Selective Sweeps

Wolfgang Stephan[1]
Leibniz-Institute for Evolution and Biodiversity Science, 10115 Berlin, Germany

ABSTRACT For almost 20 years, many inference methods have been developed to detect selective sweeps and localize the targets of directional selection in the genome. These methods are based on population genetic models that describe the effect of a beneficial allele (*e.g.*, a new mutation) on linked neutral variation (driven by directional selection from a single copy to fixation). Here, I discuss these models, ranging from selective sweeps in a panmictic population of constant size to evolutionary traffic when simultaneous sweeps at multiple loci interfere, and emphasize the important role of demography and population structure in data analysis. In the past 10 years, soft sweeps that may arise after an environmental change from directional selection on standing variation have become a focus of population genetic research. In contrast to selective sweeps, they are caused by beneficial alleles that were neutrally segregating in a population before the environmental change or were present at a mutation-selection balance in appreciable frequency.

KEYWORDS genetic hitchhiking; selective sweeps; background selection; demography

WHEN a strongly advantageous mutation occurs and spreads in a population by directional selection, it is inevitable that the frequency of linked neutral (or weakly selected) variants increases. In a seminal paper entitled *The hitch-hiking effect of a favorable gene*, Maynard Smith and Haigh (1974) described this process and termed it *genetic hitchhiking*. They showed that, in very large populations, a single hitchhiking event can reduce genetic variation near the site of selection in the genome.

The hitchhiking effect can most easily be envisioned in nonrecombining organisms (bacteria), in which it was actually first studied and called *periodic selection* (Atwood *et al.* 1951). Suppose a new, selectively favored, mutation arises in a haplotype that carries a given set of neutral nucleotide variants. If the favored mutation goes to fixation, the neutral variants linked to the selected mutation will also spread ("hitchhike") to fixation, while the other variants in the region will get lost. As a consequence, at the time of fixation of the beneficial allele, genetic variation on the entire haplotype is completely eliminated. However, as Maynard Smith and Haigh (1974) demonstrated, in the presence of recombination, the size of the region of reduced variation may be limited to a relatively small fraction of the genome.

The basic hitchhiking model analyzed by Maynard Smith and Haigh (1974) is shown in Figure 1. Initially, when a beneficial allele arises by mutation there are three different haplotypes present in the population: two of them are polymorphic at the focal neutral locus (with alleles A and a) and monomorphic at a selected locus nearby, while the third haplotype carries the beneficial allele at the selected locus and one of the neutral alleles (here A) at the other locus. After fixation of the beneficial allele, only one haplotype exists in the population if no recombination event has occurred between the neutral and selected loci (lower left side of the panel); in this case, variation at the neutral locus is eliminated at the time of fixation through the hitchhiking effect. In contrast, if recombination has occurred during the fixation process of the beneficial allele, the neutral locus remains polymorphic, and thus two haplotypes are present in the population (lower right side). After fixation of the beneficial allele, the neutral locus remains polymorphic if it can escape hitchhiking. The chance of this happening increases with the recombination rate c and with the time available for recombination to occur. The latter is proportional to $1/s$, where s is the selection coefficient of the beneficial allele. Combining this suggests that c/s is the crucial parameter of Maynard Smith and Haigh's hitchhiking model and that this parameter determines the size of the region of reduced variation in the genome of sexual species.

doi: https://doi.org/10.1534/genetics.118.301319
Manuscript received May 8, 2018; accepted for publication July 10, 2018
[1]Address for correspondence: Leibniz-Institute for Evolution and Biodiversity Science, Natural History Museum, Invalidenstrasse 43, 10115 Berlin, Germany. E-mail: stephan@bio.lmu.de

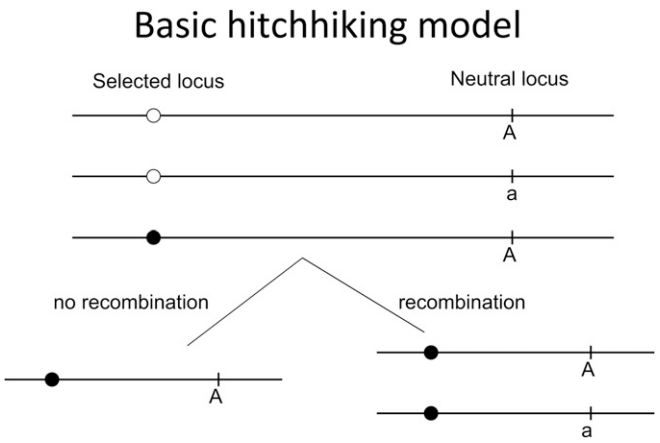

Figure 1 Basic hitchhiking model. The upper part of the figure shows the three haplotypes present in a population when a beneficial mutation (filled circle) occurs at the selected locus. The wildtype allele at the selected locus is indicated by an open circle. At the neutral locus two alleles A and a are present. The haplotypes after the fixation of the beneficial allele are depicted in the lower part of the figure. If no recombination occurs during the fixation process one haplotype is present (left side). With recombination the neutral locus stays polymorphic and two haplotypes remain (right side).

The hitchhiking model was stimulated by Lewontin's (1974) observation that allozyme variability levels are only weakly related to population size, which contradicts the predictions of the neutral theory. Maynard Smith and Haigh's analysis suggests that the observed pattern of genetic variability in a species would depend more on the frequency of hitchhiking events than genetic drift determined by effective population size. Readers further interested in the historical context of the hitchhiking model, and the study of neutral variation linked to selection in general, are referred to the recent review by Charlesworth and Charlesworth (2018).

In the 1970s, population geneticists were conscious that selection at a locus may have effects on another locus, but were thinking almost entirely about fitness interactions between two selected loci, and not about effects when the variants at one locus are neutral (D. Charlesworth, personal communication). Maynard Smith and Haigh's work was therefore original and exciting for the community or provoked strong rebuttals (see Ohta and Kimura 1975).

Genetic hitchhiking, or, in other words, the study of selection via linked neutral variation, has become an important concept in evolutionary genetics, because the observed patterns of neutral variation may be used to infer where selective events (whose genomic location is usually unknown) have occurred along the genome. This may then lead to the identification and characterization of genes that are important in adaptation, or of alleles that cause fitness differences between individuals or populations. Identifying selection events along the genome through genetic hitchhiking may also be useful for fine-scale analysis in other mapping studies (*e.g.*, of quantitative trait loci). Furthermore, evolutionary rates, such as rates of adaptive substitutions, can be inferred from such analyses as well as estimates of the strength of selection.

In this paper, I will briefly describe the extensions of the basic hitchhiking model and its application to statistical tools for detecting positive selection in the genome from patterns of variation at neutral sites, which in the meantime have developed to a large volume.

Selective Sweeps *vs.* Background Selection

The hitchhiking effect was revisited in the late 1980s to explain patterns of reduced variation in restriction-fragment-length-polymorphism data, which were collected in Charles H. Langley's laboratory. These patterns were found in genomic regions of low recombination rates around centromeres and telomeres (Aguadé *et al.* 1989; Stephan and Langley 1989) and also on the fourth chromosome of *Drosophila* (Berry *et al.* 1991). Begun and Aquadro (1992) corroborated these results by showing that levels of DNA variation correlate with recombination rates across much of the *D. melanogaster* genome, whereas average divergence to *Drosophila simulans* was hardly affected by recombination. Given these data, the deterministic hitchhiking model of Maynard Smith and Haigh (1974) was extended by Kaplan *et al.* (1989), who analyzed a stochastic version of the process (including genetic drift) by means of coalescent theory. Furthermore, Stephan *et al.* (1992) and Wiehe and Stephan (1993) studied the recurrent hitchhiking case using the diffusion equation method. In the recurrent case, hitchhiking events occur at multiple loci across the genome. Alternative approximations of the hitchhiking model were provided by Barton (1998) and Gillespie (2000).

The concept of genetic hitchhiking is now very broadly used in population genetics and describes any situation in which changes in allele frequencies caused by selection affect the frequencies of neutral variants at linked sites in the genome. This includes any type of selection that is sufficiently strong. At the same time, and more specifically, for genetic hitchhiking caused by directional selection (as considered by Maynard Smith and Haigh) the term *selective sweep*, which was introduced by Berry *et al.* (1991), is now generally used.

Since the publication of the paper entitled *The effect of deleterious mutations on neutral variation* by Charlesworth *et al.* (1993), two major population genetic models have competed in explaining the observed reduction of nucleotide variation in genomic regions of reduced recombination rates. According to both models, the level of neutral (or nearly neutral) variation can be reduced below classical neutral expectation by selection against the steady input of deleterious mutations (so-called *background selection*) or by recurrent selective sweeps. The discovery of reduced levels of variation in genomic regions of restricted crossing-over, and the ensuing controversy over its interpretation, initiated an important phase in molecular population genetics. Since these observations were not limited to *Drosophila*, but were also found – at least to some extent – in other organisms such as humans (Nachman *et al.* 1998; Hellmann *et al.* 2003) and plants (Kraft *et al.* 1998; Stephan and Langley 1998), the development

Box 1 Overview of Models of Selective Sweeps and Statistical Tests for Sweep Detection

Model of
- a single selective sweep: describes the effect of a strongly beneficial allele on linked neutral variation (see Figure 1)
- recurrent selective sweeps: describes the cumulative effect of single sweeps occurring sequentially at multiple loci along the genome
- competing selective sweeps: like the model of recurrent selective sweeps but sweeps may occur simultaneously and therefore interfere with each other
- soft sweep: involves beneficial alleles present initially (at the time of environmental change) as >1 copy

Frequently used tests for sweep detection:
- CLR test: detects a single selective sweep (without controlling for demography)
- GOF test: used in combination with the CLR test to control for demography
- SweepFinder: CLR test that is applicable to the whole genome and also controls for background selection
- SweeD test: computationally advanced test based on the SweepFinder algorithm
- iHS test: haplotype-based test that detects an ongoing selective sweep

of new methods distinguishing the relative contributions of background selection and selective sweeps was a major activity in those years (until ~2000).

Despite substantial efforts from many theorists and empiricists, fundamental questions on the relation of background selection and selective sweeps are still open. However, since these issues are not a major subject here, the reader is referred to recent work of Comeron (2014, 2017) and Elyashiv *et al.* (2016). In this article, I describe a major shift in focus of the basic model of a selective sweep to models of selective sweeps with an emphasis on an application to data from genomic regions of normal recombination. This shift that occurred ~2000 is accompanied by the advent of population genomics, which allowed collecting large sets of polymorphism data along the entire genome of recombining species (including regions of normal recombination rates). In the following, complexities, such as demography, population structure, and confounding selection regimes, are introduced successively into the basic model of Maynard Smith and Haigh (1974), to make the models more realistic and applicable to data. An overview of the various models and statistical methods is provided in Box 1.

Single Selective Sweeps

The simplest case: a single selective sweep in a local population of constant size

We consider a locus under positive directional selection in a local population. We assume that a beneficial allele occurs at some time in the past and goes to fixation. The sudden occurrence of this allele may be caused by mutation, migration from another subpopulation or may be due to a very rare allele in the standing variation after an environment change. According to Maynard Smith and Haigh's model, nucleotide diversity vanishes in recombining chromosomal regions at the site of selection immediately after the fixation of the beneficial allele and increases as a function of the ratio of the recombination rate c (between the neutral and selected sites) and the selection coefficient s. For finite populations of large constant size, the results obtained by coalescent (Kaplan *et al.* 1989) and diffusion (Stephan *et al.* 1992) approximations are in excellent agreement with Maynard Smith and Haigh's deterministic predictions. This is because the derivations in these aforementioned papers require that selection is very strong ($N_e s >> 1$), such that fixation occurs very quickly relative to one unit of the time scale of the effective population size N_e, which corresponds to N_e generations.

Later research found two other important signatures of the selective sweep model: (i) shifts in the site frequency spectrum (SFS) of polymorphisms such as an excess of low-frequency (Braverman *et al.* 1995) and high-frequency (Fay and Wu 2000) derived alleles, and (ii) characteristic patterns of linkage disequilibrium (LD), such as an elevated level of LD during the fixation process and a complete break-down of LD across the selected site after fixation (Kim and Nielsen 2004; Stephan *et al.* 2006).

These features of the selective sweep model have been used to infer signatures of strong positive directional selection in the genomes of recombining organisms. Kim and Stephan (2002) developed a composite-likelihood ratio (CLR) test to detect local reductions of nucleotide diversity along a recombining chromosome, and to predict the strength and location of the target of selection. The CLR test compares the probability of the observed polymorphism data under the standard neutral model (*i.e.*, constant population size) with the probability of the data under the model of a selective sweep. Since the null and the alternative hypotheses in the CLR test are explicitly modeled, the interpretation of the test results is straightforward. On the other hand, it is important to note that the null hypothesis of the test is formulated based on the standard neutral model. This means that a violation of the assumptions of the null hypothesis may influence the

results and favor the alternative hypothesis. Therefore, the application of the CLR test is not appropriate for detecting selective sweeps when severe demographic events have occurred in the history of a population. In this case an additional approach may be used to distinguish sweeps from demography (in particular bottlenecks), which is described in the next section.

The first genome scans for selective sweeps were performed on humans (Akey *et al.* 2002), *D. melanogaster* (Harr *et al.* 2002; Schlötterer 2002; Glinka *et al.* 2003; Orengo and Aguadé 2004), and mice (Ihle *et al.* 2006). The data consisted of large numbers (∼100) of mostly randomly chosen loci along chromosomal regions of normal recombination rates. Both microsatellite and nucleotide variation were investigated. In all these studies, a surprisingly large number of loci showed signatures of selective sweeps, such as reduced variation or high differentiation between subpopulations. This immediately suggested that not all of these signatures were due to directional selection. As the populations that were examined were mostly derived, demographic factors (in particular bottlenecks during the colonization of new habitats) needed to be taken into account.

Selective sweep in a population undergoing demographic changes

Jensen *et al.* (2005) showed that the CLR test is not robust in the case of recent strong bottlenecks. Under this scenario, the false-positive rate may be as high as 80%, depending on the severity of the bottleneck. They proposed to use in addition to the CLR test a goodness-of-fit (GOF) approach to distinguish between the true positives that come from the rejection of the standard neutral scenario because of a sweep, and the false positives that come from the rejection of the standard neutral model due to demography. The combined CLR and GOF tests have been used extensively to analyze subgenomic data, *i.e.*, data from local genomic regions (reviewed elsewhere, *e.g.*, Pavlidis *et al.* 2008 and Stephan 2010a).

The availability of whole-genome or chromosome-wide SNP data, mainly from the HapMap Project (International HapMap Consortium 2003), motivated Nielsen *et al.* (2005) to develop a more general method (called SweepFinder) that could also be applied to genome-wide data. This test is based on the CLR approach of Kim and Stephan (2002). However, it differs from the latter in that the null hypothesis is not derived from the standard neutral model, but estimated from the empirical background distribution of the data. It therefore may take deviations from the constant-population size neutral model into account, at least to some extent.

Although SweepFinder may be robust against some demographic scenarios that have been investigated by Nielsen *et al.* (2005), simulations have shown that this does not hold in general, especially in cases of recent severe bottlenecks (Pavlidis *et al.* 2008). We have therefore incorporated LD information into the methods for detecting targets of positive directional selection that thus far have been based on the SFS alone. As suggested by the simulations of Jensen *et al.* (2007), the statistic ω proposed by Kim and Nielsen (2004) may be very powerful in distinguishing demographic from selective scenarios. Indeed, analyzing the correlation of SweepFinder and ω has enabled us to separate selection from demography even for rather deep bottlenecks (Pavlidis *et al.* 2010).

The more recent developments have focused on advancing the computational power of the sweep tests. The approach of Boitard *et al.* (2009), which is based on hidden Markov models and machine-learning, appears to perform better than the original methods by Kim and Stephan (2002) and Nielsen *et al.* (2005) in the case of bottlenecks. A computationally advanced CLR-based test is SweeD (Pavlidis *et al.* 2013). It includes a demographic model with an arbitrary number of instantaneous changes in population size and is applicable to large datasets. Another recent sweep test is OmegaPlus, which is a very fast algorithm that utilizes only LD information (Alachiotis *et al.* 2012). Finally, the algorithm just published by Akbari *et al.* (2018) claims that it can pinpoint the causative mutation of a single sweep in a large genomic region without prior knowledge of demography or functional annotations of mutations.

Selective sweep in a substructured population

In a panmictic population, the fixation of a strongly advantageous allele may occur very rapidly on the time scale of the effective population size N_e (Kaplan *et al.* 1989; Stephan *et al.* 1992). In contrast, in a subdivided population this process may take much longer, especially when migration is reduced (Slatkin and Wiehe 1998; Whitlock 2003; Kim and Maruki 2011). As a consequence, the hitchhiking process may usually not be complete, but ongoing (in the total population). Incomplete sweeps are often observed in humans (Nielsen *et al.* 2007), because in this case limited migration and also strong population size expansion slow down fixation of beneficial alleles.

Theoretical predictions of the effect of sweeps on genetic differentiation have been obtained by several authors. In the case of reduced migration a sweep in a subdivided population due to sequential fixation of the beneficial allele increases differentiation if neutral variation near the selected site is relatively homogenous across subpopulations initially (Slatkin and Wiehe 1998; Bierne 2010). On the other hand, if the subpopulations are initially differentiated, hitchhiking of the same beneficial allele will decrease F_{ST} (Santiago and Caballero 2005).

The CLR method has been extended to substructured populations by Chen *et al.* (2010). However, population size changes have not been considered in this approach (called XP-CLR test).

If a selective sweep is ongoing, the hitchhiking haplotype is expected to be rather long due to strong LD (see above). This feature of the hitchhiking effect has been exploited in model-free, haplotype-based tests, such as the iHS test (Sabeti *et al.* 2002; Voight *et al.* 2006; Tang *et al.* 2007). The decay of the

haplotype length due to recombination is slower if the haplotypes are driven by positive selection.

Yet another class of tests compares polymorphism data from two or more subpopulations to find evidence for local adaptation. Different selection pressures between demes may lead to strong genetic differentiation that can be measured by F_{ST}. Bayesian approaches have been used to reveal genomic regions that have experienced recent strong positive directional selection and hence large F_{ST}, although these methods may lead to overestimation of F_{ST} and should only be used after careful analysis of the population structure (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler *et al.* 2008).

Joint inference of demographic and selective forces

In the inference of selective sweeps described above, we have assumed that the demographic history of a population is not confounded by forms of weak selection that concern the whole genome. This, however, is only a rough approximation, especially for populations with large effective size. For this reason it would be desirable to infer the demographic and selective history jointly, which means that all selective processes and demography are analyzed simultaneously. In the case of background selection and selective sweeps, both of which reduce levels of genetic diversity and are hard to distinguish (particularly in regions of low recombination, Stephan 2010b), an early attempt has been made to study their joint effects (Kim and Stephan 2000). However, progress along this line has been made only very recently. Comeron (2014) proposed background selection as a sensible null hypothesis for investigating the presence of other forms of linked selection, such as directional selection, and Elyashiv *et al.* (2016) modeled the joint effects of background selection and selective sweeps. Both applied their approaches to *D. melanogaster* data, utilizing the detailed genetic maps available for this species. Furthermore, Huber *et al.* (2016) extended SweepFinder to detect selective sweeps while controlling for background selection.

Selective Sweeps at Multiple Loci

In this section, I will first discuss recurrent selective sweeps; *i.e.*, sweeps that occur sequentially at multiple selected loci, because at any time at most one beneficial allele is assumed to be on the way to fixation. Then I will describe the models on competing sweeps, also called evolutionary traffic; *i.e.*, we will allow for interference between simultaneously occurring sweeps.

Recurrent selective sweeps

Given the relatively high rate of selective sweeps at individual loci for species with large effective population size such as *D. melanogaster*, the question arises whether a model of recurrent sweeps is more appropriate in describing the data than a model of single sweeps at individual loci. The model of selective sweeps at individual loci described above can be extended to multiple loci in a straightforward way by assuming that sweeps occur along the genome independently according to a time-homogeneous Poisson process at rate ν per site per generation (Kaplan *et al.* 1989). Using this assumption, Wiehe and Stephan (1993) derived a simple formula quantifying the expected level of equilibrium nucleotide diversity, π, along the genome given the recombination rate, ρ, per generation per nucleotide site and the intensity of selection, $\alpha = 2N_e s$, where s is the average selection coefficient of strong beneficial substitutions in the genome:

$$\pi = \pi_0 \frac{\rho}{\rho + \kappa\alpha\nu}. \quad (1)$$

Here, π_0 is the neutral equilibrium level of diversity and $\kappa = 0.075$ is a constant.

In contrast to the single-sweep model, simulations have shown that the targets of selection are difficult to localize based on this recurrent hitchhiking model (Pavlidis *et al.* 2010). The frequency of advantageous substitutions, on the other hand, can be estimated rather accurately (Jensen *et al.* 2008). The above equation suggests that the parameters α and ν cannot be estimated individually but only as a product (Stephan 1995). This would mean that frequent weak beneficial substitutions and rare strongly selected substitutions predict similar average effects on linked neutral variation. However, utilizing the insight that rare strong selection increases the variance of the common summary statistics of nucleotide heterozygosity along the genome relative to ubiquitous weak selection, the ABC approach of Jensen *et al.* (2008) allows distinguishing between these alternatives and the estimation of α and ν separately.

Campos *et al.* (2017) considered models of both background selection and recurrent selective sweeps, including gene conversion in addition to crossing-over. Using *Drosophila* data, they showed that gene conversion may have large effects on the parameter estimates of these models.

The effect of recurrent selective sweeps on the SFS of neutral polymorphisms has been analyzed by Kim (2006). He showed that the excess of high-frequency derived alleles, a hallmark of single sweeps (Fay and Wu 2000; Przeworski 2002), disappears under recurrent selective sweeps.

Competing selective sweeps

Next, I will discuss the case in which selective sweeps along the genome do not occur sequentially, but interfere with each other. Such evolutionary traffic of interfering positive fixations has been described by several authors (Barton 1995; Kirby and Stephan 1996; Yu and Etheridge 2010; Bossert and Pfaffelhuber 2016), but the impact on linked neutral variation is not well understood. To my knowledge, only two studies have modeled genetic hitchhiking in the presence of interference between partially linked beneficial alleles on their way to fixation. Using full-forward simulations and analytical approximations, Kim and Stephan (2003) found that interference between linked beneficial alleles causes a reduction of their fixation probability. The hitchhiking effect on neutral

variation for a given substitution also decreases slightly due to interference. As a result, the strength of recurrent selective sweeps is weakened. However, this effect is significant only in chromosomal regions of low recombination rates (*e.g.*, around the centromeres in *Drosophila*). Therefore, the results on recurrent sweeps derived for the case that at most one beneficial allele is on the way to fixation are still largely valid, at least in chromosomal regions of normal recombination.

Chevin *et al.* (2008) explicitly modeled the case of two closely linked, selected loci, and one neutral locus for infinitely large populations using ordinary differential equations. Similar to Kim and Stephan (2003), they also observed a weaker hitchhiking effect than for a single sweep of comparable selection strength. Most interestingly, the interference of both fixation processes may lead for some initial conditions and, in some parameter ranges, to an excess of intermediate-frequency variants in the genomic region between the selected sites, which may be interpreted falsely as a sign of balancing selection. The reason is that, when the beneficial alleles arise on different chromosomes, they need to recombine into one chromosome to go to fixation, which can take a long time and thus increase genetic variation. This phenomenon is related to the case of associative overdominance (Frydenberg 1963; Zhao and Charlesworth 2016), which was the first process studied in which selection on sites in genomes affects neutral variants (Sved 1968; see also the review by Charlesworth and Charlesworth 2018).

Soft Sweeps

As explained above, a selective sweep arises if a beneficial allele occurs at some recent time in the past and goes to fixation. The sudden occurrence of the beneficial allele may be caused by mutation, migration from another subpopulation, or may be due to a very rare allele in the standing variation after an environmental change. In contrast, a different case of linked selection occurs if the driving favorable allele is from the standing variation, but not in very low frequency when the environmental shift occurs. This may lead to a so-called *soft sweep* (Hermisson and Pennings 2005).

This process is different from a selective sweep because, in a soft sweep, the driving beneficial allele was present in a population before the environment changed; *i.e.*, it was segregating neutrally or at an appreciable frequency under a mutation-selection balance, and was thus present on more than one haplotype. Several authors (Orr and Betancourt 2001; Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski *et al.* 2005) examined the frequency at which an allele must be segregating before the shift in selection pressure such that a soft sweep arises. This frequency depends on the size of several relatively unknown parameters, which makes it difficult to predict theoretically how often soft sweeps occur.

In practice, another problem is to distinguish soft sweeps from classical selective sweeps in the data because of multiple confounding effects. Jensen (2014) discussed a widely cited example of selection on standing variation describing this point, the *Eda* locus in sticklebacks (Colosimo *et al.* 2005):

> With evidence for selection reducing armor plating in freshwater populations compared with the ancestral heavily plated marine populations, the authors sequenced marine individuals to estimate the allele frequency of the freshwater adaptive low plate morphs, with estimates ranging from 0.2 to 3.8%. While the low plate morph is likely deleterious in marine populations (potentially suggesting that it is at mutation-selection balance), migration from the marine environment may indeed serve as an important source of variation for local freshwater selective sweeps. However, as noted by the authors, it is difficult to separate this hypothesis from that of local freshwater adaptation on new mutations, followed by back migration of locally adapted alleles into the marine population (Jensen 2014).

On the other hand, Hermisson and Pennings (2017) reported cases of soft sweeps in various organisms, including eukaryotes. Perhaps the best example is lactase persistence in humans. At the lactase gene *LCT*, more than a single haplotype has been found in some local African populations, which would indeed be indicative of a soft sweep.

Some authors have pushed the idea that soft sweeps may also arise from multiple adaptive mutations (where the mutations are meant to occur at the same nucleotide site). As I discussed elsewhere (Stephan 2016), this would require extremely large effective population sizes and/or nucleotide mutation rates. Readers interested in this model and its applications to data are referred to the work of Petrov and colleagues, who claim to have evidence for soft sweeps from multiple beneficial mutations (*e.g.*, Karasov *et al.* 2010).

Conclusions

There has been much progress in identifying selective sweeps underlying a range of adaptations. In particular, in organisms with large effective population sizes, such as *D. melanogaster*, the evidence for sweeps is quite striking (reviewed in Stephan 2010a). There is also agreement that sweeps may be detected with reasonably high confidence if the demographic history of a population is taken into account, except in the case of some complex demographies such as recent severe population size bottlenecks (Pavlidis *et al.* 2010). On the experimental side, however, the search for causative nucleotide changes that led to selective sweeps has started only recently (Saminadin-Peter *et al.* 2012; Voigt *et al.* 2015; Catalán *et al.* 2016), although sweep mapping may lead to quite accurate identification of the targets of selection. Clearly, there is a lot of room for future research activities in this latter area. One should keep in mind that all of the methods discussed can only create hypotheses about the regions under selection, which should be tested by manipulative experiments.

On the theoretical side, although the theoretical advances in the detection of positive selection in genomes are impressive, some aspects need further attention. First, the evolution of interacting selective sweeps (the traffic model) is still

largely unexplored. We still lack predictions of this model about the distribution of variation across the genome around the selected sites and on the SFS. Second, efforts of estimating demography and weak selection jointly have so far not led to computer programs that are applicable to data. Third, population subdivision is not well incorporated into the sweep approaches yet (except in F_{ST}-based methods such as BayeScan), but the mathematical treatment of this case is very difficult (Greven *et al.* 2016).

Acknowledgments

I thank the editors of *GENETICS* for inviting me to contribute this *Perspectives* article. I am particularly grateful to Adam Wilkins and Deborah Charlesworth, as well as three reviewers, for their constructive suggestions on a previous version of this paper. My current research is supported by grant STE 325/17 from the Priority Program 1819 of the Deutsche Forschungsgemeinschaft (DFG).

Literature Cited

Aguadé, M., N. Miyashita, and C. H. Langley, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics 122: 607–615.

Akbari, A., J. J. Vitti, A. Iranmehr, M. Bakhtiari, P. C. Sabeti *et al.*, 2018 Identifying the favored mutation in a positive selective sweep. Nat. Methods 15: 279–282. https://doi.org/10.1038/nmeth.4606

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12: 1805–1814. https://doi.org/10.1101/gr.631202

Alachiotis, N., A. Stamatakis, and P. Pavlidis, 2012 OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics 28: 2274–2275. https://doi.org/10.1093/bioinformatics/bts419

Atwood, K. C., L. K. Schneider, and F. J. Ryan, 1951 Periodic selection in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 37: 146–155. https://doi.org/10.1073/pnas.37.3.146

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics 140: 821–841.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133. https://doi.org/10.1017/S0016672398003462

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13: 969–980.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rate in *D. melanogaster*. Nature 356: 519–520. https://doi.org/10.1038/356519a0

Berry, A. J., J. W. Ajioka, and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 129: 1111–1119.

Bierne, N., 2010 The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. Evolution 64: 3254–3272. https://doi.org/10.1111/j.1558-5646.2010.01050.x

Boitard, S., C. Schlötterer, and A. Futschik, 2009 Detecting selective sweeps: a new approach based on hidden Markov models. Genetics 181: 1567–1578. https://doi.org/10.1534/genetics.108.100032

Bossert, S., and P. Pfaffelhuber, 2016 The fixation probability and time for a doubly beneficial mutant. arXiv:1610.06613

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796.

Campos, J. L., L. Zhao, and B. Charlesworth, 2017 Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. Proc. Natl. Acad. Sci. USA 114: E4762–E4771. https://doi.org/10.1073/pnas.1619434114

Catalán, A., A. Glaser-Schmitt, E. Argyridou, P. Duchen, and J. Parsch, 2016 An indel polymorphism in the *MtnA 3′* untranslated region is associated with gene expression variation and local adaptation in *Drosophila melanogaster*. PLoS Genet. 12: e1005987. https://doi.org/10.1371/journal.pgen.1005987

Charlesworth, B., and D. Charlesworth, 2018 Neutral variation in the context of selection. Mol. Biol. Evol. 35: 1359–1361. https://doi.org/10.1093/molbev/msy062

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

Chen, H., N. Patterson, and D. Reich, 2010 Population differentiation as a test for selective sweeps. Genome Res. 20: 393–402. https://doi.org/10.1101/gr.100545.109

Chevin, L.-M., S. Billiard, and F. Hospital, 2008 Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. Genetics 180: 301–316. https://doi.org/10.1534/genetics.108.089706

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villareal, M. Dickson *et al.*, 2005 Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307: 1928–1933. https://doi.org/10.1126/science.1107239

Comeron, J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 10: e1004434. https://doi.org/10.1371/journal.pgen.1004434

Comeron, J. M., 2017 Background selection as a null hypothesis in population genomics: insights and challenges from Drosophila studies. Philos. Trans. R. Soc. Lond. B Biol. Sci. 372: 20160471. https://doi.org/10.1098/rstb.2016.0471

Elyashiv, E., S. Sattah, T. T. Hu, A. Strutsovsky, G. McVicker *et al.*, 2016 A genomic map of the effects of linked selection in Drosophila. PLoS Genet. 12: e1006130. https://doi.org/10.1371/journal.pgen.1006130

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

Foll, M., and 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. https://doi.org/10.1534/genetics.108.092221

Frydenberg, O., 1963 Population studies of a lethal mutant in *Drosophila melanogaster*. I. Behaviour in populations with discrete generations. Hereditas 50: 89–116. https://doi.org/10.1111/j.1601-5223.1963.tb01896.x

Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155: 909–919.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*. Genetics 165: 1269–1278.

Greven, A., P. Pfaffelhuber, C. Pokalyuk, and A. Wakolbinger, 2016 The fixation time of a strongly beneficial allele in a structured population. Electron. J. Probab. 21: 1–42. https://doi.org/10.1214/16-EJP3355

Harr, B., M. Kauer, and C. Schlötterer, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 99: 12949–12954 [corrigenda: Proc. Natl. Acad. Sci. USA 100: 3004 (2003)]. https://doi.org/10.1073/pnas.202336899

Hellmann, I., I. Ebersberger, S. Ptak, S. Pääbo, and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. 72: 1527–1535. https://doi.org/10.1086/375657

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352. https://doi.org/10.1534/genetics.104.036947

Hermisson, J., and P. S. Pennings, 2017 Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol. 8: 700–716. https://doi.org/10.1111/2041-210X.12808

Huber, C. D., M. DeGiorgio, I. Hellmann, and R. Nielsen, 2016 Detecting recent selective sweeps while controlling for mutation rate and background selection. Mol. Ecol. 25: 142–156. https://doi.org/10.1111/mec.13351

Ihle, S., I. Ravaoarimanana, M. Thomas, and D. Tautz, 2006 An analysis of signatures of selective sweeps in natural populations of the house mouse. Mol. Biol. Evol. 23: 790–797. https://doi.org/10.1093/molbev/msj096

Innan, H., and Y. Kim, 2004 Patterns of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA 101: 10667–10672. https://doi.org/10.1073/pnas.0401720101

International HapMap Consortium, 2003 The international HapMap Project. Nature 426: 789–796. https://doi.org/10.1038/nature02168

Jensen, J. D., 2014 On the unfounded enthusiasm for soft selective sweeps. Nat. Commun. 5: 5281. https://doi.org/10.1038/ncomms6281

Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–1410. https://doi.org/10.1534/genetics.104.038224

Jensen, J. D., K. R. Thornton, C. D. Bustamante, and C. F. Aquadro, 2007 On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176: 2371–2379. https://doi.org/10.1534/genetics.106.069450

Jensen, J. D., K. R. Thornton, and P. Andolfatto, 2008 An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. PLoS Genet. 4: e1000198. https://doi.org/10.1371/journal.pgen.1000198

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The 'hitchhiking effect' revisited. Genetics 123: 887–899.

Karasov, T., P. W. Messer, and D. A. Petrov, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PLoS Genet. 6: e1000924. https://doi.org/10.1371/journal.pgen.1000924

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics 172: 1967–1978. https://doi.org/10.1534/genetics.105.048447

Kim, Y., and T. Maruki, 2011 Hitchhiking effect of a beneficial mutation spreading in a subdivided population. Genetics 189: 213–226. https://doi.org/10.1534/genetics.111.130203

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524. https://doi.org/10.1534/genetics.103.025387

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 155: 1415–1427.

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

Kim, Y., and W. Stephan, 2003 Selective sweeps in the presence of interference among partially linked loci. Genetics 164: 389–398.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of the *white* gene of *Drosophila melanogaster*. Genetics 144: 635–645.

Kraft, T., T. Säll, I. Magnusson-Rading, N.-O. Nilsson, and C. Halldén, 1998 Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics 150: 1239–1244.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35. https://doi.org/10.1017/S0016672300014634

Nachman, M. W., V. L. Bauer, S. L. Crowell, and C. F. Aquadro, 1998 DNA variability and recombination rates at *X*-linked loci in humans. Genetics 150: 1133–1141.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575. https://doi.org/10.1101/gr.4252305

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. 8: 857–868. https://doi.org/10.1038/nrg2187

Ohta, T., and M. Kimura, 1975 The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). Genet. Res. 25: 313–326. https://doi.org/10.1017/S0016672300015731

Orengo, D. J., and M. Aguadé, 2004 Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics 167: 1759–1766. https://doi.org/10.1534/genetics.104.028969

Orr, H. A., and A. J. Betancourt, 2001 Haldane's sieve and adaptation from standing genetic variation. Genetics 157: 875–884.

Pavlidis, P., S. Hutter, and W. Stephan, 2008 A population genomic approach to map recent positive selection in model species. Mol. Ecol. 17: 3585–3598.

Pavlidis, P., J. D. Jensen, and W. Stephan, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907–922. https://doi.org/10.1534/genetics.110.116459

Pavlidis, P., D. Zivkovic, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30: 2224–2234. https://doi.org/10.1093/molbev/mst112

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

Przeworski, M., G. Coop, and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution 59: 2312–2323. https://doi.org/10.1554/05-273.1

Riebler, A., L. Held, and W. Stephan, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics 178: 1817–1829. https://doi.org/10.1534/genetics.107.081281

Sabeti, P., D. E. Reich, J. M. Higgins, H. Z. P. Levine, J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837. https://doi.org/10.1038/nature01140

Saminadin-Peter, S. S., C. Kemkemer, P. Pavlidis, and J. Parsch, 2012 Selective sweep of a *cis* regulatory sequence in a non-African population of *Drosophila melanogaster*. Mol. Biol. Evol. 29: 1167–1174. https://doi.org/10.1093/molbev/msr284

Santiago, E., and A. Caballero, 2005 Variation after a selective sweep in a subdivided population. Genetics 169: 475–483. https://doi.org/10.1534/genetics.104.032813

Schlötterer, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 160: 753–763.

Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. 71: 155–160. https://doi.org/10.1017/S001667239800319X

Stephan, W., 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Mol. Biol. Evol. 12: 959–962.

Stephan, W., 2010a Detecting strong positive selection in the genome. Mol. Ecol. Resour. 10: 863–872. https://doi.org/10.1111/j.1755-0998.2010.02869.x

Stephan, W., 2010b Genetic hitchhiking versus background selection: the controversy and its implications. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1245–1253

Stephan, W., 2016 Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol. Ecol. 25: 79–88. https://doi.org/10.1111/mec.13288

Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics 121: 89–99.

Stephan, W., and C. H. Langley, 1998 DNA polymorphism in Lycopersicon and crossing-over per physical length. Genetics 150: 1585–1593.

Stephan, W., T. H. E. Wiehe, and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. 41: 237–254. https://doi.org/10.1016/0040-5809(92)90045-U

Stephan, W., Y. S. Song, and C. H. Langley, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172: 2647–2663. https://doi.org/10.1534/genetics.105.050179

Sved, J. A., 1968 The stability of linked systems of loci with a small population size. Genetics 59: 543–563.

Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e171. https://doi.org/10.1371/journal.pbio.0050171

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72 (erratum: PLoS Biol. 4: e154); [corrigenda: PLoS Biol. 5: e147 (2007)]. https://doi.org/10.1371/journal.pbio.0040072

Voigt, S., S. Laurent, M. Litovchenko, and W. Stephan, 2015 Positive selection at the *polyhomeotic* locus leads to reduced thermosensitivity of gene expression in temperate *Drosophila melanogaster*. Genetics 200: 591–599. https://doi.org/10.1534/genetics.115.177030

Whitlock, M. C., 2003 Fixation probability and time in subdivided populations. Genetics 164: 767–779.

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. 10: 842–854.

Yu, F., and A. M. Etheridge, 2010 The fixation probability of two competing beneficial mutations. Theor. Popul. Biol. 78: 36–45. https://doi.org/10.1016/j.tpb.2010.04.001

Zhao, L., and B. Charlesworth, 2016 Resolving the conflict between associative overdominance and background selection. Genetics 203: 1315–1334. https://doi.org/10.1534/genetics.116.188912

Communicating editor: A. S. Wilkins