Euphytica (2008) 161:1–17
DOI 10.1007/s10681-007-9465-8

Using molecular markers for detecting domestication, improvement, and adaptation genes

Bruce Walsh

Received: 1 December 2006 / Accepted: 15 May 2007 / Published online: 14 June 2007

Abstract Development of statistical tests to detect selection (strictly speaking, departures from the neutral equilibrium model) has been an active area of research in population genetics over the last 15 years. With the advent of dense genome sequencing of many domesticated crops, some of this machinery (which heretofore has been largely restricted to human genetics and evolutionary biology) is starting to be applied in the search for genes under recent selection in crop species. We review the population genetics of signatures of selection and formal tests of selection, with discussions as to how these apply in the search for domestication and improvement genes in crops and for adaptation genes in their wild relatives. Plant domestication has specific features, such as complex demography, selfing, and selection of alleles starting at intermediate frequencies, that compromise many of the standard tests, and hence the full power of tests for selection has yet to be realized.

Keywords Selective sweeps · Detecting selection · Genomic scans

B. Walsh (✉)
Department of Ecology and Evolutionary Biology,
University of Arizona, Tucson, AZ 85721, USA
e-mail: jbwalsh@u.arizona.edu

Domestication, improvement, and adaptation genes

Localizing genes critical to the domestication of modern crops from their ancestral wild relatives and genes subsequently involved in significant improvement following the initial domestication has both agronomic and intellectual importance. How might such genes be found? If one wishes to specify the traits thought to be involved in domestication or improvement, standard QTL approaches can be used, either by examining specific candidate genes or by applying more general genomic scans using line crosses or association studies (e.g., Lynch and Walsh 1997). In many cases of domestication, improvement, or adaptation, the actual traits that have been modified are unknown, especially if these are not morphological, but rather physiological or biochemical traits. How can we logically search for domestication/improvement genes when the actual traits are unspecified? Further, even if a candidate region is segregating variation that influences a proposed trait, how can one formally show that this region was actually involved in a domestication or improvement event?

The answer comes from a rather rich population genetic literature on tests of departures from the neutral equilibrium model (drift and mutation in a random-mating population of constant size). Here we review the basic logic behind such tests, examine a few key tests in detail, and look at some of the

limitations and open questions raised by this approach as concerns their use in the search for plant domestication and adaptation genes. Our intended audience for this review are plant breeders with (at best) a passing knowledge of this area. A number of reviews of tests of selection have appeared (a partial, but not exhaustive, list includes Kreitman 2000; Nielsen 2001; Ford 2002; Schlotterer 2003; Storz 2005; Wright and Gaut 2005; Nielsen 2005; Sabeti et al. 2006; Biswas and Akey 2006). Our aim here is not to compete with these previous reviews, but rather introduce this area to an important audience that may not have had significant exposure to this developing field. We have tried (perhaps unsuccessfully) to walk a fine line between introducing the key concepts while still presenting enough of the formal theory to allow the more inquisitive reader to follow some of the more technical developments.

While QTL/association studies look for marker-trait associations at a target locus, tests of signatures of selection look directly at the pattern of molecular variation at a target locus, ignoring any trait information. Hence, one does not need to specify a trait (or traits) as would be required in a search for marker-trait associations. While both QTL scans and scans for selection rely on linkage and a dense set of highly variable markers, we can regard them as complementary approaches. A QTL scan might detect candidate genes influencing a trait, while tests of selection on these genes can directly look for past (or present) signatures of selection at these sites.

There is a rich population genetics literature on tests of whether an observed pattern of polymorphism and/or observed between-population/species difference can be accounted for with the standard neutral model. Rejection of this hypothesis offers the *possibility* that selection may play a role, but (as we will see) other forces (such as changes in population size and population subdivision) can also cause strong deviations from the neutral equilibrium model. Tests for departures from neutrality roughly fall into three categories, depending on whether they require only within-species (polymorphism) data, or only between-species (divergence) data, or both.

Examples of tests based solely on the pattern and amount of within-species polymorphism include Tajima's D, Fu and Li's D^* and F^*, Fu's W and F_S, and Fay and Wu's H (all discussed below). These tests can detect on-going or recent selection, but are also strongly influenced by demographic history (such as a recent population bottleneck).

Two of the most widely applied tests, the McDonald–Kreitman and the Hudson, Kreitman and Aguade (or HKA) tests, use both within- and between-species variation (polymorphism and divergence). These tests also allow one to detect on-going or recent selection. The final approach are those tests based solely on phylogenetic comparisons between species. These only detect signatures of rather historical (and repeated) selection, and we will not discuss these further (see Nielsen and Yang 1998; Nielsen 2005 for a review).

The structure of this paper is as follows. We first review some basic features of the neutral model, and then consider the typical signature (a selective sweep) left in a region under recent directional selection. Next, we review recent attempts to look for signatures of selection caused by domestication. This is followed with a review of several of the most popular formal tests for selection and limitations of these approaches for the detection of domestication genes. We conclude by discussing several unresolved issues in applying these approaches in plant breeding.

Within- and between-species patterns of variation under strict neutrality

In order to look for departures in the pattern of within- or between-species variation due to selection, we first need to consider what these patterns are expected to look like under strict neutrality. Hence, we start with a brief digression reviewing classic results from the neutral theory (e.g., Kimura 1983).

The effect of finite population size (genetic drift) is to remove genetic variation, with the decay of existing variation being fastest in small populations. Conversely, while drift is removing variation, mutation is creating it. Hence, with a constant population size, a balance between these two forces results in a standing level of variation occurring within a population. The parameter that describes how much equilibrium variation is expected within a population is $\theta = 4N_e\mu$, where μ is the (neutral) per-generation mutation rate for the locus under consideration and N_e the effective population size.

To see where this parameter comes from, we turn to a very powerful approach for thinking about drift,

namely the *coalescent process* (Rosenberg and Nordborg 2002 provide a nice introduction to coalescent theory, while Hein et al. 2005 provide a more detailed treatment). Here one follows lineages instead of alleles, and we think of the lineages merging, or coalescing, as we move back in time, with all of the current lineages eventually tracing back to a single most recent common ancestor (MRCA). If the time until the MRCA is very short, there is little chance for mutation to generate new variation within the resulting daughter lineages, while if the time is long, considerable variation can arise. Specifically, if the time back to the MRCA for two sequences (*A* and *B*) being compared is *t* generations, then we expect (on average) $t\mu$ mutations to arise as we move from the sequence in the MRCA to sequence *A* and likewise $t\mu$ mutations when moving from the MRCA to sequence *B*, or an expected total of $2t\mu$ mutations between the two sequences. Since mutations occur at random, the actual number of mutational differences between these two sequences follows a Poisson distribution with success parameter $2t\mu$. If N_e is the effective size of a population, then for two randomly drawn sequences, the expected time back to their MRCA under pure drift is just $2N_e$ (for a diploid). Hence, the expected number of mutational differences between these two sequences is $2t\mu = 4N_e\mu = \theta$.

More generally, we can compute the distribution of variation within a sample of k sequences from a population of size N by considering the time until the two most recent lineages within the sample coalesce, now leaving $k-1$ distinct lineages and so forth until we coalesce all of the k sampled lineages back to a single MRCA (Fig. 1). The first coalescent time (t_k, moving from k to $k-1$ lineages) follows a geometric distribution with success parameter $k(k-1)/4N$ (Hein et al. 2005), and hence an expected time of $4N/[k(k-1)]$. This leaves $k-1$ lineages, and the time t_{k-1} back to the next MRCA (leaving $k-2$ lineages) follows a geometric with parameter $(k-1)(k-2)/4N$. One continues until all of the lineages coalescent to a single MRCA, and the resulting joint distribution of coalescent times (under pure drift) is entirely determined by the sample size k and effective population size N_e (which more generally replaces N in our above discussion). Onto this distribution of coalescent times one overlays the particular mutation model being assumed (e.g., infinite alleles, infinite sites, step-wise) to obtain a complete probabilistic model of the expected variation in the sample.

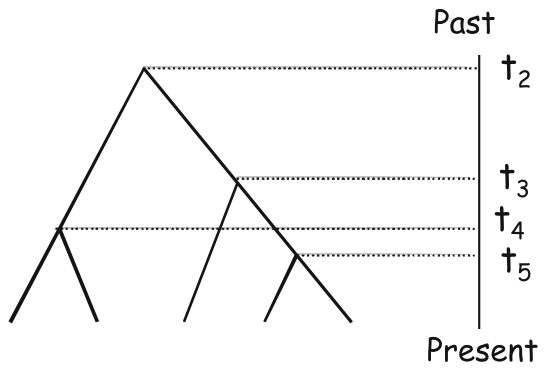

Fig. 1 The expected coalescent for a sample of five sequences from a (diploid) population of size N. t_5 denotes the first coalescence event, the time in which we move from five distinct lineages to four distinct lineages because we reach the MRCA of the two most recent. The expected time for this is $4N/[k(k-1)] = 4N/(5*4) = N/5$. Likewise, the time to move from four distinct lineages to three distinct lineages, t_4 has expected value $4N/(4*3) = N/3$. Similarly, t_3 has expected value $N/1.5$, and t_2 a value of $2N$, for a total time to the MRCA of all sampled sequences of N (1/5 + 1/3 + 2/3 + 2) = 3.2N

Another consequence of drift is that populations will diverge over time as new mutations arise and are fixed by chance, creating a between-line (or population/species) divergence. For a diploid population of size N, on average $2N\mu$ new mutations are created each generation. Under the assumption of neutrality, each of these new mutations has probability $1/(2N)$ of being fixed. Thus, the expected number of new (neutral) mutations arising each generation that are destined to become fixed is just $2N\mu \cdot 1/(2N) = \mu$. Hence, under pure drift, the expected number of fixed differences within a line after t generations is just $t\mu$, giving the expected amount of divergence from two lines separated t generations ago as $d(t) = 2t\mu$. Between- and within-population variation thus behaves differently under drift, with the amount of within-population variation a function of $\theta = 4N_e\mu$, while the between-population is a function of $t\mu$, independent of population size.

Hitch-hiking, linkage drag, and selective sweeps

Against this neutral background, what is the nature of a signal that selection would leave? When selection

rapidly increases the frequency of an allele, linked sites also hitch-hike along for the ride (Maynard Smith and Haigh 1974). Plant breeders are aware of this phenomenon, namely linkage drag (Brinkman and Frey 1977), wherein a favorable introgressed region may drag along unfavorable linked genes. If the introgression is sufficiently rapid, the amount of linkage drag can be considerable. Likewise, when selection (natural or artificial) favors a particular allele, sites linked to that allele are also dragged along to fixation, resulting in a region around the selected site showing reduced genetic variation relative to the rest of the genome. Such a *selective sweep* occurs because the effect of selection is to reduce the effective population size at linked regions. This results in decreased times to the MRCA, and hence less polymorphism, due to a shortening of the coalescence times relative to pure drift. In the extreme case wherein the favorable allele starts as a single copy that is rapidly fixed, sites tightly linked to that region will also descend from this initial haplotype containing the favorable allele. The more rapid the fixation, the more reduced the level of variation around the favored site and the larger the size of the region influenced by the sweep. It is important to note, however, that while linkage may reduce the levels of standing variation through their reductions in N_e (and hence $\theta = 4N_e\mu$), linkage has essentially no effect on the average substitution rate at linked neutral sites. This is because (as discussed above), the per generation rate of divergence between neutral sites is simply the mutation rate, *independent* of population size.

Thus, one signal for selection is a reduced level of polymorphism relative to the rest of the genome, something that could be detected by either scoring a number of markers around a candidate gene or using a dense marker screen of the entire genome, the equivalents of candidate gene and genomic scan approaches (respectively) that one uses in QTL mapping. However, a significant reduction in the level of polymorphism (by itself) is by no means sufficient to indicate selection, as this could simply reflect a reduction in the neutral mutation rate in that region. Further, even under a neutral model, a plot of the polymorphism level over a chromsome can often show significant dips, simply due to sampling effects (Jensen at al. 2005). Thus, while regions of significantly decreased polymorphism are certainly suggestive, more formal tests are required, as we will review shortly.

As shown in Fig. 2, directional selection results in a reduction in the level of polymorphism at sites linked to the region under selection. Conversely, a region under overdominant selection will show an increase in the amount of polymorphism at linked neutral sites. Both of these observations can be thought of in terms of *time*. Under a selective sweep, markers tightly linked to a selected locus have a more recent common ancestor than the rest of the genome, while under overdominant selection, linked sites have a deeper (older) common ancestor. Figure 3 illustrates this idea by contrasting the coalescent times under pure drift, directional, and overdominant selection.

There is a final selective force that can cause a decrease in the level of polymorphism, namely

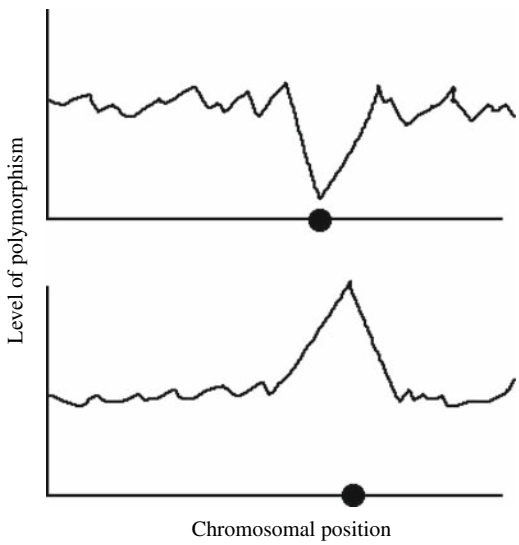

Fig. 2 The impact of selection on a particular site (filled circle) on variability at surrounding neutral sites. The vertical axis plots heterozygosity, the horizontal axis genome location. The upper graph shows the effect of a selective sweep, which results in a decrease in the levels of linked neutral polymorphisms around the selected site. The width of this window of reduction is a function of recombination (smaller c = larger width), selection (larger s = larger width) and time of the sweep (longer the time since the sweep, the smaller the width). Plots such as this using real data are generated by computing the variation in a window (of, say 100–1,000 bases) that we slide along the genome. The lower graph shows that balancing selection results in an *increase* in the level of linked neutral polymorphisms

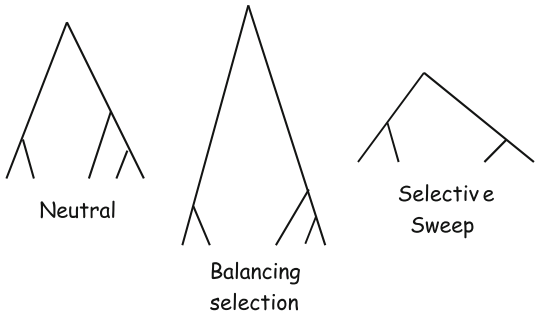

Fig. 3 Coalescent times under pure drift and two types of selection. Under *balancing selection* (by which we formally mean selective overdominance where the fitness of the heterozygote exceeds that of the homozygotes), the time back to the most recent common ancestor (MRCA) is longer than under pure drift. Under directional selection (often called a *selective sweep*), an allele sweeps through a population far quicker than under drift and hence has a more recent MRCA

background selection (Charlesworth et al. 1993, 1995). Here, selection *against* newly arising deleterious mutations also reduces the effective population size in a linked region around the selected site. Highly deleterious alleles have little impact, as such mutations are removed almost immediately. However, *slightly* deleterious mutations may drift up to low (but not rare) frequencies, and their removal has a larger impact. While the effect for any single removal may be minor, one expects significantly more deleterious mutations arising within a region than beneficial ones, and hence background selection can potentially have a rather significant effect. Further, it is very difficult to distinguish between selective sweeps (selection for new alleles) and frequent background selection (selection against new alleles), although we discuss some approaches for this below. Of special interest to us is that the effects of background selection can be quite significant in highly selfing plant populations (Charlesworth et al. 1993), due to their low effective recombination rates.

The nature and persistence of the signal left by a selective sweep has been extensively investigated by population geneticists. As one might except, the stronger the selection, the quicker an allele becomes fixed, and the larger the linked region influenced by the sweep. Likewise, as recombination decreases, the length of the sweep-influenced region increases. Specifically, Kaplan et al. (1989) showed that an approximation for the distance d at which a neutral site can be influenced by a sweep is a function of the strength of selection s and the recombination fraction c,

$$d \simeq 0.01 \frac{s}{c} \tag{1a}$$

Equation 1a thus allows for an estimate of the historical value of s given an observed size d and an estimate of the local recombination fraction,

$$s \simeq 100 \cdot d \cdot c \tag{1b}$$

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

The time for the sweep to occur, namely the time to fixation of a favorable allele, is approximately $2\ln(2N)/s$ generations, assuming that the favorable allele starts out as a single new mutation (Stephan et al. 1992). Assuming one starts with a single new favorable mutant, a crude approximate for the time t (in generations) for fixation can be expressed as

$$t \sim \frac{2\ln(2N)}{s} \sim \frac{2\ln(2N)}{100 \cdot d \cdot c} = 0.02 \frac{\ln(2N)}{d \cdot c} \tag{1c}$$

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

scenario has significant implications for the type of signature left by a selective sweep, making them much more difficult to detect. We will return to this point later.

Signatures of selective sweeps in maize and rice

Several groups have reported potential loci under selection in maize (Wang et al. 1999; Whitt et al. 2002; Vigouroux et al. 2002; Palaisa et al. 2004; Clark et al. 2004; Wright et al. 2005; Yamasaki et al. 2005) and rice (Olsen et al. 2006). Perhaps the most convincing example is the work of Doebley on the maize gene *teosinte branched 1*, henceforth *tb1* (Wang et al. 1999; Clark et al. 2004, 2006). This gene was originally detected in QTL mapping studies in crosses between maize and teosinte and subsequently shown to be the previously characterized locus *tb1*, which has major effects on plant architecture (Doebley et al. 1995). Given its obvious role as a candidate domestication gene, Wang et al. (1999) compared the levels of polymorphism around this locus in maize with the corresponding region in teosinte. Throughout this region, maize was found to have reduced levels of polymorphism (about 75%) relative to teosinte, but this is consistent with a bottleneck during domestication influencing all loci in modern maize (Tenaillon et al. 2004). More importantly, they observed a significant decrease in the amount of polymorphism in the 5′ NTR region of maize (but not teosinte) *tb1*, suggesting a selective sweep influenced this region. Surprisingly, the sweep did not influence the coding region, suggesting that the selected site was in the 5′ regulatory region, as opposed to selection on a change in the amino acid sequence of *tb1*. Clark et al. (2004) examined the 5′ *tb1* region in more detail, finding evidence for a sweep influencing a region of 60–90 kb in the 5′NTR. Both Wang et al. and Clark et al. controlled for the possibility of this reduction in neutral polymorphism being caused by reduced mutation rates in this region by comparing polymorphism levels with a close relative (teosinte).

Wang et al. applied Eq. 1a to estimate the average strength of selection on this site, obtaining $s \simeq 0.05$. This value of s implies an expected time for selection to fix the domestication allele of around 300–1,000 years, indicating a fairly long period of domestication.

While *tb1* is a potential (and very likely) gene involved in very early domestication, a possible example involving a gene selected after initial domestication (during the improvement phase) is the *Waxy* gene in rice (Olsen et al. 2006). Sticky (glutinous) rice results from low amylose levels, which are typical of temperate japonica variety groups, and this has been shown to be due to a splice mutant in the *Waxy* gene. Olsen et al. observed a region of 250 kb around *Waxy* with greatly reduced levels of polymorphism compared to control populations (lines of nonsticky rice). The size of this region (using the local recombination rate) gives the estimated strength of selection acting on this site as $s \simeq 4.6$, implying selection much stronger than on *tb1* during maize domestication. Further, while the sweep around *tb1* did not even influence the coding region of its gene, the *Waxy* sweep covers 39 rice genes. As we discuss below, if a number of sweeps covering large regions occur during the course of domestication, this can have significant evolutionary effects on the entire genome.

While a reduction in the level of polymorphism at neutral markers is consistent with a selective sweep, this simple observation (by itself) is not a formal test for selection. Hence, we now consider such tests.

Tests based strictly on within-population variation

The logic behind polymorphism-based tests, in a nutshell, is *time*. If a locus has been under positive selection, it will have a younger MRCA relative to a sequence under pure drift. Conversely, if a locus is experiencing overdominant selection, two random sequences will, on average, have a more distant MRCA relative to pure drift (Fig. 3). This difference in time to MRCA has consequences for levels of standing polymorphism (the shorter the time back to the MRCA, the less the polymorphism). The time back to the MRCA also influences the length of a region under linkage disequilibrium. The longer the time, the shorter the expected block of disequilibrium around a gene. Hence, reduced level of polymorphism and/or longer blocks of disequilibrium relative to a neutral model are both *potential* signals of directional selection. Finally, selection shifts the frequency spectrum of alleles (the number of alleles in each frequency category), either producing too

many rare alleles or too many alleles at intermediate frequencies relative to pure drift, and many polymorphism-based tests look for such departures. A second feature of time that is exploited in more recent tests of selection is that, under drift, alleles at higher frequency are older and hence should have (on average) shorter blocks of linkage disequilibrium around them than younger alleles.

There are a large number of tests based on comparing different features of the current sequence variation around a locus (such as number of alleles versus average pair-wise divergence between alleles). Two sequence evolution frameworks are generally used as the basis for such comparison: the *infinite alleles* and *infinite sites* models. The key assumption of both models is that each mutation generates a new sequence, and hence leaves a unique signature. Such is *not* the case when using microsatellite (STR) markers, as these follow a step-wise mutation model, with two mutations potentially recovering the original state. When analyzing STR markers, this very different mutation process must be explicitly modeled into the analysis, unless one is willing to assume back mutation is negligible.

Returning to the infinite alleles versus infinite sites frameworks, how do these two basic sequence evolution models differ? Given a DNA sequence, an infinite alleles framework would treat each *haplotype* as a different allele (under the assumption of no intra-genic recombination), while the infinite sites framework looks at each position in the sequence separately. Figure 4 illustrates this difference. Here, in a sample of five sequences, there are three haplotypes (and hence three alleles in the infinite alleles framework). In an infinite sites framework, looking over the six sites, we find that only two of these sites are segregating.

Polymorphism-based tests compare the frequency of alleles with their expectations under the neutral model. Two typical departures are seen: (i) an excess of common alleles and a deficiency of rare alleles (alleles younger than expected) and (ii) a deficiency of common alleles and an excess of rare alleles (alleles older than expected). Pattern (i) would be expected under directional selection, when the coalescent times have been shrunk (relative to what is expected under drift) by a selective sweep. Pattern (ii) would be expected under overdominant selection, where the coalescent times are longer than expected

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

Fig. 4 Differences in interpretation under an infinite sites versus infinite alleles model. Consider five sequences (each row representing a sequence). There are three haplotypes (rows 1, 3; rows 2, 4; row 5) and hence three alleles under the infinite alleles model. However, there are only two segregating sites (columns 4 and 6, where each column represents a different position)

under drift. The problem is that these patterns can also be generated by *demographic* events. A population bottleneck and/or recent expansion can generate pattern (i), while population subdivision can generate pattern (ii). Thus, polymorphism-based tests contrast the null (strict neutral model with a single population of constant size) against a composite alternative hypothesis (selection and/or departures from a single random-mating population of constant size).

Obviously this is a serious limitation (and one well recognized in the literature). One approach to at least partly work around this concern is that demographic effects should leave a constant signature throughout the genome, while selection events leave a unique signature against this background. Hence, recent whole-genome scans of selection have performed polymorphism-based tests scanning a large, dense set of markers spanning the genome, using this information to generate a null distribution of the test statistic corrected for population history. Selection is suggested by looking at the extreme outliers against this null distribution, a rather dubious procedure we return to shortly.

Several of the early tests of neutrality are based on summary statistics from the infinite sites model (Appendix 1). The typical setting is a sample of n sequences taken from a population, with the goal of estimating $\theta = 4N_e\mu$. Three summary statistics are commonly used for this purpose under an infinite sites framework (Table 1). The first is S, the *number*

of segregating sites in the sample. The second is k, the *average pair-wise difference* between any two random sequences. The final is η, the *number of singletons*. Table 1 gives the expected values and sample variances for these summary statistics under strict neutrality. The quantities

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

which appear in the table, appear in several of the expressions for test statistics (Appendix 1). Under the neutral equilibrium model, these three measures are all simple functions of θ, and hence can be independently used to estimate it,

$$\widehat{\theta}_S = \frac{S}{a_n}, \quad \widehat{\theta}_k = k, \quad \widehat{\theta}_\eta = \frac{n-1}{n}\eta \tag{3}$$

$\widehat{\theta}_S$ is often called the *Waterson estimator* for θ (Watersons 1975). Proposed tests for neutrality contrast pairs of these estimates, with Tajima's (1989) D test comparing estimates based on S and k, while two tests (D^* and F) proposed by Fu and Li (1993) contrast estimates based on S and k with those based on η. To provide a general feel for how such tests are defined, we examine Tajima's test here in some detail and several other tests in the Appendix.

Tajima's D test

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

$$D = \frac{\widehat{\theta}_k - \widehat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} \tag{4a}$$

where the coefficients in the sample variance are

$$\alpha_D = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) - \beta_D \tag{4b}$$

and

$$\beta_D = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right) \tag{4c}$$

Tajima provides tables of critical values for D. Tajima's motivation for this test was his intuition that there is an important difference between the number of segregating sites S and the average number k of nucleotide differences. For the former, one simply counts polymorphic sites (independent of their frequencies), while the latter is a frequency-weighted measure. Hence, S is much more sensitive to changes in the frequency of rare alleles, while k is much more sensitive to changes in the frequency of intermediate alleles. A negative value of D indicates too many low frequency alleles, while a positive D indicates too many intermediate-frequency alleles. Expressed another way, Tajima's D checks whether the amount of heterozygosity is consistent with the number of polymorphisms. Under selective sweeps (and background selection and population expansion), heterozygosity should be significantly less than predicted from the number of polymorphisms, giving $E(D) < 0$.

Genome-wide polymorphism tests

As mentioned several times, polymorphism-based tests have limited scope in that if we reject the null hypothesis (neutrality), we are left with a composite alternative hypothesis that, in additional to selection, includes departures from the standard demographic assumptions (a single, randomly mating population of constant size). In light of this problem, much thought has gone into trying to

Table 1 Summary statistics of sample variation under the infinite sites model

Statistic	Expected value	Sample variance
S = number of segregating sites	$E[S] = a_n\theta$	$\sigma^2(S) = a_n\theta + b_n\theta^2$
k = average number of pair-wise differences	$E[k] = \theta$	$\sigma^2(k) = \theta\frac{n+1}{3(n-1)} + \theta^2\frac{2(n^2+n+3)}{9n(n-1)}$
η = number of singletons	$E[\eta] = \theta\frac{n}{n-1}$	$\sigma^2(\eta) = \theta\frac{n}{n-1} + \theta^2\left[\frac{2a_n}{n-1} - \frac{1}{(n-1)^2}\right]$
where	$a_n = \sum_{i=1}^{n-1}\frac{1}{i}$ and $b_n = \sum_{i=1}^{n-1}\frac{1}{i^2}$	

estimate the coalescent process under neutrality, but accounting for the population structure inherent in the data.

A number of workers have used Cavalli-Sforza's (1966) idea that all of the genome experiences the same demography (focusing here on autosomal chromosomes). Hence, markers across the genome provide useful information on the null distribution. Using this approach, one could scan a huge number of loci, under the assumption that the vast bulk are essentially neutral (i.e., not under strong directional selection), and these can be used to generate the null distribution. Outliers in this null indicate potential loci under selection. However, it is very difficult (at best!) to compute *p* values for test statistics that were chosen on the basis of being outliers without a correct model of the underlying demography. This approach is best viewed as a discovery procedure for generating an enriched set of candidate genes, not a formal test of selection.

A much better approach is to make some assumptions about the demography, and then use these to generate a neutral coalescent under this structure, from which we can obtain a null distribution for comparison. This approach is often referred to as *coalescent simulation*. In particular, given that most of the loci in a scan are likely neutral, we can use information for these to make inferences about the population parameters in the assumed demographic model. The problem is that if the assumed model is wrong (for example, the model assumes a single bottleneck, while in reality the population was formed through a series of successive bottlenecks), then the estimated demographic parameters will be biased and the critical values generated by the coalescent simulation are incorrect.

An interesting modification of the general coalescent simulation was offered Wright et al. (2005) in their scan of 774 maize genes. They assumed that loci fall into two classes, both of which experienced a bottleneck during domestication, but with loci under selection during domestication experiencing a more extreme bottleneck than non-selected loci. This results in a mixture-model likelihood for bottleneck size, which then allows Bayes theorem to be used to assign probabilities of a given locus being in the background bottleneck versus the more severe (and hence selected) bottleneck.

The ghost of Lewontin–Krakauer: genome wide F_{ST}-based scans

While many tests simply use the variation within a single population as the basis for the test statistic, one of the very first tests for selection with sequence data was proposed by Lewontin and Krakauer (1973), who looked at allele frequencies values in different populations by computing *Wright's* F_{ST} statistic. F_{ST} is basically the fraction of between-group variation. Specifically, it is the between-group variance divided by the total variance. In their test, the F_{ST} value for a candidate locus is compared with its expected neutral value. Lewontin and Krakauer reasoned that if differential (directional) selection was occurring in the different populations, this would generate a larger than expected F_{ST} value. Likewise, if overdominant selection was operating, the between-population divergence would be less than expected. While their logic was sound, their test was heavily criticized, as the null distribution under neutrality depends very strongly on details of the (unknown) population structure. As a result, their test died a quick death. However, we are now starting to see its ghost reappear in the literature (e.g., Akey et al. 2002; Kayser et al. 2003; Storz et al. 2004), wherein F_{ST} statistics are computed for a large number of loci, with outliers indicating potential loci under selection. Again, the concern is that this is simply an enrichment procedure, rather than a formal test.

One interesting partial way around this problem associated with the use of F_{ST} was suggested by Kayser et al. (2003), who looked at 322 STR loci in both Africans and Europeans. With STR markers, where a step-wise mutation model is biologically more appropriate, F_{ST} is replaced by the related measure R_{ST}. Of the 322 STR loci, Kayser et al. found that 11 showed unusually high R_{ST} values. As a check, they sequenced a nearby STR (for each of the candidates), finding that these new (and tightly linked) loci also had R_{ST} values larger than average. Vigouroux et al. (2002) also used an F_{ST} approach when screening 501 maize genes for signatures of selection. They used coalescent simulation (incorporating a founding bottleneck) to see which loci showed F_{ST} values that, given the number of segregating alleles, were significant against the coalescent simulation.

If the significance issue can be resolved, F_{ST}-based scans potentially offer a powerful tool for plant breeders, suggesting candidate loci in natural populations of a relative sampled across a series of environments in a scan for adaptation genes that can be potentially introgressed into domesticated lines.

Tests based on linkage disequilibrium

As mentioned, one feature of selective sweeps is that they have an excess of newly derived alleles at high frequency. Fay and Wu's H test (Appendix 1) is a polymorphism-based test for this specific feature. A second class of tests is offered by the following observation. Under a selective sweep, some alleles are at much higher frequencies than their age would suggest under a neutral model, and being younger these alleles have larger regions of linkage disequilibrium than expected under drift. Again, the key here is time. The more time an allele has been segregating in a population, the smaller the window of disequilibrium. If a sweep moves an allele quickly to high frequency, the amount of disequilibrium, given its frequency, should be excessive relative to a neutral model. Human geneticist have developed several tests based on this idea (Sabeti et al. 2002; Wang et al. 2006; Voight et al. 2006). These tests require very dense markers, as one must accurately measure levels of disequilibrium over very small regions, but can be very powerful. For example, Wang et al. (2006) used a massive human data set of 1.6 million SNPs, finding that 1.6% of the markers showed some signatures of positive selection. Simulation studies by Wang et al. found that their disequilibrium-based test effectively distinguishes selection from population bottlenecks and admixture (population structure). Thus, such tests (which require dense SNP markers) offer a potentially promising approach to disentangling the effects of selection from demography.

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

Joint polymorphism and divergence tests

Another important class of tests requires not only within-species sequence data, but also data on the divergence of sequence, either between very distant populations of the same species or (better) closely related species. Two widely used tests for selection, the McDonald–Kreitman and HKA tests, are of this form.

McDonald–Kreitman test

One of the most straightforward tests of selection when one has both polymorphism and divergence data was offered by McDonald and Kreitman (1991). Consider a single locus, where we contrast the polymorphism levels and divergence rates at synonymous versus replacement sites. The ratio of expected divergence between synonymous versus replacement sites is

$$\frac{d_{syn}}{d_{rep}} = \frac{2t\mu_{syn}}{2t\mu_{rep}} = \frac{\mu_{syn}}{\mu_{rep}} \tag{5a}$$

Likewise, the ratio of heterozygosity within these two classes is

$$\frac{H_{syn}}{H_{rep}} = \frac{4N_e\mu_{syn}}{4N_e\mu_{rep}} = \frac{\mu_{syn}}{\mu_{rep}} \tag{5b}$$

Hence, these two ratios have the same expected value under neutrality. We note that McDonald and Kreitman provide a more general derivation of Eq. 5a, replacing $4N_e$ (the equilibrium value) by T_{tot}, the total time on all of the within-species coalescent branches, so that any effects of demography cancel. Hence, the McDonald–Kreitman is *not affected by population demography* (Nielsen 2001). Given the constancy of these ratios under general neutrality, the McDonald–Kreitman test is performed by contrasting polymorphism versus divergence data at synonymous versus replacement sites in the gene in question through a simple contingency table analysis (χ^2 or Fisher's exact test).

At present, the McDonald–Kreitman test is best considered a test for a defined set of candidate genes. To use this test in a genomic-scan framework requires largely complete genomic sequences for the target

and a related species. For several crop species, this requirement is on the near-horizon. Issues still remain, however, as to the appropriate test statistic given the expected thousands of comparisons in a full genomic scan. Likewise, as *tb1* illustrates, many (indeed, perhaps the majority) of sites of selection are likely to be outside of coding regions, and hence are not covered by the McDonald–Kreitman test.

Hudson–Kreitman–Aguade (HKA) Test

Hudson et al. (1987) proposed the first test to jointly use information from the standing levels of polymorphisms within a species and the amount of divergence between species. The result was the very popular *HKA test*. Full details of the test are given in the Appendix, but the basic structure is as follows: one considers L loci in two species (or divergent populations) A and B. For any given locus one can estimate the amount of polymorphism in A, the amount in B, and the divergence between A and B, resulting in $3L$ estimates. Under drift, for each of the L loci, the ratio of θ values should be a constant,

$$\frac{\theta_{i,B}}{\theta_{i,A}} = \frac{4N_e(B)\mu_i}{4N_e(A)\mu_i} = \frac{N_e(B)}{N_e(A)} = \alpha$$

as the common mutation rate cancels, so that $\theta_{i,B} = \alpha\theta_{i,A}$. Thus, $L + 1$ parameters ($L\theta_i$ values and α) describe the within-species polymorphism, while the θ_i values and one additional parameter (scaled divergence time T) describe (under strict drift) the between-species divergence. Hence, one has $3L$ variation estimates which are (under drift) fully described by $L + 2$ parameters. The HKA estimates these $L + 2$ parameters and then performs a goodness-of-fit test with the $3L$ observations, with a significant departure of fit indicating failure of the strict neutral model. Unlike the McDonald–Kreitman test, the HKA test is strongly dependent on demography, and our previous comments about attempts to adjust for this apply.

Constraints on detecting selection

Even if they have experienced very strong selection, domestication genes may not leave a strong signal at linked neutral markers. What exactly are the optimal conditions for detecting selection and are these met with many crop species? At the genetic level, there are at least four factors that impact the signal of recent selection: level of polymorphism in the genome of the ancestral species, frequency of the favored allele at the start of selection, amount of local recombination around the selected site, and levels of selfing versus outcrossing.

For starters, high levels of polymorphism in the ancestral species are required. If the ancestral species has low levels of polymorphisms at the start of selection (perhaps from it passing through bottlenecks and/or being under selection itself), then the reduction in polymorphism around the selective site leaves a much weaker signal and is harder to detect. Thus, for some crops it may be very difficult to detect signatures of selection, even under strong selection. For example, Hamblin et al. (2006) found that the genome-wide background variation in Sorghum is too low to reliably detect signatures of selection given the marker density used by these investigators. Increasing the marker density may provide a partial way around this lack of sufficient variation, but this is certainly not guaranteed. As one anonymous reviewer pointed out, however, while attempts using polymorphism data may not be successful, linkage disequilibrium approaches might be able to detect signals. Clearly, this is an important open area for research.

A more subtle complication results from the frequency of favorable alleles at the start of the domestication process. A typical selective sweep is generally thought to occur following the introduction of a single favorable new mutation. Hence, there is only one founding haplotype at the time of selection. It should be kept in mind, however, that selection on domestication alleles is akin to a sudden shift in the environment, with many of these alleles pre-existing in the population before domestication. If the frequency of any such allele is >0.05, multiple haplotypes are likely present, resulting in considerable variation around the selective site even after fixation, and a very weak (if any) signal (Innan and Kim 2004; Teshima et al. 2006). Hence, there is the very real possibility than many important domestication genes will not have left a detectable signature in the pattern of linked neutral variation. Further, if these alleles were pre-existing at the time of selection (and hence of indeterminate age), there is no clear expectation as

to levels of linkage disequilibrium around selected sites.

Another factor that influences the signal of a selective sweep is the amount of effective recombination, which is a function of both the local recombination rate around a selected site as well as the amount of selfing (high levels of selfing reducing the effective recombination rate). High effective levels of recombination result in a shorter window of influence around the selected site (resulting in shorter regions of disequilibrium). While this is a plus for fine-mapping of potential genes, it also means that a dense marker scan around a putative region is required, otherwise a potential signal might be missed. In contrast, a low effective recombination rate around a selected site will result 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. The signal of a sweep (reduction in the level of neutral polymorphisms) arises because of a reduction in the effective population size around the selected site. This reduction in the effective population size has important evolutionary consequences, as the efficiency of selection on linked genes is reduced within the region influenced by the sweep. Within the region influenced by the sweep, deleterious alleles have a higher probability of fixation, while favorable alleles have a reduced probability of fixation compared to sites outside of the sweep. Thus, in species with high effective recombination rates, only a small region is potentially influenced by a sweep, but in species with low effective recombination rates (such as would occur with high levels of selfing), the sweep may influence the behavior of a number of genes beyond the selected site.

For example, the sweep around the *Waxy* gene in rice covers over 250 KB and influences close to 40 genes (Olsen et al. 2006). In a highly selfing species, sweeps can have consequences for the behavior of numerous genes beyond the target gene. 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. Lu et al. (2006) compared the genomes of *Oryza sativa* ssp. *indica* and *japonica* with their ancestral relative *O. rufipogon*. They found significantly more amino acid substitutions during domestication than expected based on the divergence of wild species. Further, many of these substitutions involve radical changes in amino acids (such as changes in electric charge), and they estimated that roughly 25% of the amino acid differences between *indica* and *japonica* were likely deleterious. Lu et al. suggest that excessive reductions in N_e due to selective-sweeps covering much of the genome during selection for domestication greatly reduced the efficiency of natural selection in removing deleterious alleles.

Caveats and unanswered questions

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

A more formal approach is to use the large amount of marker information in a genomic scan to estimate parameters for some assumed demographic model, which can then be used in coalescent simulations to generate null distributions for the test statistic. The concern with this method is that most coalescent simulations are based on simple demographic models (such as a single bottleneck in a randomly mating population). However, the true demographic situation during domestication is likely to be much more

complex, with migration between a number of subpopulations and (often) a considerable amount of selfing in addition to random outcrossing. Hence, the appropriate demographic model to underlie a coalescent simulation can be very difficult to ascertain. Further, even with a lack of demographic issues, most tests have been developed for fully outcrossing populations, and special modification may be required to correctly account for the presence of significant selfing in many crops. Such extensions have to be developed in order for plant breeders to fully exploit the power of selection scans. Demographics can also have important complications for association mapping. Population subdivision can introduced correlations between alleles that are unlinked, which can lead to markers being associated within an unlinked QTL and hence faultly marker-trait associations.

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

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

Acknowledgments Many thanks to the two careful reviewers, and Associate Editor H.-P. Piepho for their detailed comments that significantly improved the manuscript. This paper was initially presented at the 2006 Biometrics in Plant Breeding meeting in Zagreb, Croatia.

Appendix: Details of frequency spectrum-based tests

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

Table 1 provides three different estimators of θ under the infinite-sites model. Tajima's D is based on the contrast between two of these, but this leaves two other contrasts, which Fu and Li (1993) used as the basis for two new tests. Their D^* *test* compares the segregating sites (S) versus singletons (η) estimator of θ,

$$D^* = \frac{\widehat{\theta}_S - \widehat{\theta}_\eta}{\sqrt{\alpha_* S + \beta_* S^2}} \tag{A1a}$$

$$\alpha_* = \frac{1}{a_n}\left(\frac{n+1}{n} - \frac{1}{a_n}\right) - \beta_* \tag{A1b}$$

$$\beta_* = \frac{1}{a_n^2 + b_n}\left(\frac{b_n}{a_n^2} - \frac{2}{n}\left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n}\right) - \frac{1}{n^2}\right). \tag{A1c}$$

In contrast, their F^* *test* compares the average pairwise divergence (k) versus singletons (η) estimator of θ,

$$F^* = \frac{\widehat{\theta}_k - \widehat{\theta}_\eta}{\sqrt{\alpha_F S + \beta_F S^2}} \tag{A2a}$$

$$\alpha_F = \frac{1}{a_n}\left(\frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)}\right) - \beta_F \tag{A2b}$$

$$\beta_F = \frac{1}{a_n^2 + b_n}\left(\frac{2n^4 + 110n^2 - 255n + 153}{9n^2(n-1)} + \frac{2(n-1)a_n}{n^2} - \frac{8b_n}{n}\right). \tag{A2c}$$

These expression are from Simonsen et al. (1995), with Eq. A2c correcting a typo in the original Fu and Li paper. Critical values for both tests are tabulated by Fu

and Li (1993). While these tests are fairly widely used, Simonsen et al. (1995) found that they are not as powerful as Tajima's test for detecting a selective sweep or population structure departures (bottlenecks or population subdivision). However, Fu (1997) found that both tests have more power than Tajima's D for detecting signals of background selection.

Fu's W and F_S tests

Fu (1996, 1997) proposed several more refined tests for specific settings, such as too few alleles or too many alleles. These tests use the infinite alleles (as opposed to infinite sites) framework for sequence analysis (see Fig. 4). To develop these, we first need to introduce Ewen's Sampling Formula (Evens 1972), which gives the probability (under the infinite alleles model) that we see K alleles (haplotypes) in a sample of size n as

$$\Pr(K = k) = \frac{|S_n^k|\theta^k}{S_n(\theta)} \tag{A3a}$$

where

$$S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1) \tag{A3b}$$

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

$$\Pr(K = 1) = \frac{(n-1)!}{(\theta+1)(\theta+2)\cdots(\theta+n-1)} \tag{A4}$$

Using Eq. A3a, the mean and variance for the number of alleles can be found to be

$$E(K) = 1 + \theta \cdot \sum_{j=2}^{n} \frac{1}{\theta + j - 1} \tag{A5a}$$

and

$$\sigma^2(K) = \theta \cdot \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2} \tag{A5b}$$

Fu's W *test* (1996) is based on Ewen's sampling formula, and is as follows. Suppose we have an estimate $\widehat{\theta}$ of θ and we observe k alleles in our sample. The probability of seeing k (or fewer) alleles in our sample under the null hypothesis is just

$$W = \Pr(K \le k) = \sum_{i=1}^{k} \Pr(K = k|\widehat{\theta}) = \sum_{i=1}^{k} \frac{|S_n^k|\widehat{\theta}^k}{S_n(\widehat{\theta})} \tag{A6}$$

The W test uses the Watterson estimator $\widehat{\theta} = S/a_n$ so that

$$S_n(\widehat{\theta}) = S/a_n(S/a_n+1)(S/a_n+2)\cdots(S/a_n+n-1)$$

This is a test for a deficiency of rare alleles, and hence W is a one-sided test statistic. Fu (1996) showed that the W test is more powerful that Tajima's D and Fu and Li's D^* and F^* tests for detecting samples from a structured population (as also occurs with overdominant selection).

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

$$S' = \Pr(K \ge m) = \sum_{i=m}^{n} \frac{|S_n^m|\widehat{\theta}^m}{S_n(\widehat{\theta})} \tag{A7a}$$

but now using $\widehat{\theta} = k$, the estimator of θ based on average number of pair-wise differences. Fu notes that S' is not an optimal test statistic because its critical points are often too close to zero. Because of this, the test statistic S is the logit of S',

$$F_S = \ln\left(\frac{S'}{1-S'}\right) \tag{A7b}$$

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

Fay and Wu's *H* test

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

$$\widehat{\theta}_k = 2\sum_{i=1}^{n-1} \frac{S_i i(n-i)}{n(n-1)} \tag{A8a}$$

while an estimate of θ weighted by homozygosity is

$$\widehat{\theta}_H = 2\sum_{i=1}^{n-1} \frac{S_i i^2}{n(n-1)} \tag{A8b}$$

Fay and Wu's *H test* is given by the scaled difference of $\widehat{\theta}_H - \widehat{\theta}_k$.

Given that Fay and Wu's test weights derived allele at high frequency, a significant *H* and *D* test is consistent with a selective sweep, while a significant *D* test, but *not* a significant *H* test suggests background selection or demographic features more likely account for the departure from neutrality. While widely used, the *H* test is not without problems. While it is largely insensitive to population bottlenecks, it is highly sensitive to population structure. Further, the power of *H* rapidly decreases over time following a sweep, while the *D* test retains substantial power over a much longer time after a sweep (Przeworski 2002).

Hudson–Kreitman–Aguade (HKA) test

Consider two species (or distant populations) *A* and *B* that are at mutation-drift equilibrium with population sizes $N_A = N$ and $N_B = \alpha N$, respectively. Further assume they separated $T = \tau/(2N)$ generations ago from a common population of size $N^* = (N_A + N_B)/2 = N(1+\alpha)/2$, the average of the two current population sizes. Suppose $i = 1, \cdots, L$ unlinked loci are examined in both species. The amount of polymorphism for locus i in *A* is a function of $\theta_i = 4N_e\mu_i$, while the amount of polymorphism for the same locus in *B* is a function of $4N_B\mu_i = 4(\alpha N_e)\mu_i = \alpha\theta_i$. The resulting summary statistics used are LS_i^A values, for the number of segregating sites at locus i in *A*, another $LS^B{}_i$ for the same loci in *B*, and *L* D_i values, for the amounts of divergence (measured by the average number of differences between a random gamete from *A* and a random gamete from *B*). Given these $3L$ summary statistics, the HKA test statistic X^2 is given by

$$X^2 = \sum_{i=1}^{L} \frac{\left(S_i^A - \widehat{E}(S_i^A)\right)^2}{\widehat{Var}(S_i^A)} + \sum_{i=1}^{L} \frac{\left(S_i^B - \widehat{E}(S_i^B)\right)^2}{\widehat{Var}(S_i^B)} + \sum_{i=1}^{L} \frac{\left(D_i - \widehat{E}(D_i i)\right)^2}{\widehat{Var}(D_i)} \tag{A9}$$

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

$$\widehat{E}(S_i^A) = \widehat{\theta}_i a_{n_A}, \quad \widehat{E}(S_i^A) = \widehat{\alpha}\widehat{\theta}_i a_{n_B} \tag{A10a}$$

$$\widehat{Var}(S_i^A) = \widehat{\theta}_i a_{n_A} + \widehat{\theta}_i{}^2 b_{n_A}, \quad \widehat{Var}(S_i^B) = \widehat{\alpha} \quad \widehat{\theta}_i a_{n_A} + \widehat{\alpha}^2\widehat{\theta}_i{}^2 b_{n_B} \tag{A10b}$$

$$\widehat{D_i} = \widehat{\theta}_i\left(\widehat{T} + \frac{1+\widehat{\alpha}}{2}\right) \tag{A10c}$$

$$\widehat{Var}(D_i) = \widehat{\theta}_i\left(\widehat{T} + \frac{1+\widehat{\alpha}}{2}\right) + \left(\frac{\widehat{\theta}_i(1+\widehat{\alpha})}{2}\right)^2 \tag{A10d}$$

and a_n and b_n are given by Eq. 2. Equations A10a and (A10b) follow from Eq. 3, while Eq. A10c follows by re-writing

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

where the first term is the between-population divergence due to new mutations and the second term the divergence from partitioning of the polymorphism $4N^*\mu_i$ in the ancestral population. Thus, the HKA test has $L + 2$ parameters to estimate, the $L\theta_i^x$ values and two demographic parameters, T and α. The HKA test estimates these parameters and then (using Eq. A10) computes the goodness of fit X^2 statistic (Eq. A9), which is approximately χ^2 distributed with $3L - (L + 2) = 2L - 2$ degrees of freedom.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22:437–446

Brinkman MA, Frey KJ (1977) Yield component analysis of oat isolines that produce different grain yields. Crop Sci 17:165–168

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc Royal Soc London Ser B 164:362–379

Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene *tb1*. PNAS 101:700–707

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene *tb1* has pleotropic effects on plant and inflorescent architecture. Nat Genet 38:594–597

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303

Charlesworth D, Charlesworth B, Morgan MT (1995) The pattern of neutral molecular variation under the background selection model. Genetics 141:1619–1632

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

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V et al (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. Mol Ecol 11:1245–1262

Fu Y-X (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143:557–570

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

Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Hamblin MT, Casa AM, Su H, Murray SC, Paterson AH, Auadro CF, Kresovich S (2006) Challenges of detecting directional selection after a bottleneck: lessons from *Sorhum bicolor*. Genetics 173:953–964

Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. PNAS 101:10667–10672

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–1410

Jensen MA, Charlesworth B, Kreitman M (2002) Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D simulans*. Genetics 160:493–507

Kaplan NL, Hudson RR, Langley CH (1989) The ''hitchhiking effect'' revisited. Genetics 123:887–899

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

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777

Kimura M (1983) The neutral theory of molecular evolution. Cambridge Univ Press, UK

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Ann Rev Gemoics Hum Genet 1:539–559

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195

Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet 22:126–131

Lynch M, Walsh B (1997) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genet Res 23:23–35

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. Heredity 86:641–647

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD (2006) Selection under domestication:

evidence for a sweep in the rice *Waxy* genomic region. Genetics 173:975–983

Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. PNAS 101:9885–9890

Perlitz M, Stephan W (1997) The mean and variance of the number of segregating sites since the last hitchhiking event. J Math Biol 36:1–23

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189

Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. Genetics 164:1667–1676

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380 –390

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312:1614–1620

Schlštterer C (2003) Hitchhiking mapping: functional genomics from the population genetics perspective. Trends Genet 19:32–38

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429

Stephan W, Wiehr THE, Lenz M (1992) The effect of strongly selected substitutions on neutral polymorphisms: analytic results based on diffusion theory. Theor Pop Biol 41: 237–254

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688

Storz JG, Payseur BA, Nachman MW (2004) Genomic scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21:1800–1811

Tajima F (1989) Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus deomography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21:1214–1225

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genoem Res 16:702 –712

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. PNAS 99:9650–9655

Voight BF, Kudaravlli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:446–458

Wang R-L, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398:236–239

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA 103:135–140

Waterson GA (1975) On the number of segregation sites. Theor Popul Biol 7:256–276

Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES IV (2002) Genetic diversity and selection in the maize starch pathway. PNAS 99:12959–12962

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872