New Phytologist

Review

*Tansley review*

# The population genomics of plant adaptation

Author for correspondence:
*Mathieu Siol*
*Tel: +1 416 9785603*
*Email: mathieu.siol@utoronto.ca*

Received: *30 April 2010*
Accepted: *18 June 2010*

Mathieu Siol, Stephen I. Wright and Spencer C. H. Barrett

Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada

## Contents



*New Phytologist* (2010) **188**: 313–332
**doi**: 10.1111/j.1469-8137.2010.03401.x

**Key words:** demographic history, DNA sequence polymorphism, effective population size, local adaptation, natural selection, plant populations, population genomics, standard neutral model.

## Summary

There has been an enormous increase in the amount of data on DNA sequence polymorphism available for many organisms in the last decade. New sequencing technologies provide great potential for investigating natural selection in plants using population genomic approaches. However, plant populations frequently show significant departures from the assumptions of standard models used to detect selection and many forms of directional selection do not fit with classical population genetics theory. Here, we explore the extent to which plant populations show departures from standard model assumptions, and the implications this has for detecting selection on molecular variation. A growing number of multilocus studies of nucleotide variation suggest that changes in population size, particularly bottlenecks, and strong subdivision may be common in plants. This demographic variation presents important challenges for models used to infer selection. In addition, selection from standing genetic variation and multiple independent adaptive substitutions can further complicate efforts to understand the nature of selection. We discuss emerging patterns from plant studies and propose that, rather than treating population history as a nuisance variable when testing for selection, the interaction between demography and selection is of fundamental importance for evolutionary studies of plant populations using molecular data.

New Phytologist

## I. Introduction

Molecular population genetics is being invigorated by the ever-growing amount of nucleotide sequence data available. As a result, during the last two decades considerable efforts have been devoted to designing and applying analytical methods for detecting the footprint of natural selection at the molecular level. Finding genomic regions under selection is one of the first steps required to bridge the gap between the genotype and phenotype of adaptive traits, and is thus crucial for understanding the process of adaptation. Multilocus DNA sequence data also provide opportunities to gain detailed insight into population history and structure using explicit models that incorporate demographic features of populations. This presents an important challenge because both selection and population history have important influences on the amount and patterns of genetic variation. Studies of selection should ideally incorporate the confounding effects of demographic history, but studies of population history typically assume the absence of selection. Our review highlights this problem and discusses the progress and prospects for jointly inferring the role of population history and selection in plant populations.

Methods developed in the last 20 yr to test for selection on molecular variation mostly stem from the neutral theory of molecular evolution (Kimura, 1968, 1983). In a nutshell, the neutral theory posits that: the fate of segregating polymorphism is effectively determined by genetic drift, as most variation is neutral with regard to natural selection; fixed differences in alleles between species (divergence) are mostly neutral, with a negligible contribution from adaptive substitutions, and neutral loci are not affected by the effects of linked selection. Although this theory has stimulated much debate since its inception (Gillespie, 2000, 2001), it soon became widely used as a null hypothesis in molecular population genetics against which to test for selection. However, several crucial assumptions of the standard neutral model (hereafter SNM), namely no population structure, a constant population size and random mating make it a composite hypothesis (Nielsen, 2001; Garrigan *et al.*, 2010). Thus, the mere rejection of neutrality does not point unambiguously to an effect of selection, but could also result from the violation of one (or several) of the aforementioned assumptions.

In parallel with attempts to test for the action of natural selection, considerable progress has been made in fitting explicit coalescent models to DNA sequence data for inferring demographic history (Hudson, 2002; Gutenkunst *et al.*, 2009; Kuhner, 2009). These approaches allow for important inferences about the amount and timing of changes in population size, the extent of gene flow among populations and species (Hey & Nielsen, 2004; Kuhner, 2006; Hey, 2010), and the geographic structuring of populations (Charlesworth *et al.*, 2003). These approaches have the potential to provide important quantitative insights into the process of speciation, the connectedness of populations, and the role of environmental factors, such as past climates, in influencing historical population dynamics.

Increasingly, methods to test for selection are being developed that explicitly take demography into account (Kim & Stephan, 2002; Jensen *et al.*, 2005; Nielsen *et al.*, 2005; De Mita *et al.*, 2007; Eyre-Walker & Keightley, 2009). However, frequent and severe bottlenecks or extensive population subdivision are likely to strongly influence the power and ability to detect selection, and our understanding of these influences on testing for selection is still rather limited. For example, using simulations Städler *et al.* (2009) demonstrated how population subdivision can modify patterns of polymorphism and therefore affect the efficacy of tests of selection. Furthermore, departures from the classical model of directional selection also influence our ability to detect selection when it occurs. Here, we specifically explore the extent to which plant populations may be especially susceptible to violations of the assumptions of the SNM, and investigate the consequences that this may have for inferences of natural selection on molecular variation.

First, we outline the different methods that have been devised to detect the traces left by natural selection at the molecular level. We devote some effort to comparing these methods because our ability to detect selection at the molecular level depends critically on the types of data used and how robust the methods are to the underlying assumptions. We then explore violations of standard assumptions of the SNM and review recent evidence from multilocus data indicating that plant populations are indeed often susceptible to these violations. We also consider progress that has been made in developing methods to account for these violations. Finally, we conclude with a discussion of prospects and future directions in the field of plant population genomics, taking into account the increasing amount of data soon to be generated for a growing number of diverse species by next-generation sequencing techniques (Shendure & Ji, 2008; Wang *et al.*, 2009).

## II. Methods to detect selection at the molecular level

Natural selection has several types of effects on patterns of nucleotide variation, including on the level and structure of polymorphism, the amount of linkage disequilibrium (LD) around selected regions, the degree of population differentiation and the proportion and frequency of nonsynonymous substitutions (Table 1). The approaches used to examine DNA sequence variation can be distinguished by those that aim to detect the footprint of selection on linked neutral

**Table 1** An incomplete list of approaches for detecting selection on DNA sequences (see the text for further discussion)

| Test category | Signature detected | Limitations |
|---|---|---|
| Level of diversity | Unusually low or high genetic diversity around the selected locus | High sensitivity to demographic assumptions |
| Site frequency spectrum (SFS) based-test | Modification in the relative proportions of low and high frequency mutations in the selected region | High sensitivity to demographic assumptions. High rate of false positives |
| Linkage disequilibrium (LD) | A rise in frequency of long haplotypes created by the increased LD around the selected region | Spurious signal of selection created by population structure. LD levels decrease rapidly after selective sweep is complete |
| Synonymous/nonsynonymous mutations | Differences between the ratio of nonsynonymous to synonymous polymorphism and nonsynonymous to synonymous divergence | Cannot distinguish between past and current selection. Slightly deleterious mutations inflate polymorphism. Spurious signal of selection with population expansion and bottlenecks if there are slightly deleterious mutations |
| Population differentiation | Increased or decreased population differentiation of a genomic region relative to the rest of the genome | Hierarchical genetic substructure creates false positives. Importance of the sampling scheme |

sites, and those that infer the action of selection directly on the sites themselves.

## 1. Level and structure of polymorphism – effects of linked selection

Under classical models of positive directional selection, a new advantageous allele quickly spreads to fixation. As a result of hitchhiking effects, the variation at adjacent regions is reduced, as neutral alleles linked with the selected mutation become fixed (Maynard Smith & Haigh, 1974). Therefore, strong positive selection leaves a highly characteristic signature in the molecular data involving a reduction in diversity around the selected locus (see Wang *et al.*, 1999 for a well-studied example in maize). By contrast, balancing selection caused by overdominance or negative frequency-dependent selection generates a peak of diversity near the site under selection, as has been shown for plant self-incompatibility loci (Ruggiero *et al.*, 2008; Schierup & Vekemans, 2008) and disease resistance genes (Tian *et al.*, 2002). The most prominent test using this information is the Hudson–Kreitman–Aguadé (HKA) test (Hudson *et al.*, 1987). This test and its derivatives (Wright & Charlesworth, 2004; Innan, 2006) use polymorphism data from several loci and correct for differences in mutation rate by incorporating divergence information. Under neutrality, polymorphism and divergence are proportional because they both depend on the neutral mutation rate. Any excess or deficit in diversity could be indicative of the action of selection (balancing selection and positive selection, respectively) on at least one of the loci.

Several widely used neutrality tests rely on information given by the site frequency spectrum (SFS), which summarizes the allele frequencies of polymorphisms in the sample and whose shape is strongly affected by different forms of selection (Tajima, 1989; Fu & Li, 1993; Fay & Wu, 2000). For example, under a selective sweep there can be an excess of new low-frequency mutations following the fixation of an advantageous allele (Braverman *et al.*, 1995). In addition, with recombination the SFS following a sweep exhibits an excess of high-frequency derived alleles compared with the neutral SFS (Fay & Wu, 2000), because neutral alleles become partially swept to fixation. By contrast, under balancing selection the SFS tends to be enriched in intermediate frequency alleles. Tests based on the SFS are among the most widely implemented, primarily because only polymorphism data is required, without the need for close outgroup sequences to control for mutation rates.

More advanced methods to detect selection have also been devised, such as the composite likelihood ratio test (CLRT) of Kim & Stephan (2002) or the goodness-of-fit (GOF) test of Jensen *et al.* (2005), which both use an explicit model of positive selection. It has been shown (Thornton & Jensen, 2007) that the application of these methods on a set of preselected loci showing extreme patterns of variation (as is often typical in population genetic studies) creates an ascertainment bias resulting in a high rate of false positives (i.e. loci inferred to be under selection when they are actually neutrally evolving). This ascertainment bias can be partly corrected (Thornton & Jensen, 2007). Nevertheless, a conceptual advantage of these approaches is that rather than simply rejecting the standard neutral model they allow for explicit comparisons of models of selection and neutrality.

Another typical signature of strong positive selection is an excess of LD between polymorphisms around selected loci (Maynard Smith & Haigh, 1974). Several tests have been devised that incorporate LD information to detect selection (Hudson *et al.*, 1994; Kelly, 1997; Depaulis & Veuille, 1998; Andolfatto *et al.*, 1999; Sabeti *et al.*, 2002; Kim &

Nielsen, 2004). However, as noted by Przeworski (2002) and McVean (2007) the LD signature left by selective sweeps tends to dissipate very quickly once the selected mutation has reached fixation. Therefore, methods aimed at detecting complete sweeps using LD have a fairly narrow time window during which the power is sufficient.

In addition to detecting the fixation of advantageous mutations, researchers have also been interested in developing methods to detect the ongoing spread of an advantageous allele, known as a partial selective sweep. These methods also use LD information (Hudson *et al.*, 1994; Sabeti *et al.*, 2002; Voight *et al.*, 2006) and are based on the principle that the sudden rise in frequency of a selected mutation leaves less time for recombination to break up the haplotype carrying the mutation than if the mutation was neutral. As a result, the observation of a high-frequency haplotype exhibiting an unusually long-ranging LD is a strong clue indicating the action of directional selection.

The final category of test is based on the concept of genetic hitchhiking applied to subdivided populations and traces back to Lewontin & Krakauer (1973). The idea is once again to detect outlier loci, but this time the quantity of interest is the level of differentiation exhibited between populations ($F_{ST}$). The rationale is that if selection favours different alleles in different populations, this should increase the allele frequency differences between populations (and therefore $F_{ST}$) compared with neutral loci (Charlesworth *et al.*, 2003). On the other hand, if selection favours the same allele in different populations, a lower level of differentiation is expected than genetic drift acting alone. The main problem is therefore to determine the expected $F_{ST}$ distribution under neutrality. Beaumont & Nichols (1996) and Vitalis *et al.* (2001) used coalescent simulations to determine the expected $F_{ST}$ distribution. Recent Bayesian approaches involve more realistic scenarios in which the migration rate can differ between pairs of subpopulations (Beaumont & Balding, 2004; Foll & Gaggiotti, 2008).

## 2. Comparison of polymorphism and divergence for different classes of mutations

In addition to tests of the effect of linked selection on neutral diversity, comparisons of different classes of mutation allow for direct tests of selection at functional sites. The basic premise to this approach was first proposed by McDonald & Kreitman (1991, MK test) and is based on a comparison of two types of mutations both within (polymorphism) and between (divergence) species. Typically, synonymous and nonsynonymous mutations are compared, although in principle the test is applicable for any set of two categories for which one is neutral (Andolfatto, 2008). Under the neutral theory of molecular evolution, synonymous mutations are neutral whereas nonsynonymous mutations are either strongly deleterious or neutral. Under this model, deleterious nonsynonymous mutations contribute negligibly to polymorphism (they are readily eliminated by purifying selection) and the ratio of nonsynonymous ($P_A$) to synonymous ($P_S$) polymorphism ($f = P_A/P_S$) therefore reflects the proportion of new mutations that are neutral. Under complete neutrality, we expect the ratio of nonsynonymous ($D_A$) to synonymous ($D_S$) divergence ($D_A/D_S$) to be equal to $f$ because the ratio for both polymorphism and divergence is a simple function of the fraction of neutral nonsynonymous mutations. However, if some of the nonsynonymous mutations are advantageous, there will be an excess of nonsynonymous divergence, and we can estimate the proportion of substitutions fixed by positive selection as $\alpha = 1 - \frac{D_S P_A}{D_A P_S}$ (Charlesworth, 1994; Smith & Eyre-Walker, 2002). The MK test itself consists of applying a Fisher's exact test to the contingency table with entries $P_A$, $P_S$, $D_A$ and $D_S$; the idea being to determine whether the type of mutations (synonymous vs nonsynonymous) and their status (polymorphism vs divergence) are independent. If independence is rejected it indicates a departure from neutrality.

An important assumption underlying the MK test is that the fraction of nonsynonymous mutations that are nonneutral are strongly deleterious. However, in practice, a substantial fraction of nonsynonymous mutations might be slightly deleterious rather than strongly deleterious. The fate of nonsynonymous mutations is determined by both purifying selection and genetic drift. The result is that these mutations will be counted as polymorphism and sometimes reach fixation, although they will contribute more to polymorphism than to divergence, therefore biasing both estimates of $f$ and $\alpha$. The common method to reduce this bias has been to exclude rare polymorphisms from the analysis, because most weakly deleterious mutations will segregate at low frequency (Fay *et al.*, 2001; Sella *et al.*, 2009). Recently, several studies have developed likelihood methods to estimate the full distribution of fitness effects of deleterious amino acid changes using polymorphism and divergence data (Boyko *et al.*, 2008; Eyre-Walker & Keightley, 2009). These methods allow for an estimate of $\alpha$ that fully accounts for the presence of slightly deleterious mutations.

Finally, the comparison of synonymous and nonsynonymous mutations can readily be extended to a phylogenetic context. The key quantity of interest here is $\omega = d_N/d_S$ where $d_N$ and $d_S$ are the nonsynonymous and synonymous substitution rates, respectively (for a review see Yang & Bielawski, 2000). The idea is quite simple, if there is no selection, synonymous and nonsynonymous substitutions should occur at the same rate and $\omega$ should equal 1. Under negative selection $\omega < 1$ and under positive selection $\omega > 1$. The likelihood framework allows estimation of $\omega$ and refinement of the model to various degrees. For example, $\omega$ can be allowed to vary among the branches of a phylogeny to assess if selection has been more important in one lineage than another, or among sites along the

sequence, such that only some sites would be affected by positive selection.

## III. Population size changes

One of the core assumptions of the SNM is constant population size, yet changes in population size are common in plant populations (Harper, 1977; Silvertown & Charlesworth, 2001). Population size changes can have a number of important effects on genetic variation that complicate inferences of selection (Tenaillon *et al.*, 2004; Haddrill *et al.*, 2005; Wright & Gaut, 2005; Teshima *et al.*, 2006). First, changes in population size, particularly those resulting from population bottlenecks, increase the variance in levels of diversity among genes. This has the effect of increasing the number of false positive tests of genetic hitchhiking when the standard neutral model is assumed (Wright & Gaut, 2005; Andolfatto, 2008). Second, both bottlenecks and population expansion can skew the SFS in similar ways to natural selection, generating genome-wide departures from the SNM. Third, changes in population size will influence levels of LD (Wall *et al.*, 2002). Therefore, molecular signatures characteristic of positive selection can also be generated by changes in population size.

How are different tests of selection likely to be affected by changes in population size? Overall MK-based tests are expected to be less sensitive to demographic assumptions than SFS or LD-based methods (McDonald & Kreitman, 1991). This follows from the fact that synonymous and nonsynonymous mutations are interspersed throughout the genome and should be affected in the same way by demographic events (Nielsen, 2005). However, an important assumption of the MK approach is that the fraction $f$ of neutral mutations is constant over the timescale in which both polymorphism and divergence are being estimated. Indeed, it has been shown that artifactual evidence of adaptive evolution can be obtained with the MK test if some nonsynonymous mutations are slightly deleterious and there has been a population expansion or a bottleneck during divergence (Ohta, 1993; Eyre-Walker, 2002). Moreover, the removal of low-frequency polymorphisms aggravates this problem because it makes the MK test more sensitive to changes in effective population size (Eyre-Walker, 2002; Charlesworth & Eyre-Walker, 2008). Simulation studies also demonstrate that bottlenecks reduce the power to detect adaptive substitutions (Eyre-Walker & Keightley, 2009). Thus, the fraction of adaptive substitutions can be overestimated when significant population expansion occurs and underestimated if there is a recent population bottleneck.

Many plant species are self-compatible and/or capable of clonal reproduction, and this allows new populations to be founded by a very small number of individuals, sometimes only one, creating the potential for severe population bottlenecks during colonization events (Baker, 1955; Pannell & Barrett, 1998; Foxe *et al.*, 2008). Similarly, founder events during speciation may also lead to strong population bottlenecks and, depending on the time since speciation, this could have important effects on patterns of neutral diversity (Gottlieb, 1973; Jakobsson *et al.*, 2006). Although the general role of founder events in speciation has been questioned in recent years (Barton & Charlesworth, 1984; Coyne & Orr, 2004), two common modes of plant speciation, namely reproductive isolation resulting from the evolution of selfing and allopolyploid speciation, can involve origins from a small number of individuals (Jakobsson *et al.*, 2006; Foxe *et al.*, 2009). Given recent evidence that a significant percentage of plant speciation events involve polyploidy (Wood *et al.*, 2009), there is thus the potential for many species to be recovering from severe speciation bottlenecks, although multiple origins of polyploids may not be uncommon (Soltis & Soltis, 1993). Finally, a major focus of studies of selection in plants has been on cultivated species, and for these lineages the domestication process is almost invariably accompanied by a loss of genetic variation through bottlenecks and strong artificial selection (Gaut & Clegg, 1993; Thuillet *et al.*, 2005; Wright *et al.*, 2005; Caicedo *et al.*, 2007; Haudry *et al.*, 2007).

A growing number of studies of nucleotide variation using coalescent models provide quantitative evidence for strong signatures of recent size changes in plant populations (Table 2). These studies take advantage of the development of coalescent methods to fit the data to demographic and speciation parameters. The basic approach involves varying the parameters associated with ancestral and present-day population sizes, and fitting the data to these parameters. Evidence for population bottlenecks associated with the evolution of selfing (Foxe *et al.*, 2009; Ness *et al.*, 2010) and allopolyploid speciation (Jakobsson *et al.*, 2006) are consistent with the notion that founder events are likely to play an important role in many plant speciation events, especially in groups capable of long-distance dispersal.

Glacial cycles can also cause colonization bottlenecks (*Arabidopsis lyrata*, see Ross-Ibarra *et al.*, 2008) as well as rapid population expansion (*Populus balsamifera*, Keller *et al.*, 2010). Detailed surveys involving very large samples have shown a strong signal of a recent founder event in North American populations of *Arabidopsis thaliana*, with stronger patterns of relatedness over extensive geographic regions compared with European populations (Platt *et al.*, 2010). Studies of European *A. thaliana* are consistent with an advancing wave of colonization from east to west following glaciation (François *et al.*, 2008). In domesticated species, there is strong evidence for population bottlenecks of varying severity from near-complete loss of variation in wheat (Thuillet *et al.*, 2005; Haudry *et al.*, 2007), to minimal signs of population bottlenecks in alfalfa (Muller *et al.*,

**Table 2** Studies of selection on multilocus sequence data that fit demographic models with population size changes

| Species | Approach for parameter inference | Patterns of genetic variation compared with Standard Neutral Model | Coalescent inference | Population subdivision | Reference |
|---|---|---|---|---|---|
| *Arabidopsis lyrata* | ABC | Reduced diversity in postglacial populations, excess linkage disequilibrium, excess intermediate-frequency variants | Severe population bottlenecks in most populations | High population structure species-wide | Ross-Ibarra *et al.* (2008) |
| *Arabidopsis suecica* | ABC | Very low polymorphism | Polyploid speciation from a single founding individual | – | Jakobsson *et al.* (2006) |
| *Arabidopsis thaliana* | ABC | Excess of rare variants | Population expansion following glacial episode | Strong isolation-by-distance | François *et al.* (2008) |
| *Capsella bursa-pastoris* | IM | Haplotype sharing with *C. rubella* | Population bottleneck following polyploid speciation, followed by population growth and introgression from *C. rubella* | – | Slotte *et al.* (2008) |
| *Capsella rubella* | MIMAR | Strong loss of variation and increase in linkage disequilibrium relative to outcrossing congener *Capsella grandiflora* | Severe population bottleneck associated with the evolution of selfing | – | Foxe *et al.* (2009) |
| *Eichhornia paniculata* | MIMAR | Low diversity in selfing populations and excess of rare variants | Bottleneck associated with the colonization of the Caribbean | High population structure | Ness *et al.* (2010) |
| Sunflowers (*Helianthus annuus* and *Helianthus petiolaris*) | IM | Excess of rare variants | Population growth | – | Strasburg & Rieseberg (2008) |
| *Medicago truncatula* | ABC | Excess of rare and high frequency variants | Population expansion | High population structure | De Mita *et al.* (2007) |
| Norway Spruce (*Picea abies*) | Coalescent simulations | Excess of rare variants | Ancient bottleneck followed by moderate expansion | Substantial population structure | Heuertz *et al.* (2006) |
| Spruces from Tibetan plateau (four *Picea* species) | IM/ABC | Excess of low frequency variants overall with excess of high-frequency variants for *P. schrenkiana* and *P. purpurea* | *P. likiangensis* and *P. wilsonii* compatible with SNM, *P.schrenkiana* bottleneck, *P. purpurea* population growth | High population structure | Li *et al.* (2010) |
| Scots Pine (*Pinus sylvestris*) | Coalescent simulations | Excess of rare variants | Moderate population bottleneck in northern populations | Low population structure | Pyhäjärvi *et al.* (2007) |
| *Populus tremula* | ABC | Excess of rare and high frequency variants | Bottleneck | – | Ingvarsson (2008) |
| Balsam Poplar (*Populus balsamifera*) | LAMARC | Excess of low-frequency variants | Population expansion following a glacial episode | Three main genetic clusters | Keller *et al.* (2010) |
| Douglas Fir (*Pseudotsuga menziesii*) | ABC | Excess of rare variants (perhaps followed by a recent bottleneck) | Population expansion | Low population structure | Eckert *et al.* (2009) |

**Table 2** (*Continued*)

| Species | Approach for parameter inference | Patterns of genetic variation compared with Standard Neutral Model | Coalescent inference | Population subdivision | Reference |
|---|---|---|---|---|---|
| Tomato (*Solanum chilense* and *S. peruvianum*) | Isolation model (similar to IM except with no subsequent migration) | Quite similar to SNM, when high diversity subpopulations pooled | Population growth in *S. peruvianum*, stable population in *S. chilense* | – | Städler *et al.* (2008) |
| Wheat (*Triticum turgidum*, *Triticum dicoccum*) | Coalescent simulations | Low polymorphism in domesticated populations compared to the wild progenitor | Severe bottleneck following domestication | – | Haudry *et al.* (2007) |
| Maize (*Zea mays*) | Coalescent-based likelihood | Reduced variation, excess linkage disequilibrium, excess of high frequency variants | Recent bottleneck following domestication from Teosinte (*Z. mays* ssp. *parviglumis*) | – | Wright *et al.* (2005) |

Information regarding the level of population subdivision is indicated when available. ABC, Approximate Bayesian computation (see Fig. 1 and Marjoram & Tavare, 2006); IM, Isolation–Migration model (see Hey & Nielsen, 2004); MIMAR, MCMC estimation of the Isolation–Migration model Allowing for Recombination (see Becquet & Przeworski, 2007); LAMARC, Likelihood Analysis with Metropolis Algorithm using Random Coalescence (see Kuhner, 2006)

2006). Although not exhaustive, Table 2 emphasizes how prevalent historical changes in population size are in many plant species, particularly those that are annual and self-compatible. Table 2 also shows evidence of bottlenecks for long-lived plant species such as trees, where even ancient bottlenecks can influence present-day patterns of polymorphism. With more comparative datasets of this kind, it will be interesting to quantitatively compare the extent of historical population size fluctuations and effective sizes among plant species that vary in life history and mating system.

As a result of growing recognition of the importance of population size changes, there is now increased effort to incorporate the basic ingredients of demography in building more realistic null models. The underlying idea is that whereas selection will only affect particular genes and the adjacent linked regions, demography affects the entire genome more or less evenly. Therefore, if one has a plausible demographic scenario for the populations of interest, it is – at least in principle – possible to simulate what the polymorphism pattern under this scenario is likely to be and look for outliers putatively under selection (see Box 1 for an explanation of this principle). This has been rendered possible by the increased availability of highly flexible simulation tools such as Hudson's (2002) ms software. Most of these simulation tools use coalescent modelling of the genealogical history of the sample backward in time (Hudson, 1991), although a fast and efficient simulation program simulating entire populations forward in time has also been developed (Hernandez, 2008). More specifically, recent studies have

Demography can strongly influence the distribution of most diversity statistics or neutrality test statistics such as Tajima's *D* or Fay and Wu's *H*. Therefore, if one uses the standard neutral model (SNM) to assess if the locus is an outlier when the true demographic model substantially deviates from SNM, it can generate a high proportion of false positives.




The problem is therefore to generate the distribution of the statistic(s) of interest under a demographic model congruent with the observed data. Coalescent simulations can easily accommodate some level of complication such as population expansion, bottlenecks, splitting and so forth. Once the parameters are estimated, the distribution of the statistic of interest can be simulated under the most likely demographic model and then an assessment is made as to how unusual a locus is with regard to this distribution. This approach can be powerful although time consuming if, for example, the demographic parameters have to be estimated through ABC (see text and Fig. 1). Furthermore, it will be conservative when a large proportion of the genome is under selection.

**Box 1** Model-based approach for the detection of outliers in DNA sequence data.

New Phytologist

aimed at detecting selection while explicitly modelling population size changes. As an example, Li & Stephan (2006) fitted a complex demographic model for *Drosophila melanogaster* populations, including a population expansion following the spread out of Africa and a bottleneck in Europe, using coalescent simulations conditioned on the observed joint SFS (see Section IV Population subdivision for more detail on the joint SFS) and proceeded to detect outliers.

Several attempts have been made to combine demographic fits of population size change with tests of selection in plant populations. For example, Wright *et al.* (2005) modelled the divergence of two populations (teosinte and maize) and estimated the bottleneck severity parameter ($k$) that best explained the maize data. Using a likelihood approach, they showed that a model allowing an additional class of genes under a more severe bottleneck was more likely than a model assuming a single bottleneck parameter for all genes, consistent with the idea that a subset of loci were under directional selection. Each locus was then given a posterior probability of being in the selected class, providing a ranked order list of candidate selected genes. Similarly, De Mita *et al.* (2007) calibrated a population expansion model in *Medicago truncatula* using a set of 24 reference loci through Approximate Bayesian Computation (ABC see Fig. 1). They then tested how a few candidate loci departed from the 'neutral envelope' simulated from the demographic model they identified.

It is important to appreciate that these approaches are only as good as the demographic model that is inferred. When outliers are identified they may be the result of a poor fit to the true demographic history rather than because of selection. An alternative and perhaps more powerful approach is to use many genes dispersed throughout the genome, and use a semi-nonparametric approach to identify regions with patterns of polymorphism that depart significantly from the rest of the genome (Nielsen *et al.*, 2005). This method explicitly quantifies the departure of one genomic region from patterns of diversity (e.g. the SFS) across the genome, allowing for a more empirical measure of unusual patterns of local diversity within the genome. Although significance levels still require that a demographic model is specified, the method is quite robust to uncertainty in the underlying model, mitigating the dependence of results on exact inference of demographic history. However, as with any method for identifying unusual loci with an empirical distribution, this approach will tend to miss regions under selection if a substantial part of the genome is affected by recurrent selective sweeps (Sella *et al.*, 2009). Although crucial, the assumption that selection must not be pervasive for this sort of test to have power is rarely mentioned explicitly. Nevertheless, this approach may provide one of the most robust means of identifying selected regions as genome-wide polymorphism datasets become increasingly available for plant populations.

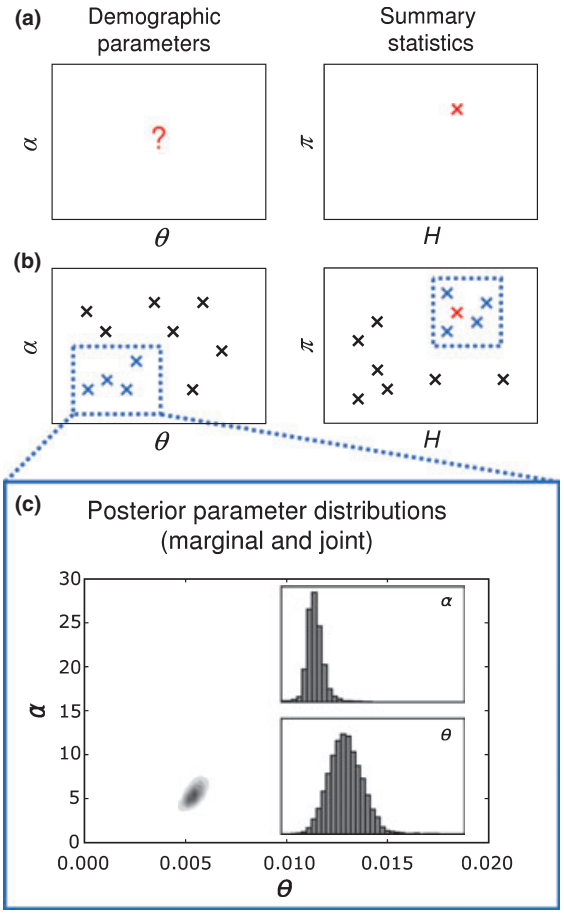


**Fig. 1** Estimation of demographic parameters through the use of Approximate Bayesian Computation (ABC). Suppose we want to estimate the parameters for a demographic model that is hypothesized to have given rise to the observed data. In the example, the model is determined by two parameters (which could be the population growth rate ($\alpha$) and the population mutation rate ($\theta$), for example, if the underlying model is assumed to be a model with a single expanding population) (a). Draw values for each parameter from prior distributions then simulate under the demographic model using these values (b). Compute a set of summary statistics (here we suppose there are two summary statistics Fay and Wu's $H$ and $\pi$, but there can be any number) on the simulated data (black crosses) and see how they compare with the same statistics calculated from the observed data (red cross). Simulated data within close distance of the observed data (blue crosses) are retained and the parameters can be estimated from the approximate posterior distribution obtained from the retained simulations (c). The total procedure can be iterated using parameter values from the posterior distribution estimated during the previous round. The joint posterior distribution describes the probability density of all parameters, taking into account all potential associations. Marginal posterior distributions can be computed for any parameters by integrating over all other parameters. A number of improvements from the initial rejection-sampling procedure have helped to make ABC applications faster and more accurate in their approximation of the posterior distribution. These are beyond the scope of this paper (for more detailed references see Beaumont *et al.*, 2002; Wegmann *et al.*, 2009; Blum & François, 2010; Leuenberger & Wegmann, 2010). Figure inspired by S. De Mita, with thanks.

## IV. Population subdivision

Plants may be particularly susceptible to the effects of population structure because of their immobile habit, tendency to mate with near neighbours, and local dispersal of the majority of seeds in seed crops. Two major concerns arise when considering the effects of population structure on inferences of selection. First, as with population size changes discussed above, population subdivision creates departures from neutral expectations, and therefore increases the rate of false positives when scanning for selected regions. Second, restricted gene flow (Levin & Kerster, 1974) and/or contrasting selection pressures resulting in local adaptation (Linhart & Grant, 1996) across the species range may slow or prevent the global spread of advantageous alleles (Charlesworth *et al.*, 2003). These effects can hinder the ability to detect selection, particularly in species-wide samples, where individuals are sampled extensively across the species distribution.

### 1. Models of population subdivision

In contrast to the efforts made to incorporate population size changes into studies of selection on nucleotide diversity, the fit of explicit models of population subdivision to data is still in its infancy. This problem is partly caused by the vast range of possible parameter space that needs to be considered in such models. Nevertheless, progress has been made in predicting the effects of population subdivision on neutral diversity under several limiting assumptions. One of the most common models of population subdivision is Wright's island model, which assumes equal migration rates and population sizes across a constant number of subpopulations, or demes (Wright, 1931). The properties of the island model for a range of deme numbers from two to infinity have been considered in these models.

Theory and simulation studies using the island model emphasize the importance of sampling schemes when considering the effect of subdivision on patterns of genetic variation. Perhaps counter-intuitively, samples taken from a single subpopulation under this model often exhibit a high variance in the amount of diversity, increased LD and highly skewed allele frequencies because of the immigration of unusual alleles (Städler *et al.*, 2008). This situation is accentuated as the rate of migration decreases, as migration events generate distinct haplotypes. By contrast, 'scattered' samples consisting of a single sample per deme for many demes are more likely to approximate neutral coalescent processes, particularly with a large number of demes (Wakeley, 2003). 'Pooled' samples, consisting of more than one sample per population for multiple populations, create patterns that are intermediate between the two. Careful consideration is required in plant species with broad geographical ranges as to the most suitable sampling scheme for molecular studies.

The results obtained for the island model of migration are not restricted to this form of population subdivision. De & Durrett (2007) modelled a stepping-stone model of population structure, where migration is more likely to occur between local populations. They found that local population samples created strongly skewed SFS and a strong excess of LD, potentially generating spurious signatures of selection. Recent theoretical work suggests that models with a large numbers of demes and those with more biologically realistic forms of population structure may converge with results from the island model (Matsen & Wakeley, 2006). However, when population size changes and/or extinction and recolonization processes (metapopulation dynamics) are added to these models, skewed allele frequencies also become a feature of scattered samples (Pannell, 2003; Städler *et al.*, 2009).

Biologically realistic models of population structure are not only problematic for standard tests of hitchhiking at neutral sites, but they can also influence tests that have been traditionally thought to be more robust to demographic assumptions. For example, metapopulation processes have been shown to increase the variation across loci in levels of differentiation, which could lead to an excess of false positives when using population structure statistics to test for local adaptation (Pannell, 2003). Moreover, in situations where population structure is hierarchical, for example, where samples are obtained from several populations within each of several broad geographic regions, a naive use of $F_{ST}$-based tests of local adaptation results in a large proportion of false positives (Excoffier *et al.*, 2009). Finally, using MK approaches Gossmann *et al.* (2010) found that under a two-deme island model a pooled sample of alleles from both populations generated a spurious signature of positive selection, whereas a single-deme sample under this model did not. However, where a large number of demes are sampled (many-demes limit) MK-based inferences on the strength of selection are robust to subdivision, either with scattered or within-population samples (Wakeley, 2003). In general, models suggest that sampling broadly from many demes will provide the best approach for inferring historical patterns of selection across the genome.

### 2. The extent of subdivision in plant populations

Concerns about the effect of subdivision on inferences of selection present a number of pressing questions to workers interested in the population genomics of plant adaptation. To what extent is subdivision strong enough in plant populations to create problems for inferring selection at the molecular level? Do most species conform to the 'many-deme' or 'few-deme' models of population structure? How extensive is gene flow in plant populations? Despite extensive work on measuring population differentiation in plants both at the 'ecotype' level through common garden and

transplant studies (Langlet, 1971; Linhart & Grant, 1996) and at marker loci (Hamrick & Godt, 1996), we are still some way from being able to answer these questions with confidence.

Levels of population differentiation are typically quantified using a variant of Wright's $F_{ST}$ parameter, which measures the proportion of variation in a sample that is distributed among populations. However, it is important to realize that estimates of $F_{ST}$ (and its relatives such as $G_{ST}$ and others) applied to genetic variation data are not strictly measures of differentiation (Charlesworth, 1998; Jost, 2008, 2009). This is because these measures are highly influenced by the amount of within-population diversity of the markers that are used. Putting aside these mathematical misconceptions, a number of additional caveats should be borne in mind. Under an idealized island model, $F_{ST}$ is a simple function of effective population size and the migration rate. As a result, it has been commonly used to estimate rates of gene flow among populations. However, departures from the island model assumptions are likely to be common in plants, making quantitative inferences of gene flow difficult (Whitlock & McCauley, 1999). In the extreme case, recently diverged populations with no gene flow will have low values of $F_{ST}$, causing an erroneous inference of high migration rates. For example, Ross-Ibarra *et al.* (2008) used a coalescent model of divergence with no gene flow to pairs of *A. lyrata* populations from North America and Europe. This provided a good fit to simulations of the observed data, even among population pairs with low $F_{ST}$ values. Thus, in this case a fit to the island model implies a rate of gene flow greater than one migrant per generation, whereas the data are more consistent with a model of no gene flow since divergence *c.* 6000 yr ago. Given the common occurrence of range expansion and contraction following glaciation, recent divergence with low levels of gene flow would appear to be a reasonable alternative hypothesis to explain their data.

Despite these caveats, interspecific comparisons of levels of molecular differentiation from various types of markers are consistent with expectations based on mating systems and predicted differences in gene flow. Outcrossing populations typically exhibit lower levels of differentiation than selfing populations, and local samples show less differentiation than those sampled over a broader geographical area (Morjan & Rieseberg, 2004). Multilocus estimates of population differentiation in plants using single nucleotide polymorphisms (SNPs) generally display comparable levels of differentiation to previous studies of $F_{ST}$ using other markers (average $F_{ST}$ = 0.32, Morjan & Rieseberg, 2004). In addition to quantifying differentiation by $F_{ST}$ using populations as units, new cluster-based approaches that assign individuals to populations by minimizing levels of LD have been widely implemented (Pritchard *et al.*, 2000; Gao *et al.*, 2007; Huelsenbeck & Andolfatto, 2007). The general picture to emerge from these studies suggests that plant populations typically cluster into broader regional groupings, and it is not uncommon to find a multilevel hierarchy of geographic structuring revealed by varying the number of clusters and/or treating regional populations separately (Nordborg *et al.*, 2005; Ross-Ibarra *et al.*, 2008; Ness *et al.*, 2010).

## 3. Accounting for subdivision in tests of selection

When testing for selection, several conflicting sampling solutions have been proposed to account for population structure. On one hand, scattered population samples from many populations, ignoring within-population diversity, may best approximate a neutral coalescent process under a broad range of models (Wakeley, 2003; Städler *et al.*, 2008). However, scattered samples do not allow for the investigation of local adaptation and for this goal within-population samples are required (Siol *et al.*, 2008; Bomblies *et al.*, 2010; Turner *et al.*, 2010). This stems from the fact that local adaptation results in levels of differentiation around genes under selection that is greater than expected for neutrally evolving regions. Furthermore, taking samples from multiple populations does not rule out hierarchical population structure; in an extreme example where the species is split into two geographic clusters it could reflect sampling from two demes, leading to genome-wide departures from neutrality (Excoffier *et al.*, 2009). Similarly, if a species is structured as an ancestral, refugial or source population and an advancing wave of colonizing populations (François *et al.*, 2008), it is not clear that a scattered sample will best reflect the history of selection. Combining both within and between population samples should allow for in-depth characterization of population history. Furthermore, integrating data from multiple within-population samples affords the most powerful approach for modelling both population history and selection. Of course, this requires considerable sequencing effort and cost.

A significant advance for selection models with structured populations is the use of the multidimensional allele frequency spectrum (or joint frequency spectrum, Li & Stephan, 2006). This is a natural extension of the SFS discussed in the first section, and describes the joint distribution of polymorphisms across populations. Fig. 2 shows the joint frequency spectrum for two populations having diverged from a common ancestral population. However, the principle can be extended to any number of populations by using a *P*-dimensional matrix. The advantage of the multidimensional SFS is that it provides a more complete summary of the data than traditional SFS summary statistics or $F_{ST}$, which can all be calculated from the multidimensional SFS.

The use of the multidimensional SFS was first introduced by Li & Stephan (2006) in the context of demographic model fitting in a two-population scenario involving

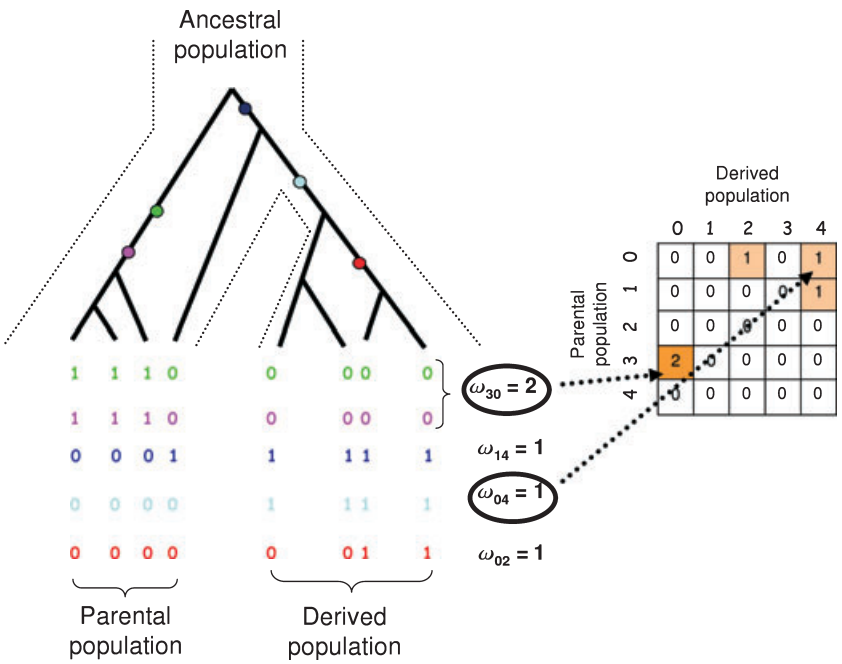


**Fig. 2** The joint frequency spectrum of mutations for two populations derived from the same ancestral population represents the cell counts of the matrix on the right-hand side. In this case, the matrix is of dimension 5 × 5 as each allele can be at a frequency 0 to 4 in each population. The 0s and 1s under the coalescent tree stand for ancestral and derived alleles, respectively; five mutations are considered. So in this example, there are two single nucleotide polymorphisms (SNPs) for which the derived allele is segregating at a frequency of 3 in the parental population, while it is fixed for the ancestral allele in the derived population (so we note $\omega_{30} = 2$). The principle can be extended to any number of populations, for example, for three populations a three-dimensional matrix can be built whose entries record the number of SNPs for which the derived allele was found at frequency *i* in population 1, *j* in population 2 and *k* in population 3. Using the same notation, the cell $\omega_{301}$ of the matrix records the number of sites for which the derived allele is at frequency 3 in population 1, absent from population 2, and at frequency 1 in population 3.

*Drosophila* (and see Hernandez *et al.*, 2007; Gutenkunst *et al.*, 2009; Nielsen *et al.*, 2009). Gutenkunst *et al.* (2009) used a diffusion approximation to fit demographic parameters to the multipopulation SFS. Even though the diffusion framework is in theory applicable to any number of populations, in practice computational issues associated with solving the multidimensional diffusion equation limit its implementation to three. However, simulation approaches could be used to extend to any number of populations, and the use of the multidimensional SFS represents an improvement in our ability to fit realistic demographic scenarios, and use information present in the data more efficiently to test for selection in a demographic context.

Nielsen *et al.* (2009) used information encapsulated in the two-dimensional SFS to propose a new test of neutrality (which they termed the G2D test) that they applied to human genetic data to identify loci subject to local adaptation. A feature of this test is that the null hypothesis is directly derived from the background pattern of variation in the data, similar to the authors' previous work on single populations (Nielsen *et al.*, 2005). This approach avoids relying on a potentially mis-specified population genetic model. More specifically, the test quantifies the fit of the multi-dimensional SFS for a particular genomic region with the global multidimensional SFS observed throughout the genome through the calculation of a (composite) likelihood ratio test. The critical values of the test statistic are determined using coalescent simulations under the demographic model identified from the genome-wide data. It should be noted that although the authors use the composite likelihood ratio test in the case of a two-dimensional frequency spectrum, their approach is readily applicable to higher-dimensional problems, the limiting factor being once again computational feasibility.

The potential to detect the footprint of selection using the G2D test remains to be investigated for a range of demographic scenarios. However, as noted by Nielsen *et al.* (2009), the test should be sensitive to any deviations from neutrality, therefore it should be able to detect various modifications of the multidimensional SFS shape according to the form of natural selection, including purifying selection and local positive selection. It would be interesting to know under what circumstances there is enough power to detect different forms of selection. As an example, Fig. 3 shows the effect of a selective sweep in a derived bottlenecked population. The scenario is similar to the one considered in Thornton & Jensen (2007) and Innan & Kim (2008).

Thornton & Jensen (2007) considered a number of summary statistics and concluded that under this type of scenario $F_{ST}$ was the most powerful statistic for identifying outlier loci compared with statistics based on the frequency spectrum. However, they did not consider using the full joint-frequency spectrum of the two populations. Fig. 3 suggests that the net effect of the selective sweep is to decrease the proportion of shared polymorphisms and to increase the proportion of fixed differences between populations. Whether the signal is strong enough to be detected as statistically significant depends on parameters such as divergence time, migration rate between the populations and the intensity and duration of the bottleneck.

Some progress towards identifying genes under positive selection using structured populations has been achieved using large-scale plant population genomics data. For example, Toomajian *et al.* (2006) used a nonparametric approach to show high haplotype sharing at two independently derived loss-of function alleles at the flowering time gene *FRI* in European populations of *A. thaliana*. Similarly, Turner *et al.* (2010) used two pairs of local populations of *A. lyrata* to screen for candidate loci thought to be involved with local adaptation to serpentine soils. Although these kinds of approaches lack explicit demographic models and are thus nonparametric, the outlier loci that are identified should be enriched for the targets of selection. Integrating the results from such approaches with functional data (e.g. quantitative trait loci (QTL) mapping, association mapping, gene annotation) will provide a powerful approach for the identification of targets of positive selection in plant genomes.

## V. Local adaptation, standing genetic variation, quantitative traits and multiple adaptive substitutions

A substantial amount of adaptation in plant populations may arise from variation that departs from the idealized

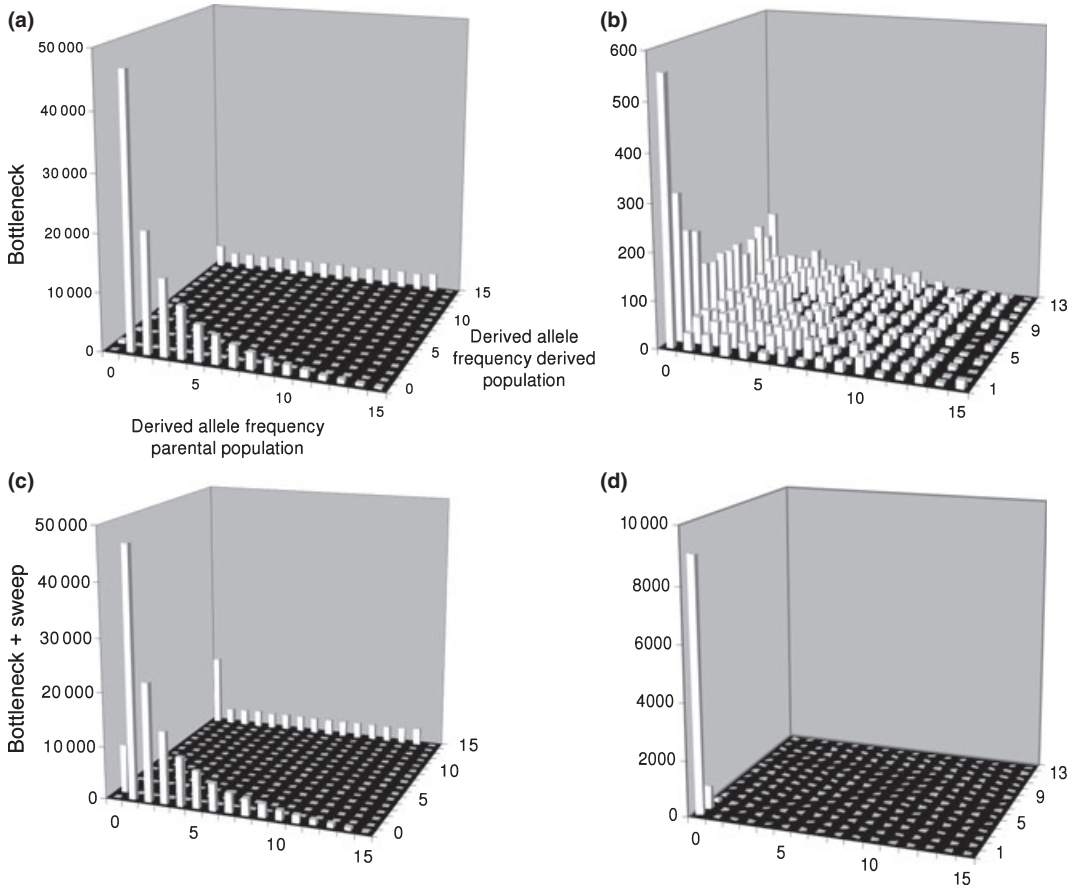


**Fig. 3** The joint frequency spectrum following a bottleneck (a,b) and a bottleneck accompanied by a selective sweep in the derived population (c, d). (a) and (c) represent the full joint frequency spectrum, whereas (b) and (d) zoom in on the sites for which the derived population is polymorphic. The heights of the bars represent the absolute number of polymorphic site segregating at particular frequencies. The selective sweep increases the proportion of fixed differences (a,c) and reduces the number of shared polymorphisms (b,d). We conducted the simulations using the code written by Thornton & Jensen (2007).

model of positive selection. Under the standard model of a selective sweep, a new beneficial mutation arises in a population as a single copy and increases in frequency owing to natural selection of constant strength and direction (Maynard Smith & Haigh, 1974). The extent to which this is typical of most adaptive events remains to be determined, but it is likely that a significant fraction of adaptive evolution does not proceed in this way. First, many adaptations may originate from standing genetic variation that has been present in a population for some time before the new selective episode that assembles the adaptation being considered (this is referred to as a soft selective sweep; Orr & Betancourt, 2001; Hermisson & Pennings, 2005; Pritchard *et al.*, 2010). Second, well-developed population subdivision can slow the spread of an advantageous mutation, making it more likely that an alternative adaptive mutation will occur in a distinct local population before spread of the first advantageous allele through gene flow. Third, fluctuation in the strength and mode of selection across space (diversifying selection resulting in local adaptation) and time violates a simple model of constant selection resulting in the rapid spread of a new beneficial mutation across the species (Harder & Johnson, 2009). Finally, even though there is a growing literature identifying mutations of major effect on phenotype, many adaptive traits are likely to be polygenic, especially those associated with life history.

The quantitative genetics perspective on adaptation is quite different from what has been described so far in our review. Indeed, adaptation is most commonly viewed as the outcome of selection operating at many loci for a given trait (Fisher, 1930; Lynch & Walsh, 1998). The consequences of quantitative inheritance on the traces left by positive selection at the sequence level have been surprisingly underinvestigated. A few studies (Latta, 1998; Le Corre & Kremer, 2003; Chevin & Hospital, 2008) have started to fill the gap by demonstrating that the dynamics of a beneficial mutation affecting a quantitative trait depends not only on its own selection coefficient (the parameter encapsulating the beneficial or deleterious effect of a particular mutation), but also on the genetic variation for this trait at other loci. These studies highlight the fact that strong selection on a quantitative phenotype may not necessarily translate to strong selection on a single locus influencing the trait.

Selection from standing genetic variation may be particularly likely under conditions of rapid environmental change or in the colonization of new environments, such as when invasive species are introduced to new regions (Barrett *et al.*, 2008). Under these circumstances the timescale involved may limit the introduction of new beneficial mutations. Innan & Kim (2004) studied the case of a domestication event, where a previously neutral or slightly deleterious trait in the wild progenitor is strongly favoured by artificial selection (e.g. selection for nonshattering habit in domesticated cereals; Glémin & Bataillon, 2009). Another instance investigated by the same authors (Innan & Kim, 2008) is the local colonization of a novel environment from an ancestral population following a bottleneck (the scenario is depicted in Fig. 2 where the parental population is in the environment of origin and the derived population experiences different selective pressures). The take-home message from these analyses and others (Hermisson & Pennings, 2005; Przeworski *et al.*, 2005; Pennings & Hermisson, 2006) is that the 'typical' signatures of positive selection (reduced levels of polymorphisms in linked regions, increased LD and skewed SFS) exhibit more variance and that many loci under selection are likely to go undetected, depending on the selection coefficient and the initial frequency of the mutation when selection commenced.

Although there are still relatively few examples of adaptive mutations that have been cloned and characterized in plants, a number of those that have been identified suggest that more complex models of adaptation may be the norm. For example, a recent study of trichome evolution in *A. lyrata* demonstrated parallel loss of trichomes in Swedish and Russian populations, through independent loss-of-function mutations in the *glb1* gene (Kivimaki *et al.*, 2007). Similarly, variation in flowering time in *A. thaliana* is mediated, in part, by numerous independent loss-of-function alleles with different geographic distributions and constitutes one of the most well-studied examples of loss-of-function mutations with large phenotypic effects (Alonso-Blanco *et al.*, 2005). Large numbers of independent loss-of-function mutations have similarly been identified in studies of candidate plant disease-resistance genes (Gos & Wright, 2008). Finally, in *Petunia* loss-of-function alleles involved in flower colour have arisen several times independently and have mediated a shift in the types of pollinators attracted to populations (Hoballah *et al.*, 2007). These results suggest that the rate of adaptive mutation may exceed the rate of migration, particularly for loss-of-function changes.

Finally, given the common occurrence of hybridization in plants, gene introgression is likely to be another important source of adaptive genetic variation. Although this possibility was noted early on by Stebbins (1971) and introgression has been well-documented in plants (Baack & Rieseberg, 2007), it has proven more difficult to establish introgression for adaptive alleles. A convincing example concerns regulatory genes controlling the shape of florets that have been introgressed from *Senecio squalidus* to *Senecio vulgaris* and which enhance pollinator attraction (Kim *et al.*, 2008; Chapman & Abbott, 2010). Also, in sunflowers herbivore resistance has been transferred from *Helianthus debilis* to *Helianthus annuus* (Whitney *et al.*, 2006). It is probable that the relative paucity of well-studied examples of adaptive introgression does not accurately reflect the true frequency of such events in plant adaptation.

All of these findings highlight the fact that the signature of positive selection may often be more local and complex

than is generally assumed in standard population genetic models. However, some recent progress has been made in developing methods to better detect selection from standing genetic variation. Innan & Kim (2008) demonstrated that pairwise comparisons of ancestral and derived populations greatly increase the power to detect selection on standing variation. Thus, methods using the joint SFS of ancestral and derived populations (Fig. 2) will likely provide increased power to detect selection following an environmental change or after a colonization event. This also emphasizes the importance of targeted, local population samples in conjunction with further development of methods such as those using LD to identify the targets of recent positive selection (Pennings & Hermisson, 2006; Toomajian *et al.*, 2006) and those based on between-population differentiation (Thornton & Jensen, 2007; Ross-Ibarra *et al.*, 2008; Chen *et al.*, 2010). Thus, while scattered samples from many populations may provide the closest match to standard neutral expectations, local sampling combined with explicit demographic models will also be crucial for the realistic understanding of selection dynamics in structured populations.

## VI. Demographic context of selection and future directions

During the first phase of plant molecular population genetics involving one or a small number of genes, rejection of the standard neutral model (SNM) was most often interpreted as resulting from selection rather than because of departures from demographic assumptions (Wright & Gaut, 2005). Since then important progress has been made in developing methods to fit demographic models to population genomic data, and in attempts to 'control for demography' in searching for the footprint of selection at the molecular level. In comparison with other groups of organisms, multilocus population genetic studies of plants, while still sparse, have provided surprisingly little definitive evidence for positive selection at the genome level. In particular, few studies have identified genes putatively under selection using patterns of neutral variation. The failure to detect genes under selection may be in part result from inherent features of many plants (e.g. immobility, hermaphroditism, clonal propagation) that make them especially vulnerable to demographic violations of the SNM assumptions and to departures from standard models of selective sweeps.

Given the evidence for the prevalence of population structure and the dynamic nature of population size in many plant species, it is likely that population history itself plays an important role in the nature, direction and efficacy of natural selection. Low levels of gene flow enhance the potential for local adaptation (Ronce & Kirkpatrick, 2001), while severe population bottlenecks and small effective population size ($N_e$, Charlesworth, 2009) are expected to reduce the efficacy of positive and negative selection. Recently expanding populations may be subject to high rates of adaptive evolution owing to range expansion (Karasov *et al.*, 2010), but may also be susceptible to bottleneck effects in the newly colonized area that could limit adaptive potential. Thus, understanding demographic history provides more than simply a way to generate the appropriate null model in testing for selection, but is also essential for formulating appropriate hypotheses and models for the detailed action of natural selection.

A key framework for understanding the influence of population history and subdivision on selection is through consideration of the many factors influencing effective population size, a crucial parameter in population genetics theory determining the intensity of genetic drift (Wakeley, 2008). Population genetic theory predicts that in species characterized by low $N_e$, a larger proportion of slightly deleterious and slightly advantageous mutations will be effectively neutral. This stems from the fact that the fate of a selected mutation is determined by two parameters, $N_e$, which determines the intensity of genetic drift, and $s$, the coefficient of selection. More precisely, mutations for which the product $N_e s$ is approximately equal to 1 behave as if they are neutral. As a result, in low-$N_e$ species the efficacy of selection is reduced and the fate of weakly selected mutations is determined more by genetic drift (Ellegren, 2009). Furthermore, in such species the input of mutations will also be lower and beneficial mutations therefore arise less frequently. Depending on the shape of the distribution of fitness effects for deleterious and beneficial mutations, a moderate difference in effective population size could potentially lead to a substantial change in the number of effectively neutral mutations and thus affect the efficacy of natural selection (Kassen & Bataillon, 2006; Bachtrog, 2008; Woolfit, 2009). Thus, low effective populations sizes could influence the intensity of selection on molecular variation.

In general agreement with these basic predictions, it appears that organisms exhibiting higher rates of adaptive evolution and purifying selection may generally be those for which $N_e$ tends to be large (Ellegren, 2009). Fig. 4 illustrates the estimated level of adaptive substitutions for diverse species, using MK-based approaches (Boyko *et al.*, 2008; Eyre-Walker & Keightley, 2009), against the logarithm of their effective population sizes, estimated using neutral polymorphism. At the broadest taxonomic scale there seems to be a correlation between $N_e$ and $\alpha$, similar to the relation that has been found between $N_e$ and the level of purifying selection (see Fig. 1 in Wright & Andolfatto, 2008). However, this figure highlights that many plant species show evidence for relatively low effective population sizes compared with other model systems and relatively few provide evidence for significant adaptive evolution (Bustamante *et al.*, 2002; Nordborg *et al.*, 2005; Schmid

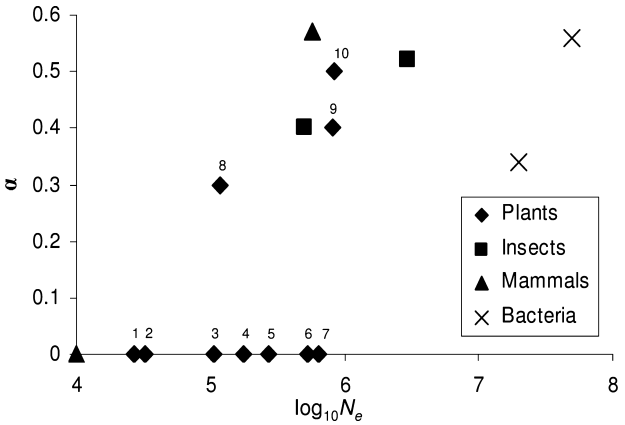


**Fig. 4** Proportion of adaptively driven substitutions for different species plotted against the logarithm of effective population size ($\log_{10}N_e$) estimated from $\theta_W$ (Nordborg *et al.*, 2005; Charlesworth & Eyre-Walker, 2006; Bachtrog, 2008; Eyre-Walker & Keightley, 2009; Gossmann *et al.*, 2010; Halligan *et al.*, 2010; Ingvarsson, 2010; Slotte *et al.*, 2010) and estimates of per generation mutation rate (Charlesworth & Eyre-Walker, 2006; Ingvarsson, 2008; Keightley *et al.*, 2009; Halligan *et al.*, 2010; Ossowski *et al.*, 2010). The plant species are: 1, *Populus balsamifera*; 2, *Schiedea globosa*; 3, *Boechera stricta*; 4, *Oryza rufipogon*; 5, *Arabidopsis thaliana*; 6, *Zea mays*; 7, *Arabidopsis lyrata*; 8, *Populus tremula*; 9, *Capsella grandiflora*; 10, *Helianthus annuus*.

*et al.*, 2005; Kim *et al.*, 2007; Foxe *et al.*, 2008; Gossmann *et al.*, 2010 although see Strasburg *et al.*, 2009; Ingvarsson, 2010; Slotte *et al.*, 2010). It should be noted that these estimates of $N_e$ assume the standard equilibrium neutral model, and thus do not fully account for recent demographic history or population subdivision. In this context is notable that there is little evidence for significant population structure or bottlenecks in the three plant species (*Capsella grandiflora*, *Populus tremula*, *Helianthus annuus* see Table 2) in which there is evidence for high rates of adaptive protein evolution. However, more data points from species with varying demographic histories are clearly needed to better resolve the interaction between demography and adaptive protein evolution in plants.

Disentangling the relative influence of population size, demographic changes and population structure will be difficult because many species with small population sizes also exhibit strong population differentiation and size fluctuations (Nordborg *et al.*, 2005; Muller *et al.*, 2008; Ross-Ibarra *et al.*, 2008; Liti *et al.*, 2009). Furthermore, low rates of species-wide adaptive substitution do not necessarily imply low rates of adaptive evolution. Widespread local adaptation can only be inferred using targeted population samples and currently MK tests do not have a clear analogue at the within-population level, although recent studies that have compared levels of population differentiation at coding vs noncoding regions appear promising (Coop *et al.*, 2009). Finally, as noted by Karasov *et al.* (2010), $N_e$ estimates from neutral polymorphisms represent the harmonic average over a very long period of time and are thus sensitive to periods of low population size. This can lead to very different estimates of effective population size using levels of neutral variability compared with demographic approaches, which may more accurately reflect current effective population size (see Charlesworth, 2009). Thus, while the patterns shown in Fig. 4 highlight the possible importance of effective population size on rates of adaptive evolution, changes in population size and population subdivision likely have a confounding influence.

Exciting as the past decade has been in giving us new insights into the genomic structure of plant populations, the advent of so-called 'next-generation' sequencing holds even more promise. These new techniques generate quantities of data that are orders of magnitude greater than classic sequencing methods and they are now being increasingly applied in the field of population genomics (Simmons *et al.*, 2008; Keightley *et al.*, 2009; Hohenlohe *et al.*, 2010). The continuous decline in sequencing costs, increase in coverage and length of reads, along with the development of powerful *de novo* assembly algorithms for species without a reference genome should allow a broader diversity of plants to be investigated in the near future, including studies of nonmodel species. In particular, it will soon be possible to assay species with diverse life-history traits spanning a much larger range of $N_e$ values, population histories and patterns of subdivision.

Evolutionary analysis of genomic data is still in its infancy and many formidable challenges face the field of evolutionary bioinformatics (for a thorough review, see Pool *et al.*, 2010). The first involves the sheer number of sequences that must be dealt with, which imposes a strong constraint on bioinformatic automation and computational demand. The comparison of observed patterns of variation at thousands of loci makes it all the more difficult to avoid false positives, and inclusion of sequencing errors (appearing as rare SNPs) can skew diversity estimates and the SFS, perhaps leading to spurious inferences. One possible solution is removing rare variants (Turner *et al.*, 2010), but for many analyses low frequency SNPs are of direct interest when testing for the action of selection.

It thus appears that for the first time in population genetics history, the limiting factor is the availability of methods and models and not the data on which to address evolutionary questions. However, such methods are beginning to appear (Jiang *et al.*, 2009; Haubold *et al.*, 2010) and more will surely follow. Even if the challenges are daunting, there are grounds for optimism. The parallel improvement of next-generation sequencing techniques and computational and analytical tools should allow large-scale interspecific comparisons of the historical and contemporary context in which selection operates at the molecular level. These approaches will yield important insights into the interactions

between demography and adaptive evolution in plant populations.

## Acknowledgements

We thank our colleagues Aneil Agrawal, Asher Cutter, Rob Ness and John Stinchcombe for valuable discussions on population genomics and for help with references. Thanks also to Stéphane De Mita and Tanja Slotte who provided advice on an earlier version of this manuscript. MS was supported by a post-doctoral fellowship from the Canada Research Chair's program to SCHB; SIW and SCHB acknowledge support from NSERC Discovery Grants.

## References

**Alonso-Blanco C, Mendez-Vigo B, Koornneef M. 2005**. From phenotypic to molecular polymorphisms involved in naturally occurring variation of plant development. *International Journal of Developmental Biology* **49**: 717–732.

**Andolfatto P. 2008**. Controlling type-I error of the McDonald–Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* **180**: 1767–1771.

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

**Baack EJ, Rieseberg LH. 2007**. A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics & Development* **17**: 513–518.

**Bachtrog D. 2008**. Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evolutionary Biology* **8**: 334–348.

**Baker HG. 1955**. Self-compatibility and establishment after long-distance dispersal. *Evolution* **9**: 347–349.

**Barrett SCH, Colautti RI, Eckert CG. 2008**. Plant reproductive systems and evolution during biological invasion. *Molecular Ecology* **17**: 373–383.

**Barton NH, Charlesworth B. 1984**. Genetic revolutions, founder effects, and speciation. *Annual Review of Ecology, Evolution and Systematics* **15**: 133–164.

**Beaumont MA, Balding DJ. 2004**. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**: 969–980.

**Beaumont MA, Nichols RA. 1996**. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B Biological Sciences* **263**: 1619–1626.

**Beaumont MA, Zhang W, Balding DJ. 2002**. Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

**Becquet C, Przeworski M. 2007**. A new approach to estimate parameters of speciation models with application to apes. *Genome Research* **17**: 1505–1519.

**Blum M, François O. 2010**. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing* **20**: 63–73.

**Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D. 2010**. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genetics* **6**: e1000890.

**Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR *et al.* 2008**. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* **4**: e1000083.

**Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995**. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783–796.

**Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002**. The cost of inbreeding in *Arabidopsis*. *Nature* **416**: 531–534.

**Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR *et al.* 2007**. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genetics* **3**: 1745–1756.

**Chapman MA, Abbott RJ. 2010**. Introgression of fitness genes across a ploidy barrier. *New Phytologist* **186**: 63–71.

**Charlesworth B. 1994**. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* **63**: 213–227.

**Charlesworth B. 1998**. Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* **15**: 538–543.

**Charlesworth B. 2009**. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**: 195–205.

**Charlesworth B, Charlesworth D, Barton NH. 2003**. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution and Systematics* **34**: 99–125.

**Charlesworth J, Eyre-Walker A. 2006**. The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution* **23**: 1348–1356.

**Charlesworth J, Eyre-Walker A. 2008**. The McDonald–Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution* **25**: 1007–1015.

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

**Chevin LM, Hospital F. 2008**. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* **180**: 1645–1660.

**Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009**. The role of geography in human adaptation. *PLoS Genetics* **5**: e1000500.

**Coyne JA, Orr HA. 2004**. *Speciation*. Sunderland, MA, USA: Sinauer Associates, Inc.

**De A, Durrett R. 2007**. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics* **176**: 969–981.

**De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki R, Bataillon T. 2007**. Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in Nod factor signaling in *Medicago truncatula*. *Genetics* **177**: 2123–2133.

**Depaulis F, Veuille M. 1998**. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* **15**: 1788–1790.

**Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB. 2009**. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas Fir (*Pseudotsuga menziesii* var. *menziesii*). *Genetics* **183**: 289–298.

**Ellegren H. 2009**. A selection model of molecular evolution incorporating the effective population size. *Evolution* **63**: 301–305.

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

**Eyre-Walker A. 2002**. Changing effective population size and the McDonald–Kreitman test. *Genetics* **162**: 2017–2024.

**Eyre-Walker A, Keightley PD. 2009**. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* **26**: 2097–2108.

**Fay JC, Wu CI. 2000**. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

**Fay JC, Wyckoff GJ, Wu CI. 2001**. Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.

**Fisher RA. 1930**. *The genetical theory of natural selection*. Oxford, UK: Clarendon Press.

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

**Foxe JP, Dar VU, Zheng H, Nordborg M, Gaut BS, Wright SI. 2008**. Selection on amino acid substitutions in *Arabidopsis*. *Molecular Biology and Evolution* **25**: 1375–1383.

**Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009**. Recent speciation associated with the evolution of selfing in *Capsella*. *Proceedings of the National Academy of Sciences, USA* **106**: 5241–5245.

**François O, Blum MG, Jakobsson M, Rosenberg NA. 2008**. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics* **4**: e1000075.

**Fu YX, Li WH. 1993**. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

**Gao H, Williamson S, Bustamante CD. 2007**. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* **176**: 1635–1651.

**Garrigan D, Lewontin RC, Wakeley J. 2010**. Measuring the sensitivity of single-locus "neutrality tests" using a direct perturbation approach. *Molecular Biology and Evolution* **27**: 73–89.

**Gaut BS, Clegg MT. 1993**. Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). *Genetics* **135**: 1091–1097.

**Gillespie JH. 2000**. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**: 909–919.

**Gillespie JH. 2001**. Is the population size of a species relevant to its evolution? *Evolution* **55**: 2161–2169.

**Glémin S, Bataillon T. 2009**. A comparative view of the evolution of grasses under domestication. *New Phytologist* **183**: 273–290.

**Gos G, Wright SI. 2008**. Conditional neutrality at two adjacent NBS–LRR disease resistance loci in natural populations of *Arabidopsis lyrata*. *Molecular Ecology* **17**: 4953–4962.

**Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010**. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*. doi: 10.1093/molbev/msq1079

**Gottlieb LD. 1973**. Genetic differentiation, sympatric speciation and the origin of a diploid species of *Stephanomeria*. *American Journal of Botany* **60**: 545–553.

**Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009**. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: e1000695.

**Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005**. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**: 790–799.

**Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010**. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics* **6**: e1000825.

**Hamrick JL, Godt MJW. 1996**. Effect of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London B Biological Sciences* **351**: 1291–1298.

**Harder LD, Johnson SD. 2009**. Darwin's beautiful contrivances: evolutionary and functional evidence for floral adaptation. *New Phytologist* **183**: 530–545.

**Harper JL. 1977**. *Population biology of plants*. London, UK: Academic Press.

**Haubold B, Pfaffelhuber P, Lynch M. 2010**. mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology* **19**(Suppl. 1): 277–284.

**Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glémin S *et al.* 2007**. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Molecular Biology and Evolution* **24**: 1506–1517.

**Hermisson J, Pennings PS. 2005**. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.

**Hernandez RD. 2008**. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786–2787.

**Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, Nazareth L, Indap A, Bourquin T, McPherson J *et al.* 2007**. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* **316**: 240–243.

**Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006**. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105.

**Hey J. 2010**. Isolation with migration models for more than two populations. *Molecular Biology and Evolution* **27**: 905–920.

**Hey J, Nielsen R. 2004**. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.

**Hoballah ME, Gubitz T, Stuurman J, Broger L, Barone M, Mandel T, Dell'Olivo A, Arnold M, Kuhlemeier C. 2007**. Single gene-mediated shift in pollinator attraction in *Petunia*. *Plant Cell* **19**: 779–790.

**Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010**. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6**: e1000862.

**Hudson RR. 1991**. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**: 1–49.

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

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

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

**Huelsenbeck JP, Andolfatto P. 2007**. Inference of population structure under a Dirichlet process model. *Genetics* **175**: 1787–1802.

**Ingvarsson PK. 2008**. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180**: 329–340.

**Ingvarsson PK. 2010**. Natural selection on synonymous and nonsynonymous mutations shape patterns of polymorphism in *Populus tremula*. *Molecular Biology and Evolution* **27**: 650–660.

**Innan H. 2006**. Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173**: 1725–1733.

**Innan H, Kim Y. 2004**. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences, USA* **101**: 10667–10672.

**Innan H, Kim Y. 2008**. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics* **179**: 1713–1720.

**Jakobsson M, Hagenblad J, Tavare S, Sall T, Hallden C, Lind-Hallden C, Nordborg M. 2006**. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Molecular Biology and Evolution* **23**: 1217–1231.

**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.

**Jiang R, Tavare S, Marjoram P. 2009**. Population genetic inference from resequencing data. *Genetics* **181**: 187–197.

**Jost L. 2008**. $G_{ST}$ and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.

**Jost L. 2009**. *D* vs. $G_{ST}$: response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology* **18**: 2088–2091.

**Karasov T, Messer PW, Petrov DA. 2010**. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genetics* **6**: e1000924.

**Kassen R, Bataillon T. 2006**. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics* **38**: 484–488.

**Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009**. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* **19**: 1195–1201.

**Keller SR, Olson MS, Silim S, Schroeder W, Tiffin P. 2010**. Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*. *Molecular Ecology* **19**: 1212–1226.

**Kelly JK. 1997**. A test of neutrality based on interlocus associations. *Genetics* **146**: 1197–1206.

**Kim M, Cui ML, Cubas P, Gillies A, Lee K, Chapman MA, Abbott RJ, Coen E. 2008**. Regulatory genes control a key morphological and ecological trait transferred between species. *Science* **322**: 1116–1119.

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

**Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007**. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **39**: 1151–1155.

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

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

**Kimura M. 1983**. *The neutral theory of molecular evolution*. Cambridge, UK: Cambridge University Press.

**Kivimaki M, Karkkainen K, Gaudeul M, Loe G, Agren J. 2007**. Gene, phenotype and function: GLABROUS1 and resistance to herbivory in natural populations of *Arabidopsis lyrata*. *Molecular Ecology* **16**: 453–462.

**Kuhner MK. 2006**. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**: 768–770.

**Kuhner MK. 2009**. Coalescent genealogy samplers: windows into population history. *Trends in Ecology and Evolution* **24**: 86–93.

**Langlet O. 1971**. Two hundred years of genecology. *Taxon* **20**: 653–722.

**Latta RG. 1998**. Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist* **151**: 283–292.

**Le Corre V, Kremer A. 2003**. Genetic variability at neutral markers, quantiative trait loci in a subdivided population under selection. *Genetics* **164**: 1205–1219.

**Leuenberger C, Wegmann D. 2010**. Bayesian computation and model selection without likelihoods. *Genetics* **184**: 243–252.

**Levin DA, Kerster HW. 1974**. Gene flow in seed plants. *Evolutionary Biology (New York)* **7**: 139–220.

**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.

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

**Li Y, Stocks M, Hemmila S, Kallman T, Zhu H, Zhou Y, Chen J, Liu J, Lascoux M. 2010**. Demographic histories of four spruce (*Picea*) species of the Qinghai–Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution* **27**: 1001–1014.

**Linhart YB, Grant MC. 1996**. Evolutionary significance of local genetic differentiation in plants. *Annual Review of Ecology and Systematics* **27**: 237–277.

**Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V *et al.* 2009**. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

**Lynch M, Walsh B. 1998**. *Genetics and analysis of quantitative traits*. Sunderland, MA, USA: Sinauer Associates.

**Marjoram P, Tavare S. 2006**. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics* **7**: 759–770.

**Matsen FA, Wakeley J. 2006**. Convergence to the island-model coalescent process in populations with restricted migration. *Genetics* **172**: 701–708.

**Maynard Smith J, Haigh J. 1974**. The hitch-hiking effect of a favourable gene. *Genetical Research* **23**: 23–35.

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

**McVean G. 2007**. The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**: 1395–1406.

**Morjan CL, Rieseberg LH. 2004**. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology* **13**: 1341–1356.

**Muller MH, Leppala J, Savolainen O. 2008**. Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. *Heredity* **100**: 47–58.

**Muller MH, Poncet C, Prosperi JM, Santoni S, Ronfort J. 2006**. Domestication history in the *Medicago sativa* species complex: inferences from nuclear sequence polymorphism. *Molecular Ecology* **15**: 1589–1602.

**Ness RW, Wright SI, Barrett SCH. 2010**. Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylous plant *Eichhornia paniculata*. *Genetics* **184**: 381–392.

**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. *Annual Review of Genetics* **39**: 197–218.

**Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A *et al.* 2009**. Darwinian and demographic forces affecting human protein coding genes. *Genome Research* **19**: 838–849.

**Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005**. Genomic scans for selective sweeps using SNP data. *Genome Research* **15**: 1566–1575.

**Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R *et al.* 2005**. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**: e196.

**Ohta T. 1993**. Amino acid substitution at the *Adh* locus of *Drosophila* is facilitated by small population size. *Proceedings of the National Academy of Sciences, USA* **90**: 4548–4551.

**Orr HA, Betancourt AJ. 2001**. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157**: 875–884.

**Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010**. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**: 92–94.

**Pannell JR. 2003**. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57**: 949–961.

**Pannell JR, Barrett SCH. 1998**. Baker's Law revisited: reproductive assurance in a metapopulation. *Evolution* **52**: 657–668.

**Pennings PS, Hermisson J. 2006**. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genetics* **2**: e186.

**Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Agren J, Bossdorf O, Byers D, Donohue K *et al.* 2010**. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics* **6**: e1000843.

**Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010**. Population genetic inference from genomic sequence variation. *Genome Research* **20**: 291–300.

**Pritchard JK, Pickrell JK, Coop G. 2010**. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**: R208–R215.

**Pritchard JK, Stephens M, Donelly P. 2000**. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

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

**Przeworski M, Coop G, Wall JD. 2005**. The signature of positive selection on standing genetic variation. *Evolution* **59**: 2312–2323.

**Pyhäjärvi T, Garcia-Gil MR, Knurr T, Mikkonen M, Wachowiak W, Savolainen O. 2007**. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* **177**: 1713–1724.

**Ronce O, Kirkpatrick M. 2001**. When sources become sinks: migrational meltdown in heterogeneous habitats. *Evolution* **55**: 1520–1531.

**Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS. 2008**. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* **3**: e2411.

**Ruggiero MV, Jacquemin B, Castric V, Vekemans X. 2008**. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. *Genetical Research* **90**: 37–46.

**Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ *et al.* 2002**. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.

**Schierup MH, Vekemans X. 2008**. Genomic consequences of selection on self-incompatibility genes. *Current Opinion in Plant Biology* **11**: 116–122.

**Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T. 2005**. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**: 1601–1615.

**Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009**. Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics* **5**: e1000495.

**Shendure J, Ji H. 2008**. Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

**Silvertown J, Charlesworth D. 2001**. *Introduction to plant population biology*. Oxford, UK: Blackwell Science.

**Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF. 2008**. Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biology* **6**: e177.

**Siol M, Prosperi JM, Bonnin I, Ronfort J. 2008**. How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. *Heredity* **100**: 517–525.

**Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010**. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution*. doi: 10.1093/molbev/msq1062

**Slotte T, Huang H, Lascoux M, Ceplitis A. 2008**. Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and Evolution* **25**: 1472–1481.

**Smith NG, Eyre-Walker A. 2002**. Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022–1024.

**Soltis DE, Soltis PS. 1993**. Molecular data and the dynamic nature of polyploidy. *Critical Reviews in Plant Sciences* **12**: 243–273.

**Städler T, Arunyawat U, Stephan W. 2008**. Population genetics of speciation in two closely related wild tomatoes (*Solanum* section *Lycopersicon*). *Genetics* **178**: 339–350.

**Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009**. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.

**Stebbins GL. 1971**. *Chromosomal evolution in higher plants*. London, UK: Edward Arnold.

**Strasburg JL, Rieseberg LH. 2008**. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris* – large effective population sizes and rates of long-term gene flow. *Evolution* **62**: 1936–1950.

**Strasburg JL, Scotti–Saintagne C, Scotti I, Lai Z, Rieseberg LH. 2009**. Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. *Molecular Biology and Evolution* **26**: 1341–1355.

**Tajima F. 1989**. Statistical method 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 demography: a multilocus investigation of the domestication process in maize. *Molecular Biology and Evolution* **21**: 1214–1225.

**Teshima KM, Coop G, Przeworski M. 2006**. How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**: 702–712.

**Thornton KR, Jensen JD. 2007**. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* **175**: 737–750.

**Thuillet AC, Bataillon T, Poirier S, Santoni S, David JL. 2005**. Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics* **169**: 1589–1599.

**Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002**. Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **99**: 11525–11530.

**Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, Zhao K, Calabrese P, Dean C, Nordborg M. 2006**. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biology* **4**: e137.

**Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010**. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* **42**: 260–263.

**Vitalis R, Dawson K, Boursot P. 2001**. Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**: 1811–1823.

**Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006**. A map of recent positive selection in the human genome. *PLoS Biology* **4**: e72.

**Wakeley J. 2003**. Polymorphism and divergence for island-model species. *Genetics* **163**: 411–420.

**Wakeley J. 2008**. *Coalescent theory: an introduction*. Greenwood Village, CO, USA: Roberts and Company Publishers.

**Wall JD, Andolfatto P, Przeworski M. 2002**. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**: 203–216.

**Wang Z, Gerstein M, Snyder M. 2009**. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63.

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

**Wegmann D, Leuenberger C, Excoffier L. 2009**. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**: 1207–1218.

**Whitlock MC, McCauley DE. 1999**. Indirect measures of gene flow and migration: $F_{ST}$ not equal to $1/(4Nm + 1)$. *Heredity* **82**(Pt 2): 117–125.

**Whitney KD, Randell RA, Rieseberg LH. 2006**. Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *American Naturalist* **167**: 794–807.

**Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009**. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* **106**: 13875–13879.

**Woolfit M. 2009**. Effective population size and the rate and pattern of nucleotide substitutions. *Biology Letters* **5**: 417–420.

**Wright S. 1931**. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

**Wright SI, Andolfatto P. 2008**. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annual Review of Ecology, Evolution and Systematics* **39**: 193–213.

**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.

**Wright SI, Charlesworth B. 2004**. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.

**Wright SI, Gaut BS. 2005**. Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* **22**: 506–519.

**Yang Z, Bielawski JP. 2000**. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**: 496–503.