Molecular Ecology (2008) **17**, 3585–3598 doi: 10.1111/j.1365-294X.2008.03852.x

INVITED REVIEW

A population genomic approach to map recent positive selection in model species

P. PAVLIDIS, S. HUTTER and W. STEPHAN
Section of Evolutionary Biology, Department of Biology, University of Munich, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

Abstract

Based on nearly complete genome sequences from a variety of organisms data on naturally occurring genetic variation on the scale of hundreds of loci to entire genomes have been collected in recent years. In parallel, new statistical tests have been developed to infer evidence of recent positive selection from these data and to localize the target regions of selection in the genome. These methods have now been successfully applied to *Drosophila melanogaster*, humans, mice and a few plant species. In genomic regions of normal recombination rates, the targets of positive selection have been mapped down to the level of individual genes.

Keywords: adaptation, genome scan, genome variation, model species, selective sweep, statistical tests

Received 22 February 2008; revision received 20 May 2008; accepted 28 May 2008

Introduction

Searching for strong positive selection in the genomes of individuals of a natural population has been the focus of a multitude of studies over the past 5 years (Harr *et al*. 2002; Kim & Stephan 2002; Glinka *et al*. 2003; Akey *et al*. 2004; Orengo & Aguadé 2004). The goals of these studies have been (i) to provide evidence of positive selection, (ii) estimate the strength of selection, and (iii) localize the targets of selection. A long-term goal is that the genes that experienced recent, strongly positive selection could be identified and the associated functions and phenotypes characterized.

In general, these studies followed a two-tier approach: at first, levels of DNA polymorphism are measured for a very large number of loci on a *genome-wide* scale within populations. [For humans, the best-studied species, continuous single nucleotide polymorphism (SNP) data are also available along the entire genome, though with some varying density.] The goal of this initial step is to identify loci that display patterns of variability suggesting recent positive selection. Some studies employed microsatellite markers to measure polymorphism and looked for regions of depleted variability as an indicator of a selective sweep due to genetic hitch-hiking in the region (see Box 1). Other studies analysed SNP by directly sequencing small fragments of DNA at multiple loci, which allows for the estimation of properties of the site frequency spectrum (SFS) of SNPs and linkage disequilibrium. While this approach might seem straightforward, the actual definition of a candidate locus can be challenging, especially in populations that have undergone demographic perturbations. Most studies up to now have employed rather simple methods such as outlier analysis, in order to select candidate loci (e.g. Kauer *et al*. 2003; Ometto *et al*. 2005). Only recently more sophisticated methods have been developed for analysing genome-wide polymorphism data, including tests based on the background SFS (Nielsen *et al*. 2005), F_{ST} (Beaumont & Balding 2004; Riebler *et al*. 2008) and explicit modelling of the population history (Li & Stephan 2006).

As a second step following the identification of a candidate locus, polymorphism patterns of the surrounding region are obtained by fine-scale sequencing. The resulting high-density SNP data are then used for tests of deviation from neutral expectations (including the standard tests of Hudson *et al*. 1987; Tajima 1989; Fay & Wu 2000). In addition, however, specific tests for positive selection in these *subgenomic* regions such as the CLR-GOF (Kim & Stephan 2002; Jensen *et al*. 2005) tests are used; they can also estimate the strength of selection and the approximate location of the beneficial mutation within the region. In the following, we

Correspondence: Pavlos Pavlidis, Fax: +49 89 2180 74 104;
Email: pavlidis@zi.biologie.uni-muenchen.de

Box 1 The hitch-hiking effect

When a strongly beneficial mutation occurs and spreads in a population, 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, which they termed *genetic hitch-hiking*. They show that in very large populations hitch-hiking can drastically reduce genetic variation near the site of selection (thus, causing a selective sweep).

According to Maynard Smith and Haigh's deterministic model, in recombining chromosomal regions diversity vanishes at the site of selection immediately after the fixation of the beneficial allele and is predicted to increase as a function of the distance to the selected site (scaled by the selection coefficient). This result is also roughly correct in finite populations (Kaplan *et al*. 1989; Stephan *et al*. 1992). Further signatures of the hitch-hiking effect include (i) shifts in the site frequency spectrum of polymorphisms such as an excess of low- and high-frequency derived alleles (Braverman *et al*. 1995; Fay & Wu 2000), and (ii) distinct patterns of linkage disequilibrium such as an elevated level of linkage disequilibrium in the early phase of the fixation process of a beneficial mutation (Kim & Nielsen 2004; Stephan *et al*. 2006). In a suite of statistical tests, these properties of the hitch-hiking effect have been used to map recent, strongly positive directional selection along recombining chromosomes of several species.

describe these new tests and show that they have been successfully used to identify the targets of recent, strongly positive selection. If the rate of local recombination is not too low, individual genes or even regions within a gene can be mapped with this approach.

Methods for detecting selective sweeps

Subgenomic data

Composite-likelihood ratio test. Using predictions of the hitch-hiking model (Maynard Smith & Haigh 1974; see Box 1), Kim & Stephan (2002) developed a composite-likelihood ratio (CLR) test to detect local reductions of nucleotide variation along a recombining chromosome and to predict the strength and the location of a selective sweep. The CLR test compares the probability of the observed polymorphism data under the standard neutral model with the probability of the data under a model of selective sweep. Under the standard neutral model the expected number of sites where the derived variant is in the frequency interval $(p, p + dp)$ in the population (the SFS) is given by

$$\phi_0(p)dp = \frac{\theta}{p}dp \qquad \text{(eqn 1)}$$

(Fu 1995; Ewens 2004). Fay & Wu (2000) have shown that immediately after a hitch-hiking event this distribution is transformed approximately to

$$\phi_1(p) = \begin{cases} \frac{\theta}{p} - \frac{\theta}{C} & \text{for } 0 < p < C \\ 0 & \text{for } C \le p \le 1 - C, \\ \frac{\theta}{C} & \text{for } 1 - C < p < 1 \end{cases} \qquad \text{(eqn 2)}$$

where the parameter C depends on the strength of selection $\alpha = 2Ns$ and the recombination rate r between the neutral site and the site where the beneficial mutation has occurred (Kim & Stephan 2002). N is the effective population size and s the selection coefficient.

The probability of observing a site where k derived alleles are found in a sample of n sequences is obtained by binomial sampling as

$$P_{n,k} = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k}\, \phi(p)dp, \qquad \text{(eqn 3)}$$

where $\phi(p) = \phi_0(p)$ applies under the standard neutral model and $\phi(p) = \phi_A(p)$ under the hitch-hiking model. Kim & Stephan (2002) compare the two hypotheses:

- (H_0) The observed allelic class at each position of the subgenomic region under consideration is derived from a standard neutral model.
- (H_A) The observed allelic class at each position of the subgenomic region is due to a selective sweep that occurred at some position X of the fragment and is characterized by the selection parameter α.

The probabilities of the data under these hypotheses are calculated as the product of the probabilities of all sites of the fragment under consideration (eqn 3) using the densities $\phi_0(p)$ and $\phi_A(p)$, respectively. The maximum log-likelihood-ratio statistic Λ_{CLR} is then given by

$$\Lambda_{CLR} = \log \frac{\max P(\text{Data} \mid H_A)}{P(\text{Data} \mid H_o)} \qquad \text{(eqn 4)}$$

where max refers to the maximization of $P(\text{Data} \mid H_A)$ with respect to the parameters X and α.

Since the null and alternative hypotheses that are compared in the CLR test are explicitly modelled, the interpretation of the test results is rather simple. That means that the expectation of the SFS is well formulated under both evolutionary scenarios. On the other hand, it is important to realize that the null hypothesis of the test is based on the standard neutral model. That means that any violation of the assumptions of the null hypothesis may influence the results and favour the alternative hypothesis (Jensen *et al*. 2005; Thornton & Jensen 2007). Therefore, the application of the CLR test is not appropriate for detecting selective events when severe demographic events (especially bottlenecks) have occurred in the recent history of the population. In such cases a combination of the CLR test and the following approach may be used.

Distinguishing between selective sweeps and demography. Jensen *et al*. (2005) showed that the CLR test is not robust in the cases of structured populations or recent bottlenecks. Under these scenarios, the false-positive rate may be as high as 80% (Jensen *et al*. 2005). They proposed a goodness-of-fit (GOF) test to distinguish between the true positives that come from the rejection of the standard neutral scenario because of a selective sweep event, and the false positives that come from the rejection of the standard neutral hypothesis due to demographic factors.

The GOF test is based on the hypothesis that nonselective evolutionary processes influence the frequency spectrum globally (e.g. the whole region under investigation) and not locally as a selective sweep does. This assumption is adopted widely, although it is possible that a recent strong bottleneck combined with recombination may create local patterns that resemble those of a selective sweep (Barton 1998; Thornton & Jensen 2007).

The GOF approach tests whether the observed data are drawn from a selective sweep model. Thus, the latter represents the null hypothesis H_0. The alternative hypothesis H_A claims that the data are not drawn from a selective sweep scenario. Thus, for H_A an alternative model is not specified, except that it is assumed that the evolutionary forces in action affect the whole region under investigation. The likelihood of the alternative model is calculated as

$$P(\text{Data}\,|\,H_A) = \prod_{i=1}^{\ell} P(Y = y_i\,|\,H_A) = \prod_{i=1}^{\ell} \binom{n}{y_i} P_i^{y_i}(1-p_i)^{n-y_i} \qquad \text{(eqn 5)}$$

The composite-maximum-likelihood estimates of p_i are given by the empirical frequencies $p_i = (k/n)$, where y_i is the number of sequences that carry the derived allele at site i, and ℓ is the length of the region under study. The proposed GOF statistic is then formulated as

$$\Lambda_{GOF} = \log \frac{\max P(\text{Data}\,|\,H_A)}{P(\text{Data}\,|\,H_o)} \qquad \text{(eqn 6)}$$

For the null hypothesis H_0, the maximization refers to the X and α parameters. For the alternative hypothesis H_A, the maximization is related to the estimates of p_i as mentioned above. Λ_{GOF} values cannot directly reveal the favourable model even if, intuitively, small Λ_{GOF} values support the selective sweep hypothesis. More importantly, it is difficult to predict the effect of various evolutionary forces on the value of Λ_{GOF}. This is because the alternative hypothesis lacks the function $\phi(\cdot)$, which is specific to the evolutionary model. Simulations under a selective sweep scenario are employed in order to obtain the null distribution of the Λ_{GOF} statistic. The parameters X and α are estimated using the CLR approach of Kim & Stephan (2002). If the P value for Λ_{GOF} is smaller than a cut-off value P_c, then H_0 is rejected, otherwise it is accepted. Jensen *et al*. (2005) suggest a cut-off value of 0.15.

Simulating neutral data under various bottleneck scenarios allows for the estimation of the false-positive rate of the GOF approach under various values of P_c. We may use $P_{0.05}$ such that the false-positive rate of the GOF test would be 0.05. Simulations, however, show that both the false-positive rate (using a certain cut-off P_c) and the value $P_{0.05}$ depend on the demographic scenario. Thus, results obtained using the GOF test should be interpreted carefully, when there is evidence that the population has experienced recent demographic changes, especially bottlenecks. It should be noted that there is not a single $P_{0.05}$ value appropriate for all the demographic changes. This is illustrated in Fig. 1 using two different bottleneck scenarios.

The combined CLR and GOF tests are used extensively in subgenomic scans for the detection of selective sweeps (see below). Subgenomic data sets are usually obtained by resequencing short fragments of DNA segments. Subsequently, a particular 'interesting' region that shows evidence for a selective sweep may be selected for fine-scale sequencing and parameters like the position of the sweep or the strength of selection are estimated from the data. However, the preselection of interesting regions creates an ascertainment scheme that has been shown to result in high false-positive rates (Thornton & Jensen 2007). Both the CLR and GOF tests are not robust to this combination of ascertainment and demography. Thornton & Jensen (2007) propose to control the false-positive rate by using the null distribution of the Λ_{CLR} statistic that is both generated from the correct demographic model and conditional on the ascertainment scheme. This strategy can be applied when the demographic model is known or can be estimated from the data.

The source codes of the CLR and the GOF tests as well as their documentation are freely available online or may be requested from Yuseob Kim and Jeffrey Jensen, respectively.

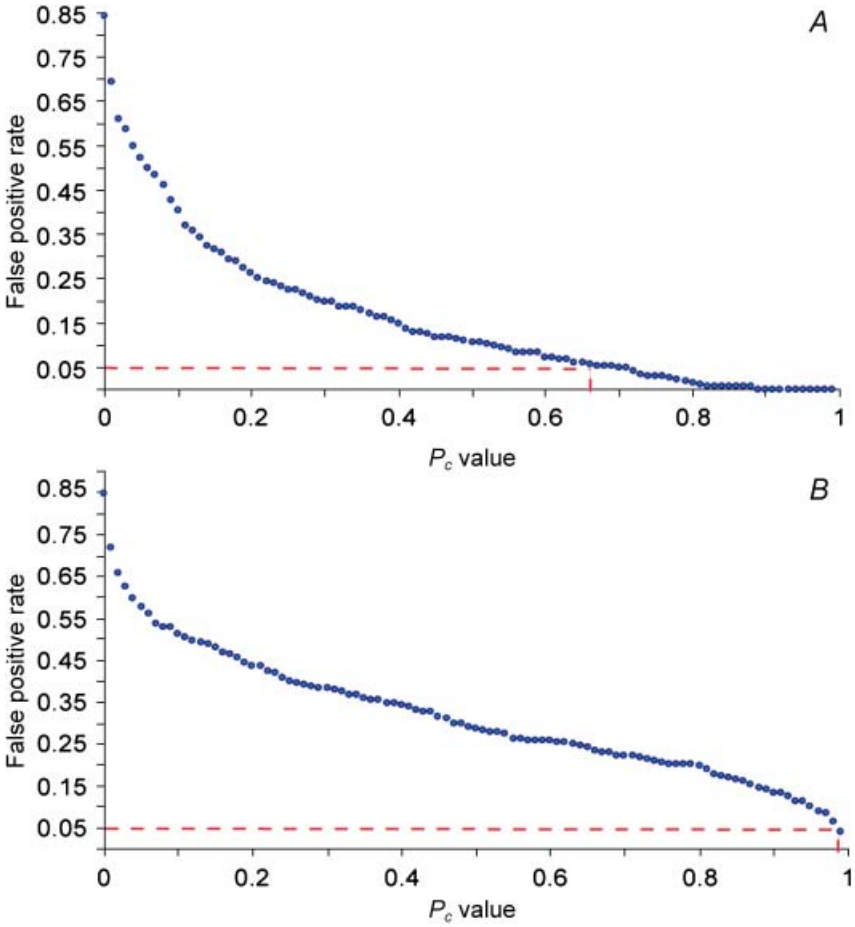

Fig. 1 The false-positive rate as a function of the P_c value for two different bottleneck scenarios. (a) The bottleneck scenario is characterized by the following parameters: sample size $n = 15$, population mutation rate $\theta = 75$, population recombination rate $4Nr = 1000$, population growth rate $G = 184.20$. This corresponds to the ratio $(N_0/N_{tb}) = 100$, when the bottleneck occurred at $t_b = 0.1\ N$ generations in the past (i.e. backwards in time, the population experienced an exponential contraction to 1% of its present size). Subsequently (backwards in time), it increased instantaneously to the present size as it is described in Jensen *et al*. (2005). The dashed line denotes that $P_{0.05} \approx 0.67$. (b) This more complicated bottleneck scenario is characterized by a population that contracted according to the ratio $(N_0/N_{tb}) \approx 500$ at $t_b = 0.140\ N$ generations in the past in a stepwise fashion. The bottleneck ended at $t_1 = 0.148\ N$ generations in the past. Then the population increased instantaneously according to the ratio $(N_0/N_{t1}) \approx 0.124$ and eventually it decreased again to $(N_0/N_{t2}) \approx 0.625$ at the time $t_2 = 0.556$. This demographic scenario was inferred in Li & Stephan (2006) for the European *Drosophila melanogaster* population. The dashed line denotes that $P_{0.05} \approx 0.98$.

Genome-wide data

SweepFinder. 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 method for the detection of selective sweeps, which would allow for an analysis of genome-wide data. Full genomic scans, however, face several challenges. First, the confounding effects of demography obscure the detection of selective events in similar ways as in subgenomic scans. Second, data usually consist of SNPs that were initially identified in an ascertainment process, which may be quite complicated in some cases and can generate biases that should be taken into account.

The test *SweepFinder* proposed by Nielsen *et al*. (2005) is a CLR test that is based on the ideas of the CLR approach of Kim & Stephan (2002). However, it differs from the latter one in that the null hypothesis is not derived from a specific evolutionary model, but estimated from the empirical background distribution of the data. The idea behind the use of the background distribution is similar to the arguments presented in Jensen *et al*. (2005) for formulating the alternative hypothesis. That means that the nonselective evolutionary processes that shape SFS affect the whole genome. Additionally, the method relies on the assumption that a class of neutral DNA exists in the genome.

SweepFinder is also based on the principles of the hitchhiking theory. That is, when a beneficial mutation occurs on a chromosome and goes to fixation, variation at linked neutral loci is reduced as the beneficial mutation spreads through the population. A selective sweep is modelled by assuming that each ancestral lineage escapes the sweep with a probability p_e, which is given as a function of the recombination distance from the selected site and $a = (r/s)\ln(2N)$.. Given that some lineages have escaped the selective sweep by recombination, the method calculates the probability to observe a mutant allele of frequency B. In order to calculate this quantity, the method estimates the number of ancestral lineages that carry the neutral mutation after the end of the selective phase and assumes that the SFS after the selective sweep is the same as at present (i.e. at the time of sampling).

Similarly to Kim & Stephan (2002), *SweepFinder* uses a CLR approach to choose between a neutral and a selective model. The alternative hypothesis H_A states that a beneficial

mutation has occurred at some position X. The likelihood of H_A is calculated as the product of the site probabilities (p_B^*) for all the sites and maximized with regard to the parameters X and α. When only polymorphic sites are included in the data set the method is properly standardized. The null hypothesis is formulated as the probability to observe the data given the empirical frequency spectrum. That means that if the probability of a specific allelic class is f_i, $i = 1, \ldots, n-1$, in the case of an unfolded spectrum and the allelic class at position j is given by ξ_j, then the likelihood of H_0 is equal to

$$L_{H_o} = \prod_{j=1}^{\ell} f_{\xi j}. \quad \text{(eqn 7)}$$

Obviously, L_{H_O} depends only on the empirical frequency spectrum.

Similarly to Kim & Stephan (2002), the CLR statistic Λ_{SF} is given by

$$\Lambda_{SF} = \log \frac{\max P(\text{Data} \mid H_A)}{P(\text{Data} \mid H_o)} \quad \text{(eqn 8)}$$

The null distribution of the statistic Λ_{SF} is obtained by using the specific demographic scenario that might have shaped the observed data. Even if the method is robust against several demographic scenarios that have been investigated in Nielsen *et al*. (2005), our simulations have shown that this does not hold in general, especially in cases of severe and recent bottlenecks. Additionally, it is unknown how the method behaves in cases that SNP data is retrieved from the ascertainment schemes described in Thornton & Jensen (2007). Thus, these factors should be included when the null distribution of the statistic Λ_{SF} is constructed and from this a threshold value is calculated. The method is robust against multiple testing.

SweepFinder may also be applied to subgenomic data. In this case, the program offers the flexibility to employ a user-specified frequency spectrum instead of calculating it from the data. This may be useful when the genomic region under study is not representative of the whole genome.

The source code and the documentation of *SweepFinder* are available from Rasmus Nielsen's webpage (http://www.binf.ku.dk/~rasmus/webpage/sf.html). The program is written in C and tested successfully on 32-bit and 64-bit machines. The simulations for the calculation of the threshold of Λ_{SF} may also be done on computer clusters.

Joint inference of demography and positive selection. While the CLR and GOF approaches do not use explicit demographic models, Li & Stephan (2006) describe a statistical method to detect footprints of selection in chromosome- or genome-wide data (multiple loci), while taking fluctuations of the population size into account (Li & Stephan 2006). They analyse X chromosomal SNP data from a Zimbabwe and a European *Drosophila* population. Initially, they infer the demographic scenario of the African population (from the ancestral range). This is characterized by a stepwise expansion such that population size changed instantaneously some generations ago. The European population is derived from the African population thereby undergoing a recent severe bottleneck. The parameters of this model are estimated by applying maximum-likelihood techniques based on the SFS for the African population and the joint SFS for the European population. In the analysis it is assumed that there is no recombination within loci (which are only about 500 bp long in this data set), but the loci are partially linked.

Performing simulations for the whole X chromosome, and considering that the SFSs between the loci are independent given their genealogy, they inferred the parameters of the demographic scenario by maximum likelihood.

Li & Stephan (2006) avoid the problem of inefficient sampling of genealogies by calculating the likelihood as a function of the expected branch lengths that may produce the observed pattern

$$L_k = P(\text{SFS} \mid \hat{G}_k) = \prod_{i=1}^{n_k - 1} P(\zeta_{ik} \mid E(l_{ik})). \quad \text{(eqn 9)}$$

However, this is just an approximation and its accuracy has still to be demonstrated.

After estimating the demography, Li & Stephan (2006) perform a sliding window analysis to find genomic regions which are affected by the action of strong positive selection. They conduct a likelihood-ratio test that employs two hypotheses. The null hypothesis considers that the population has experienced the inferred demographic scenario, and the alternative one assumes that the forces that shape the data consist of the inferred demographic scenario together with a selective sweep. In order to overcome the problem of inefficient sampling of genealogies for the loci that belong to the sliding window, they consider a compact frequency spectrum. In this approach, all high frequency variants are pooled together and, hence, the number of inconsistent coalescent trees is diminished (Li & Stephan 2005).

It is encouraging and promising that methods that incorporate demographic events explicitly in the inference of selection are being developed. Even if the CLR-GOF and the *SweepFinder* approaches do that only indirectly, demographic models can be incorporated in the estimation of the null distributions of the relevant statistics. Simulations have shown that this strategy can control the false-positive rate (Thornton & Jensen 2007; P. Pavlidis, unpublished results).

Li & Stephan (2006) implemented a software package called MOSY (http://www.zi.biologie.uni-muenchen.de/~li/mosy/) to detect recent selective sweeps and estimate parameters in populations of varying size.

Methods for detecting selection based on genetic differentiation between populations

F_{ST}-based methods. Bayesian approaches have been shown to be powerful for quantifying differentiation between populations (Balding & Nichols 1995) and for the estimation of demographic events (Beaumont 2003). More recently, Bayesian methods have also been applied to whole-genome data (multiple loci) in order to reveal genetic regions that have experienced selective sweeps (Beaumont & Balding 2004; Riebler *et al*. 2008). These methods combine information from multiple populations. Thus, they are able to use data produced from recent genotyping (e.g. International HapMap Consortium 2005) and sequencing projects (e.g. Glinka *et al*. 2003). Additionally, they can infer both positive and balancing selection. Since Riebler *et al*. (2008) extended the method introduced by Beaumont & Balding (2004), we discuss here the Riebler *et al*. approach.

This approach infers selective events using the F_{ST} value of a population for a given locus in a hierarchical two-level Bayesian framework. F_{ST} represents the probability that two randomly chosen alleles from the locus in the same subpopulation are identical because of common ancestry within the subpopulation (Crow & Kimura 1970). In a coalescent framework, F_{ST} may be seen as the probability that a coalescent event precedes a migration event (going backwards in time) (Hudson 1998). That means that F_{ST} values may be used for inferring balancing or positive selection since positive selection may increase the F_{ST} value and balancing selection decreases it.

In the two-level model of Riebler *et al*. (2008), the lower level expresses the likelihood for the allele-frequency counts as a function of F_{ST} using a multinomial Dirichlet model (Beaumont 2003; Beaumont & Balding 2004). The higher level describes the logistic regression of locus-specific, population-specific and locus-by-population-specific effects on F_{ST}. The advancement of the Riebler *et al*. (2008) approach consists in a reparameterization of the original framework of Beaumont & Balding (2004) and the subsequent use of an auxiliary Bayesian variable that indicates whether a locus is neutral or not.

It should be noted, however, that both the Beaumont & Balding (2004) and the Riebler *et al*. (2008) approaches are based on haplotype statistics since distinct haplotypes (e.g. sequences) are treated as alleles. As a consequence, the calculation of genetic differentiation based on F_{ST} loses information when many haplotypes in the sample are unique. This may be the case when the sample size is small, the mutation rate high and/or the sequence of a locus long. The source code, C executable files and R programs, is available from Andrea Riebler (andrea.riebler@ifspm.uzh.ch).

Haplotype-based methods. All methods discussed thus far are designed to detect complete sweeps within a panmictic population or, in the case of population structure, within a subpopulation. To discover incomplete sweeps (i.e. sweeps that are ongoing within a subpopulation or sweeps that are complete within one subpopulation, but not with regard to the total population), haplotype-based methods have been developed. These methods analyse the length of haplotypes around a given locus of interest, which is thought to be the target of selection.

If a selective sweep is ongoing in a subpopulation, the hitch-hiking haplotype is expected to be rather long (see Box 1). This feature of the hitch-hiking effect has been exploited by Sabeti *et al*. (2002) who developed the so-called long-range haplotype (LRH). A slight modification of this is the *iHS* statistic (Voight *et al*. 2006). A disadvantage of these approaches is that they lose power when the beneficial allele is close to fixation. To overcome this problem, Tang *et al*. (2007) and Sabeti *et al*. (2007) apply the ideas of the haplotype-based tests not to a single (local) subpopulation but contrast the haplotype profiles between subpopulations.

Until now, little is known about the power and robustness of haplotype-based methods. Additional research is needed to investigate the false-positive rate of these methods under various demographic scenarios or migration models when more than one subpopulation is involved.

Footprints of recent positive selection inferred from genome-wide and subgenomic data

In this section, we review the studies in which scans of genome-wide or subgenomic DNA variability were performed. In some of them, the statistical tests discussed earlier were applied to find regions for recent positive selection. Here, we focus on the model organism *Drosophila melanogaster*, but also discuss studies carried out in other species.

Drosophila melanogaster

The fruit fly *D. melanogaster* was one of the first multicellular species to have its genome fully sequenced (Adams *et al*. 2000). As the availability of an annotated genome greatly facilitates scans for positive selection, it has since been the focus of many such studies. *D. melanogaster* was originally a tropical species that originated in sub-Saharan Africa. After the last glaciation it moved to the more temperate zones of Eurasia approximately 15 000 years ago (David & Capy 1988; Lachaise *et al*. 1988). As a human commensal, *D. melanogaster* is nowadays found all over the world. This provides the interesting opportunity to compare ancestral African populations to derived

non-African (also called cosmopolitan) populations that are supposed to have undergone adaptations to their new environments.

Harr *et al*. (2002) typed microsatellite variability in a genomic region of 274 kb on the X chromosome in two African and eight non-African populations of *D. melanogaster*. The surveyed chromosomal segment was chosen because of a priori evidence that positive selection might have acted in European populations (Schlötterer *et al*. 1997). Additionally, the study included the scan of a 578-kb large putatively neutrally evolving autosomal region, which served as a control. The authors could isolate two distinct regions on the X that show patterns of microsatellite variability indicative of the action of positive selection in the derived populations.

A slightly different approach using microsatellites was conducted by Kauer *et al*. (2003). These authors typed 205 loci that were randomly distributed on chromosome 3 and the X. Variability was measured in two African and several European populations of *D. melanogaster* and candidate loci residing in regions under positive selection were again defined as those, which showed significantly reduced diversity in the derived populations compared to the ancestral ones. Based on this criterion, the authors identified 33 subgenomic regions that are supposedly subject to positive selection. A drawback of the two aforementioned studies is the ignorance of demography. It is nowadays well established that non-African populations of *D. melanogaster* have undergone population bottleneck(s) during their range expansion (e.g. Andolfatto 2001; Glinka *et al*. 2003; Baudry *et al*. 2004). Since such events reduce genetic variability much like selective sweeps, many of the obtained results could represent false positives. In fact, the application of the CLR-GOF tests to the two sweep regions of Harr *et al*. (2002) shows that positive selection is not needed to explain the patterns of polymorphism observed in the derived populations (Thornton *et al*. 2007).

The first genome scan using SNP markers was performed by Glinka *et al*. (2003). Polymorphism was analysed in an African and a European population by sequencing short fragments of noncoding DNA located in introns or intergenic regions. The 105 fragments were approximately 500 bp in length and distributed across the X chromosome in regions of intermediate recombination rates. This data set was later expanded to over 250 loci resulting in an average distance between loci of under 50 kb for the larger part of the chromosome (Ometto *et al*. 2005). This level of resolution was chosen, because theoretical studies suggested that signatures of a sweep should extend to approximately this length given the recombination rates in *D. melanogaster* (Kim & Stephan 2002). Ometto *et al.* tried to identify candidate loci for the European population by estimating bottleneck parameters and defining those loci as candidates where reduction in polymorphism could not be explained by the bottleneck alone. The resulting list of subgenomic regions was the starting point for fine-scale sequencing studies that provided further evidence for positive selection in the regions surrounding the selected loci by application of the CLR-GOF approach.

One of these regions encompassed the gene *polyhomeotic-proximal* (*ph-p*). This region showed a signature of positive selection in the African population (Beisswanger & Stephan 2008). The sweep at *ph-p* was localized in the large intron or the proximal 5′ flanking region of this gene (within an interval of < 3 kb) and is relatively old dating to about 50 000 years ago, i.e. to a time before *D. melanogaster* migrated out of Africa. The huge valley of reduced variation that was found in the European population (Beisswanger *et al*. 2006) is a mere consequence of the sweep identified in Africa. The gene *ph-p* and its duplicate *ph-d* code for proteins of the Polycomb group and are thus involved in transcriptional repression of hundreds of genes and perhaps whole gene networks. Although the *ph* duplication is at least 25–30 million years old, the duplicates are nearly identical over large parts of the gene, suggesting that frequent gene conversion (concerted evolution) occurred during evolutionary time. Despite this homogenizing force, the functions of the *ph* genes have begun to diverge: there is no clear evidence that the distal and proximal products bind to and modify chromatin in different ways; however, both transcriptional units are differentially regulated at the mRNA level. The observed sweep was mapped to a narrow region of *ph-p* containing several regulatory elements that are absent in *ph-d*. This indicates that strong positive selection has been driving the functional divergence of these gene duplicates. To our knowledge, this is the first case of neofunctionalization driven by positive selection in the presence of gene conversion.

Another successful application of the CLR-GOF tests was reported by Glinka *et al*. (2006). Using the same approach as Beisswanger *et al*. (2006), they found a huge valley of reduced variation in the European population (of about 80 kb) and a much narrower one in Africa, encompassing the *brinker* gene (*brk*). This gene is a transcriptional repressor in the decapentaplegic signalling pathway (regulating epidermal cell fates). In the African sample, the *P* values confidently fall in a range consistent with a selective sweep and inconsistent with the demographic models examined by Jensen *et al*. (2005). The case of the European population, however, is less clear with regard to the GOF test.

Recently, SNP variability was also analysed at a large number of loci on chromosome 3 for the aforementioned two populations (Hutter *et al*. 2007) to estimate X-autosome diversity and the sex ratio. Tests are currently applied to this new data (Nielsen *et al*. 2005; Li & Stephan 2006) in order to localize candidate regions of selection.

A further study investigating patterns of noncoding SNPs was published by Orengo & Aguadé (2004). The

authors concentrated on the X chromosome and analysed 109 short noncoding DNA fragments with an average distance of ~200 kb in a Spanish population. This work also spawned a follow-up study that detected a region presumably under selection using fine-scale sequencing (Orengo & Aguadé 2007). Here, the authors resequenced a region of 20 kb around the gene *phantom* (*phm*) to localize the target of selection. Application of the CLR-GOF tests to the completely sequenced region placed the position of the target of selection within the transcriptional unit of *phm.* This gene codes for CYP306A1, a cytochrome P450 enzyme in the ecdysteroidogenic pathway.

Using a similar experimental design as Harr *et al*. (2002), Bauer DuMont & Aquadro (2005) scanned a 60-kb large X chromosomal region using microsatellite markers surrounding the *Notch* locus for an ancestral and three derived populations. Here also, previous studies showed evidence for positive selection in the region (Bauer DuMont *et al*. 2004). Two more recent surveys have expanded the microsatellite data set beyond the initial study. Pool *et al*. (2006) investigated a 330-kb large chromosomal stretch just upstream of *Notch*, while Jensen *et al*. (2007) scanned microsatellite variability in a 260-kb region located immediately downstream of the initial study. The latter two studies have been done without prior knowledge about selection in the corresponding genomic regions. Interestingly, signals indicative of positive selection were found in all three microsatellite data sets.

Resequencing approximately 14 kb around the *Notch* locus, Bauer DuMont & Aquadro (2005) found evidence for a recent selective sweep downstream of *Notch* within or between the open reading frames of *CG18508* and *Fcp3C* (*Follicle cell protein 3C*) in non-African populations (from the USA and China). However, while the CLR test produced a highly significant result, the *P* value of the GOF test was relatively low (0.114). The ancestral African population (Zimbabwe) did not show a signature of a sweep. *Fcp3C* codes for a protein involved in the formation of follicle cuticles. These proteins often evolve rapidly due to positive selection (Swanson & Vacquier 2002).

Pool *et al*. (2006) localized a selective sweep to a 361-bp window within the 5′ regulatory region of the *roughest* gene (*rst*) in a Zimbabwe population based on the CLR-GOF methods. Estimation of the age of the sweep suggested that the selected fixation occurred prior to the migration of *D. melanogaster* out of Africa. As for *ph-p*, the sweep signal detected in the non-African populations is thus only a consequence of the sweep in Africa. *rst* codes for a membrane-spanning protein involved in developmental processes (e.g. of the eye or muscles).

Finally, Jensen *et al*. (2007) generated SNP data in a 25-kb region encompassing the gene *diminutive* (*dm*) for populations from China and Zimbabwe. They detected strong evidence for a sweep in the African and non-African populations within or near *dm*. These results are consistent with the known role of *dm* as a positive regulator of body size (Craymer & Roy 1980) and the observed clinal pattern of variation of this trait (Gockel *et al*. 2002; Calboli *et al*. 2003).

Men and mice

Naturally, finding genes under positive selection in the human genome is of great interest to the scientific community. The Duffy blood group locus (*FY*) in humans has been analysed by the CLR-GOF tests (Jensen *et al*. 2005) based on the sequencing data of Hamblin *et al*. (2002). Sequencing was done in short segments (totalling 12 kb on average per line) within a region of about 34 kb around *FY*. The Hausa population from sub-Saharan Africa showed a clear signature of a selective sweep at this gene, while for the samples from China, Italy and Pakistan neutrality could not be rejected. The likely target of positive selection is the *FY*O* mutation, a single noncoding base change in the 5′ flanking region of *FY* that eliminates transcription of the *FY* mRNA in erythroid cells. This allele that confers resistance to malaria due to *Plasmodium vivax* is fixed in sub-Saharan Africa but essentially absent elsewhere.

Another human locus analysed by resequencing and application of the CLR-GOF tests is the *AIM1* gene (Soejima *et al*. 2006). This gene encodes a melanocyte differentiation antigen. Reduced variation at this 40-kb large gene (in particular, in a completely sequenced region of 7.5 kb around the missense polymorphism L374F) was found in individuals of European descent. The results suggest that positive selection has recently acted on this part of the melanogenic gene and that an advantageous haplotype spread rapidly in Europe.

Both *FY* and *AIM1* were not detected by genome scans but were identified a priori as candidate genes on the basis of functional information and some population genetic surveys. For example, the characteristic geographical distribution of the three major alleles *FY*A*, *FY*B* and *FY*O* for the Duffy locus was previously known. Similarly, striking patterns of genetic variability were discovered at two other loci that were suspected to be under selection, the lactase (*LCT*) locus and the glucose-6-phosphate dehydrogenase gene (*G6PD*). *LCT* shows a signature of an incomplete (ongoing) sweep in Europeans as a result of positive selection during the past 10 000 years after the emergence of dairy farming (Bersaglieri *et al*. 2004). A similar pattern of variation was found at *G6PD* (Tishkoff *et al*. 2001). Here deficiency alleles conferring resistance to malaria show a signature of positive selection.

Numerous studies have also undertaken genome scans to find signals of recent positive selection in the human genome. Akey *et al*. (2004) resequenced a total of 132

protein-coding genes in 24 Americans of African ancestry and 23 American of European decent. The final data set consisted of a total of 12 890 SNPs. They then used statistics based on the SFS to find genes deviating from neutral expectations. As it was clear that both populations had undergone population size changes in the past, the authors estimated the most likely demographic history using the genome-wide data. This was then used as a null model to assess the statistical significance for each gene separately. After correction for multiple testing, eight genes showed patterns of polymorphism that could not be explained by demography alone in the European-American sample and were therefore candidates for positive selection. The strongest signal was found in a region spanning 115 kb on chromosome 7 that contained four of these genes. Among them are the genes *TRPV5* and *TRPV6* which have been shown to play an important role in the absorption efficiency of calcium (Nijenhuis *et al*. 2003; van de Graaf *et al*. 2003). The authors suggest that these genes were under selection in Europeans in order to fully profit from a milk-rich diet similar to *LCT*.

What currently sets apart humans from all other species is the availability of large-scale SNP databases such as SeattleSNP (http://pga.gs.washington.edu), Perlegen (Hinds *et al*. 2005) or HapMap (International HapMap Consortium 2003). These data sets include a vast number of SNPs typed in many individuals coming from different populations. The second generation of the HapMap database, for example, contains 3.1 million SNPs for a total of 270 individuals from four populations (International HapMap Consortium 2007). This provides an exceptional opportunity for the application of genome-scan type of analyses. In fact, there are many studies that have already made use of the available data employing a multitude of different statistical methods.

Due to the high density and large size of such SNP data sets, numerous different candidate genes for adaptation have been detected. In a systematic screen based on HapMap data, Voight *et al*. (2006) found evidence for incomplete sweeps using the *iHS* statistic (including the 17q21 inversion in Europeans, many cytochrome P450 genes in all populations analysed, skin pigmentation genes in non-African populations and the olfactory-receptor genes on chromosome 11 in Africans). The cytochrome P450 genes are important in the detoxification of xenobiotic compounds. The observation of positive selection at the skin pigmentation genes indicates that alleles that confer lighter skin colour have a selective advantage in geographical regions that are further away from the tropics. However, there is a *caveat* to gene mapping in humans using genome scans without prior knowledge on the targets of selection. Since variation in humans is low and demography rather complex, the gene regions identified by these haplotype tests may be very large (up to hundreds of kilobases) and contain several genes (up to 10 or more). The genes discussed above may therefore not be accurately identified and further scrutinizing of the genomic regions is needed. Sabeti *et al*. (2007) have recently proposed a heuristic method to fine-map the target of selection that may be promising for future studies.

Using *SweepFinder* (Nielsen *et al*. 2005), Williamson *et al*. (2007) searched the Perlegen SNP data (Hinds *et al*. 2005) for complete sweeps. They found 101 genomic regions that experienced a recent complete selective sweep (such that the target of the sweep is within 100 kb of a known gene). These regions encompass similar categories of genes that were also found by Voight *et al.* including those for olfactory receptors, pigmentation and immunity (not, however, *LCT*, which served as a negative control). As in Voight *et al*. (2006) there was wide variation among populations (Americans of African, Asian, and European descent) in the predicted target regions of selection, and the inferred target regions were relatively large in comparison to African *D. melanogaster* populations.

Finally, using evidence based on population structure, Akey *et al*. (2002) identified genomic regions of increased F_{ST} in a study of 26 530 SNPs from African-Americans, European-Americans and East Asians. One of the regions contains the *CFTR* gene, associated with cystic fibrosis. Similarly, Weir *et al*. (2005) analysed the Perlegen data and found several regions with significantly increased population differentiation, including the *LCT* locus. Further examples are reported in the main publication of the International HapMap Consortium (2005).

The house mouse, *Mus musculus*, has also been the subject of genome scans for positive selection (Ihle *et al*. 2006). In this study the authors genotyped roughly 200 microsatellite loci distributed across the genome. Variability for these loci was assessed in three populations of *M. m. domesticus* along with a population of the sister subspecies *M. m. musculus*. Around 60 individuals were typed for each of these populations. In analogy to the microsatellite studies in *D. melanogaster*, candidates for positive selection were defined as those loci that showed reduced variability in pairwise comparisons among the four investigated populations. A locus with signs of positive selection in two of the surveyed *M. m. domesticus* populations was then further studied at the SNP level, but unfortunately the results were inconclusive.

In mice the CLR-GOF tests have been successfully applied to the genomic region encompassing *MKK7*, a gene coding for mitogen-activated protein-kinase-kinase (Harr *et al*. 2006). This region was identified by a survey of gene expression variation among subspecies of the *M. musculus* complex using microarrays (and qRT-PCR). In *M. m. domesticus* testes *MKK7* is significantly up-regulated. Sequencing of several short fragments within a 47.5-kb region around this gene showed a valley of reduced variation

of about 20 kb, which contained *MKK7* and also two downstream genes. In *M. m. musculus*, this pattern was not observed. Application of the *CLR* and *GOF* tests (producing *P* values of 0.0025 and 0.24, respectively) suggests that this reduction of variation is likely due to positive selection. *MKK7* is known to be involved in modulating a kinase signaling cascade in a stress response pathway, which in *Caenorhabditis elegans* was shown to play a role in innate immunity.

Plants

Plants exhibit extensive morphological and functional variation, much of which is thought to be adaptive (Wright & Gaut 2005). Since plants are sessile organisms, local processes may be particularly important in shaping genetic diversity. Yet, studies of genetic variation in plants have largely ignored local sampling in outcrossing species, making it difficult to infer the action of positive selection. This has changed only very recently with the multilocus surveys of local populations of the wild ancestor of maize (Moeller *et al*. 2007) and tomato (Arunyawat *et al*. 2007). These two studies illustrate the important point that local population sampling may provide information about the demographic history that needs to be relatively well known before the action of past selection can be inferred.

In plants, the genomic approach to adaptation and natural selection that is highlighted in this review is most advanced for *Arabidopsis lyrata*. *A. lyrata* has become a model system for plant molecular population genetics, in part because it has large population sizes (Savolainen *et al*. 2000; Ramos-Onsins *et al*. 2004), thereby facilitating the detection of natural selection. Based on analyses of genetic differentiation between populations, positive selection has been found at several genes, including *GLABROUS1* (which is essential for the initiation of trichome formation). Differentiation at the *GLABROUS1* gene was higher than at neutral marker loci, which suggests that trichome production is subject to divergent selection (Kärkkäinen *et al*. 2004). Flowering time also appears to be under strong selection, with large differences in day-length requirements between northern and southern populations (Riihimäki *et al*. 2005). In a systematic search for adaptive evolution (S. Wright, York University, Toronto, Canada, personal communication), DNA sequence polymorphism for 71 *A. lyrata* plants coming from six different natural populations was obtained. Diversity was analysed at 77 mostly exonic fragments with a length of up to 800 bp. Based on these data the authors modelled the demographic history of the species sample which then served as a null model for tests using F_{ST}. Eight genes showed a signal of elevated F_{ST} in at least one pairwise population comparison, indicating the action of adaptive differentiation. After correction for multiple testing, however, only the flowering time gene *FCA* remained statistically significant.

Among plants, domesticated crop species are interesting candidates for the application of genome scans for positive selection. Since these species have undergone strong and recent artificial selection, one would expect to detect signatures of selective sweeps around genes that are involved in the control of desired phenotypic traits selected for during domestication. Maize (*Zea mays* ssp. *mays*) has been a focal species in such research. In a study investigating microsatellite variability (Vigouroux *et al*. 2002) a total of 501 loci were typed in 50 accessions representing modern-maize landraces along with 27 accessions of teosinte (*Z. mays* ssp. *parviglumis*), the wild ancestor of cultivated maize, and 23 accessions of *Z. mays* ssp. *mexicana*, which frequently forms hybrids with maize in the wild. Loci that showed a reduction in microsatellite variability within maize along with an increased F_{ST} when compared to teosinte were designated as putative targets of selection. Since domestication is associated with a strong population bottleneck the authors estimated parameters for the demographic history of maize. This demographic scenario was then used to assess deviations from neutral expectations. 15 loci showed a significant signal of non-neutral evolution, and one of the candidates, a gene encoding a MADS box transcriptional regulator, was further studied at the SNP level. This gene is of particular interest, as it is located in close proximity to a quantitative trait locus associated with the different structure of ears between maize and its wild ancestor (Doebley & Stec 1993). The DNA polymorphism data revealed lower levels of diversity than expected along with significant tests (Hudson *et al*. 1987), indicative of positive selection in the region.

In a more recent study, Wright *et al*. (2005) typed SNPs for 774 genes for 14 inbred lines of maize and 16 inbred lines of teosinte. In order to find genes under selection in cultivated maize the authors developed a likelihood-ratio test based on two different demographic models. At first, parameters for the most likely population bottleneck associated with domestication are obtained using the full data set. Then, a bottleneck of ten-fold intensity is modelled. This model mimics the loss of diversity through positive selection in addition to demography. Using this test the authors find that there is statistical support for the presence of two classes of genes. One class of genes fits the domestication population bottleneck model, while the other class has undergone a bottleneck of 10-fold intensity and therefore seems to additionally have been under positive selection. Overall, 2–4% of all genes seem to have been selected during domestication. The list of candidates that show the highest posterior probability of belonging to the selected class contains many genes with putative functions in plant growth and genes involved in amino acid biosynthesis.

Discussion

The availability of large-scale data for population genetics studies has offered a possibility for the precise identification of the footprints of hitch-hiking events in the genome. While up to now this kind of analysis has been only performed in few model species the increasing accessibility of DNA sequencing technology will allow the generation of population genetic data also in nonmodel organisms. The constant accumulation of such data in more and more species and populations allows us to address questions, such as (i) are genes involved in certain functions more subject to adaptive evolution than others? (ii) is positive selection more frequent in populations that have to adapt to new environments? and (iii) what is the rate at which adaptive substitutions occur? The currently available data can already give us hints in answering these questions.

Functional categories of genes under selection

The genes that were identified by selection mapping in natural populations of *D. melanogaster*, mice and plants appear to fall into three functional categories: genes in sensory pathways (i.e. genes involved in the development of the eye, skin or hairs), genes determining body size, and defence/immunity genes. Although the number of genes detected so far is small, the emerging pattern confirms the working hypothesis that most genes identified on the basis of selective sweeps play a role in ecological adaptation. Among the genes mentioned above, it appears that only *ph-p* (encoding part of a universal transcription repressor) may not have a specific function related to the environment. On the other hand, it is remarkable that genes involved in temperature adaptation and energy metabolism have not yet been identified in flies by the selective sweep method.

For the genes that experienced positive selection in humans, further categories (or subcategories) can be defined. For example, the genes responding to the selection pressures in the transition to novel food sources with the advent of agriculture form a new category (including *LCT*). Furthermore, olfactory and pigmentation genes are important subcategories of the genes involved in sensory perception (Nielsen *et al*. 2007).

It should be noted, however, that the identification of a specific gene or function might not always be possible. There is accumulating evidence that selection also affects noncoding portions of the genome (e.g. Andolfatto 2005; Bush & Lahn 2005). As the biological role of these noncoding regions is poorly understood up to now, the assessment of the functional consequences of positive selection on such loci poses an additional challenge.

Geographical distribution and rate of adaptive evolution

In both flies and humans the signatures of selection are to some extent population specific and thus suggestive of local adaptation. Voight *et al*. (2006) found that the strongest signals of selection are from human populations in Africa (Yoruba). Williamson *et al*. (2007), however, detected more evidence for sweeps in Chinese and European-American populations than in the African-American population. These contradictory results may be due to the fact that the power to detect selective sweeps is lower in the African-American sample.

In *D. melanogaster*, in five of the six cases discussed above, both African and non-African samples were analysed and in four of them the sweep originated in Africa. This result is not consistent with the hypothesis that the novel environments of flies encountered imposed new selective pressures, which in turn led to an increased rate of local selective sweeps. Whether this result is a consequence of a lack of power is unclear at present. Nonetheless, it is consistent with the analysis of Li & Stephan (2006) who found no difference in the rate of adaptive substitution between African and European populations in an X chromosome-wide analysis. This issue needs to be revisited as soon as more data are available.

The estimated rates of adaptive substitutions obtained by Li & Stephan (2006) agree surprisingly well with earlier estimates based on DNA sequence divergence between *Drosophila simulans* and *Drosophila yakuba* (Smith & Eyre-Walker 2002). However, the latter study estimates the rate of adaptive substitutions over a long time period and also includes weak selection into account. As Li & Stephan (2006) only estimate the rates of relatively young and strong selection events, this might indicate an acceleration of adaptive evolution in recent times.

The reliability of outlier methods

Despite the availability of whole-genome data, identifying the true signature of positive selection still remains a challenge. Empirical approaches use polymorphism data to detect candidate genes as outliers by calculating summaries of the data from a large number of loci. Their success depends on (i) the extent that positive selection has shaped the genome, and (ii) how well the statistics are differentiated between the neutral and selected loci. If positive selection has acted upon a large part of the genome, empirical approaches may miss many genes that have undergone adaptive evolution (Hahn 2008). Furthermore, simulation studies (Teshima *et al*. 2006) suggest that the efficiency of outlier methods is higher when the population size remains constant and directional selection acts on new and codominant variants.

Conclusion

Of the goals set up at the beginning of this research endeavour, several have been met. Statistical tools are now available to find evidence of recent, strongly positive selection and to estimate the strength and targets of selection from a huge amount of DNA sequence data. Of most practical importance is the ability to map the target genes of selection because this may open up new ways to study adaptation and understand disease factors in humans. However, to make progress in these directions, it is important that the genes under selection are also functionally analysed. For many of the genes identified by selection it is not clear what the function is. For most of them we have only a vague notion why the gene was under recent selection and what the molecular variant under selection is. Here additional studies, such as QTL analysis relating the selection mapping approach to specific phenotypes and transgenic experiments for finding the molecular variants targeted by selection will be important research directions for the future.

Acknowledgements

This research was supported by grants from the Volkswagen-Foundation (I/78815 and 824234-1) and the Deutsche Forschungsmeinschaft (Ste 325/6 and 325/7).

References

Adams MD, Celniker SE, Holt RA *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

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.

Andolfatto P (2001) Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution*, **18**, 279–290.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**, 1149–1152.

Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution*, **24**, 2310–2322.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Barton NH (1998) The effect of hitch-hiking on neutral genealogies. *Genetical Research*, **72**, 123–133.

Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila melanogaster* have a unique origin. *Molecular Biology and Evolution*, **21**, 1482–1491.

Bauer DuMont V, Aquadro CF (2005) Multiple signatures of positive selection downstream of *Notch* on the X chromosome in *Drosophila melanogaster*. *Genetics*, **171**, 639–653.

Bauer DuMont V, Fay JC, Calabrese PP, Aquadro CF (2004) DNA variability and divergence at the notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics*, **167**, 171–185.

Beaumont MA (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

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

Beisswanger S, Stephan W (2008) Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **105**, 5447–5452.

Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics*, **172**, 265–274.

Bersaglieri T, Sabeti PC, Patterson N *et al*. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, **74**, 1111–1120.

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.

Bush EC, Lahn BT (2005) Selective constraint on noncoding regions of hominid genomes. *PLoS Computational Biology*, **1**, e73.

Calboli FC, Gilchrist GW, Partridge L (2003) Different cell size and cell number contribution in two newly established and one ancient body size cline of *Drosophila subobscura*. *Evolution: International Journal of Organic Evolution*, **57**, 566–573.

Craymer L, Roy E (1980) Report of new mutants — *Drosophila melanogaster*. *Drosophilia Information Service*, **55**, 200–204.

Crow J, Kimura M (1970). *An Introduction to Population Genetics Theory*. Burgess Publishing, Minneapolis, Minnesota.

David JR, Capy P (1988) Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics*, **4**, 106–111.

Doebley J, Stec A (1993) Inheritance of the morphological differences between maize and teosinte: comparison of results for two F_2 populations. *Genetics*, **134**, 559–570.

Ewens WJ (2004) *Mathematical Population Genetics I. Theoretical Introduction*, 2nd edn. Springer, New York.

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

Fu YX (1995) Statistical properties of segregating sites. *Theoretical Population Biology*, **48**, 172–197.

Glinka S, De Lorenzo D, Stephan W (2006) Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **23**, 1869–1878.

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

Gockel J, Robinson SJ, Kennington WJ, Goldstein DB, Partridge L (2002) Quantitative genetic analysis of natural variation in body size in *Drosophila melanogaster*. *Heredity*, **89**, 145–153.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution: International Journal of Organic Evolution*, **62**, 255–265.

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. *American Journal of Human Genetics*, **70**, 369–383.

Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **99**, 12949–12954.

Harr B, Voolstra C, Heinen TJ *et al*. (2006) A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse *Mus musculus domesticus*. *Journal of Evolutionary Biology*, **19**, 1486–1496.

Hinds DA, Stuve LL, Nilsen GB *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.

Hudson R (1998) Island models and the coalescent process. *Molecular Ecology*, **7**, 413–418.

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics*, **177**, 469–480.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D (2006) An analysis of signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution*, **23**, 790–797.

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

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

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, Bauer DuMont VL, Ashmore AB, Gutierrez A, Aquadro CF (2007) Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics*, **177**, 1071–1085.

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

Kärkkäinen K, Loe G, Agren J (2004) Population structure in *Arabidopsis lyrata*: evidence for divergent selection on trichome production. *Evolution: International Journal of Organic Evolution*, **58**, 2831–2836.

Kauer MO, Dieringer D, Schlötterer C (2003) A microsatellite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*. *Genetics*, **165**, 1137–1148.

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

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

Lachaise D, Cariou M-L, David J *et al*. (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evolutionary Biology*, **22**, 159–225.

Li H, Stephan W (2005) Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics*, **171**, 377–384.

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.

Moeller DA, Tenaillon MI, Tiffin P (2007) Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics*, **176**, 1799–1809.

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, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Nijenhuis T, Hoenderop JG, Nilius B, Bindels RJ (2003) (Patho)physiological implications of the novel epithelial Ca2+ channels TRPV5 and TRPV6. *Pflügers Archiv: European Journal of Physiology*, **446**, 401–409.

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution*, **22**, 2119–2130.

Orengo DJ, Aguadé M (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.

Orengo DJ, Aguadé M (2007) Genome scans of variation and adaptive change: extended analysis of a candidate locus close to the phantom gene region in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **24**, 1122–1129.

Pool JE, Bauer DuMont V, Mueller JL, Aquadro CF (2006) A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. *Genetics*, **172**, 1093–1105.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M (2004) Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics*, **166**, 373–388.

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

Riihimäki M, Podolsky R, Kuittinen H, Koelewijn H, Savolainen O (2005) Studying genetics of adaptive variation in model organisms: flowering time variation in *Arabidopsis lyrata*. *Genetica*, **123**, 63–74.

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

Sabeti PC, Varilly P, Fry B *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Savolainen O, Langley CH, Lazzaro BP, Fr H (2000) Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **17**, 645–655.

Schlötterer C, Vogl C, Tautz D (1997) Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. *Genetics*, **146**, 309–320.

Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. *Nature*, **415**, 1022–1024.

Soejima M, Tachida H, Ishida T, Sano A, Koda Y (2006) Evidence for recent positive selection at the human AIM1 locus in a European population. *Molecular Biology and Evolution*, **23**, 179–188.

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.

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nature Reviews Genetics*, **3**, 137–144.

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.

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

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*, **175**, 737–750.

Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007) Progress and prospects in mapping recent selection in the genome. *Heredity*, **98**, 340–348.

Tishkoff SA, Varkonyi R, Cahinhinan N *et al*. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, **293**, 455–462.

van de Graaf SF, Hoenderop JG, Gkika D *et al*. (2003) Functional expression of the epithelial Ca(2+) channels (TRPV5 and TRPV6) requires association of the S100A10-annexin 2 complex. *EMBO Journal*, **22**, 1478–1487.

Vigouroux Y, Jaqueth JS, Matsuoka Y *et al*. (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution*, **19**, 1251–1260.

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

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**, 1468–1476.

Williamson SH, Hubisz MJ, Clark AG *et al*. (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, **3**, e90.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, **22**, 506–519.

Wright SI, Bi IV, Schroeder SG *et al*. (2005) The effects of artificial selection on the maize genome. *Science*, **308**, 1310–1314.

The authors have common interests in theoretical and molecular population genetics and, in particular, in the detection of natural selection at the genome level. They focus on the development of statistical inference methods, and carry out genomic studies of DNA sequence and gene expression variation in Drosophila and wild tomatoes.
