. 2002; Kauer et al. 2003; Schöfl and Schlötterer 2004) proposed several straightforward tests using information based on K unlinked genes sampled from L populations. While their tests were applied to STR data, this same approach can be applied to SNP data in sliding windows. ANOVA is used to obtain site- and population-specific values to correct (respectively) for variation in mutation rates over loci and variation in N_e over populations, and then test if a specific locus-population combination is unusual.

Consider a set of STRs, and let V_{ij} denote the variance in repeat copy number at locus i in population j, with $v_{ij} = \ln(V_{ij})$. The effects of locus-specific mutation rates are accounted for by averaging locus i over all populations ($v_{i\cdot}$), while population-specific effects are accommodated by using the average $v_{\cdot k}$ of population k over all loci. Under the assumption of no locus-by-population effects, the expected log-variance can be written as a simple ANOVA model,

$$\widehat{v}_{ij} = v_{i\cdot} + v_{\cdot j} - v_{\cdot\cdot} + e_{ij} \tag{9.14}$$

A sweep in population j near the i-th STR is indicated by a significant deficiency from the predicted value $\widehat{v}_{ij}$, which can be tested using a t-like statistic (see Schlötterer et al. 1997 for details).

A somewhat related approach, the **Log RV** statistic, was proposed by Schlötterer (2002). The locus-specific mutation rate cancels in the ratio $RV = V_{ij}/V_{ik}$ for the same locus (i) in two different populations (j and k), leaving only the ratio of effective population sizes. (The **Log RH** statistic of Kauer et al. (2003) uses the ratio of heterozygosities, as opposed to copy number variance.) Simulation studies showed that the log of these ratios is very close to a normal under many demographic situations (although it does appear sensitive to extreme bottlenecks). While similar in spirit to Equation 9.14, the Log RV and Log RH statistics are outlier approaches, computing all pair-wise values and using outliers as potential sites of

selection. Modifications have been proposed for two linked STRs (Harr et al. 2002) and for a follow-up statistic for any particular STR that shows signs of selection (Wiehe et al. 2007).

Tests of Sweeps Using Bottleneck Models

In domesticated species, might can imagine a founding bottleneck that reduces variation across all loci relative to the ancestral source population. However, *in addition* to this common bottleneck, a *further* reduction is associated with genes selected during domestication (assuming these generate hard sweeps). This further reduction can be thought of as an additional bottleneck beyond the common founding bottleneck. This idea leads to a more formal maximum-likelihood (ML) based test. Data for multiple loci from two (or more) populations are first used to estimate a common bottleneck value for loci in one population (relative to another). One then tests whether model fit is improved by allowing a subset of these loci to experience an additional bottleneck (as would happen with a sweep).

This approach was first considered by Galtier et al. (2000). They assumed a population in mutation-drift equilibrium, where $\theta_i = 4N_e\mu_i$ measures of variation at the i-th locus (Chapter 2). At some time T in the past (scaled in terms of $2N_e$ generations), a bottleneck occurred, and after some passage of time in the bottleneck, the population quickly expands back to its previous size. The authors made the clever observation that while a bottleneck is a reduction in population size, the net result is that the number of distinct lineages going into a bottleneck is far greater than the number that survive the bottleneck. Motivated by this observation, they introduced a strength measure S for the bottleneck, which is the expected amount of time required to lose the same number of lineages in a model of constant population size. They then assumed for some time 0 until T, a standard neutral coalescent process is occurring, with mutations and coalescent events. From time $T < t < T + S$, only coalescences are allowed (mutation is effectively turned off), while from time $t > T + S$ mutation is turned back on again. Given a population sample of segregating sites over k loci, the method of Griffiths and Tavare (1994a,b) can be used to obtain maximum likelihood estimates of T, S, and θ_i for $1 \leq i \leq k$. They then construct a model where the bottleneck is potentially different for each locus, in which case T and S are now locus-specific, and one estimates T_i, S_i, and θ_i for $1 \leq i \leq k$. A standard likelihood ratio test (LW Appendix 4) determines if the second model provides a better fit.

Example 9.6. Wright et al. (2005) used a multiple bottleneck model in their search for genes under selection in maize. The authors used SNP data on 774 genes from 14 maize and 16 teosinte inbred lines. Collectively, the sampled maize lines contained roughly 60% of the variation as the teonsite lines, showing a strong bottleneck signal across the entire maize genome. The authors quantified the strength of the bottleneck by $b = N_b/d$, the ratio of the size of the population in the bottleneck divided by its duration. Using just the teosinte data, θ_i was obtained using Watterson's estimator (Equation 4.3a), and these were used in a likelihood function $L(b|S)$ to estimate b based on the observed number of segregating sites S for each locus in maize (this is just the probabilty of the data S given b, $p(S|b)$, see LW Appendix 4). For the base model with a single b over all loci, the total likelihood (assuming loose linkage between loci) is just the product of $L(b|S_i)$ over the sampled loci. The resulting maximum likelihood estimate (MLE) was $\widehat{b} = 2.45$.

The authors then fit a second model that assumed two classes of loci exist in the populations: a fraction $(1-q)$ experiencing a bottleneck of strength b_1 and a fraction q experiencing a much stronger bottleneck of strength $b_2 < b_1$, giving the resulting likelihood for locus i as

$$L(f, b_1, b_2|S_i) = (1-q)L(b_1|S_i) + q\,L(b_2|S_i) \tag{9.15a}$$

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

$$\Pr(b_2 \mid S_i) = \frac{q\, P(S_i \mid b_2)}{(1-q)\, P(S_i \mid b_1) + q\, P(S_i \mid b_2)} = \frac{q\, L(b_2 \mid S_i)}{(1-q)\, L(b_1 \mid S_i) + q\, L(b_2 \mid S_i)} \tag{9.15b}$$

This gives a posterior probability for a particular locus (here i) being in the strong bottleneck class (b_2). This same approach for posterior prediction reappears in Example 10.15, where we are looking for actual sites in a protein under selection given divergence data.

Tests of Sweeps Using Genomic Spatial Information: CLRT-GOF

While bottleneck approaches are rather elegant, they ignore any information from the expected spatial pattern of variation — an expected decrease as one moves closer to the site of selection. Further, they require samples from two (or more) populations, which might be impractical in some cases. Kim and Stephan (2002) proposed a likelihood-based test to use spatial pattern of variation from a single population sample to not only detect a sweep, but to also localize its position and estimate the strength of selection.

The basic structure of their test (and several extensions) is as follows. Suppose m linked segregating sites from a local chromosomal region of interest are scored in n samples. Using an outgroup, we can polarize any segregating alleles, determining which are derived. The resulting data are the number k_i of derived alleles at site i where $1 \le k_i \le n-1$. Building on Equation 2.36c, the probability of observing k_i given the sample size n and the vector $\boldsymbol{\Theta}$ of model parameters is just

$$\Pr(k_i \mid n, \boldsymbol{\Theta}) = \binom{n}{k_i} \int_{1/(2N)}^{1-1/(2N)} x^{k_i} (1-x)^{n-k_i} \phi_i(x \mid \boldsymbol{\Theta})\, dx \tag{9.16a}$$

where $\phi_i(x \mid \boldsymbol{\Theta})$ is the frequency spectrum for site i under the model ($\boldsymbol{\Theta}$) of interest. Equation 9.16a is also the likelihood for $\boldsymbol{\Theta}$ given the data, $L(\boldsymbol{\Theta} \mid k_i, n)$. Under the equilibrium neutral model, the Watterson distribution (Equation 2.34a) is used for ϕ_i, where $\boldsymbol{\Theta} = \theta_i$, the scaled mutation rate for site i. If one starts with a Watterson distribution and then has a sweep at a linked site, ϕ_i is now given by Equation 8.13, with parameters $\boldsymbol{\Theta} = (\theta_i, f_{s,i})$, where $f_{s,i}$ measures the strength of the sweep at site i. From Table 8.1,

$$f_{s,i} = (4N_e s)^{-c_i/(2hs)} = e^{-c_i\, \zeta} \tag{9.16b}$$

with $\zeta = (1/2hs)\ln(4N_e s)$ and c_i the recombination fraction between site i and the site of the sweep. Usually one assumes that $\theta_i = \theta$ is the same for all sites, in which case under the null model all sites follow the same Watterson distribution (no expected spatial pattern), while under the alternative of a sweep, the spatial location of a site relative to the assumed sweep site changes its frequency spectrum in a very defined manner.

Since Equation 9.16a gives a proper likelihood for each site, one might imagine that the total likelihood $L(\boldsymbol{\Theta} \mid \mathbf{k}, n)$ given the vector $\mathbf{k}$ of the k_i values for each site is just the product of the site likelihoods,

$$L(\boldsymbol{\Theta} \mid \mathbf{k}, n) = \prod_{i=1}^{m} L(\boldsymbol{\Theta} \mid k_i, n) \tag{9.16c}$$

and the MLE for $\boldsymbol{\Theta}$ is given by the value that maximizes $L(\boldsymbol{\Theta} \mid \mathbf{k}, n)$. However, this product is *not* a proper likelihood, as adjacent sites are correlated (due to shared history). The resulting product of the site likelihoods is thus a **composite likelihood** (also called a **quasi-maximum likelihood** or **pseudo-likelihood**), an approximation of the total likelihood. Kim and Stephan contrast the maximum value of this composite likelihood under the Watterson distribution (using Equation 2.35a for ϕ) with the maximum obtained under a sweep model (using Equation 8.13 for ϕ). Since f_s changes with distance c from any particular site (Equation 9.16b), the location giving maximum value of f_s corresponds to the estimated position of the selected site. They compare the ratio of the two composite likelihoods, $L_{sweep}/L_{neutral}$, corresponding to

$$\Lambda_{CLR} = \frac{\max L(\theta, f_s(\zeta, c_i) \mid \mathbf{k})}{\max L(\theta, f_s = 0 \mid \mathbf{k})} \tag{9.16d}$$

They call this approach the **CLR** test (or **CLRT**), for **composite likelihood ratio test**. Since this is *not* a strict likelihood ratio, large-sample approximations (LW Appendix 4) for its distribution are *not* valid, and the critical values must be obtained by simulation. Boitard et al. (2009; also see Kern and Haussler 2010) propose the use of hidden markov models to account for the correlations among markers that are ignored under a composite-likelihood framework.

Jensen et al. (2005) found that the CLR test was *not* robust to population structure nor recent bottlenecks. While neither of these demographic processes is expected to generate the same spatial pattern as seen under a sweep, they *can* generate an excess of sites segregating rare derived alleles, so that some features of the frequency spectrum can be similar to a sweep, and this in turn could improve the fit relative to the equilibrium Watterson distribution. To distinguish sweeps from false signals generated by demography and/or population structure, they proposed that any significant CLR result then be subjected to a goodness-of-fit (**GOF**) test to see how well it fits a sweep model. They do so by comparing the maximum of the composite likelihood function under the sweep model against the maximum of the likelihood under a more general model where the population frequency p_i (of a derived allele) is unique for each site. Under this alternative, the likelihood function for site i is

$$L_A(k_i) = \binom{n}{k_i} p_i^{k_i} (1 - p_i)^{n - k_i} \tag{9.17a}$$

Again, the total composite likelihood is the just product of Equation 9.17a over all sites. Assuming each site is independent, the MLE for p_i is just $\widehat{p_i} = k_i/n$ (LW Appendix 4), giving the maximum of the composite likelihood under the alternative as

$$\max\left(CL_A\right) = \prod_{k=1}^{n-1} \left[\binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} \right]^{n_k} \tag{9.17b}$$

where n_k is the number of sites in the sample with k of n sampled alleles being derived. The GOF test is the ratio

$$\Lambda_{GOF} = \ln\left(\frac{\max\left(CL_A\right)}{\max\left(CL_0\right)}\right) \tag{9.17c}$$

where max (CL_0) is the value of the sweep composite likelihood function at the MLEs for the sweep parameters. Again, since these are not true likelihoods, large-sample distribution theory can not be used to assess significant. Rather, Jensen et al. used the MLEs for the sweep parameters to generate a large number of data sets under the null (the sweep model), using these to compute max (CL_A) and hence a distribution of Λ_{GOF} under the null. Support for a sweep is indicated when (1) the CLR test gives a significant result and (2) the GOF test is *not* significant. If the latter is significant, there is not support for a sweep, rather demographic features are likely the cause of departures from the neutral equilibrium model. Jensen et al. found that this approach was much more robust to population structure and demography, giving false positives only when very serve bottlenecks have occurred. This improved control over the false positive rate comes at the cost of decreased power (Jensen et al. 2006; Boitard et al. 2009).

Kim and Nielsen (2004) extend the CLRT to include information on linkage disequilibrium using their ω statistic (Equation 9.37), which contrasts measures of LD among all pairs of loci on the same side of a putative sweep with those for pairs of loci on opposite sides of the sweep. While incorporation of LD information did improve estimation, the gain was rather modest. Li and Stephan (2005, 2006) develop a true maximum likelihood method that uses only a subset of the frequency spectrum (the **compact frequency spectrum**, recording the number of singletons, doubles, and sites greater than two) to estimate the position of the sweep.

Tests of Sweeps Using Genomic Spatial Information: "SweepFinder"

Nielsen et al. (2005) proposed a modification of the CLRT approach, replacing the Watterson distribution by an empirical distribution $\mathbf{p} = (\widehat{p}_1, \cdots, \widehat{p}_{n-1})$, where $\mathbf{p}$ is estimated by using a reference sample of m presumed neutral sites, with $\widehat{p}_k = n_k/m$, where n_k is the number of sites with exactly k copies of the derived allele. Their idea is that the Watterson distribution assumes an equilibrium neutral population, while using the actual distribution observed in the population of interest (at presumed neutral sites) to a large extent can accommodate any demographically-induced departures. In addition to partly accounting for the underlying demographic model, the use of an empirical site-frequency spectrum should also at least partly correct for SNP ascertainment bias. While rather an elegant approach, the delicate issue is being able to find an appropriate (and large) set of presumed neutral sites. This approach goes by the name **SweepFinder**, and the resulting likelihood function is derived in Example 9.7. While Nielsen et al.'s (2005) simulations showed this approach is more robust that the CLR test, demography can still influence the test statistic. In particular, intermediate bottlenecks (size reduced to 5 to 10% of the original N_e) seem to be the most problematic (Williamson et al. 2007; Stephan 2010a). Hence, significance values should be obtained by simulating this procedure under a set of assumed demographic models.

Example 9.7. While using the same basic logic as the CLRT, namely constructing a likelihood model where the frequency spectrum is a function of distance from the site, there is a bit more bookkeeping required to obtain the likelihood function. The task is to translate a background pattern $\mathbf{p}$ before a sweep into a pattern $\mathbf{p}_s$ after the sweep. Nielsen et al. approach this problem by considering the lineages in a sample of size n. Suppose ℓ of these represent lineages that have escaped the sweep (and hence will be reflective of the normal background distribution $\mathbf{p}$), while the remainder $n - \ell$ will be lineages that did not escape, and hence will either all contain the derived allele (if it was associated with the initial favorable mutation) or all lack it. From Chapter 8, the probability that any sample sequence escaped a sweep is $1 - f_s$. Assuming a strong sweep, the probability for a site at distance c from the sweep that ℓ out of n sample sequences are lineages that escaped the sweep is binomial with success probability

$1 - f_s$,

$$P_e(\ell) = \binom{n}{\ell}(1-f_s)^\ell f_s^{n-\ell} = \binom{n}{\ell}\left(1-e^{-c\,\zeta}\right)^\ell \left(e^{-c\,\zeta}\right)^{n-\ell} \tag{9.18a}$$

Suppose that ℓ lineages from our sample escaped the sweep and thus $n-\ell$ did not, so that there are $M = \min(\ell+1, n)$ lineages (the ℓ distinct lineages and the single lineage associated with the sweep). Conditioned on ℓ, we need to compute $p_{s,i}$ the probability of seeing i/n derived alleles in our sample. We then average this over $P_e(\ell)$ to obtain the likelihood for a site. The probability of seeing j derived alleles in a sample of m lineages at the start of the sweep is

$$\Pr(j \mid M) = \sum_{i=j}^{n-1} p_i \frac{\binom{i}{j}\binom{n-i}{M-j}}{\binom{n}{M}} \tag{9.18b}$$

where the second term is from the hypergeometric distribution: Assuming i copies of the derived allele in a sample of n initial lineages, this gives the probability that M draws (without replacement) gives j copies in our sample. Averaging over the probability p_i that i copies of the derived allele were in this ancestral sample gives us $\Pr(j \mid M)$. Given that there are j lineages of the M carrying the derived allele, the probability the derived allele is in the lineage that did not escape is just j/M. For a derived allele to have i copies at a site following the sweep, it either was linked to the favorable allele and hence was present in $j-(n+\ell+1)$ of the $\ell+1$ lineages at the start of the sweep, leading to j copies in the sample, or it was not associated with the favorable mutation, so that j copies in the initial $\ell+1$ lineages leads to j copies in the sample. Putting these together gives the likelihood as

$$\begin{aligned} p_{S,i} = P_e(n)\,p_i + \sum_{\ell=0}^{n-1} P_e(\ell) \Bigg(& \Pr\Big[j-[n+\ell+1] \,\Big|\, \ell+1\Big] \frac{j-(n+\ell+1)}{\ell+1} \\ & + \Pr\Big[\ell+1-j \,\Big|\, \ell+1\Big] \frac{\ell+1-j}{\ell+1} \Bigg) \end{aligned} \tag{9.18c}$$

These values now replace those given by Equation 9.16a to construct the likelihood under a sweep. It is useful to note where parameters appear. The sweep strength and location appears through f_s in the $P_e(\ell)$ term, the number of lineages that escaped the sweep. The empirical background spectrum $\mathbf{p}$ appears in the $\Pr(j \mid M)$ terms that populate the original sample of lineages at the start of the sweep.

Tests of Sweeps Using Genomic Spatial Information: XP-CLR

Chen et al. (2010) introduced a test that is similar in spirit to the CLRT, but uses the spatial pattern of *allele frequency differences* between two populations (a reference and a candidate), as opposed to heterozygosity data from a single population. Their **cross-population composite likelihood ratio test**, or **XP-CLR** for short, is constructed as follows. They assume a biallelic-locus, such as a SNP, and score the frequency of an allele in the reference and candidate populations, with selection assumed to have occurred in the latter (but not the former). First, they model the neutral divergence in allele frequency between two isolated populations that originally shared a common ancestor. To simplify matters, they use a Brownian motion (or Wiener process) approximation for drift (Appendix 1), which simply means using a normal for the expected distribution of allele frequencies. For pure drift, the expected change $m(x)$ in an allele at frequency x is zero, while the expected variance of the change is $v(x) = x(1-$

$x)/(2N_e)$. Building the Wiener process by using Equation A1.31a, the expected frequency x of an allele at generation t given it started at frequency p_0 is approximately normally-distributed, with

$$x(t) \sim \mathrm{N}\left(p_0, t\,\frac{p_0(1-p_0)}{2N_e}\right) \tag{9.19a}$$

More generally, we can write the variance as $\beta\, p_0(1-p_0)$, where the β term accounts for the population history (allowing for variation in population size, etc.), so that if p_2 is the allele frequency in some reference population, then under a pure drift model the distribution of frequencies in the candidate population should follow

$$x \sim \mathrm{N}\left(p_2, \sigma^2\right), \quad \text{where} \quad \sigma^2 = \beta\, p_2(1-p_2) \ll 1 \tag{9.19b}$$

The last condition follows because this approximation only works well for σ^2 small. Because of the shared population history, all neutral genes should have (roughly) the same β value, which one can estimate directly from the data. If k out of n sampled chromosomes in the control population contain the allele, then the likelihood (for β) for this site is given by using Equation 9.16a, but with $\phi(x)$ now replaced by the normal density function with parameters given by Equation 9.19b. Chen et al. introduce an additional refinement. Instead of simply multiplying all individual likelihoods together, they form their composite likelihood by multiplying the *weighted* product of the likelihoods, downweighting SNPs that are in LD in the reference population.

Now consider the effect of a sweep in the candidate population. Their argument follows the same logic leading to the shift in the frequency spectrum under a sweep (Equation 8.13), but now the focus is the shift in the distribution of the allele frequency in the candidate population. Suppose a single mutation immediately under selection arises that is linked to the SNP being followed. If the SNP allele frequency is x, then (Table 8.1) with probability x, the SNP allele increases to frequency $f_s + x(1-f_s)$, while with probability $1-x$ it decreases to $x(1-f_s)$, where f_s is again given by Equation 9.16b. Akin to Equation 8.13, Chen et al. show that the distribution of x in the candidate population shifts to a mixture of two normals, one representing SNP alleles not initially associated with the new favorable mutation, and hence driven towards frequency zero, and the second where the SNP allele was initially associated with the mutation and driven close to frequency one, giving

$$x \sim \left(\frac{f_s - x}{f_s^2}\right) I_{[0,f_s]}\, \mathrm{N}\left(f_s p_2, f_s^2\sigma^2\right) + \left(\frac{x + f_s - 1}{f_s^2}\right) I_{[1-f_s,1]}\, \mathrm{N}\left(f_s p_2 + 1 - f_s, f_s^2\sigma^2\right) \tag{9.20}$$

where $I_{[a,b]}$ denotes an indicator function which is one when x is in the interval (a,b) and zero otherwise. This distribution now replaces $\phi(x)$ in Equation 9.16a to give the likelihood under selection for this site. As with the neutral likelihoods, a composite likelihood function is formed (again weighted by LD between SNPs in the reference population). The formal test consists of the ratio of the maximum of composite likelihood under the sweep model (Equation 9.20) to the maximum of the composite likelihood under the neutral model (Equation 9.19b).

What makes Chen et al.'s method potentially compelling is that the empirical null distributions over a wide range of demographic models were essentially identical. Further, this approach had greater power than the CLRT in their simulation studies, as well as showing robustness to variation in recombination rates. Peng et al. (2011) present an interesting application of this model, localizing human genes for adaptation to high altitudes by contrasting a Himalayan Tibetan population (as the candidate) with Han Chinese from Beijing (as the reference).

The various CLR tests are highly parametric, which offers more power if the model is correct, but concerns about robustness if the model is wrong. A nonparameteric approach

using spatial information on both heterozygosity and between-population divergence was suggested by Oleksyk et al. (2008). Their application considers two populations, and used a series of sliding windows. Within each window, three statistics were computed: the average heterozygosity for both populations and the variance in F_{ST} values within each window (the latter because a spread in F_{ST} values is expected as one moves from near a selected site to away from it). To assess significance, this same procedure was replicated for all three statistic by sampling SNPs at random from the genome (removing any spatial information). The authors used 100,000 resamples to generate a null distribution for each sized window, plotting the resulting p values for each statistic as a scan along the chromosome. Regions with two (or more) significant statistics were candidates for selection. While conceptually simple, the concern is that unusual neutral genealogies also contain spatial information, and randomly choosing SNPs for the permutation test destroys these correlations, inflating the false-positive rate.

Ascertaiment Issues

Since many of the above likelihood models exploiting genomic spatial information are computationally demanding, they are typically employed *following* a general scan of a genome for some signature of selection, such as a regions of depressed variation, or site-frequency tests (such as negative Tajima's D or positive Fay and Wu's H values, discussed in the next section). Choosing which region(s) to perform the likelihood tests based on the appearance of these special features creates a strong ascertainment bias that dramatically shifts the null distribution. (Note that this is different from SNP ascertainment bias arising from nonrandom choice of SNPs at the start of the analysis). The coalescent process can be noisy and regions with unusual underlying genealogies (such as strong compression of nodes) can occur by chance even under the equilibrium neutral model. This is especially true when a large number of sites are sampled, presenting more draws from the same underlying process, some of which will be realizations that are a bit extreme.

Thornton and Jensen (2007) outline an approach to adjust for both ascertainment and non-equilibrium population structure when simulating the neutral null distribution. First, genome scan data is used to estimate the parameters for an appropriate demographic model (such as the time and duration of a bottleneck). Next, these are used in neutral coalescent simulations. Normally, this would be the null distribution. However, we also have to model the ascertainment process. For example, suppose the lowermost two percent of regions of reduced variation are chosen for follow-up CLR tests. The appropriate null would be constructed by also sampling such low-variation regions from the unconditional null, and then using these as the appropriate ascertained null distribution. Thornton and Jensen note that increasing the length of region of analysis is a good general strategy for increasing power. Likewise, they and Teshima et al. (2006) both note that measures of either diversity (such reduced heterozygosity vales) or population differentiation (such as F_{ST} values) seem more reliable for identifying outliers than frequency-spectrum based approaches (such as Tajima's D or Fay and Wu's H).

Model Fragility: Demography, Mutation, Recombination, and Gene Conversion

The concern one must always have in mind when using a sophisticated model is its *robustness* to errors in the underlying assumptions. Simple approaches often have modest power but considerable robustness, while highly specialized models can be quite powerful when the data fit the assumptions, but may be quite fragile when they do not. Given the constant concern about nonequilibrium and/or structured populations, and hence generation of the appropriate null model, the robustness of any method to demography is critical. One simple approach was suggested by Akey et al. (2004): look for signals that are **demographically**

robust, showing significant evidence for selective over a wide range of possible demographic models.

While fragility to demographic assumptions is generally well-appreciated, less so are the effects of genomic assumptions, such as a constant mutation and/or recombination rate across the region. We have already stressed that variation in mutation rates can generate peaks and valleys in the background patterns of neutral variation. These patterns can easily be declared as signals of selection by most of the above likelihood tests. The exception is XP-CLT, which being based on between-population differences, controls for this to some extent. Likewise, although the recombination rate r between a neutral site and the selected target appears in most of the likelihood models, operationally one assumes a constant rate c per nucleotide, with $r = cd$, the distance d in nucleotides. Recombination rates can vary dramatically over very fine scales (Coop and Przeworski 2007). Finally, gene conversion (the nonreciprocal transfer of information between two sites) also needs to be considered (Andolfatto and Nordborg 1998). Operationally, a conversion event is identical to a double-recombinant involving a short region of DNA, and can often occur in regions of low recombination. Conversion events can disrupt the signal excepted from a sweep. For example, Glinka et al. (2006) observed a sharp peak of variation in the middle of a valley of depressed variation around the *unc-119* gene in European population of *Drosophila*. They reasoned that a sweep plus two conversion events generated this unusual pattern.

TESTS BASED ON THE SITE-FREQUENCY SPECTRUM

Under the infinite-sites model, a sequence is treated as a series of L sites, with each new mutation assumed to occur at a new site (Chapter 4). At mutation-drift equilibrium, most features of this model, including the site-frequency spectrum (SFS) are fully specified by the parameter $\theta = 4N_e\mu$, the population-size scaled mutation rate. Since our focus here is testing diversity over a *region* of interest, in this section θ refers to the regional value (μ is the total rate over the region), or L times the average estimate based on a per-nucleotide basis (Chapter 4). Depending on the nature of the data, an observed frequency spectrum is viewed as either folded or unfolded (Chapter 2). An unfolded spectrum considers the frequency of the derived allele (Equation 2.35a), and such data are said to be polarized (typically using an outgroup to distinguish between ancestral and derived = mutant alleles). The folded spectrum (Equation 2.35b) uses the frequency of the minor allele at a site, ignoring whether it is ancestral or derived.

As detailed in Chapter 8, both a hard sweep and long-term balancing selection are expected to perturb a starting site-frequency spectrum into some new distribution. A hard sweep increases the frequency of both sites with rare derived alleles and sites with high-frequency derived alleles (Figure 8.5). With a folded frequency spectrum, these jointly appear as an increase in sites with rare alleles. Conversely, long-term balancing selection is expected to increase the number of sites with intermediate-frequency alleles. While widely used, the common problem with site-frequency spectrum tests is that a nonequilibrium population (such as during recovery following a bottleneck) or one with structure also causes the frequency spectrum of neutral alleles to depart from the benchmark Watterson distribution. Thus, a significant amount of our discussion is how to deal with these concerns.

Summary Statistics Based on Infinite-Sites Models

As introduced in Chapter 4, there are a variety of summary statistics that can be used to estimate $\theta = 4N_e\mu$ under the infinite-sited model. For a sample of L sites (generally nucleotides), suppose there are S segregating sites, S_1 sites with singletons (a single copy of one allele), and let Π be the average number of pairwise differences between two random sequences.

If our goal is to estimate θ on a *per-nucleotide* basis, so that μ is the per-nucleotide mutation rate, then (as in Chapter 4), we would consider the fraction of segregating sites S/L, the nucleotide diversity $\pi = \Pi/L$, and the fraction of sites that are singletons S_1/L. When looking for genomic regions under selection, our goal shifts to $\theta_L = 4N_e\mu_L$, the corresponding value for the region, where $\mu_L = L\mu$ is the total mutation rate over the L sites. Since all of the focus in this section is on specific regions, in an effort to keep the notation simple we will drop the subscript and just use θ and μ to denote the region-side values (θ_L, μ_L).

While S and Π have the same values for polarized and unpolarized data, the number of singletons can be slightly different. To distinguish these, let S_1 denote the number of sites with a single copy of the derived allele and S_1^* the number of sites with a single copy of the minor allele ($S_1^* \geq S_1$, but the two are usually very close). All of these summary statistics yield estimates of θ, with

$$\widehat{\theta}_S = \frac{S}{a_n}, \qquad \widehat{\theta}_\Pi = \Pi, \qquad \widehat{\theta}_1 = S_1^*, \qquad \widehat{\theta}_{1^*} = \frac{n-1}{n} S_1^* \tag{9.21a}$$

where $a_n = \sum_{j=1}^{n-1} 1/j$ (Equation 4.3b). These correspond (respectively) to the Watterson estimator (Equation 4.3a, which is also commonly denoted by θ_W), Tajima's estimator (Equation 4.1), our previous singleton estimator (Equation 4.6a), and the corresponding singleton estimator using folded data. The sampling variances for these estimates are given by Equations 4.4a ($\widehat{\theta}_S$), 4.2 ($\widehat{\theta}_\Pi$), and 4.6b ($\widehat{\theta}_1$). These expressions for the variance are functions of both θ and θ^2, and are typically (e.g., Tajima 1989) computed by replacing

$$\theta \;\text{ by }\; S/a_n \qquad \text{and} \qquad \theta^2 \;\text{ by }\; \frac{S(S-1)}{a_n^2 + b_n} \tag{9.21b}$$

where $b_n = \sum_{j=1}^{n-1} 1/j^2$ (Equation 4.4b).

Example 9.8. Suppose we sample 10 alleles from a population and observe a total of 12 segregating sites ($S = 12$), an average of 4 differences between alleles ($\Pi = 4$), and 3 segregating sites that have only a single copy of the minor allele ($S_1^* = 3$). What are the estimates of θ based on these three summary statistics?

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

$$\widehat{\theta}_\Pi = \Pi = 4, \quad \text{and} \quad \widehat{\theta}_{1^*} = \frac{n}{n-1} S_1^* = \frac{10}{9} \cdot 3 = 3.33$$

The idea behind site-frequency tests is to compare two different estimates of θ based on information from different parts of the site-frequency spectrum. When infinite-sites model holds and the population is at mutation-drift equilibrium, these estimates should be within sampling error of each other, while they can be significant different when the neutral equilibrium model does not hold. Table 9.1 summarizes the various site-frequency test statistics discussed here, all of which have the form

$$t = \frac{\widehat{\theta}_i - \widehat{\theta}_j}{\sigma(\widehat{\theta}_i - \widehat{\theta}_j)} \tag{9.21c}$$

Table 9.1. Summary of the site-frequency tests presented in this chapter, which contrast estimates of θ based on different parts of the site-frequency spectrum. The estimators $\widehat{\theta}_S$, $\widehat{\theta}_\Pi$, $\widehat{\theta}_1$, and $\widehat{\theta}_{1^*}$ are given by Equation 9.21a, while the estimators $\widehat{\theta}_H$ (Equation 9.28a) and $\widehat{\theta}_L$ (Equation 9.29a) are developed below. The column labeled spectrum indicates whether the test requires unfolded data, with alleles designated as ancestral or derived. Further details are given in the text.

Test	Contrast	Spectrum	Signal
Tajima's D	$\widehat{\theta}_S$ vs. $\widehat{\theta}_\Pi$	Folded	< 0: Excess of rare alleles Sweep or population bottleneck > 0: Excess of intermediate-frequency alleles Balancing selection or population structure
Fu and Li's D	$\widehat{\theta}_S$ vs. $\widehat{\theta}_1$	Unfolded	Same as for Tajima's D
Fu and Li's D^*	$\widehat{\theta}_S$ vs. $\widehat{\theta}_{1^*}$	Folded	Same as for Tajima's D
Fu and Li's F	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_1$	Unfolded	Same as for Tajima's D
Fu and Li's F^*	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_{1^*}$	Folded	Same as for Tajima's D
Fay and Wu's H	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_H$	Unfolded	< 0: Excess of high-frequency derived alleles Sweep or allelic surfing
Zeng et al.'s E	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_L$	Unfolded	< 0: Excess of low- v. high-frequency derived alleles Signal of a recent *past* sweep

When applying any of these tests, care must be taken to avoid SNP ascertainment bias. If the process by which SNPs are chosen is biased by their frequency (SNP discovery panels are generally biased in favor of intermediate-frequency sites), this results in a biased estimate of the frequency spectrum, potentially compromising tests. Likewise, when tests are based on the unfolded SFS, errors introduced by incorrect polarity assignment (incorrectly assigning a derived allele ancestral status, and vice versa) can be serious, if not fatal (Baudry and Depaulis 2003; Hernandez et al. 2007). Finally, sequencing errors result in an increase in singletons, which can bias tests away from neutrality (Johnson and Slatkin 2008; Achaz 2008). Errors on the order of one per 10 to 100 singletons can significantly bias results.

Example 9.9. All of the tests summarized in Table 9.1 follow from a general family of estimators of θ based on the Waterson distribution for a sample of n sequences with L sites (Equation 2.35). The expected number n_i of segregating sites with i copies of the derived (unfolded) or minor (folded) allele is,

$$E(n_i) = \begin{cases} \dfrac{\theta}{i}, & i \text{ copies of the derived allele,} \quad 1 \leq i \leq n-1 \\ \\ \dfrac{\theta}{i}\dfrac{n}{n-i}, & i \text{ copies of the minor allele,} \quad 1 \leq i \leq [n/2] \end{cases} \tag{9.22a}$$

where $\theta = 4N_e\mu$ is the scaled mutation rate for the entire region and $[n/2]$ denotes largest integer below or equal to $n/2$. Hence, an estimator for θ using only the number in the ith class is just

$$\widehat{\theta}_i = \begin{cases} i \cdot n_i, & i \text{ copies of the derived allele,} \quad 1 \leq i \leq n-1 \\ \\ \dfrac{i \cdot (n-i)}{n}\, n_i, & i \text{ copies of the minor allele,} \quad 1 \leq i \leq [n/2] \end{cases} \tag{9.22b}$$

Nawa and Tajima (2008) suggest that a plot of $\widehat{\theta}_i$ versus i can be helpful for visualizing departures from the neutral FS, although values for large i may be more problematic as the variance of $\widehat{\theta}_i$ dramatically increases with i (as fewer expected observed weighted by larger numbers). Following Zeng et al. (2006), consider any summary statistic g of the unfolded site-frequency spectrum of the form

$$g = \sum_{i=1}^{n-1} c_i \, n_i \tag{9.23a}$$

From Equation 9.22a,

$$E(g) = \sum_{i=1}^{n-1} c_i \, \frac{\theta}{i} = \theta \, u(n), \quad \text{where} \quad u(n) = \sum_{i=1}^{n-1} \frac{c_i}{i} \tag{9.23b}$$

Thus, a family of estimators for θ based on an arbitrary vector $(c_1, \cdots, c_{n-1})$ of weights is given by

$$\widehat{\theta}_g = \frac{g}{u(n)} \tag{9.23c}$$

where $u(n)$ is a function of the sample size n and the chosen weights c_i, and g is the observed value of the statistic. The choice of weights allows one to tailor statistics to use different parts of the frequency spectrum when estimating θ. Taking $c_i = 1$ gives $g = S$ and $u(n) = a_n$, recovering the Watterson estimator. Taking $c_i = i(n-i)$, so that $u(n) = \sum_{i=1}^{n-1} i(n-i)/i = n(n-1)/2$ computes the average pairwise difference Π. As with S, Π is symmetric with respect to i and $n-i$, so that both folded and unfolded data return the same estimate. Taking $c_1 = 1, c_{i>1} = 0$ gives $g = S_1$ (the number of derived singletons), recovering our $\widehat{\theta}_1$ estimator. Similarly, for a folded frequency spectrum,

$$g = \sum_{i=1}^{[n/2]} c_i \, n_i, \qquad \widehat{\theta}_g = \frac{g}{f(n)}, \qquad f(n) = \sum_{i=1}^{[n/2]} c_i \, \frac{n}{i(n-i)} \tag{9.23d}$$

Considering folded singletons (S_1^*), $f(n) = n/(n-1)$, recovering $\widehat{\theta}_{1^*}$. Achaz (2009) extends these ideas and presents general expressions for obtaining the variance of any estimator, and the covariance between any two estimators, providing all of the machinery to develop general tests in the form of Equation 9.21c using any feature of interest in the SFS.

Tajima's D Test

The first, and most widely used, site-frequency spectrum test is **Tajima's D** (1989), which contrasts θ estimates based on segregating sites (S) and average pairwise difference (Π),

$$D = \frac{\widehat{\theta}_\Pi - \widehat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} \tag{9.24a}$$

where

$$\alpha_D = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) - \beta_D \tag{9.24b}$$

$$\beta_D = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right) \tag{9.24c}$$

Being based on S and Π, this test does not require unfolded data. Tajima's motivation was that θ_S and θ_Π measure different features of the frequency spectrum. The number of segregating sites S (and thus θ_S) simply counts polymorphic sites independent of their frequencies, making it more sensitive to changes in the frequencies of rare alleles (as small changes can cause sites segregating rare alleles to either enter, or drop out of, the sample). Conversely, the average pairwise difference Π (and thus θ_Π) is a frequency-weighted measure and more sensitive to changes in the frequencies of intermediate alleles. A negative value of D indicates too many low-frequency sites, a positive D too many intermediate-frequency sites. Expressed another way, D is a test for whether the amount of heterozygosity per site is consistent with the number of polymorphic sites expected under the equilibrium neutral model. Under selective sweeps (and population expansion), heterozygosity should be significantly less than predicted from the number of polymorphisms. As with all site-frequency spectrum tests, the distribution of D is *critically* dependent on the neutral equilibrium assumption. Tajima obtained upper D_{max} and lower D_{min} bounds on D, so that $(D - D_{min})/(D_{max} - D_{min})$ lies in (0,1). On this modified scale, Tajima showed that D is well approximated by a Beta distribution (Equation A2.40) with mean zero and variance one. The distribution for D for a neutral allele but in a population of changing size is significantly different (Innan and Stephan 2000; Živković and Wiehe 2008).

Example 9.10. Two interesting applications of the D test were offered by Tajima (1989). First, he considered Aquadro and Greenberg's (1983) data for 900 base pairs in the mitochondrial DNA of seven humans. They observed 45 segregating sites and an average number of nucleotide differences between all pairs of 15.38. Hence,

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

$$\widehat{\theta}_S = \frac{S}{a_n} = \frac{45}{2.45} = 18.38, \qquad \widehat{\theta}_\Pi = \Pi = 15.38$$

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

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

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

Table 2 of Tajima (1989) gives the 95% confidence interval for 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, he used the data of Miyashita and Langley (1988) who 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 $\Pi = 0.94$, which gives $D = -2.0709$. Given that the 95% confidence interval under neutrality is -1.795 to 2.055, the site-frequency spectrum associated with this locus shows evidence of either directional selection or a population bottleneck.

Since the minimal value of D varies with the number of segregating sites S, Schaeffer (2002) proposed a standardized $D' = D/D_{min}$, that adjusts for this, allowing for more fair

comparisons of D across loci. He notes that the minimum pairwise difference given S is

$$\Pi_{min} = S\,\frac{2(n-1)}{n^2} \tag{9.25}$$

D_{min} is computed from Equation 9.24a with Π_{min} replacing $\widehat{\theta}_{\Pi}$. Finally, Achaz (2008) noted that estimates of both S and π can be biased by sequencing errors, which introduce an excess of singletons, skewing D towards more negative values. His **Y** and **Y*** tests are modifications of Tajima's D by computing $\widehat{\theta}_{\Pi}$ and $\widehat{\theta}_{S}$ after removing singletons. The Y test applies for a unfolded SFS (only derived singletons removed), while the Y^* test is for folded SFS (any singleton removed).

Fu and Li's D* and F* Tests

Fu and Li (1993) introduced tests based on the two other contrasts of the three infinite-sites θ estimators given by Equation 9.21a. Both use the number of singleton sites, with variants using either folded (S_1^* sites with a single copy of the minor allele) or unfolded (S_1 sites with a single copy of the derived allele) singletons. Using these statistics gives rare-alleles based estimates of θ, which are then contrasted with estimates based on either S or Π. Using folded data, this gives rise to their D^* and F^* tests that we consider here. Their exact counterparts for unfolded data (using S_1 in place of S_1^*) are their D and F tests, which are not discussed further. (Given the widespread use of Tajima's D, when we simply reference a "D test", this always refers to Tajima's test.) Fu and Li's **D* test** compares the segregating sites (S) versus folded singletons (S_1^*) estimators of θ,

$$D^* = \frac{\widehat{\theta}_S - \widehat{\theta}_{1^*}}{\sqrt{\alpha_* S + \beta_* S^2}} \tag{9.26a}$$

$$\alpha_* = \frac{1}{a_n}\left(\frac{n+1}{n} - \frac{1}{a_n}\right) - \beta_* \tag{9.26b}$$

$$\beta_* = \frac{1}{a_n^2 + b_n}\left(\frac{b_n}{a_n^2} - \frac{2}{n}\left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n}\right) - \frac{1}{n^2}\right) \tag{9.26c}$$

Their **F* test** compares the average pair-wise divergence (Π) versus folded singletons (S_1^*) estimators,

$$F^* = \frac{\widehat{\theta}_{\Pi} - \widehat{\theta}_{1^*}}{\sqrt{\alpha_F S + \beta_F S^2}} \tag{9.27a}$$

$$\alpha_F = \frac{1}{a_n}\left(\frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)}\right) - \beta_F \tag{9.27b}$$

$$\beta_F = \frac{1}{a_n^2 + b_n}\left(\frac{2n^4 + 110n^2 - 255n + 153}{9n^2(n-1)} + \frac{2(n-1)a_n}{n^2} - \frac{8b_n}{n}\right) \tag{9.27c}$$

These expression are from Simonsen et al. (1995), with Equation 9.27c correcting a typo in the original Fu and Li paper. Critical values (assuming no recombination in the region) 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 demographic features (bottlenecks or population subdivision).

Fay and Wu's H Test

The first test to use the full power of the unfolded frequency spectrum was proposed by Fay and Wu (2000), who noted that a hard sweep results in an excess of sites with high frequency

derived alleles. However, this signature is rather fleeting (Figure 9.4). This excess forms the basis for their **H test**, which disproportionately weights sites containing derived alleles at high frequencies, using $c_i = i^2$. From Equation 9.23b, these weights imply $u(n) = n(n-1)/2$, and Equation 9.23c gives

$$\widehat{\theta}_H = 2\sum_{i=1}^{n-1} \frac{i^2}{n(n-1)}\, n_i \tag{9.28a}$$

The H test is the scaled difference between their estimator and than for average pairwise differences,

$$H = \frac{\widehat{\theta}_\Pi - \widehat{\theta}_H}{\sigma(H)} \tag{9.28b}$$

Zeng et al. (2006) obtain the sampling variance as

$$\sigma^2(H) = \left[\frac{n-2}{6(n-1)}\right]\theta + \left[\frac{18n^2(3n+2)b_{n+1} - (88n^3 + 9n^2 - 13n + 6)}{9n(n-1)^2}\right]\theta^2 \tag{9.28c}$$

As above, in computing Equation 9.28c, Equation 9.21b is used for θ and θ^2. Since Π is a measure of the intermediate-frequency sites, H is a contrast between high- and intermediate-frequency variation, with a negative H indicating an excess of sites with a high-frequency of derived alleles. Jointly negative (and significant) values of D and H are consistent with a selective sweep, indicating an excess of rare alleles and an excess of high-frequency derived alleles. One caution when applying the H test is that is it extremely sensitive to polarity errors (Baudry and Depaulis 2003).

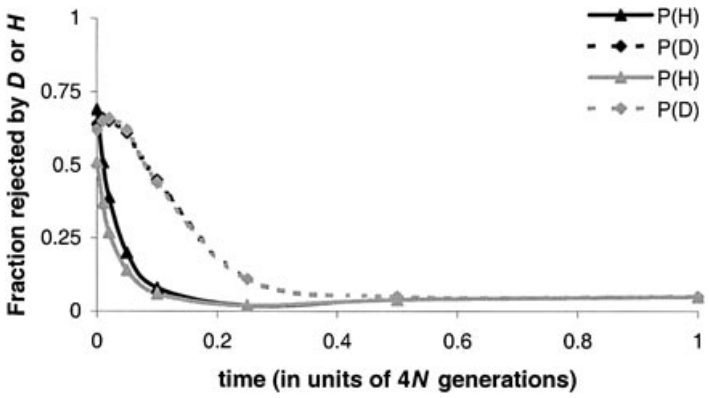


Figure 9.4. The power of the H and D tests to detect signatures of a recent sweep is very fleeting. The power of H (which is based on high-frequency derived alleles) in particular falls rapidly after a sweep (as high frequency alleles are fixed), essentially having power only within $0.2N_e$ generations following a sweep. D, based on an excess of rare alleles, can pick up new mutations entering following the sweep, and has power over about $0.5N_e$ generations following a sweep. A value of $c/s = 0.01$ was used and power is graphed for two different values of $4N_e s$ (20,000 and 2000, the dark and grey lines, respectively). After Przeworski (2002).

Przeworski (2002) showed that both the D and H tests have moderate power immediately after a sweep, but that the power of the H test rapidly dissipates (within $0.2N_e$ generations) as the high frequency alleles become fixed (Figure 9.4). The D test retains power

a bit longer (roughly $0.5N_e$ generations), as it is sensitive to new mutations generating rare alleles immediately after the sweep. As a result of using different signatures in the diversity data, one can easily have situations where, following a sweep, one test is highly significant while the other is not. Even for a strong hard sweep, neither D nor H may be significantly negative, depending on the time since the sweep was completed (Figure 9.4).

Zeng et al.'s E Test

A variant of the H test was proposed by Zeng et al. (2006). They noted that the most powerful contrasts between regions of the unfolded frequency spectrum following selection should be the high and low frequency sites. However, most contrasts involve comparison against θ_Π, which is a measure of intermediate-frequency alleles, while θ_S is sensitive to low frequency alleles and θ_H to high-frequency derived alleles. They introduced the estimator θ_L based on the weights $c_i = i$ that places more emphasis on high-frequency sites that θ_S (but not as much as does θ_H). For these weighs, Equation 9.23b implies $u(n) = n - 1$, and hence Equation 9.23c gives

$$\widehat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\, n_i \tag{9.29a}$$

Zeng et al.'s **E test** contrasts the high and low frequency regions of the frequency spectrum,

$$E = \frac{\widehat{\theta}_L - \widehat{\theta}_S}{\sigma(E)} \tag{9.29b}$$

where

$$\sigma^2(E) = \left[\frac{n}{2(n-1)} - \frac{1}{a_n}\right]\theta + \left[\frac{b_n}{a_n^2} + 2\left(\frac{n}{n-1}\right)^2 b_n - \frac{2(nb_n - n + 1)}{(n-1)a_n} - \frac{3n+1}{n-1}\right]\theta^2 \tag{9.29c}$$

Again the variance is computing by replacing θ and θ^2 by Equation 9.21b. A negative E indicates an excess of low- versus high-frequency sites relative to expectations of the equilibrium neutral model. This occurs *immediately after* a sweep, as the excess of high-frequency alleles is lost by drift, while at the same time mutation is generating an excess of low-frequency sites which have yet to drift up their neutral equilibrium values. The unique feature of E is in indicating a signal *after* a sweep, persisting up to $2N_e$ generations (and hence much longer than H).

Adjusting the Null to Account for Nonequilibrium Populations

The site-frequency tests summarized in Table 9.1 critically depend on the Watterson distribution as the null model, as all are in the family θ estimators given by Equation 9.23. As such, they are especially susceptible to false positives when samples are from a population not satisfying the equilibrium assumptions (a panmictic population in mutation-drift equilibrium). Four strategies have been proposed to address this concern. The first three are standard approaches to refine the null to better suit the sampled population: using the empirical distribution of test statistics from a set of genes in the sample (the outlier approach), coalescent simulations using marker-based estimates of demographic parameters, and using the empirical site-frequency spectrum at a reference location as the null. The final strategy is **support via a preponderance of evidence**, considering joint signatures from a number of different tests, and will be discussed separately.

While approaches attempting to account for nonequilibrium populations offer improvements over tests based on the standard neutral model, they still do not guarantee that significant signatures represent true regions influenced by positive selection. Because of this,

the current operational use for many of these tests (such as Tajima's D) are as *convenient summary statistics* for features in a region of interest (such as an excess of rare alleles).

The empirical distribution of a test statistic over a large number of genes sampled from the target population can potentially provide useful information. Under the equilibrium-neutral model, the test statistics reviewed in Table 9.1 should have mean zero, while the empirical distribution shows whether tests trend away from this expectation in the target population. Figure 9.5 gives examples of the empirical distribution of D for 201 genes in two different human populations. For these genes in African-Americans, the mean D is negative, while it is positive in European-Americans. A gene whose negative D value is significant under the equilibrium-neutral model is likely to be even more significant in this European-American population (given the trend towards positive D), but problematic in this sample of African-Americans (given that random tests trend toward negative D). While the mean of this distribution can be informative, one cannot simply use it to adjust test statistics for individual genes. This is because departures from the standard neutral model often *inflate the variance* of test statistics (Nielsen 2001). Thus, even when the mean of the empirical distribution is zero, the variance under the standard model may be too narrow, and hence significance can be overstated. Finally, the empirical distribution is largely shaped by common demographic features influencing all genes. Allelic surfing of neutral alleles does not leave a genome-wide signature, and sites experiencing surfing can easily create outliers, mimicking signatures of selection.

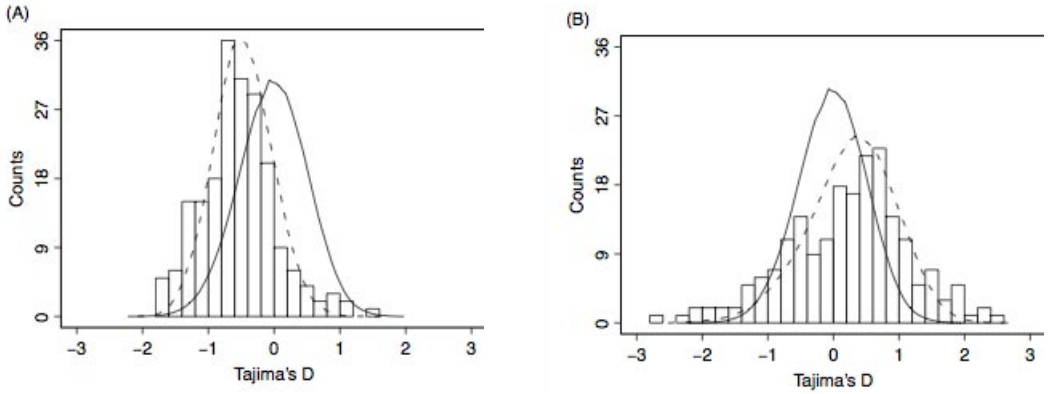


Figure 9.5. Distribution of Tajima's D for 201 genes in African-American (A) and European-American (B) samples. The empirical distribution is given by the histogram, the solid line gives the simulated values under the equilibrium neutral model, and the dashed lined the simulated distribution under the best fitting demographic model. For (A) this is exponential growth starting 50,000 years ago, while for (B) this is a bottleneck starting 40,000 years ago. After Ronald and Akey (2005).

The second approach to account for a nonequilibrium population is to use genomic data to infer demographic parameters (such as the size and duration of any past bottleneck), which are then used as the basis for a coalescent simulations (Chapter 2). This generates a more appropriate null distribution of the test statistic for the target population (e.g., Figure 9.5). An example is Schaffner et al. (2005), who used human data to find the best-fitting model over a rather rich parameter space, including population structure, bottleneck times, and even variation in recombination rates. Tenaillon et al. (2004) performed a similar analysis on the bottleneck during the formation of maize from teosinte. With estimated demographic parameters in hand, coalescent programs such as MSMS (Ewing and Hermisson 2010), GENOME

(Liang et al. 2007), cosi (Schaffner et al. 2005), or MS (Hudson 2002) can be used. Again, this approach only corrects for demographic features that leave a common signal over the entire genome, so that sites that experienced allelic surfing can still generate false signals of selection even after this correction.

The final approach is to use the empirical site-frequency spectrum vector $\mathbf{p}$ from a reference set (as opposed to the Watterson distribution) as the null (Nielsen et al. 2005, 2009). Here p_i is the fraction of sites in the reference set with i copies of the allele (derived or minor, for the unfolded and folded spectra, respectively). A standard goodness-of-fit test (such as the G-test, LW Appendix 2) is then used to assess whether the spectrum $n_1, \cdots n_{n-1}$ in a candidate region is consistent with the multinominal probabilities given by $\mathbf{p}$. One can also compare different parts of the spectrum, such as looking for an excess of low-frequency alleles or high-frequency derived alleles relative to this standard. Nielsen et al. use this approach for their **MWU-low** and **MWU-high** tests, where MWU stands for the Mann-Whitney U test (a common nonparametric test for comparing two groups, e.g., Conover 1999). One major reservation with these nonparametric approaches is the choice of the reference set of sites for the neutral background spectrum. Even if these are neutral, local effects such as differences in the mutation rates (and hence θ) and more subtle local effects such as differences in the background recombination rates that influence the levels of standing variation (Chapter 8) can result in the target sites (even if strictly neutral) differing from the distribution at reference sites.

Support via a Peponderance of Evidence

A common strategy in the literature to support a claim for selection is to show that a number of different tests are all highly significant. To this aim, a number of authors have proposed compound tests based on the joint distribution of two (or more) summary statistics of selection (Zeng et al. 2006, 2007; Pavlidis et al. 2010; Lin et al. 2011). Others have advocated **meta-analysis**, combining the significance values over multiple tests (Appendix 4). This can be accomplished several ways. Utsumomiya et al. (2103) use Stouffer's Z score (Equation A4.2) to combine p values from different tests at each region. Randhawa et al. (2014) used a slightly different approach where, for a given test, a standardized rank score $R/(n+1)$ is computed for each of the n SNPs (R the rank, from lowest to highest, of the p value of the test). The resulting scores (for a given test) for each SNP range from $1/(n+1)$ to $1-1/(n+1)$, which are then probit-transformed (Equation 14.2) and averaged over all of the tests to obtain a Z score for each particular SNP. Grossman et al. (2010, 2013) constructed a likelihood based on a composite of multiple signals (their **composite of multiple signals**, or **CMS** approach). While composite tests likely do not return proper significance values, they can still have significant utility. Grossman et al. noted that CMS offers a significantly narrow region for a selected site, increasing resolution up to 100 fold. Further, given that different tests are optimal over different time scales during a sweep (Table 8.2), a composite test offers to possibility of having power over a larger time range.

While seemingly logical, there are a number of subtle concerns with this approach. First, although the tests reviewed in Table 9.1 highlight different features of the site-frequency spectrum, they are generally still correlated (Fu 1997). Hence, when choosing a region because of an abnormal D value, one might expect to find other abnormal site frequency values as well, even if the region is neutral. This also holds for other types of tests, such as parametric sweep-based approaches (CLR, Sweepfinder) and haplotype-based tests (discussed below). When a region is ascertained by having an usual test statistic value, this skews the distribution of other tests as well. For example, a region of low recombination can amplify random departures from the neutral equilibrium model. This point was stressed by O'Reilly et al. (2008), who noted that significant selection tests in humans are disproportionately found in regions of low recombination. On one hand, this makes sense, as regions with low c values

are expected to have stronger signals from sweeps (Chapter 8). However, regions of low recombination create longer correlations among adjacent sites as well, so than an extreme realization of a local coalescence from the neutral equilibrium model extends over a larger region. This is one reason to be very cautious of tests that look for localized runs of a particular statistic. One such example is Carlson's et al. (2005) **Continuous Regions of Tajima's D Reduction** (or **CRTR**) test. Such a run is expected under a sweep, but it is also expected around an unlikely (but not exceptional) neutral genealogy in a region of low recombination.

The strongest preponderance of evidence support comes from *completely independent* tests, such as site-frequency data from one population coupled with an abnormal F_{ST} value for that site between populations. Even in these cases, the skeptic can suggest that most of the signal is coming from an unusual event in a single population, but an event that could be an outlier from a neutral drift process. For example, if one catches a surfing allele in one population, it has the potential for a number of selection-like signals and will also give a large F_{ST} value relative to other populations where it has not surfed.

Recombination Makes Site-Frequency Tests Conservative

A final comment on frequency spectrum tests is that, ignoring demographic concerns, they are likely conservative in many settings. In particular, Wall (1999) notes that site frequency spectrum tests all have the assumption of no recombination within the region of interest. While recombination does not bias the expected values for various statistics, it does *reduce* their variances (Wall 1999; Rozas et al. 1999). This occurs because the observed values represent the average across several correlated genealogies (Depaulis et al. 2003). As a result, when recombination *does* occur within a region, tests are *conservative*, with the actual p value less than the zero-recombination values tabulated by the original authors of the various tests. As a result of the conservative nature of tests under recombination, they are often significantly *underpowered*, using more stringent critical values than necessary. Wall found this effect to be significant when the rate of recombination is on the order of the total regional mutation rate, as is often the case (Table 4.1). Coalescent simulations allowing for recombination can significantly improve the power of tests by obtaining more accurate p values. As discussed in Chapter 4, the four gamete test (Hudson and Kaplan 1985) can be used to detect recombination in a sample, and their R_M statistic estimates the minimal number of recombinants in the sample, which can then be incorporated into an appropriate coalescent simulation (e.g., Depaulis et al. 2005).

HAPLOTYPE-BASED TESTS

While powerful in some settings, the site-frequency spectrum does not contain all the information in a sample of sequences, as it ignores their **haplotype structure** — the nature of the association (linkage disequilibrium, LD) among segregating sites. Put another way, knowledge of the number of segregating sites S only places rough bounds on the number k of haplotypes, in that if a region is polymorphic ($S \geq 1$), then for a sample of size n, $2 \leq k \leq \min(S, n)$. **Strong haplotype structure** occurs when there is a deficiency in the number of haplotypes (roughly analogous to S under infinite-sites) and/or **haplotype diversity** H (the probability that two random haplotypes from the sample for different, analogous to Π under infinite-sites), and/or an excess of high-frequency haplotypes (roughly analogous to Fay and Wu's H test). All of these are signatures of excessive LD within a region, with the number of haplotypes being *underdispersed* relate to the number of polymorphic sites. Such signatures are generated by any process that gives the coalescent long internal branches (relative to the equilibrium neutral model), such as a partial sweep (the favorable allele is not yet fixed), recovery from a moderate bottleneck, balancing selection, or population struc-

ture. Conversely, we can have the opposite (*overdispersion* of haplotypes), with an excess of haplotypes, excess haplotype diversity, and an excess of rare-frequency haplotypes (relative to S). Such signals are generated when the coalescent genealogy is star-like, as would occur near the conclusion of a hard sweep or recovery from an extreme population bottleneck. However, in these settings LD summary statistics typically have low power, as S is small (most variation is removed), so that while haplotype overdispersion occurs, the signal is often weak.

Table 9.2. Haplotype-based signals of positive selection under different types of sweeps.

Completed/Nearly Completed Hard Sweep
Overdispersion of Haplotype structure relative to S
Excess number of haplotypes
Excess haplotype diversity
Excess of high-frequency haplotypes
LD structure
High LD on either side of selected site, little across site
Partial Sweep/Recent Balancing Selection
Strong Haplotype structure
Deficiency in number of haplotypes
Deficiency in haplotype diversity
Excess of low-frequency haplotypes
LD structure
Alleles with long haplotypes at excessive frequencies
Allele age
Alleles with long haplotypes at excessive frequencies
Soft Sweep
Moderate Haplotype structure
A few dominant haplotypes
LD structure
High pairwise LD across entire region

Another classic LD signature of ongoing selection are **long haplotypes**, regions of LD far longer than expected given the observed frequency of an allele (recall from Chapter 2 that higher frequency alleles are *older* under neutrality, and hence have experienced more recombination). Finally, there is a characteristic LD structure around the selected site following a completed sweep. For a soft sweep, an excess of pairwise LD is expected *throughout* this region, even when the site-frequency spectrum shows little change (Pennings and Hermisson 2006b). For a hard sweep, a different pattern is seen with strong LD between sites on the same side of the sweep, but no LD *across* the sites (Figure 8.7).

Haplotype-based tests focusing on these different signals of LD/ haplotype structure fall into three categories (Table 9.2). The first are those based on the infinite-alleles model, such as the number of unique haplotypes and their frequency distribution within a sample. These are the analogs of site-frequency tests but now under an infinite-alleles framework, focusing on haplotypes instead of sites. The second are tests based on summary statistics of all pairwise linkage disequilibrium over sites within a region. The final class essentially uses linkage information to determine the age of an allele, either by looking at sequence variation within a haplotype (such as variability at tightly linked STRs) or by the decay of LD as one moves away from a core sequence.

Finally, it is worth stressing that haplotype (and LD) structure can provide signals of selection that are missed by site-frequency and hard-sweep tests, and thus offer more

power in some settings, especially for the detection of partial and soft sweeps. Age-of-allele tests (particularly in the form of detecting long haplotypes) are perhaps the most powerful approach for detecting an *ongoing* sweep, but usually have little to no power once the sweep is close to completion. Conversely, tests based on pairwise LD summary statistics offer significant power (albeit over a very short time window) for the detection of a just-completed sweeps.

Example 9.11. Hudson et al. (1994) used a sample of 41 homozygous lines (making haplotypes easy to score) of *Drosophila melanogaster* from California and Spain to survey variation at the superoxide dismutate (*Sod*) gene. For these data, neither Tajima's D or Fu and Li's D^* were significant. However, the haplotype data told a very different story. They observed 19 slow and 22 fast alleles, as judged by isozyme mobility. Using these to define two allelic classes, they found at all 19 slow alleles were identical in sequence through a 1410-bp region surrounding the fast/slow site, while the 22 fast alleles consisted of 10 different haplotypes. They used coalescence simulations conditional on the observed number of segregating sites and sample size to show that this is a significant decrease in variation of slow haplotypes relative to their frequency, suggesting the slow allele has experienced a recent, and rapid, expansion, as might occur under positive selection. Other *Drosophila* examples where haplotype-based tests gave a strong signal, but site-frequency tests were not significant, include Kirby and Stephan (1995), who found very strong haplotype structure at the *white* locus, but nonsignificant D and D^* tests. Andolfatto et al. (1999) examined a 1.4 kb region spanning the breakpoint of a naturally-occurring chromosome inversion, also finding highly significant haplotype structure but a nonsignificant D value. Finally, Rozas et al. (2001) examined a 1.3 kb region around the *rp49* gene in *D. simulans*, which encodes ribosomal protein 49. Tajima's D, Fu and Li's D and F (the polarized versions of their D^* and F^*, based on the number of derived singletons), and Fay and Wu's H were all nonsignificant over three populations (Montblanc and Maputo plus a composite of both) while a number of measures of haplotype structure (diversity, number of haplotypes, most frequent haplotype) were all significant in most populations.

Defining and Inferring Haplotypes

If one considers a sufficiently large stretch of DNA, every sequence is a unique haplotype, so just how are haplotypes defined? The question depends on both the test being used and the feature(s) of linkage disequilibrium of interest. If interested in number and diversity of haplotypes in an infinite-alleles framework, the unit of analysis is a sufficiently small region, ideally with no recombination observed in the sample. The four gamete test of Hudson and Kaplan (1985) can be used to detect for recombination in the sample (Chapter 4), helping to define the size of a region (for example, by setting the size of a sliding window moving through a larger region). Practically, one may be constrained to find regions with sufficient haplotype diversity for analysis given either the marker density and/or background levels of variation, so that small amounts of recombination may appear in the sample. For tests based on the average pair-wise disequilibrium among all sites within a region, one actually wants some (but not too much) recombination. Finally, tests based on long haplotypes require a **core haplotype** (a set of a few tightly-linked SNPs) showing no recombination to provide distinct allelic classes, with the disequilibrium patterns within each class (i.e., as one moves away from the core) generating the tests. Again, recombination (outside of the core) is critical to these tests.

How are haplotypes determined/inferred? In the simplest case, one has haploid sequence data, which can include X chromosome data from males (and Z chromosome data

from species with heterogametic females) and mitochondrial/chloroplast sequences. One can also have effectively haploid data, such as sequences from fully inbred lines (Example 9.11). Usually, however, haplotypes have to be *inferred* from sequence data. Ideally, one has *trio* data — both parents and their offspring. In such settings, since parent and offspring data are not independent, they cannot be pooled to provide a random sample from the population of interest. The result is that the effective number of sequences in the analysis is less than the number observed. More generally, for unrelated individuals from a population, haplotypes are inferred from unphased data by a variety of methods (reviewed in Stephens and Scheet 2005). The most popular program is PHASE and its descendants (Stephens et al. 2001; Stephens and Scheet 2005). Surprisingly little discussion/analysis appears in the literature as to whether these reconstructions are biased by selection or other demographic departures. Given this concern, it is not unreasonable for one to feel a little uneasy when using inferred haplotypes in tests of selection.

The Ewens-Watterson Test

Ewens (1972) and Watterson (1977, 1978) proposed the first formal tests of selection using molecular data, comparing the fit of the observed allele-frequency spectrum with the conditional distribution given the observed number of alleles (Equation 2.33a). Ewens suggested using the summary statistic

$$I = -\sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \ln\left(\frac{n_i}{n}\right) \tag{9.30a}$$

which is calculated using only the non-zero values of n_i (the number of alleles present as exactly i copies in the sample). His motivation for this statistic was as a general measure of spread (information) in the data. Watterson (1977, 1978) showed that using the sample homozygosity

$$h = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \tag{9.30b}$$

was a better choice for improved power to detect departures under weak-overdominance (the selection model *du jour* of the time). Comparing the statistic given by Equation 9.30b with its value under the equilibrium neutral model is known as **Watterson's test** or the **homozygosity test**. Watterson proposed to assess significance by taking a large number of draws from Equation 2.33a (using the observed number k of alleles in the sample) to generate a null distribution from values of h to compare against its value in the original sample. The same approach can also be used for Ewens' statistic. Advances in computing speed lead Slatkin (1994, 1996) to propose an **exact Watterson test**, wherein one computes all possible h values over the set given by the constraint in Equation 2.32 (this same approach is the basis for Fisher's exact test for contingency tables). The resulting value P_h is computed as

$$P_h = \sum_{\eta^* \text{ such that } h(\eta^*) \leq h} \Pr(n_1, n_2, \cdots, n_k \,|\, n,\, k) \tag{9.31a}$$

namely, that the sum is over all those *configurations* $\eta^* = (n_1, \cdots n_k)$, constrained by Equation 2.32, that give a value of Equation 9.30a that is the same, or smaller than, the observed value h in the sample. Slatkin also suggested a second exact test, wherein one computes the probability over all possible configurations directly,

$$P_E = \sum_{\eta^* \text{ such that } \Pr(\eta^* \,|\, n,k) \leq \Pr(\eta \,|\, n,k)} \Pr(n_1, n_2, \cdots, n_k \,|\, n, k) \tag{9.31b}$$

where η is the observed frequency spectrum. The difference here is that the sum is over a different set of η^*. In Equation 9.31a it is over those η^* that given smaller homozygosity values, while in Equation 9.31b it is over those η^* values that given smaller probabilities than the observed frequency spectrum. Slaktin found that the resulting p values for both tests are very similar for small n, but can be rather different for large n. Provided no recombination within the region of analysis, one can simply replace allele by haplotype, so that all of the above tests can be applied to haplotype data.

Other Infinite-Allele Tests: Conditioning on $\widehat{\theta}$

Watterson-type tests all use the *conditional* allele-frequency spectrum, where the observed number of alleles k is used in Equation 2.33a to generate the null distribution. What about tests based on k itself? These use the sampling distributions given by either Equations 2.30a or 2.33b, and require an estimate of θ. Fu (1996, 1997) used this approach to test whether a sample contains too many, or too few, alleles (haplotypes) relative to the neutral equilibrium model. His **W test** (1996) uses Ewens' sampling formula (Equation 2.30a) with θ replaced by the Watterson estimator $\widehat{\theta}_S$ (Equation 4.3a), and returns the probability of seeing k (or fewer) alleles in the sample as

$$W = \Pr(K \leq k) = \sum_{i=1}^{k} \Pr(K = i \,|\, \widehat{\theta}_S, n) = \sum_{i=1}^{k} \frac{S_n^i\,(\widehat{\theta}_S)^i}{S_n(\widehat{\theta}_S)} \tag{9.32}$$

where S_n^i is the coefficient on θ_S^i in the polynominal

$$S_n(\,\widehat{\theta}_S) = \widehat{\theta}_S\,(\,\widehat{\theta}_S + 1)(\,\widehat{\theta}_S + 2)\cdots(\,\widehat{\theta}_S + n - 1)$$

This is a test for a *deficiency of rare alleles/haplotypes*, and hence is *one-sided*. Fu showed that the W test is more powerful that Tajima's D or Fu and Li's D^* and F^* tests (Table 9.1) for detecting signals of balancing selection or a structured population. Indeed, Strobeck (1987) proposed essentially the same test as Fu (using $\widehat{\theta}_\Pi$ in place of $\widehat{\theta}_S$) as method for detecting population structure, rather than selection.

Fu's $\mathbf{F}_S$ **test** (1997) is the compliment of W, testing for an *excess of rare alleles/haplotypes*. It starts by computing the probability of seeing k *or more* alleles/haplotypes in a sample,

$$S' = \Pr(K \geq k) = \sum_{i=k}^{n} \frac{S_n^i\,(\widehat{\theta}_\Pi)^i}{S_n(\widehat{\theta}_\Pi)} \tag{9.33a}$$

but now using $\widehat{\theta}_\Pi$, the estimator of θ based on average number of pairwise differences (which is more sensitive to sites with intermediate allele frequencies). Fu notes that S' is not an optimal test statistic because its critical values are often too close to zero. Because of this, the test uses the logistic transformation,

$$F_S = \ln\left(\frac{S'}{1 - S'}\right) \tag{9.33b}$$

As with W, this is also a one-sided test. F_S is negative when there is an excess of rare alleles/haplotypes (as would occur with a selective sweep or population expansion), with a sufficiently large negative value being evidence for selection/expansion. Fu (1997) showed that F_S is more powerful that Tajima's D and Fu and Li's D^* and F^* tests for detecting selective sweeps or population expansion following a bottleneck.

Other Infinite-Allele Tests: Conditioning on S

While elegant in using exact results from the allele-frequency sampling distributions, the above tests based on excessive k values *do not* return exact p values, as there is error in estimating θ, making Equations 2.30a and 2.33b only approximations. Hudson et al. (1994) and Depaulis and Veuille (1998) note that while θ is unknown, the number of segregating sites S is directly observed. Hence, one can generate coalescence genealogies and then randomly place the S segregating sites over these (at a rate proportional to the branch lengths within the coalescent), generating a distribution of haplotypes in the final sample. This procedure generates draws under the neutral equilibrium model conditioned on the observed number of segregating sites. In effect, these tests examine the sequence data from *both* the infinite-alleles and infinite-sites perspectives. While the number of alleles k is a sufficient statistic under the infinite-alleles model, S *is not* — conditioning on S, the distribution still has a dependence on θ, although this is often weak (Griffiths 1982; Markovtsova et al. 2001; Wall and Hudson 2001; Depaulis et al. 2001; Depaulis et al. 2005; Innan et al. 2005).

Hudson et al. (1995) pioneered this conditioning approach with their **haplotype test**, also referred to as the **haplotype partition** (**HP**) or Hudson's haplotype test (**HHT**). Its initial form (motivated by the observations discussed in Example 9.11) was a rather open-ended test: given a sample of n sequences with S segregating sites, what is the probability of observing i sequences with j or fewer alleles. This is akin to Slatkin's exact test (Equation 9.31b), in asking how likely is a given configuration. The difference here is that Slatkin's test conditions on the observed number of alleles k, while Hudson's test conditions on S. Hudson's test is typically implemented by asking if there is an excess of the most frequent haplotype (Depaulis et al. 2005; Innan et al. 2005). Suppose there are n samples and the highest frequency haplotype occurs j times. Hudson's test is

$$\Pr(n_i \geq 1 \text{ for } i \geq j \,|\, S)$$

where, as above, n_i is the number of alleles presenting in exacty i copies in the sample. This is just the probability that the most common haplotype is found j or more times in the sample. Note that is can also be directly computed conditional on k using the appropriate partition in Equation 9.31b. Andolfatto et al. (1999) extend this test by allowing a sliding window of variable size to transverse the region of interest, developing a correction for multiple tests (the different windows). Again, hypothesis testing is done using the null generated from a coalescent with S segregating sites.

Depaulis and Veuille (1998) also use conditioning on S in their number of haplotypes **K** and the diversity of haplotypes **H tests**, where

$$H = 1 - \sum_{i=1}^{K} p_i^2, \qquad p_i = \text{ frequency of } i\text{th haplotype} \tag{9.34a}$$

is a measure of haplotype diversity. Note that the range on H is

$$\frac{2(n-1)}{n^2} \leq H \leq 1 - \frac{1}{n}, \tag{9.34b}$$

with the lower range set by the sample consisting of just two haplotypes, one with $n-1$ copies, the other a singleton ($n_1 = 1, n_{k-1} = n$), while the upper range is set by all of the haplotypes present as singletons ($n_1 = n$). Critical values for these statistic (conditioned on n and S) were generated by coalescent simulations and are tabulated by Depaulis and Veuille. Finally, Innan et al. (2005) propose a **haplotype configuration test** (**HCT**), based on

the configuration of the haplotype (allele) frequency spectrum. Again, this is as in Stalkin's exact test (Equation 9.31b), but now the conditioning is on S (as opposed to k),

$$P_E = \sum_{\eta^* \text{ such that } \Pr(\eta^* \mid n,k) \leq \Pr(\eta \mid n,S)} \Pr(n_1, n_2, \cdots, n_k \mid S) \tag{9.35}$$

These probabilties can easily be generated using the constant S coalescent simulation approach discussed above. Depaulis et al. (2005) and Innan et al. (2005) discuss all of these haplotype-frequency spectrum approaches in greater detail, while power issues are discussed by Ramos-Onsins and Rozas (2002) and Depaulis et al. (2003, 2005).

Other approaches based on haplotype number have also been suggested, although more as heuristics and summary statistics than as formal tests. Przeworski (2002) suggests standardizing the number of haplotypes by the number of segregating sites, using $k' = k/(S+1)$. The smaller k', the more the LD, as specific combinations of segregating sites are locked into a small number of haplotypes. Przeworski notes that while k' tends to decrease as one approaches a selected site *during* an ongoing hard sweep, k' can actually be greater than expected under the equilibrium neutral model *after* a sweep is completed. This occurs because high-frequency variants are fixed and new mutations arise, most of which are singletons and hence form a large collection of unique haplotypes. She notes that excluding singletons when computing both k and S gives k' much more stability, being sharply decreased at the completion of a hard sweep, and then increasing back to its neutral expectation in a relatively short time ($\sim N_e$ generations) after the sweep.

Finally, what is the effect of recombination on these various infinite-alleles-based tests? Recall that with site-frequency spectrum tests, recombination has a uniform effect, making all tests conservative, and hence underpowered. This is *not* the case for haplotype-based tests. Recombination creates new alleles, which inflates the number of haplotypes, the diversity of haplotypes, the number of rare haplotypes, and decreases the frequencies of the most common haplotypes. Thus, depending on the test, recombination can either make it conservative or it can falsely deflate the p value (Rozas et al. 2001; Wall and Hudson 2001; Depaulis et al. 2001; Depaulis et al. 2005). Tests for low number of haplotypes, low haplotype diversity, and excessive high frequency of the most common haplotype are all conservative under recombination. Conversely, tests for an excessive of rare haplotypes, excess haplotype diversity and excess haplotype number, all have their p values reduced by recombination (making these test anti-conservative). Incorporating recombination into the coalescent simulations greatly improves the power for conservative tests and creates more accurate p values for anti-conservative tests. However, using incorrect values for recombination can significantly bias the tests. Under the coalescent framework, recombination is measured by $R = 4N_e c$, the population-scaled rate. Depaulis et al. (2005) offer the following suggestion. First, estimate Hudson's minimal number of recombinants R_M in the sample from the four gamete test (Chapter 4), then choose R in the coalescent simulations as the value which gives five percent (or less) of the samples showing R_M or more recombinations.

Pairwise Disequilibrium Tests: Kelly's Z_{nS} and Kim and Nielsen's ω_{max}

Positive selection can produce two very different patterns of disequilibrium around a site. For soft and partial sweeps, there an excess of LD throughout a region, while for hard sweeps, LD is found on either side of the selected region, but not through it (Chapter 8). To test for these patterns, two different averages of pairwise disequilibrium within a region have been proposed. Both start with Hill and Robertson's (1968) scaled measure of the disequilibrium expressed as a correlation coefficient,

$$r_{ij}^2 = \frac{D_{ij}^2}{p_i(1-p_i)q_i(1-q_i)} \tag{9.36a}$$

where D_{ij} is the disequilibrium between sites i and j, p_i is the frequency of the major allele at site one, q_i at site two. For a general measure of the average amount of LD throughout a region with S segregating sites, Kelly (1997) proposed the average over all of the pairwise squared correlations,

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{ij}^2 \tag{9.36b}$$

This is often computed over windows of various sizes, so that (for example) Z_{n5} and Z_{n8} denote values for windows with five and eight segregating sites. Kelly showed that values of Z_{nS} are largely determined by the final coalescent time in the sample (the time for the last two lineages to coalesce into the ancestral lineage for the entire sample). The longer this time, the larger the value of Z_{nS}. It is smallest under a star genealogy, as most of the coalescence events occurred at roughly the same time. Thus, a small value of Z_{nS} is consistent with a hard sweep or an extreme bottleneck. In such cases, there is usually reduced amount of site polymorphism, which in turn reduces the power of Z_{nS}. Conversely, with a partial or soft sweep, Z_{nS} increases. Critical values of Z_{nS} are determined by coalescent simulations conditional on S (as discussed above). When recombination is ignored, the one-sided test of excessive Z_{nS} values is conservative (as recombination lowers Z_{nS}), while the test of a deficiency in Z_{nS} values is anti-conservative. Pennings and Hermisson (2006b) suggest that LD-based tests, such as Z_{nS}, may have the most power to detect a *very recent* soft sweep.

Kim and Nielsen (2004) propose a different measure of pairwise disequilibrium designed for the particular pattern of LD following a hard sweep (Figure 8.7), with LD on either side, but not *across* the site (this LD signal dissipates rather quickly, roughly on the order $\sim 0.1N_e$ generations immediately following a sweep, Przeworski 2002; Jensen et al. 2007). Based on this disjoint LD pattern, they proposed a test statistic ω comparing LD within versus across the left and right sides of a sliding window. Suppose there are ℓ sites $(1, \cdots, \ell)$ on the left (L) side of the putative selected region and $S - \ell$ (sites $\ell + 1, \cdots, S$) on the right (R). Define

$$\omega = C_{S,\ell} \frac{\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in R} r_{ij}^2}{\sum_{i \in L, j \in R} r_{ij}^2}, \qquad C_{S,\ell} = \frac{1/(\ell(S-\ell)}{\binom{\ell}{2} + \binom{S-\ell}{2}} \tag{9.37}$$

where the combinatorial term $C_{S,\ell}$ is a function of the number of sites contrasted over the three comparisons (within L, within R, between L and R). Under the distinct signal of LD from a completed hard sweep, one expects a large value for the numerator (strong LD within either side) and small denominator value (little LD across sides) around a sweep, giving a large value of ω. Since r^2 is sensitive to small allele frequencies, polymorphic sites showing singletons are best ignored when computing ω. Kim and Nielsen's test statistic ω_{max} is the maximum value of Equation 9.37, a function of both window size and window position, with critical values determined via a coalescent simulation. Pavlidis et al. (2010) offer some improvements which allow this test to be scaled up to scan an entire genome efficiently.

Jensen et al. (2007) found that this approach had promise for a very vexing situation, detecting sweeps that start in nonequilibrium populations. They found that population structure does not appear able to cause this LD pattern, but that a very strong bottleneck (one percent of the population surviving) can give modest ω_{max} scores. The effects of recombination on this test (i.e., recombination within the test window) are a bit unclear. On one hand, by reducing r_{ij}^2 values within each side, recombination should be conservative. Conversely, by also reducing r_{ij}^2 values between sides under the null, this can inflate p. Kim and Nielsen found in their simulation studies that while assuming the incorrect value for the scaled recombination rate R decreases power, ω_{max} was more robust to incorrect R values than Z_{nS}.

To summarize these two test statistics, Z_{nS} has power to detect partial sweeps and recently completed soft sweeps, but poor power for recently completed hard sweeps. Conversely, ω_{max} has good power to detect a recently completed hard sweep, and also robustness against many demographic concerns, although a serve bottleneck can also generate modest ω_{max} scores. While most LD/haplotype tests have more power *during* a sweep, these two statistics have power during a very short time window ($\sim 0.1N_e$ generations) immediately *following* a successful sweep.

Contrasting Frequency and Intra-allelic Variation Estimates of Haplotype Age

Assuming an allele arose as a single mutation, it initially was on a single haplotype background in complete LD with tightly linked markers. As the allele ages, the fraction of copies associated with the original background decays through recombination. Likewise, new mutations at tighty-linked sites arise, with the number of segregating SNPs and copy number variation at STRs both increasing with age. Hence, the diversity of haplotypes associated with that allele provides information about its age (reviewed by Slatkin and Rannala 1997, 2000). A common approach (especially in human genetics) is to contrast one (or more) of these **intra-allelic variation** estimates of age with the estimate of age based on allele frequency (e.g., Example 9.12). Figure 2.1 shows that, under the equilibrium neutral model, *a common allele is an old allele*. As such, only a small fraction of initial haplotypes should remain intact, and there should be ample variation associated with the haplotypes in which the allele is embedded. While a few formal tests have been proposed as to whether there is too little (directional selection) or too much (balancing selection) intra-allelic variation (Slatkin 2000, 2008; Slatkin and Bertorelle 2001), our goal here is to review some of these age estimators, as discrepancies between them are often offered as evidence supporting selection.

Example 9.12. The mutation *CCR5-δ32* destroys the *CCR5* receptor, which is used by the HIV virus to enter the call, leading to significant resistance against HIV infection. This deletion occurs at frequencies up to 14% in Eurasia, but is absent in Africans, Native Americans and East Asians. Assuming a frequency of $x = 0.10$ and an effective population size $N_e = 5000$ for Caucasians, Stephens et al. (1998) used Equation 2.12 to estimate the age of this allele as

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

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

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

Stephens et al. took these great disparities between age estimates as an indication of strong selection on the δ mutation, generating much a higher frequency that expected from its age under a pure drift model.

While Equation 2.12 gave a simple expression for the expected age of an allele as a function of its frequency, it is a bit misleading, due to its very large variance. A better

estimator follows from Slatkin and Rannala (2000), based on their approximation for the probability distribution of age as a function of frequency under the equilibrium neutral model (also see Watterson 1976; Griffiths and Tavare 1998). Letting τ denote time in $2N_e$ generations, then for an allele whose frequency is p in a sample of size n, the probability that the true age t is less than τ is given by

$$\Pr(t \leq \tau) \simeq (1-p)^{-1+n/(1+n\tau/2)} \tag{9.38a}$$

Taking the derivative with respect to τ recovers the probability distribution, and hence the likelihood function (LW Appendix 4). The resulting distribution is very skewed, with a long heavy tail towards increased age. As a result, the mode of the distribution (the MLE, LW Appendix 4) is less, and usually significantly so, than mean. In particular, the MLE (in units of $2N_e$) for t becomes

$$\text{MLE}(t) = -\ln(1-p) - \frac{2}{n} \tag{9.38b}$$

Unlike Equation 2.12, the maximum likelihood estimator accounts for the sample size n used to estimate p. A $100(1-\alpha)\%$ confidence interval for age is given by $(\tau_{\alpha/2}, \tau_{1-\alpha/2})$ where τ_x is the solution of

$$\Pr(t \leq \tau_x) = (1-p)^{-1+n/(1+n\tau_x/2)} = x \tag{9.38c}$$

For large n, $n/(1+n\tau_x/2) \simeq 2/\tau$, giving a (large n) solution to Equation 9.38c as

$$\tau_x \simeq \frac{2}{1+\ln(x)/\ln(1-p)} \tag{9.38d}$$

For small n and/or small x, Equation 9.38d serves as a starting point for numerically solving Equation 9.38c. The weakness of this estimator is that it is *extremely* sensitive to demography (Slaktin 2000; Slaktin and Rannala 2000).

Example 9.13. Recall Example 9.12, where we showed that the estimated age of the *CCR5-δ32* mutation (roughly 5100 generations from Equation 2.12) was incompatible with an independent estimate of its age based on intra-allelic variation (roughly 28 generations). Is this conclusion changed if Equation 9.38 is used instead of Equation 2.12? In Example 9.12 we used $p = 0.10$ and $N_e = 5000$. Taking $n = 4000$ as the approximate sample size of Europeans used to estimate the allele frequency, Equation 9.38b gives the MLE as

$$\text{MLE}(t) = -2N_e\left(\ln(1-p) + \frac{2}{n}\right) = -10,000\left(\ln(0.9) + \frac{2}{4000}\right) = 1048 \text{ generations}$$

which, while considerably smaller than the estimate based on Equation 9.1, is still not compatible with the allelic-variation estimate. Perhaps, however, the variation-based estimator is still within the 95% confidence interval. Values of $\tau_{0.025}$ and $\tau_{0.975}$ are obtained by solving Equation 9.38c. Using the approximation offered by Equation 9.38d, gives starting values of 0.055 and 1.61. Plotting Equation 9.38c around these values gives the exact answers, $\tau_{0.025} = 0.067$ and $\tau_{0.975} = 1.61$, which translates into 670 to 16,100 generations ($2N_e\tau$). Despite this very wide range, the lower value still greatly exceeds 28 generations, so that the frequency of this allele is not consistent with an equilibrium neutral model.

As Example 9.12 shows, potential evidence for selection is offered when a frequency-based estimator of age is too large relative to an estimate based on within-haplotype variation.

We briefly consider three such comparisons: The persistence of disequilibrium with a marker at known recombination fraction c, the number of segregating sites, and the variance at STRs. The last two are assumed to be scored in a sufficiently close region around the allele allowing recombination to be ignored.

Estimating age from the persistence of LD between the allele and a linked marker (Example 9.12) was first proposed by Serre et al. (1990). Let A denote a marker allele at a site closely-linked to the target allele we are trying to age. The marker allele is typically chosen because the target allele is over-represented on A-bearing haplotypes, suggesting the initial mutation arose on such a background. Let $x(t)$ denote the current frequency of haplotype A given the target allele is present. Assuming a single origin of the mutation and that it occurred on an A background, $x(0) = 1$. Further assuming that the frequency of haplotypes carrying A in the ancestral population is y (assumed to be the same as the present-day value), the decay in LD by recombination is

$$x(t) - y = (1-c)^t(x(0) - y) = (1-c)^t(1-y) \tag{9.39a}$$

solving for t gives

$$t = \frac{1}{\ln(1-c)} \ln\left(\frac{x-y}{1-y}\right) \tag{9.39b}$$

This approach is very closed related to LD mapping (LW Chapter 14), but in the latter, one assumes that t is known and solves for c in Equation 9.39a. Equation 9.39b is a slightly more accurate approach that we used in Example 9.13, as it accounts for the population-wide frequency of the marker allele A. Risch et al. (1995) and McPeek and Strahs (1999) discusses how multiple linked loci are simultaneously used to estimate t. The delicate issue in applying Equation 9.39b is that it is sensitive to c. Using a value of c less than the true value overestimates the age, while age is underestimated if too large a c value is used.

A second approach using intra-allelic diversity to estimate age is based on the number of segregating sites in a tightly-linked region around the target allele (Example 8.4). Under a hard selective sweep (and hence an approximate star genealogy for the sample sequences), Equation 8.14 provides a rough estimate of age as $t = S/\mu n$, where n is the sample size and S the number of segregating sites. Alternatively, one could use the fact that S follows a Poisson distribution to obtain both a ML estimator and confidence intervals.

Example 9.14. Meiklejohn et al. (2004) observed 3 segregating sites in a population sample of n = 26 sequences from a region spanning the *janB* allele in *D. simulans*. They assumed a total mutation rate for this region of $\mu = 1.73 \times 10^{-5}$ per year. For a mutation arising t generations ago, the expected number of mutations on a single sampled allele is $\mu\, t$. Under the approximation that the coalescence has a star-shaped structure (very little shared coalescence time among sampled alleles), each of the n lineages in the sample is essentially independent of the others, giving the expected number of segregating sites in a sample of n alleles as $\lambda = n\,\mu\, t$, with the distribution of observed S values within the sample following a Poisson distribution with parameter λ. The resulting MLE for t given this data is simply the value that maximizes the likelihood given $S = 3$,

$$\Pr(S = 3 \,|\, t\,\mu) = \frac{\lambda^3}{3!}\, \exp(-\lambda) \quad \text{where} \quad \lambda = n\,\mu\, t = 0.0004498\, t$$

Plotting the above expression as a function of t shows the maximum (and hence the MLE) occurs at t = 6667 years (which is essentially identical to the estimate given by Equation 8.14). The 95% confidence interval follows from the Possion, by finding those values $t_{0.025}$ and $t_{0.975}$ such that

$$\Pr(S \leq 3 \,|\, t_{0.025}\,\mu) = 0.025 \quad \text{and} \quad \Pr(S \leq 3 \,|\, t_{0.975}\,\mu) = 0.975$$

where (from the Possion)

$$\Pr(S \leq 3 \,|\, \lambda) = \exp(-\lambda)\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!}\right)$$

Numerically solving these gives $\Pr(S \leq 3 \,|\, \lambda = 8.77) = 0.025$ and $\Pr(S \leq 3 \,|\, \lambda = 1.09) = 0.975$. Since $8.77 = 0.0004498\,t$, $t \sim 19,500$ years. Likewise, the lower limit is ~ 1400 years.

While Examples 8.4 and 9.14 both assume a very defined genealogy, in reality, the shape of the genealogy is unknown and a function of both time t of origin plus the effective population size N_e. As emphasized by Slatkin and Rannala (2000), there is considerable variation in this genealogy even under an equilibrium neutral model which introduces considerable noise into the above estimators. In particular, the assumption of a star genealogy creates a process with too little variation, and hence confidence intervals that are too narrow. Slatkin and Rannala caution that intra-allelic age estimates generally tend to have a downward bias, making alleles appear too young. Given that most contrasts of frequency and intra-allelic variation declare selection when the frequency estimate is much larger than the variation-based estimator, caution is always in order when using this approach.

Finally, the copy number variance at an STR also provides an estimator of age. As above, a new mutation arises on a haplotype carrying a specific allele at a linked STR. As that haplotype spreads and ages, mutations arise at the STR (due to its high mutation rate), generating a variance in copy number at the STR for the haplotype that carries the mutation. Thomas et al. (1998) note that the average squared distance between copy number over all sampled haplotypes and the ancestral haplotype has an expectation of μt, where μ is the microsatellite mutation rate. Much more sophisticated analyses can also be used (e.g., Wilson et al. 2003).

Haplotype Homozygosity and Garud et al.'s $\mathbf{H}_{12}$ and $\mathbf{H}_2$ Tests

Several tests are based on detecting significant differences in the amount of LD associated with alternative alleles at the same site. To proceed, DNA sequences for a region of interest are first partitioned into different allelic classes. In the simplest case (alleles defined by a single SNP), the two alternative nucleotides define two different allelic classes. More generally, different allelic classes can be defined by a core haplotype consisting of set of very tightly linked SNPs that show both no recombination in the sample (Hudson and Kaplan's $R_M = 0$) and are also in complete LD (scaled $|\, D' \,| = 1$).

Haplotype homozygosity (**HH**, **HAPH**) is the probability that two randomly-chosen copies of an allele are identical (homozygous) for all markers within a specified region. For a very small region around the allelic-defining site, this should be close to 100 percent, decaying away as one moves either right or left from the core. The simplest measure of HH is just a modification of the Depaulis and Veuille H statistic (9.34a), where p_i is the fraction of the i haplotype, and

$$H_1 = \sum_i p_i^2 \tag{9.40a}$$

Some variants of this statistic replace p_i^2 by $(p_i + 1/k)^2$ where k is the number of haplotypes (e.g., Kemper et al. 2014). As mentioned in Chapter 8, Garud et al. showed that a simple modification of this statistic gives a test that can detect *both* hard and soft sweeps. Their $\mathbf{H}_{12}$ test statistic combines the two largest haplotype classes into a single one,

$$H_{12} = (p_1 + p_2)^2 + \sum_{i>3} p_i^2 = H_1 + 2p_1p_2 \tag{9.40b}$$

Their logic is that a soft sweep results in not one, but several dominant haplotypes. If the sweep is not too soft, then the first two haplotypes, both presumable harboring the favored allele, will together comprise most of the haplotype variation. In the case of a hard sweep, the second-most frequent haplotype still be sufficiently rare that $H_{12} \sim H_1$. The authors applied this approach to *Drosophila*, looking at windows with a fixed number of SNPs and adjusting for local recombination rate, and using coalescent simulations to generate values under the null of neutrality. The authors considered a second modified HH statistic, namely the homozygosity with the largest class removed

$$H_2 = \sum_{i>1} p_i^2 \tag{9.40c}$$

Under a hard-sweep with its single dominant haplotype, H_2 should be considerably smaller than H_1, while under a soft-sweep the drop-off in value from H_1 to H_2 should be much less dramatic. Based on this observations, the ratio H_2/H_1 forms the basis of a test as to whether a detected sweep is hard or soft.

Long Haplotype Tests

As was the case for considering the detailed the spatial structure around a sweep (as opposed to a simple summary statistic of loss of variation), more information on LD can be extracted by looking at the pattern of HH as one moves away from a core site. While the choice of a cutoff value for HH is rather arbitrary, Sabeti et al. (2002) defined **extended haplotype homozygosity** (**EHH**) as the length of a region around the allele where HH is five percent or greater (Figure 9.6). While alleles with excessive values of EHH are produced by partial sweeps, simply scanning for sites with large EHH values is not a sufficient indicator of selection, as a localized decrease in the recombination rate inflates the EHH value. The formal use of EHH as a selection-detecting statistic thus requires an internal control. Sabeti et al. proposed considering the **relative extended haplotype homozygosity** (**rEHH**) of a particular allele, the ratio of the EHH value for that allele divided by the average EHH values for all other alleles at that site. For allele i,

$$rEHH_i = \frac{EHH_i}{\text{ave}(EHH_j) \text{ for } j \neq i} \tag{9.41}$$

where ave(EHH_j) denotes the average EHH values for all other alleles at this site. For the biallelic case, this is simply the ratio of the EHH values for the two alleles. By contrasting different alleles at the same site, most concerns about local variation in the recombination rates are ameliorated. However, if there are haplotype-specific recombination rates (for example, insertion of a mobile element can reduce local recombination rates, e.g., Macpherson et al. 2008), then this test may be compromised. One consequence of comparing different alleles at a site is that as one allele approaches fixation, the power of the test disappears, as there are too few individuals in the comparison class to produce a meaningful statistic. As a result, the rEHH test has a rather small time window of detection, namely for partial sweeps where the allele has not yet become fixed (a rough rule is that the frequency of the favored allele is 70 percent or less). However, within such a time window, this test is among the most powerful for detecting selection. A large rEHH value, however, is not sufficient, as some rare alleles (potentially being very young) are expected to have large rEHH values. To detect selection, Sabeti et al. plot rEHH values versus allele frequencies and look for outliers. Coalescent simulations are then performed under different neutral models (with different demographic assumptions) to assess significance. Wang et al.'s (2006) **linkage disequilibrium decay** (**LLD**) test is a modification of EHH that does not requiring phasing. It does this

by ignoring some of the information, using only data from individuals who are homozygotes (e.g., A/A and a/a) for the alleles at the target locus. This allows one to quickly score recombinants as one moves away from the core, with the rate of LD decay estimated using a regression approach.

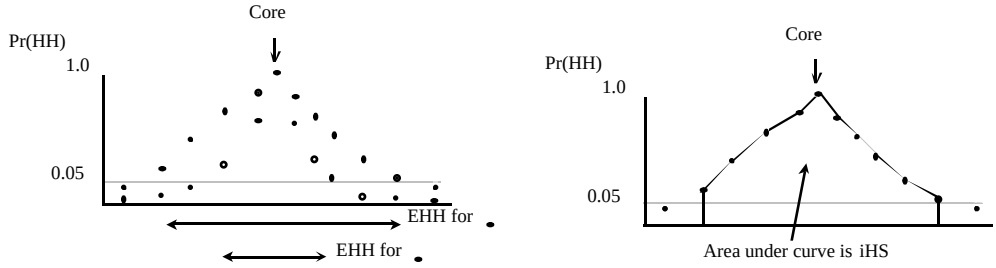
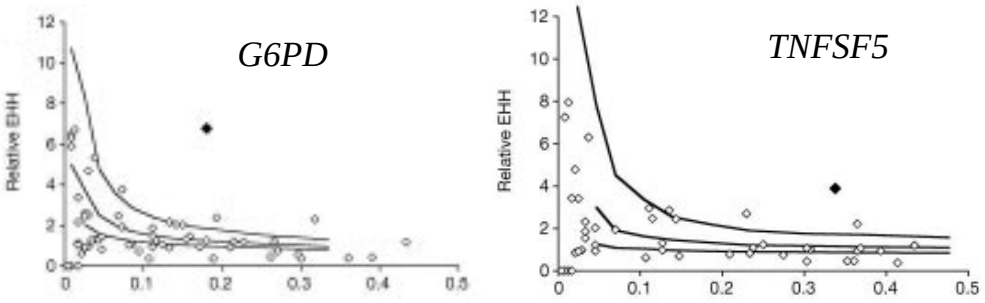


Figure 9.6. The values of haplotype homozygosity (HH) at a series of SNP markers (open and filled circles) around the site (the core) defining the different allelic classes within a region. **Left:** The extended haplotype homozygosity (EHH) for an allelic class is the length of the region around the core where the HH value (the probability that all markers within this region are identical, and so the haplotypes are homozygous) is five percent or greater (above the dotted line). The figure plots HH values for two different allelic classes, with the open and filled circles giving the HH values for these two different alleles at the same SNP markers. The allele corresponding to the filled circles has a larger EHH value. **Right:** A potentially more informative measure is given by the **integrated EHH score** iEHH, the total area under the HH curve over the region spanned by the EHH. For ease of presentation, only the values corresponding to the allele with the larger EHH value (SNPs with filled circles) is plotted.

Example 9.14. As a proof-of-concept of the rEHH method, Sabeti et al. (2002) looked for signatures of selection at two genes, *G6PD* and the *CD40* ligand gene (*TNFSF5*), that carry segregating alleles strongly suspected to increase malaria resistance.



Standard site-frequency tests (Taijma's D, Fu and Li's D^*, Fay and Wu's H) were all not significant. However, site-frequency spectrum signals are largest around the time of fixation of a favorable allele, and much weaker while the allele is still at modest frequency. The figure above displays rEHH versus allele frequency for both the candidate allele (solid diamond) along with values for alleles at other randomly-chosen autosomal loci (open diamonds). Lines indicate the empirical 95th, 75th, and 50th percentages. Simulation results gave similar values, showing both candidate alleles to be extreme outliers. Figure after Sabeti et al. (2002).

Variant tests based on length of shared haplotypes have been proposed by a number of authors (e.g., Toomajian et al. 2003, 2006; Wang et al. 2006; Hanchard et al. 2006). Perhaps the most powerful modification is due to Voight et al. (2006), which extracts more LD information than simply the size of the EHH, and corrects for differences in the local recombination rate and the target allele frequency. Their approach uses polarized data, so let p denote the frequency of the derived (D) SNP and $1-p$ the frequency of the ancestral (A) allele. To extract more information from the plot of HH versus distance from the core site, they compute an **integrated EHH score** (**iHS**), the area under the curve drawn by connecting the adjacent values for the SNPs within the EHH (Figure 9.6). They define the (unstandarized) integrated iHS score as the log of the ratio of the iHS score for the ancestral allele divided by that for the derived allele,

$$iHS_{us} = \ln\left(\frac{iHS_A}{iHS_D}\right) \tag{9.42a}$$

A negative score occurs when the iHS value for the derived allele exceeds the ancestral allele, with the converse when the ancestral allele has the larger iHS score. One can either keep the sign or use the absolute iHS score, $|\,iHS\,|$. Voight et al. standardize the (signed) iHS_{us} score by defining the statistic

$$iHS = \frac{\ln\left(\frac{iHS_A}{iHS_D}\right) - E_p\left[\ln\left(\frac{iHS_A}{iHS_D}\right)\right]}{\mathrm{SD}_p\left[\ln\left(\frac{iHS_A}{iHS_D}\right)\right]} \tag{9.42b}$$

The expectations in Equation 9.42b are taken over all iHS_{us} values for SNPs whose derived allele frequency is p (operationally, within the same bin containing the current p value, e.g., say 0.45 to 0.5). Standardizing the score with respect to p automatically incorporates any relationship between the iHS_{us} score and the allele frequency. The authors note that this approach seems fairly robust to demographic departures from the equilibrium neutral model, especially at extreme values of the standardized score. Despite this, Voight et al. correctly do not assign significance values to individual iHS values, but rather use large (absolute) scores as a screening method for potential sites under selection.

Tests based on comparing the haplotype lengths of alternative alleles lose all power as a favorable allele approaches fixation. However, if the favored allele is only fixed in a single population, a *between* population comparison of haplotype length still has power immediately following fixation. This approach was proposed by both Tang et al. (2007) and Sabeti et al. (2007), and follows the same logic as Equation 9.42a/b. However, instead of contrasting the EHH or iHS score for alternate alleles in the same population, they contrast values for the *same* allele in *different* populations. Sabeti et al. refer to this as the **cross population extended haplotype homozygosity** (or **XP-EHH**) test. A similar test (with a few subtle differences) was proposed by Tang et al., who defined the analog of Equation 9.42b as their **ln(*Rsb*)** statistic. A final cross-population comparison test was suggested by Kimura et al. (2007), who considered two ratios of haplotype homozygosity in a control versus test population. The first is the ratio of haplotype homozygosities in the two populations (***rHH***), while the second is the ratio ***rMHH*** of haplotype homozygosities restricted to the most frequency haplotype in the control population (i.e., the probability that two random draws of sequences are both the most common haplotype as defined by the control). For a recently-fixed allele, the expectation is a low $rMHH$ value (high population divergence), but a high rHH value (population-specific decrease in haplotype diversity). Simulation studies with neutral models (under limited demographic conditions) suggested that the combination of low $rMHH$ and high rHH values is rather unusual. Again, this is an outlier-based approach with outliers on an $rMHH$ by rHH plot of the genomic data suggesting potential targets of selection.

Table 9.3. Summary of tests reviewed in this chapter that use haplotype/LD information to indicate positive selection. AFS(k) denotes the equilibrium neutral distribution for the allele frequency spectrum (distribution of number of haplotype classes within a sample), conditioned on k (Equation 2.33a). S denotes the number of segregating sites, HH denotes the haplotype homozygosity (the probablity that two randomly-chosen sequences are identical in some defined region), and EHH is the extended haplotype homozygosity (the length of the region for a specificed allele overwhich $HH > 0.05$.

Tests based on the allele-frequency spectrum (AFS)
Watterson's Test: Observed allelic homozygosity vs. expected homozygosity under AFS(k)
Slatkin's exact Test: Observed AFS vs. expected AFS(k)
Innan et al's *HCT*: Observed AFS vs. expected AFS conditioned on observed S
Hudson's *HP* test: Frequency of most common haplotype given S
Fu's *W*: Test for deficiency of rare haplotypes given S
Fu's F_s: Test for excess of rate haplotypes given $\widehat{\theta}_{\Pi}$ (average pair-wise difference estimator)
Depaulis and Veuille's *K*: Observed number of haplotypes given S
Depaulis and Veuille's *H*: Observed haplotype diversity given S
Garud et al.'s H_{12} : Observed haplotype diversity lumping two most frequent classes
Garud et al.'s H_2: Observed haplotype diversity ignoring the most frequent class
Tests based on averages of pairwise disequilibria
Kelly's Z_{nS}: Average of all pairwise disequilibria between all sites in a region
Kim and Nielson's ω_{max}: Pairwise LD among sites within vs. between sides of a region
Tests based on frequency estimates of age vs. allelic diversity estimates of age
Age estimated by decay of LD between allele and a linked marker
Age estimated by number of segregating sites S within an allelic haplotype class
Age estimated by copy number variance at tightly-linked STRs in the allelic class
Tests contrasting haplotype lengths of alternative alleles in the same population
Sabeti's *rEHH*: ratio of the haplotype length (EHH) of two alternative alleles
Wang's *LDD*: Rate of linkage disequilibrium decay, modification of EHH
Voight's *iHS*: ratio of area under the EHH curve for ancestral vs. derived alleles
Tests contrasting haplotype lengths of the same allele in two populations
Sabeti's *XP-EHH* , Tang's *ln(Rsb)*: ratio of area under the EHH curve in different populations
Kimura's *rHH* vs. *rMHH* plot: ratios of overall HH , HH based on most frequency haplotype

Summary: Tests Based on Haplotype/LD Information

As summarized in Table 9.2, different sweeps (hard, partial, soft) leave different haplotype signals. Given the diversity of such signals, it is not surprising that there are a rich number of haplotype-based tests to detect these different features (Table 9.3). LD-based tests are generally regarded as the *most powerful for detecting partial sweeps*. Signatures from such a sweep include long haplotypes at excessive frequencies, alleles that are at too high a frequency given other estimates of their age, an excess of one or a few haplotypes, and a reduction in haplotype diversity. Under a partial sweep, site-frequency spectrum tests often perform poorly, as the distortion in the frequency spectrum is often not sufficiently powerful. Thus, LD-based tests often a powerful tool for looking at selection that is *currently underway*. A partial sweep can be generated under a number of rather different settings. First, we could be sampling a population where an sweep (hard or soft) is ongoing. Second, the population could be experiencing selection for an allele that is recently under balancing

selection, with its frequency increasing towards its equilibrium value. The signature under balancing selection is very different during the partial sweep versus long-term equilibrium phases (Charlesworth 2006). Finally, alleles may routinely experience fluctuating selection, wherein the nature of selection changes dramatically over time. A partial sweep can be generated by such an allele that has recently significantly increased in frequency due to it being transiently favored.

In addition to their unique role in detecting partial sweeps, LD summary statistics can also offer significant power to detect *just-completed sweeps*. Under a hard-sweep, the unusual pattern of high LD on either side, but not across, a selected site can be detected using Kim and Nielson's ω statistic (Equation 9.37). However, this statistic has no power to detect a soft sweep. Conversely, Kelly's Z_{ns} statistic (average pairwise LD throughout a region, Equation 9.36b) can detect a recently-completed soft sweep, but has no power for a just-completed hard sweep.

As with both parametric sweep-based tests and tests using the site-frequency spectrum, LD/haplotype-based tests can also generate false positives for neutral alleles in nonequilibrium populations. The standard tools used with site-frequency tests of outlier analysis to suggest regions of interest and coalescent simulations based on marker-based demographic estimates can also be used here, with the same caveats. As mentioned, both of these approaches use corrections based on genome-wide patterns thus do not adjust for allelic surfing. This is especially troublesome as the model species most surveyed for recent selection — humans, cosmopolitan human commensal *Drosophila* (*melanogaster* and *simulans*), and *Arabidopsis* — are all know to have undergone massive spatial expansions over the last 100,000 years, making them prime candidates to experience surfing.

SEARCHES FOR SELECTION IN HUMANS AND DOMESTICATED ORGANISMS

The search for selection is motivated by both gene-specific and genome-wide questions. At the level of an individual gene, we would like to understand how an ecological challenge is met by a molecular solution, which we expect to be highly idiosyncratic for each particular case. On a genomic scale, we are interested in general trends of adaptation: What is the relative importance of regulatory versus structural (amino-acid) changes? Are genes of major effect more important than genes of minor effect? The growing consensus on the former is that regulatory changes may be the predominant route for adaptation (e.g., Grossman et al. 2013), but whether their contribution is roughly equal to structural changes or significantly greater remains an open question. The latter question is still rather open (e.g., Tian et al. 2009), and any current understanding is certainly based on a nonrandom sample, as methods for detecting sweeps are strongly biased towards hard sweeps, namely genes of major effect under strong selection.

At present, only a limited number of species have had aggressive genome-wide scans for genes under recent/ongoing selection. For natural populations, the most extensive work has been done on humans, *Drosophila*, and *Arabidopsis*. Given that we know a great detail about the genetics, genomics, and molecular biology of these species, their choice is not surprising. Moreover, all have undergone major expansions into a wide range of new habitats over the last 100,000 years. For humans, the movement out of Africa into more temperate climates coupled with the transition from hunter-gatherer to agriculture, along with increases in population density, generated novel environmental pressures. The commensal *D. melanogaster* and *D. simulans* followed humans into these new environments, while in the northern hemisphere, *Arabidopsis* underwent significant range expansion following the end of the ice age. The environmental challenges faced by these species leads us to expect a history of recent selection. Since numerous examples from *Drosophila* and *Arabidopsis* have been discussed

over the last two chapters, our focus will be on humans.

The other group of organisms where extensive searches of recent selection have been conducted are domesticated plants and animals. Domestication represents a major change in the environment, and hence the opportunity for significant selection within a very recent time period (roughly within the past 10,000 years). For example, the domestication of several different grain-producing crops required the transformation of wild progenitors with hard-coated seeds that easily scattered off the plant for dispersal (**seed shattering**) to domesticated varieties with soft coated seeds that remained on the plant for easier harvesting.

Recent/Current Selection in Humans

Early searches in humans looked for molecular signals in genes either believed, or very strongly suspected, to be under selection in particular environments. Examples include disease-resistance genes such as *Duffy* (*FY*) and *G6PD*, dietary genes such as lactase (*LCT*), and climate-related genes such as *MC1R* (influencing skin color). Ronald and Akey (2005) and Harris and Meyer (2006) review these and other candidate genes, which show strong signals (such as skewed site-frequency spectra, long haplotypes, and/or excessive F_{ST}), adding support to the belief they have experienced recent selection. Taking an estimate of long-term population size in humans as $N_e \sim 5,000$, signals of recent hard sweeps should persist for no more than $2Ne$ generations (Table 8.1), or roughly covering no longer than the past 250,000 years.

Starting with the advent of dense-SNP maps, and continuing as whole-genome sequencing became economically feasible, candidate gene studies were replaced by **genomic scans**, searching the genome without any preconception of what sites might be under selection. Biswas and Akey (2006) reviewed six such scans that used different statistics to infer positive selection. In total, roughly 2300 genes were found with signals of selection in at least one study, but the overlap between studies was quite small. For example, consider the overlap between Wang et al. (2006) and Voight et al. (2006) who used long-range haplotype tests (LDD and iHS, respectively, Table 9.3), and Carlson et al. (2005) who used outliers in Tajima's D (Table 9.1). As tabulated below (after Biswas and Akey), diagonal elements represent the number of sites declared to be under positive selection in that study, and off–diagonal elements the number shared between the two studies. For example, of the 455 sites detected by Voight, 125 (27%) were also seen by Wang. Conversely, of the 176 sites with outliers in D, only 6% (11) of these were also detected by Voight, while 27% (47) were detected by Wang.

	Wang (LD)	Voight (LD)	Carlson (D)
Wang	1799	125	47
Voight		455	11
Carlson			176

More recent summaries (Akey et al 2009; Fu and Akey 2013) echo this general lack of cross-study replication for most detected sweeps. A number of reasons can generate these discrepancies. The first is clearly power. When power is low, many sites under positive selection are missed, and the overlap between studies is expected to be small. If two studies both have 50% power, replication (both studies declaring a site to be significant) is expected only 25% of the time. If power is much less, say 20%, then only four percent of the sites will be replicated. Second, as we have stressed throughout this chapter, different tests detect selection over different time scales. A partial sweep gives an LD signal, but not likely a site-frequency signal, while the converse is true for a recently completed sweep. Thus, LD-based tests (such as EHH and iHS) are *expected* to detect different genes than site-frequency tests. Finally, selection can also be rather *local*, so that only a subset of populations might be experiencing selection. If one such population is included in one study, but not in another, lack of replication oc-

curs. Pickrell et al. (2009) found evidence of significant local adaptation (population-specific changes) in a survey of 53 populations, although Hofer et al. (2009) note that the striking differences in allele frequencies between human populations could have easily arisen as a consequence of population expansion (and the accompanying allelic-surfing).

The take-home message is that genomic scans reveal numerous potential sites under selection, but these are usually not replicated across studies. This could simply be a consequence of low power, but as is the case for QTL and association-mapping studies, an initial exciting finding is only the start, not the conclusion, in the search for genes under selection. Unlike association studies, where support is offered by independent replication, the concept of independent replication in the search for genes under selection is more problematic. Finding similar support in two independent studies drawn from the same population is comforting, but not a formal validation, as demographic features could have generated the signal, and no amount of resampling will remove its effect. Similarly, lack of replication between two different populations can easily be explained by differential selection pressures. Support for positive selection would be offered by observing differences in fitness among alternative alleles at a candidate gene, but failure to do so is not damning. Very small fitness differences below the level of detection in most reasonable studies can still be critically important (Chapter 5). Likewise, the nature of selection changes with the environment, and a signal could be the result of a past environmental effect that is no longer important.

Further tempering the above results is the recent realization that classic sweeps appear to be rare in humans (Hernandez et al. 2011; Lohmueller et al. 2011; Alves et al. 2012) and that several studies are consistent with polygenic sweeps (Hancock et al. 2010a, b; Fumagalli et al. 2011; Turchin's et al. 2012), both points discussed in Chapter 8. Human geneticists thus face the very real possibility that only a biased set of events are being detected (along with a potentially large number of false positives), giving a very distorted view of the genetics of human adaptation.

Domestication Versus Improvement Genes

While domestication is often perceived as the result of gradual, or even unintentional, selection over some period of time, this not need be the case. Some events are very sharp and deliberate, such as the creation of novel forms for human use by species crosses. A classic example is the mule, the sterile offspring of a male donkey and a female horse. Likewise, a number of crops are allopolyploids resulting from the cross of two (or more) different progenitor species. At the other extreme is the genetic transformation of an ancient progenitor into a modern variety, such as the dramatic changes in plant architectural from teosinte to modern maize (LW Figure 5.2). The sites of these genetic changes are called **domestication genes**. The threshold beyond which a wild species is said to be domesticated can be hard to assess, but one operational definition is that domesticated varieties survive very poorly in a natural setting, due to the loss of traits that improve fitness in the wild, but decrease it in the domesticated environment. As best stated by Zeder et al. (2006), "domestication is a unique form of mutualism", leaving both genetic and archaeological signals (see Zeder et al. for several interesting examples). The genetics of domestication is a rapidly growing field, which we only touch on briefly here (see Doebley et al. 2006 and Gross and Olsen 2010 for overviews). Additional genetic perspectives are offered by Diamond (2002) and Purugganan and Fuller (2009).

A domesticated species may have multiple independent origins, with each distinct lineage potentially having its own unique set of domestication genes. Maize (Matsuoka et al. 2002), emmer and einkom wheats (*Triticum turgidum* and *T. monococcum*, Zohary 1999), potatoes (*Solanum tuberosum*, Spooner et al. 2005) and peanuts (*Arachis hypogaea*, Kochert 1996) all appear to have a single origin of domestication. A caveat with many of these findings (which are often based on seeing a monophyletic clade when using neutral markers) is that

simulations by Allaby et al. (2008) showed monophyletic clades can still be produced in crops with multiple origins, provide there was a rather protracted period of domestication. Other crops, such as barley (*Hordeum vulgare*, Zohary 1999) and *Phaseolus* beans (Gepts et al. 1986), show evidence of multiple (i.e., separate) domestication events. The situation is a bit more complex in Asian rice (*Oryza sativa*), whose *indica* and *japonica* varieties have been regarded as independent domestication events (Londo et al. 2006; Sang and Ge 2007). However, it now appears that *japonica* was domesticated first from its wild progenitor *O. rufipogon* in the Pearl River basin in southern China, with *indica* being subsequently developed with crosses between *japonica* and *rufipogon* strains from South and Southeast Asia (Huang et al. 2012). This nested relatedness explains why these two major divisions of Asian rice share a number of key domestication alleles (such as *sh4* which reduces grain shattering) that appear to have arisen just once (Sang and Ge 2007; He et al. 2011).

The distinction between genes involved in domestication events and those involved in subsequent improvement of varieties can be rather subtle. Operationally, domestication alleles are assumed to be present in all descendant varieties from the domestication event, while **improvement** (or **diversification**) genes were further selected in only a subset of varieties. These could be the result of deliberate selection, such as for sticky rice, or could be the result of local selection conditions that lead to the formation of **landraces** (locally adapted varieties). Detecting genes involved in domestication is intellectually interesting as model systems of adaptation (Ross-Ibarra et al. 2007), while knowledge of improvement genes may give the breeder insight into achieving specific objectives. Genes under selection in specialized landraces may be especially informative. For example, genes with signatures of selection that are limited to (say) varieties in high drought conditions can suggest important target genes to improve drought tolerance in current elite germplasms. An example of this is Kane and Riesenberg (2007), who used the log RH and log RV tests (Equation 9.14) to search for sunflower (*Helianthus annuus*) genes whose signatures of selection are restricted to populations in drought and/or high salt environments.

Finding Domestication and Improvement Genes in Crops

One standard approach for finding domestication/improvement genes is QTL mapping in a cross between the wild ancestor (provided it still exists) and the domesticated/improved variety. Such a strategy relies on knowing which *traits* are important. Classic examples are the *teosinte branched1* (*tb1*) and *barren stalk1* (*ba1*) genes for plant architecture in teosinte/maize (Doebley et al. 1995; Gallavotti et al. 2004), *teosinte glume architecture1* (*tga1*) for naked grains in maize (Wang et al. 2005), *fw2.2* for tomato fruit size (Frary et al. 2000), and *sh4* and *qSH1* for reduced seed shattering in rice (Li et al, 2006; Konishi et al. 2006). Given the obvious success of the QTL mapping approach, what role do signatures of selection play in the search for domestication/improvement genes? First, showing that QTL-detected regions were under selection provides independent support for their role in domestication. Second, one can estimate the average strength of selection, and hence obtain some indication of the required time to either fix, or substantially increase, its frequency during domestication. Finally, and perhaps most importantly, scans for selection are *trait-independent* searches. While changes in some morphological traits may be rather obvious candidates for the basis for domestication/improvement (and hence characters for QTL/association mapping), more subtle physiological changes may be less obvious. It is worth noting that Hufford et al. (2012) found the majority of scan-detected regions in maize showed stronger selection signals than those for QTL regions associated with major morphological differences related to domestication.

Domestication can offer important insight into the genetics of adaptation. For example, how important are loss-of-function versus gain-of-function mutations? An interesting example is the *Q* gene in bread wheat (*Triticum aestivum*). This is a critical domestication gene,

allowing modern (nonhulled) wheats to be easily threshed (separation of the seed from chaff). A phylogeneic analysis by Simons et al. (2006) indicated that the *Q* gene had a single origin, and appears to be a gain-of-function, rather than a loss-of-function, mutation. In crops, Doebley et al. (2006) and Gross and Olsen (2010) note that loss-of-function mutations are rare among known domestication genes, but not uncommon among improvement genes. Recent surverys of chickens and pigs (Rubin et al 2010, 2011) found very little evidence that loss-of-function mutations are common in animal domestication/improvement genes.

A second general question is the relative importance of regulatory changes (Zuckerkandle 1968; King and Wilson 1975; Carroll 2008). One interesting study bearing on this is the regulatory *CAULIFLOWER* (*BoCal*) gene and its role in different varieties of domesticated cabbage (*Brassica oleracea*). Closely related to wild and domesticated cabbage (*B. oleracea oleracea*) are kale (*B. oleracea acephala*), cauliflower (*B. oleracea botrytis*), and broccoli (*B. oleracea italica*). The last two show significant modification of their inflorescence structures during domestication, while kale and cabbage have normal floral structures. Purugganan et al. (2000) showed that a nonsense allele (due to a premature termination codon) of *BoCal* is fixed in both cauliflower and broccoli, but also present in wild cabbage and kale. Strong haplotype structure is seen, with a reduction in nucleotide diversity around this gene relative to other sites. A sample of cauliflower and broccoli alleles showed significantly negative Tajima's D and Fu and Li's D^* (Table 9.1), all consistent with recent positive selection. Neither D nor D^* was significant in a sample of kale and wild cabbage alleles. The nonsense (loss of function) allele appears to have a single origin and is a regulatory mutation. However, the presence of this allele (at lower frequencies) in normal flowering populations of wild cabbage and kale show that it is not sufficient by itself for the inflorescence modification that arose during the domestication. Studies in *A. thaliana* indicate that mutations in a second gene *AP1* (which is also found in *B. oleracea*) are required to produce the cauliflower phenotype in *thaliana*.

Domestication and Improvement Genes in Rice

As perhaps the most important single food item in the world, it is not surprising that rice has been widely studied for domestication/improvement genes. A key change during domestication of Asian rice was moving from a reasonably outcrossed species to a highly selfing one. Selfing reduces the effective recombination rate, resulting in the effects of a sweep extending over a larger region of the genome. In particular, if t is the rate of selfing, Nordborg (2000) found that the effective recombination rate c^* is well approximated by

$$c^* \simeq c\left(1 - \frac{t}{2-t}\right) \tag{9.43}$$

For modern Asian rice, $t \simeq 0.99$, giving a roughly fifty-fold decrease in the effective recombination rate. This reduction in the recombination rate, when combined with small genome size (less than 400 Mb), implies a significant impact on most of the rice genome is expected if even a modest number of sweeps occurred during domestication (Example 8.12). Caicedo et al. (2007) noted that domesticated rice shows a genome-wide excess of high-frequency derived alleles, which is not consistent with a simple founding bottleneck but is consisted with sweeps impacting much of the genome. Both He et al. (2011) and Huang et al. (2012) detected numerous regions of reduced diversity over a panel of domesticated lines relative to wild *O. rufipogon* populations, many of which exceeded 200 kb.

An example of such a long region of depressed variation is seen around the *Waxy* gene, where a splice mutant results in low amylose levels, giving "Sticky" (glutinous) rice (reviewed by Olsen et al. 2006). This is an improvement trait, largely restricted to temperate *japonica* varieties. There is a massive signature around this gene, with a 97% reduction in nucleotide diversity ($\pi = 0.0002$ versus normal levels of $\pi = 0.0064$ in wild accessions).

The sweep signature spans 250 kb, reducing diversity at close to 40 genes. Further, there is a strong EHH signal around *Waxy*, and alleles from temperate *japonica* lines show a highly negative Tajima's D. Using an estimate of $c = 3.7 \times 10^{-7}$ per bp (Inukai et al. 2000), Olsen used Equation 8.6b to estimate the strength of selection as

$$s \simeq \frac{3.7 \times 10^{-7} \cdot 250,000}{0.02} = 4.6$$

Using the effective recombination rate (Equation 9.43) reduces this estimate of s to around 0.1 (assuming a high selfing rate, $t = 0.99$).

Domestication and Improvement Genes in Maize

Maize is the king of crops when it comes to both genomic scans and specific searches for selection at candidate genes, and excellent overviews are given by Doebley (2004) and Tian et al. (2009). The first demonstration of selection on a putative domestication gene comes from maize, and is *teosinte branched 1*, or *tb1* for short (Wang et al. 1999; Clark et al. 2004, 2006). Given its obvious role from QTL studies as a candidate domestication gene, Wang et al. (1999) compared the levels of nucleotide diversity 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 variation in the 5′ NTR region of maize (but not teosinte) *tb1*, suggesting a selective sweep influenced this region (Figure 8.2). 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. Wang et al. applied Equation 8.6b to estimate the average strength of selection as $s \simeq 0.05$. This value of s implies an expected time for selection to fix the *tb1* allele of around 300 to 1000 years, indicating a fairly long period of domestication. Recently, Studer et al. (2011) localized the selected site to an insertion of the *Hopscotch* retrotransposon roughly 64 kb upstream of *tb1*, which results in up-regulation in the amount of *tb1* transcripts. Interestingly, this insertion predates domestication by at least 10,000 years and hence standing variation was exploited during domestication.

Other examples of selection signals on putative domestication genes include the *c1* gene that regulates anthocyanin production and hence the transition from colorless to colored kernels in early maize (Hanson et al. 1996), genes in the starch pathway (Whitt et al. 2002), *Y1* for yellow kernels (Palaisa et al. 2004), *barren stalk 1* (Gallavotti et al. 2004), the *tga1* gene for naked seeds (Wang et al. 2005), and MADS-box regulators of plant floral development (Zhao et al. 2010).

Modest-scale genomic scans have been performed in maize by Vigouroux et al. (2002), Yamasaki et al. (2005), Wright et al. (2005), and Hufford et al. (2007). Based on finding that two to four percent of 774 sampled genes showed signatures of selective, Wright et al. suggested that over 1200 maize genes have likely been influenced by artificial selection during maize domestication/improvement. Hufford et al. inferred (based on analysis of 30 of Wright et al.'s candidates) that roughly 40 percent of these are domestication genes and the other 60 percent improvement. They did not find that regulatory genes (such as transcription factors) were over-represented among these candidates. However, a more recent study by Zhao et al. (2010) sequenced 32 MADS-box genes (transcription factors) and 32 randomly-chosen loci, finding 8 MADS-box genes were targets for domestication and an additional one for improvement, while two of the random genes were domestication targets and an additional four improvement. Thus, in this study regulatory changes were overrepresented among

selected sites. Hufford et al. (2007) also noticed that candidate genes were significantly overrepresented in expression in ear tissue relative to vegetative tissues.

A more comprehensive scan was performed by Hufford et al. (2012), who examined 35 improved lines, 23 landraces, and 17 wild relative lines, using the XP-CLR test (Equation 9.20). Recall that this likelihood-based test compares the genomic spatial F_{ST} pattern in a selected line relative to an unselected control and returns an estimate of the strength of selection during the sweep. Domestication genes were detected by contrasting landraces (selected lines) with wild relatives (control), while improvement genes were located by contrasting improved lines against landraces. Taking regions with the highest 10% of test scores returned 484 potential domestication genes and 695 improvement genes. The average selection coefficients for these groups was $s = 0.015$ for domestication and $s = 0.003$ for improvement. Relative to random genes, domestication candidates showed greater changes in gene expression from their teosinte ancestor, tending to have higher levels of expression and more stability in expression over maize lines. Divergence in gene expression between teosinte and maize was further studied by Swanson-Wagner et al. (2012), who found that the regions detected by Hufford et al. were significantly enriched for both differences in expression and altered coexpression profiles relative to random genes from the maize genome.

An especially interesting study on maize domestication was Jaenicke-Després et al. (2003), who used ancient maize ears as a time machine to look at the fixation of domestication alleles. Five maize cobs from the Ocampo Caves in NE Mexico were carbon dated, with two being around 4300 years old, and the other three between 2300 and 2800 years old. Six ancient cobs from Tularosa Cave in New Mexico were also examined, two of which dated to around 1900 years, with the remaining four around 650 to 900 years. DNA was able to be extracted from all cobs, and all showed the modern maize allele at *tb1*. A second domestication gene *pbf* (involved in seed storage protein production) had the modern allele in all cobs as well. The final domestication gene examined was *sugary 1* (*su1*), which is involved in starch expression in the kernels. Here the pattern was mixed. The alleles M1 and M2 at this locus are found in 30 and 62% (respectively) of modern maize lines, whereas both are around 7% in teosinte. All the cobs from Mexico were homozygous for M2, while the four younger cobs from New Mexico homozygous for M1. However, the two older cobs from New Mexico were heterozygotes, M1/M2 and M1/T1, where the T1 allele is not seen in modern maize and found in only around four percent of current teosinte populations. Thus, it appears that much of the domestication was completed by 4000 years ago, but (at least in New Mexico populations) as of around 2000 years ago, allelic selection was not yet finished.

Finally, a cautionary tale in the search for domestication genes is offered by *Shrunken2*, *Sh2* (Whitt et al. 2002; Manicacci et al. 2006). *Sh2* is involved in endosperm starch biosynthesis, and was suggested as a target domestication gene from QTL studies showing that a seed-weight QTL in a maize-teosinte cross covers the *Sh2* region. More careful analysis by Whitt et al. and especially Manicacci et al. showed similar levels of nucleotide diversity in both maize and teosinte. A comparison with two sister species suggested that a sweep in the 3′ region of *Sh2* occured in teosinte that predates domestication. The wild ancestors of our current crops are themselves subject to selection, so caution is in order if one declares selection by contrasting diversity in a domesticated variety with that in a sister species of the progenitor.

Silkmoths and Flies

When one envisions domesticated animals, usually pets or farm animals come to mind, not insects. However, insect populations have been domesticated as well. Xia et al. (2009) sequenced the genomes of 29 lines of domesticated silkmoths (*Bombyx mori*) and 11 wild progenitors (*B. mandarina*). Their analysis clearly shows a single domestication event giving rise to *mori*, with only a mild bottleneck (90% of the ancestral diversity is maintained). Using a joint statistic based on reduction in diversity ($\widehat{\theta}_{mori}/\widehat{\theta}_{mand}$) within a region coupled

with a low Tajima's D score, they identified slightly over 1000 regions of interest, spanning just under 3% of the genome. This suggested around 350 protein-coding genes within these regions as potential candidate for domestication genes (focusing on structural, as opposed to regulatory, changes). Of these, 159 showed differential expression between *mori* and its wild relative, 90 of which are expressed in the silk gland, midgut, or testis. Two of the candidate genes in the silk gland were related to counterparts in *Drosophila* involved in transcriptional regulation of the glue genes.

The selection pressures during *Bombyx* domestication were likely both deliberate (increased silk production) and unintentional (easier handling and better survival under cultivation). Such domestication selection pressures for growth and survival under laboratory conditions can potentially occur in any organism under long-term captivity, and *Drosophila* is no exception. Montgomery et al. (2010) examined the behavior of eight STRs over time in series of replicate populations of different sizes with very tight control over the effective population sizes through complete knowledge of the population pedigree (meticulously accomplished by each generation choosing every set of parents in the population). The loss of heterozygosity at all STRs was about 12% faster than predicted from the pedigree-generated inbreeding coefficients, while the between-population divergence (F_{ST}) and temporal within-population changes in allele frequencies were also significantly greatly (by 25% and 33%, respectively) than predicted under drift. The authors interpreted these results as support for multiple ongoing partial sweeps throughout the genome influencing the dynamics at linked STRs, suggesting ongoing selection to adapt to the domestication conditions.

Constraints on Finding Domestication/Improvement Genes Through Selective Signals

While a number of putative domestication genes have been located, there are reasons to suspect that many have been missed. Soft and polygenic sweeps are certainly a complication, as we assume that domestication presents a fairly dramatic change in the environment, and thus standing variation is likely to be exploited, potentially giving a very reduced signal (Chapter 8). Indeed, one might imagine a bias in the current domesticated species, in that ancient farmers used those species that they could most easily exploit. This likely favored species with at least some standing variation for traits of interest, with these beating out competing species for which farmers had to wait for new mutations to show improvement. However, even if selection occurs on a preexisting allele, this may be partially offset by a strong bottleneck during domestication. If only one or a few lineages carrying the favorable allele survive, this creates a signal closer to a hard, than a soft, sweep. Another complicating factor is that domestication and improvement events can be largely polygenic. For example, comparisons of dairy versus beef cattle largely rely on differences in allele frequencies, with few fixed differences (Example 9.3, Figure 9.1). One exception to this are some breeds of dogs, where dramatic changes in just a few traits can underlie breed differences (Example 9.5, Figure 9.3).

Several other features complicate the detection and localization of domestication genes. If the ancestral species had low levels of polymorphisms at the start of selection (perhaps from passing through bottlenecks and/or being under selection themselves), then the reduction in polymorphism around the selective site leaves a much weaker signal. Thus, for some species it may be very difficult to detect signatures of selection, even for sites under strong selection. Hamblin et al. (2006) found that the genome-wide background variation in Sorghum (*Sorghum bicolor*) was too low to reliably detect signatures of selection given the markers and density they used. Wild accessions of Sorghum had levels of nucleotide diversity around $\pi = 0.0027$, far lower than teosinte, while domesticated varieties had even lower levels ($\pi = 0.0008$). The combination of low background levels of variation in the progenitor coupled with an obvious strong bottleneck during domestication makes detection of regions under selection challenging, but not impossible (Casa et al. 2006).

Another factor that influences the signal of a selective sweep is the effective amount of recombination c^*, which is a function of both the local recombination rate c and the amount of selfing (Equation 9.43). The strength of selection associated with a sweep can be significantly overestimated in selfing species unless one corrects for the effective recombination rate (using c^* in place of c). High effective levels of recombination result in a shorter window of influence around the selected site. While a plus for fine-mapping of potential genes, this setting requires a denser marker scan around a putative region, otherwise a potential signal might be missed. In contrast, a low effective recombination rate (such as with selfers) results in a large window of influence, making such a site easier to detect but much harder to localize.

The size of a sweep also has important implications beyond our ability to detect and localize sites of recent selection. Signals of a sweep arises because of a reduction in the effective population size around the selected site. This reduction has important evolutionary consequences, as the efficiency of selection at linked genes is reduced within the region influenced by the sweep (Chapter 8). Deleterious alleles are more likely, and favorable alleles less likely, to become fixed compared to sites outside of the sweep. In species with high effective recombination rates, only a small genomic region (and hence few genes) are influenced by a sweep. As we saw for the *Waxy* gene, in a highly selfing species, sweeps can have consequences for the behavior of numerous genes beyond the target. In a species experiencing a number of large sweeps during domestication, deleterious mutations can accumulate during the domestication process due to reduced selection resulting from the decrease in effective population size, above and beyond the general bottleneck that occurs across the genome due to domestication. There is at least some suggestive evidence of this occurring in rice (Example 8.12) and it expected to be more of a concern in selfing species.

Relative Strenghts of Selection on Domestication vs. Improvement Genes

An unresolved question concerns the relative strength of selection on domestication versus improvement genes, a contrast first discussed by Olsen et al. (2006). Based on local estimates of recombination rates and the length of depressed variation around the candidate genes, two domestication genes in maize had estimates of s between 0.02 (*tga1*, Wang et al. 2005) and 0.05 (*tb1*). However, the improvement gene *Y1* has a 600 kb sweep, giving an estimated strength of selection of $s = 1.2$ (Palaisa et al. 2004), and (as previously discussed) the strength of selection on the rice improvement gene *Waxy* (correcting for the effective recombination rate) is $s = 0.1$. This small initial sample suggested stronger selection on improvement genes. Conversely, in the Hufford et al. (2012) survey of the maize genome, regions involved in domesticated had a average estimated value of $s = 0.015$, while regions associated with improvement had an average estimated value of $s = 0.003$. One potential reason for the significant decrease in improvement genes was that the authors lumped together both tropical and temperate landraces, which reduces the average strength of selection if landraces were differentially selected. If selection is too intense (especially when selfing can occur), considerable linkage drag allows 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 have been a good thing if selection during domestication was relatively weak.

Achaz, G. 2008. Testing for neutrality in samples with sequencing errors. *Genetics* 179: 1409–1424. [9]

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249–258. [9]

Akey, J. M. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.* 19: 711–722. [9]

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2: e286. [9]

Akey, J. M., A. L. Ruhe, D. T. Akey, A. K. Wong, C. F. Connelly, J. Madeoy, T. J. Nicholas, and M. W. Neff. 2010. Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci. USA* 107: 1160–1165. [9]

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* 12: 1805–1814. [9]

Allaby, R. G., D. Q. Fuller, and T. A. Brown. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl. Acad. Sci. USA* 105: 13982–13986. [9]

Alves, I., A. S. Hanulova, M. Foll, and L. Excoffier. 2012. Genomic data reveal a complex making of humans. *PLoS Genet.* 8: e1002837. [9]

Andolfatto, P., and M. Nordborg. 1998. The effect of gene conversion on intralocus associations. *Genetics* 148: 1397–1399. [9]

Andolfatto, P., J. D. Wall, and M. Kreitman. 1999. Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster. Genetics* 153: 1297–1311. [9]

Aquadro, C. F., and B. D. Greenberg. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* 103: 287–312. [9]

Bamshad, M., and S. P. Wooding. 2003. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* 4: 99–111. [9]

Baudry, E., and F. Depaulis. 2003. Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165: 1619–1622. [9]

Bazin, E. K. J. Dawson, and M. A. Beaumont. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* 185: 587–602. [9]

Beaumont, M. A. 2005. Adaptation and speciation: what can F_{st} tell us? *Trends Ecol Evol.* 20: 435–440. [9]

Beaumont, M. A., and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13: 969–989. [9]

Beaumont, M. A., and R. A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond.* B 263: 1619–1626. [9]

Biswas, S. and J. M. Akey. 2006. Genomic insights into positive selection. *Trends in Genetics* 22: 437-446. [9]

Boitard, S., C. Schlötterer, and A. Futschik. 2009. Detecting selective sweeps: A new approach based on hidden Markov models. *Genetics* 181: 1567–1578. [9]

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd, and L. L. Cavalli-Sforza. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms *Proc. Natl. Acad. Sci. USA* 1991 88: 839–843. [9]

Caicedo, A. L., S. H. Williamson, R. D. Hernandez, A. Boyko, A. Fledel-Alon, T. L. York, N. R. Polato, K. M. Olsen, R. Nielsen, S. R. McCouch, C. D. Bustamante, and M. D. Purugganan. 2007. Genome-wide patters of nucleotide polymorphism in domesticated rice. *PLoS Genetics* 3: e163. [9]

Carlson, C. S., D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston, M. J. Rieder, and D. A. Nickerson. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15: 1553–1565. [9]

Carroll, S. B. 2008. Evo-Dev and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134: 25–36. [9]

Casa, A. M., S. E. Mitchell, J. D. Jense, M. T. Hamblin, A. H. Paterson, C. F. Aqudaro, and S. Kresovich. 2006. Evidence for a selective sweep on chromosome 1 of cultivated sorghum. *Crop Sci.* 46: S27–S40. [9]

Cavalli-Sforza, L. L. 1966. Population structure and human evolution. *Proc. Royal Soc. London. B* 164: 362–379. [9]

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2: e64. [9]

Chen, H., N. Patterson, and D. Reich. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20: 393–402. [9]

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Gen. Res.* 15: 1496–1502. [9]

Clark, R. M., E. Linton, J. Messing, and J. F. Doebley. 2004. Pattern of diversity in the genomic region near the maize domestication gene *tb1*. *Proc. Natl. Acad. Sci. USA* 101: 700–707. [9]

Clark, R.M., T.N. Wagler, P. Quijada, and J. Doebley, 2006. A distant upstream enhancer at the maize domestication gene *tb1* has pleotropic effects on plant and inflorescent architecture. *Nature Genet.* 38: 594-597. [9]

Conover, W. J., 1999. *Practial nonparametric statistics,* 3rd Ed. Wiley, NY. [9]

Cook, L. M., and D. A. Jones. 1996. The *medionigra* gene in the moth *Panaxia dominula*: the case for selection. *Phil. Trans. R. Soc. B.* 351: 1623–1634. [9]

Coop, G., and M. Przeworski. 2007. An evolutionary view of human recombination. *Nat. Rev. Genet.* 8: 23–34. [9]

Coque, M., and A. Gallias. 2006. Genomic regions involved in response to grain yield selection at high and low nitrogen fertilization in maize. *Theor. Appl. Gene.* 112: 1205–1220. [9]

Decker, J. E., D. A. Vasco, S. D. McKay, M. C. McClure, M. M. Rolf, J.W. Kim, S.L. Northcutt , S. Bauch, B. W. Woodward, R. D. Schnable, and J. F. Taylor. 2012. A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (*Bos taurus*) genome to selection on complex traits. *BMC Genom.* 13: 606. [9]

De Koeyer, D. L., R. L. Phillips, and D. D. Stuthman. 2001. Allelic shifts and quantitative trait loci in a recurrent selection population of oat. *Crop Sci.* 41: 1228–1234. [9]

De Kovel, C. G. F. 2006. The power of allele frequency comparisons to detect the footprint of selection in natural and experimental populations. *Genet. Sel. Evol.* 38: 3–23. [9]

Depaulis, F., S. Mousset, and M. Veuille. 2001. Haplotype tests using coalescent simulations conditional on the number of segregating sites. *Mol. Biol. Evol.* 18: 1136–1138. [9]

Depaulis, F., S. Mousset, and M. Veuille. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *J. Mol. Evol.* 57: S190–S200. [9]

Depaulis, F., S. Mousset, and M. Veuille. 2005. Detecting selective sweeps with haplotype tests: Hitchhiking and haplotype tests. *in* D. Numinsky (ed) *Selective sweep*, pp 34–54. Springer, New York. [9]

Depaulis, F., and M. Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol* 15: 1788–1790. [9]

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. *Biometrics* 55: 997–1004. [9]

Diamond, J. 2002. Evolution, consequences and future of plant and animal domestication. *Nature* 418: 700–707. [9]

Doebley, J. 2004. The genetis of maize evolution. *Ann. Rev. Genet.* 38: 37–59. [9]

Doebley, J. F., B. S. Gaut, and B. D. Smith. 2006. The molecular genetics of crop domestication. *Cell* 127: 1309–1321. [9]

Doebley, J., A. Stec, and C. Gustus. 1995. *teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141: 333–346. [9]

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza. 2004. Mutations arising in the wave front of an expanding population. *Proc. Natl. Acad. Sci. USA* 101: 975–979. [9].

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112. [9]

Ewing, G., and J. Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26 : 2064–2065. [9]

Excoffier, L., M. Foll, and R. J. Petit. 2009. Genetic consequences of range expansions. *Ann. Rev. Ecol. Evol. Syst.* 40: 481—501. [9]

Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity* 103: 285–298. [9]

Excoffier, L., and N. Ray. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol. Evol.* 23: 347–351. [9]

Fay, J. C., and C.-I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413. [9]

Fisher, R. A., and E. B. Ford. 1947. The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity* 1: 143–173. [9]

Foll, M., and O. Gaggiotti. 2008. A genome-scane method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180: 977–993. [9]

Ford, M. J. 2002. Applications of selective neutrality tests to molecular ecology. *Mol. Ecology* 11: 1245–1262. [9]

Frary, A., T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knapp, B. Cong, J. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley. 2000. *fw2.2:* A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289: 85–88. [9]

Fu, W., and J. M. Akey. 2013. Selection and adaptation in the human genome. *Ann. Rev. Genomics Hum. Genet.* 14: 467–489. [8, 9]

Fu, Y.-X. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* 143: 557–570. [9]

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini, and R. Nielsen. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics* 7: e1002355. [9]

Fu, Y.-X. 1997. Statistical tests of neutrality of neutrality against population growth, hitchhiking and background selection. *Genetics* 147: 915–925. [9]

Fu, Y.-X., and W.-H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709. [9]

Gallavotti, A., Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley, M. E. Pé, and R. J. Schmidt. 2004. The role of *barrent stalk1* in the architecture of maize. *Nature* 432: 630–635. [9]

Galtier, N., F. Depaulis, and N. H. Barton. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155: 981–987. [9]

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov. 2014. Recent selective sweeps in *Drosophila* were abundant and primarily soft. *ArXiv* 1303.0906. FLAG UPDATE REF [8, 9]

Gepts, P., T. C. Osborn, K. Rashka, and A. Bliss. 1986. Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgris*): evidence for multiple centers of domestication. *Econ. Bot.* 40: 451–486. [9]

Gibson, J. B., N. Lewis, A. Adenca, and S. R. Wilson. 1979. Selection for ethanol tolerance in two populations of *Drosophila melanogaster* segregating alcohol dehydrogenase allozymes. *Aust. J. Biol. Sci.* 32: 387–398. [9]

Glinka, S., D. De Lorenzo, and W. Stephan. 2006. Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster. Mol. Biol. Evol.* 23: 1869–1878. [9]

Goldringer, I. and T. Bataillon. 2004. On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics* 168: 563 – 568. [9]

Goodman, S. J. 1997. R_{st} Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Mol. Ecol.* 6: 881–885. [9]

Griffiths, R. C. 1982. The number of alleles and segregating sites in a sample from the infinite-alleles model. *Adv. Appl. Prob.* 14: 225–239. [9]

Griffiths, R. C., and S. Tavare. 1994a. Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* 46: 131–159. [9]

Griffiths, R. C., and S. Tavare. 1994b. Sampling theory for neutral alleles in varying environments. *Philos. Trans. R. Soc. Lond. B* 344–403–410. [9]

Griffiths R. C., and S. Tavare. 1998. The age of a mutation in a general coalescent tree. *Stoch. Models* 14:273—95 [9]

Gross, B. L., and K. M. Olsen. 2010. Genetic perspectives on crop domestication. *Trends Plant Sci.* 15: 529–537. [9]

Grossman, S. R., K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. K. Bamezai, A. V. S. Hill, F. O. Vannberg, J. L. Rinn, 1000 Genomes Project, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152: 703–713. [9]

Grossman, S. R., I. Shlyakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327: 883–886. [9]

Guinand, B., C. Lemaire, and F. Bonhomme. 2004. How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including fishes. *J. Sea Research* 51: 167–182. [9]

Hallatschek, O. 2011. The noisy edge of traveling waves. *Proc. Natl. Acad. Sci. USA* 108: 1783–1787. [9]

Hallatschek, O., P. Hersen, S. Ramanathan, and D. R. Nelson. 2007. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl. Acad. Sci. USA* 104: 19926–19930. [9]

Hallatschek, O., and D. R. Nelson. 2008. Gene surfing in expanding populations. *Theor. Pop. Biol.* 73: 158–170. [9]

Hallatschek, O., and D. R. Nelson. 2009. Population genetics and range expansion. *Physics Today* 62: 42–48. [9]

Hamblin, M. T., A. M. Casa, H. Su, S. C. Murray, A. H. Paterson, C. F. Auadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from *Sorhum bicolor*. *Genetics* 173: 953–964. [9]

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, Re. Sukernik, G. Utermann, J. Pritchard, G. Coop, and A. Di Rienzo. 2010a. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc. Natl. Acad Sci. USA* 107: 8924–8930. [9]

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky, and A. Di Rienzo. 2010b. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Phil. Trans. R. Soc. B* 365: 2459–2468. [9]

Hanchard, N. A., K. A. Rockett, C. Spencer, G. Coop, M. Pinder, M. Jallow, M. Kimer, G. McVean, R. Mott, and D. P. Kwiatkowski. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* 78: 153–159. [9]

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman, E. H. Coe, and J. F. Doebley. 1996. Evolution of anthocyanin biosynthesis in maize kernels: The role of regulator and enzymatic loci. *Genetics* 143: 1395–1407. [9]

Harr, B., M. Kauer, and C. Schlötterer. 2002. Hitchhiking mapping: a population-based fine mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 99: 12949–12954. [9]

Harris, E. E., and D . Meyer. 2006. The molecular signature of selection underlying human adaptations. *Yearbook of Phy. Antro.* 49: 89–130. [9]

Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPatlan, I. MacLeod, L. Sethuraman, and M. E. Goodard. 2008. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.* 40: 176–184. [9]

He, Z., W. Zhai, H. Wen, T. Tang, Y. Wang, X. Lu, A. J. Greenberg, R. R. Hudson, C.-I. Wu, and S. Shi. 2011. Two evolutionary histories in the genomes of rice: the roles of domestication genes. *PLoS Genetics* 7: e1002100. [9]

Hermisson, J. 2009. Who believes in whole-genome scans for selection? *Heredity* 103: 283–284. [9]

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, and M. Przeworski. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331: 920–924. [9]

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Bio. Evol.* 24: 1782–1800. [9]

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 226–231. [9]

Hofer, T., N. Ray, D. Wegmann, and L. Excoffier. 2009. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann. Hum. Gen.* 73: 95–08. [9]

Huang, X., and 34 others. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490: 497–501. [8, 9]

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18: 337–338. [9]

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* 136: 1329–1340. [9]

Hudson, R. R., and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164. [9]

Hufford, K. M., P. Canaran, D. H. Ware, M. D. McMullen, and B. S. Gaut. 2007. Patterns of selection and tissue-specific expression among maize domestication and crop improvement loci. *Plant Physiol.* 144: 1642–1653. [9]

Hufford, M. B. and 24 others. 2012. Comparative population genomics of maize domestication and improvement. *Nature Gen.* 44: 808–813. [9]

Hughes, A. L. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99: 364–373. [9]

Innan, H., and W. Stephan. 2000. The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana. Genetics* 155: 2015–2019. [9]

Innan, H., K. Zhang, P. Marjoram, S. Taveré, and N. A. Rosenberg. 2005. Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* 169: 1763–1777. [9]

Inukai, T., A. Sako, H.-Y. Hirano, and Y. Sano. 2000. Analysis of intragenic recombination at *wx* in rice: correlation between the molecular and genetic maps within the locus. *Genome* 43: 589–596. [9]

Jaenicke-Després, V., E. S. Buckler, B. D. Smith, M. T. P. Gilbert, A. Cooper, J. Doebley, and S. Pääbo. 2003. Early allelic selection in maize as revealed by ancient DNA. *Science* 302: 1206–1208. [9]

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170: 1401–1410. [9]

Jensen, J. D., K. R. Thornton, C. D. Bustamante, and C. F. Aquadro. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* 176: 2371–2379. [9]

Johansson, A. M., M. E. Pettersson, P. B. Siegel, and Ö. Carlboorg. 2010. Genome-wide effects of long-term divergent selection. *PLoS Genet.* 6: e1001188. [9]

Johnson, P. L., and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25: 199-–206. [9]

Jones, D. A. 1989. 50 years of studying the scarlet tiger moth. *Trends Ecol. Evol.* 4: 298–301. [9]

Kane, N. C., and L. H. Riesenberg. 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus. Genetics* 175: 1823–1834. [9]

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisted. *Genetics* 123: 887–899. [9]

Kauer, M. O., D. Dieringer, and C. Schlötterer. 2003. A microsatellite variability screen for positive selection associated with the "Out of Africa" habitat expansion of *Drosophila melanogaster*. *Genetics* 165: 1137–1148. [9]

Kayser, M., S. Brauer, and M. Stoneking. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol. Biol. Evol* 20: 893–900. [11]

Keightley, P. D., and G. Bulfield. 1993. Detection of quantitative trait loci from frequency changes of marker alleles under selection *Genet. Res.* 62:195–203. [9]

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. *Genetics* 146: 1197–1206. [9]

Kemper, K. E., S. J. Saxton, S. Bolormaa, B. J. Hayes, and M.E. Goddard. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genom.* 15: 246. [8, 9]

Kern, A. D., and D. Haussler. 2010. A population genetic hidden Markov model for detecting genomic regions under selection. *Mol. Biol. Evol.* 27: 1673–1685. [9]

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513–1524. [9]

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777. [9]

Kimura, R., A. Fujimoto, K. Tokunaga, and J. Ohashi. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS One* 3: e286. [9]

King, M., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116. [9]

Kirby, D. A., and W. Stephan. 1995. Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster. Genetics* 141: 1483–1490. [9]

Klopfstein, S. M. Currat, and L. Excoffier. 2006. The fate of mutations surfing on the wave of a range expansion. *Mol. Biol. Evol.* 23: 482–490. [9]

Kochert, G., H. T. Stalker, M. Gimenes, L. Galgaro, C. R. Lopes and K. Moore. 1996. RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *Am. J. Botany* 83: 1282–1291. [9]

Kohn, M. H., H.-J. Pelz, and R. K. Wayne. 2000. Natural selection mapping of the warfarin-resistance gene. *Proc. Natl. Acad. Sci. USA* 97: 7911-7915. [9]

Konishi, S., T. Izawa, S. Y. Lin, K. Ebana, Y. Fukuta, T. Sasaki, and M. Yano. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* 312: 1293–1396. [9]

Kreitman, M. 2000. Methods to detect selection in populations with applications to the human. *Ann. Rev. Genomics Hum. Genet.* 1: 539–559. [9]

Labate, J. A., K. R. Lamkey, M. Lee, and W. L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. *Theor. Appl. Gene.* 99: 1166–1178. [9]

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics* 74: 175–195. [9]

Li, C., A. Zhou, and T. Sang. 2006. Rice domestication by reducing shattering. *Science* 311: 1936–1939. [9]

Li, H., and W. Stephan. 2005. Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171: 377–384. [9]

Li, H., and W. Stephan. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila. PLoS Genet.* 2: e166. [9]

Li, Y. F., J. C. Costello, A. K. Holloway, and M. W. Hahn. 2008. "Reverse ecology" and the power of population genomics. *Evolution* 62: 2984–2994. [9]

Liang, L., S. Zöllner, and G. R. Abecasis. 2007. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics* 23: 1565–1567. [9]

Lin, K., H. Li, C. Schlötterer, and A. Futschik. 2011. Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics* 187: 229–244. [9]

Lohmueller, K. E., A. Albrechtsen, Y.Li, S. Y. Kim, T.Korneliussen, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, A. F. Feder, N. Grarup, R. Jørgensen, T. Jiang, D. R. Witte, A. Sandbaek, I. Hellmann, T. Laurtitzen, T. Hansen, O. Pederson, J. Wang, and R. Nielsen. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* 7: e1002326. [9]

Londo, J. P., Y.-C. Chiang, K.-H. Hung, T.-Y. Chiang, and B. A. Schaal. 2006. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. USA* 103: 9578–9583. [9]

Macpherson, J. M., J. González, D. M. Witten, J. C. Davis, N. A. Rosenberg, A. E. Hirsh, and D. A. Petrov. 2008. Nonadaptive explanations for signatures of partial sweeps in *Drosphila. Mol. Biol. Evol.* 25: 1025–1042. [9]

Manicacci, D., M. Falque, S. Le Guillou, B. Peégu, A.-M. Henry, M. Le Guilloux, C. Damerval, and D. De Vienne. 2007. Maize *Sh2* gene is constrainted by natural selection but escaped domestication. *J. Evol. Biol.* 20: 503–516. [9]

Markovtsova, L., P. Marjoram, and S. Taveré. 2001. On a test of Depaulis and Veuille. *Mol. Biol. Evol.* 18: 1132–1133. [9]

Matsuoka, Y., Y. Vigouroux , M. M. Goodman, J. Sanchez G., E. Buckler, and J. Doebley. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA* 99: 6080–6084. [9]

McPeek, M. S. and A. Strahs. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Amer. J. Hum. Gene.* 65: 858–875. [9]

Meiklejohn, C. D., Y. Kim, D. L. Hartl, and J. Parsch. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* 168: 265–279. [9]

Miyashita, N., and C. H. Langley. 1988. Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster. Genetics* 120: 199–212. [9]

Montgomery, M. E., L. M. Woodworth, P. R. England, D. A. Briscoe, and R. Frankham. 2010. Widespread selective sweeps affecting microsatellites in *Drosophila* populations adapting to captivity: Implications for captive breeding programs. *Biol. Conserv.* 143: 1842–1849. [9]

Mueller, L. D., B. A. Wilcox, P. R. Ehrlich, D. G. Heckel, and D. D. Murphy. 1985. A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha*. *Genetics* 110: 495–511. [9]

Nawa, N., and F. Tajima. 2008. Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. *Genes Genet. Syst.* 83: 353–360. [9]

Nei, M., and T. Maruyama. 1975. Lewontin-Krakauer test for neutral genes. *Genetics* 80: 395. [9]

Nei, M., Y. Suzuki, and M. Nozawa. 2010. The neutral theory of molecular evolution in the genomic era. *Ann. Rev. Genom. Human Genet.* 11: 265–289. [9]

Nei, M., and F. Tajima. 1981. Genetic drift and estimation of effective population size. *Genetics* 98: 253–259. [9]

Nicolas, F. W., and A. Robertson. 1976. The effect of selection on the standardized variance of gene frequency. *Theor. Appl. Gene.* 48: 263–266. [9]

Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86: 641–647. [9]

Nielsen, R. 2005. Molecular signatures of natural selection. *Ann. Rev. Genet.* 39: 197–218. [9]

Nielsen, R., M. J. Hubisz, and A. G. Clark. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168: 237–2382. [9]

Nielsen, R., M. J. Hubsiz, I. Hellmann, D. Torgerson, A. M. Andrés, A. Albrechtsen, R. Gutenkunst, M. D. Adams, M. Cargill, A. Boyko, A. Indap, C. D. Bustamante, and A. G. Clark. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19: 838–849. [9]

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15: 1566–1575. [9]

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154: 923–929. [9]

Nuzhdin, S. V., P. D. Keightley, and E. G. Pasyukova. 1993. The use of retrotransposons as markers for mapping genes responsible for fitness differences between related *Drosophila melanogaster* strains. *Genet. Res.* 62: 125–133. [9]

Nuzhdin, S. V., and E. G. Pasyukova. 1991. New approach to polygene mapping. Location of genes controlling fitness level in related *Drosophila melanogaster* stocks. *Genetika* (Russ) 27: 849–859. [9]

Novembre, J., and A. Di Rienzo. 2009. Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.* 10: 745–755. [9]

O'Hare, B. 2005. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. *Proc. Royal Soc. London B.* 272: 211–217. [9]

Oleksyk, T. K., M. W. Smith, and S. J. O'Brien. 2010. Genome-wide scans for footprints of natural selection. *Phil. Trans. R. Soc. B* 365: 185–205. [9]

Oleksyk, T. K., K. Zhao, F. M. De La Vega, D. A. Gilbert, S. J. O'Brien, and M. W. Smith. 2008. Identifying selected regions from heterzygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* 3: e1712. [9]

Ollivier, L., L. A. Messer, M. F. Rothschild, and C. Legault. 1997. The use of selection experiments for detecting quantitative trait loci. *Genet. Res.* 69: 227–232. [9]

Olsen, K. M., A. L. Caicedo, N. Polato, A. McClung, S. McCouch, and M. D. Purugganan. 2006. Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. *Genetics* 173: 975–983. [9]

O'Reilly, P. F., E. Birney, and D. J. Balding. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection *Genome Res.* 18: 1304–1313. [9]

Palaisa, K., M. Morgante, S. Tingey, and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. USA* 101: 9885–9890. [9]

Pavlidis, P., J. D. Jensen, and W. Stephan. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* 185: 907–922. [9]

Peng, Y., Z. Yang., H. Zhang, C. Cui, X. Qi, X. Luo, X. Tao, T. Wu, Ouzhuluobu, Basang, Ciwangsangbu, Danengduojie, H. Chen, H. Shi, and B. Su. 2011. Genetic variations in Tibetan populations and high-altitude adapation at the Himalayas. *Mol. Biol. Evol.* 28: 1075–1081. [9]

Pennings, P. S., and J. Hermisson. 2006b. Soft sweeps III – The signature of positive selection from recurrent mutation. *PLoS Genetics* 2: e186. [9]

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivsan, G. S. Barsh, R. M. Meyers, M. W. Feldman, and J. K. Prichard. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19: 826–837. [9]

Prasad, A., R. D. Schnabel, S. D. McKay, B. Murdoch, P. Stothard, D. Kolbehdari, Z. Wang, J. F. Taylor, and S. S. Moore. 2008. Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. *Anim. Genet.* 39: 597–605. [9]

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179–1189. [9]

Pollinger, J. P., C. D. Bustamante, A. Fledel-Alon, S. Schmutz, M. M. Gray, and R. K. Wayne. 2005. Selective sweep mapping of genes with large phenotypic effects. *Genome* 15: 1809–1819. [9]

Ptak, S. E., and M. Przeworski. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Gen.* 18: 559–563. [9]

Purugganan, M. D., A. L. Boyles, and J. I. Suddith. 2000. Variation and selection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea. Genetics* 155: 855–862. [9]

Purugganan, M. D., D. Q. Fuller. 2009. The nature of selection during plant domestication. *Nature* 457: 843–848. [9]

Ramos-Onsins, S. E., and J. Rozas. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* 19: 2092–2100. [9]

Randhawa, I. A., M. S. Khatkar, P. C. Thomson, and H. W. Raadsma. 2014. Composite selection signals can l ocalize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics* 15: 34. [9]

Risch, N., D. de Leon, L. Ozelius, P. Kramer, L. Almasy, B. Singer, S. Fahn, X. Breakefield and S. Bressman. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat. Gene.* 9: 152–159. [9]

Riebler, A., L. Held, and W. Stephan. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178: 1817–1829. [9]

Robertson, A. 1975a. Remarks on the Lewontin-Krakauer test. *Genetics* 80: 396. [9]

Robertson, A. 1975b. Gene frequency distributions as a test of selective neutrality. *Genetics* 81: 775–785. [9]

Rockman, M. V., M. W. Hahn, N. Soranzo, D. B. Goldstein, and G. A. Wray. 2003. Positive selection on a human-specific transcription factor binding site regulating *IL4* expresion. *Curr. Biol.* 13: 2118–2123. [9]

Ronald, J., and J. M. Akey. 2005. Genome-wide scans for loci under selection in humans. *Hum. Genomics* 2: 113–125. [9]

Ross-Ibarra, J., P. L. Morrell, and B. S. Gaut. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. USA* 104: 8641–8648 [9]

Ross-Ibarra, J., S. I. Wright, J. P. Foxe, A. Kawabe, L. DeRose-Wilson, G. Gos, D. Charlesworth, and B. S. Gaut. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* 3: e2411. [9]

Rozas, J., M. Gullaud, G. Blandin, and M. Aguadé. 2001. DNA variation at the *rt49* geene region of *Drosophila simulans:* evolutionary inferences from an unusual haplotype structure. *Genetics* 158: 1147–1155. [9]

Rozas, J., C. Segarra, G. Ribó, and M. Aguadé. 1999. Molecular population genetics of the *rp49* gene region in different chromosomal inversions of *Drosophila subobscura*. *Genetics* 151: 189–202. [9]

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837. [9]

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander 2006. Positive natural selection in the human lineage. *Science* 312: 1614-1620. [9]

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Costsapas, X. Xie, E. H. Bryne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, and The International HapMap Consortium. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913–919. [9]

Sandoval-Castellanos, E. 2010. Testing temporal changes in allele frequencies : a simulation approach. *Genet. Res.* 92: 309–320. [9]

Sang. T., and S. Ge. 2007. Genetics and phylogenetics of rice domestication. *Curr. Opin. Genet. Devel.* 17: 533–538. [9]

Sattah, S., E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Gen.* 7: e1001302. [9]

Schaeffer, S. W. 2002. Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura. Genet. Res.* 80: 163–175. [9]

Schaffer, H. E., D. Yardley, and W. W. Anderson. 1977. Drift or selection: A statistical test of gene frequency variation over generations. *Genetics* 87: 371–379. [9]

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15: 1576–1583. [9]

Schöft, G., and C. Schlötterer. 2004. Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African *D. simulans. Mol. Biol. Evol.* 21: 1384–1390. [9]

Schlötterer, C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160: 753–763. [9]

Schlötterer. 2003. Hitchhiking mapping – functional genomics from the population genetics perspective. *Trends Gene.* 19: 32–38. [9]

Schlötterer, C., C. Vogl, and D. Tautz. 1997. Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. *Genetics* 146: 309–320. [9]

Serre, J. L., B. Simon-Bouy, E. Mornet, B. Jaume-Roig, A. Balassopoulou, M. Schwartz, A. Taillandier, J. Boué and A. Boué. 1990. Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Human Gen.* 84: 449–454. [9]

Simons, K. J., J. P. Fellers, H. N. Trick, Z. Zhang, Y.-S. Tai, B. S. Gill, and J. D. Aris. 2006. Molecular characterization of the major wheat domestication gene *Q*. *Genetics* 172: 547–555. [9]

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413–429. [9]

Siol, M., S. I. Wright, and S. C. H. Barrett. 2010. The population genomics of plant adaptation. *New Phytol.* 188.: 313–332. [9]

Slatkin, M. 1994. An exact test for neutrality based on the Ewens sampling distribution. *Genet. Res.* 64: 71–74. [9]

Slatkin, M. 1995a. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462. [9]

Slatkin, M. 1996. A correction to the exact test based on the Ewens sampling distribution. *Genet. Res. Camb.* 68: 259–260. [9]

Slatkin, M. 2000. Allele age and a test for selection on rare alleles. *Philos. Trans. R. Soc. Lond.* B 355: 1663–1668. [9]

Slatkin, M. 2008. A Bayesian method for jointly estimating allele age and selection intensity. *Genet. Res.* 90: 119–128. [9]

Slatkin, M., and G. Bertorelle. 2001. The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* 158: 865–687. [9]

Slatkin, M., and B. Rannala. 1997. Estimating the age of alleles by the use of intraallelic variability. *Am. J. Hum. Genet.* 60: 447–458. [9]

Slatkin, M., and B. Rannala. 2000. Estimating allele age. *Ann. Rev. Genomics Hum. Genet.* 1: 225–249. [9]

Spooner, D. M., K. McLean, G. Ramsay, R. Wuagh, and G. J. Bryan. 2005. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proc. Natl. Acad. Sci. USA* 102: 14694–14699. [9]

Stephan, W. 2010a. Detecting strong positive selection in the genome. *Mol. Ecol. Res.* 10: 863–872. [9]

Stephens, J. C., and 38 others. 1998. Dating the origin of the *CCR5-Delta32* AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* 62: 1507—1515. [9]

Stephens, M. N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68: 978–989. [9]

Stephens, M., and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Gen.* 7: 449–462. [9]

Stinchcombe, J. R., and H. E. Hoekstra. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100: 158–170. [9]

Strobeck, C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117: 149 – 153. [9]

Stuber. C. W., and R. H. Moll. 1972. Frequency changes of isozyme alleles in a selection experiment for grain yield in maize (*Zea maize* L.). *Crop Sci.* 12: 337–340. [9]

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley. 2011. Identification of a functional transposon insertion in the maize domesticaion gene *tb1*. *Nature Gen.* 43: 1160–1165. [9]

Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14: 671-688. [9]

Stuber, C. W., R. H. Moll, M. M. Goodman, H. E. Schaffer, and B. S. Weir. 1980. Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays* L.) *Genetics* 95: 225–236. [9]

Swanson-Wagner, R., R. Briskine, R. Schaefer, M. B. Huffrd, J. Ross-Ibarra, C. L. Myers, P. Tiffin, and N. M. Springer. 2012. Reshaping of the maize transcritptome by domestication. *Proc. Natl. Acad. Sci. USA* 109: 11878–11883. [9]

Tajima, F. 1989. Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595. [9]

Tang, K., K. R. Thornton, and M. Stoneking. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Bio.* 5: e171. [9]

Taylor, M. F. J., Y. Shen, and M. E. Kreitman. 1995. A population genetic test of selection at the molecular level. *Science* 270: 1497–1499. [9]

Templeton, A. R. 1974. Analysis of selection in populations observed over a sequence of consecutive generations. *Theor. Appl. Genet.* 45: 179–191. [9]

Tenaillon, M. I., J. U'Ren, O. Tenaillon, and B. S. Gaut. 2004. Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* 21: 1214–1225. [9]

Teshmia, K. M., G. Coop, and M. Przeworski. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16: 702–712. [9]

Thomas, M. G., K. Skorecki, H. Ben-Ami, T. Parfitt, N. Bradman, and D. B. Goldstein. 1998. Origins of old testament priests. *Nature* 395: 138–140. [9]

Thornton, K. R., and J. D. Jensen. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175: 737–750. [9]

Thornton, K. R., J,. D. Jensen, C. Becquest, and P. Andolfatto. 2007. Progress and prospects in mapping recent selection in the genome. *Heredity* 98: 340—348. [9]

Toomajian, C., R. S. Ajioka, L. B. Jorde, J. P. Kushner, and M. Kreitman. 2003. A method for detecting recent selection in the human genome from allele age estimates. *Genetics* 165: 287–297. [9]

Toomajian, C., T. T. Hu, M. J. Aranzana, C. Lister, C. Tang., H. Zheng, K. Zhao, P. Calabrese, C. Dean, and M. Nordborg. 2006. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Bio.* 4: e137. [9]

Travis, J. M. J., T. Munkemuller, O. J. Burton, C. Dytham, and K. Johst. 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Mol. Biol. Evol.* 24: 2334–2343. [9]

Tsakas, S., and C. B. Krimbas. 1976. Testing the heterogeneity of F values: A suggestion and a correction. *Genetics* 84: 399–401. [9]

Turchin, M. C., C. W. K. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, J. N. Hirschhorn, and Genetic Investigation of ANthropometric Traits (GIANT) Consortium. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Gene.* 44 1015–1019. [9]

Utsunomiya, Y. T., A. M. P. O'Brien, T. S. Sonstegard, C. P. Van Tassell, A. S. do Carmo, G. Meszaros, J. Soelkner, and J. F. Garcia. 2013. Detecting Loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PloS One* 8: e64280. [9]

Vigouroux, Y., M. McMullen, C. T. Hittinger, K. Houchins, L. Schulz, S. Kresovich, Y. Matsuoka, and J. Doebley. 2002. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl. Acad. Sci. USA* 99: 9650–9655. [9]

Vitalis, R., K. Dawson, and P. Boursot. 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158: 181–1182. [9]

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. *PLoS Bio.* 4: e72. [9]

Wall, J. D. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res.* 74: 65–79. [9]

Wall, J. D., and R. R. Hudson. 2001. Coalescent simulations and statistical tests of neutrality. *Mol. Biol. Evol.* 18: 1134–1135. [9]

Wang, E. T., G. Kodama, P. Baldi, and R. K. Moyzis. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* 103: 135–140. [9]

Wang, H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigourous, M. Faller, K. Bomblies, L. Lukens, and J. F. Doebley. 2005. The origin of the naked grains of maize. *Nature* 436: 714–719. [9]

Wang, R.-L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. *Nature* 398: 236–239. [9]

Wang, X., W. E. Grus, and J. Zhang. 2006. Gene losses during human origins. *PLoS Biol.* 4: e52. [9]

Waples, R. S. 1989a. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121: 379–391. [9]

Waples, R. S. 1989b. Temporal variation in allele frequencies: testing the right hypothesis. *Evolution* 43: 1236–1251. [9]

Watterson G. A. 1976. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.* 10: 239–253. [9]

Watterson, G. A. 1977. Heterosis or neutrality? *Genetics* 85: 789–814. [9]

Watterson, G. A. 1978. The homozygosity test for neturality. *Genetics* 38: 405–417. [9]

Watterson, G. A. 1982. Testing selection at a single locus. *Biometrics* 38: 323–331. [9]

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut, and E. S. Buckler IV. 2002. Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci. USA* 99: 12959–12962. [9]

Wiehe, T., V. Nolte, D. Zivkovic, and S. Schlötterer. 2007. Identification of selective sweeps using a dynamically adjusted number of linked microsaellites. *Genetics* 175: 207–218. [9]

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3: e90. [9]

Wilson, I. J., M. E. Weale, and D. J. Balding. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. Roy. Stat. Soc. A* 166: 155–201. [9]

Wilson, S. R. 1980. Analyzing gene-frequency data when the effective population size is finite. *Genetics* 95: 489–502. [9]

Wisser, R. J., S. C. Murray, J. M. Kolkman, H. Ceballos, and R. J. Nelson. 2008. Selection mapping of loci of quantitative disease resistance in a diverse maize population. *Genetics* 180: 583–599. [9]

Wootton, J. C., X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu, D. I. Bauruch, A. J. Magill, and X. Su. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418: 320–323. [9]

Wright, S. 1948. On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* 2: 279–294. [9]

Wright, S. 1951. The genetical structure of populations. *Ann. Eugen.* 15: 323–354. [9]

Wright, S. I., J. V. Bi, S. G. Schroeder, M. Yamasaski, J. D. Doebley, M. D. McMullen, and B. S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* 308: 1310–1314 [9]

Wright, S. I., and B. S. Gaut. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Mol. Bio. Evol.* 22: 506-519. [9]

Xia, Q. and 57 others. 2009. Complete resequencing of 40 genomes reveals domesticiation events and genes in silkmorm (*Bombyx*). *Science* 326: 433–446. [9]

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley, B. S. Gaut, and M. D. McMullen. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* 17: 2859–2872. [9]

Yi, X. and 70 others. 2010. Sequencing of 50 human exomes reveals adapatation to high altitude. *Science* 329: 75–78. [9]

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1431–1439. [9]

Zeng, K., S. Shi, and C.-I. Wu. 2007. Compound tests for the detection of hitchhiking under positive selection. *Mol. BIol. Evol.* 24: 1898–1908. [9]

Zeder, M. A., E. Emshwiller, B. D. Smith, and D. G. Bradley. 2006. Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* 22: 139–155. [9]

Zhao, O., A. L. Weber, M. D. McMullen, K. Guill, and J. Doebley. 2010. MADS-box genes in maize: frequent targets of selection during domestication. *Genet. Res.* 93: 65–75. [9]

Ŝivković, D., and T. Wiehe. 2008. Second-order moments of segregating sites under variable population size. *Genetics* 180: 341–357. [9]

Zohary, D. 1999. Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. *Genet. Res. Crop Evol.* 46: 133-142. [9]

Zuckerkandl, E. 1968. Hemoglobins, Haeckel's "Biogenic Law" , and molecular aspects of development. *In* Rich and N. Davidson (eds), *Structural chemistry and molecular biology*, pp. 256–274. Freeman, San Francisco. [9]