Detecting Natural Selection in Genomic Data

Detecting natural selection in the genome

Joseph J. Vitti[1,2,*], Sharon R. Grossman[2,3,4], & Pardis C. Sabeti[1,2,*]
1 Department of Organismic & Evolutionary Biology, Harvard University, Cambridge MA 02138
2 Broad Institute of MIT and Harvard, Cambridge MA 02142
3 Department of Systems Biology, Harvard Medical School, Boston MA 02115
4 Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139
* Corresponding authors: jvitti@fas.harvard.edu, psabeti@oeb.harvard.edu

Contents

Keywords

population genetics, adaptation, selective sweeps, genome scans, evolutionary genomics

ABSTRACT

The past fifty years have seen the development and application of numerous statistical methods to identify genomic regions that appear to be shaped by natural selection. These methods have been used to investigate the macro- and microevolution of a broad range of organisms, including humans. Here we provide a comprehensive outline of these methods, explaining their conceptual motivations and statistical interpretations. We highlight areas of recent and future development in evolutionary genomics methods, and discuss ongoing challenges for researchers employing such tests. In particular, we emphasize the importance of functional follow-up studies to characterize putative selected alleles, and the use of selection scans as hypothesis-generating tools for investigating evolutionary histories.

INTRODUCTION

As humans and other organisms moved to inhabit every part of the world, they were exposed and adapted to myriad new environments, diets and pathogens, leading to the great diversity we observe today. Uncovering the mechanism of this diversification has fascinated scientists and

non-scientists alike for years. In 1858, Darwin and Wallace gave grounds for species evolution when they articulated the principle of natural selection, the idea that beneficial traits – those that improve an individual's chances to survive and reproduce – will tend to become more frequent in populations over time.

Scientists have continued to search for evidence of evolution and for the specific adaptations that underlie it. Animal and plant breeders were some of the first to identify traits that are evolving, as they witnessed dramatic changes in their breeds through artificial selection. Haldane uncovered the first adaptive trait in humans when he observed that many diseases of red blood cells seemed to be distributed in regions where malaria was endemic (48). Haldane's 'malaria hypothesis' was confirmed by Allison a few years later, when he demonstrated that the sickle cell mutation in the *Hemoglobin-B* gene (*HBB*) was the target of selection for malaria resistance (4).

The ability to assess evidence for selection at the genetic level represented a breakthrough for this pursuit. Population genetic methods and datasets provide a statistically rigorous way to carry out analysis; in this way, the field of evolutionary genetics represents an antidote to the preponderance of speculative 'just-so-stories' that some biologists have lamented (42). Moreover, it demonstrates the full realization of the modern synthesis: Darwinian concepts of selection have been rendered quantitative and measurable in real populations, thanks to methodological and technological advances (1).

Through evolutionary genetics, many different traits became uncovered, from lactase persistence and skin pigmentation in humans (89, 123), to coat color in field mice (80), to armored plates in stickleback fish (63). With ongoing advancements in genomic technology, we can now go further, from testing evidence for selection on specific traits hypothesized to be adaptive, to uncovering candidate regions through genome scans. This transition from hypothesis-testing to hypothesis-generating science has been made possible both by the new data (e.g. genome sequences from increasing numbers of species, genome-wide variation data) and by increasingly sophisticated tools that allow us to make sense of this deluge of data, and to fine-map evidence of selection to individual candidate variants.

Identifying such candidates is significant not only because they demonstrate evolution and shed light on species histories, but also because they represent biologically meaningful variation. Because selection operates at the level of the phenotype, alleles showing evidence of selection are likely to be of functional relevance. Thus, alleles implicated in selection studies are often linked to resistance to various infectious diseases, as pathogens are believed to represent the strongest selective pressure acting on humans, (40), or to non-infectious genetic diseases such as those associated with autoimmune diseases or metabolic disorders (54).

Further breakthroughs in genomic annotation, genome manipulation technology, and high-throughput molecular biology, are beginning to allow researchers to progress from candidate variants to functionally elucidated instances of evolution. Taken together, all these advancements present a path forward for realizing the full potential of evolutionary genomics to shed light on species histories and to uncover biologically meaningful variation.

Modes of Selection

Natural selection is based on the simple observation that fitness-enhancing traits – i.e., those that improve an organism's chance of survival or reproductive success – are more like to be passed on to that organism's offspring, and therefore will increase in prevalence in the population over time. More generally, selection refers to any non-random, differential propagation of alleles as a consequence of their phenotypic effect. There are many specific modes of selection that have been described, some of which share conceptual overlap, and some of which are referred to by multiple names. In this section, we briefly define the different modes of selection that we will employ in our discussion, following the terminology outlined by Nielsen (84).

Most simply, selection may act in a directional manner, in which an allele is favored and so propagated (positive selection) or disfavored (negative selection, also called purifying selection). Random mutations are more likely to be deleterious than beneficial, so many novel alleles are immediately subject to negative selection – and so become removed from the gene pool before they can achieve detectable frequency within the population. This ongoing removal of deleterious mutations is a form of negative selection referred to as background selection, and is inferred to be occurring in highly conserved genomic regions – i.e., regions where substitutions and polymorphism are not observed, presumably because most novel variants are deleterious and are quickly removed from the population (19).

More subtle configurations of positive and negative selection give rise to other common evolutionary trends, particularly (though not exclusively) in diploid and polyploid organisms, where the phenotype depends on the interaction of multiple alleles at the same locus. One such phenomenon is balancing selection, in which multiple alleles are maintained at an appreciable frequency within the gene pool. This may happen as the result of heterozygote advantage (i.e. overdominance) or frequency-dependent selection, for example (20). If the alleles being maintained conduce to opposing phenotypic effects – for example, if large and small body sizes are maintained within the population, to the exclusion of intermediate sizes – then the trend is often further described as diversifying or disruptive selection. By contrast, when intermediate phenotypic values are favored – whether by balancing selection of codominant alleles, or positive selection of alleles that underlie intermediate phenotypes, the trend is called stabilizing selection.

This diversity of modes of selection notwithstanding, much research in recent years has focused on the development of genomic methods to identify positive selection. One reason for this emphasis on positive selection is practical: whereas negative selection is primarily observable in highly conserved regions, and balancing selection's effect on the genome is often subtle, positive selection leaves a more conspicuous footprint on the genome that can be detected using a number of different approaches. Another reason for interest in positive selection is theoretical: positive selection is understood to be the primary mechanism of adaptation (i.e. the genesis of phenotypes that are apt for a specific environment or niche), which in turn poses great theoretical interest to researchers (1).

Here we discuss the various approaches that have been used to identify positive selection, while also indicating the ways that these methods may be used to detect and classify instances of other modes of selection (**Table 1**). These approaches typically use summary statistics to compare

observed data with expectations under the null hypothesis of selective neutrality (see sidebar, "Selection and Neutrality").

We begin by discussing methods based on comparisons of different species and their relative rates of genetic change. These methods are most often used to identify selective events that took place within the deep past, and reflect macroevolutionary trends that occur as the result of selection between, rather than within, species. We then turn our attention to population genetic methods used to identify microevolutionary selective events within species. Loci identified by these latter methods are believed to underlie local adaptations in humans following the out-of-Africa migration, and thus have become the subject of much research towards understanding human evolution and history (109).

DETECTING SELECTION AT THE MACROEVOLUTIONARY LEVEL

Methods to detect selection at the macroevolutionary level typically hinge on comparisons of homologous traits or sequences among related taxa (**Figure 1A**). These methods identify sequences that are likely to be functional (either because they code for proteins, or because they are conserved among different species), and search for lineage-specific accelerations in the rate of evolution. Such accelerations are indicated by an excess of mutations compared to the baseline mutation rate, which can either be calculated from the rate of synonymous mutations (which are generally considered neutral), or from the overall rate of substitutions between species.

Gene-Based methods

Perhaps the best-known statistic for detecting selection is K_a/K_s, also known as d_N/d_S or ω (**Figure 1B**). This statistic compares the rate of non-synonymous substitutions per site (i.e., per potential non-synonymous change) to the rate of synonymous substitutions per site (i.e., per potential synonymous change) (60). Because synonymous changes are assumed to be functionally neutral ('silent'), their substitution rate provides a baseline against which the rate of amino acid alterations can be interpreted. A relative excess of non-synonymous substitutions indicates ongoing (or recently ended) positive selection favoring novel protein structures (or else a cessation of negative selection against protein alterations – see section, "Challenges in Detecting Selection" below). This is summarized by a value of K_a/K_s greater than 1, whereas smaller values indicate ongoing negative selection against deleterious mutations and the consequent preservation of protein structure. These methods may also be applied across an entire open reading frame, or some subdivision thereof (down to an individual codon), as different regions of a protein may be subject to different selective pressures (134). Various models for calculating synonymous and non-synonymous substitution rates take into account the different probabilities of different mutations (e.g., transitions are more likely than transversions), as well as the possibility of unobserved changes (e.g., if one species undergoes two sequential mutations at the same site) and codon usage bias (43).

The McDonald-Kreitman Test (MKT) builds upon this method by utilizing not only interspecies divergence data, but also intraspecies diversity data (77). Essentially, the MKT compares two K_a/K_s values, one between species and one within species. Under neutrality, these rates should be equal, given constant rates of mutation and substitution. If the between-species ratio significantly exceeds the within-species ratio, the null hypothesis can be rejected, suggesting positive

selection between species. Conversely, a larger within-species value suggests balancing selection or else a surplus of maladaptive variants (e.g. recessive disease alleles) under weak negative selection within the species (see "Detecting Selection at the Microevolutionary Level").

Other Rate-Based methods

Similar to the MKT, the Hudson-Kreitman-Aguadé Test (HKA) uses both divergence and diversity data to compare relative rates of change (**Figure 1C**). Specifically, the HKA examines the ratios of fixed interspecific differences (D – i.e., substitutions) to within-species polymorphisms (P) across loci (59). The test hinges on the supposition that, for a neutral site, both D and P are functions of the site's mutation rate, which is assumed to have been roughly constant at least since the point of species divergence. Using a goodness-of-fit test (e.g. χ^2), one can check individual sites for deviation from the D/P ratio at a neutral site, which allows rejection of the null hypothesis and therefore can be interpreted as evidence for selection. Relatively large D/P values indicate either that change contributing to speciation was accelerated (directional selection between species) or that diversity within the species is reduced (directional selection within species – see "Detecting Selection at the Microevolutionary Level"). Relatively small values suggest balancing selection between species.

One advantage of the HKA approach is that it can be applied to any genetic region, not just those that code for proteins. In practice, however, the rate of neutral evolution in protein-coding regions is much easier to determine (i.e., by examining the synonymous substitution rate). The variability of the mutation rate across different loci, coupled with a lack of any a priori understanding of which sites (or, indeed, what percentage of sites) are neutral, has historically made application of the HKA test challenging (138). In recent years, however, researchers have expanded this approach in a maximum likelihood framework to allow more efficient multilocus comparisons (133). By examining multiple sites, one can derive the expected neutral D/P ratio for a lineage, while accounting for variation in the mutation rate.

Other studies have used comparative genomic data to identify elements in the genome that are highly conserved between disparate species, but show a significantly accelerated rate of substitution in a particular species or lineage (14, 99, 101). For example, the gene *HAR1F*, a non-coding RNA expressed during brain development, is highly conserved between chimpanzees and other vertebrate, but has 40 times more substitutions in humans than expected under neutrality (100). This approach has been used to identify several hundred human-specific and primate-specific regions (76). Similar relative-rate methods have also been employed in understanding bacterial evolution (114).

Phenotypic methods

The idea of comparing related species and identifying striking differences can also be applied to phenotypes. Traits that are conserved across many closely related species (and thus likely to be functional) but show extreme differentiation in just one or a few of these are strong candidates for natural selection (108). This approach has been used recently in comparative studies of gene expression (13, 96). The gene *SDR16C5*, for example, regulates the metabolism of retinol, a form of vitamin A that is common in tree exudates. Slow lorises and marmosets, which feed on tree bark, show highly elevated expression levels of *SDR16C5* in the liver compared to their

close evolutionary cousins, suggesting selection on regulatory elements as a preventative measure against vitamin A toxicity (96).

Alleles or traits that repeatedly arise in independent lineages suggest the action of convergent evolution. This signature has been observed in morphological traits, for instance the loss of pelvic structures in stickleback fish (18) and wing pigmentation patterns in *Drosophila* (104). It is also seen in viral and bacterial evolution, in particular in the emergence of drug resistance (12, 58).

DETECTING SELECTION AT THE MICROEVOLUTIONARY LEVEL

Positive selection causes a beneficial allele to 'sweep' to high prevalence or fixation (100% prevalence) rapidly within the population. When a beneficial allele and surrounding variants on the same haplotype reach high prevalence together, it produces a population-wide reduction in genetic diversity (sometimes referred to as heterozygosity, polymorphism, or variability) surrounding the causal allele (117). This reduction in diversity, which persists until recombination and mutation restore diversity to the population at the selected locus, is the hallmark of a selective sweep (**Figure 2A**). There are various ways of quantifying and detecting this signal, which we discuss in the following two sections. We then discuss methods based on the environmentally specific nature of selection, which compare populations where selection is or is not hypothesized to be at play. We then turn our attention to methods that combine multiple variants in a region or distinct tests to provide greater power and resolution.

Frequency Spectrum-Based Methods

As a selected allele and its nearby 'hitchhiker' genetic region sweep towards fixation, they shift the distribution of alleles in the population (**Figure 2B**). As discussed above, a sweep will cause a population-wide reduction in genetic diversity surrounding the selected locus. New mutations will appear on this homogenous background, but they will initially be rare, since they have only recently appeared in the population. This creates a surplus of rare alleles (i.e., many sites near the selected variant will have alleles that segregate at low frequencies). Although the frequency spectrum will shift back to baseline over time, the distortion will persist for thousands of generations (several hundred thousand years, in the case of humans). Tajima's D was the first, and most commonly used, test to detect this signal (120).

Tajima's D quantifies this phenomenon by comparing the number of pairwise differences between individuals with the total number of segregating polymorphisms. If this difference is smaller (i.e., more negative) than expected under a neutral model, it suggests a shift towards rare polymorphisms, which increase the number of segregating sites but do not significantly affect the number of pairwise differences. Several variations on this method have been developed to take into account the polarity of each allele (i.e. which one is derived compared to an evolutionary outgroup) and to measure the abundance of rare alleles in different ways (38, 39).

Selective sweeps also distort the frequency spectrum by increasing the frequency of derived alleles. Under genetic drift, it takes many generations to bring neutral mutations to moderate or high prevalence. However, in a selective sweep, any derived alleles that reside near the causal allele will also 'hitchhike' to high frequency. Using a similar approach as Tajima's D, Fay and Wu's H compares the number of pairwise differences between individuals to the number of

individuals homozygous for the derived allele (33). Small values of H suggest positive selection in the region examined.

Site frequency spectrum analysis can also be very useful for other modes of selection, like balancing selection, where an excess of intermediate frequency alleles will distort metrics like Tajima's D (20). Andrés et al. sought long-term balancing selection in the human genome by leveraging frequency spectrum methods together with a modification on the HKA test to detect an excess of diversity in regions linked to the selected variants (5). Long-term balancing selection will result greater coalescence time than expected under neutrality, and thus fewer rare alleles.

Linkage Disequilibrium-Based Methods

As it sweeps through the population, a selected allele will persist in strong linkage disequilibrium (LD) with its neighboring 'hitchhiker' variants until recombination causes these associations to break down. Together, the causal allele and its linked neighbors define a haplotype. Thus, a third suite of methods for detecting positive selection looks for extended regions of strong LD (or, equivalently, long haplotypes) relative to their prevalence within a population (**Figure 2C**). The thought is that such regions must have swept to high prevalence quickly, or else recombination would have caused LD to break down and the haplotype to shorten.

LD-based approaches are particularly useful for identifying variants that have undergone a partial or incomplete selective sweep, in which a new mutation has risen to a modest frequency in the population, rather than reaching fixation. This is useful in many species including humans, as most novel alleles since the out-of-Africa migration with realistic selection coefficients are unlikely to have yet reached fixation (**Figure 3).** For example, the causal allele of lactase persistence, which has a dominant effect, is expected to take roughly 50,000 years to reach fixation, far longer than it has been in existence. Thus despite offering one the strongest known selective advantages (with a selection coefficient estimated at 0.039) in humans, the allele frequency is only 80% in Europeans (123). Beneficial mutations that arose more recently or were under less extreme selective pressure are even more likely to remain polymorphic in the selected population, and many will never reach fixation, since selective pressures can change greatly over tens of thousands of years. LD-based approaches can also be used to identify short-term balancing selection, where the signal is comparable to that of an incomplete sweep. For example, a number of papers have demonstrated long-haplotype signals at the sickle cell mutation in West Africa (51, 52).

One suite of widely-used LD-based tests for selection centers around the extended haplotype homozygosity statistic (110). One defines EHH from a core region (e.g., a putatively selected allele) to a specified distance out in both directions, and calculates the probability that any two randomly chosen chromosomes within the population carrying the core region are identical by descent for the entire region. Thus, as one travels further from the core region, EHH decreases, reflecting the action of recombination whittling down the haplotype within the population. The Long-Range Haplotype test (LRH) compares a haplotype's frequency to its relative EHH at various distances, looking for haplotypes that are extended as well as common, suggesting that they rose to high prevalence quickly enough that recombination has not had time to break down the haplotype. Zhang et al. adapted this test by focusing on derived alleles (which are believed

to be more likely candidates for selective sweeps), as well as introducing a genome-wide score (139).

The integrated haplotype score, iHS (129) is an influential variation on EHH. This statistic takes the area under the curve defined by EHH as one travels further in genetic distance from the core region, thus capturing the intuition that both extreme EHH for a short distance and moderate EHH for a longer distance are suggestive of the action of selection. Another variation is the cross-population extended haplotype homozygosity statistic, XP-EHH (111). This method compares haplotype lengths between populations in order to control for local variation in recombination rates. These two methods are complementary in terms of their scope: whereas iHS has more power to detect incomplete sweeps, XP-EHH is useful when the sweep is near fixation within one population (98).

Other LD-based tests include the linkage disequilibrium decay (LDD) test, which circumvents the need to determine haplotypes (i.e. by phasing) by limiting its scope to homozygous SNP sites and inferring the fraction of recombinant chromosomes at adjacent polymorphisms (130). Recently, Wiener and Pong-Wong developed a new test that fits a regression to heterozygosity data as a function of genomic position: selection is inferred based on the goodness-of-fit to the reduction in heterozygosity as predicted in a selective sweep (132). The strength of this test is that whereas traditional LD-based approaches are designed for analysis of SNP data, their regression test can be used with any genetic marker.

In recent years, a number of researchers have adapted identity-by-descent (IBD) analyses to selection mapping, invoking essentially the same conceptual motivations as earlier EHH-based approaches (15, 50). IBD analyses, which have been employed in a number of population history analyses (140), similarly search for regions in which a set of individual share a long stretch of DNA, a pattern that presumably can only be due to shared ancestry. Although IBD- and EHH-based methods look for the same pattern in genomic data, differences in their computational implementation give IBD-based approaches the advantage of being able to detect selection on standing variation (see "Selection on Standing Variation and Soft Sweeps") with greater power than EHH-based approaches (3).

Population Differentiation-Based Methods

An allele's selective valence is dependent on the particular environment in which it exists. Different populations are subject to different environmental pressures, and as a result, the traits that would be adaptive in each may be different. If selection is acting on a locus within one population, but not within other related populations, then the allele frequencies at that locus among the populations can differ significantly (**Figure 2D**). This principle is the foundation of a set of tests that rely on population differentiation to detect evidence of selection.

The most commonly used metric for population differentiation is Wright's fixation index, F_{st}, which compares the variance of allele frequencies – often estimated using measures of genetic diversity – within and between populations (57). Comparatively large values of F_{st} at a locus (i.e. relative to neutral regions) indicate stark differentiation between populations, which is suggestive of directional selection. Comparatively small values, on the other hand, indicate that the populations being compared are homogenous, which may be indicative of balancing or

directional selection in both. Unlike other methods, population differentiation-based approaches can detect many types of selection, including classic sweeps, sweeps on standing variants, and negative selection. In recent years, a number of alternative statistics and variations on F_{st} have also been proposed (for review, see (78)).

F_{st}-based tests for selection have a long history, originating with the Lewontin-Krakauer Test (LKT) in 1973 (72). This method uses the (then limited) available data to estimate F_{st} at multiple loci within *n* populations, and evaluates the neutrality of this distribution based either on its goodness-of-fit to a χ^2 distribution with *n*-1 degrees of freedom, or else based on the comparison of this distribution's variance with a theoretical predicted value. The production of large genetic datasets in recent years has made feasible a more robust application of this test, in which researchers compare the genome-wide distribution of F_{st} to individual loci (2).

Although such 'outlier approaches' are believed to mitigate the effect of demographic events – operating on the understanding that such events affect the genome in its totality, whereas selection acts in a locus-specific manner – certain patterns of migration and mutation within subpopulations can still produce false-positives (81). To correct for these effects, new variations on this test have also been developed that incorporate explicit, user-specified assumptions about demographic history (11, 31, 127). Bonhomme et al.'s T_{FLK} statistic, for example, modifies the Lewontin-Krakauer Test (labeled T_{LK} by the authors) to incorporate a kinship matrix, derived from prespecified neutral loci, to account for historical population branching (11). Another line of development reinterprets the F_{st} metric within a Bayesian framework, often implemented via Markov chain Monte Carlo algorithms (9, 36, 107). These approaches utilize F_{st}-based statistics to estimate the posterior probability of a given allele being under selection.

Other metrics that derive from F_{st} improve its computational power by incorporating more data. This data comes from either a greater number of populations, or a greater number of allelic sites. Following the former strategy, the locus-specific branch length metric (LSBL) uses pairwise calculations of F_{st} from three or more populations to isolate population-specific changes in allele frequency relative to a broader genetic context (115). On the other hand, the cross-population composite likelihood ratio of allele frequency differentiation (XP-CLR) extends F_{st} to many loci (21). This method, which is analogous to the XP-EHH method discussed above, identifies genetic regions where changes in allele frequency over many sites occurred too quickly (as assessed by the size of the affected region, which would gradually return to a neutral distribution over time) to be due to genetic drift. More recently, Fariello et al. introduced a new statistic, hapFLK, that examines differentiation among populations based on haplotypes rather than individual alleles (32).

Composite Methods

As the above discussion suggests, each test is tailored towards slightly different signals, and has its own strengths and limitations. Accordingly, researchers will sometimes combine multiple metrics into composite tests in order to provide greater power and/or spatial resolution. These tests come in two distinct formations, both of which are typically referred to as 'composite.'

First, some methods will form a composite score for a genetic region, rather than a single genetic marker, by combining individual scores at all the markers within the region. The motivation for

such an approach is that, while false-positives may occur at any one site by chance, a contiguous region of positive markers is much more likely to be a bona fide signal (16). Indeed, because selective sweeps affect whole haplotypes, one assumes that the signal of selection will extend across a region. Thus, composite methods that incorporate the same test across multiple sites improve power and reduce false discovery rate. Several of the previously discussed tests, including iHS, XP-EHH, and XP-CLR, employ such window-based analyses.

One exemplar of this approach is Kim and Stephan's composite likelihood ratio (CLR) test, which evaluates the probability of a selective event being responsible for a surplus of derived alleles (i.e., a skew of the unfolded site frequency spectrum) across multiple sites (67). Subsequent variations also incorporated LD-based data (66) and a goodness-of-fit test to help distinguish selection from demographic events (62). These tests calculate a composite likelihood by multiplying marginal likelihoods for each site considered within a sequence, and then compare the composite likelihood under a model where a sweep has occurred to the composite likelihood under a model where no sweep has occurred. In the above tests, the null hypothesis was calculated based on a population genetic model, which Nielsen et al. further modified by deriving the null hypothesis from background patterns of variation in the data itself (88). In a later, separate composite test, Nielsen et al. created a two-dimensional site frequency spectrum using allele frequencies from two populations; analysis of this table involved the combination of population differentiation-based signatures (i.e., F_{st}) with measures for high frequency derived alleles and excesses of low frequency alleles (85).

Whereas these methods combine the results of one test for many variants, other composite methods combine the results of many tests at a single site. The purpose of these composite methods is to utilize complementary information from different signals in order to provide better spatial resolution (**Figure 2E**).

One such line of composite test development began with the unification of Tajima's D and Fay and Wu's H, each of which are sensitive to different demographic processes (136). The authors later observed that, by limiting themselves to site frequency spectrum-based methods, they limit the power of their test in the presence of high recombination rates, and opted to further incorporate the Ewens-Watterson test, which compares the population's Hardy-Weinberg homozygosity to that predicted under a neutral model (30, 131), and is largely insensitive to recombination (137). Another composite test of this sort was developed by Grossman et al. (44). This test, called composite of multiple signals test (CMS), incorporates metrics from all three suites of methods discussed here. Specifically, CMS integrates F_{st} with iHS and XP-EHH, as well as two new site frequency spectra-based tests that the authors developed: ΔDAF tests for derived alleles that are at high frequency relative to other populations, and ΔiHH measures the absolute, rather than relative, length of the haplotype.

MORE COMPLEX MODELS OF SELECTION

While the sweep model has been a useful approach for identifying evidence of selection in diverse species, many selective events in humans and other organisms may not adhere to this model, and devising new tests to identify different forms of sweeps continues to be an area of active research (56, 103). In the selective sweep model, a novel allele at a single locus immediately confers a fitness benefit. Two ways to update the model, then, are to delay the

fitness benefit and to allow for multiple loci. In what follows, we discuss these two possibilities. We then turn our attention to ways that existent tests have been modified to identify different targets of selection.

Selection on Standing Variation and Soft Sweeps

Because mutations happen randomly, and not in response to specific selective pressures, alleles may arise at a time when they are not immediately beneficial. Such neutral alleles might reach a moderate frequency within the population simply as the result of genetic drift. If environmental pressures later change to make such a variant beneficial, the scenario is termed selection on standing variation. Notably, a standing variant in the EDA signaling pathway present in seawater fish has been shown to be under positive selection in freshwater stickleback fish. The variant, which is largely hidden in the heterozygous state in seawater populations, has emerged to cause loss of scales in multiple distinct freshwater populations (23).

Selection on standing variation is likely to occur in two scenarios: when the selection coefficient and mutation rate are both high, or when the selection coefficient is weak (92). This latter possibility suggests a potential application to complex organisms like humans in particular. Selection on standing variation imprints the genome in ways that are comparable to selection on novel variants (8), but can be more subtle, and therefore more difficult to detect. For example, LD between the standing variant and its neighbors will persist as in a hard sweep; compared to a hard sweep, however, the resulting trough in diversity will be shallower, owing to the fact that the standing variant will have time to recombine and associate with different haplotype backgrounds (105). This fact will also distort the frequency spectrum in a distinctive manner: compared to a hard sweep, selection on standing variation will create a greater number of linked neutral sites that have alleles at intermediate frequency (105). As the distinction between signatures of hard sweeps and selection on standing variation may be subtle, Peter et al. offer an Approximate Bayesian Computation (ABC) framework for distinguishing standing variants from de novo mutations (97).

A special instance of selection on standing variation occurs when the standing variant (or another allele that performs the same biological function) appears on multiple distinguishable haplotype backgrounds – for example, as a result of recurrent mutation or migration. This phenomenon is called a ‘soft sweep’ (55, 92, 93). Although the term ‘soft sweep’ is sometimes mistakenly used to indicate selection on standing variation more broadly, the two should be distinguished, as the selective signature that these trends leave, and consequently the methods developed to detect them, will differ (102).

Through computational simulations, Hermisson and Pennings demonstrated that the signature of a soft sweep should be in many ways comparable to that of a hard sweep (93). While frequency-based methods do not have predictive power for soft sweeps – owing to the fact that soft sweeps may involve an arbitrary number of distinct haplotypes – LD-based methods are able to detect the signatures of soft sweeps, albeit with diminished power. Similar to a hard sweep, the locus under selection is situated at the bottom of a trough of genetic diversity. These results suggest that computational methods to identify soft sweeps are within reach; it remains for researchers to fine-tune current LD-based methods to detect them.

Polygenic Networks and Ecological Methods

All of the methods discussed thus far assume that selection acts on one or a few sites at a time. However, given the known importance of polygenic networks and of epistatic interactions, researchers have suggested that selection may more often act on multiple sites in tandem, causing coordinated and distributed shifts in allele frequencies (53, 102).

One way to identify polygenic groups under selection is to incorporate ecological information. By binning related populations according to presumably relevant variables (e.g. habitat, climate, mode of sustenance, etc.) one can seek shifts in allele frequency shared across ecologically similar populations. Joost et al. formalize this approach as the 'Spatial Analysis Method' (SAM), using multiple univariate logistic regressions to test for association between allele frequencies and environmental variables (64). Jones et al. use a similar approach in their comparison of marine and freshwater sticklebacks from globally distributed populations to identify loci consistently associated with habitat (63), and Hancock et al. perform a similar analysis to identify ecologically relevant loci in humans (53).

An important limitation of ecological approaches is their reliance on user-specified variables (102). These methods run the risk of being biased by the information put in or left out. Polygenic selection can be detected without the risk of this bias by examining shared functional sets like quantitative trait loci (QTLs), where multiple genetic regions contribute to a single trait. Selection acting on a network of QTLs can be inferred based on a significant bias in their directionality – that is, the tendency of a locus to either amplify or lessen the magnitude of the phenotype (90). Whereas under neutrality, the distribution of positive or negative QTLs may be random, an overrepresentation of one or the other type of loci within a lineage is suggestive of selection. Fraser et al. developed a framework in which this test can be applied in a genome-wide scan, focusing on regulatory elements (i.e. expression QTLs, or eQTLs) in mice (37). Similarly, Simonson et al. performed a genome-wide scan with attention to genic networks known to be involved in oxygen-carry capacity to reveal adaptation to high altitudes in a Tibetan population (116).

Alternative Targets of Selection

Most natural selection studies to date have focused on genetic changes at the single nucleotide level, primarily because they have been the most accessible from a technological standpoint, through advances in protein analysis and SNP genotyping. Given their mutation mechanism, which typically creates simple biallelic changes of unique origin, they also can be more easily incorporated into statistical tests for selection. The focus on SNPs has been further motivated by that fact that point mutations are not only frequent generators of novel alleles for selection to act on themselves, but can also tag nearby hidden variation due to LD.

Many other genetic alteration that affects an organism's phenotype may be subject to selection, including copy number variants (CNVs) (113), microsatellites (46), chromosomal rearrangements (e.g. indels, inversions, translocations) (35), polygenic networks (discussed above), and epigenetic annotations (125). One of the first elucidated examples of selection was the thalassemias, CNVs of alpha and beta-globin genes, that, along with sickle cell anemia, confer resistance to malaria (6, 135). More recently, increased CNV counts of the gene for amylase has also been demonstrated to be associated with diets containing larger amounts of

starch (95). Another example is a major inversion on chromosome 17 in humans that was shown to be associated with greater reproductive success in an Icelandic population (119) and contains population genetic evidence of positive selection. Structural variants (SV) like CNVs and inversions, are often subject to negative selection (especially those that may cause frameshifts in protein-coding regions) (75), or can lead to relaxed evolutionary constraint through gene duplication (69). The many tests for selection described above may be applied to SVs, although the broad diversity of variants under the umbrella term 'SVs' and the large effects they can have on genomic architecture makes the systematic detection of selected variants challenging (61).

The recent discovery that certain epigenetic arrangements are heritable across many generations also raises the possibility of selection acting on the epigenome (29). Such neo-Lamarckian selection has been detected in orchids using SAM (91). It remains to be clarified to what extent such modes of selection are prevalent, but is an area of active interest.

CHALLENGES IN APPLYING STATISTICAL TESTS FOR SELECTION

While each approach has its own particular strengths and limitations, there are a number of challenges that are shared among these tests, particularly in the interpretation of their significance.

While a neutrality test may allow rejection of the null hypothesis, there are many possible explanations besides selection for the genomic results observed. Demographic events (e.g. migration, expansions, bottlenecks) can often create selection-mimicking signals, for example. Historically, most studies have aimed to rule out this possibility by comparing locus-specific to genome-wide data, as demographic events are understood to affect the genome in its totality, whereas selection acts in a more targeted manner (17). In recent years, however, some have questioned this 'outlier approach,' arguing that if selection is pervasive (as in *Drosophila*, for example: (73)) then distributed patterns of genetic hitchhiking would be misinterpreted as reflecting demographic events (47). More generally, the recognition that the effects of selection and demography may be interconnected have led some to adopt other approaches, such as explicitly estimating demographic parameters, including population structure, through various computational frameworks and incorporating these into subsequent analyses (e.g. (31); for review, see (74)). Another, related issue is that false-positives can be produced when tests implicate neutral variants in strong LD with a causal allele (122).

Even when these confounding effects can be ruled out, the interpretation of selection may not be straightforward. For example, rate-based tests implicate regions where evolutionary change has been accelerated: this may be due to positive selection of novel variants, but the relaxation of selective constraint (i.e., of purifying selection) over a region may have the same effect. Distinguishing between these possibilities involves case-by-case analysis. In a study of the evolution of CNVs in humans, for example, Nguyen et al. ruled out positive selection in regions where they observed an inverse relationship between rates of change and rates of recombination (83). More generally, however, functional analysis of candidate regions can help adjudicate between these two possibilities: if the derived variant has no potentially fitness-enhancing variation of function, relative to the ancestral, then the relaxation of selective constraint is the more likely explanation.

Another recurrent challenge for researchers is accounting for systematic biases that may be present in genomic data. The majority of selection studies to date have utilized SNP data, which is collected using genotyping arrays designed to detect known polymorphisms. The practical limitations of SNP discovery protocols means that low-frequency alleles may go undetected, in which case they will be excluded from these arrays. These arrays can therefore generate data that may be unrepresentative of the full extent of genetic diversity, a phenomenon known as ascertainment bias (22). This sampling of the data can artificially distort allele frequency measures as well as derivative statistics including LD. When the SNP discovery protocol is known, statistical measures can be taken to counteract the effect of ascertainment bias (86, 87, 106). In addition, genotyping assays that incorporate variable intensity oligonucleotide (VINO) probes can be used to mitigate the number of polymorphisms overlooked due to ascertainment bias (25).

Another salient issue for researchers investigating natural selection, particularly for those studying it in humans, is the potential for misinterpretation of results and their societal significance. By attending to linguistic subtleties and employing caution in disseminating results, researchers can prevent unethical application of evolutionary research (128).

FROM GENOME SCANS TO EVOLUTIONARY HYPOTHESES

The ultimate validation of genomic metrics of selection is the demonstration that putative selective variants have phenotypic effects. Functional understanding of candidate region begins with fine-mapping that region so as to localize the signal. Until recently, localizing signals of selection was a major challenge, but new composite methods and full genome sequence data provide stronger resolution that can allow researchers to identify tractable regions for functional scrutiny (44).

Once individual alleles have been identified for experimentation, researchers can measure the effect of said alleles as compared to their wildtype analogues. Genomic annotation can be informative for experimental design by suggesting the most probable types of traits that a variant may affect, or the types of cells where a variant is most commonly expressed.

Phenotypic screening may then proceed through an association study of various traits in the organism in question, for example, although background genetic variation can introduce noise into the data. To correct for this, researchers may instead compare the derived and ancestral variants against the same genetic background introduced into a cell line in vitro, or model organisms in vivo. Even in such situations, however, the possibility that a variant will have pleiotropic effects makes it difficult to discern whether a functional follow-up study correctly identifies the selective significance of the variant in question (7). While exhaustive phenotype screens are not possible, researchers can bolster the strength of their evidence by screening through as comprehensive a list of possible effects as possible. For example, Enard et al. introduced two human-specific amino acid substitutions in the *FOXP2* gene into mice and screened approximately three hundred traits, ultimately finding that only a small fraction of these (those involving the structure and function of cortico-basal ganglia circuits) were significantly different between humanized and wildtype mice (27).

Creating a case for selection necessitates a combination of genomic and functional evidence. With the availability of large population genetic datasets, statistical methods to interpret that data, and increasingly sophisticated technologies for transgenesis and other functional methods, researchers are moving into a new era of natural selection studies, in which both the causes and effects of changes to the genomes of humans and other organisms can be modeled and understood.

SUMMARY POINTS

The development of genotyping and sequencing technologies has allowed for the full realization and application of methods to investigate selection based on theory from the fields of comparative genomics and population genetics.

Methods to detect selection in the genome may be categorized by their effective time scale (i.e. macro- vs. micro-evolutionary) as well as by the types of data they utilize (i.e. interspecies divergence data, intraspecies diversity data, or a combination of these).

Tests to detect selection at the macroevolutionary level make interspecific comparisons, often aided by phylogenetic considerations, of the rates of change at the nucleotide level, and look for genetic regions in species that have experienced accelerated change.

Tests for microevolutionary selection come in a broad range of formats, but often aim to detect regions of reduced genetic diversity, which is indicative of a selective sweep. Other tests compare populations where selection is or is not hypothesized to be at play, and measure the extent of differentiation between them. Combining tests of multiple types can increase power and resolution.

An active area of research is developing tests for modes of selection that do not adhere to the selective sweep model. Among these are polygenic selection and soft sweeps.

Genomic evidence for selection is suggestive, not conclusive. A combination of genomic and functional evidence constitute the current standard for the field.

FUTURE ISSUES

How can we make tests for alternative selective modes (soft sweeps, polygenic selection, etc.) more robust?

How can we accurately quantify the prevalence of selection, and the relative contribution of drift, in humans and other organisms?

Can we develop high-throughput assays for functional analysis and validation of candidate variants?

LITERATURE CITED

1. Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19(5):711–722.
2. Akey JM, Zhang G, Zhang K, Jin L, and Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12(12):1805–1814.
3. Albrechtsen A, Moltke I, and Nielsen R. 2010. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186(1):295–308.
4. Allison AC. 1954. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *Br Med J* 1(4857):290–294.
5. Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, et al. 2009. Targets of balancing selection in the human genome. *Mol. Biol. Evol.* 26(12):2755–2764.
6. Barrai I, Rosito A, Cappellozza G, Cristofori G, Vullo C, Scapoli C, and Barbujani G. 1984. Beta-thalassemia in the Po Delta: selection, geography, and population structure. *Am J Hum Genet* 36(5):1121–1134.
7. Barrett RDH and Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* 12(11):767–780.
8. Barrett RDH and Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol. Evol. (Amst.)* 23(1):38–44.
9. Beaumont MA and Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13(4):969–980.

10. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, and Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74(6):1111–1120.
11. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, and Sancristobal M. 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186(1):241–262.
12. Boucher CAB, O'Sullivan E, Mulder JW, Ramautarsing C, Kellam P, Darby G, Lange JMA, Goudsmit J, and Larder BA. 1992. Ordered Appearance of Zidovudine Resistance Mutations during Treatment of 18 Human Immunodeficiency Virus-Positive Subjects. *J Infect Dis.* 165(1):105–110.
13. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
14. Burbano HA, Green RE, Maricic T, Lalueza-Fox C, De la Rasilla M, Rosas A, Kelso J, Pollard KS, Lachmann M, and Pääbo S. 2012. Analysis of human accelerated DNA regions using archaic hominin genomes. *PLoS ONE* 7(3):e32877.
15. Cai Z, Camp NJ, Cannon-Albright L, and Thomas A. 2011. Identification of regions of positive selection using Shared Genomic Segment analysis. *Eur. J. Hum. Genet.* 19(6):667–671.
16. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, and Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15(11):1553–1565.

17. Cavalli-Sforza LL. 1966. Population Structure and Human Evolution. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164(995):362–379.

18. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, et al. 2010. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* 327(5963):302–305.

19. Charlesworth B, Morgan MT, and Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

20. Charlesworth D. 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genet* 2(4):e64.

21. Chen H, Patterson N, and Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20(3):393–402.

22. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, and Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15(11):1496–1502.

23. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G Jr, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, and Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* 307(5717):1928–1933.

24. Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, Brown DL, et al. 2009. Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. *J. Invest. Dermatol.* 129(2):392–405.

25. Didion JP, Yang H, Sheppard K, Fu C-P, McMillan L, De Villena FP-M, and Churchill GA. 2012. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 13:34.

26. Egea R, Casillas S, and Barbadilla A. 2008. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* 36(Web Server issue):W157–162.

27. Enard W, Gehre S, Hammerschmidt K, Hölter SM, Blass T, Somel M, Brückner MK, et al. 2009. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* 137(5):961–971.

28. Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, and Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418(6900):869–872.

29. Eva Jablonka B and Raz G. 2009. Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution. *The Quarterly Review of Biology* 84(2):131–176.

30. Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3(1):87–112.

31. Excoffier L, Hofer T, and Foll M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* 103(4):285–298.

32. Fariello MI, Boitard S, Naya H, San Cristobal M, and Servin B. 2013. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics*:

33. Fay JC and Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405–1413.

34. Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* 27(9):343–349.

35. Feuk L, Carson AR, and Scherer SW. 2006. Structural variation in the human genome. *Nature Reviews Genetics* 7(2):85–97.

36. Foll M and Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180(2):977–993.

37. Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, and Schadt EE. 2011. Systematic detection of polygenic cis-regulatory evolution. *PLoS Genet.* 7(3):e1002023.

38. Fu YX and Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.

39. Fu Y-X. 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics* 147(2):915–925.

40. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, and Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7(11):e1002355.

41. Ge R-L, Simonson TS, Cooksey RC, Tanna U, Qin G, Huff CD, Witherspoon DJ, et al. 2012. Metabolic insight into mechanisms of high-altitude adaptation in Tibetans. *Mol. Genet. Metab.* 106(2):244–247.

42. Gould S. 1978. Sociobiology: The art of storytelling. *New Scientist* 80(1129):530–33.

43. Graur D and Li W-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer.

44. Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327(5967):883–886.

45. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, et al. 2013. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell* 152(4):703–713.

46. Haasl RJ and Payseur BA. 2012. Microsatellites as Targets of Natural Selection. *Mol. Biol. Evol.*:

47. Hahn MW. 2008. Toward a Selection Theory of Molecular Evolution. *Evolution* 62(2):255–265.

48. Haldane JBS. 2006. Disease and Evolution. In *Malaria: Genetic and Evolutionary Aspects*, pp. 175–187. Springer US.

49. Hamblin MT and Di Rienzo A. 2000. Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. *The American Journal of Human Genetics* 66(5):1669–1679.

50. Han L and Abney M. 2012. Using identity by descent estimation with dense genotype data to detect positive selection. *Eur. J. Hum. Genet.*:

51. Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, and Kwiatkowski DP. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* 78(1):153–159.

52. Hanchard N, Elzein A, Trafford C, Rockett K, Pinder M, Jallow M, Harding R, Kwiatkowski D, and McKenzie C. 2007. Classical sickle beta-globin haplotypes exhibit

a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genetics* 8(1):52.

53. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, et al. 2010. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc. Natl. Acad. Sci. U.S.A.* 107 Suppl 2:8924–8930.
54. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, and Di Rienzo A. 2008. Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. *PLoS Genet* 4(2):e32.
55. Hermisson J and Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169(4):2335–2352.
56. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, and Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331(6019):920–924.
57. Holsinger KE and Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* 10(9):639–650.
58. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, et al. 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat Genet* 40(8):987–993.
59. Hudson RR, Kreitman M, and Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1):153–159.
60. Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18(9):486.

61. Iskow RC, Gokcumen O, and Lee C. 2012. Exploring the role of copy number variants in human adaptation. *Trends Genet.* 28(6):245–257.

62. Jensen JD, Kim Y, DuMont VB, Aquadro CF, and Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170(3):1401–1410.

63. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55–61.

64. Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, and Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16(18):3955–3969.

65. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, et al. 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* 152(4):691–702.

66. Kim Y and Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167(3):1513–1524.

67. Kim Y and Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160(2):765–777.

68. Kimura M. 1985. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

69. Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* 279(1749):5048–5057.

70. Kreitman M and Akashi H. 1995. Molecular Evidence for Natural Selection. *Annual Review of Ecology and Systematics* 26(1):403–422.

71. Kwiatkowski DP. 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *Am J Hum Genet* 77(2):171–192.

72. Lewontin RC and Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.

73. Li H and Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genet.* 2(10):e166.

74. Li J, Li H, Jakobsson M, Li S, Sjödin P, and Lascoux M. 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol. Ecol.* 21(1):28–44.

75. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, et al. 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature Biotechnology* 29(8):723–730.

76. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.

77. McDonald JH and Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328):652–654.

78. Meirmans PG and Hedrick PW. 2011. Assessing population structure: FST and related measures. *Molecular Ecology Resources* 11(1):5–18.

79. Miller LH, Mason SJ, Clyde DF, and McGinniss MH. 1976. The Resistance Factor to Plasmodium vivax in Blacks. *New England Journal of Medicine* 295(6):302–304.

80. Mullen LM, Vignieri SN, Gore JA, and Hoekstra HE. 2009. Adaptive basis of geographic variation: genetic, phenotypic and environmental differences among beach mouse populations. *Proc Biol Sci* 276(1674):3809–3818.

81. Nei M and Maruyama T. 1975. Letters to the editors: Lewontin-Krakauer test for neutral genes. *Genetics* 80(2):395.

82. Nei M, Suzuki Y, and Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* 11:265–289.

83. Nguyen D-Q, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, and Ponting CP. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res.* 18(11):1711–1723.

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

85. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, et al. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19(5):838–849.

86. Nielsen R, Hubisz MJ, and Clark AG. 2004. Reconstituting the Frequency Spectrum of Ascertained Single-Nucleotide Polymorphism Data. *Genetics* 168(4):2373–2382.

87. Nielsen R and Signorovitch J. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology* 63(3):245–255.

88. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, and Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15(11):1566–1575.

89. Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, and Shriver MD. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* 24(3):710–722.

90. Orr HA. 1998. Testing Natural Selection vs. Genetic Drift in Phenotypic Evolution Using Quantitative Trait Locus Data. *Genetics* 149(4):2099–2104.

91. Paun O, Bateman RM, Fay MF, Hedrén M, Civeyrel L, and Chase MW. 2010. Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). *Mol. Biol. Evol.* 27(11):2465–2473.

92. Pennings PS and Hermisson J. 2006. Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* 23(5):1076–1084.

93. Pennings PS and Hermisson J. 2006. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* 2(12):e186.

94. Pérez-Morga D, Vanhollebeke B, Paturiaux-Hanocq F, Nolan DP, Lins L, Homblé F, Vanhamme L, et al. 2005. Apolipoprotein L-I Promotes Trypanosome Lysis by Forming Pores in Lysosomal Membranes. *Science* 309(5733):469–472.

95. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39(10):1256–1260.

96. Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, Michelini K, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22(4):602–610.

97. Peter BM, Huerta-Sanchez E, and Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8(10):e1003011.

98. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19(5):826–837.

99. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, et al. 2006. Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet* 2(10):e168.

100. Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443(7108):167–172.

101. Prabhakar S, Noonan JP, Pääbo S, and Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314(5800):786.

102. Pritchard JK, Pickrell JK, and Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20(4):R208–215.

103. Pritchard JK and Di Rienzo A. 2010. Adaptation - not by sweeps alone. *Nat. Rev. Genet.* 11(10):665–667.

104. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh S-D, True JR, and Carroll SB. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 440(7087):1050–1053.

105. Przeworski M, Coop G, and Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.

106. Ramírez-Soriano A and Nielsen R. 2009. Correcting Estimators of θ and Tajima's D for Ascertainment Biases Caused by the Single-Nucleotide Polymorphism Discovery Process. *Genetics* 181(2):701–710.

107. Riebler A, Held L, and Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178(3):1817–1829.

108. Romero IG, Ruvinsky I, and Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* 13(7):505–516.

109. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, and Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312(5780):1614–1620.

110. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832–837.

111. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913–918.

112. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, Lambert C, et al. 2012. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.* 13(1):R1.

113. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305(5683):525–528.

114. Shapiro BJ and Alm EJ. 2008. Comparing patterns of natural selection across species using selective signatures. *PLoS Genet.* 4(2):e23.

115. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, and Jones KW. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* 1(4):274–286.

116. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, et al. 2010. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science* 329(5987):72–75.

117. Smith JM and Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research* 23(01):23–35.

118. Smith NGC and Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415(6875):1022–1024.

119. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, et al. 2005. A common inversion under selection in Europeans. *Nature Genetics* 37(2):129–137.

120. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

121. Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135(2):599–607.

122. Teshima KM, Coop G, and Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16(6):702–712.

123. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39(1):31–40.

124. Tsetskhladze ZR, Canfield VA, Ang KC, Wentzel SM, Reid KP, Berg AS, Johnson SL, Kawakami K, and Cheng KC. 2012. Functional Assessment of Human Coding Mutations Affecting Skin Pigmentation Using Zebrafish. *PLoS One* 7(10):

125. Turner BM. 2009. Epigenetic responses to environmental change and their evolutionary implications. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 364(1534):3403–3418.

126. Vanhamme L, Paturiaux-Hanocq F, Poelvoorde P, Nolan DP, Lins L, Abbeele JVD, Pays A, et al. 2003. Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* 422(6927):83–87.

127. Vitalis R, Dawson K, and Boursot P. 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158(4):1811–1823.

128. Vitti JJ, Cho MK, Tishkoff SA, and Sabeti PC. 2012. Human evolutionary genomics: ethical and interpretive issues. *Trends Genet.* 28(3):137–145.

129. Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72.

130. Wang ET, Kodama G, Baldi P, and Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl. Acad. Sci. U.S.A.* 103(1):135–140.

131. Watterson GA. 1978. The homozygosity test of neutrality. *Genetics* 88(2):405–417.

132. Wiener P and Pong-Wong R. 2011. A regression-based approach to selection mapping. *J. Hered.* 102(3):294–305.

133. Wright SI and Charlesworth B. 2004. The HKA Test Revisited. *Genetics* 168(2):1071–1076.

134. Yang and Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol. (Amst.)* 15(12):496–503.

135. Yokoyama S. 1983. SELECTION FOR THE α-THALASSEMIA GENES. *Genetics* 103(1):143–148.

136. Zeng K, Fu Y-X, Shi S, and Wu C-I. 2006. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics* 174(3):1431–1439.

137. Zeng K, Shi S, and Wu C-I. 2007. Compound tests for the detection of hitchhiking under positive selection. *Mol. Biol. Evol.* 24(8):1898–1908.

138. Zhai W, Nielsen R, and Slatkin M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Mol. Biol. Evol.* 26(2):273–283.

139. Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, et al. 2006. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics* 22(17):2122–2128.

140. Zhuang Z, Gusev A, Cho J, and Pe'er I. 2012. Detecting Identity by Descent and Homozygosity Mapping in Whole-Exome Sequencing Data. *PLoS ONE* 7(10):e47618.

ACRONYMS AND DEFINITIONS

SNP: Single nucleotide polymorphism; individual base pair sites in the genome of an organism where multiple variants exist.
Homologue: Traits or sequences that are similar in disparate groups due to common ancestry.
Non-synonymous: A change in the protein-coding region of a gene that alters the amino acid encoded.
Synonymous: A change in the protein-coding region of a gene that does not change the amino acid encoded.
Heterozygote advantage: A trend where the fitness of a heterozygote is greater than that of homozygotes. Also referred to as **overdominance**.
Frequency-dependent selection: A trend where the fitness of a given genotype is correlated with its prevalence in the population (e.g., if an allele is advantageous when it is rare).
Codominance: Condition in which multiple alleles are dominant; the heterozygote expresses phenotypes associated with both alleles.
Microsatellite: Genetic regions consisting of repeating sequences of two to six base pairs. Also referred to as Short Tandem Repeats (**STRs**) or Simple Sequence Repeats (**SSRs**).
Derived: An allele that arises by a novel mutation and has not achieved fixation in a population (as contrasted with an ancestral allele).
Ancestral: An allele that was pre-existing in a population, from which a derived allele may arise.
Genetic drift: Change in allele frequencies over time due to chance (random sampling).
Linkage disequilibrium (LD): Tendency of certain variants on the same chromosome to be co-inherited at above chance rates (e.g. due to selection or founder effects).
Codon bias: The tendency of an organism's genome to more commonly have a certain codon for a given amino acid than any of its synonymous counterparts.
Singleton: An allele that appears only once in a sample set.
Pleiotropy: A trend where one genotype affects multiple phenotypes.
Structural variants (SVs): alterations in the genome affecting relatively large chromosomal regions, including deletions and insertions (indels), translocations, inversions, and duplications.
Copy number variants (CNVs): A form of structural variant in which multiple copies of a genetic region exist.
Epigenome: Annotations to the DNA molecule that alter patterns of gene expression, but do not change the sequence.

SIDEBARS:

SELECTION AND NEUTRALITY
Kimura's neutral theory of molecular evolution held that the vast majority of genetic change is attributable to genetic drift, rather than Darwinian selection (68). As researchers began to develop methods to distinguish neutral from adaptive change in the genome, however, many came to reject the stronger versions of the neutral theory, and turned their attention towards quantifying the relative contributions of drift and selection to molecular evolution (118) (70)

Importantly, however, the neutral theory enabled the development of tests for selection by assisting in the sophistication of models of genetic drift. In many tests for selection (neutrality

tests), researchers compare empirical data against data generated by simulations of drift, which serve as a null hypothesis. Other neutrality tests may use background rates of change, inferred from whole-genome analyses, to furnish a null hypothesis.

In this review, we focus our discussion on the wide range of tests for selection that have been developed and their applications. We discuss the importance for researchers conducting natural selection studies to examine alternate hypotheses and to pursue functional characterization. Readers interested in the selectionist-neutralist debate are encouraged to consult recent reviews on the subject (7, 34, 82).

TABLES:

Table 1: AN OVERVIEW OF COMMON APPROACHES FOR DETECTING SELECTION

	Approach	**Intuition**	**Representative Tests**	**Refs.**
METHODS FOR MACROEVOLUTION	Gene-based methods	Synonymous substitutions are (assumed to be) selectively neutral - thus, they tell us about the 'background rate' of evolution. We can compare the rate of non-synonymous substitutions to infer how selection is acting.	K_a/K_s (also known as d_N/d_S or ω)	(43, 60)
			McDonald Kreitman Test (MKT)	(26, 77)
	Other rate-based methods	Levels of polymorphism and divergence should be correlated (since both are primarily functions of the mutation rate) unless selection causes one to exceed the other.	Hudson-Kreitman-Aguadé Test (HKA)	(59, 133)
		Regions that undergo accelerated change in one lineage, but are relatively conserved in related lineages, are probable candidate for selection.	Identification of Accelerated Regions	(14, 76, 99, 101, 114)
METHODS FOR MICROEVOLUTION	Frequency-based methods	Selective sweeps bring a genetic region to high prevalence in a population, bringing nearby linked derived alleles (i.e. mutations) to high frequency. From this background, new alleles arise but are initially at low frequency (hence many rare alleles)	Ewens-Watterson Test	(30, 131)
			Tajima's D and derivatives	(38, 39, 120, 121)
			Fay and Wu's H	(33)
	Linkage disequilibrium-based methods	Selective sweeps bring a genetic region to high prevalence in a population, including the causal variant and its neighbors, which define a haplotype. This haplotype persists in the population until recombination breaks it down.	Long Range Haplotype Test (LRH)	(110, 139)
			Integrated Haplotype Score (iHS)	(129)
			Cross-Population Extended Haplotype Homozygosity (XP-EHH)	(111)

			Linkage Disequilibrium Decay (LDD)	(130)
			Identity-by-Descent (IBD) Analyses	(15, 50)
	Population differentiation -based methods	Selection acting on an allele in one population, but not in another, will create a marked difference in the frequency of that allele between the two populations. This effect of differentiation will stand out against the differentiation between populations with respect to neutral (i.e., non-selected) alleles.	Lewontin-Krakauer Test	(11, 31, 72, 127)
			Locus-Specific Branch Length (LSBL)	(115)
			hapFLK	(32)
	Composite methods	Combining test scores for multiple regions can reduce the rate of false positives.	Composite Likelihood Ratio (CLR)	(66, 67, 85, 88)
			Cross-Population Composite Likelihood Ratio (XP-CLR)	(21)
		Combining multiple tests for the same region can improve resolution and distinguish causal variants. Different test can provide complementary information.	DH Test	(136, 137)
			Composite of Multiple Signals (CMS)	(44)

Table 2: USING SELECTION SCANS TO STUDY HUMAN EVOLUTION

Gene under selection	**Population(s)**	**Genomic evidence**	**Functional evidence**	**Putative Adaptive Role**	**Refs**
FOXP2	All (selection predates out-of-Africa migration)	Accelerated evolution in coding region, D, H	Mouse transgenic	Affects development of cortico-basal ganglia circuits; thought to be	(27, 28)

				involved in mechanics of speech	
LCT	Northern Europeans, East Africans (pastoralist societies)	EHH, iHS; F_{st} analysis	Human association study, in vitro lactase expression assay	Confers lactase persistence; allows digestion of lactose into adulthood	(10, 123)
EDAR	East Asians	CMS	Human association study, mouse transgenic	Affects patterns of hair and sweat glands	(44, 65)
TLR5	West Africans	CMS	In vitro assay of NF- κ B pathway activation	Modulates immune response to bacterial flagellin	(44, 45)
DARC	African populations in malaria-endemic regions	F_{st}	Human association study	Heterozygosis reduces susceptibility to malaria	(49, 79)
APOL1	African populations in trypanosome-endemic regions	CMS	In vitro assay of response to trypanosome invasion	Modulates susceptibility to trypanosomiasis	(44, 94, 126)
HBB	African populations in malaria-endemic regions	LRH	Human association study	Heterozygosis reduces susceptibility to malaria	(4, 71, 110)
EPAS1, EGLN, et al.	Tibetans	iHS, XP-EHH	Human association study	Selected variants decrease hemoglobin concentration and modulate hypoxia response	(41, 116)
SLC24A5, SLC45A2	Europeans	F_{st} analysis	Human association study, in vitro assay of melanocyte cultures, zebrafish	Decreases melanin pigmentation in skin	(24, 89, 124)

			transgenic		
CBARA1, VAV3, et al.	Ethiopian highland popoulations	LSBL, iHS, XP-EHH	Human association study	Selected variants decrease hemoglobin concentration and modulate hypoxia response	(112)

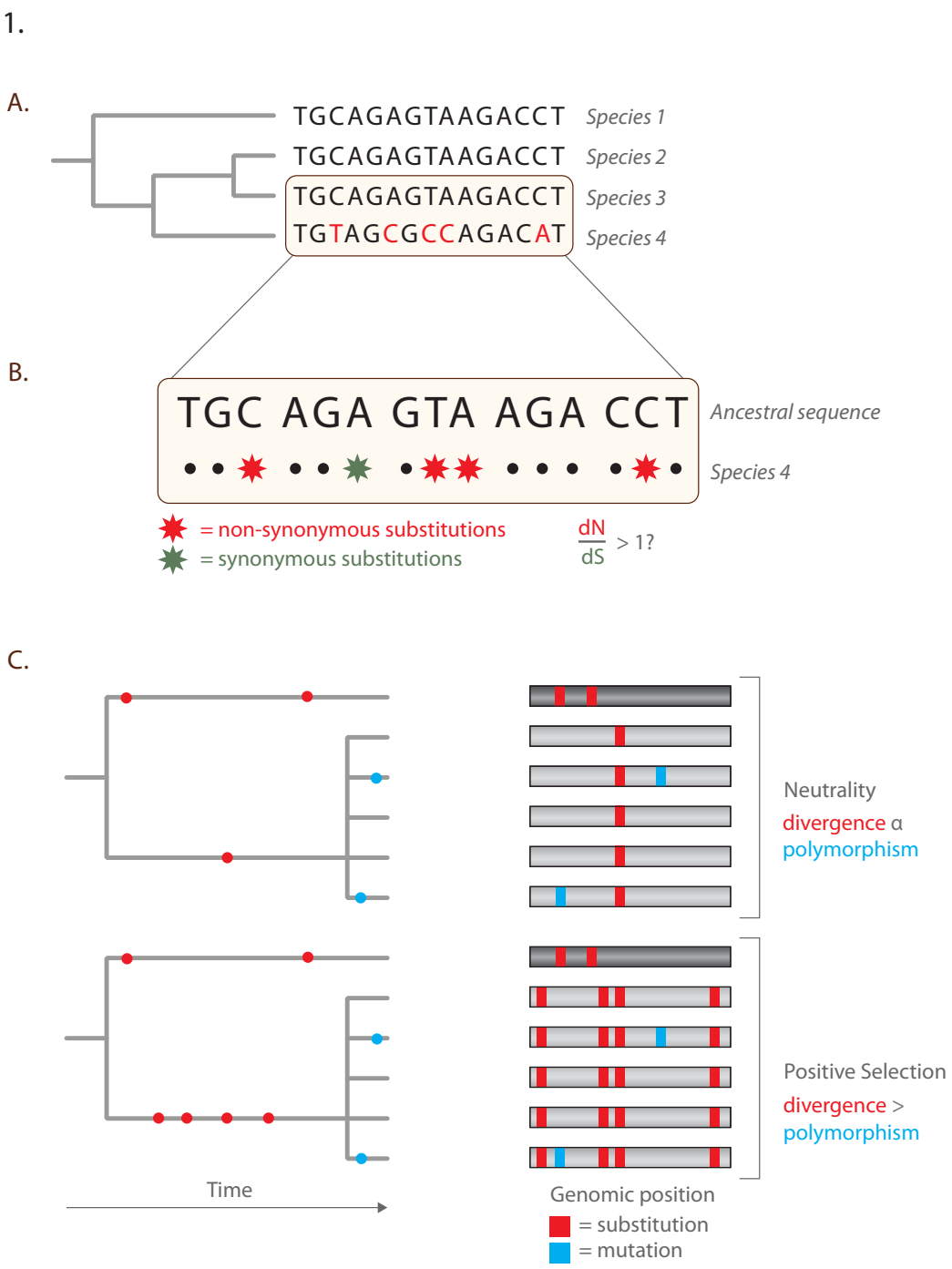

Figure 1: Methods for detecting selection at the macroevolutionary level
A) Traits that are conserved across many clades of a phylogeny, but that show extreme differentiation in one or a few lineages, are likely candidates for selection.
B) Metrics like Ka/Ks compare the rate of non-synonymous (i.e., amino acid-altering) substitutions in a lineage to the rate of synonymous substitutions, whichare assumed to be selectively neutral.
C) The McDonald-Kreitman Test and Hudson-Kreitman-Aguade Test hinge on the intuition that levels of interspecies divergence and intraspecies polymorphism are both governed by the mutation rate, and so will be correlated unless selection or some other force (e.g. fluctuations in population size) is at play.

2.

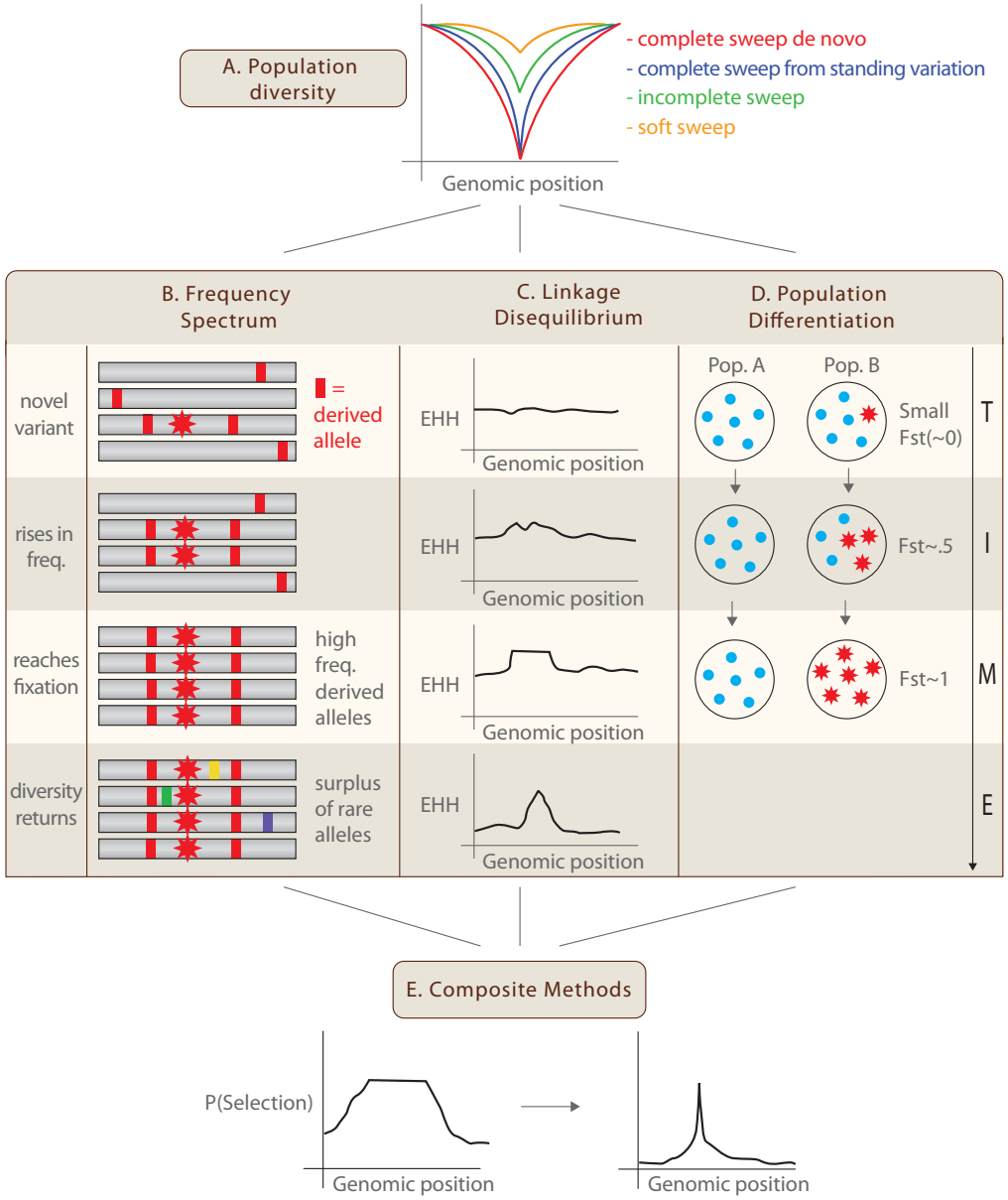

Figure 2: Methods for detecting selective sweeps at the microevolutionary level
A) Beneficial mutations will bring nearby 'hitchhiker' variants to high frequency, causing a population-wide reduction in diversity surrounding the selected locus. This trough in diversity may be shallower and/or narrower if the sweep is incomplete, or if the mutation is not subject to immediate selection (i.e., selection on standing variation or soft sweep).
B) A beneficial mutation will bring nearby derived alleles to high frequency. After the sweep is complete, novel mutations against a homogenous background will create a surplus of rare alleles.
C) A selective sweep will cause EHH to rise across the haplotype containing the selected allele. The plateau of high EHH will begin to break down when novel mutations and recombination gradually restore diversity to the population.
D) Differences in allele frequencies, reflecting the population-specific action of selection, will cause F_{st} between two populations to increase.
E) Composite methods that integrate information from multiple signals of selection can provide finer resolution and help pinpoint causal variants.

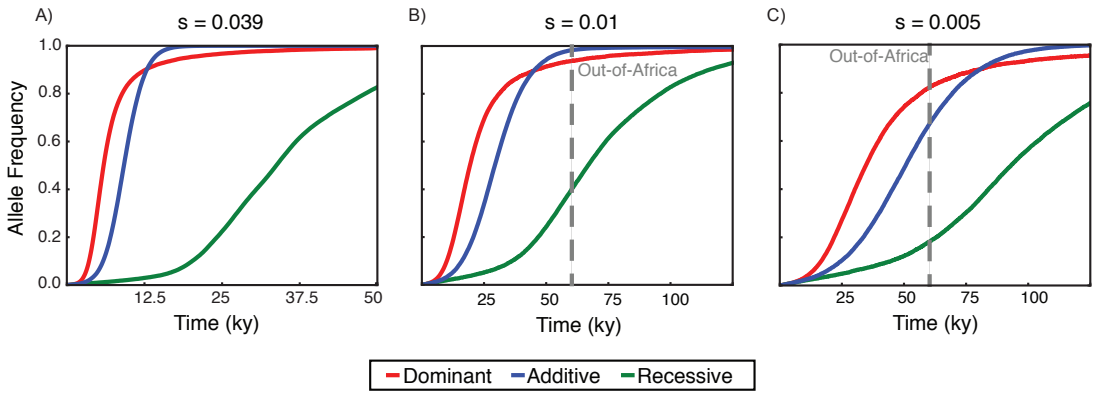

Figure 3: Trajectories of beneficial alleles with realistic selection coefficients simulated in human populations

The fate of a beneficial allele depends on many factors, including the strength of selection and the extent of the allele's phenotypic influence (i.e., whether it is dominant or recessive). Most alleles with realistic selection coefficients that have arisen since the Out-of-Africa migration are expected not to have yet reached fixation in their respective populations.