Molecular Population Genetics and the Search for Adaptive Evolution in Plants

Stephen I. Wright,† and Brandon S. Gaut**

*Department of Ecology and Evolutionary Biology, University of California, Irvine; and †Department of Biology, York University, Toronto, Canada

The first papers on plant molecular population genetics were published approximately 10 years ago. Since that time, well over 50 additional studies of plant nucleotide polymorphism have been published, and many of these studies focused on detecting the signature of balancing or positive selection at a locus. In this review, we discuss some of the theoretical and statistical issues surrounding the detection of selection, with focus on plant populations, and we also summarize the empirical plant molecular population genetics literature. At face value, the literature suggests that a history of balancing or positive selection in plant genes is rampant. In two well-studied taxa (maize and *Arabidopsis*) over 20% of studied genes have been interpreted as containing the signature of selection. We argue that this is probably an overstatement of the prevalence of natural selection in plant genomes, for two reasons. First, demographic effects are difficult to incorporate and have generally not been well integrated into the plant population genetics literature. Second, the genes studied to date are not a random sample, so selected genes may be overrepresented. The next generation of studies in plant molecular population genetics requires additional sampling of local populations, explicit comparisons among loci, and improved theoretical methods to control for demography. Eventually, candidate loci should be confirmed by explicit consideration of phenotypic effects.

Introduction

Plant species exhibit great morphological and functional variation, much of which is thought to be adaptive. To understand the process of adaptation for any single trait requires synthesis over several disparate levels of information. These levels range from knowledge about mutations that provide the raw material for adaptation, to measuring genetic diversity for the adaptive trait, to dissecting the developmental and phenotypic effects of genetic variants, to exploring the relationship between genetic variants and the environment. Each level requires expertise in distinct disciplines, and for each there are important lapses in our understanding.

Given the complexity of the process of adaptation, it should be no surprise that there are few plant studies that integrate across both the adaptive phenotype and its genotype. There are, however, a few notable exceptions. One example is the *Mimulus* work of Bradshaw, Schemske and colleagues (Bradshaw et al. 1998; Bradshaw and Schemske 2003; Ramsey, Bradshaw, and Schemske 2003). These researchers have dissected the genetic difference in flower color between two *Mimulus* species and have further demonstrated the effect of this gene on pollinator visitation. Another example is *Ipomea purperea*, in which the adaptive effects of flower color polymorphism have been studied at ecological, molecular, and evolutionary levels (Clegg and Durbin 2003; Durbin et al. 2003; Zufall and Rausher 2004). A third is *tb1*, a gene that governs lateral branching in maize (*Zea mays* ssp. *mays*) (Doebley, Stec, and Hubbard 1997; Wang et al. 1999; Clark et al. 2004). The phenotypic effects of variation at this gene are well understood, as is its evolutionary history as a target of artificial selection. In addition to combining information across both genotype and phenotype, these outstanding examples have at least three other things in common: (1) they represent a "top-down" approach; that is, they began with phenotypic observations (e.g., flower color and shape, plant habit) and worked down to the gene level, (2) in each system, it took years to isolate genes and integrate information over multiple levels of information, and (3) despite their importance, knowledge about the adaptive process for each trait is incomplete. Obviously, thorough study of plant adaptation is laborious and difficult, even for traits with a relatively simple genetic basis.

These successes accentuate the fact that remarkably little is known about plant adaptation at the genetic level. We have little understanding about the nature of the biochemical pathways, genes, and mutations subject to adaptive evolution or, indeed, the extent and strength of adaptation in plant populations. To address these questions generally, we need to complement "top-down" methods, which rely on a priori selection of traits of interest, with a "bottom-up" approach. In principle, molecular population genetics, which is based on population samples of DNA sequences, provides the basis for such an approach.

The era of empirical molecular population genetics based on sequence data began approximately 20 years ago (Kreitman 1983) with the promise that the approach would yield insights into the nature and importance of natural selection and also identify genes targeted by adaptive events. One reason for initial optimism toward molecular population genetics was that DNA polymorphism data are unambiguous and irreducible; they represent complete genotypic information, of which other marker-based technologies represent a subset. Another reason was that sequence data integrate information over a long time period (on average, $4N$ generations, where N is the effective population size), thus, providing powerful historical insights. Coupled with theoretical breakthroughs, such as coalescent theory (Hudson 1991) and related tests of neutral equilibrium (Hudson, Kreitman, and Aguade 1987; Tajima 1989), DNA polymorphism data promised to provide new insights into the frequency and molecular signature of adaptation.

Key words: Adaptive evolution, balancing selection, linkage disequilibrium, neutral equilibrium, positive selection.

E-mail: bgaut@uci.edu.

Mol. Biol. Evol. 22(3):506–519. 2005
doi:10.1093/molbev/msi035
Advance Access publication November 3, 2004

The first papers on plant molecular population genetics were published approximately 10 years ago. Since that time, well over 50 additional studies of plant nucleotide polymorphism have appeared, and plant sequence diversity data are increasing at a rapid rate. It is, therefore, a convenient time to summarize what has been learned in the past 10 years and to point toward the future. We begin by briefly considering some of the theoretical considerations for detecting selection from sequence data and then summarize the plant empirical literature. Throughout, we place particular emphasis on the question: What insights can molecular population genetics provide about the process of plant adaptation?

The Search for Selection: Theoretical Issues

Tests of the Neutral Equilibrium Model and the Power to Detect Selection

The neutral equilibrium model (NE) has a rich history as a null hypothesis in molecular population genetics. The NE model grew from the neutral theory of Kimura (1968), which posits that the vast majority of segregating polymorphisms are selectively neutral, and, thus, their evolutionary fate is determined by genetic drift. In addition to neutrality, the NE model makes other assumptions, including random mating and constant long-term population size. The model provides broad scope for generating readily testable null hypotheses. The search for selection has, thus, become largely an exercise in comparing expectations of the NE model to extant patterns of genetic variation.

When investigating plant adaptation with molecular population genetics, we are concerned principally with two types of natural selection: positive selection, which represents the fixation of a favorable mutation, and balancing selection, which is the long-term selective maintenance of multiple alleles. Both types of selection produce characteristic patterns of nucleotide diversity, which can be detected as departures from the expectations of NE. Several statistics have been developed to test the NE model and infer the action of selection based on different features of sequence data. The first feature is the amount of diversity: strong positive selection causes a reduction in levels of nucleotide diversity (Maynard Smith and Haigh 1974), whereas long-term balancing selection elevates diversity (Kaplan, Darden, and Hudson 1988). However, it is impossible to reject NE at a single locus based on diversity alone without explicit information about the mutation rate; thus, multilocus data are usually required to detect diversity effects. With multilocus data, for example, one can test for a significant difference among loci in the amount of diversity within species relative to divergence between species, using the HKA test (Hudson, Kreitman, and Aguade 1987). Rejection of a multilocus HKA test is often cited as evidence of selection, although demographic history can also lead to rejection (see below).

The second feature of the data is the frequency distribution of polymorphisms. Selection skews the population frequency of genetic variants relative to NE expectations. Tajima's D statistic is one of several tests that measures this frequency spectrum (Tajima 1989); under NE, the mean Tajima's D statistic is expected to be zero. After a positive selection event, the frequency spectrum is skewed; there is an excess of rare polymorphisms as new variants accumulate after a selective sweep (Braverman et al. 1995). A negative value of Tajima's D statistic indicates an excess of rare variants relative to NE expectations and is, thus, consistent with the possibility of recent positive selection. With recombination, positive selection can also be detected as an excess of high-frequency derived (nonancestral) mutations, leading to a negative value of Fay and Wu's H statistic (Fay and Wu 2000) or a skew in the pairwise mismatch distribution (Mousset, Derome, and Veuille 2004). In contrast to positive selection, balancing selection retains genetic variants, so that there can be an excess of intermediate frequency variants (Hudson and Kaplan 1988), and Tajima's D statistic can be elevated towards a positive value.

A third feature of the data is the degree of association between polymorphisms, or linkage disequibrium (LD). In the presence of recombination between selected and neutral sites, a complete selective sweep can increase LD (Kim and Nielsen 2004; Przeworski 2002) because of the sampling of only a subset of the ancestral haplotypes. Similarly, a partial selective sweep, in which the sampled allele has not reached fixation, can also cause strong elevation of LD (Hudson 1990; Sabeti et al. 2002) because of the rapid increase in frequency of a single haplotype.

Recently, methods have been developed that combine multiple features of the data—such as the level of the diversity, the frequency spectrum, and LD patterns—to characterize properties of natural selection and to make inferences about the strength and timing of selection (Jensen, Charlesworth, and Kreitman 2002; Kim and Stephan 2002, 2003; Przeworski 2003; Kim and Nielsen 2004). These approaches have been helpful in two respects. First, they have provided insight into the effects of positive selection. For example, Kim and Stephan (2002) have shown that the effect of positive selection on nucleotide diversity is often *not* centered directly on the site that has been the target of natural selection. Second, and more importantly, they provide empiricists with methods to infer parameters about selection events, such as the timing, strength, and location of selection, in some cases with confidence intervals on those parameters. For example, Przeworski's (2003) method was used to estimate a selection coefficient of approximately 3% on a putative positively selected disease-resistance gene in *Zea* (Tiffin, Hacker, and Gaut 2004). By modeling natural selection explicitly, these approaches represent a promising avenue for inferring the strength and timing of selection beyond simple rejection of the standard neutral model.

Factors Affecting the Ability to Detect Positive Selection

The power to reject NE and infer selection depends on a number of important factors. The chance to detect positive selection depends critically on the strength of selection, the time since fixation of the beneficial mutation, and the amount of recombination between the selected and neutral sites (Braverman et al. 1995; Przeworski 2002, 2003). The reason for this is straightforward: mutations arise after a positive selection event, recombination breaks down associations between variants, and the signature of

a selective sweep disappears rapidly over time. As a result, old advantageous events at a locus are difficult to detect. For some tests of positive selection, notably Fay and Wu's H test (Fay and Wu 2000) and tests based on LD, simulations have shown that, given reasonable estimates of the rate of beneficial mutation ($s = 0.005$ to 0.05) and a constant rate of selective sweeps through time, the power to detect selection at randomly chosen loci is almost zero (Przeworski 2002). These tests for positive selection have power only for candidate genes thought a priori to be under strong and recent selection (e.g., Kim and Nielsen 2004), because a priori evidence increases the chance that selection is within the narrow time window in which there is a reasonable opportunity to distinguish a sweep from neutrality (Przeworski 2002). These results have profound influence on empirical research in molecular population genetics. In general, many positive selection events in the history of a species are not likely to be identified by population genetic approaches; instead, only a subset of relatively recent events are likely to be identified.

Nonetheless, if there has been a recent increase in the rate of positive selection, it may be possible to detect a great many positive selection events. Domesticated plants are expected to be good models for detecting positive selection because they have undergone recent and drastic selection events at (presumably) numerous loci. Innan and Kim (2004) have published a theoretical treatment of positive selection during a domestication event. They show that the power to detect selection after domestication depends strongly on p, the frequency of the favorable allele in the ancestral (wild) population. If p is low—that is, the beneficial mutation was present at low frequency in the wild population—the power to detect selection during domestication can be quite high. As p increases, the power to detect selection decreases sharply because the favorable mutation has had time to recombine onto several genetic backgrounds in the ancestor. Innan and Kim (2004) use this result to suggest that only a fraction (the fraction with low p) of selected loci can be identified in domesticated plants. However, this argument depends critically on p. If most selected alleles were deleterious in the wild, as seems probable for many agronomic traits, then theses alleles had low p in the ancestor, and the identification of many, if not most, artificially selected loci is likely. On the other hand, if the selected alleles are neutral in the ancestor, they may drift to high p, making selection difficult to detect. A study in the wild ancestors of maize raises this possibility because cryptic genetic variation (i.e., variation without phenotypic effect) was documented within wild populations for a trait that may have been selected during maize domestication (Lauter and Doebley 2002).

In addition to examining patterns of diversity at individual loci to infer selection, the genomic distribution of polymorphism can also be used to make inferences about selection. If positive selection occurs at equal rates across the genome, theory predicts that diversity should be lower in regions of low recombination (Braverman et al. 1995) because the size of the affected region is strongly influenced by recombination rate. This should generate a correlation between recombination and genetic diversity (Begun and Aquadro 1992) and also a reduction in diversity in highly self-fertilizing populations, where the effective rate of recombination is reduced. However, ongoing purifying selection across the genome (background selection) also predicts a similar correlation (Charlesworth, Morgan, and Charlesworth 1993; Charlesworth, Charlesworth, and Morgan 1995), and distinguishing the effects of positive selection has been difficult.

Factors Affecting the Ability to Detect Balancing Selection

In contrast to positive selection, balanced polymorphisms can be maintained for indefinite periods of time. As a result, identification of balanced polymorphisms is less dependent on time, provided the polymorphism has reached long-term equilibrium. However, the degree to which diversity is affected and, thus, the chance of detecting the action of selection, depends critically on the rate of recombination (Hudson and Kaplan 1988). If recombination rates are high, the selected site becomes uncoupled from surrounding neutral sites, and the regions surrounding the selected site could have neutral patterns of diversity. Given the central importance of recombination on the ability to detect balancing selection, the degree to which it can be inferred may differ among species with contrasting rates of recombination.

Nordborg and Innan (2003) have shown that the power to detect a balanced polymorphism depends on the ratio of the population-mutation rate, $\theta = 4N_e\mu$ (where N_e is the effective population size and μ is the mutation rate) to the population-recombination rate, $4N_e r$ (where r is the rate of recombination). The population-mutation rate determines the amount of neutral diversity in a population, and the population-recombination rate determines the degree of LD; their ratio is a measure of the relative level of diversity to recombination. In a highly outcrossing species such as maize, this ratio is expected to be high (e.g., Tenaillon et al. 2001), and balanced polymorphisms are likely to persist over only very small distances. In contrast, highly selfing species such as *Arabidopsis thaliana* can have substantially lower effective rates of recombination, despite substantial amounts of diversity (Hagenblad and Nordborg 2002; Wright, Lauga, and Charlesworth 2003); thus, the effect of a balanced polymorphism can extend over a much larger region. In principle, it should therefore be easier to identify regions of elevated diversity subject to balancing selection in *Arabidopsis* than maize.

In accordance with this prediction, a number of loci exhibit evidence for balancing selection in *Arabidopsis* (see below), but no clear cases have been identified in maize. However, one must consider what evolutionary forces give rise to a balanced polymorphism in a selfing species such as *Arabidopsis*. Given high levels of homozygosity in selfers, one form of balancing selection, heterozygote advantage, should be ineffective. Thus, the relatively frequent identification of balancing selection in *Arabidopsis* must be driven by other processes that favor the retention of two or more alleles. Possible mechanisms include frequency-dependent selection and fluctuating

Table 1
Some Plant Taxa Studied for SNP Diversity, with a Summary of Number of Loci Studied and the Number of Loci Interpreted to Have Evolved Under Positive or Balancing Selection

Organism[a]	Breeding System/History	Loci[b]	Selected[c]	Proportion Under Selection	Comments
Arabidopsis lyrata	Obligate outcrossing	13	1	0.08	Subdivision and introgression interpreted as most likely causes of multilocus departures from neutrality
Arabidopsis thaliana	Selfing	48+ (606)[d]	18	~ 0.38	Several loci under balancing selection; diversity varies dramatically over physical distance along chromosomes
Gossypium hirsutum (cotton)	Allopolyploid	2	0	0.00	Two subgenomes contain different levels of diversity
Hordeum vulgare (wild barley)	Selfing	9	2	0.22	Perhaps as many as four loci under selection
Leavenworthia	Mixed mating system	2	1	0.50	Mating system has large effects on within-population diversity
Lycopersicon sp.	Mixed mating system	5	1 or 2	0.20	Primary focus on mating system effects, but one locus likely under balancing selection; another locus potentially hitchhiked
Mimulus sps.	Mixed mating systems	2	0	0.00	Diversity is low in selfing populations; directional introgression between species
Oryza sativa (rice)	Selfing/domesticate	15	1	0.07	Only two loci tested for selection
Pinus sylvestris	Outcrossing	3	0	0.00	Surprisingly low levels of nucleotide polymorphism
Pennisetum glaucum	Selfing/domesticate	1	0	0.00	Low levels of polymorphism relative to maize caused in part by slower mutation rate
Sorghum bicolor (sorghum)	Selfing/domesticate	95	?	?	Differences in diversity among loci not caused by mutation alone, but the number of loci affected by selection is unclear
Triticum aestivum (wheat) and relatives	Allopolyploid	2	0	0.00	Extreme *D* statistic values interpreted as evidence for multiple polyploid origins
Zea mays ssp. *mays* (maize)	Outcrossing/domesticate	50+ (18)[d]	11	~0.24	Artificial selection on genes of agronomic interest
Zea mays ssp. *parviglumis*	Outcrossing/wild	~25	1	0.04	Thus far, the primary motivation is to estimate maize predomestication (ancestral) polymorphism
Zea perennis	Autopolyploid	6	1	0.16	Investigation of the evolutionary pressures on an autopolyploid
Zea diploperennis	outcrossing	9	2	0.22	Diploid congener of maize

[a] Citations available in table 1 of Supplementary Material online.

[b] Recently, researchers have been sampling a targeted locus and linked regions around the locus (Clark et al. 2004; Palaisa et al. 2004), making it difficult to get an exact count of the number of independent loci studied.

[c] These numbers include only genes with polymorphism data that reject the neutral equilibrium model by some criterion and for which the favored interpretation by the authors is selection (as opposed to demographic effects).

[d] Parentheses denote additional loci for which tests for selection were not applied. In *Arabidopsis*, SNP diversity was described for 606 loci. In maize, an additional 18 loci were sampled from elite breeding germplasm, where selection could not be differentiated from breeding effects.

environmental conditions (Stahl et al. 1999). Furthermore, local adaptation can give an apparent signal of "balancing selection" in species-wide samples, and many, perhaps most cases of balancing selection in *Arabidopsis* may be driven by this effect (Nordborg and Innan 2003). There is, however, at least one extreme case of balancing selection that has been detected in outcrossing species: the self-incompatibility (SI) locus (Boyes et al. 1997; Charlesworth et al. 2003). Balancing selection may be relatively easy to detect in this locus in part because of a strong reduction in recombination in the SI region (Charlesworth et al. 2003; Takebayashi et al. 2003 [but see Awadalla and Charlesworth {1999}]).

The Effects of Demography

In addition to the issue of power, several other factors complicate tests for selection in plant populations. First, very few plant taxa are expected a priori to fit NE. Among other things, the NE model assumes random mating and a constant long-term population size. Departures from these assumptions can increase the variance in diversity across loci, generate a skew in the frequency spectrum at an average locus, and cause excess linkage disequilibrium, even in the absence of natural selection (see below). For example, many taxa studied thus far are selfers (table 1). Selfing taxa have a technological advantage (direct sequencing of PCR products is easier in the absence of heterozygotes) but clearly do not fit the random mating assumption of the NE model. Selfing alone does not drive deviations from NE once corrections have been made for the reduced effective recombination rate (Nordborg 2000), and this has been demonstrated empirically for patterns of linkage disequilibrium in *Arabidopsis* (Hagenblad and Nordborg 2002; Nordborg et al. 2002; Wright, Lauga, and Charlesworth 2003). Nonetheless, the weedy life history of many selfing species such as *Arabidopsis* may include frequent extinction, recolonization, and expansion; such a history may often lead to violations from the constant-size equilibrium assumptions (Wakeley and Aliacar 2001). Similarly, crops such as maize and barley have experienced population bottlenecks during domestication (e.g.,

Buckler, Thornsberry, and Kresovich 2001), thus, violating the assumption of constant long-term population size.

Second, population subdivision can also cause departures from standard neutral expectations, particularly if the sampling is effectively "admixed"—that is, a mixture of samples from distinct historical populations. Thus far, most empirical plant studies have employed "species-wide" sampling strategies to test selection. The samples typically contain five to 25 sequences from individuals throughout the species range, with multiple individuals rarely sampled from a single population. The motivation behind this sampling is to examine species-wide diversity, but this sampling strategy necessarily limits conclusions about population subdivision and population-scale events, such as local adaptation.

At present, it is not clear whether "species-wide" sampling is the best strategy to search for selection. Wakeley and Aliacar (2001) have shown that samples collected from only a single population can have dramatic departures from NE if there has been migration from other populations. Under models that assume a large number of historical populations, "species-wide" samples may be the most appropriate for testing NE in structured populations (Wakeley and Aliacar 2001). However, in a species-wide sample, it is difficult to rule out the possibility of the presence of multiple samples from individual historical populations, particularly in species such as *Arabidopsis* that may be prone to colonization and extinction events. The possibility of the presence of recent introgression from related species (Sweigart and Willis 2003; Wright, Lauga, and Charlesworth 2003; Ramos-Onsins et al. 2004) exacerbates the effects of population subdivision even further; introgressed alleles are likely relatively rare and highly divergent, thereby potentially skewing the frequency spectrum radically. As a result, processes such as population subdivision, introgression, and colonization can generate a broad range of diversity patterns, some of which mimic the effects of natural selection.

We performed simulations to illustrate the effect of demography on summary statistics such as θ and Tajima's *D* statistic. Figure 1 shows the effects that population subdivision and population size changes can have on the tails of the distribution of these summary statistics. The probability of observing values of θ at the upper and lower tails of the distribution can be dramatically inflated over the expectations of the neutral model (Nielsen 2001; Przeworski 2002). For the demographic conditions explored in figure 1, the effect is particularly noticeable in samples taken from a single population in a subdivided species and from a strong recent bottleneck. Under such conditions, the probability values obtained for tests such as the HKA test can be very misleading. In contrast, population expansion can reduce the variance in θ, leading to a reduction in power for neutrality tests based on NE. The distribution of Tajima's *D* statistic can also be severely inflated; in the case of population subdivision with restricted sampling, this can lead to both too many positive and negative values of Tajima's *D* statistic and a huge overestimation of the amount of selection if NE is assumed.

To summarize: most plant species studied to date do not fit the demographic assumptions of the NE model, and thus rejection of the neutral model with sequence polymorphism data is not unexpected. This leads to an inherent difficulty: when NE is rejected, is it because of demography or selection? Selection affects a single gene and its linked regions. In contrast, demography affects all genes within the genome to some extent. Thus, loosely speaking, the search for adaptive events becomes a search for genes that fulfill two criteria: (1) patterns of sequence diversity that deviate from the NE model and (2) patterns of sequence diversity that are extreme for genes within that species. The latter criterion makes it clear that adaptation is best inferred in the context of multilocus patterns of nucleotide variation, and we now turn to summarizing sequence diversity data across plant taxa.

Ten Years of Data

Patterns and Levels of Diversity Across Species

Since the first publications of nucleotide diversity in plants (Shattuck-Eidens et al. 1990; Gaut and Clegg 1993*a*), the vast majority of molecular population genetic studies have focused on maize, *Arabidopsis*, and their congeners (table 1). Other systems that have been studied at several loci include barley (Morrell, Lundy, and Clegg 2003; Piffanelli et al. 2004), tomato (Baudry et al. 2001), sorghum (Hamblin et al. 2004), and rice (Olsen and Purugganan 2002; Garris, McCouch, and Kresovich 2003). Several other taxa have been examined at a handful of loci, such as *Mimulus* species (Sweigart and Willis 2003), *Leavenworthia* species (Liu, Zhang, and Charlesworth 1998; Filatov and Charlesworth 1999; Liu, Charlesworth, and Kreitman 1999), cotton (Small, Ryburn, and Wendel 1999; Small and Wendel 2002), and pine (Dvornyk et al. 2002; Garcia-Gil, Mikkonen, and Savolainen 2003). The motivations for these studies are diverse: in some cases, researchers explicitly sought evidence for adaptive evolution at a particular gene; in others the purpose was to infer population history. Others have assessed the effect of mating system on patterns of nucleotide diversity (Liu, Charlesworth, and Kreitman 1999; Savolainen et al. 2000; Baudry et al. 2001; Charlesworth and Wright 2001; Wright, Lauga, and Charlesworth 2003). No matter the motivation, in most cases researchers performed tests of neutrality and interpreted their results in terms of selection and other factors.

Sequence diversity is often summarized by θ_w, Watterson's (1975) estimator of the population-mutation parameter θ. Under NE, θ_w estimates the population-mutation parameter, but otherwise it is perhaps best viewed as a simple summary of diversity. As seen in table 2, the average value of θ_w per silent site varies roughly approximately sevenfold across the plant species in our sample, with the outcrossing wild plants *Zea mays* ssp. *parviglumis* and *A. lyrata* ssp *petraea* containing the highest levels of genetic diversity measured to date. Note, however, that all of these plants have substantially higher levels of nucleotide diversity than do humans, where θ_w at silent sites is approximately 0.001 (Zwick, Cutler, and Chakravarti 2000; Frisse et al. 2001).

At least four factors contribute to variation in mean θ_w across plant species. First, as mentioned above, if

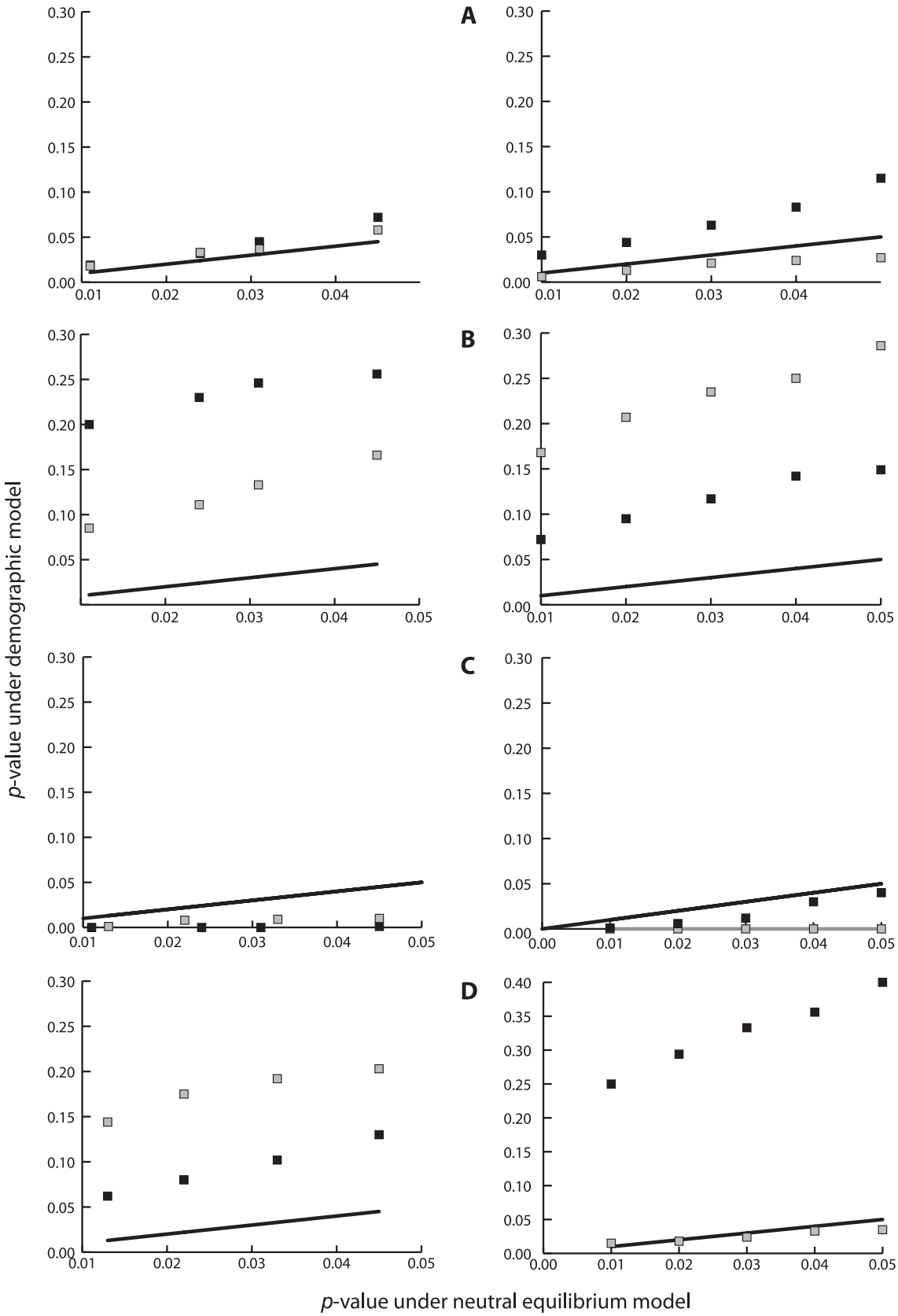

FIG. 1.—Simulation results showing the probability of observing values of θ_w (left column) and Tajima's D (right column) under different demographic scenarios. Demographic models included (*A*) Subdivision with 10 populations, sampling unequally from each population, with a migration parameter $4Nm = 1$. (*B*) Ten populations, with all 30 samples from one of the 10 populations and $4Nm = 1$. (*C*) Recent population expansion, with a 10-fold increase in effective size N generations ago. *D*) Recent population bottleneck, with a 10-fold reduction in effective population

Table 2
Summaries of Diversity and Frequency Spectrum from a Representative Set of Loci for Seven Taxa Based on "Species-Wide" Sampling Strategies

	θ_w (per Silent Site)			Tajima's *D* Statistic		
Species	Number of Genes	Average	Standard Deviation	Number of Genes	Average	Range
Arabidopsis thaliana[a]	33	0.0161	0.0122	37	−0.41	−2.03 to 1.86
Arabidopsis lyrata ssp. *lyrata*[b]	8	0.0040	0.0030	8	−0.48	−1.76 to 1.49
Arabidopsis lyrata ssp. *petraea*[b]	13	0.0247	0.0194	13	0.36	−1.08 to 2.17
maize[c]	29	0.0149	0.0118	25	−0.0094	−1.52 to 1.62
Z. mays ssp. *parviglumis*[d]	18	0.0247	0.0122	18	−0.66	−1.82 to 0.39
Sorghum bicolor[e]	52	0.0034	0.0061	73	−0.0015	−1.89 to 2.39
Hordeum vulgare[f]	9	0.0109	0.0081	9	−0.38	−1.95 to 1.79

[a] Data sources: Shepard and Purugganan 2003; Wright, Lauga, and Charlesworth 2003, and references therein.
[b] Data sources: Wright, Lauga, and Charlesworth 2003; Ramos-Onsins et al. 2004.
[c] Data sources: Shattuck-Eidens et al. 1990; Tenaillon et al. 2001; Tiffin 2004; Tiffin and Gaut 2001; Zhang et al. 2002.
[d] Data sources: Tenaillon et al. 2004; Zhang et al. 2002.
[e] Data source: Hamblin et al. 2004.
[f] Data source: Morrell, Lundy, and Clegg 2003.

species have population structure, the level of diversity within a species can be function of the degree of sampling among populations. Differences in θ_w across the species in table 2 may be exaggerated if sampling is not equivalent across species. Furthermore, polymorphism within a species can vary tremendously with sampling; average diversity is approximately 10-fold lower in the *adh1* region of maize when sampling is based on elite maize inbreds as opposed to a "species-wide" sample (Jung et al. 2004). Second, demography affects θ. This effect is obvious in the reduction of θ_w in maize relative to its congener (table 2); in part, this loss reflects a population bottleneck during domestication (Eyre-Walker et al. 1998; Hilton and Gaut 1998; Tenaillon et al. 2004). There is an even more dramatic reduction in diversity in North American *A. lyrata* ssp. *lyrata* relative to its European congener *A. lyrata* ssp. *petraea*, presumably caused by colonization followed by a long-term reduction in effective population size (Wright, Lauga, and Charlesworth 2003; Ramos-Onsins et al. 2004). Third, θ can vary among species because μ differs among species. For example, θ_w is higher at the maize *adh1* locus than at the pearl millet *adh1* locus, in part because maize has higher mutation rates (Gaut and Clegg 1993*b*). In contrast, there is evidence for a slightly elevated mutation rate in *A. thaliana* relative to *A. lyrata* (Wright, Lauga, and Charlesworth 2002). Nonetheless, nucleotide polymorphism is significantly higher in *A. lyrata* (Wright, Lauga, and Charlesworth 2003; Ramos-Onsins et al. 2004), suggesting that differences in demographic history and mating system overcome differences in mutation rate. Finally, loci with extremely high or low diversity as a consequence of selection (either long-term balancing selection or a recent selective sweep) can unduly affect average θ_w. For example, the *Zea mays* ssp. *parviglumis* data in table 2 contain one potentially swept locus (Zhang et al. 2002), and, thus, θ_w may prove to be higher in an equivalent collection of "neutral" genes. Similarly, several *A. thaliana* loci have been studied because they were expected to have high or low levels of genetic variation, and, hence, the average is biased (e.g., Shepard and Purugganan 2003).

For each species in table 2, the standard deviation in θ_w across loci is similar to average θ_w. In the sample of 33 *A. thaliana* genes in table 2, θ_w ranges 20-fold from the *fah1* gene ($\theta_w = 0.0025$) (Aguade 2001) to an antigen receptor gene ($\theta_w = 0.0489$) (Shepard and Purugganan 2003). One striking feature of *Arabidopsis* sequence diversity is that θ_w varies dramatically over small physical distances. For example, 11 linked loci (with mean length of approximately 700 bp) vary up to approximately sixfold in θ_w in the 40-kb CLAVATA region (Shepard and Purugganan 2003), and 14 loci vary up to approximately 30-fold in silent diversity in the 140-kb MAM region (Haubold et al. 2002). Such variation in θ_w is accentuated, in part, because each of these regions contains a locus for which high diversity may be maintained by balancing selection. Nonetheless, the observed variation in θ_w among genetically linked loci may be unexpected under NE, although this needs to be modeled explicitly with careful consideration of local recombination and mutation rates.

One explanation for a wide variance in θ_w is unequal mutation rates across loci. However, several multilocus studies using these and other data sets have yielded significant HKA tests, suggesting that differences in

←

size at time 0.4*N* generations ago. For each graph, values on the *y*-axis are the probabilities of observing values of θ_w and Tajima's *D* statistic equal to or more extreme than the critical values under NE; these critical values are provided on the *x*-axis. The solid line represents the probability of the critical value under the standard neutral model and has a slope of 1.0. The light squares represent critical values for the lower tail, and the dark squares represent the upper tail. For example, under scenario (*B*), the probability of observing a value of Tajima's *D* statistic greater than or equal to the 5% critical value from NE is as high as 25%, suggesting a very high proportion of "significant" results when NE is assumed. All simulations were run using Hudson's program ms (Hudson 2002). For each demographic model, 1,000 simulations were run with a sample size of 30 individuals. The parameters of each model were chosen so that the mean value of θ_w across all samples from all models was equal to 10.

mutation rate across loci are not sufficient to explain the variance in θ_w (Wright, Lauga, and Charlesworth 2003; Hamblin et al. 2004; Ramos-Onsins et al. 2004). Loci at the ends of the range of θ_w may represent genes that have experienced a recent selective sweep or loci subjected to long-term balancing selection. However, as demonstrated in figure 1, the range could also be a product of demographic effects that inflate the variance in θ_w across loci, including within a single chromosomal region such as CLAVATA. A central challenge is to fit a demographic model to multilocus data and ask: Is demography sufficient to explain the properties of the multilocus data, or must selection be invoked to explain the level and pattern of diversity across loci? Thus far, we are aware of few papers in the plant literature that address this question explicitly (Morrell, Lundy, and Clegg 2003; Tenaillon et al. 2004).

Like θ_w, the average Tajima's *D* statistic also varies among plant species (table 2); it ranges from approximately −0.70 in *Zea mays* ssp. *parviglumis* to 0.36 in *A. lyrata* ssp. *petreaea*. The most striking feature is that the average *D* statistic is negative for six of seven species and strongly negative for four of seven species. The one exception with a positive value, *A. lyrata* ssp. *petraea*, was unique in that the data included multiple samples from within multiple populations; such sampling could inflate the *D* statistic. One possible explanation for the negative *D* statistic across most taxa is sequencing error, which inflates the number of low frequency polymorphisms. It is impossible to gauge the impact of this problem, but it could be substantial for some genes in some taxa. A more likely explanation for the negative *D* statistic is demography or selection on some of the variation surveyed. When selection is weak, deleterious variants can rise to an appreciable but low population frequency (Akashi 1999), thereby contributing to negative *D* statistic values. There are also many possible demographic explanations for negative *D* values, such as an expanding population (figure 1), rare introgression, or extinction and recolonization events (Wakeley and Aliacar 2001; Ptak and Przeworski 2002).

The main focus of this review is selection and adaptation, and, thus, it is important to reiterate that the distribution of statistics such as θ_w and *D* have an impact on the search for adaptive evolution. For example, the frequency skew in *Arabidopsis* favors the identification of genes under balancing selection, as opposed to selective sweeps, for two reasons. First, the strong bias toward low frequency variants (table 2) mimics the pattern expected after a selective sweep (Tajima 1989; Fay and Wu 2000). Thus, a selective sweep may be difficult to differentiate from the demographic process that shape diversity in "neutral" loci. Second, as noted above, low effective recombination rates in selfers ensure that sites under balancing selection are in linkage disequilibrium (LD) with sites in close physical proximity (Nordborg et al. 2002). High LD leads to a pronounced peak of variation around sites under balancing selection (Kreitman and Hudson 1991). Such peaks of variation may be relatively easy to identify as regions of high θ_w (Tian et al. 2002).

Adaptive Evolution Inferred from Population Genetic Data

For approximately 10 years, plant population genetic researchers have used SNP data and test of neutrality to infer selection on a gene or a linked region. Many of these studies did not have the benefit of large, multilocus data sets for comparison. Nonetheless, several studies have concluded that genes bear the signature of either a selective sweep or a long-term balancing selection. Here, we review these conclusions in the two most studied plant taxa to date—-Arabidopsis (*Arabidopsis thaliana*) and maize (*Zea mays* ssp. *mays*)— and also discuss work in barley that highlights important issues in the search for adaptation.

Arabidopsis thaliana

A large number of papers have inferred the action of selection on *Arabidopsis* genes (table 1); more than 30% of genes sampled to date have been interpreted as having been affected by either balancing or directional selection. Extending this proportion to approximately 26,000 *Arabidopsis* genes in the genome (Arabidopsis Genome Initiative 2000), these results suggest that approximately 7,800 genes (= 26,000 × 30%) have been subject to selection (either directly or through linkage). Three considerations suggest that this number is a gross overestimate. First, theory suggests that statistical tests can detect only that subset of genes with recent positive selection events (Przeworski 2002), and, thus, 30% seems an unrealistically large proportion. Second, the genes for empirical studies were not sampled randomly; many genes were studied because they were hypothesized to be under selection. This sampling bias must be remembered whenever summary statistics —such as nucleotide diversity (table 2), *D* (table 2), or selection coefficients (Bustamante et al. 2002)—are reported. For comparison's sake, SNP surveys on randomly chosen genes and genomic regions, such as the 606 regions reported by Schmid et al. (2003), are a welcome addition.

Third, the proportion of selected *Arabidopsis* genes is overestimated because there have yet to be attempts to correct for demography in tests for selection. As noted above, the frequency distribution in *A. thaliana* is skewed toward low-frequency variants. In some cases, a highly skewed frequency spectrum has been equated with a selective sweep. However, researchers have invoked selection less often as the underlying skew, which is likely a product of demographic history, has become more apparent. Another bias was based on the observation that diversity grouped into two distinct haplotypes at some loci (Hanfstingl et al. 1994). Occasionally, this pattern was taken as evidence for balancing selection before it was noted that this distribution of haplotypes is expected under the neutral coalescent model for regions of low recombination (Aguade 2001).

At least 18 genes have been interpreted as having been subject to a selective sweep or balancing selection in *A. thaliana* (table 1). Beyond population genetic data, is there any evidence that these genes have been involved in adaptive processes? For most of these genes, the molecular function is known, and there is some plausible reason for

the gene to have been under selection. However, with only one or two exceptions, it is not known whether selection has been on the studied gene or a linked region, and there is also little supporting information about the broader process of adaptation.

To date, the most complete stories of adaptation in *Arabidopsis* come from genes that confer resistance against herbivores and pathogens. Several of these loci exhibit a history consistent with balancing selection (Tian et al. 2002; Kroymann et al. 2003; Mauricio et al. 2003; Rose et al. 2004). For at least two of the balanced loci, there are reasonable explanations of the mechanism that maintains the polymorphism. In *GS-Elong*, a locus that modifies glucosinolates that mediate herbivore resistance, there is an apparent trade-off that results in a balance between the presence and absence of the *MAM2* gene. Presence of *MAM2* confers resistance to a generalist herbivore, without an obvious physiological (allocation) cost. However, presence may also have an ecological cost related to interactions with specialist herbivores (Kroymann et al. 2003). Thus, the selective pressure on the presence and absence of the gene likely depends on the community of herbivores.

An even more compelling story is that of *RPM1*, which has an absence/presence polymorphism maintained by balancing selection. For this locus, there is a fitness cost to containing the gene in the absence of pathogen but a fitness advantage in the presence of pathogen (Tian et al. 2002, 2003). The fitness cost associated with maintaining the locus has been clearly demonstrated (Tian et al. 2003), but the question remains as to the evolutionary forces that maintain the two alleles. Because heterozygote advantage is unlikely, it seems reasonable to postulate that the polymorphism is maintained by fluctuating selection pressure in the presence and absence of the pathogen. In our opinion, the *RPM1* work is the best example to date of identifying the pattern of an adaptive event using molecular population genetic approaches (Tian et al. 2002) and then extending that observation to phenotype and fitness.

Zea mays ssp. mays

Maize is another taxon for which there have been numerous putative examples of adaptation based on molecular population-genetic data. One of the advantages of maize as a study system is that many genes are expected to have been the target of artificial selection. Another advantage is that selective sweeps can be detected both by standard tests of the NE model and by contrasting nucleotide diversity between maize and its wild ancestor (Hanson et al. 1996; Vigouroux et al. 2002; Whitt et al. 2002; Tenaillon et al. 2004). Selected genes are expected to demonstrate a much greater reduction in diversity between the wild taxon and its domesticate relative to nonselected genes.

Thus far, approximately 50 genes have been surveyed for nucleotide polymorphism in maize. There is evidence of a selective sweep in 12 of the genes, a proportion of approximately 24%. The "selected" genes exhibit a marked decrease in genetic diversity in maize. For a large subset of these genes, maize retains only approximately 15% of genetic diversity found in the wild ancestor; in contrast "nonselected" genes retain approximately 80% (Whitt et al. 2002; Zhang et al. 2002; Tenaillon et al. 2004). However, the same caveats apply to maize as to *Arabidopsis*—that is, the proportion of selected genes is probably biased upward because of sampling biases. In contrast, a genome scan of microsatellite variation in maize yielded a smaller estimate of the proportion of loci under selection in maize (3%) (Vigouroux et al. 2002). It is also important to note that there have been explicit attempts to model demographic effects on maize genetic diversity (Eyre-Walker et al. 1998; Hilton and Gaut 1998; Vigouroux et al. 2002; Tenaillon et al. 2004) and demography has been incorporated in tests for selection (Tenaillon et al. 2004). It remains to be seen what proportion of genes have been affected by artificial selection during domestication and germplasm improvement.

What is known about the putatively selected maize genes? In general, adaptation is better substantiated in maize than in most *Arabidopsis* candidates. For example, Doebley and colleagues' (Doebley, Stec, and Hubbard 1997; Wang et al. 1999; Clark et al. 2004) work on *tb1* is one of the few examples where information about adaptation extends from gene to phenotype. There is also reasonable phenotypic data on at least three other genes: *c1*, *d8*, and *y1*. The *c1* gene regulates anthocyanin biosynthesis; alleles that mediate the purple kernel phenotype are found at different frequencies in maize and teosinte (Hanson et al. 1996). At the *d8* locus, segregating genetic variation is associated with flowering time (Thornsberry et al. 2001), suggesting that flowering time is the selected phenotype. Similarly, nucleotide polymorphism at the *y1* locus associates with endosperm coloration (Palaisa et al. 2003), and there is convincing evidence that this locus was under artificial selection (Palaisa et al. 2004). Another interesting study of adaptation focused on the starch biosynthesis pathway genes. Three of six genes in this pathway appear to have been selected during domestication or improvement (Whitt et al. 2002). Although the function of these genes is well known, the phenotypic effects of selection at these loci have not yet been demonstrated precisely.

A great many interesting questions remain about maize that are important both to maize breeders and to students of domestication. What genes were selected during the history of domestication and improvement? What proportion of the genome do they represent? What are their phenotypic effects? Can existing genetic variants from either wild taxon or the domesticate further benefit agriculture? Continued comparisons of sequence diversity between maize and its wild ancestor may be able to answer some of these questions.

Wild Barley

A recent study in wild barley (*Hordeum vulgare*) used geographic information to infer selection. Morrell, Lundy, and Clegg (2003) surveyed nucleotide diversity at nine loci, based on individuals sampled across a geographic range from Israel to Afghanistan. Previous studies have shown a deep divergence in *adh3* sequences between individuals from the eastern and western edges of the

range. In contrast, genetic diversity at two other loci (*adh1* and *adh2*) demonstrated no obvious geographic pattern (Lin, Brown, and Clegg 2001; Lin, Morrell, and Clegg 2002). The study of Morrell, Lundy, and Clegg (2003) sought to characterize the evolutionary forces that produced these disjunct spatial patterns.

To do this, they assigned individuals to one of three regions: the west, the east, or the Zagros Mountains. Using a computer-intensive maximum-likelihood approach (Kuhner et al. 2002), migration rates were estimated among regions for each locus. The results were surprising. Over all loci, the estimated migration rates were high enough to expect homogenization of genetic variation across regions. Nonetheless, at least two loci (*adh3* and *G3pdh*) demonstrated marked geographic heterogeneities in the distribution of genetic diversity. Two other loci (*adh1* and *Pepc*) harbored very little diversity, with no apparent geographic pattern. Taken at face value, these data are difficult to reconcile without invoking selection, but the nature and extent of selection are unclear. It is likely that *adh3* and *G3pdh* exhibit spatial heterogeneity because of local adaptation, but it is also possible that the low diversity loci (*adh1* and *pepc*) experienced selective sweeps across the entire geographic range. Overall, selection appears to have shaped diversity in at least two (2/9=22%), and perhaps four (4/9=44%), loci over the time period and geographic range spanned by the sequence data.

The barley study is notable for two reasons. The first reason is the use of the spatial distribution of genetic variants to argue for selection. Spatial information has been used relatively rarely in plant molecular population genetics thus far (but see Delye et al. [2004] and Sweigart and Willis [2003]). The second reason is that selection was inferred largely from pattern matching (i.e., migration and geographic distribution), as opposed to explicit tests of NE. Clearly, additional work needs to be done. Formal tests need to be developed to rule out the possibility that differences in levels differentiation across loci are the result of a single demographic model. For the loci with spatial structure, additional sampling from local populations will be required to drive home the conclusion that these loci have been under selection in local populations. Finally, there is as yet no phenotype associated with any of the putative adaptive events, and, thus, there is much to learn about the process and effect of adaptation.

Conclusions: The Future of Plant Molecular Population Genetics

Thus far, reports of selection on plant genes are numerous. Summarizing across data in highly studied taxa, authors have concluded that as many as 20% to 30% of loci studied to date have been under selection (table 1). These values are certainly biased upward, both by the choice of genes and by other factors, such as poor consideration of sampling and demographic effects.

Much remains to be learned about selection. For any candidate locus identified by molecular population-genetic approaches, some key questions remained to be answered: (1) Have demographic effects been considered properly or is demography misleading the inference of selection? (2) After greater scrutiny, such as increased sampling, do the genes remain candidates? (3) If so, how does the gene contribute to an adaptive phenotype?

To date, demographic processes have been poorly addressed in the plant literature, in part because careful consideration of demography requires large, multilocus data sets. However, in at least two taxa (*Arabidopsis* and maize), there will soon be reports of nucleotide diversity in over a thousand loci. In another taxon (rice), plans are in place to sequence several hundred loci. These data sets will be of tremendous value because they will provide a "genomic distribution" of population-genetic statistics such as θ_W and Tajima's *D*. These distributions will reflect processes that have affected all genes, and they will thus provide a basis to better understand species and demographic history. Yet, it is important to recognize that these distributions will not be a panacea for inferring selection. By definition, 5% of the observations will fall in the 5% tails of the distribution. One hopes that most selected genes fall into these tails, but many (if not most) of the genes in the tails likely represent deviations around the mean caused by nonselective factors. Thus, the question: Which of the genes in the tails represent interesting adaptive events? This question cannot be answered without careful investigation of the demographic process and without distinct models that incorporate selection. Therefore, new statistical approaches and models are requirements for future progress. The inference of demographic history from multilocus data is itself still in its infancy.

As noted above, most empirical plant population-genetic inferences have been based on "species-wide" samples of nucleotide sequences; very few plant studies have focused on comparing sequence diversity among local populations. The next generation of empirical studies of plant molecular population genetics needs to include explicit sampling of distinct populations, for two reasons. First, population sampling will provide insights into migration, colonization, and other demographic factors. Such insights will prove important not only for understanding the evolutionary process but also for the design of association studies. More importantly, population sampling may provide additional insights into adaptive events. Because the time frame for detecting positive selection is short, local adaptation may occur on an appropriate timescale for detection. For many candidate-selected genes, additional sampling in local populations may provide valuable insights into the nature and strength of selection.

An additional fruitful avenue of research is demonstrated by the study of genes in the maize starch biosynthesis pathway (Whitt et al. 2002). By comparing diversity for a group of genes that contribute to a particular pathway, one can provide detailed and relatively complete information about the genes that have contributed to an adaptive shift for a particular trait. Further, comparing diversity patterns between genes of different functional categories provides an important avenue for controlling genome-wide effects of demographic history.

If positive selection occurs at a high rate, as suggested by empirical studies to date, there should be a significant

correlation between rates of recombination and diversity throughout the genome (Begun and Aquadro 1992). Thus far, there is little evidence for an effect of recombination on sequence diversity in plants (Baudry et al. 2001; Tenaillon et al. 2001), although there may be an effect for SSR and RFLP diversity (Dvorak, Luo, and Yang 1998; Tenaillon et al. 2002). The lack of correlation with sequence diversity is inconsistent with a model of rampant positive selection, but there are possible confounding factors, such as a reduction of gene density in regions of low recombination or a lack of uniform rates of positive selection across the genome. Overall, the role of positive selection in shaping diversity across the genome is unclear and difficult to distinguish from background selection, particularly in genomes with negative average values of Tajima's *D* statistic. Nonetheless, genome-wide patterns of diversity merit increased attention.

With few exceptions, selected loci identified by molecular population genetics have little corroborating evidence from phenotypes or fitness assays. To further the study of adaptation, there is a continuing need to move from the "bottom-up" — that is, from population genetics to phenotypic effects—just as there have been efforts to move "top-down." At present, there are two ways to establish the link between genotype and phenotype that may prove useful for the study of adaptation. The first is association studies. Association mapping can be fraught with statistical and biological difficulties, but it already boasts some success (Palaisa et al. 2003; Wilson et al. 2004). The success and failure of the approach will likely differ dramatically among traits and taxa. The second method is the comparison of alleles via complementation or transgenic analysis. Transgenics have been particularly useful in fitness assays in field trials (Tian et al. 2003; Kessler and Baldwin 2004). Zufall and Rausher's (2004) transformation of *Ipomea* alleles into *Arabidopsis* null–mutants represents another approach that may be useful for linking genotype to phenotype. By using population-genetics approaches to develop a list of candidate-selected genes and then returning to phenotypic analyses, one can reduce the limitations of each method and also minimize the number of false positives.

Acknowledgments

The authors thank P. Tiffin, P. Morrell, M. Clegg, M. Nordborg. and M. Przeworski for comments. This work supported by NSF grants DEB-0316157 and DBI-0321467.

Literature Cited

Aguade, M. 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. Mol. Biol. Evol. **18**:1–9.

Akashi, H. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151**:221–238.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Awadalla, P., and D. Charlesworth. 1999. Recombination and selection at *Brassica* self-incompatibility loci. Genetics **152**: 413–425.

Baudry, E., C. Kerdelhue, H. Innan, and W. Stephan. 2001. Species and recombination effects on DNA variability in the tomato genus. Genetics **158**:1725–1735.

Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature **356**:519–520.

Boyes, D. C., M. E. Nasrallah, J. Vrebalov, and J. B. Nasrallah. 1997. The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9**:237–247.

Bradshaw, H. D. Jr., K. G. Otto, B. E. Frewen, J. K. McKay, and D. W. Schemske. 1998. Quantitative trait loci affecting differences in floral morphology between two species of monkeyflower (*Mimulus*). Genetics **149**:367–382.

Bradshaw, H. D., and D. W. Schemske. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. Nature **426**:176–178.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**: 783–796.

Buckler, E. S., J. M. Thornsberry, and S. Kresovich. 2001. Molecular diversity, structure and domestication of grasses. Genet. Res. **77**:213–218.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in *Arabidopsis*. Nature **416**:531–534.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

Charlesworth, D., C. Bartolome, M. H. Schierup, and B. K. Mable. 2003. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Mol. Biol. Evol. **20**:1741–1753.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. Genetics **141**:1619–1632.

Charlesworth, D., and S. I. Wright. 2001. Breeding systems and genome evolution. Curr. Opin. Genet. Dev. **11**:685–690.

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.

Clegg, M. T., and M. L. Durbin. 2003. Tracing floral adaptations from ecology to molecules. Nat. Rev. Genet. **4**:206–215.

Delye, C., C. Straub, S. Michel, and V. Le Corre. 2004. Nucleotide variability at the acetyl coenzyme A carboxylase gene and the signature of herbicide selection in the grass weed *Alopecurus myosuroides* (Huds.). Mol. Biol. Evol. **21**:884–892.

Doebley, J., A. Stec, and L. Hubbard. 1997. The evolution of apical dominance in maize. Nature **386**:485–488.

Durbin, M. L., K. E. Lundy, P. L. Morrell, C. L. Torres-Martinez, and M. T. Clegg. 2003. Genes that determine flower color: the role of regulatory changes in the evolution of phenotypic adaptations. Mol. Phylogenet. Evol. **29**:507–518.

Dvorak, J., M.-C. Luo, and Z.-L. Yang. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics **148**:423–434.

Dvornyk, V., A. Sirvio, M. Mikkonen, and O. Savolainen. 2002. Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Mol. Biol. Evol. **19**:179–188.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman, and B. S. Gaut. 1998. Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95**:4441–4446.

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

Filatov, D. A., and D. Charlesworth. 1999. DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia *PgiC* locus. Genetics **153**:1423–1434.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. DiRienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69**:831–843.

Garcia-Gil, M. R., M. Mikkonen, and O. Savolainen. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol. Ecol. **12**:1195–1206.

Garris, A. J., S. R. McCouch, and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics **165**:759–769.

Gaut, B. S., and M. T. Clegg. 1993*a*. Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA **90**:5095–5099.

———. 1993*b*. Nucleotide polymorphsim in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics **135**: 1091–1097.

Hagenblad, J., and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. Genetics **61**:289–298.

Hamblin, M. T., S. E. Mitchell, G. M. White, J. Gallego, R. Kukatla, R.A. Wing, A. H. Paterson, and S. Kresovich. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics **167**:471–483.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa, W. Rudiger, and F. M. Ausubel. 1994. Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138**:811–828.

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

Haubold, B., J. Kroymann, A. Ratzka, T. Mitchell-Olds, and T. Wiehe. 2002. Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. Genetics **161**:1269–1278.

Hilton, H., and B. S. Gaut. 1998. Speciation and domestication in maize and its wild relatives: evidence from the *globulin-1* gene. Genetics **150**:863–872.

Hudson, R. R. 1990 Gene genealogies and the coalescent process. Pp. 1–44 *in* J. A. D. Futuyma, ed. Oxford surveys in evolutionary biology. Oxford University Press, New York.

———. 1991. Gene genealogies and the coalescent process. Oxford Surv. in Evol. Biol. **7**:1–44.

———. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337–338.

Hudson, R. R., and N. L. Kaplan. 1988. The coalescent process in models with selection and recombination. Genetics **120**: 831–840.

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Innan, H., and Y. Kim. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA **101**:10667–10672.

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

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey, M. Morgante, and A. Rafalski. 2004. Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. Theor. Appl. Genet.

Kaplan, N. L., T. Darden, and R. R. Hudson. 1988. The coalescent process in models with selection. Genetics **120**:819–829.

Kessler, A., and I. T. Baldwin. 2004. Herbivore-induced plant vaccination. Part I. The orchestration of plant defenses in nature and their fitness consequences in the wild tobacco *Nicotiana attenuata*. Plant J. **38**:639–649.

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

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

———. 2003. Selective sweeps in the presence of interference among partially linked loci. Genetics **164**:389–398.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624–626.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304**:412–417.

Kreitman, M., and R. R. Hudson. 1991. Inferring the evolutionary histories of *Adh* and the *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127**:565–582.

Kroymann, J., S. Donnerhacke, D. Schnabelrauch, and T. Mitchell-Olds. 2003. Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus. Proc. Natl. Acad. Sci. USA **100**(Suppl 2):14587–14592.

Kuhner, M. K., P. Beerli, J. Yamato, and J. Felsenstein. 2002. LAMARC: likelihood analysis with metropolis algorithm using random coalescence. Version 1.1. University of Washington, Seattle.

Lauter, N., and J. Doebley. 2002. Genetic variation for phenotypically invariant traits detected in teosinte: implications for the evolution of novel forms. Genetics **160**:333–342.

Lin, J. Z., A. H. Brown, and M. T. Clegg. 2001. Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). Proc. Natl. Acad. Sci. USA **98**:531–536.

Lin, J. Z., P. L. Morrell, and M. T. Clegg. 2002. The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp. *spontaneum*). Genetics **162**: 2007–2015.

Liu, F., D. Charlesworth, and M. Kreitman. 1999. The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. Genetics **151**:343–357.

Liu, F., L. Zhang, and D. Charlesworth. 1998. Genetic diversity in *Leavenworthia* populations with different inbreeding levels. Proc. R. Soc. Lond. B Biol. Sci. **265**:293–301.

Mauricio, R., E. A. Stahl, T. Korves, D. Tian, M. Kreitman, and J. Bergelson. 2003. Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. Genetics **163**:735–746.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet Res **23**:23–25.

Morrell, P. L., K. E. Lundy, and M. T. Clegg. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. Proc. Natl. Acad. Sci. USA **100**:10812–10817.

Mousset, S., N. Derome, and M. Veuille. 2004. A test of neutrality and constant population size based on the mismatch distribution. Mol. Biol. Evol. **21**:724–731.

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

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

Nordborg, M., J. O. Borevitz, J. Bergelson et al. (12 co-authors). 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **30**:190–193.

Nordborg, M., and H. Innan. 2003. The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics **163**:1201–1213.

Olsen, K. M., and M. D. Purugganan. 2002. Molecular evidence on the origin and evolution of glutinous rice. Genetics **162**:941–950.

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.

Palaisa, K. A., M. Morgante, M. Williams, and A. Rafalski. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell **15**:1795–1806.

Piffanelli P., L. Ramsay, R. Waugh, A. Benabdelmouna, A. D'Hont, K. Hollricher, J. H. Jorgensen, P. Schulze-Lefert, and R. Panstruga. 2004. A barley cultivation- associated polymorphism conveys resistance to powdery mildew. Nature **430**: 887–891.

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

———. 2003. Estimating the time since the fixation of a beneficial allele. Genetics **164**:1667–1676.

Ptak, S. E., and M. Przeworski. 2002. Evidence for population growth is confounded by fine-scale population structure. Trends Genet. **18**:559–563.

Ramos-Onsins, S. E., B. E. Stranger, T. Mitchell-Olds, and M. Aguade. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics **166**:373–388.

Ramsey, J., H. D. Bradshaw Jr., and D. W. Schemske. 2003. Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (Phrymaceae). Evol. Int. J. Org. Evol. **57**:1520–1534.

Rose, L. E., P. D. Bittner-Eddy, C. H. Langley, E. B. Holub, R. W. Michelmore, and J. L. Beynon. 2004. The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in *Arabidopsis thaliana*. Genetics **166**:1517–1527.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. (14 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature **419**:832–837.

Savolainen, O., C. H. Langley, B. P. Lazzaro, and H. Fr. 2000. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol. Biol. Evol. **17**:645–655.

Schmid, K. J., T. R. Sorensen, R. Stracke, O. Torjek, T. Altmann, T. Mitchell-Olds, and B. Weisshaar. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res **13**:1250–1257.

Shattuck-Eidens, D. M., R. N. Bell, S. L. Neuhausen, and T. Helentjaris. 1990. DNA sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. Genetics **126**:207–217.

Shepard, K. A., and M. D. Purugganan. 2003. Molecular population genetics of the *Arabidopsis CLAVATA2* region: the genomic scale of variation and selection in a selfing species. Genetics **263**:1083–1095.

Small, R. L., J. A. Ryburn, and J. F. Wendel. 1999. Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Mol. Biol. Evol. **16**:491–501.

Small, R. L., and J. F. Wendel. 2002. Differential evolutionary dynamics of duplicated paralogous Adh loci in allotetraploid cotton (*Gossypium*). Mol. Biol. Evol. **19**:597–607.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. Nature **400**:667–671.

Sweigart, A. L., and J. H. Willis. 2003. Patterns of nucleotide diversity in two species of *Mimulus* are affected by mating system and asymmetric introgression. Evol. Int. J. Org. Evol. **57**:2490–2506.

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

Takebayashi, N., P. B. Brewer, E. Newbigin, and M. K. Uyenoyama. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. **20**:1778–1794.

Tenaillon, M. I., M. C. Sawkins, L. K. Anderson, S. M. Stack, J. Doebley, and B. S. Gaut. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics **162**:1401–1413.

Tenaillon, M., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA **98**:9161–9166.

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.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. Nat. Genet. **28**:286–289.

Tian, D., H. Araki, E. Stahl, J. Bergelson, and M. Kreitman. 2002. Signature of balancing selection in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **99**:11525–11530.

Tian, D., M. B. Traw, J. Q. Chen, M. Kreitman, and J. Bergelson. 2003. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. Nature **423**:74–77.

Tiffin, P. 2004. Comparative evolutionary histories of chitinase genes in the genus *Zea* and family Poaceae. Genetics **167**: 1331–1340.

Tiffin, P., and B. S. Gaut. 2001. Molecular evolution of the wound-induced serine protease inhibitor *wip1* in *Zea* and related genera. Mol. Biol. Evol. **18**:2092–2101.

Tiffin, P., R. Hacker, and B. S. Gaut. 2004. Population genetic evidence for rapid changes in intraspecific diversity and allelic cycling of a specialist defense gene in *Zea*. Genetics **168**: 425–434.

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.

Wakeley, J., and N. Aliacar. 2001. Gene genealogies in a metapopulation. Genetics **159**:893–905.

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

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:188–193.

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

Wilson, L. M., S. R. Whitt, A. M. Ibanez, T. R. Rocheford, M. M. Goodman, and E.S. Buckler. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell **16**:2719–2733.

Wright, S. I., B. Lauga, and D. Charlesworth. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol. Biol. Evol. **19**:1407–1420.

———. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol. Ecol. **12**:1247–1263.

Zhang, L., A. S. Peek, D. Dunams, and B. S. Gaut. 2002. Population genetics of duplicated disease-defense genes, *hm1* and *hm2*, in maize (*Zea mays* ssp. *mays* L.) and its wild ancestor (*Zea mays* ssp. *parviglumis*). Genetics **162**:851–860.

Zufall, R. A., and M. D. Rausher. 2004. Genetic changes associated with floral adaptation restrict future evolutionary potential. Nature **428**:847–850.

Zwick, M. E., D. J. Cutler, and A. Chakravarti. 2000. Patterns of genetic variation in mendelian and complex traits. Annu. Rev. Genomics Hum. Genet. **1**:387–407.

William Martin, Associate Editor

Accepted October 26, 2004