MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 79–88 doi: 10.1111/mec.13288

DETECTING SELECTION IN NATURAL POPULATIONS: MAKING SENSE OF GENOME SCANS AND TOWARDS ALTERNATIVE SOLUTIONS

Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation

WOLFGANG STEPHAN*†

*Biocenter, Department of Biology, Ludwig-Maximilian University Munich, Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany, †Museum für Naturkunde, Berlin, Germany

Abstract

In the past 15 years, numerous methods have been developed to detect selective sweeps underlying adaptations. These methods are based on relatively simple population genetic models, including one or two loci at which positive directional selection occurs, and one or two marker loci at which the impact of selection on linked neutral variation is quantified. Information about the phenotype under selection is not included in these models (except for fitness). In contrast, in the quantitative genetic models of adaptation, selection acts on one or more phenotypic traits, such that a genotype–phenotype map is required to bridge the gap to population genetics theory. Here I describe the range of population genetic models from selective sweeps in a panmictic population of constant size to evolutionary traffic when simultaneous sweeps at multiple loci interfere, and I also consider the case of polygenic selection characterized by subtle allele frequency shifts at many loci. Furthermore, I present an overview of the statistical tests that have been proposed based on these population genetics models to detect evidence for positive selection in the genome.

Keywords: adaptation, evolution, evolutionary theory, natural selection, population genetics, quantitative genetics

Received 30 April 2015; revision accepted 19 June 2015

Introduction

In the past 15 years, numerous methods have been developed to find evidence for positive selection in the genomes of natural populations (Kim & Stephan 2002; Akey *et al.* 2004; Nielsen *et al.* 2005, and many more). The specific aims of these studies have been to (i) detect signatures of positive directional selection in the genomes of natural (sexually recombining) populations, (ii) estimate the strength of selection and (iii) localize the targets of selection. The ultimate goal has been to identify the genes targeted by selection and to characterize their associated functions and phenotypes.

The population genetic models underlying these methods consider mostly single loci at which positive selection acts, with occasional extensions to multilocus models. In more recent time, however, due to the advance of genomewide association studies (GWAS), polygenic selection was also studied using quantitative genetics models that are formulated in terms of allele frequencies. That is, in contrast to the population genetics, single- and multilocus cases, in these latter types of models selection acts on a phenotypic trait, and a genotype–phenotype map is assumed to bridge the gap to population genetics (de Vladar & Barton 2014).

In most cases considered here, neutral or weakly selected marker loci are included that are linked to the genetic loci under selection. In other words, we consider the selection models in the context of genetic hitchhiking (Maynard Smith & Haigh 1974). Based on this principle, genomic signatures of selection may be inferred. I will describe the models underlying the statistical approaches to detect evidence for positive selection and briefly sketch the inference methods based on them. For any type of adaptive scenario, I will start

Correspondence: Wolfgang Stephan, Fax: 0049 89 2180 74104; E-mail: stephan@bio.lmu.de

with a basic model and then introduce complexities, such as demography, population structure and confounding selection regimes, to make the models more realistic and applicable to data. I will consider three classes of models: (i) single-locus models of selection, (ii) multilocus models and (iii) models of polygenic selection.

Selective sweeps at individual loci

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 rare allele in the standing variation after an environment change. If this beneficial allele is strongly selected, it is inevitable that the frequency of linked neutral (or weakly selected) variants increases. In a seminal paper, Maynard Smith & Haigh (1974) described this process and termed it genetic hitchhiking. They showed that in very large populations, hitchhiking can drastically reduce genetic variation at or near the site of selection, which is now also called 'selective sweep' (Berry *et al.* 1991).

In the recent literature, selective sweeps arising from new beneficial mutations have been called hard sweeps and those due to beneficial alleles from the standing variation or due to multiple new mutations at the same site as soft sweeps (Hermisson & Pennings 2005). However, in this review I do not make this distinction. As I am concentrating on initially rare alleles leading to sweeps, it is difficult to distinguish these processes in eukaryotic recombining species. Furthermore, the soft sweep model assuming multiple new beneficial mutations requires extremely large values of $N_e\mu$ where N_e is effective population size and μ the beneficial nucleotide mutation rate (Jensen 2014). Readers interested in this model and its application to data are referred to the work of Petrov and colleagues who claim to have evidence for sweeps due to multiple beneficial mutations (e.g. Karasov *et al.* 2010).

According to Maynard Smith and Haigh's deterministic 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 distance to the selected site r and the selection coefficient s. For finite populations of large, constant size, the results obtained by coalescent (Kaplan *et al.* 1989) and diffusion approximations (Stephan *et al.* 1992) are in excellent agreement with Maynard Smith and Haigh's deterministic predictions. This is because the derivations in these aforementioned studies require that selection is very strong ($N_e s \gg 1$) such that fixation occurs nearly instantaneously on the timescale of the effective population size N_e.

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

These features of the hitchhiking model have been used to infer signatures of strong positive directional selection in the genomes of recombining organisms. Kim & 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 above model of genetic hitchhiking. As the null and the alternative hypotheses in the CLR test are explicitly modelled, 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 favour 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 next.

Selective sweep in a population of varying size

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 analyse

subgenomic data, that is data from local genomic regions (reviewed elsewhere; see Pavlidis *et al.* 2008; Stephan 2010a or elsewhere in this volume).

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 genomewide data. This test is based on the CLR approach of Kim and Stephan. However, it differs from the latter one 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 & Nielsen (2004) may be very powerful in distinguishing demographic from selective scenarios. Indeed, analysing the correlation of SweepFinder and ω has enabled us to separate selection from demography for rather deep bottlenecks (Pavlidis *et al.* 2010). However, inferring positive selection in strongly bottlenecked populations remains difficult (Poh *et al.* 2014). Currently, the most 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.

Selective sweep in a substructured population

In a panmictic population, the fixation of a strongly advantageous allele occurs very rapidly on the timescale 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 & Wiehe 1998; Whitlock 2003; Kim & 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.

Theoretical predictions on the effect of hitchhiking on genetic differentiation have been obtained by several authors. In the case of reduced migration, hitchhiking 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 & Wiehe 1998; Bierne 2010). On the other hand, if the subpopulations are initially relatively differentiated, hitchhiking of the same beneficial allele will decrease F_{ST} (Santiago & 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 (Stephan *et al.* 2006). This feature of the hitchhiking effect has been exploited in model-free, haplotype-based tests (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. Because of a lack of theoretical predictions, these tests are usually employed in a statistical outlier approach.

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} (Beaumont & Balding 2004; Foll & Gaggiotti 2008; Riebler *et al.* 2008). If positive selection is not strong (as, for instance, for polygenic selection), correlation methods may be used (discussed below).

On the 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 weak selection. This, however, is only a rough approximation, especially for populations with large effective size, such as *Drosophila melanogaster* (Pool *et al.* 2012) or *D. ananassae* (Das *et al.* 2004). For this reason, it would be desirable to infer the demographic and selective history jointly, which means that all selective processes (of positive and negative selection) and demography are analysed simultaneously. Although the likelihood-based and Bayesian methods have been greatly improved in recent years, such an undertaking is currently not possible. However, first attempts have been made to combine some of the demographic and selective forces that make similar predictions on the observed patterns of genetic variation, such as the SFS or average levels of diversity. For

example, population size expansion and purifying selection lead to an excess of low-frequency derived alleles on a genomewide scale. Thus, Zivkovic *et al.* (2015) used diffusion theory to derive properties of the SFS under the joint action of drift, demography and weak purifying selection. They address questions such as this: if neutrality is incorrectly assumed when there is selection, which effect does this have on the estimation of the demographic parameters? Similarly, as the effects of background selection and selective sweeps reduce the levels of genetic diversity but are hard to distinguish (particularly in regions of low recombination; Stephan 2010b), attempts have been made to study their joint effect (Kim & Stephan 2000). This may lead to new insights into the way and extent these forces interact (G. Sella, personal communication).

Selective sweeps at multiple loci

In this section, I will first discuss recurrent selective sweeps, that is 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 review the models on competing sweeps, also called evolutionary traffic; that is, 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* (as estimated by several authors, including Li & Stephan (2006)), 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 genetic hitchhiking at individual loci described above can be extended to multiple loci in a straightforward way by assuming that hitchhiking events 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 & 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:

$$\pi = \pi_0 \frac{\rho}{\rho + \kappa\alpha\nu} \qquad \text{eqn 1}$$

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 relative to ubiquitous weak selection, the ABC approach of Jensen *et al.* (2008) allows distinguishing between these alternatives and the estimation α and ν separately.

The effect of recurrent selective sweeps on the SFS of neutral polymorphisms has been analysed by Kim (2006). He showed that the excess of high-frequency derived alleles, a hallmark of single sweeps (Fay & 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 & Stephan 1996; Yu & Etheridge 2010), but the impact on linked neutral variation is not well understood. To my knowledge, only two studies have modelled 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 & Stephan (2003) found that interference between linked beneficial alleles causes a reduction in their fixation probability. The hitchhiking effect on neutral variation for a given substitution also slightly decreases 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 modelled the case of two closely linked, selected loci and one neutral locus for infinitely large populations using ordinary differential equations (ODEs). Similar to Kim & 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 falsely be interpreted 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 that can take a long time and thus increase genetic variation.

Quantitative genetic models of adaptation

Quantitative genetic models of adaptation date back to the time before the genetic mechanisms of inheritance were discovered (Orr 2005). In contrast to the models discussed above, the quantitative genetic models contain an explicit description of metric phenotypes. These phenotypes are characterized by a distribution of gradual differences among individuals of a population. The distribution of a trait is smooth if the number of genes controlling a trait is large. In varying environments, different phenotypes may be favoured. This leads to a change in the population mean phenotype that is known to depend on additive genetic variance. When a population deviates from its optimum, mutations are favoured according to their effect size and distance to the optimum.

Selective sweeps in quantitative genetic models of adaptation

We first ask the question whether selective sweeps that are the hallmark of models of positive directional selection at individual sites are also observed in quantitative genetic models of adaptation. Per definition, selective sweeps only play a role in polygenic selection if the beneficial alleles are driven to fixation from very low frequency by strong selection. However, such events occur very rarely, as first hypothesized by Chevin & Hospital (2008). These authors presented a model for the footprint of positive directional selection at a quantitative trait locus (QTL) in the presence of a fixed amount of background genetic variation due to other loci. Their approach is based on Lande's (1983) model that consists of a locus of major effect on the trait and treats the remaining loci of minor effect as genetic background. Their analysis predicts that QTL of adaptive traits under stabilizing selection, which is thought to be the most common form of selection on quantitative traits, exhibits patterns of selective sweeps only very rarely.

Pavlidis *et al.* (2012) analysed an explicit multilocus model with two to eight loci controlling an additive quantitative trait under stabilizing selection (with and without genetic drift). Using simulations, they showed that multilocus response to selection often prevents trajectories from going to fixation, particularly for the symmetric viability model. They also found that the probability of fixation strongly depends on the genetic architecture of the trait, in that a larger number of loci led to fewer fixations. To understand these results in greater depth, Wollstein & Stephan (2014) analysed a two-locus model of symmetric viability selection. They observed that about 16% of the trajectories may lead to fixation if the initial allele frequencies are sampled from the neutral site frequency spectrum. However, if the population is pre-adapted when it undergoes an environmental change (i.e. sits in one of the equilibria of the model), the fixation probability decreases dramatically. In other two-locus models with general viabilities or an optimum shift, the proportion of adaptive fixations may increase to more than 20%. Similarly, genetic drift and loose linkage lead to a higher probability of fixation. Thus, these analyses showed that selective sweeps may occur in quantitative genetic models of adaptation in appreciable frequency. The restriction to only two loci and the impact of the initial conditions on the results, however, did not allow us to draw fully convincing conclusions.

Subtle to moderate allele frequency changes at many loci

Several authors have argued verbally that adaptation in natural populations occurs not by sweeps alone, but involves subtle allele frequency shifts at many loci controlling polygenic traits (e.g. Pritchard *et al.* 2010). Suppose that a population is well adapted when a sudden environmental change happens, which shifts the optimum of one or more traits. If there is sufficient heritable variation present in the population for the trait in question, then the population can adapt rapidly to the new environment. The speed of adaptation depends on the initial genetic variance and the strength of selection.

To understand these conjectures in a quantitative way, we have recently analysed a model that captures the polygenic response to stabilizing selection and mutation after an optimum shift (Jain & Stephan 2015). We considered a single trait z that is determined additively (no dominance or epistasis) by n di-allelic loci, where γ_i is the allelic effect at locus i. If the frequency of the '+' allele at locus i is denoted by x_i, then the mean phenotype $\bar{z}$, and the genetic variance v are given by

$$\bar{z} = \sum_{i=1}^{n} \gamma_i (2x_i - 1) \qquad \text{eqn 2}$$

and

$$v = 2 \sum_{i=1}^{n} \gamma_i^2 x_i (1 - x_i) \qquad \text{eqn 3}$$

(Barton 1986). We also assume that the fitness of an individual with trait value z follows a Gaussian distribution as z deviates from the optimum z_0

$$W(z) = \exp\left[-\frac{1}{2} s (z - z_0)^2\right], \qquad \text{eqn 4}$$

where s is the selection coefficient. Then, in an infinitely large, randomly mating population, the change in allele frequency at locus i due to selection and symmetric mutation is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i = -\frac{s}{2} \gamma_i x_i (1 - x_i)[2\Delta z + \gamma_i (1 - 2x_i)] + \mu(1 - 2x_i), \qquad \text{eqn 5}$$

where t is time, $i = 1, \ldots, n$, μ is the mutation rate, and $\Delta z = \bar{z} - z_0$. Because Δz contains a sum over allele frequencies, eqn (5) represents a system of coupled ODEs. This model is sufficiently realistic to analyse GWAS data. However, if needed, one could go further and model the mutation process in a more realistic way (for instance, by relaxing the symmetry assumption and assuming locus-specific mutation rates). This model is also more general than the two-locus model that Wollstein & Stephan (2014) analysed. It contains n loci and its selection term can be derived from the symmetric viability selection model under loose linkage. Most importantly, it lets us define the initial conditions in a natural way.

The selection part of the eqn (5) was first established by Wright (1935) and later Barton (1986) introduced the mutation term. de Vladar & Barton (2014) presented an analytical treatment of the equilibrium properties of the model and performed extensive numerical calculations. They found that the alleles may be classified into those with effects smaller than a threshold value $2\sqrt{2\mu/s}$ and those with larger sizes. At equilibrium with no deviation from the optimum, the frequency of the alleles of small effect is $\frac{1}{2}$, whereas the large-effect alleles are in a mutation–selection balance near zero (approximately at $2\mu/s\gamma^2$) or one at $1–2\mu/s\gamma^2$ if $2\mu \ll s\gamma^2$. We used these equilibrium results for our analyses described next.

So far, we studied the short-term response of a trait to selection. We assumed that the population is in equilibrium at a local peak with no deviation from the optimum that is at z_0. That means that the allele frequencies at the large-effect loci are in a mutation–selection balance either near zero or one, and the frequencies of the small-effect loci are at $\frac{1}{2}$, as described above. Then, suddenly the optimum is shifted to another value z_f. To calculate the response of the system after this optimum shift, note that in the biologically important case in which most loci have small effects on the trait (i.e. smaller than the threshold value), the full model defined by eqn (5) can be approximated for short times as (Jain & Stephan 2015)

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i = -s\gamma_i x_i (1 - x_i)\Delta z, \qquad \text{eqn 6}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \Delta z = -sv\Delta z. \qquad \text{eqn 7}$$

When most effects are small, it can be shown that the variance $v(t)$ may be approximated by the initial variance $v(0)$ as the small-effect loci contribute most to the variance (Jain & Stephan 2015). Under these circumstances, the ODEs (6) and (7) can be solved as

$$\Delta z(t) = \Delta z(0) e^{-sv(0)t} \qquad \text{eqn 8}$$

and

$$x_i(t) = \frac{x_i(0)}{x_i(0) + (1 - x_i(0)) e^{\gamma_i \Delta z(0)/v(0)(1 - e^{-sv(0)t})}} \qquad \text{eqn 9}$$

Two points are remarkable about these equations. First, the mean phenotype approaches the new optimum in about $(sv\ (0))^{-1}$ generations. If the initial variance is large (for instance, when the number of loci contributing to the trait is large and/or many loci are of small effect with initial frequency $\frac{1}{2}$), this may be very quick (in the order of 10–100 generations). Second, as the right-hand side of eqn (9) is smaller than one, the shift of the mean phenotype to the optimum is caused by subtle to intermediate changes of the allele frequencies. In contrast, numerical solution of the ODEs (5) (including the long-term behaviour) shows that single-locus selective sweeps are not predicted in this parameter range of the model.

Next, we discuss the response to a sudden optimum shift in a situation when most effects are large. Again, we assume that the population is at equilibrium with no deviation from the optimum z_0 when a sudden shift of the optimum to z_f occurs. In this parameter range, most of the initial allele frequencies are either close to zero or one (as defined above). Thus, the variance and skewness may change appreciably during the selection response, and the constant-variance approximation applied above is not suitable (Jain & Stephan 2015). However, it can be shown that the short-term dynamics may be described using a few loci of large effect. Allele frequencies at these loci approach fixation as in the case of selective sweeps described above. Numerical analysis

shows that the sweeps may occur sequentially as in the case of recurrent sweeps, such that the sweeps occur in a defined order from the largest effect locus to lower-effect loci. Yet, in some cases sweeps occur also simultaneously.

Detecting allele frequency shifts caused by polygenic selection

Loci involved in local adaptation can potentially be identified by an unusual correlation between allele frequencies and relevant ecological variables or by extreme allele frequency differences between geographic regions. However, both approaches are confounded by various factors, in particular when the alleles have relatively weak phenotypic effects on adaptive traits and are therefore under weak selection: (i) such comparisons are complicated by correlations of allele frequencies across populations at neutral loci due to shared population history and gene flow, and (ii) the observed frequency shifts between populations resulting from initial responses to selection on a trait cannot be expected to be extreme if selection is weak and environmental changes are very recent. Some work has been done to overcome these difficulties. For example, concerning the correlation-based methods (point i), Graham Coop and collaborators proposed a Bayesian method that calculates a set of standardized allele frequencies that allows investigators to apply tests to multiple populations while accounting for sampling and covariance due to population history. Using these standardized frequencies, one can construct a test to detect SNPs that deviate strongly from neutral population structure (Coop *et al.* 2010). However, this approach only works if there exist relatively large extended gradients of ecological variables, which may not be the case in rapid adaptation, a process that occurs by definition on a fast timescale. This problem of correlation-based modelling is not alleviated either in the latest method that Berg & Coop (2014) proposed. Shortly after a population occupies a new habitat, we expect that the allele frequency shifts between the parental and derived populations are relatively small. This also means that available software, such as BayeScan-like methods (Foll & Gaggiotti 2008; Riebler *et al.* 2008), is not able to detect significant frequency shifts between populations (point ii).

For detecting small allele frequency shifts after environmental changes in fast adapting populations, it may be better to use the following method. To model the colonization of a new niche after an environmental change, assume that a small fraction k of a large population migrates into a new niche and that the remaining part of the large population continues to exist. This split occurs at time t_s. Following Innan & Kim (2008) who proposed such a model, we refer to the population before t_s as ancestral population and after t_s as parental population. The effective sizes of the ancestral and parental populations are N_A and N_P, respectively, which are assumed to be constant (but this could be changed). In contrast, the derived population, which is established at time t_s, may undergo a population size change. For example, it may increase from effective size kN_A at time t_s to N_D at present (see Fig. 1). Gene flow may be included into this model at a rate m_D into the derived population and at rate m_P in the reverse direction.

Next, we consider a locus that contributes to a quantitative trait and assume that selection on this trait suddenly changed in the derived population (but not in the parental population). This leads to a shift of allele frequencies at this locus from values observed as standing variation in the parental population to some new ones. Our goal is to detect these frequency shifts in DNA polymorphism data, although they may be subtle. To do this, we assume that selection on the trait affects causative variants and also SNPs partially linked to these causative variants. That is, one should include statistics from the entire toolbox of molecular population genetics. We will use the standard statistics, such as nucleotide diversity π and θ, haplotype heterozygosity H and Tajima's (1989) D. We apply them to each population separately. Furthermore, assuming that we have joint samples from the parental and derived populations available, we consider for each of these statistics the difference between the two populations or their ratio.

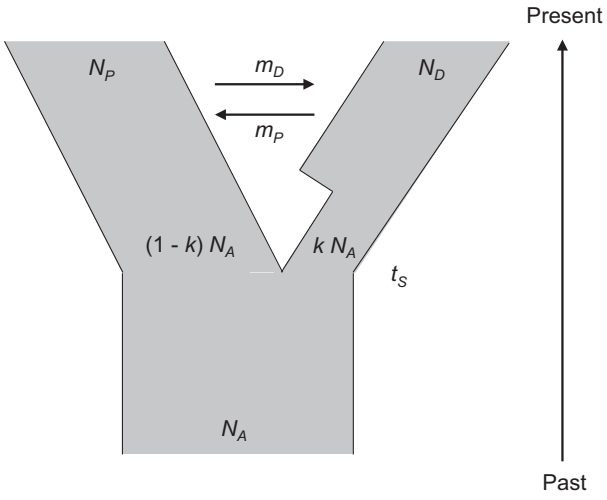

Fig. 1 This figure shows an ancestral population of constant effective size N_A that splits at time t_s into a large parental population of size $(1–k)N_A$ and a small population of size kN_A. The latter (derived) population undergoes a bottleneck and increases subsequently to size N_D, while the parental population stays at a constant size $N_P = (1–k)N_A$. Migration between the parental and derived populations is possible.

Furthermore, following Innan & Kim (2008), we will utilize F_{ST}-based methods, which measure the differentiation between the parental and derived populations. The F_{ST}-based methods were the most powerful approaches in their study. In the absence of migration between these populations, Innan and Kim showed that selection may cause a drastic change in the pattern of polymorphism in the derived population, but not in the parental population. This difference was well picked up by F_{ST}, if the frequency shift was large (such as those leading to soft sweeps).

In addition to these procedures, we could analyse LD between neutral polymorphisms partially linked to a selected site. Including LD into the detection of subtle frequency differences appears to be a promising approach as it may 'amplify' small allele frequency shifts that have occurred after an environment change. This is illustrated in the following example. In our recent work on identifying molecular variants associated with cold tolerance in *Drosophila melanogaster*, we fine-mapped a QTL for chill coma recovery time (a proxy for cold tolerance) and then subjected the resulting genomic region to a population genetic analysis and a study of gene expression variation (Wilches *et al.* 2014). The gene *brinker* was found to be induced by cold stress and to contribute to the observed differences between African and European populations in gene expression. Interestingly, a small group of SNPs located in the regulatory region upstream of *brinker* and associated with *brinker* expression variation shows a significant peak of LD, whereas LD in the rest of the analysed 60-kb genomic region does not deviate from the genomewide average in the European population and LD in the African population was very low in the entire region. In contrast, the shift in the SNP frequencies was moderate (only about 20%). If this effect is general, LD could play a similarly important role in the detection of selective shifts of allele frequencies as in the detection of selective sweeps (Kim & Nielsen 2004; Voight *et al.* 2006; Pavlidis *et al.* 2010).

From a theoretical point of view, the difference to the classical selective sweep case is that not a single haplotype is favoured after the occurrence of a new selected mutation or the influx of a selected migrant allele, but that at the time of an environmental change, a whole group of haplotypes is selected that then is linked to the favoured allele and hence begins rising in frequency. Thus, if the frequency distribution of the haplotypes in the selected group is different from the distribution of the remaining haplotypes, we should observe 'extended' haplotypes as in the case of selective sweeps in the initial period of the selected phase (Stephan *et al.* 2006; Voight *et al.* 2006).

Of course, the 'amplification' effect of LD between the African and European populations described above may also be partly due to demography, in particular due to increased genetic drift during bottlenecks that occurred in the colonization of Europe. In the example motivating this work, the amplification effect is indeed larger than expected. (Expected differences of LD between the parental African and the derived European populations can be calculated assuming large constant population sizes; Stephan *et al.* 2006). Because of the likely impact of demography on LD, it is necessary that the method for detecting allele frequency shifts between populations will be analysed for nonequilibrium scenarios (in particular, population size bottlenecks and expansions).

Open questions

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 *Drosophila melanogaster*, the evidence for sweeps is quite striking (reviewed in Stephan 2010a; for a later reference, see Sattath *et al.* 2011). 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; Poh *et al.* 2014). On the experimental side, however, the search for causative nucleotide changes that led to selective sweeps has only recently started (Saminadin-Peter *et al.* 2012; Voigt *et al.* 2015), although sweep mapping may lead to quite accurate identification of the targets of selection. Clearly, there is much room for future research activities in this latter area.

On the theoretical side, although the theoretical advances in the detection of positive selection in genomes during the past 15 years 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, integrating background selection into the model of selective sweeps has not proceeded beyond the initial efforts that are already 15 years old. Fourth, population subdivision is not well incorporated into the sweep approaches yet (except in F_{ST}-based methods such as BayeScan). Fifth, and perhaps most important, there is an urgent need for more theoretical modelling and more powerful statistical methods to analyse the genomic signatures of polygenic traits for populations that are subdivided and undergo varying population sizes.

Acknowledgements

I thank Charles Aquadro, Joachim Hermisson and an anonymous reviewer for their comments on this manuscript. My current research is supported by grant STE 325/17-1 from the Priority Program 1819 of the Deutsche Forschungsgemeinschaft (DFG).

References

Akey JM, Eberle MA, Rieder MJ *et al.* (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, **2**, e286.

Barton N (1986) The maintenance of polygenic variation through a balance between mutation and stabilizing selection. *Genetical Research*, **47**, 209–216.

Barton NH (1995) Linkage and the limits to natural selection. *Genetics*, **140**, 821–841.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics*, **10**, e1004412.

Berry AJ, Ajioka JW, Kreitman M (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.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783–796.

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

Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics*, **180**, 1645–1660.

Chevin LM, Billiard S, Hospital F (2008) Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. *Genetics*, **180**, 301–316.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Das A, Mohanty S, Stephan W (2004) Inferring population structure and demography of *Drosophila ananassae* from multilocus data. *Genetics*, **168**, 1975–1985.

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

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

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

Innan H, Kim Y (2008) Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*, **179**, 1713–1720.

International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796.

Jain K, Stephan W (2015) Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. *G3-Genes Genomes Genetics*, **5**, 1065–1074.

Jensen JD (2014) On the unfounded enthusiasm for soft selective sweeps. *Nature Communications*, **5**, 5281.

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

Jensen JD, Thornton KR, Bustamante CD, Aquadro CF (2007) On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, **176**, 2371–2379.

Jensen JD, Thornton KR, Andolfatto P (2008) An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics*, **4**, e1000198.

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

Karasov T, Messer PW, Petrov DA (2010) Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genetics*, **6**, e1000924.

Kim Y (2006) Allele frequency distribution under recurrent selective sweeps. *Genetics*, **172**, 1967–1978.

Kim Y, Maruki T (2011) Hitchhiking effect of a beneficial mutation spreading in a subdivided population. *Genetics*, **189**, 213–226.

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

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

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

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

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

Lande R (1983) The response to selection on major and minor mutations affecting a metrical trait. *Heredity*, **50**, 47–65.

Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genetics*, **2**, e166.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nielsen R, Hellmann I, Hubisz M *et al.* (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Orr HA (2005) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, **6**, 119–127.

Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. *Molecular Ecology*, **17**, 3585–3598.

Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, **185**, 907–922.

Pavlidis P, Metzler D, Stephan W (2012) Selective sweeps in multi-locus models of quantitative traits. *Genetics*, **192**, 225–239.

Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.

Poh YP, Domingues VS, Hoekstra HE, Jensen JD (2014) On the prospect of identifying adaptive loci in recently bottlenecked populations. *PLoS ONE*, **9**, e110579.

Pool JE, Corbett-Detig RB, Sugin RP *et al.* (2012) Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genetics*, **8**, e1003080.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

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

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

Sabeti P, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Saminadin-Peter SS, Kemkemer C, Pavlidis P, Parsch J (2012) Selective sweep of a *cis*- regulatory sequence in a non-African population of *Drosophila melanogaster*. *Molecular Biology and Evolution*, **29**, 1167–1174.

Santiago E, Caballero A (2005) Variation after a selective sweep in a subdivided population. *Genetics*, **169**, 475–483.

Sattath S, Elyavish E, Kolodny O *et al.* (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genetics*, **7**, e1001302.

Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. *Genetical Research*, **7**, 155–160.

Stephan W (1995) An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Molecular Biology and Evolution*, **12**, 959–962.

Stephan W (2010a) Detecting strong positive selection in the genome. *Molecular Ecology Resources*, **10**, 863–872.

Stephan W (2010b) Genetic hitchhiking *versus* background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B*, **365**, 1245–1253.

Stephan W, Wiehe THE, Lenz MW (1992) The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology*, **41**, 237–254.

Stephan W, Song YS, Langley CH (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, **172**, 2647–2663.

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

Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, **5**, e171.

de Vladar HP, Barton N (2014) Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics*, **197**, 749–767.

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

Voigt S, Laurent S, Litovchenko M, Stephan W (2015) Positive selection at the *polyhomeotic* locus led to decreased thermosensitivity of gene expression in temperate *Drosophila melanogaster*. *Genetics*, **200**, 591–599.

Whitlock MC (2003) Fixation probability and time in subdivided populations. *Genetics*, **164**, 767–779.

Wiehe THE, Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution*, **10**, 842–854.

Wilches R, Voigt S, Laurent S *et al.* (2014) Fine-mapping and selective sweep analysis of QTL for cold tolerance in *Drosophila melanogaster*. *G3-Genes Genomes Genetics*, **4**, 1635–1645.

Wollstein A, Stephan W (2014) Adaptive fixation in two-locus models of stabilizing selection and genetic drift. *Genetics*, **198**, 685–697.

Wright S (1935) Evolution in populations in approximate equilibrium. *Journal of Genetics*, **30**, 257–266.

Yu F, Etheridge AM (2010) The fixation probability of two competing beneficial mutations. *Theoretical Population Biology*, **78**, 36–45.

Zivkovic D, Steinrücken M, Song YS, Stephan W (2015) Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, **200**, 601–617.

W.S. wrote the paper.
