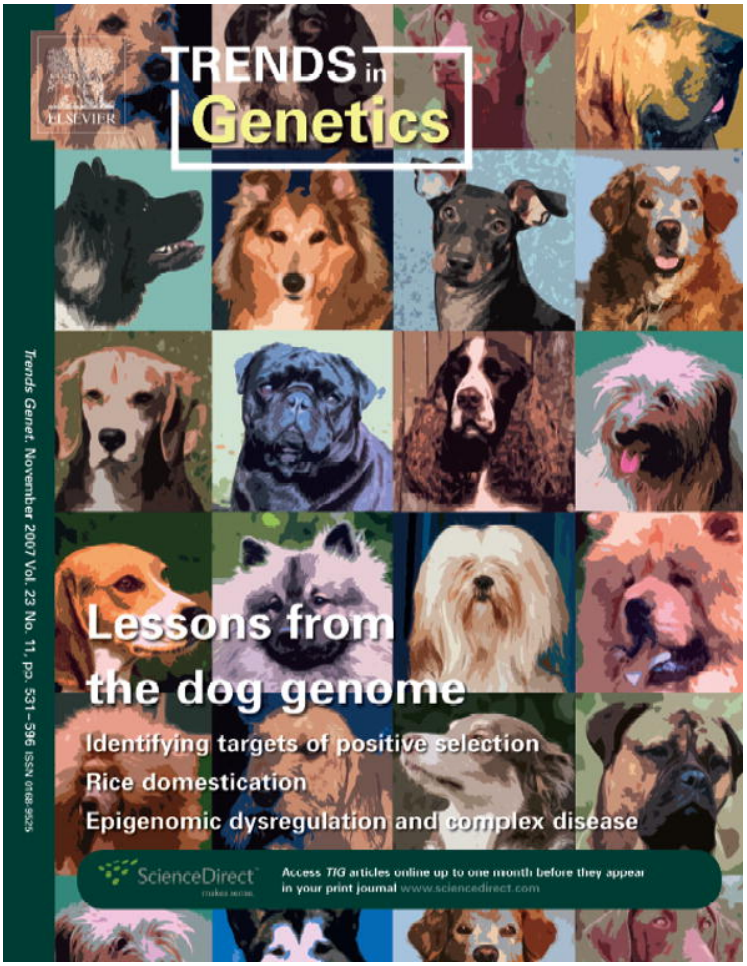

Review TRENDS in Genetics Vol.23 No.11

Approaches for identifying targets of positive selection

Jeffrey D. Jensen[1,2], Alex Wong[1] and Charles F. Aquadro[1]

[1] Department of Molecular Biology and Genetics, Biotechnology Building, Cornell University, Ithaca, NY 14853, USA
[2] Section of Ecology, Behavior and Evolution, University of California San Diego, La Jolla, CA 92037, USA

Despite significant advancements in both empirical and theoretical population genetics throughout the past century, fundamental questions about the evolutionary forces that shape genomic diversity remain unresolved. Perhaps foremost among these are the strength and frequency of adaptive evolution. To quantify these parameters, statistical tools are needed that are capable of effectively identifying targets of positive selection throughout the genome in an unbiased manner, and functional approaches are needed that are capable of connecting these identified genotypes with the resulting adaptively significant phenotypes. Here we review recent advancements in both statistical and empirical methodology, and discuss important challenges and opportunities that remain as researchers continue to uncouple the relative importance of stochastic and deterministic factors in the evolution of natural populations.

Understanding the molecular genetic basis of adaptive evolution

A central goal of evolutionary biology is to understand the process of adaptation. How do species respond to factors such as climatic change, predation and pathogens, or to strong artificial selection induced through domestication? Molecular methods have allowed researchers to begin to identify specific genes that underlie adaptive morphological variation. Striking examples include pelage color in beach mice [1], size and shape of beaks in Galapagos finches [2] and plant and fruit architecture in maize [3]. Although powerful, this candidate adaptive phenotype approach is limited to identified adaptations, and might provide a biased view of the underlying types of genes and variation favored by positive selection. Recent developments in the analysis of intraspecific polymorphism and interspecific divergence increasingly are making an alternative approach to the study of adaptation possible: namely, screening the genome for 'footprints' of adaptation (e.g. [4,5]), followed by an identification of the affected phenotype, and finally functional verification of the predicted effect of the variation. Because of the inherent importance of these questions in evolutionary biology and the recent availability of large amounts of sequence data, several recent reviews have dealt with methods for detecting positive selection (e.g. [6–10]). Our aim here is somewhat different, however, because we do not deal with statistical and functional methods independently, but rather focus on how these separate classes of approaches can be combined to move from genomes worth of sequence data to a specific list of functionally verified mutations that

Glossary

Empirical Bayes procedure: a statistical approach whereby codons under positive selection are identified on the basis of maximum likelihood estimates of the parameters of interest.
Equilibrium population: a panmictic population of continuously constant size.
Heterozygosity: a measure of the variability of a population, often the fraction of individuals in a population that are heterozygous for a particular locus.
Haplotype: a set of polymorphisms that are statistically associated.
High frequency derived alleles: polymorphisms that occupy the upper bounds of the site frequency spectrum, and are of a type differing from the ancestral state.
Hitchhiking effect: resulting from a selective sweep, which affects linked neutral variation in a characteristic way through the fixation of the selected haplotype.
Incomplete sweep: an advantageous mutation that is sweeping through the population, but has not yet been fixed.
Linkage disequilibrium: a nonrandom association between two or more loci.
Maximum likelihood estimates (MLEs): parameter values that maximize the likelihood (probability) of the observed data under a particular model.
Model-based comparison: often in the form of a likelihood ratio test; a statistical test of the goodness-of-fit between two models.
Nonequilibrium population: a population in which the equilibrium assumptions of panmixia and/or constant size are violated.
Nonsynonymous mutation: in a coding sequence, a nucleotide change that alters the amino acid encoded for by a codon.
Omega (ω): the ratio between the rate of nonsynonymous and synonymous substitution; $\omega > 1$ is evidence for positive selection assuming that synonymous mutations do not impact fitness.
Parametric bootstrap: a parametric model is fit to the data, and then samples are drawn from this model in order to quantify confidence in a particular estimated parameter.
Partially sequenced regions: a common approach in empirical population genetics, in which polymorphism data is generated for a series of small fragments spanning a large genomic region.
Population bottleneck: a temporary reduction in the census size of a population.
Rare alleles: polymorphisms that occupy the lower bounds of the site frequency spectrum, often as singletons.
Recurrent sweeps: a model in which sweeps are considered to occur at a given rate, rather than at a fixed time in the past.
Segregating variant: often termed a polymorphism; a mutation that is not fixed in the population.
Selective sweep: the process by which a selectively advantageous mutation is increased in frequency, which results in a hitchhiking effect.
Site-directed mutagenesis: introduction of one or more mutations at specified positions in a gene of interest.
Standard neutral model: a common null model, in which all mutations are assumed to be selectively neutral, and the population is at equilibrium.
Substitution: a mutation that becomes fixed in a species, in contrast to a segregating variant or polymorphism.
Synonymous mutation: in a coding sequence, a nucleotide change that does not alter the amino acid encoded for by a codon.
Targeted gene replacement: substitution of one allele for another in a genome, at the same genomic location; allows for comparisons between alleles in a homogeneous genetic background without position effects.
Unfolded site frequency spectrum: the observed count of derived mutations in each of the n-1 frequency classes, where n is the number of alleles sampled.

Corresponding author: Aquadro, C.F. (cfa1@cornell.edu).
Available online 23 October 2007.

 0168-9525/$ – see front matter doi:10.1016/j.tig.2007.08.009

have been important in the adaptive history of a given population. Thus, we first review several statistical methods for localizing positive selection from intra- and interspecific DNA sequence data and then we discuss examples of the types of functional studies available to test these inferences. We do not review those more general tests of an equilibrium neutral model that could indicate selection but do not identify the target sites of adaptation (Nielsen [6] provides an overview of these approaches).

Broadly speaking, statistical inferences of selection can use polymorphism data, divergence data, or a combination of both. Polymorphism based methods involve sampling multiple copies of an orthologous genomic region within populations, often together with a copy from a closely related species to define the ancestral and derived states for variation. The unfolded site frequency spectrum (see Glossary) is then evaluated across the sampled region to identify patterns consistent with recent or historical positive selection. This class of tests is appropriate for detecting single, recent selective sweeps, and is primarily concerned with the effects of selection on linked neutral variation (the so-called 'hitchhiking effect').

An additional important, and often unstated, assumption of these methods is that the rate of sweeps (see 'recurrent sweeps' in the glossary) is great enough to result in recent (and thus detectable) selection in the genome at the time of sampling, but not so great as to obscure detectable patterns of variation owing to overlapping, or even competing, selective sweeps. Several methods have been proposed for quantifying the rate and strength of sweeps using polymorphism and divergence data (e.g. [11–13]), the observed correlation between nucleotide diversity and recombination rate (e.g. [14–18]), and multilocus resequencing data (e.g. [19]). These analyses are relatively consistent in predicting a moderate prevalence of adaptive fixations across the genome, particularly for *Drosophila*. Although these studies will not be discussed here as they generally do not make specific site predictions, recurrent sweep models raise the important point that the number of detectable sweeps in a genome at a given point in time is a function of the underlying distribution of rates and selection coefficients. If selection is on average strong and rare, there will not be many sweeps recent enough to be detectable based on patterns of polymorphism (i.e. less than 0.1 $4N$ generations since fixation) [20]; whereas if they are on average weak and frequent, only small genomic regions will be affected by each sweep and thus might be missed by genomic scans.

Divergence-based approaches often involve sampling single individuals from each of a wide range of species, and then testing for sites that have changed more often than expected across the species tree. Although divergence-based tests do not have such a rigid time restriction associated with power compared with polymorphism-based methods, they frequently only detect recurrent selection and as such will miss single adaptive events in one or a few of the sampled species. Additionally, divergence-based tests rely on evidence of positive selection at the target itself. Historically, these tests have been almost exclusively applicable to coding regions, and often rely on knowledge of the underlying species phylogeny.

Both polymorphism and divergence approaches generate specific predictions that relate to the target of adaptation and are amenable to functional testing by available genetic, molecular and biochemical tools. There are only a limited number of studies that have connected statistical inference and functional verification. However, with recent advances in both computational and genetic methodologies, the number of such analyses will increase, which will provide a more unbiased view of the molecular genetic basis for adaptation.

Polymorphism-based methods

Mutation, drift, selection and population history interact to shape and maintain levels of variation, such that the contribution of any one factor can be difficult to determine. Model-based approaches are commonly used to tease apart their relative effects. The most commonly employed model assumes that all observed variation is selectively neutral and the population is at equilibrium, generally referred to as the standard neutral model (for a review, see Ref. [21]). Under the long list of assumptions for this model, the site frequency spectrum is well described and, thus, so are the commonly employed tests of neutrality and selection.

The second model considered here is the 'hitchhiking' model, in which a single, new, beneficial mutation occurs in an equilibrium population and is swept by selection to fixation, which leaves a 'footprint' of the selective sweep in the pattern of polymorphisms that flank the target of selection. Others have investigated extensions of this model, which include selection that favors a segregating variant (e.g. [22–25]), incomplete sweeps (e.g. [26–27]), sweeps in nonequilibrium populations (e.g. [28–29]), and recurrent sweeps in a given region of the genome (e.g. [18,30]).

There are also a variety of neutral, nonequilibrium models that will be relevant to varying degrees for any given population. The problem of nonequilibrium models that replicate patterns of variation associated with the hitchhiking model has been well described (e.g. [10,20,31–35]). Population bottlenecks in particular are capable of replicating many aspects of the site frequency spectrum that are commonly associated with positive selection (e.g. for a helpful visual depiction see Figure 3 of Ref. [10]). Progress has been made in estimating and incorporating these demographic parameters into the evaluated models to achieve more appropriate comparisons (i.e. the estimated nonequilibrium model under neutrality versus the estimated nonequilibrium model in which some proportion of loci are under positive selection; [29,35]).

Several recent tests have sought to capitalize on population data not only to identify genomic regions recently shaped by positive selection, but also to identify the specific sites under selection. The fixation of a beneficial mutation will alter patterns of polymorphism at linked neutral loci [36–38]. From this basic premise, more specific predictions have been proposed: these include a depression of heterozygosity relative to divergence [39] and an excess of rare alleles near the target of selection, because all new mutations that arise 'post-sweep' will be present initially as singletons [31,40–42]. However, in recombining regions, these patterns will not extend

indefinitely across a chromosome. Moving away from the target, recombination can rescue ancestral polymorphism, which creates an excess of high frequency derived alleles in the flanking regions [43]. Because of this extended haplotype structure, a selective sweep will create strong linkage disequilibrium (LD) in flanking regions but with little or no LD that spans the target of a recent fixation [44–46].

Exploiting patterns in the frequency distribution, Kim and Stephan [47] proposed an explicit test of the selection model that also generates maximum likelihood estimates of the strength and target of selection. By assuming that sites are independent, they calculated the probability of observing the given number of derived alleles at each site under both the equilibrium neutral and the equilibrium sweep models, then took the product of these probabilities under each model, and compared them using a likelihood ratio test. This method has reasonable power to localize the target of a recent sweep, although this power decays rapidly as the time since fixation grows owing to the accumulation of new mutations that obscure predicted patterns [47], consistent with other analyses (e.g. [20]).

Although this test is commonly applied to partially sequenced regions owing to the relatively coarse scale of localization from genome scan studies, the confidence interval on the estimate of the location of the target is large when the target has not been sequenced (Figure 1). Jensen and co-workers (unpublished data) have shown that a parametric bootstrap of the estimated selection parameters, and thus confidence intervals of the predicted target location, can be obtained from the null distribution of the goodness-of-fit test [34] to improve upon the performance of the approach by Kim and Stephan [47]. For an example of the construction of these confidence intervals, see Glinka *et al.* [48].

Thus, polymorphism data can be valuable for localizing targets of recent adaptation, because there are several specific predictions concerning the effects of a selective sweep on the site frequency spectrum and on spatial patterns of variation. However, a number of caveats are worth mentioning that are associated with polymorphism-based methods, and their consequences vary between organisms owing to differing demographic histories, levels of variation, extent of linkage disequilibrium and rates of crossing-over (for more details see Box 1).

Divergence-based methods

The past several years have seen a proliferation of methods which use sequence data from multiple species to detect and localize positive selection [49–56]. Such methods have been used to infer selection on single genes, and, more recently, on a genome-wide scale (e.g. [57]). As more genomes of closely related organisms are fully sequenced, we envisage methods that use interspecific comparisons to detect positive selection will become increasingly popular.

Divergence-based methods typically assume that some specified class of sites evolves under little or no functional constraint. Positive selection is inferred if a particular class of sites evolves significantly more rapidly than the unconstrained class. For coding sequences, comparisons between the number or rate of nonsynonymous

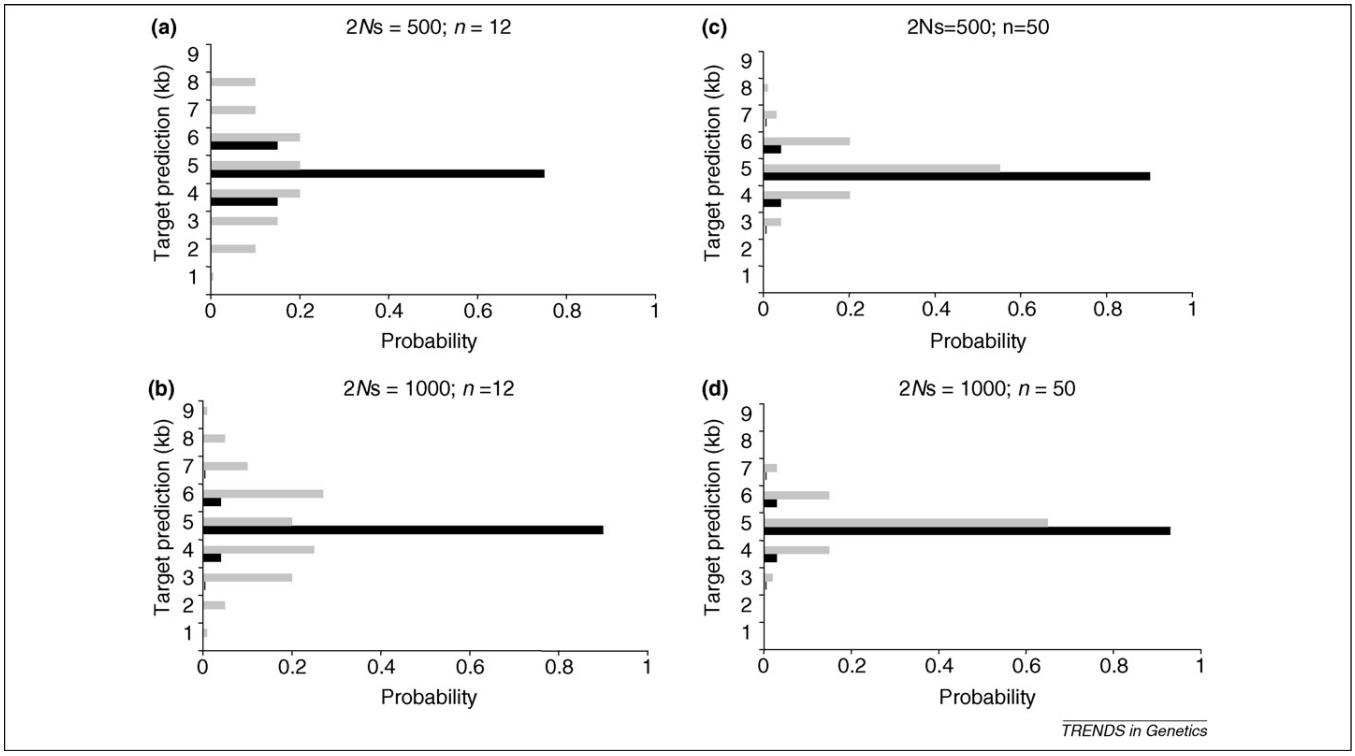

Figure 1. Statistical target prediction. Probability of the target of selection (X) being placed within a 1 kb window centered around positions 1 kb, 2 kb, 3 kb, etc., for two large values of $2Ns$ (1000 and 500) and **(a,b)** common ($n = 12$) and **(c,d)** large ($n = 50$) sample sizes, for a 10 kb region in which 5 kb has been sequenced in five, evenly spaced, 1 kb fragments. In black are the partial datasets in which the true target has been sequenced, and in gray the partial datasets in which it has not been sampled. In all cases $4Nr = 1000$, $\theta = 75$ and $X = 5.0$ kb. Although target prediction is relatively accurate for partial sequencing when the site of selection has been sampled, the predictions are nearly uniform when it has not been sampled. This result suggests that complete sequencing is much more efficient.

Box 1. Empirical examples of polymorphism-based approaches to the localization of selective sweeps

Owing to severe bottlenecks associated with domestication, as well as very strong and recent artificial selection, crop plants tend to occupy the lower bounds of the variation spectrum. For example, a Simple Sequence Repeat (SSR) -based genome-wide diversity screen in *Sorghum bicolor* was recently followed up with direct sequencing, to investigate regions with unusual, sweep-like genealogies [70]. Consistent with the screen, evidence of positive selection was found around the SSR-locus *Xcup15*. However, Sorghum tends to have extremely low diversity overall (perhaps one-quarter of the amount of variation observed in landraces of maize [71,72]), and thus the power to localize the target of selection based on patterns of linked neutral variation is diminished. As such, their statistical target prediction has a very large confidence interval (on the order of 40 kb), spanning several genes. Interestingly, only a single fixed change was observed between wild and cultivated sorghum in the segments of this 40 kb window that were sequenced. Particularly tantalizing is the fact that this variant lies in the Protein Phosphatase 2C (*PP2C*) gene, which belongs to a gene family implicated in abscisic acid signal transduction, the regulation of flower development [73], and seed germination [74]. There are many such informative examples to be found in plant population genetics. Studies in maize have been particularly fruitful, and are reviewed by Doebley *et al.* [75,76].

Several studies have been conducted in *Drosophila melanogaster*, and the relatively higher levels of variation in *Drosophila* compared with humans and many domesticated plants and animals can provide increased power to localize targets of selection. For example, Beisswanger *et al.* [77] recently performed a follow up analysis based on a genome scan presented by Glinka *et al.* [78] around the *wapl* region, by sequencing roughly 6 kb of data distributed as 12 fragments across a 110 kb region. The confidence intervals are estimated to span ~68 kb (Jensen and co-workers, unpublished data). Additionally, using the parametric bootstrap simulations, groupings of target predictions are observed where sequence data exists, which suggests that sparse data of this sort introduces a bias in the localization procedure (Jensen *et al.*, unpublished data). Given the high level of variation in this organism, confidence intervals of this size make the identification of a small number of testable sites extremely difficult. More complete sequencing, however, can lead to more precise target predictions. For example, Bauer DuMont and Aquadro [79] examined a putatively swept region downstream of the *Notch* locus. Contiguous sequence was generated for 15 chromosomes for a 10.5 kb region, with resulting confidence intervals of the predicted target spanning 3 kb. Given the limited number of fixed ancestral versus derived changes in this smaller window, they searched for transcriptional binding sites and identified a fixed-derived substitution within the active site of a putative *Caudal* transcription factor binding domain that is predicted only in the non-African populations (i.e. those in which the strongest evidence of a sweep is observed) (Figure I).

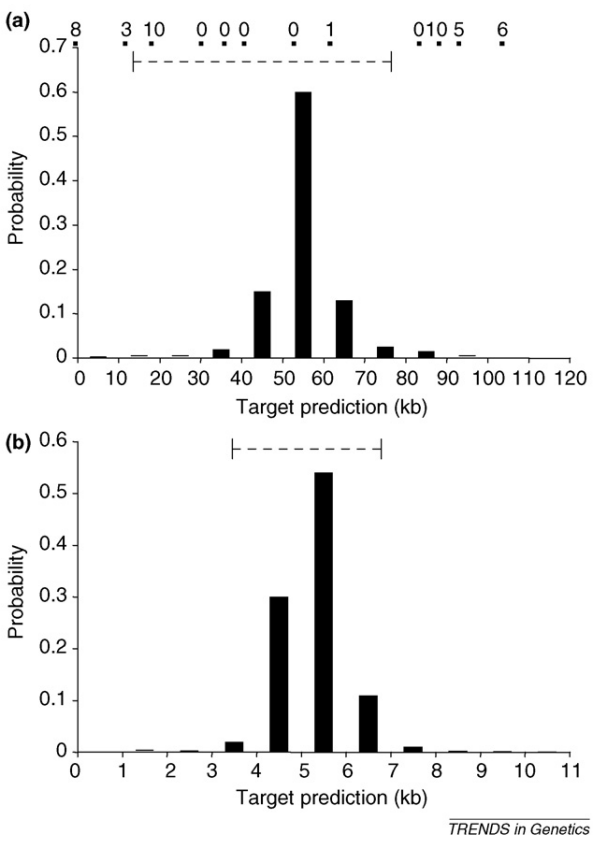

Figure I. Empirical examples of target estimation. **(a)** The probability distribution of location of the maximum likelihood estimate (MLE) of the target of selection (X) for 1000 selection simulations obtained through parametric bootstrap using the parameters specified for the *D. melanogaster* Zimbabwe data (the population showing the strongest evidence of a sweep) from Beisswanger *et al.* [77] The y-axis represents the probability that the target site prediction is placed within the location window given on the x-axis. The sweep was simulated as although it had just ended ($\tau = 0$), a value that should allow the test to perform much better than using their inferred ancestral sweep value. The broken lines indicate the 95% confidence interval on their estimate of the target ($X = 49.8$ kb). The positions of their sequenced fragments are indicated with black bars, and the number of segregating sites observed in each region are given above the x-axis. **(b)** The probability distribution of the MLE of X for 1000 selection simulations obtained through parametric bootstrap using the parameters specified from the *D. melanogaster Notch* region analysis of Bauer DuMont and Aquadro [79]. The broken lines indicate the 95% confidence interval on their estimate of the target ($X = 5.426$ kb). Note that the entire 10.5 kb region was completely sequenced in all lines ($n = 15$) for this USA population sample, and 147 segregating sites were observed. Note the difference in scale between panels (a) and (b). Combined with the simulation results of Figure 1, these results strongly suggest that the complete sequencing of putatively swept regions is far more accurate and efficient than partial sequencing, despite the preliminary savings in both time and cost associated with the latter.

substitutions (d_N) and the number or rate of synonymous substitutions (d_S) provide a good measure of the strength and character of selection. A d_N:d_S ratio (henceforth ω) of <1 is indicative of functional constraint; $\omega = 1$ is expected in the absence of constraint (as in pseudogenes) and; $\omega > 1$ is usually interpreted as evidence for positive selection on amino acid changes. Other, conceptually similar approaches have been extended to different kinds of data. For example, Wong *et al.* [51] have developed a maximum likelihood test to evaluate rates of change between codons of amino acids with different physico-chemical properties.

Many proteins are expected to be subject to strong purifying selection to preserve core function or structure. Thus, the examination of ω over an entire gene is extremely conservative, and unlikely to detect positive selection acting on a few sites. For proteins with known domains or active sites, the assumption of a single ω can be relaxed by grouping together different sites *a priori* on the basis of function. Using such an approach, Hughes and Nei [58] showed that residues in the antigen binding cleft of class I major histocompatability complex (MHC) molecules are subject to positive selection, whereas the rest of the protein

is relatively conserved (see [59] for another example). Recent methods have allowed variation in ω between sites without requiring *a priori* specification of which sites evolve under which ω class. Methods for estimating the distribution of ω in the presence of site-to-site variation have been proposed in counting-based (e.g. [50,60]) and model-based frameworks [49,52,55]. Here we focus on model-based methods, particularly maximum likelihood approaches, because these are used most frequently and have sufficient power to be of use to the small to medium sized datasets that typify most current applications.

In maximum likelihood methods, such as the popular computer package Phylogenetic Analysis by Maximum Likelihood (PAML) of Yang and colleagues (e.g. [49]), two distinct inferential steps are required to identify sites subject to positive selection. In the first step, it must be shown that a given alignment contains any sites likely to be under positive selection, regardless of the precise identity of those sites. This is accomplished by means of model comparison: an inference of selection requires that a model that includes selection performs better than does a model that does not allow selection (the null model). This null model allows different sites to have different values of ω, but does not permit $\omega > 1$ (i.e. positive selection is disallowed). The alternative model adds a class of sites with $\omega > 1$, allowing positive selection. The two models are compared by a likelihood ratio test. If the data fit the alternative model better than they do the null, then the action of positive selection can be inferred. In most implementations of these models, the same codons belong to the positively selected class in all lineages. As such, positive selection will only be inferred if it acts on the same set of codons on many or all branches of the phylogeny.

The second inferential step involves the localization of positively selected sites. The problem can be stated as follows: given the alternative model, which includes a class of sites with $\omega > 1$, which sites are likely to belong to the selected class? For each site, the posterior probability can be calculated under each ω class in the maximum likelihood model. A high posterior probability under the $\omega > 1$ class is suggestive of positive selection at that site. Simplistically, sites that have undergone a particularly high rate of amino acid change (relative to the background synonymous sites) will be identified as the most likely targets of selection. Many studies have used PAML and related applications to detect and localize sites subject to positive selection. In several cases, residues inferred as targets of recurrent positive selection fall into protein domains already suspected of undergoing adaptive evolution. Swanson *et al.* [61], for example, validated this method by using human class I MHC molecules. They found evidence for positive selection at several sites surrounding the antigen binding cleft, but not elsewhere in the protein.

Although divergence-based methods show great promise for detecting site specific acceleration, several important questions and caveats remain that must be taken into account when applying these approaches (Box 2).

The fact that a major proportion of most (eukaryotic) genomes is noncoding, yet contains important regulatory sequences (not to mention noncoding RNA genes), motivates a desire to be able to detect the targets of adaptation here as well as at protein-coding genes. This remains an important challenge, however, in part owing to the fact that simple classes of noncoding sites (analogous to synonymous versus nonsyonymous coding sites) are not known. Wong and Nielsen [53] have developed a method to infer and localize positive selection on noncoding regions by comparing noncoding divergence to synonymous site divergence, but the method has not yet been widely applied. Andolfatto [62] has, however, recently argued for a high frequency of adaptive noncoding fixations between species of *Drosophila* based on a contrast of intraspecific polymorphism versus interspecific divergence at noncoding versus synonymous sites in *D. melanogaster*. His approach does not make specific predictions, however, as to the precise noncoding sites that are the targets of positive selection.

Box 2. Caveats of divergence-based approaches

The Bayesian method used by PAML to localize selection to particular sites is an 'empirical Bayes' approach, because the parameters used to calculate the posterior probabilities are estimated by maximum likelihood. Early implementations of the empirical Bayes method in PAML did not take uncertainty in the maximum likelihood estimates of these parameters into account, and is hence referred to as a 'naïve empirical Bayes' (NEB) method. Recent approaches address uncertainty to varying degrees. PAML now uses a 'Bayes Empirical Bayes' (BEB) approach to the identification of positively selected sites, which does address uncertainty in the maximum likelihood estimate (MLE) of ω [54]. MrBayes goes further by adopting a fully Bayesian approach to the identification of sites under positive selection, with uncertainty of all parameter estimates (e.g. tree topology, rate estimates, branch lengths) taken into account. Comparisons between the NEB, BEB and fully Bayesian methods for identifying selected sites suggest that, in some cases (particularly where MLEs are extreme), incorporating uncertainty in parameter estimates is vital for accuracy [52,54].

Care must also be taken to choose an appropriate model of sequence evolution, as violations of the model assumptions can lead to erroneous inferences of positive selection and/or misidentification of selected sites. For example, the models implemented in PAML allow variation in ω, but assume a single synonymous substitution rate for all sites across the gene. Although in some cases this assumption might be warranted, there is clear evidence for synonymous rate variation in several datasets [80]. Failure to incorporate variation in d_S might result in mistakenly inferring positive selection at sites with individually low d_S but an average d_N [80]. Another important assumption that is sometimes overlooked is that a single tree underlies all sites in the alignment. This assumption should hold for 'well behaved' species, but might be violated in population samples, in organisms where horizontal transfer occurs, or in species affected by ancestral lineage sorting (e.g. [81–82]). Moderate to high levels of recombination can result in an unacceptably high rate of false positives in the inference of selection [83], although the empirical Bayes procedure appears to be less sensitive to the effects of recombination [83]. Problems associated with recombination might be mitigated by explicitly modeling recombination within the sample [56]. Fully Bayesian methods can also represent an improvement by incorporating uncertainty in the tree topology, although the impact of recombination on such methods has not been investigated.

Finally, as divergence based methods are applied to fully sequenced genomes, issues of data quality and alignment confidence will gain importance. No studies have explicitly addressed the impact of sequencing errors or inaccurate alignments on inferences of positive selection. However, it seems likely that false positives are a risk, especially for problematic alignments.

Functionally verifying predicted targets of positive selection

A statistical inference of positive selection indicates, at best, that selection has occurred in a genomic region, but does not indicate why. In cases where a gene (or noncoding region) under selection has a known function, a plausible hypothesis concerning the possible phenotypic consequences of selection can sometimes be constructed, but sometimes even this much is not possible. Functional studies are required to determine the biochemical, physiological, and fitness consequences of a putatively selected change.

Assessment of the functional consequences of evolutionary change is fraught with both conceptual and technical difficulties. In some cases, it will be particularly difficult to determine which phenotype or phenotypes to examine, and under what ecological conditions they should be examined. A single selective event, such as one that is uncovered in a genome scan, might have occurred in response to specific historical conditions that are unknown to the investigator. Although the nature of the genomic region targeted by selection might give clues as to potential phenotypes of interest, this need not be the case. In other examples, prior knowledge (or sheer luck) can enable the formation of a reasonable functional hypothesis that can be tested in a laboratory setting. There are several *in vitro* and *in vivo* methods used to address the functional consequences of evolutionary change.

By '*in vitro*' methods, we refer to the functional characterization of different forms of a gene, for example, alleles from within a species, from different species, or mutagenized alleles, outside of the context of a whole organism. Recombinant proteins can be generated in bacteria, in cultured eukaryotic cells, or in a cell free system, and assayed for activity on specified substrates. Such methods are one of the few options for organisms for which transgenic technologies are not available. Moreover, they tend to be faster and cheaper than studies on whole organisms, and can therefore be a valuable prelude to further *in vivo* work. A main drawback to *in vitro* studies, however, is that there might be only an indirect connection to organismal phenotypes.

Several recent studies have used *in vitro* methods to demonstrate that putatively selected changes lead to differences in protein function [63–67]. Ivarsson *et al.* [65] and Norrgard *et al.* [66] conducted extensive site-directed

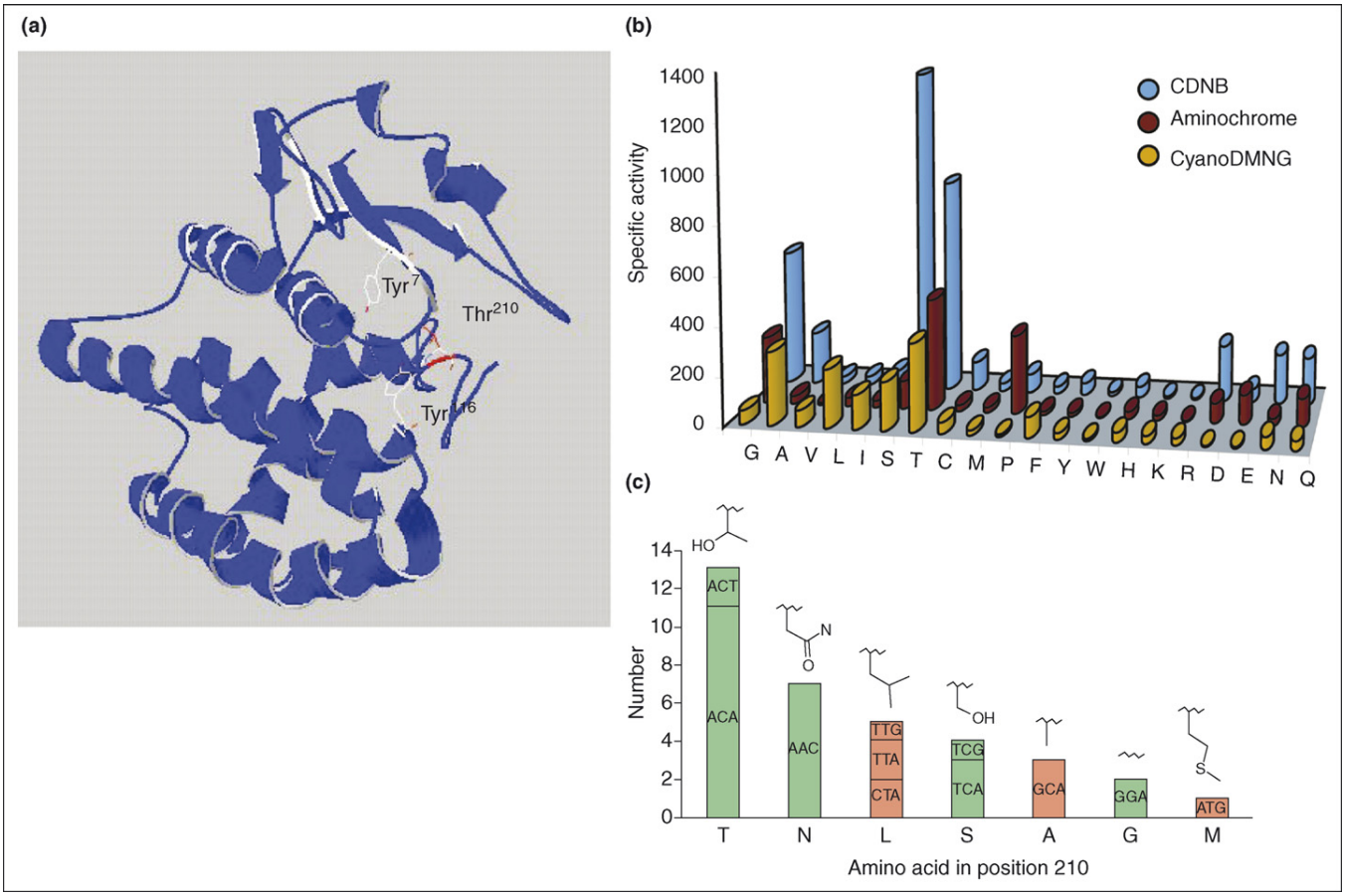

Figure 2. Site-directed mutagenesis of a putatively selected residue in glutathione S-transferase (GST) demonstrates shifts in substrate specificity. Panel **(a)** shows the crystal structure of human GST Mu2–2 (from entry 3GTU in the Protein Data Bank). Side chains are shown for the active site residues Tyr^{7} and Tyr^{116}, and for the putatively selected residue Thr^{210} (in red) [65]. Panel **(b)** shows enzyme activity for all twenty possible substitutions at position 210 of GST M2–2, using three different substrates. 1-chloro-2,4-dinitrobenzene (CDNB) and 2-cyano-1,3-dimethyl-1-nitrosoguanidine (cyanoDMNG) were used to assess the effects of substitutions on substitution reactions, and aminochrome was used to investigate addition reactions. Threonine is the naturally occurring amino acid at position 210. Note that different substitutions can show markedly different activities towards different substrates; compare, for example, proline (P) to glycine (G). Panel **(c)** shows the distribution of amino acids at position 210 in a panel of 35 mammalian GSTs. Polar residues are indicated in green, and hydrophobic residues are indicated in red. {Panels (b) and (c) have been modified, with permission, from [66]}.

mutagenesis at sites in vertebrate Mu-class glutathione transferases predicted to be under positive selection by divergence-based analyses. Glutathione transferases constitute a large family of detoxification enzymes, and thus might undergo positive selection in response to toxins in novel food sources or produced by novel pathogens. Norrgard *et al.* [66] found that changes at one highly variable site resulted in altered substrate specificity, with 1000-fold changes in activity towards some substrates, consistent with the hypothesis that functional diversification underlies rapid amino acid evolution in this gene family (Figure 2). Similarly, Sawyer *et al.* [67] used chimeric forms of the ubiquitin ligase TRIM5α to demonstrate a role for positively selected residues in response to HIV-1. In cultured feline cells, expression of wild type human TRIM5α confers only a weak ability to resist infection by HIV-1, whereas rhesus TRIM5α restricts the virus effectively. Divergence analyses of multiple primate species localized positive selection on TRIM5α primarily to a small 11–13 amino acid patch. Remarkably, a chimeric human TRIM5α that bears the rhesus 'patch' gains substantial restrictive capability, whereas transfer of the human patch to the rhesus protein reduces its activity against HIV-1.

A more direct assessment of the impact of putatively selected changes on organismal phenotypes can be gained from *in vivo* studies. The power to detect such an impact can be maximized by expressing alternate alleles, which differ at one or a few sites inferred to be under positive selection, in identical genetic backgrounds. Relevant phenotypes of individuals that bear alternative alleles can then be measured, bearing in mind the caveats described above with respect to appropriateness of laboratory conditions for phenotypic assessment. Such experiments are possible only in organisms for which transgenic technologies are available; furthermore, targeted gene replacement is usually preferable to random insertion of a transgene, owing to position effects. Association or quantitative trait locus studies can be of use; however, even large studies will be hampered by noncausative variants in linkage disequilibrium with the putatively selected change(s). Alternatively, use of a related model organism sometimes allows for the recapitulation of a phenotype observed in a nonmodel target species (e.g. [2]).

We know of no studies that have identified a locus under positive selection in a genome scan and that have used transgenic methods to examine *in vivo* the phenotypic consequences of selection. However, detection of positive selection has motivated functional studies using other methods. Schlenke and Begun [68], for example, identified selective sweeps in homologous genomic regions of California populations of *D. melanogaster* and *D. simulans*, but not in African (ancestral-like) populations. The sweeps are associated with independent transposable element insertions that increase transcription of a gene, *Cyp6g1*, that is implicated in DDT resistance. Lines that bear the insertions show marginally (*D. simulans*) or substantially (*D. melanogaster*) increased resistance to DDT, consistent with the hypothesis that the insertions swept to near-fixation owing to their impact on insecticide resistance. The evidence is not unambiguous, however, as the resistance studies were carried out on inbred African and Californian lines, such that variation elsewhere in the genome might contribute to resistance. Complete elimination of such background variation requires, instead, the use of transgenic technologies. Targeted gene replacement techniques offer a promising in-road, although accurate empirical estimation of fitness differences between alleles remains a difficult task (see Box 3).

Box 3. Targeted gene replacement

Targeted gene replacement techniques that allow for the comparison of alternate alleles at the same genomic location in otherwise identical genetic backgrounds are currently available in a handful of model eukaryotes, including mouse, yeast and *D. melanogaster* [84,85]. Greenberg *et al.* [86] used precise gene replacement to examine the phenotypic differences between alleles of a putative pre-mating isolation locus, desaturase-2 (*dsat2*), in *D. melanogaster*. *Dsat2* was initially identified through mapping studies, rather than as a target of selection; however, patterns of nucleotide polymorphism are suggestive of a recent selective sweep in derived worldwide populations, but not in ancestral African populations. Further application of targeted gene replacement methods to assess functional differences between naturally occurring alleles should prove to be a valuable tool for the functional verification of statistical inferences of positive selection (see also [87]).

Although the power to detect the phenotypic consequences of allelic variants can be maximized by targeted gene replacement in a homogeneous genetic background, direct measurement of the impact of natural variants on organismal fitness remains a considerable challenge. In nature, alternate alleles are present in many genetic backgrounds, with much opportunity for epistatic effects to modify the impact of a single mutation. An accurate estimate of the fitness consequences of a putatively selected change thus requires phenotypic measurement of alternate alleles in a randomized genetic background, such that the average fitness effect of an allele can be measured (e.g. [88]). No currently available technology is ideally suited to this task, although mutually supportive results from gene replacement and QTL or association studies might provide a strong argument in favor of a fitness effect.

Conclusions and future directions

The past decade has seen a significant gain in our understanding of the evolutionary and functional basis of adaptation, and of the role demography has played in shaping genomic diversity. Increasingly sophisticated statistical methods have enhanced capacity for data generation (both DNA sequence polymorphism and *de novo* sequencing of new species, including that at the whole-genome level), and more efficient and precise methods for functional analysis at the RNA, protein, cellular and organismal levels have all contributed to this deeper insight. However, because of limitations owing to sample size, demographic effects, and the specification of appropriate models, it is likely that we have just scratched the surface (Box 4).

Even in cases where strong signals of adaptation at the molecular level have been found, connecting the molecular function to organismal context has often remained elusive. *Alcohol dehydrogenase* (*Adh*) in *Drosophila* is a good example, for which evidence of both balancing selection within *D. melanogaster* and of adaptive fixations between species exist. Expression and kinetic differences in the allelic products have also been studied, but the specific selective forces acting in natural populations remain speculative at best [69]. The discovery of footprints of

Box 4. Outstanding questions

Determining an optimal sample size

Although the statistical power to detect departures from an equilibrium neutral model have long been known to be a function of sample size, studies that use sample sizes of only ten or so allelic copies of a gene have prevailed until recently (e.g. [41]). Tests such as Tajima's *D* have little power without sample sizes of upwards of 50 alleles [41]. Whether increased sample sizes will help resolve footprints of adaptive processes in more genes remains to be seen. One interesting example consists of male reproductive proteins (accessory gland proteins), for which tantalizing hints of selection for protein diversification exist, but for which statistically significant departures from neutrality have been limited to date for many studies (e.g. [89]).

Investigating alternative models of selection

Evidence for adaptive selection acting on several types of variation (synonymous variation, nonsynonymous variation, and putative regulatory sequences) in a roughly 15 kb region overlapping the 3′ end of the *Notch* locus [79] raises the question of how seriously genome scans of smaller regions sampled every ten to hundreds of kb across the genome have underestimated the frequency of adaptive fixations (e.g. [78]). The ability of polymorphism studies to detect recent footprints of selection also is largely restricted to the recent fixation of new, or previously very rare, adaptive mutations. The detection of positive selection that acts on previously segregating variation (owing perhaps to a change in environment and thus selection pressures) remains a formidable challenge for which very large sample sizes might be the only hope, given the very weakly detectable effect of such fixations on flanking variation. We have focused here on adaptive fixations, but the roles of various types of balancing selection (e.g. frequency dependent selection, temporally varying selection, heterozygote advantage, spatially varying selection) are relatively unexplored, in part owing to the extensive sampling over time and space that are required. The exponential increase in sequencing capacity (and dramatic drop in cost) should allow for more powerful tests of these types of fitness consequences of naturally occurring variation.

Disentangling demographic effects

The challenge of distinguishing nonequilibrium demographic effects from adaptive processes remains, although the incorporation of higher dimension metrics of variation (that capture more of the structure of linkage disequilibrium) shows promise [46]. The success of these methods of course will demand significantly larger sample sizes for the accurate description of patterns of linkage disequilibrium. It is also true that one could avoid many of the confounding effects of nonequilibrium demography by focusing on populations thought to be closer to equilibrium (or at least not highly structured or recently bottlenecked). However, many particularly interesting biological questions center on the impact of colonization of new environments, which is usually associated with the strong reduction of population size owing to the founding event.

adaptation from genome scans of polymorphism and/or divergence are somewhat like Quantitative Trait Locus mapping experiments; signals of selection, however, often only hint at the target locus, and sometimes provide limited insight into the phenotypic trait on which selection is acting. The contribution of these approaches to the functional annotation of genomes, by virtue of inferences not just of what is the same among genomes but what is different, has a bright future, but one for which substantive challenges remain that ensure a rich interplay between empirical, theoretical, statistical, computational, and functional approaches. Recognition of the ecological context will be increasingly important, and will bring an even deeper richness to the study of adaptation at the molecular level.

Acknowledgements

The authors would like to thank Nadia Singh, Kevin Thornton, Dan Merl and the Aquadro laboratory for helpful comments and discussion. This work was supported in part by grants to J.D.J. (National Science Foundation [NSF] biological informatics post-doctoral fellowship), A.W. (Howard Hughes Medical Institute pre-doctoral fellowship, and National Science Foundation Doctoral Dissertation Improvement Grant 0508152), and C.F.A. (NIH GM36431 and NSF/NIH grant DMS-0201037 to R. Durrett, C.F.A. and R. Nielsen).

References

1 Hoekstra, H.E. *et al.* (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* 313, 101–104
2 Abzhanov, A. *et al.* (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's Finches. *Nature* 442, 563–567
3 Wright, S.I. *et al.* (2005) The effects of artificial selection on the maize genome. *Science* 308, 1310–1314
4 Schlotterer, C. (2003) Hitchhiking mapping - functional genomics from the population genetics perspective. *Trends Genet.* 19, 32–38
5 Payseur, B.A. and Cutter, A.D. (2006) Integrating patterns of polymorphism at SNPs and STRs. *Trends Genet.* 22, 424–429
6 Nielsen, R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* 39, 197–218
7 Biswas, S. and Akey, J.M. (2006) Genomic insights into positive selection. *Trends Genet.* 22, 437–446
8 Sabeti, P.C. *et al.* (2006) Positive selection in the human lineage. *Science* 312, 1614–1620
9 Nachman, M.W. (2005) The genetic basis of adaptation: lessons from concealing coloration in pocket mice. *Genetica* 123, 126–136
10 Thornton, K.R. *et al.* (2007) Progress and prospects in mapping recent selection in the genome. *Heredity* 98, 340–348
11 Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176
12 Smith, N.G. and Eyre Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415, 1022–1024
13 Sawyer, S.A. *et al.* (2007) Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 6504–6510
14 Wiehe, T.H. and Stephan, W. (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10, 842–854
15 Stephan, W. (1995) An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* 12, 959–962
16 Andolfatto, P. (2001) Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* 11, 635–641
17 Innan, H. and Kim, Y. (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10667–10672
18 Kim, Y. (2006) Allele frequency distribution under recurrent selective sweeps. *Genetics* 172, 1967–1978
19 Li, H. and Stephan, W. (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genet.* 2, e166
20 Przeworski, M. (2002) The signature of positive selection at randomly chosen loci. *Genetics* 160, 1179–1189
21 Rosenberg, N.A. and Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380–390
22 Orr, H.A. and Betancourt, A.J. (2001) Haldane's sieve and adaptation from the standing genetic variation. *Genetics* 157, 875–884
23 Hermisson, J. and Pennings, P.S. (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169, 2335–2352
24 Przeworski, M. *et al.* (2005) The signature of positive selection on standing genetic variation. *Evolution Int. J. Org. Evolution* 59, 2312–2323
25 Pennings, P.S. and Hermisson, J. (2006) Soft Sweeps II – molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* 23, 1076–1084
26 Parsch, J. *et al.* (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

27 Meiklejohn, C.D. *et al.* (2004) Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* 168, 265–279
28 Teshima, K.M. and Przeworski, M. (2006) Directional positive selection on an allele of arbitrary dominance. *Genetics* 172, 713–718
29 Thornton, K.R. and Jensen, J.D. (2007) Controlling the false positive rate in multilocus genome scans for selection. *Genetics* 175, 737–750
30 Durrett, R. and Schweinsberg, J. (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Proc. Appl.* 115, 1628–1657
31 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis. *Genetics* 123, 585–595
32 Fu, Y-X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* 133, 693–709
33 Wakeley, J. and Aliacar, N. (2001) Gene genealogies in a metapopulation. *Genetics* 159, 893–905
34 Jensen, J.D. *et al.* (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170, 1401–1410
35 Thornton, K. and Andolfatto, P. (2006) Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in non-African populations of *Drosophila melanogaster*. *Genetics* 172, 1607–1619
36 Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favorable gene. *Genet. Res.* 23, 23–35
37 Kaplan, N.L. *et al.* (1989) The "hitchhiking effect" revisited. *Genetics* 123, 887–899
38 Stephan, W. *et al.* (1992) The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* 41, 237–254
39 Hudson, R.R. *et al.* (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159
40 Braverman, J.M. *et al.* (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140, 783–796
41 Simonsen, K.L. *et al.* (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141, 413–429
42 Fu, Y-X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915–925
43 Fay, J.C. and Wu, C-I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413
44 Kim, Y. and Nielsen, R. (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167, 1513–1524
45 Stephan, W. *et al.* (2006) Hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 172, 2647–2663
46 Jensen, J.D. *et al.* (2007) On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. *Genetics* 176, 2371–2379
47 Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765–777
48 Glinka, S. *et al.* (2006) Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Mol. Biol. Evol.* 23, 1869–1878
49 Yang, Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449
50 Suzuki, Y. *et al.* (2001) ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* 17, 660–661
51 Wong, W.S. *et al.* (2006) Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* 16, 148
52 Huelsenbeck, J.P. and Dyer, K.A. (2004) Bayesian estimation of positively selected sites. *J. Mol. Evol.* 58, 661–672
53 Wong, W.S. and Nielsen, R. (2004) Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167, 949–958
54 Yang, Z. *et al.* (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118
55 Pond, S.L. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679
56 Wilson, D.J. and McVean, G. (2006) Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172, 1411–1425
57 Clark, A.G. *et al.* (2003) Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302, 1960–1963
58 Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167–170
59 Suzuki, Y. (2006) Natural selection on the influenza virus genome. *Mol. Biol. Evol.* 23, 1902–1911
60 Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328
61 Swanson, W.J. *et al.* (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. U. S. A.* 98, 2509–2514
62 Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152
63 Zhang, J. *et al.* (2004) Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16246–16250
64 Benderoth, M. *et al.* (2004) Positive selection driving diversification in plant secondary metabolism. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9118–9123
65 Ivarsson, Y. *et al.* (2003) Identification of residues in glutathione transferase capable of driving functional diversification in evolution. A novel approach to protein redesign. *J. Biol. Chem.* 278, 8733–8738
66 Norrgard, M.A. *et al.* (2006) Alternative mutations of a positively selected residue elicit gain or loss of functionalities in enzyme evolution. *Proc. Natl. Acad. Sci. U. S. A.* 103, 4876–4881
67 Sawyer, S.L. *et al.* (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2832–2837
68 Schlenke, T.A. and Begun, D.J. (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1626–1631
69 Powell, J.R. (1997) *Progress and Prospects in Evolutionary Biology: the Drosophila Model*. Oxford University Press
70 Casa, A.M. *et al.* (2006) Evidence for a selective sweep on chromosome 1 of cultivated sorghum. *The Plant Genome*, a suppl. to *Crop Sci.* 46, S27–S40
71 Hamblin, M.T. *et al.* (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. *Genetics* 171, 1247–1256
72 Hamblin, M.T. *et al.* (2006) Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics* 173, 953–964
73 Schweighofer, A. *et al.* (2004) Plant PP2C phosphatases: emerging function in stress signaling. *Trends Plant Sci.* 9, 236–243
74 Yoshida, T. *et al.* (2006) *ABA-Hypersensitive Germination3* encodes a protein phosphatase 2C (AtPP2CA) that strongly regulates abscisic acid signaling during germination among *Arabidopsis* protein phosphatase 2Cs. *Plant Physiol.* 140, 115–126
75 Doebley, J. (2004) The genetics of maize evolution. *Annu. Rev. Genet.* 38, 37–59
76 Doebley, J.F. *et al.* (2006) The molecular genetics of crop domestication. *Cell* 127, 1309–1321
77 Beisswanger, S. *et al.* (2005) Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* 172, 265–274
78 Glinka, S. *et al.* (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* 165, 1269–1278
79 DuMont, V.B. and Aquadro, C.F. (2005) Multiple signatures of positive selection downstream of notch on the X chromosome in *Drosophila melanogaster*. *Genetics* 171, 639–653
80 Pond, S.L. and Muse, S.V. (2005) Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* 22, 2375–2385
81 Wong, A. *et al.* (2006) Phylogenetic incongruence in the *Drosophila melanogaster* species group. *Mol. Phylogenet. Evol.* 43, 1138–1150
82 Pollard, D.A. *et al.* (2006) Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* 2, e173
83 Anisimova, M. *et al.* (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236
84 Rong, Y.S. and Golic, K.G. (2000) Gene targeting by homologous recombination in *Drosophila*. *Science* 288, 2013–2018
85 Venken, K.J. *et al.* (2006) P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* 314, 1747–1751

86 Greenberg, A.J. *et al.* (2003) Ecological adaptation during incipient speciation revealed by precise gene replacement. *Science* 302, 1754–1757
87 Mitchell-Olds, T. and Schmitt, J. (2006) Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature* 441, 947–952
88 Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change.* Columbia University Press
89 Kern, A.D. *et al.* (2004) Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. *Genetics* 167, 725–735