Int. J. Plant Sci. 171(9):1059–1071. 2010.

1058-5893/2010/17109-0012$15.00 DOI: 10.1086/656306

USING POPULATION GENOMICS TO DETECT SELECTION IN NATURAL POPULATIONS: KEY CONCEPTS AND METHODOLOGICAL CONSIDERATIONS

Paul A. Hohenlohe,[1,*,†] Patrick C. Phillips,* and William A. Cresko*

*Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, U.S.A.; and †Department of Zoology, Oregon State University, Corvallis, Oregon 97331, U.S.A.

Natural selection shapes patterns of genetic variation among individuals, populations, and species, and it does so differentially across genomes. The field of population genomics provides a comprehensive genome-scale view of the action of selection, even beyond traditional model organisms. However, even with nearly complete genomic sequence information, our ability to detect the signature of selection on specific genomic regions depends on choosing experimental and analytical tools appropriate to the biological situation. For example, processes that occur at different timescales, such as sorting of standing genetic variation, mutation-selection balance, or fixed interspecific divergence, have different consequences for genomic patterns of variation. Inappropriate experimental or analytical approaches may fail to detect even strong selection or falsely identify a signature of selection. Here we outline the conceptual framework of population genomics, relate genomic patterns of variation to evolutionary processes, and identify major biological factors to be considered in studies of selection. As data-gathering technology continues to advance, our ability to understand selection in natural populations will be limited more by conceptual and analytical weaknesses than by the amount of molecular data. Our aim is to bring critical biological considerations to the fore in population genomics research and to spur the development and application of analytical tools appropriate to diverse biological systems.

Keywords: coalescent, genome scan, natural selection, population genomics, selective sweep, sequence divergence.

Introduction

The ability to observe natural selection and detect its effects on specific genetic loci has been greatly advanced by studies of the entire genomes of organisms. Genomic techniques are widely used in model organisms such as humans and *Drosophila*, where a large library of resources (multiple reference genome sequences, massive SNP chips, pedigree info, inbred strains, etc.) lays the methodological foundation for detailed analysis (Akey 2009; Pritchard et al. 2010). However, next-generation sequencing technology (Shendure and Ji 2008) and a variety of array techniques (Perkel 2008) can be combined with minimal genomic resources, such as a reference genome sequence or an expressed-sequence tag database in the study organism or a related taxon, to allow genome scans for selection in nonmodel organisms. As a result, a myriad of statistical techniques for detecting the signature of selection has been developed (Pavlidis et al. 2008; Oleksyk et al. 2010). In this article, we focus on approaches for detecting selection in natural populations, using techniques that can be applied to both model and nonmodel organisms.

Tests for selection in population genomics are diverse (see box 1). These assays of genetic variation focus on a plethora of genetic patterns, from nucleotide diversity, allele frequency spectrum (AFS), haplotype structure, and linkage disequilibrium (LD) within and among populations to fixed DNA sequence divergence among related taxa. These tests detect the effect of different modes and timescales of selection under different scenarios of population structure. Consideration of these biological factors drives the experimental and analytical approach(es) that have the most power to detect selection in genomic studies. Our goal here is not to review the mechanics of statistical approaches (see Nielsen 2005; Pavlidis et al. 2008; Oleksyk et al. 2010) or delve deeply into population genomic theory. Instead we focus on the biological factors that should be considered in designing genomic experiments to test for selection and in choosing the appropriate analyses and the potential pitfalls that accompany each of them. While the theory connecting biological process to genomic pattern is complex, we seek an intuitive understanding of these connections that can inform the experimental design and analysis of genomic studies. Armed with such an understanding, researchers may take advantage of any prior information on the timescale and mode of selection, demographic fluctuations, population structure, or specific phenotypic traits under selection in the study organism.

We begin by describing the qualitative differences in evolutionary processes and patterns that are studied in population genomics compared to traditional population genetics and the implications of a population genomics perspective for understanding the effects of selection. Then we discuss major aspects of selection in natural populations, divided into categories of temporal, biological, and spatial, and the ways in which they affect patterns of variation along the genome.

[1] Author for correspondence; e-mail: hohenlo@uoregon.edu.

Manuscript received April 2010; revised manuscript received July 2010.

These factors lead to a rough decision tree for assessing which tests are most appropriate and have the most statistical power to detect selection in each particular case (fig. 1). An emerging conclusion in population genomics is that a combination of tests based on multiple aspects of genomic structure will be most able to separate the effects of selection from demography and other genetic processes (Grossman et al. 2010).

Natural Selection from a Population Genomics Perspective

Population genetics theory has provided a powerful theoretical framework for understanding the action of evolutionary forces—mutation, selection, gene flow, and genetic drift—on patterns of genetic variation (Fisher 1930; Wright 1978). As it developed over the past century, this theory has confronted empirical data with the emergence of each new technique for tracking alleles in populations: from discrete phenotypes to allozymes, microsatellites, and, ultimately, DNA sequence data. Empirical results have, in turn, spawned novel approaches for understanding the origin and maintenance of genetic variation, such as the neutral theory of molecular evolution (Kimura 1968) and the coalescent (Kingman 1982). The evidence of selection on single loci in natural populations can be seen in summary statistics (fig. 2*A*; see box 1) describing population differentiation (e.g., F_{ST}; Beaumont 2005), AFS (Tajima's *D*; Tajima 1989), or genetic sequence divergence between species (*dN*/*dS*; Nielsen and Yang 1998) at specific loci.

However, despite the robustness of classical population genetics, a limitation has been its focus on allelic variation at single genes (whether in the form of discrete alleles or nucleotide sequence variants). In the context of whole genomes, basic evolutionary forces such as natural selection act in concert with genome-scale processes—dominance and epistasis, linkage and recombination, gene and genome duplication—to produce the structure of genomic variation observed in natural populations. To be sure, population genetics theory has addressed these interactions among loci as well, viewing genes in a genome as discrete beads on a string. But a more

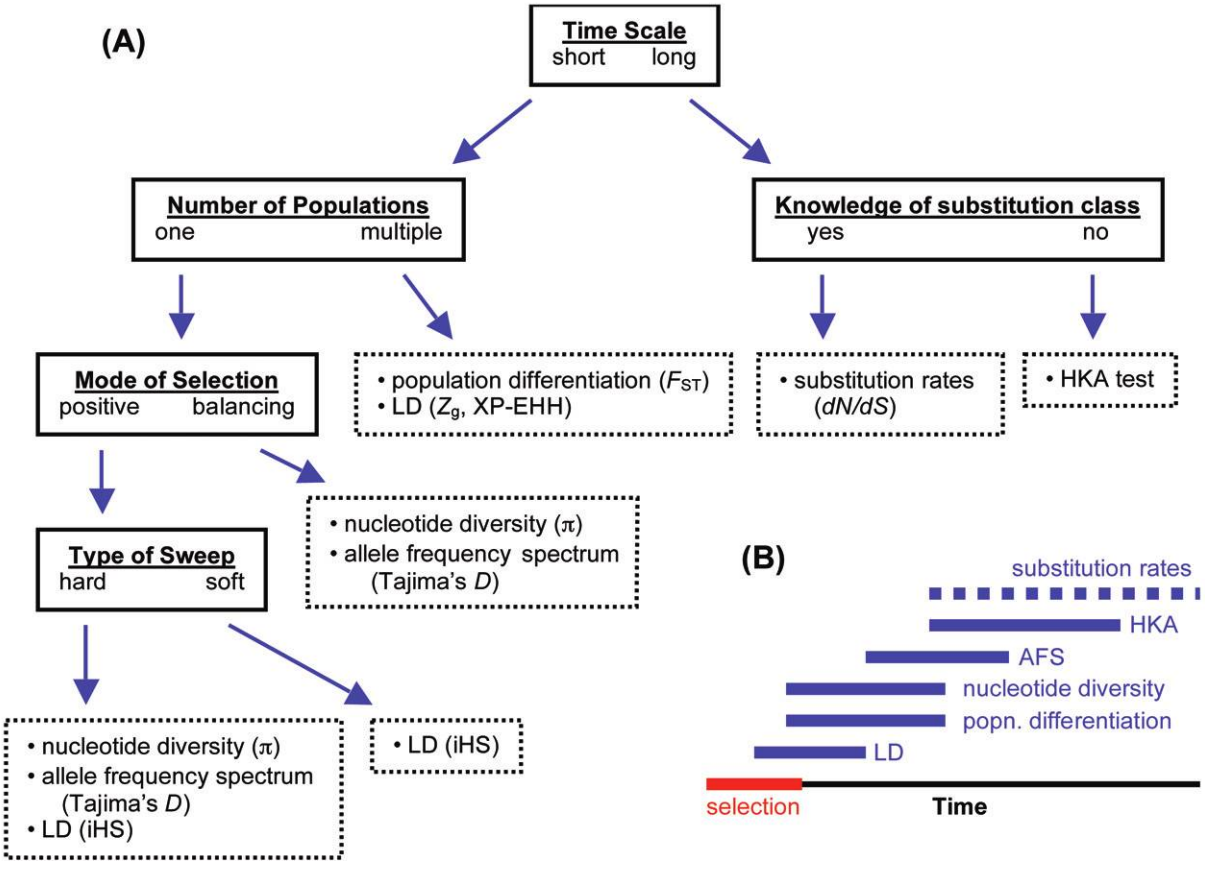

Fig. 1 *A*, Decision tree summarizing the major biological considerations in using population genomics to test for selection (solid outline) and the classes of statistical tests that are most appropriate for each case (dotted outline). See box 1 for descriptions of particular tests. *B*, Conceptual view of the timescale during which different classes of tests are best able to detect selection. A selective sweep is shown in red. Tests based on substitution rates (e.g., *dN*/*dS*) have a potentially long life span but require multiple amino acid substitutions. Time is in units of effective population size. Based on Hudson et al. (1987), Pennings and Hermisson (2006*b*), Sabeti et al. (2006), and Oleksyk et al. (2010; but note that these latter two references focused on applications to human populations).

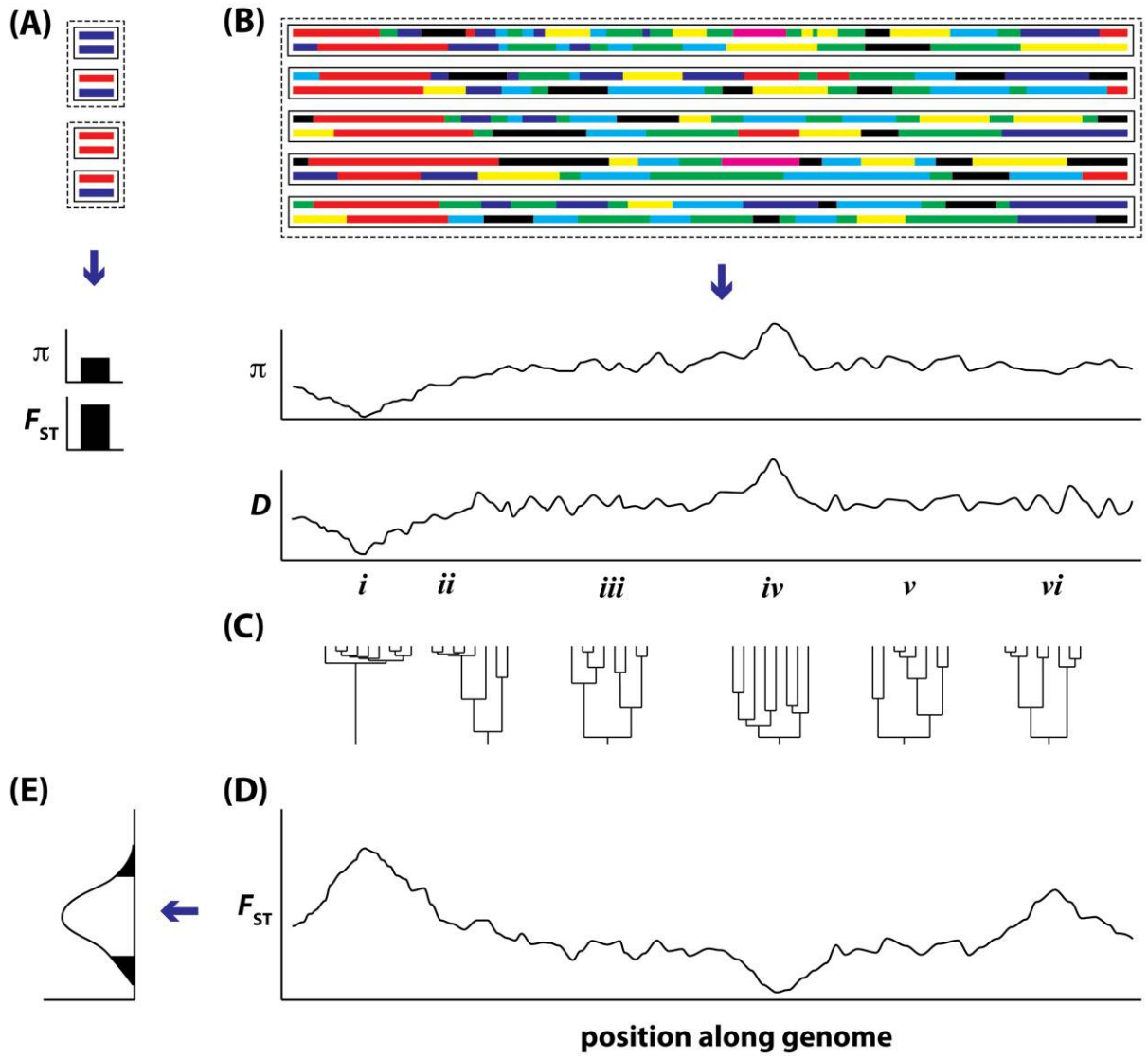

Fig. 2 A population genomics perspective. *A*, Traditional population genetics takes data on alleles (colored bars), grouped within individuals (solid boxes) and populations (dashed boxes), and calculates summary statistics to make inferences about evolution, such as nucleotide diversity (π) and population differentiation (F_{ST}). *B*, Population genomics takes data on haplotypes within a population and calculates summary statistics as continuous variables along the length of the genome, such as π and allele frequency spectrum (Tajima's *D*). The impact of different types of evolutionary processes leave different signatures in these distributions: *i*, hard selective sweep; *ii*, region linked to hard sweep; *iii*, neutral expectation; *iv*, balancing selection; *v*, neutral expectation; *vi*, soft sweep. *C*, The coalescent structure of ancestral relationships among alleles within a population also reflects these processes along the genome. *D*, Given these genomic processes within a population, statistics comparing genetic variation across populations, such as F_{ST}, can also indicate genomic patterns of selection. *E*, Collapsing the genomic distribution of a statistic into a frequency distribution provides an estimate of the genomewide average, allowing identification of statistically significant outliers (shaded regions).

informative view of a genome, perhaps, is as an axis along which statistics such as AFS and F_{ST} are continuously distributed variables (fig. 2*B*–2*D*). A critical feature of these genomic distributions is their spatial autocorrelation—correlation among measurements at neighboring genomic regions—reflecting LD among neighboring loci (Hahn 2006). The degree to which this autocorrelation itself changes along the genome is the result of selection and recombination, as well as other evolutionary forces. Inferring the evolutionary history of any single locus is complicated by the influence of its genomic neighbors. However, this genomic structure also opens the door to new tests of selection based specifically on statistics describing the local extent of LD, such as the integrated haplotype score (Voight et al. 2006) or cross-population extended haplotype homozygosity (Sabeti et al. 2007; see box 1). In certain biological situations, these tests are, in fact, the most powerful at detecting selection.

A useful conceptual approach in population genetics for connecting a range of demographic and evolutionary processes to patterns of genetic variation at single loci, especially

Box 1

Critical Population Genetic Concepts and Statistical Measures Used to Detect Selection in Population Genomics

Allele frequency spectrum (AFS): The distribution of frequencies across alleles in a sample. Tests based on AFS using DNA sequence data rely on a few related statistics, all of which are comparisons between estimates of the population genetic parameter $\theta = 4N\mu$. The statistics are calculated as the difference of two such estimates, normalized by the expected variance of the difference under a neutral model, so that values below −2 or greater than 2 roughly exceed the 95% confidence limits about the neutral expectation of 0. However, the actual mean may frequently deviate from 0 (Thornton 2005; Wares 2010). Simonsen et al. (1995) compared three measures and found Tajima's D to have the most statistical power:

Tajima's D: Normalized difference between π and S, the number of segregating sites (Tajima 1989).

Fu and Li's D^*: Normalized difference between S and the number of singletons η (alleles observed only once in a sample; Fu and Li 1993).

F^*: Normalized difference between π and η (Fu and Li 1993).

Background selection: Ongoing selection against deleterious mutations that can result in the loss of linked neutral variation (Charlesworth et al. 1993).

Balancing selection: Here we define balancing selection broadly as the class of selective forces that maintain polymorphism over time. This can include, for example, frequency-dependent selection or heterozygote advantage (Charlesworth 2006).

Coalescent theory: A theoretical framework for understanding genetic variation based on the retrospective pattern of shared ancestry among alleles in a sample (Wakeley 2009).

Divergent selection: Positive selection acting differentially between separate populations.

dN/dS: Ratio of nonsynonymous (amino acid–changing) to synonymous substitutions in a nucleotide sequence. Testing for selection based on this ratio typically uses aligned sequence data among populations or taxa and can detect selection over long timescales, although it requires multiple amino acid substitutions (i.e., recurrent selective sweeps).

F_{ST}: A statistic describing the partitioning of allelic variance within versus among populations; F_{ST} ranges from 0 (no population differentiation) to 1 (complete population differentiation). There are multiple ways of calculating F_{ST} that can occasionally have substantial effects on its value but rarely its relative magnitude among loci (Charlesworth 1998; Holsinger and Weir 2009). Commonly used population genomic tests for selection based on identifying outliers in F_{ST} are as follows:

LOSITAN (Antao et al. 2008) computer software implements the method of Beaumont and Nichols (1996) to identify F_{ST} outliers based on heterozygosity, which affects the predicted neutral distribution of F_{ST}.

ARLEQUIN (Excoffier et al. 2009) software performs the same analysis, accounting for hierarchical population structure.

BAYESFST (Beaumont and Balding 2004) assesses the significance of a locus-specific parameter that indicates selection in a model of F_{ST}.

BAYESCAN (Foll and Gaggiotti 2008) modifies the approach of Beaumont and Balding (2004) to estimate the posterior probability of a locus being subject to selection.

DETSEL (Vitalis et al. 2003) uses coalescent simulations in a simple two-population model to identify F_{ST} outliers.

Genetic draft: The loss of genetic diversity and changes in AFS at loci linked to a selected locus during a selective sweep (Gillespie 2000).

HKA test: A test of the neutral prediction for the relationship between within-population diversity and among-population divergence (Hudson et al. 1987).

Linkage disequilibrium (LD): The correlation between alleles across loci. Traditionally, LD has been calculated as a function of a pair of loci, regardless of their physical position (Slatkin 2008). This aspect of LD can be partitioned among populations in the statistic Z_g as a test of selection (Storz and Kelly 2008). Genome scans for selection also apply several of the following statistics that describe the decay of LD as a function of physical distance, also known as haplotype structure:

Extended haplotype homozygosity (EHH) measures the probability that any two randomly chosen haplotypes are identical over a given distance from a focal site (Sabeti et al. 2002).

Integrated haplotype score (iHS) integrates the area under the EHH curve (Voight et al. 2006). Huff et al. (2010) found this measure to have greater statistical power and to be more robust to complex demographics than two related alternatives.

Cross-population extended haplotype homozygosity (XP-EHH) compares EHH between two populations to test for interpopulation differences in the extent of LD (Sabeti et al. 2007).

π: A measure of nucleotide diversity, calculated as the proportion of pairwise differences in a sample; π can be estimated either within or between populations and is directly used in some calculations of F_{ST} (Charlesworth 1998).

Positive (directional) selection: Selection in which one or a class of alleles is favored.

Selective sweeps: The increase in frequency of one or a class of alleles favored by selection. **Hard sweeps** result from selection on a single allele, typically a new mutation that is favored immediately on its appearance in a population. **Soft sweeps** are selection on standing genetic variation or on variants supplied by recurrent mutation or migration during the selective phase, so that a number of different alleles are collectively favored and increase in frequency. These alleles are typically considered to be neutral or even deleterious before a shift in selective regime (Hermisson and Pennings 2005).

in the case of DNA sequence data, is the coalescent (Kingman 1982). Coalescent theory focuses retrospectively on the ancestral relatedness of samples of alleles within and among populations (Wakeley 2009). Unlike traditional phylogenetics, the goal of coalescent theory is usually not to estimate the specific relationships among a sample of sequences. Rather, coalescent theory provides a rigorous model linking evolutionary processes, such as effective population size or natural selection, to expected patterns of resultant genetic variation, such as π or Tajima's D. While harder to visualize on a single plot, statisti-

cal properties of the coalescent structure of haplotypes, such as parameters describing the distribution of coalescence times, are also continuously varying functions of genome position (fig. 2*C*; Nordborg and Innan 2003; Storz 2005). Coalescent theory can also describe genealogies across multiple populations, so that demographic processes such as migration rate are connected to observed patterns of genetic variation, such as F_{ST}, through this genealogical model (fig. 3; Slatkin 1991).

One promise of a population genomics approach is the simultaneous identification of both a genomewide average and outliers for any given statistic, whether these are traditional measures of allele frequencies or aspects of coalescent genealogy. The genomewide average is taken to provide a baseline view of neutral processes, both demographic (e.g., population size, migration rate) and genetic (e.g., mutation rate, recombination). Estimation of the genomewide distribution is an advantage of using a large number of markers spread across the genome, as opposed to a candidate-gene screen for selection, particularly when the underlying demographic processes may not be well known in advance (Wright and Gaut 2005). Outliers from the background indicate the action on specific loci of evolutionary forces such as natural selection, providing an apparently clean separation between neutral and nonneutral processes (fig. 2*E*; Luikart et al. 2003). However, this separation is not as distinct as it may appear. Natural selection, for instance, in the form of background selection (Charlesworth et al. 1993) or divergent adaptation among populations (Nosil et al. 2009), can lead to patterns of population differentiation that affect the entire genome. Conversely, the signature of neutral processes remains even on highly selected loci (Coop et al. 2009). Moreover, some demographic processes increase the variance of population genetic measures across the genome, potentially causing spurious outliers (Teshima et al. 2006; Hermisson 2009).

It can often be informative to consider the impact of selection in terms of processes typically considered to be neutral; for instance, the impact of a selective sweep on coalescence times and genetic diversity mimics the effect of a population bottleneck at the selected locus. In a population genomic sense, demographic factors such as effective population size and migration rate are continuous variables along the genome as well. Therefore, the signatures of all of these neutral and nonneutral processes on both the genomewide distribution of population genetic statistics and specific genomic regions must be considered together in making inferences from genomic data and understanding genomic patterns of variation.

Temporal Considerations

Short Term

The genetic signature of selection persists while selection is occurring and potentially long after it has resulted in fixation of selected alleles. However, that signature differs with time, and different approaches are required to detect it (Oleksyk et al. 2010). In its early stages, selection changes the allele frequency at the selected locus in a population, with detectable effects on linked neutral variation in the genomic region. In the most straightforward case, in which a single new mutation is swept to high frequency (genomic location *i* in fig. 2*B*; see hard vs. soft sweeps), the extent of LD is expected to be larger on either side of the selected locus (Schlötterer 2003; Stephan et al. 2006; McVean 2007). Identifying genomic regions that are outliers from the background in terms of the extent of LD is often the best way to detect an incomplete or very recent selective sweep (Pennings and Hermisson 2006*b*; Sabeti et al. 2006). Somewhat counterintuitively, in the case of a rapid sweep of a single mutation, LD can be positive on either side of the selected site but negative among neutral variants on either side of it (McVean 2007; Pfaffelhuber et al. 2008). Nonetheless, tests for selection based on extended haplotype structure in each direction from the selected site (Sabeti et al.

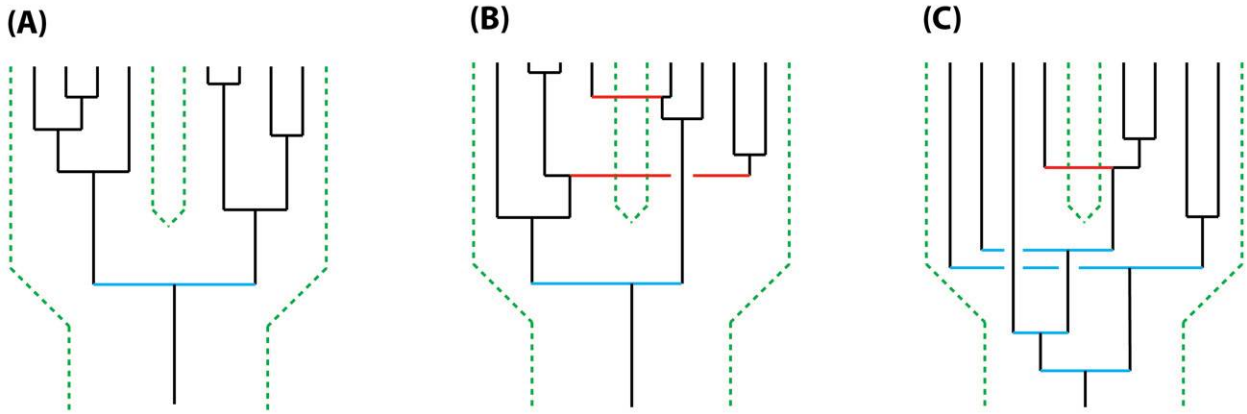

Fig. 3 Possible coalescent genealogies (solid lines) in pairs of populations (dashed green lines) that split from a common ancestor. *A*, Most coalescent events occur within populations, and only one (blue) occurs between populations, predating their split. *B*, Some migration events lead to coalescence between populations (red). *C*, Coalescent events result from migration (red) and also predate the split (blue). These figures could represent three different neutral demographic scenarios: (*A*) a long time elapsed since the populations split and/or small effective population size within each, with no migration; (*B*) a long time since the split but some migration; and (*C*) a short time since the split, with incomplete lineage sorting. Alternatively, these could represent three loci along the genome from the same two populations: (*A*) strong differential selection between the two populations, (*B*) neutral expectation, and (*C*) balancing selection.

2002, 2006, 2007) are effective at detecting the signal in LD during or soon after a sweep. This LD-based signature of selection erodes relatively quickly, on the order of ~0.1 N_e generations (fig. 1*B*; Pfaffelhuber et al. 2008).

The effect of selection is reflected in the coalescent structure of alleles at neighboring sites (fig. 2*B*, *ii*). Haplotypes linked to the favored allele that have not recombined since the selective sweep exhibit the same shallow coalescent pattern, clustering on the genealogy close to the original mutation event. For other haplotypes, in which recombination has occurred between sites *i* and *ii*, genetic diversity at site *ii* will represent a sample from the background coalescent structure (Leocard and Pardoux 2010). Thus, the mean coalescent depth at *ii* is intermediate between the selected locus *i* and the genomewide background *iii*, *v*, but the distribution of coalescence times is qualitatively different from either case.

Overall nucleotide diversity (π) in the genomic region tightly linked to the selected site is reduced compared to the genomewide background in a process called genetic draft (fig. 2*B*; Gillespie 2000; Schlötterer 2003). This reduction in genetic diversity can be used to detect a selective sweep, and this signature of selection lasts longer than that of LD (Pennings and Hermisson 2006*b*; Sabeti et al. 2006). The coalescent patterns described above also lead to an increase in the proportion of low-frequency variants close to the selected locus. This shift in AFS is reflected in a decreased value of Tajima's *D* (Tajima 1989), which is essentially the normalized difference between π and the total number of segregating sites, *S* (box 1). However, the statistical power of AFS-based tests in this situation depends on mutations arising on the haplotype containing the selected allele (i.e., during or after the selective sweep), so it may peak some time after the sweep event (fig. 1*B*; Pennings and Hermisson 2006*b*). Before this point, rapid fixation of the selected allele can lead to a broad region in which nucleotide diversity is virtually eliminated, so that sampling may not identify enough polymorphism to estimate AFS. After this point, the neutral coalescent process acting on the new mutations erodes the signature of selection in AFS (Pennings and Hermisson 2006*b*).

The biological situation can inform the design of genome scans for detecting recent selection within a population. For example, to detect selection that may be ongoing in a population that has colonized a novel habitat, LD-based tests are likely to be most informative. These tests require knowledge of gametic phase. While this can be inferred in genotype samples from diploid organisms (Scheet and Stephens 2006), the uncertainty from this process is typically not carried through the subsequent analyses. A better experimental design may be to sample gametes or individuals in the haploid phase of a life cycle, where gametic phase is unambiguous. On the other hand, if more time has elapsed since the period of selection, knowledge of gametic phase is not required for tests based on nucleotide diversity or allele frequency. In these cases, genotyping of diploids is appropriate, and it may even be possible to pool samples in the genotyping design (Lynch 2009).

Long Term

As time since selection increases, this picture changes. The mean coalescence time among alleles at the selected site increases, approaching the genomewide background, so that other statistics of nucleotide diversity and AFS also approach the genomewide average. Thus, the tests described above for selection within a single population become ineffective; instead, tests that also use data on fixed sequence divergence among taxa are required. For instance, LD-based tests rely on patterns of haplotypes present before the selective sweep. The lifetime of LD-based tests following the selective sweep can be extended by conditioning on mutations present in the ancestral population, if this can be inferred from a comparison with related populations (Pennings and Hermisson 2006*b*). This technique would be most applicable to a situation in which adaptation is being studied in a relatively small population that colonized a new habitat and the ancestral population can also be sampled. A more common approach (the Hudson-Kreitman-Aguadé test) compares fixed sequence divergence between species to diversity within a species, using a neutral model of the predicted relationship between the two (Hudson et al. 1987). Nonetheless, these tests still have a limited lifetime of utility following selection because of their implicit inferences about the ancestral population and the coalescent processes that produce observed patterns of polymorphism (Hudson et al. 1987).

At longer timescales, tests for selection must rely solely on fixed differences between populations rather than polymorphism within populations (fig. 3*A*). For DNA sequence data in which the functional status of nucleotide substitutions can be determined, one can compare the rates of nonsynonymous (potentially selected) with synonymous (putatively neutral) nucleotide substitutions (McDonald and Kreitman 1991; Nielsen and Yang 1998). The assumption is that a difference in evolutionary rate between these classes within a locus indicates selection. This approach requires a longer time period of selection, resulting in multiple amino acid substitutions, because it depends on rates estimated across multiple sites.

Biological Considerations

Mode of Selection

Two major modes of selection can be distinguished by most of the basic tests: balancing and positive selection (see box 1). Positive selection favors one allele (or class of alleles) over others, and the short-term result as described above is a reduction in genetic diversity, a shortening of coalescence intervals, and an increase in low-frequency variants. For longer-term sequence divergence, an excess of nonsynonymous substitutions and an increase in interspecific divergence relative to the expectation from intraspecific diversity indicate positive selection. Balancing selection, a collection of selective forces that maintain polymorphism, leads to generally opposite results in the short term (fig. 2*B*, *iv*). Nucleotide diversity is elevated, and coalescence times are extended (Charlesworth et al. 1997; Nordborg and Innan 2003). However, balancing selection may be harder to detect in a genome scan because the coalescent signal may be weak at the selected locus (Nordborg and Innan 2003; Charlesworth 2006). The genomic width of the signature of balancing selection due to LD may be either narrower or broader than that of positive selection, depending on population size and structure, number of alleles maintained

by balancing selection, and other factors (Charlesworth et al. 1997; Charlesworth 2006). Over the longer term, evidence for balancing selection is an excess of intraspecific nucleotide diversity relative to interspecific divergence (Hudson et al. 1987). Prime examples of loci subject to balancing selection include those related to immune function and pathogen resistance, although such loci may also exhibit a mix of strong positive and balancing selection (Chen et al. 2010). Apparent balancing selection may also result from hidden population structure, in which diversifying selection occurs across populations but these populations are lumped together in the experimental design and analysis. Such hidden population structure can be a source of false positive results in other tests for selection as well (Excoffier et al. 2009).

Hard versus Soft Sweeps

The discussion of positive selection above focused on a hard sweep (fig. 2*B*, *i*), in which a single allele is favored immediately on its appearance in the population and is swept to high frequency. Positive selection can also produce a soft sweep, in which multiple alleles from standing genetic variation are selected to high frequency following an environmental shift (fig. 2*B*, *vi*; Hermisson and Pennings 2005). Such a situation can arise under biologically realistic conditions (Pennings and Hermisson 2006*a*, 2006*b*), and adaptation from standing genetic variation may describe a substantial portion of selected loci across genomes (Pritchard et al. 2010). The distinction between hard and soft sweeps relates to a fundamental question of evolutionary genetics about the source of adaptive genetic variation: whether it represents primarily new mutations (resulting in a hard-sweep pattern) or standing genetic variation that is uncovered or becomes favored as a result of migration, environmental shifts, or changes in genetic interactions (resulting in a soft-sweep pattern).

Many of the classic expectations for the signature of a hard sweep within a population—reduced nucleotide diversity, reduced coalescence times—are less pronounced or absent in the case of a soft sweep (Raquin et al. 2008). The reason is that coalescence among the lineages of the alleles contributing to the sweep occurs both relatively recently, when the alleles are favored, and earlier, when they are selectively neutral or even deleterious. This results in greater variance in coalescence times (Przeworski et al. 2005). Accordingly, a soft sweep has little effect on the mean expectation for Tajima's *D* but can greatly increase its variance, so that soft sweeps can lead to spuriously reduced or elevated values, resembling either positive or balancing selection (Przeworski et al. 2005). In contrast, tests based on LD are more likely to detect a soft sweep and may even have greater statistical power in a soft-sweep than in a hard-sweep situation because soft sweeps leave greater levels of overall polymorphism on which calculations of LD can be based (Pennings and Hermisson 2006*b*). Similarly, F_{ST}-based tests of population differentiation should also detect soft-sweep patterns, as differentiation between populations would be elevated even if within-population diversity does not differ from the genomewide average (fig. 2*D*).

The simplest model of a soft sweep envisions a single population undergoing an environmental shift, with a class of alleles either neutral or deleterious before the shift and favored afterward (Hermisson and Pennings 2005). However, the same model applies to situations in which favored alleles are not present in standing genetic variation but rather supplied by recurrent mutation or migration (Pennings and Hermisson 2006*a*, 2006*b*). Instances of all of these types of soft sweep may be relatively common. In domesticated crops, the alleles favored under domestication may have been neutral or slightly deleterious in the ancestral wild population and thus present under mutation-drift or mutation-selection equilibrium (Purugganan et al. 2000; Innan and Kim 2004; Raquin et al. 2008). Structural variants such as gene copy number may be important in the evolution of resistance to drugs or pesticides in pathogens, and these changes may be expected to have a higher mutation rate than other types of adaptive alleles (Raymond et al. 2001; Nair et al. 2007). Loss-of-function mutations that confer an adaptive phenotype may also lead to a soft sweep by recurrent mutation due to their higher mutation rate (Cao et al. 2005).

Demographic Equilibrium versus Nonequilibrium

The impact of selection on particular loci can mimic the effects of changes in effective population size (Hill and Robertson 1966; Gillespie 2000). For example, consider the short-term effects of a hard sweep: the changes in nucleotide diversity, coalescent structure, and AFS at the selected locus, as well as the decay in LD moving along the genome, are akin to the effects of a recent bottleneck in effective population size at this locus, even if the total population size remains constant. This reflects the fact that individuals carrying the favored allele are disproportionately represented at this genomic region in subsequent generations. Soft sweeps may have a similar, if less pronounced, effect (Raquin et al. 2008). As with increased genetic drift during a population bottleneck, genetic diversity is reduced in the neighborhood of the selected locus (Gillespie 2000). Coalescent events are more likely to occur during the time when the "population size" of the selected haplotype (i.e., its frequency in the population) is small, and these coalescent events remain clustered at the root of the genealogy as this haplotype population expands. This clustering results in a shift in AFS. In addition, the reduced effective population size at this locus reduces the effective recombination rate, increasing the distance of LD from the selected locus.

Conversely, demographic processes in populations that are not at equilibrium can mimic the effects of selection. For instance, many domesticated species went through a population bottleneck at the time of domestication, leading to genomewide loss of genetic diversity; tests for selection based on loss of diversity at specific loci must take this into account (Wright et al. 2005; Doebley et al. 2006). A genomic approach disentangles the genomically localized effects of selection, which mimic demographic processes, from the genomewide background effects of demography. This disentanglement is done by focusing on outliers in a genome scan. Tests for selection that can incorporate complex demographic scenarios are best able to differentiate outliers from the expected genomewide distribution (Nielsen et al. 2005; Excoffier et al. 2009). For instance, expanding population size is expected to reduce Tajima's *D*, and declining population size is expected to increase it (Tajima 1989; Innan and Stephan 2000). Outliers in Tajima's

D from the genomewide average, rather than deviation from the expected neutral value, can indicate selection in this case (Kelley et al. 2006). However, it is the effects of nonequilibrium demographics on variance in these measures that can be more problematic for tests of selection. For instance, declining population size or a population bottleneck can greatly increase the variance in coalescent times across loci (Hermisson 2009; Wakeley 2009). Tests based on an assumption of constant population size, when a population has actually experienced fluctuations, are likely to suggest false positives on both ends of the statistical distributions (Teshima et al. 2006).

Spatial Considerations

Single versus Multiple Populations

Tests based on nucleotide diversity, AFS, and patterns of LD can be applied to short-term selection within a single population, but comparisons of these measures across multiple populations provide further tests of selection. The most common approach to multiple-population comparisons is based on F_{ST}, the partitioning of variance in allele frequency among versus within populations (fig. 2*D*; Wright 1931; Holsinger and Weir 2009). Other measures of the partitioning of genetic variance have been proposed (Schlötterer and Dieringer 2005), and LD can also be partitioned among populations in an analogous manner (Kelly 2006; Storz and Kelly 2008). If differential selection is operating on a locus between two populations, a greater proportion of variance is expected between populations, resulting in higher values of statistics that measure population structure, such as F_{ST} (Storz 2005). This is a result of positive selection within one or both populations producing shorter within-population coalescence intervals (fig. 2*B*, *i*), leading to a greater proportion of coalescent events occurring within each population rather than between them (fig. 3*A*). In contrast, balancing selection within one or both populations maintains polymorphism and pushes coalescence intervals back in time (fig. 2*B*, *iv*). Under balancing selection, the branches leading back from present-day alleles are more likely to experience a migration event, crossing from one population to the other (fig. 3*B*), or even to coalesce when the populations were panmictic (fig. 3*C*). However, F_{ST} is a ratio of allelic diversities and as such is sensitive to other processes that can affect distribution of nucleotide diversity across the genome, such as recombination and background selection (Charlesworth et al. 1997; Storz 2005). Studies focusing on F_{ST} should examine diversity within and among populations as well (Charlesworth et al. 1997).

The statistic F_{ST} is often estimated as a parameter in a specific demographic model and is thus translated into a neutral estimate of the effective migration rate among many populations, of which the observed populations are a sample (Weir and Cockerham 1984). However, F_{ST} can also be viewed from a coalescent perspective (Slatkin 1991), reflecting the distribution of coalescent events within versus among populations (fig. 3). Separation of the timescales of these two coalescent processes provides the basis for tests of selection using F_{ST} (Beaumont 2005). Again, this coalescent view illustrates the analogies between selection and demographic processes that are specific to particular loci. Differential selection at a locus is reflected in relatively fewer coalescent events between populations and more within, which is analogous to a lower migration rate between populations at the selected locus.

Replicate Populations

Strong inferences about selection can be made in a set of replicate populations that evolve in parallel across habitats or putative selective regimes. While F_{ST} outliers in a single population comparison can be the result of selection, other factors such as hidden population structure can result in false positives (Excoffier et al. 2009). However, if replicate evolved populations can be sampled, F_{ST} outliers localized to the same genetic region across separate pairwise comparisons can provide much stronger evidence for the selective significance of that region. In this case, multiple comparisons among populations provide a more complete picture of selection. For instance, genomic regions exhibiting elevated F_{ST} in multiple comparisons across habitat types but average or even reduced F_{ST} within habitat types suggest that the same alleles are responding in parallel to selection within each habitat (Kane and Rieseberg 2007; Hohenlohe et al. 2010). In contrast, genomic regions with elevated F_{ST} both within and among habitats suggest either selection on different alleles or differential selection that is uncorrelated with habitat type (Hohenlohe et al. 2010).

Replicate oceanic and freshwater populations of threespine stickleback provide an example of parallel genomic evolution between habitat types (Hohenlohe et al. 2010). By examining both population differentiation and within-population nucleotide diversity, signatures of balancing and divergent selection are evident across the genome, including on linkage group (LG) XXI (fig. 4). Balancing selection is suggested at one location (fig. 4, *i*) by significantly elevated π within populations and reduced F_{ST} (although not significantly) among populations. Stronger evidence for divergent selection between freshwater and oceanic habitats occurs at another location on LG XXI (fig. 4, *ii*), where each pairwise F_{ST} between an independently derived freshwater population and the ancestral oceanic populations is highly significant (fig. 4*A*) and nucleotide diversity within each population is significantly reduced (fig. 4*B*). Further, the response to selection at this location exhibits a parallel hard-sweep pattern, in which the same alleles are selected to high frequency in each of the freshwater populations, evidenced by the lack of any significant population differentiation among the freshwater populations at this genomic region. Other regions of the genome exhibit nonparallel sweeps, in which each freshwater population was again highly differentiated from the oceanic ancestors but also from the other freshwater populations. In these other cases, it appears that divergent selection has acted on different genetic variants in the different freshwater populations (Hohenlohe et al. 2010).

Discussion

Limits to the Detection of Selection

The biological considerations described above give some guidance as to which statistical analyses may have the greatest

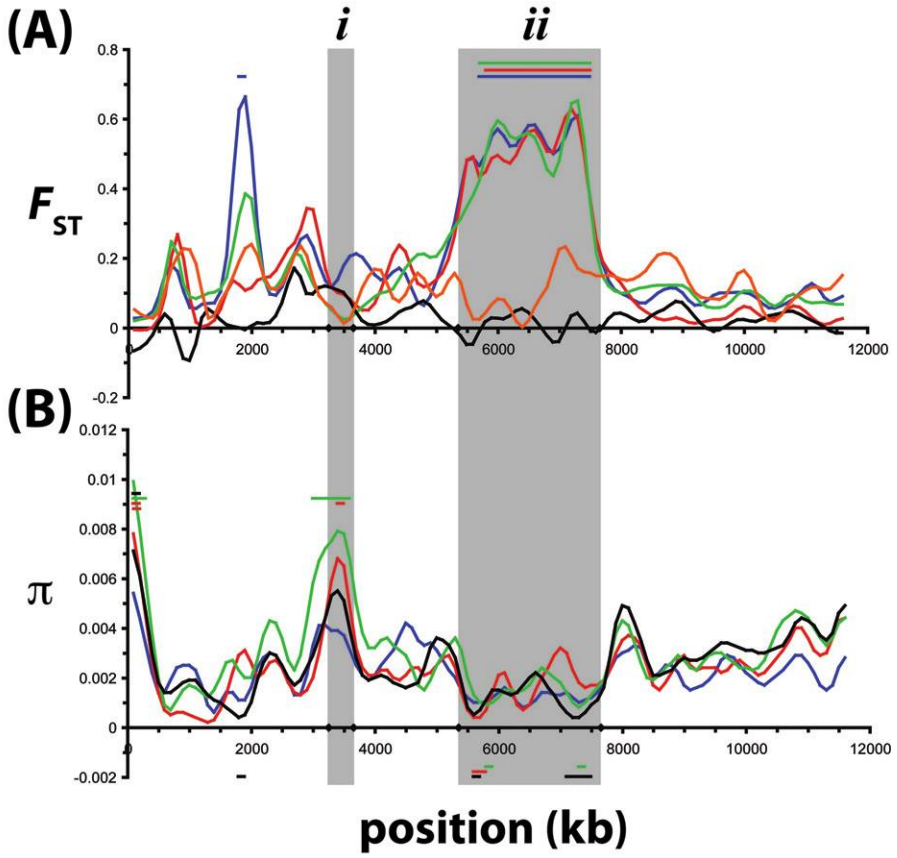

Fig. 4 Evidence of selection in replicate oceanic and freshwater populations of threespine stickleback, showing only linkage group XXI (~2.5% of the total genome). *A*, Population differentiation: pairwise F_{ST} between each of three independently derived freshwater populations and oceanic ancestors (blue, red, green), F_{ST} among the three freshwater populations (orange), and F_{ST} between the two oceanic populations (black). Bars above, colored by population, represent bootstrap P values $<10^{-5}$. *B*, Nucleotide diversity: π within the three freshwater populations (blue, red, green) and within the oceanic populations (black). Bars above and below, colored by population, represent bootstrap $P < 10^{-4}$. Shaded areas delineate putative regions of balancing selection (*i*) and of parallel divergent selection between oceanic and freshwater habitats (*ii*), based on significantly elevated π (across freshwater populations; $P < 10^{-4}$) and F_{ST} values (overall oceanic vs. freshwater; $P < 10^{-5}$), respectively. All plots were calculated from genotypes at 1169 SNPs on this linkage group in 100 individuals from five populations, using RAD sequencing (Hohenlohe et al. 2010).

statistical power under different scenarios (fig. 1*A*). However, there remain limits on the power of population genomics approaches to detect loci under selection. As described above, the signal of selection in patterns of polymorphism, coalescent structure, and LD erodes over time after selection has ended (fig. 1*B*). More distant selection can be detected in the functional significance of nucleotide substitutions (e.g., *dN*/*dS* tests), but this requires multiple substitutions to have occurred. A single selective sweep in the distant past may be impossible to detect in a genome scan, even with complete genetic information and sophisticated analyses.

Lessons can be learned from genomewide association studies, which assess the genetic basis of complex phenotypic traits. In this case, an independent estimate of the heritability of a trait gives a target for the total genetic contribution to trait variance that theoretically could be explained in an association study. In general, genomewide association studies fall well short of this target, meaning that a large proportion of heritability goes unexplained (Frazer et al. 2009). For example, the heritability of height in humans is ~0.8, meaning that ~80% of the phenotypic variance is additive genetic. Nonetheless, a trio of genomewide association studies covering ~63,000 subjects identified genetic variants accounting for only ~5% of the variation (reviewed in Maher 2008; Visscher 2008). In studies of selection, there is no analog to heritability that would provide a target of the total genetic response to selection that could potentially be explained. Sample sizes and genomic resources for selection studies in nonmodel organisms will typically be smaller and less extensive than those in humans. On the other hand, an advantage of selection studies over association studies is that the effect of selection is integrated over many generations. It is possible to detect a signature of selection on a locus, the result of multiple generations, even when the per-generation selection coefficient (i.e., the effect size of the locus) is quite small. Nonetheless, many of the same factors that keep genetic variation hidden from view in association studies (Maher 2008) are likely to apply to selection studies as well. The number of loci identified as respond-

ing to selection and contributing to adaptive evolution is likely to be an underestimate.

Soft sweeps—adaptation from standing genetic variation—can be very difficult to detect and may explain a large portion of the response to selection in natural populations (Pritchard et al. 2010). For example, while many loci have been implicated in selection scans in humans, very few of these actually show fixed differences among populations. Pritchard et al. (2010) interpret these results to emphasize the importance of polygenic adaptation, fitting with a quantitative genetic model in which shifts in allele frequency across a large number of loci, each with small effect, produce a cumulatively large phenotypic response to selection. In this case, the genetic signature of selection on each single locus may be weak (Storz 2005). As with genomewide association studies, epistatic interactions among loci can also mask individual effects, producing outcomes similar to a soft sweep (Takahashi 2009). Genotype-by-environment interactions can further complicate both genomewide association studies and selection studies, although very large sample sizes may be able to untangle them (Brachi et al. 2010).

One conclusion from these considerations is that the set of loci implicated across the genome in a particular study may still explain only a small portion of the phenotypic response to selection in the population. While identifying some loci that have responded to selection is a relatively straightforward process and proceeding from there to functional studies of candidate loci is a valuable research program, identifying all the loci across the genome that are important in a particular selective response is a much more daunting task. Accordingly, conclusions about the genomewide proportion of loci subject to selection or the distribution of their effects should be made tentatively, with these limitations in mind. Further, the tests for selection described above have different levels of statistical power to detect forms of selection and its response, such as hard versus soft sweeps, positive versus balancing selection, or dominant versus additive alleles (Teshima et al. 2006). Thus, conclusions about the relative importance or frequency of these modes of adaptive evolution are subject to bias that has not been adequately quantified.

Future Prospects

Despite these cautionary notes, it is clear that genome scans for selection have been widely successful in uncovering loci responding to selection in natural populations (Sabeti et al. 2006). Such loci provide candidate genes for functional studies and emphasize the ubiquity of selection in natural populations. In addition to studying the genomics of selection from other perspectives, such as experimental evolution and continued technological advances, there remain several avenues for future progress in strengthening population genomic approaches to detect selection (Hermisson 2009).

Two general principles should apply in most cases. First, tests based on multiple aspects of genomic structure should be applied to each data set. In addition to increasing overall statistical power (Nielsen et al. 2005; Grossman et al. 2010), this approach provides more opportunity to separate the roles of demographic and genetic factors from selection. For instance, tests of outliers in F_{ST} should be combined with examination of nucleotide diversity within and across populations because of the intricate relationship among these statistics and recombination rates (Charlesworth et al. 1997). Second, given the availability of genome-scale data, the null hypothesis for any test of selection should be derived from the genomewide distribution rather than a simple a priori neutral model. For instance, while neutrality predicts an expectation for Tajima's D of 0, the several assumptions underlying this prediction may be violated more often than not (Thornton 2005; Wares 2010).

Related to this second point, demographic history and population structure can have a disturbingly large effect on genomewide expectations and variances and on the rate of false positives (Boitard et al. 2009; Excoffier et al. 2009). To account for this issue, statistical models are required that can accommodate arbitrarily complex, nonequilibrium demographic scenarios; estimate relevant parameters from the genomewide data; and then identify outlier regions that exhibit signatures of selection. For example, Gutenkunst et al. (2010) present a method for estimating from genomic data the parameters of a population model that can include combinations of expansion, contraction, migration, and admixture.

Finally, an approach that is currently gaining strength in genomewide association studies is to incorporate the network structure of the genotype-phenotype relationship (Benfey and Mitchell-Olds 2008; Flowers et al. 2009; Schadt 2009). This is done by assessing evidence for interaction effects among loci in addition to direct effects. A key difficulty in assessing interactions is statistical power. However, the impact of epistasis on the response to selection in terms of coalescent structure and patterns of genetic variation has only just begun to be addressed (Takahashi 2009), and understanding the signature of selection in this context is the first step toward developing new analyses to detect it.

Conclusions

The population genomics of natural selection confronts a set of fundamental questions about the genetics of adaptation: What proportion of the genome contributes to adaptive genetic variation? Does genetic variation for adaptation to novel selective conditions come primarily from new mutations or from standing genetic variation? What is the relative importance of different modes of natural and sexual selection in evolution? To what extent do LD and genomic architecture limit adaptation? To date, our ability to answer these questions has been primarily limited by available genetic data (Phillips 2005).

However, the rate of technological advance in nucleotide sequencing means that it will soon be feasible to have complete genetic information—the entire genome sequence—for multiple individuals across populations of many organisms. At this point, there will be no further genetic information available to gather from present-day populations in order to elucidate past evolutionary history. Increasingly, the limits on our ability to understand the genomics of natural selection are a result of weaknesses in the available theory and analytical tools rather than gaps in molecular data. Major steps can be taken by better understanding the sensitivity of existing approaches to un-

derlying assumptions, such as constant effective population size or lack of genotype-by-environment interactions. Over the longer term, the genomics of selection will benefit from applying other branches of statistics and mathematics that have not yet been used in population genetics to genomic processes. Coalescent theory provides an example of how a new theoretical framework can both unify key concepts and provide new statistical and analytical tools. With such an infusion of novel theory, the emerging field of population genomics can make great strides in addressing the fundamental questions above.

Acknowledgments

This work was funded in part by National Science Foundation (NSF) grant IOS-0843392 to P. A. Hohenlohe; National Institutes of Health (NIH) grant 1R24GM079486-01A1 and NSF grants DEB-0919090 and IOS-0642264 to W. A. Cresko; and NSF grant DEB-0641066, NIH grant AG022500, and a Senior Scholar Award from the Ellison Medical Foundation to P. C. Phillips. We thank M. Streisfeld for assistance with references and two anonymous reviewers for helpful and detailed comments on the manuscript.

Literature Cited

Akey JM 2009 Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19:711–722.

Antao T, A Lopes, RJ Lopes, A Beja-Pereira, G Luikart 2008 LOSITAN: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. BMC Bioinform 9:323.

Beaumont MA 2005 Adaptation and speciation: what can F_{ST} tell us? Trends Ecol Evol 20:435–440.

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

Beaumont MA, RA Nichols 1996 Evaluating loci for use in the genetic analysis of population structure. Proc R Soc B 263:1619–1626.

Benfey PN, T Mitchell-Olds 2008 From genotype to phenotype: systems biology meets natural variation. Science 320:495–497.

Boitard S, C Schlötterer, A Futschik 2009 Detecting selective sweeps: a new approach based on hidden Markov models. Genetics 181: 1567–1578.

Brachi B, N Faure, M Horton, E Flahauw, A Vazquez, M Nordborg, J Bergelson, J Cuguen, F Roux 2010 Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. PLoS Genet 6:e1000940.

Cao S, Q Xu, Y Cao, K Qian, K An, Y Zhu, H Binzeng, H Zhao, B Kuai 2005 Loss-of-function mutations in *DET2* gene lead to an enhanced resistance to oxidative stress in *Arabidopsis*. Physiol Plant 123:57–66.

Charlesworth B 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol 15: 538–543.

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

Charlesworth B, M Nordborg, D Charlesworth 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70:155–174.

Charlesworth D 2006 Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.

Chen Q, Z Han, H Jiang, D Tian, S Yang 2010 Strong positive selection drives rapid diversification of *R*-genes in *Arabidopsis* relatives. J Mol Evol 70:137–148.

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

Doebley JF, BS Gaut, BD Smith 2006 The molecular genetics of crop domestication. Cell 127:1309–1321.

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

Fisher RA 1930 The genetical theory of natural selection. Clarendon, Oxford.

Flowers JM, Y Hanzawa, MC Hall, RC Moore, MD Purugganan 2009 Population genomics of the *Arabidopsis thaliana* flowering time gene network. Mol Biol Evol 26:2475–2486.

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

Frazer KA, SS Murray, NJ Schork, EJ Topol 2009 Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241–252.

Fu Y-X, W-H Li 1993 Statistical tests of neutrality of mutations. Genetics 133:693–709.

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

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

Gutenkunst RN, RD Hernandez, SH Williamson, CD Bustamante 2010 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Hahn MW 2006 Accurate inference and estimation in population genomics. Mol Biol Evol 23:911–918.

Hermisson J 2009 Who believes in whole-genome scans for selection? Heredity 103:283–284.

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

Hill WG, A Robertson 1966 The effect of linkage on limits to artificial selection. Genet Res 8:269–294.

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

Holsinger KE, BS Weir 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet 10:639–650.

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

Huff CD, HC Harpending, AR Rogers 2010 Detecting positive selection from genome scans of linkage disequilibrium. BMC Genomics 11:8.

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

Innan H, W Stephan 2000 The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. Genetics 155:2015–2019.

Kane NC, LH Rieseberg 2007 Selective sweeps reveal candidate

genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. Genetics 175:1823–1834.
Kelly JK 2006 Geographical variation in selection, from phenotypes to molecules. Am Nat 167:481–495.
Kelley JL, J Madeoy, JC Calhoun, W Swanson, JM Akey 2006 Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16:980–989.
Kimura M 1968 Evolutionary rate at the molecular level. Nature 217:624–626.
Kingman JFC 1982 The coalescent. Stoch Proc Appl 13:235–248.
Leocard S, E Pardoux 2010 Evolution of the ancestral recombination graph along the genome in case of selective sweep. J Math Biol, doi: 10.1007/s00285-009-0321-4.
Luikart G, PR England, D Tallmon, S Jordan, P Taberlet 2003 The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994.
Lynch M 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182:295–301.
Maher B 2008 The case of the missing heritability. Nature 456:18–21.
McDonald JH, M Kreitman 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654.
McVean G 2007 The structure of linkage disequilibrium around a selective sweep. Genetics 175:1395–1406.
Nair S, D Nash, D Sudimack, A Jaidee, M Barends, A-C Uhlemann, S Krishna, F Nosten, TJC Anderson 2007 Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol 24:562–573.
Nielsen R 2005 Molecular signatures of natural selection. Annu Rev Genet 39:197–218.
Nielsen R, SH Williamson, Y Kim, MJ Hubisz, AG Clark, CD Bustamante 2005 Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575.
Nielsen R, Z Yang 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–996.
Nordborg M, H Innan 2003 The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics 163:1201–1213.
Nosil P, DJ Funk, D Ortiz-Barrientos 2009 Divergent selection and heterogeneous genomic divergence. Mol Ecol 18:375–402.
Oleksyk TK, MW Smith, SJ O'Brien 2010 Genome-wide scans for footprints of natural selection. Philos Trans R Soc B 365:185–205.
Pavlidis P, S Hutter, W Stephan 2008 A population genomic approach to map recent positive selection in model species. Mol Ecol 17:3585–3598.
Pennings PS, J Hermisson 2006*a* Soft sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23:1076–1084.
——— 2006*b* Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2:e186.
Perkel J 2008 SNP genotyping: six technologies that keyed a revolution. Nat Methods 5:447–453.
Pfaffelhuber P, A Lehnert, W Stephan 2008 Linkage disequilibrium under genetic hitchhiking in finite populations. Genetics 179:527–537.
Phillips PC 2005 Testing hypotheses regarding the genetics of adaptation. Genetica 123:15–24.
Pritchard JK, JK Pickrell, G Coop 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208–R215.
Przeworski M, G Coop, JD Wall 2005 The signature of positive selection on standing genetic variation. Evolution 59:2312–2323.
Purugganan MD, AL Boyles, JI Suddith 2000 Variation and selection at the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. Genetics 155:855–862.
Raquin A-L, P Brabant, B Rhoné, F Balfourier, P Leroy, I Goldringer 2008 Soft selective sweep near a gene that increases plant height in wheat. Mol Ecol 17:741–756.
Raymond M, C Berticat, M Weill, N Pasteur, C Chevillon 2001 Insecticide resistance in the mosquito *Culex pipiens*: what have we learned about adaptation? Genetica 112–113:287–296.
Sabeti PC, DE Reich, JM Higgins, HZP Levine, DJ Richter, SF Schaffner, SB Gabriel, et al 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.
Sabeti PC, SF Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, TS Mikkelsen, DM Altshuler, ES Lander 2006 Positive natural selection in the human lineage. Science 312:1614–1620.
Sabeti PC, P Varilly, B Fry, J Lohmueller, E Hostetter, C Cotsapas, X Xie, et al 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.
Schadt EE 2009 Molecular networks as sensors and drivers of common human diseases. Nature 461:218–223.
Scheet P, M Stephens 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644.
Schlötterer C 2003 Hitchhiking mapping: functional genomics from the population genetics perspective. Trends Genet 19:32–38.
Schlötterer C, D Dieringer 2005 A novel test statistic for the identification of local selective sweeps based on microsatellite gene diversity. Pages 55–64 *in* D Nurminsky, eds. Selective sweep. Landes Bioscience, Georgetown, TX.
Shendure J, H Ji 2008 Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145.
Simonson KL, GA Churchill, CF Aquadro 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429.
Slatkin M 1991 Inbreeding coefficients and coalescence time. Genet Res 58:167–175.
——— 2008 Linkage disequilibrium: understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485.
Stephan W, YS Song, CH Langley 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172: 2647–2663.
Storz JF 2005 Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688.
Storz JF, JK Kelly 2008 Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. Genetics 180:367–379.
Tajima F 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Takahashi KR 2009 Coalescent under the evolution of coadaptation. Mol Ecol 18:5018–5029.
Teshima KM, G Coop, M Przeworski 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702–712.
Thornton K 2005 Recombination and the properties of Tajima's *D* in the context of approximate-likelihood calculation. Genetics 171: 2143–2148.
Visscher PM 2008 Sizing up human height variation. Nat Genet 40: 489–490.
Vitalis R, K Dawson, P Boursot, K Belkhir 2003 DetSel 1.0: a computer program to detect markers responding to selection. J Hered 94:429–431.
Voight BF, S Kudaravalli, X Wen, JK Pritchard 2006 A map of recent positive selection in the human genome. PLoS Biol 4:e72.
Wakeley J 2009 Coalescent theory: an introduction. Roberts, Greenwood Village, CO.
Wares JP 2010 Natural distributions of mitochondrial sequence diversity support new null hypotheses. Evolution 64:1136–1142.

Weir BS, CC Cockerham 1984 Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.

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

——— 1978 Evolution and the genetics of populations. University of Chicago Press, Chicago.

Wright SI, IV Bi, SG Schroeder, M Yamasaki, JF Doebley, MD McMullen, BS Gaut 2005 The effects of artificial selection on the maize genome. Science 308:1310–1314.

Wright SI, BS Gaut 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519.