Heredity (2007), 1–9

www.nature.com/hdy

SHORT REVIEW

Progress and prospects in mapping recent selection in the genome

KR Thornton[1,4], JD Jensen[1], C Becquet[2] and P Andolfatto[3]
[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA; [2]Department of Human Genetics, University of Chicago, Chicago, IL, USA and [3]Department of Ecology, Behavior and Evolution, Biological Sciences, University of California San Diego, La Jolla, CA, USA

One of the central goals of evolutionary biology is to understand the genetic basis of adaptive evolution. The availability of nearly complete genome sequences from a variety of organisms has facilitated the collection of data on naturally occurring genetic variation on the scale of hundreds of loci to whole genomes. Such data have changed the focus of molecular population genetics from making inferences about adaptive evolution at single loci to identifying which loci, out of hundreds to thousands, have been recent targets of natural selection. A major challenge in this effort is distinguishing the effects of selection from those of the demographic history of populations. Here we review some current progress and remaining challenges in the field.
Heredity advance online publication, 2 May 2007; doi:10.1038/sj.hdy.6800967

Keywords: genome scans; selective sweeps; adaptive evolution; demography; hitchhiking; population bottleneck

Background

Empirical population genetics seeks to use patterns of variability in natural populations in order to understand the evolutionary forces that shape levels of genetic variation in nature. Of particular interest is using population-level DNA variability data to ask when and where recent episodes of natural selection have occurred in genomes, with the ultimate goal being to identify individual adaptive mutations. With such information, we can begin to ask a number of key questions that have motivated decades of research in evolutionary biology: What gene functions typically underlie adaptations in natural populations? Are adaptations typically changes to proteins or changes to when and where genes are expressed (i.e., regulatory changes)? Does natural selection typically act on newly arising mutations, or does it often act on previously neutral or even deleterious mutations? How strong is selection underlying single adaptive mutational changes? Is adaptation frequent enough – and selection strong enough – to leave a noticeable impact on overall levels of genome variability within a species?

Adaptive evolution occurs when a population of organisms reacts to challenges posed by changes to its external environment. Thus, it follows that a particularly fruitful place to look for recent adaptations in the genome may be in groups of species that have relatively recently (on an evolutionary timescale) colonized habitats that are different from those experienced by ancestral populations. Examples include everything from anatomically modern humans to the fruitfly, *Drosophila melanogaster*, both of whom are believed to have had an African origin and relatively recently acquired a cosmopolitan distribution, to marine and freshwater forms of sticklebacks (Reusch *et al*., 2001) and mice (Ihle *et al*., 2006). Parallels between these natural systems and the domestication of plants and animals (e.g., Doebley, 2004; Pollinger *et al*., 2005) have motivated the application of population genetics approaches to understanding the genetics of domestication.

In all of these systems, adaptation to new habitats was likely to have been accompanied by a reduction in population size that affected patterns of variability throughout the whole genome in derived populations. As a motivating example, consider levels of X-linked genetic diversity and intra-locus linkage disequilibrium in recently derived (Netherlands, Europe) and putatively ancestral (Zimbabwe, Africa) populations of *D. melanogaster* (Figure 1). Sub-Saharan Africa is believed to be the ancestral range of *D. melanogaster*, and the species may have colonized Europe roughly 10 000 years ago (Lachaise *et al*., 1988). As this species was exposed to a new colder and wetter climate in Europe, it is likely that adaptation subsequently occurred (David and Capy, 1988). Zimbabwe populations of *D. melanogaster* have been widely studied as a putative ancestral population due to early observations that they are considerably more variable at the nucleotide level (Begun and Aquadro, 1992) and have lower levels of linkage disequilibrium (Haddrill *et al*., 2005a) than non-African populations. In Figure 1, it is apparent that levels of diversity in Europe are lower at the majority of loci, an effect that may be due in large part to a founder effect associated with dispersal from Africa (Li and Stephan, 2006; Thornton and Andolfatto, 2006). A qualitatively similar pattern is

Correspondence: Dr P Andolfatto, Department of Ecology, Behavior and Evolution, Biological Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.
E-mail: pandolfatto@ucsd.edu
[4]Current address: Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA.
Received 6 November 2006; revised 8 February 2007; accepted 16 February 2007

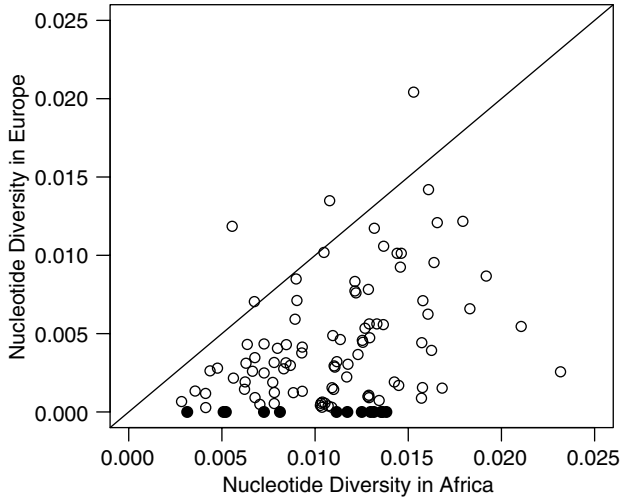

Figure 1 Genetic diversity in *D. melanogaster*. *D. melanogaster* is believed to have originated in Africa and colonized Europe approximately 10 000 years ago (Lachaise *et al*., 1988). Non-African populations have long been recognized as having reduced genetic diversity relative to Africa (Begun and Aquadro, 1992). Here, nucleotide diversity per site is compared between Zimbabwe, Africa and the Netherlands, Europe populations for 105 non-coding loci from the *D. melanogaster* X chromosome (Glinka *et al*., 2003). On average, diversity in Europe is about 40% of that in Zimbabwe. Such a genome-wide effect is probably due in large part to a reduction in population size associated with the colonization of Europe (Haddrill *et al*., 2005a; Thornton and Andolfatto, 2006). However, exposure to novel environments may have been accompanied by adaptation at some loci, and one goal of *Drosophila* population genetics has been to identify regions of the genome subject to recent selective sweeps in non-African populations (Glinka *et al*., 2003; Kauer *et al*., 2003; Bauer DuMont and Aquadro, 2005; Ometto *et al*., 2005; Pool *et al*., 2006). The filled circles are loci with no variability in Europe, which are the best candidate loci to have undergone a recent sweep.

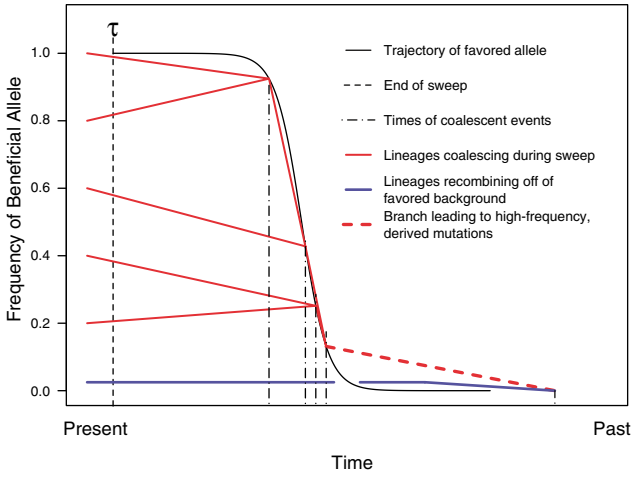

Figure 2 The selective sweep process. This figure illustrates the process of a selective sweep for a sample of six chromosomes drawn at random from a population at the present time. A beneficial allele fixed in the population at time τ. The frequency trajectory of the beneficial allele is given by the logistic curve in black (e.g., Stephan *et al*., 1992), and we superimpose the genealogy of the sample onto this trajectory. Moving backwards in time, the majority of lineages in the sample share their common ancestors during the period of time when the beneficial allele is rapidly increasing in frequency (red lines). These coalescent events occur rapidly during this period (vertical dashed lines). During the sweep, lineages may recombine off of the chromosome carrying the beneficial allele, and onto a background not carrying the beneficial allele (blue line). The most recent common ancestor of the sample is reached in the past, at a time more ancient than the beginning of the sweep. In this figure, branch length is directly proportional to time. Note that the branches that lead to single individual samples are longer than those leading to multiple individuals. Since the number of mutations depends on the product of the mutation rate and time, many mutations in the sample will therefore be rare (present in only a single individual in the sample). Mutations that fall on the dashed red branch of the tree lead to all of the descendants whose history is given by the solid red branches, which are the lineages who share a common ancestor during the sweep. Such mutations will be at high frequency (5/6) in the sample, and comparison with an outgroup species will reveal that they are derived character states. A large number of such high-frequency-derived mutations are predicted to be observed following a recent sweep in a recombining portion of the genome; however, they are also extremely short lived (Przeworski, 2002).

observed in non-African populations of humans, where reduced diversity in these populations and increased levels of linkage disequilibrium are believed to be due to a bottleneck associated with dispersal from Africa about 50 000 years ago (Tishkoff and Verrelli, 2003). Similar patterns are also observed in domesticated maize, which is less diverse than its ancestral form, *teosinte*, probably due in part to a bottleneck associated with domestication (Eyre-Walker *et al*., 1998; Wright *et al*., 2005). Given the observation of reduced diversity genome wide in these species, and of possible effects of bottlenecks associated with recently derived species (i.e., maize) or populations (i.e., *Drosophila* and humans), how does one identify which loci have been targeted by recent natural selection?

Speaking broadly, the most common approach is to ask whether patterns of polymorphism are compatible with the predictions of a particular population genetic model. Two models have been the most widely considered when interpreting patterns of variability. The first is an idealized model of a large, panmictic population experiencing no selection (hereafter, the 'standard neutral model', see Rosenberg and Nordborg, 2002). The second model is that of a selective sweep, where a beneficial mutation arises and moves quickly through the population, sweeping away linked neutral variability in closely linked regions (Maynard Smith and Haigh, 1974; Kaplan *et al*., 1989; Figure 2). Compared to the standard neutral model, the selective sweep model predicts reduced diversity, an excess of rare alleles, elevated linkage disequilibrium and an excess of high-frequency-derived alleles in regions closely linked to a site that has recently experienced a selective sweep (Figure 2). In this article, we focus our discussion on methods to detect recent positive selection using the 'hitchhiking' effect of a sweep on patterns of linked, neutral variability (Maynard Smith and Haigh, 1974; Kaplan *et al*., 1989; Figure 2). We do not consider two related, and important, classes of methods – those comparing patterns of polymorphism and divergence (e.g., Hudson *et al*., 1987; McDonald and Kreitman, 1991; Bustamante *et al*., 2003) and those seeking to estimate the strength of selection on alleles currently segregating in natural populations (Williamson *et al*., 2005; Zhu and Bustamante, 2005), both of which have been reviewed recently by Nielsen (2005).

Historically, violations of the standard neutral model in directions predicted by a selective sweep (e.g., reduced variability, distortions in the distribution of

polymorphism frequencies or high levels of linkage disequilibrium) have been taken as evidence for recent positive selection (e.g., Parsch *et al*., 2001; Harr *et al*., 2002). This approach has generally relied on comparing observed summaries of the data to the predictions of the standard neutral model, and has largely been applied on a gene-by-gene basis. This approach has two major limitations. First, when selection is truly acting, the power to detect it tends to be rather low (e.g., Simonsen *et al*., 1995; Przeworski, 2002). Second, a rejection of the null model does not imply that any particular alternative model is accepted, and we therefore cannot rule out the possibility that rejecting the standard neutral model is due to a violation of any one of a number of assumptions of that model (e.g., including the assumption of a constant population size over time or random mating). In fact, many of the predictions of the selective sweep model can be mimicked by purely demographic scenarios (see Figure 3, Table 1).

In recent years, the focus has turned to the analysis of large multi-locus data sets (e.g., Akey *et al*., 2002; Ometto *et al*., 2005; Wright *et al*., 2005), and now, with the recent completion of the human HapMap project (IHC, 2003), whole-genome genotype data. These larger datasets have fostered the development of methods designed to distinguish the effects of locus specific selection, from those of demography, which will have genome-wide effects. There are now many different tests for detecting selective sweeps from DNA sequence data. The simplest is a genome scan approach in which outlier loci are identified based on the empirical distribution of some chosen feature of the data; or similarly, outlier loci that are not compatible with some plausible demographic model (Hudson *et al*., 1987; Glinka *et al*., 2003; Ometto *et al*., 2005; Wright *et al*., 2005; Voight *et al*., 2006). Related classes of methods are those meant for regions which have been localized through other approaches, and as such are commonly used in tandem with genome scan approaches. In practice, this has often meant identifying putatively swept regions of the genome from screens of a large number of loci, and then following up with localized re-sequencing studies (Harr *et al*., 2002; Bauer DuMont and Aquadro, 2005; Beisswanger *et al*., 2006; Pool *et al*., 2006). More sophisticated *post hoc* tests employ features of the polymorphic site frequency spectrum (SFS) and patterns of linkage disequilibrium (Kim and Stephan 2002; Kim and Nielsen, 2004). Recent tests combine these two approaches, and are designed to analyze genomic-scale data and directly identify regions that have been affected by recent selective sweeps, by considering the background allele frequencies in the sample as a null model (Nielsen *et al*., 2005).

Genome scans for selection

Perhaps the most common approach to identifying loci under selection is the 'hitchhiking mapping' method (Harr *et al*., 2002; Schlotterer, 2003). Generally speaking, this method is not model based, but rather employs a two-tiered approach. First, a large number of regions of the genome are scanned for levels of variability in one or more populations, for example, using either microsatellite (Payseur *et al*., 2002; Payseur and Nachman, 2002; Schlotterer, 2003; Bauer DuMont and Aquadro, 2005) or single nucleotide polymorphism (SNP) markers (Glinka *et al*., 2003; Akey *et al*., 2004; Ometto *et al*., 2005). The data from this initial scan of variability are then summarized by some statistic (e.g., by levels of diversity, or relative levels of diversity in two populations) to estimate the genome-wide distribution of the summary statistic (an 'empirical distribution', e.g., Figure 1). Regions that show extreme values of a given statistic can be identified and investigated more thoroughly by re-sequencing (e.g., Harr *et al*., 2002; Bauer DuMont and Aquadro, 2005; Beisswanger *et al*., 2006; Pool *et al*., 2006). Model-based methods to test for selection are then typically applied to these re-sequenced regions (see below). The first step is thus a 'genome scan' for unusual regions, and the second step often comprises a specific hypothesis test about the role of selection acting on the region.

Importantly, hitchhiking mapping assumes that loci in the tails of an empirical distribution are the most likely to have undergone recent directional selection. Since this approach is not model based, strong assumptions need to be made about how frequent selection is in the genome. For example, if the frequency of selection in the genome is low relative to the density of markers surveyed, the tails of an empirical distribution will mostly contain false positives. At the other extreme, if recent selection is extremely common in the genome, limiting one's focus to the tails of the distribution will cause many recent targets of selection to be missed. Thus, the hitchhiking mapping approach will work best when selection is rare, but not so rare that recently selected markers will fail to appear in a set of genomic fragments surveyed from the genome (Teshima *et al*., 2006; Thornton and Jensen, 2007). Either way, the hitchhiking mapping approach only has power to identify regions that have experienced very recent and very strong positive selection on new mutations.

A second concern with a purely empirical approach to hitchhiking mapping is that non-equilibrium demography (such as changes in the size and structure of populations over time) typically increases the variance of summary statistics based on diversity levels, the polymorphism frequency spectrum and linkage disequilibrium (Przeworski *et al*., 2001; McVean, 2002; Lazzaro and Clark, 2003; Haddrill *et al*., 2005a), highlighting the importance of using simulations to study the power and efficiency of this approach (e.g., Teshima *et al*., 2006). As a result, it is formally possible that most outlier loci in empirical distributions are not unusual when a plausible demographic scenario is fit to the data. For example, two recent analyses of *Drosophila* SNP data found that a recent, severe bottleneck is able to account for many features observed in non-African populations, and that no individual loci were found to be incompatible with the estimated bottleneck model, after accounting for multiple tests (Ometto *et al*., 2005; Thornton and Andolfatto, 2006). While these analyses illustrate that a demographic scenario can be found which explains empirical observations, they do not address whether a model of demography-plus-selection is a better explanation for the data than demography alone. Encouragingly, models incorporating both demography and selection are now beginning to be implemented (Li and Stephan, 2006) and it remains to be seen how robust these methods are to departures from model assumptions (such as misspecification of the demographic model, the selection model and intralocus recombination).

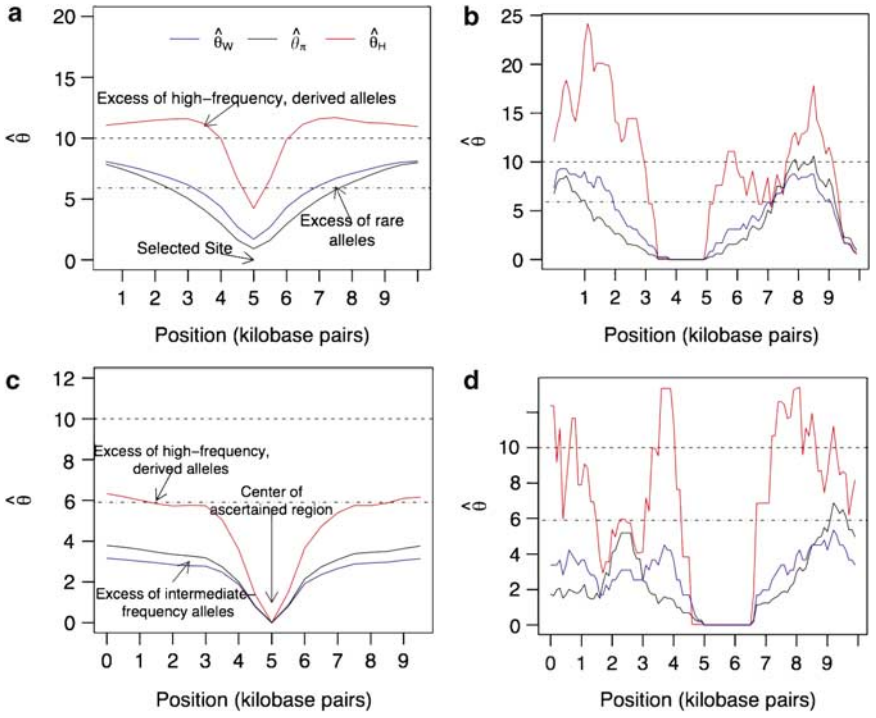

Figure 3 Patterns of variation under sweep and bottleneck models. The major challenge in distinguishing selection from demography is that demographic models can mimic the patterns of variability expected under a sweep model. This is particularly true of population bottlenecks; given that a region of low variation is observed, many properties of the genealogies of bottleneck and sweep models are identical (Barton 1998). In (**a**), the expected pattern of variability surrounding a recent selective sweep (see Figure 2) is shown for a 10 kb region. A beneficial mutation has fixed in the recent past ($\tau = 0$) at position 5000 in the middle of the region. The simulation parameters are for a sample size of 20, a population size of 10^6 and a selection coefficient of 0.001. The population recombination rate ($\rho = 4N_e r$) per site was 0.1 and the mutation rate ($\theta = 4N_e\mu$) per site was 0.01. For each position along the sequence, three estimators of θ were calculated: $\hat{\theta}_W$ (blue curve), $\hat{\theta}_\pi$ (black curve) and $\hat{\theta}_H$ (red curve), which depend on the number of mutations, the average frequencies of mutations and the number of high-frequency mutations in each window, respectively (see Table 1). The dashed line at $\hat{\theta} = 10$ represents the true value of θ for each window ($\theta = \frac{0.01}{bp} \times 1000bp = \frac{10}{region}$). The dot/dashed line at $\hat{\theta} \approx 6$ is the critical value of $\hat{\theta}_\pi$ at the 2.5% level for a large population of constant size. In other words, where the black curve is below this value in a given window, genetic diversity is significantly reduced below the expectation of the standard neutral model. This figure illustrates three important predictions of the selective sweep model. First, diversity is markedly reduced in the vicinity of the selected site, and recovers as a function of distance from the selected site. Second, the SFS is skewed toward rare alleles linked to the selected site ($\hat{\theta}_W > \hat{\theta}_\pi$, therefore Tajima's (1989) $D < 0$, see Braverman *et al*., 1995). Third, there is an excess of high-frequency-derived mutations surrounding the selected site ($\hat{\theta}_H > \hat{\theta}_\pi$, see Fay and Wu, 2000). Figure (**b**) shows a random realization of the selected sweep model simulated in (**a**). In this particular example, the selective sweep resulted a roughly 1.25 kb region of no variability. Left of the selected site, the excess of rare alleles and the 'splash' of high-frequency-derived alleles are particularly pronounced. Such signatures are similar to, for example, patterns of variation surrounding *Acp26Aa*, a putative target of a recent selective sweep in *D. melanogaster* (see Fay and Wu, 2000). The reader is referred to Figure 3 of Kim and Stephan (2002) for more examples of the pattern of spatial variability under sweep models. (**c**) and (**d**) Illustrate the effects that demography has on patterns of variability, when loci are not selected at random for sequencing. In hitchhiking mapping studies, regions identified as unusual (because they have no variability), may be selected for further investigation. Such follow-up investigation often consists of further sequencing surrounding the unusual region, followed by the application of a method to estimate the parameters of a selective sweep model (e.g., Harr *et al*., 2002; Bauer DuMont and Aquadro, 2005; Beisswanger *et al*., 2006; Pool *et al*., 2006). This ascertainment of regions results in patterns of variability mimicking what is expected from a recent selective sweep. (**c**) The expected pattern of variability in a 10 kb window surrounding a region from a bottlenecked population (simulated using parameters from Thornton and Andolfatto, 2006), for the case where the region was identified because of a 500 bp region with no variation centered at 5 kb. (**d**)A single-simulated replicate from the simulations used in the calculation of (**c**). The reason that ascertainment of invariant regions gives rise to such patterns under the bottleneck model is that the genealogies in the invariant region have reached their most recent common ancestors during the bottleneck, whereas lineages on either side have different histories (due to recombination), and reach common ancestors further back in the past. This effect of recombination is similar to what occurs during selective sweeps – lineages linked to the selected site reach their common ancestors during the sweep, whereas sites at a larger genetic distance 'escape' the sweep, and reach a common ancestor further in the past (see Figure 2). Of particular importance is that although the patterns of variability around regions of reduced variation are similar between models of selection and models of demography, they are not identical. Although both models predict that high-frequency-derived mutations are in excess, an excess of rare alleles surrounding the selected site is expected following a sweep (**a**), whereas an excess of intermediate-frequency alleles surrounds valleys of reduced variation in bottleneck models (**c**). We must note that these examples are merely illustrative, and certainly not exhaustive, given the large parameter space for both models.

Modeling concurrent selection and demography for the purposes of inference is technically challenging. Wright *et al*. (2005) have recently implemented a model-based approximation to this problem in order to map targets of artificial selection during maize domestication from the wild grass *teosinte*. They analyzed SNPs in 774 genes in an attempt to identify those which show evidence for a reduction in variation beyond that

expected under a bottleneck scenario. Their approach was to estimate parameters for two distinct 'bottlenecks' affecting patterns of variation in the maize genome. The first bottleneck is meant to model the reduction in population size known to be associated with maize domestication (Doebley, 2004). The second 'bottleneck' approximates the effects of recent selective sweeps on levels of variability (Simonsen *et al*., 1995) at some fraction of loci in the genome, associated with domestication. Using this approach to capture the effects of demography and selection simultaneously, they estimate that a significant fraction (2–4%) of genes in the maize genome are likely to have been targets of recent directional selection. Intriguingly, Wright *et al*. (2005) also show that candidate selected genes with putative functional roles in plant growth tend to cluster near quantitative trait loci that contribute to phenotypic differences between maize and *teosinte*. Also reassuring is that the co-estimated demographic bottleneck associated with domestication is a reasonable predictor of maize-*teosinte* differences in overall levels of diversity, the frequency spectrum and linkage disequilibrium.

An alternative approach in detecting selection involves the analysis of long-range haplotypes around a given locus of interest, known as core haplotypes (Sabeti *et al*., 2002). The age of each core haplotype is assessed by the decay of its association to alleles at varying distances from the locus, as measured by extended haplotype homozygosity (EHH). Core haplotypes with high EHH values and high population frequencies are taken to indicate the presence of a mutation that increased in frequency faster than expected under neutrality. Extending this approach to a genomic scale, Voight *et al*. (2006) modified the use of the EHH statistic such that the expectation and standard deviation of SNPs are estimated from the empirical distribution. Thus, their approach measures how unusual haplotypes around a given SNP are relative to the whole genome, and accounts for allele frequencies in the sample and variation in recombination rates. Applying this approach to the HapMap data set, they identify a number of genomic regions that may have experienced very recent directional selection in each of the three population groups sampled. It is worth noting that approaches based on linkage disequilibrium (and haplotype structure) are limited to detecting selection of a very specific nature (i.e., ongoing sweeps or recent balanced polymorphisms). Currently, rather little is known about the performance and robustness of this approach.

Post hoc analysis of outliers

While hitchhiking mapping approaches can quickly lead to a number of regions that may have experienced a recent selective sweep in the genome, the evidence is indirect because an explicit selection model is not being examined. Recent advances have been made in testing explicit selection models that can be used in conjunction with hitchhiking mapping approaches. One of the most widely applied of this second class of tests is the Kim and Stephan (2002) composite likelihood ratio test (CLRT). The CLRT uses the spatial pattern of polymorphism frequencies (the site frequency spectrum, hereafter SFS) to test for evidence of a selective sweep. This method uses the spatial pattern of variability in the SFS to estimate the location of the selective sweep and the magnitude of the selection coefficient. Kim and Nielsen (2004) explored various ways to extend this approach using patterns of linkage disequilibrium but found that it did not lead to substantial gains in power.

A limitation of both of these tests, however, is that they compare the standard, neutral model with a simplistic sweep model. As such, if the data happen to differ significantly from the predictions of the neutral equilibrium model (as might be expected under a number of demographic scenarios), the null model might be rejected in favor of selection – even if the likelihood of the selection model is not particularly high. Jensen *et al*. (2005) demonstrated that this test is sensitive to deviations from the assumptions of the standard neutral model, with both population substructure and bottlenecks leading to a high frequency of false-positive signals of selective sweeps.

Jensen *et al*. (2005) proposed a composite likelihood goodness-of-fit (GOF) test derived from the Kim and Stephan inference scheme. The GOF test is intended to reduce the false-positive rate of the CLRT in non-equilibrium populations. The GOF test is essentially a parametric bootstrap of the selection parameters inferred using Kim and Stephan's approach, and compares the sweep model to a generalized alternative model. The GOF statistic substantially reduces the rate of false inference of selection, when the true model is one of non-equilibrium demography. While the CLRT-GOF combination has been successfully applied to a number of data sets, with multiple loci showing evidence of positive selection (Table 2), these approaches still suffer from a number of limitations, and in particular with regard to how they are applied in practice. One of the many assumptions of all statistical tests to detect selection is that loci are randomly sampled from the genome. In practice, however, tests like the CLRT and GOF are sometimes applied to loci that have been previously identified as 'unusual' in a genome scan study, without accounting for how the locus was identified (e.g., see loci marked with an asterisk in Table 2). For example, loci may be chosen for further investigation because the initial genome scan identified regions of the genome with highly reduced variability (e.g., Glinka *et al*., 2003; Beisswanger *et al*., 2006; Ometto *et al*., 2005; Glinka *et al*., 2006). Failure to account for how loci are chosen leads to high false-positive rates – selection will be inferred from the data when the true model is selectively neutral (Thornton and Jensen, 2007). The reason for the high false-positive rate is that the ascertainment procedure identifies regions of the genome with 'sweep-like' patterns of variation, even under neutral models (Figure 3). Demographic departures, such as recent bottlenecks, increase the variance in patterns of diversity across the genome and the scale of linkage disequilibrium, making these spurious signatures of selection more common. Additionally, these methods may be sensitive to assumptions regarding mutation rates and rates of recombination, which are assumed to be constant across the given region.

Tests employing the background SFS

Nielsen *et al*. (2005) recently presented several tests aimed at detecting selective sweeps from genome-wide

Table 1 Common summary statistics for SNP data

Statistic	*Summarizes*	*Feature of data*	*Reference*
$\hat{\theta}_W$	Diversity	Number of mutations	Watterson (1975)
$\hat{\theta}_\pi$	Diversity	Intermediate-frequency mutations	Tajima (1989)
$\hat{\theta}_H$	Diversity	High-frequency-derived mutations	Fay and Wu (2000)
D	Site frequency spectrum	Proportional to $\hat{\theta}_\pi - \hat{\theta}_W$	Tajima (1989)
H	Site frequency spectrum	$\hat{\theta}_\pi - \hat{\theta}_H$	Fay and Wu (2000)

Abbreviations: SNP, single nucleotide polymorphism.
Single nucleotide polymorphism data are often summarized into single statistics, which emphasize a particular aspect of the data. There are three commonly used summaries that measure overall levels of variability in the sample. These measures emphasize mutations at different frequencies in the sample. $\hat{\theta}_W$ depends on the total number of mutations in the sample, and is strongly influenced by rare mutations. $\hat{\theta}_\pi$ depends on the average frequency of mutations in the sample, and $\hat{\theta}_H$ heavily weights high-frequency-derived mutations. Under a model of a large, panmictic population and the infinite-sites mutation model, these measures are all unbiased estimates of the population mutation rate $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per generation. Therefore, under the neutral model, the difference between any two of these estimators is expected to be approximately 0, which leads to the concept of using the difference as a test statistic of the 'standard' neutral model. Two widely used test statistics are Tajima (1989) D and Fay and Wu (2000) H statistics. When $D < 0$, there is an excess of rare alleles in the sample, as may be expected following a recent selective sweep (Braverman *et al*., 1995), in a growing population (Tajima, 1989), or in a population which experienced a bottleneck in the relatively distant past (e.g., Haddrill *et al*., 2005a, b). When $H < 0$, there is a relative excess of high-frequency-derived mutations, which may be due to a recent sweep (Fay and Wu, 2000), a recent severe bottleneck (e.g., Haddrill *et al*., 2005a, b) or hidden population structure (Przeworski, 2003).

SNP data. The methods are similar in form to Kim and Stephan in that they are based on a composite likelihood statistic calculated from the SFS, but they differ from previous methods in that the null hypothesis considered is not a specific population genetic model, but is derived from the background pattern of variation at a putatively neutral class of sites in the genome. They propose a new parametric test that has high power to detect recent, strong selective sweeps and is surprisingly robust to assumptions regarding recombination rates and demography (i.e., has low Type-I error). Similar to the Kim and Stephan approach, their parametric test also provides estimates of the location of the selective sweep(s) and the magnitude of the selection coefficient.

This method has improved upon many of the problems encountered by the previously proposed tests. It is robust to assumptions regarding recombination rate, it directly analyzes genomic scale data and thus does not depend on other approaches to pre-select candidate regions, it is computationally efficient enough to be applied on a genomic scale, and most importantly, it uses the background frequency instead of a standard, neutral model in order to define the test statistic. One limitation of this approach is that it depends on huge amounts of genomic information, of the sort that is currently only available in a limited number of organisms. This may change in the near future, as the cost associated with collecting genome-wide polymorphism decreases. A second limitation, as with all approaches based on using the background SFS, is the choice of sites in the genome used as a neutral reference. A growing body of evidence suggests that a large fraction of non-coding DNA in organisms with more streamlined genomes, like *Drosophila*, may be both selectively constrained and subject to recurrent positive selection (Bergman and Kreitman, 2001; Halligan *et al*., 2004; Kohn *et al*., 2004; Andolfatto, 2005; Haddrill *et al*., 2005b; Bachtrog and Andolfatto, 2006). These tests may be more appropriately applied to organisms with larger genome sizes, such as mammals, where constraint in non-coding DNA appears to be less pervasive in introns and intergenic regions (Shabalina *et al*., 2001; Keightley *et al*., 2005).

Conclusions and future directions

Power to detect selection in the genome

To date, the hitchhiking mapping approach is the most widely applied to identifying selective sweeps in *Drosophila* (Schlotterer, 2002; Glinka *et al*., 2003; Kauer *et al*., 2003; Ometto *et al*., 2005) and humans (Akey *et al*., 2002, 2004; Stajich and Hahn, 2005), with regions containing outlier loci then subjected to further resequencing and the application of *post hoc* tests for selection (see Table 2). While this approach is revealing potentially interesting candidate loci, the power of this approach depends on the details of both the demographic history of the species and on the model of adaptation (Teshima *et al*., 2006) and are prone to the problem of pre-ascertaining putative sweep regions (Figure 3). Further, although approaches using the 'background' SFS appear to be encouragingly robust to demographic assumptions, they do require that a null model be used in order to quantify the significance of a departure from the background site-frequency spectrum. Nielsen *et al*. (2005) show that using the SNM as the null is conservative for all demographic scenarios that they consider, and it will be of use to know if this is true in general.

Robustness to model assumptions

Many of the methods reviewed here are model based, and the issue of robustness to model assumptions is critical. The simplest model tests for a recent selective sweep in a population at demographic equilibrium (e.g., Tajima, 1989; Kim and Stephan, 2002). Although tests that are reasonably robust to demographic assumptions can be constructed (Jensen *et al*., 2005; Nielsen *et al*., 2005), the standard sweep model assumes that adaptation proceeds by selection acting immediately on new mutations. If selection acts on standing variation (which we may expect to be the case both for domesticated organisms and for populations which have recently moved into new habitats), then the effect of a selective sweep on patterns of linked, neutral variation is much weaker (Innan and Kim, 2004; Przeworski *et al*., 2005),

Table 2 Analysis of published data

Data set	*4Nr (per base)*	$\hat{\alpha}$	*CLR test* Λ_{KS} (P-*value*)	*GOF test* Λ_{GOF} (P-*value*)
Acp26A[a]	0.04	29.4	7.76 (0.033)	294.5 (0.159)
Duffy locus[b]	0.0015	90.9	8.84 (0.024)	260.7 (0.602)*
Janus/ocnus region[c]	0.02	109.6	16.60 (0.001)	1107 (0.017)
	0.065	446.6	16.61 (<0.001)	1107 (0.022)
	0.13	1009.4	16.61 (0.002)	1107 (0.029)
Jingwei gene[d]	0.034	25.3[e]	3.84 (0.120)[e]	N/A
*Sweep region 1[f]	0.005	129.4	14.00 (0.005)	675.7 (0.081)
	0.015	444.4	14.01 (0.003)	675.7 (0.110)
*Sweep region 2[f]	0.005	22.3	4.24 (0.145)	N/A
	0.015	81.1	4.24 (0.123)	N/A
AIM1 locus[g]	0.0004	220.9	14.17 (0.006)	560.2 (0.311)*
*Sweep region[h]	0.063	100	6.95 (0.045)	793.8 (0.889)*
*Downstream Notch region[i]	0.063	757	26.84 (<0.001)	1882.4 (0.114)
**Wapl* region[j]	0.005	2076	21.54 (0.02)	912 (0.87)*
**Unc*-119 region, Europe[k]	0.006	1710	NR (<0.0001)	NR (0.171)*
**Unc*-119 region, Africa[k]	0.006	2400	NR (0.03)	NR (0.326)*
MKK7[l]	NR	NR	NR (0.0025)	NR (0.24)*

Abbreviations: NR, not reported. N/A, GOF not performed because CLR test is not significant.
P-values are based on 1000 replicates of simulations under null models.
Application of the CLR and GOF tests to ten published polymorphism data sets which were argued to contain signatures of a recent selective sweep (Jensen *et al*., 2005). Two data sets (*janus/ocnus* region in *D. simulans* and *jingwei* gene in *D. teissieri*) show evidence of partial selective sweeps; in which case subsets of sampled chromosomes that exhibit strong evidence of linkage to the putative beneficial mutation (haplotype group I of *janus/ocnus* and intron-absent sequences of *jingwei*) were used. The resulting pattern of polymorphism due to hitchhiking in these subsets should be identical to that of a complete selective sweep (Meiklejohn *et al*., 2004). Of the ten data sets, two failed to reject neutrality ('sweep region 2' of Harr *et al*. (2002) and *jingwei* gene from Llopart *et al*., 2002). The eight remaining data sets that showed significantly large Λ_{KS} were subsequently analyzed using the GOF test. **P*-values that fall in a range consistent with a selective sweep, and are not consistent with the demographic models examined in Jensen *et al*. (2005). As the Kim and Stephan (2002) model is the null under this test, it is worth noting that low *P*-values are consistent with both selection and certain demographic models, whereas high *P*-values appear consistent only with the sweep model. Regions marked with an asterisk have an important caveat, in that they were detected as outliers through a genomic scan and have not taken this ascertainment in to account. Ascertainment bias of this sort has been shown to result in a high rate of spurious signals of selection (Thornton and Jensen, 2007 and see Figure 3).
[a]North Carolina population of *D. melanogaster* (Kim and Nielsen, 2004; Aguadé *et al*., 1992).
[b]Human duffy blood group locus from Hausa population (Hamblin *et al*., 2002).
[c]Haplotype group I sequences of *janus/ocnus* region sequences of *Drosophila simulans* (Meiklejohn *et al*., 2004), analyzed for three different rates of recombination (4*Nr*).
[d]Intron-absent sequences of *jingwei* gene in *Drosophila teissieri* (Llopart *et al*., 2002).
[e]Likelihood is calculated without ancestral-derived allele information (option 2 of Kim and Stephan, 2002).
[f]Sweep regions 1 and 2 of Harr *et al*. (2002) (*D. melanogaster*), each analyzed for two different rates of recombination (4*Nr*).
[g]Human AIM1 locus from mixed European population (Soejima *et al*., 2006).
[h]Sweep region located between the *white* and *kirre* genes in a Zimbabwe population of *D. melanogaster* (Pool *et al*., 2006).
[i]Sweep region downstream of the *Notch* locus in a US population of *D. melanogaster* (Bauer DuMont and Aquadro, 2005).
[j]Sweep in the *wapl* region in a Zimbabwe population of *D. melanogaster* (Beisswanger *et al*., 2006).
[k]*Unc-119* region in *D. melanogaster* (Glinka *et al*., 2006).
[l]Harr *et al*. (2006).

and hitchhiking mapping approaches will be less powerful (Teshima *et al*., 2006). Further, genome scans will have low power to detect selection if beneficial alleles are recessive (Teshima and Przeworski, 2006).

The rate of selective sweeps

The rate at which selective sweeps occur in the genome is a fundamentally important parameter that we have very little information about. An often under-stated assumption of hitchhiking mapping approaches is that positive selection occurs relatively rarely in the genome, though methods are often tested assuming that sweeps have occurred very recently (e.g., Fay and Wu, 2000; Kim and Stephan, 2002; Nielsen *et al*., 2005). If the rate of sweeps is low, there will simply be very few recent sweeps across the genome, and thus few sweeps that existing methods will have power to detect. If, on the other hand, the rate of sweeps is high, then both outlier detection and methods testing for differences from the background SFS would be comparing selected loci to one another, resulting in a great loss of power. Additionally, if the rate of recurrent hitchhiking is this great, there is an appreciable probability that sweeps are occurring on recently swept backgrounds. This multiple-sweep effect will result in very different patterns in the SFS, particularly with regard to high-frequency-derived alleles and thus linkage disequilibrium (Przeworski, 2002;

Kim and Nielsen, 2004; Kim, 2006) – patterns which existing methods rely heavily upon to detect selection.

Functional verification of putatively swept loci

Ideally, the detection of recent selection by statistical methods should eventually be verified with functional approaches. Although such experiments are technically challenging, recent results from genome scan studies are encouraging. For example, Wright *et al*. (2005) found that candidate loci for traits important to the domestication of maize were closely linked to their putatively swept regions, and Voight *et al*. (2006) study tended to identify genic regions as the targets of recent selection. In both model systems, such as *Drosophila* and in domesticated plants and animals, the ultimate verification of the methods discussed here can be tested by direct functional characterization of putatively swept regions via genetic manipulation of candidate loci (e.g., Greenberg *et al*., 2003). It will also be of interest to study whether or not loci underlying adaptive quantitative traits in natural populations (e.g., Colosimo *et al*., 2004) and in domesticated species (e.g., VanLaere *et al*., 2003) are reliably identified using existing methods for detecting selection in the genome.

Acknowledgements

We thank Doris Bachtrog and Molly Przeworski for comments on the article. JDJ was supported by National Science Foundation grant DMS-0201037 to R Durrett, CF Aquadro and R Nielsen. CB was supported by NIH GM72861 to M Przeworski. PA was supported by a Hellman Faculty Research Fellowship.

References

Aguadé M, Miyashita N, Langley CH (1992). Polymorphism and divergence in the *Mst26a* male accessory gland gene region. *Genetics* **132**: 755–777.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA *et al.* (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**: e286.

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

Andolfatto P (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149–1152.

Bachtrog D, Andolfatto P (2006). Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* **174**: 2045–2059.

Barton NH (1998). The effect of hitch-hiking on neutral genealogies. *Genet Res* **72**: 123–133.

Bauer DuMont V, Aquadro CF (2005). Multiple signatures of positive selection downstream of notch on the X chromosome in *Drosophila melanogaster*. *Genetics* **171**: 639–653.

Begun DJ, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D*. melanogaster. *Nature* **356**: 519–520.

Beisswanger S, Stephan W, De Lorenzo D (2006). Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* **172**: 265–274.

Bergman CM, Kreitman M (2001). Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**: 1335–1345.

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, Hartl DL (2003). Maximum likelihood method for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor Pop Biol* **63**: 91–103.

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D *et al*. (2004). The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol* **2**: E109.

David JR, Capy P (1988). Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet* **4**: 106–111.

Doebley J (2004). The genetics of maize evolution. *Annu Rev Genet* **38**: 37–59.

Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998). Investigation of the bottleneck leading to the domestication of maize. *Proc Natl Acad Sci USA* **95**: 4441–4446.

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

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

Glinka SL, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003). Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**: 1269–1278.

Greenberg AJ, Moran JR, Coyne JA, Wu C-I (2003). Ecological adaptation during incipient speciation revealed by precise gene replacement. *Science* **302**: 1754–1757.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005b). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol* **6**: R67.

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

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD (2004). Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res* **14**: 273–279.

Hamblin MT, Thompson EE, Di Rienzo A (2002). Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* **70**: 369–383.

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

Harr B, Voolstra C, Heinen TJAJ, Baines JF, Rottscheidt R, Ihle S *et al.* (2006). A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse Mus musculus domesticus. *J Evol Biol* **19**: 1486.

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

Ihle S, Ravaoarimanana I, Thomas M, Tautz D (2006). An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol Biol Evol* **23**: 790–797. E-pub 18 January 2006.

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

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

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

Kauer MO, Dieringer D, Schlotterer C (2003). A microsatellite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*. *Genetics* **165**: 1137–1148.

Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005). Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res* **15**: 1373–1378.
Kim Y (2006). Allele frequency distribution under recurrent selective sweeps. *Genetics* **172**: 1967–1978.
Kim Y, Nielsen R (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1513–1524.
Kim Y, Stephan W (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765–777.
Kohn MH, Fang S, Wu CI (2004). Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. *Mol Biol Evol* **21**: 374–383.
Lachaise D, Cariou M, David JR, Lemeunier F, Tsacas L, Ashburner M (1988). Historical biogeography of the *Drosophila melanogaster* species subgroup. In: Hecht MK, Wallace B, Prance GT (eds). *Evolutionary Biology*, vol. 22. Plenum Press: New York, pp. 159–225.
Lazzaro BP, Clark AG (2003). Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. *Mol Biol Evol* **20**: 914–923.
Li H, Stephan W (2006). Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2**: e166.
Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002). Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad Sci USA* **99**: 8121–8126.
Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favorable gene. *Genet Res* **23**: 23–35.
McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **354**: 114–116.
McVean GA (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* **162**: 987–991.
Meiklejohn CD, Kim Y, Hartl DL, Parsch J (2004). Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* **168**: 265–279.
Nielsen R (2005). Molecular signatures of natural selection. *Ann Rev Gen* **39**: 197–218.
Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD (2005). Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566–1575.
Ometto L, Glinka S, De Lorenzo D, Stephan W (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol* **22**: 2119–2130.
Parsch J, Meiklejohn CD, Hartl DL (2001). Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. *Genetics* **159**: 647–657.
Payseur BA, Nachman MW (2002). Gene density and human nucleotide polymorphism. *Mol Biol Evol* **19**: 336–340.
Pollinger JP, Bustamante CD, Fledel-Alon A, Schmutz S, Gray MM, Wayne RK (2005). Selective sweep mapping of genes with large phenotypic effects. *Genome Res* **15**: 1809–1819.
Pool JE, Bauer DuMont V, Mueller JL, Aquadro CF (2006). A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. *Genetics* **172**: 1093–1105.
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 Int J Org Evolution* **59**: 2312–2323.
Przeworski M, Wall JD, Andolfatto P (2001). Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol* **18**: 291–298.
Reusch TB, Wegner KM, Kalbe M (2001). Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. *Mol Evol* **10**: 2435–2445.
Rosenberg NA, Nordborg M (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**: 380–390.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.
Schlotterer C (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**: 753–763.
Schlotterer C (2003). Hitchhiking mapping – functional genomics from the population genetics perspective. *Trends Genet* **19**: 32–38.
Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS (2001). Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* **17**: 373–376.
Simonsen KL, Churchill GA, Aquadro CF (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**: 413–429.
Soejima M, Tachida H, Ishida T, Sano A, Koda Y (2006). Evidence for recent positive selection at the Human *AIM1* locus in a European population. *Mol Biol Evol* **23**: 179–188.
Stajich JE, Hahn MW (2005). Disentangling the effects of demography and selection in human history. *Mol Biol Evol* **22**: 63–73.
Stephan W, Wiehe THE, Lenz MW (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor Popul Biol* **41**: 237–254.
Tajima F (1989). Statistical method for testing the neutral mutation hypothesis. *Genetics* **123**: 437–460.
Tajima F (1989b). The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.
Teshima KM, Coop G, Przeworski M (2006). How reliable are genomic scans for selective sweeps? *Genome Res* **16**: 702–712.
Teshima KM, Przeworski M (2006). Direction positive selection on an allele of arbitrary dominance. *Genetics* **172**: 713–718.
The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**: 789–796.
Thornton KR, Andolfatto P (2006). Approximate Bayesian inference reveals evidence for a recent, sever, bottleneck in non-African populations of *Drosophila melanogaster*. *Genetics* **172**: 1607–1619.
Thornton KR, Jensen JD (2007). Controlling the false positive rate in multilocus genome scans for selection. *Genetics* **175**: 737–750.
Tishkoff SA, Verrelli BC (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* **4**: 293–340. (Review).
VanLaere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L *et al.* (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**: 832–836.
Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006). A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.
Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.
Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Pro Natl Acad Sci USA* **102**: 7882–7887.
Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD *et al.* (2005). The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.
Zhu L, Bustamante CD (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**: 1411–1421.