PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Genetic hitchhiking versus background selection: the controversy and its implications

Wolfgang Stephan

Phil. Trans. R. Soc. B 2010 **365**, 1245-1253
doi: 10.1098/rstb.2009.0278

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 1245–1253
doi:10.1098/rstb.2009.0278

Review

Genetic hitchhiking versus background selection: the controversy and its implications

Wolfgang Stephan*

Section of Evolutionary Biology, Department of Biology II, Ludwig-Maximilians University Munich, Grosshaderner Strasse 2, 82152 Planegg, Germany

The controversy on the relative importance of background selection (BGS; against deleterious mutations) and genetic hitchhiking (associated with positive directional selection) in explaining patterns of nucleotide variation in natural populations stimulated research activities for almost a decade. Despite efforts from many theorists and empiricists, fundamental questions are still open, in particular, for the population genetics of regions of reduced recombination. On the other hand, the development of the BGS and hitchhiking models and the long struggle to distinguish them, all of which seem to be a purely academic exercise, led to quite practical advances that are useful for the identification of genes involved in adaptation and domestication.

Keywords: background selection; genetic hitchhiking; reduction of nucleotide diversity in regions of restricted recombination; selective sweeps; adaptation

1. INTRODUCTION

Since publication of the seminal paper entitled 'The effect of deleterious mutations on neutral molecular variation' by B. Charlesworth, M. T. Morgan and D. Charlesworth in 1993, two major population genetic models have competed in explaining the observed reduction of nucleotide variation in genomic regions of reduced recombination (crossing-over). The level of neutral (or nearly neutral) variability can be reduced below classical neutral expectation by strong selection against recurrent deleterious alleles maintained by mutation (so-called background selection, BGS) or by substitutions of strongly beneficial alleles at linked nucleotide sites (termed genetic hitchhiking). The latter model was originally proposed by Maynard Smith & Haigh (1974), and then further developed by Kaplan *et al.* (1989), Stephan *et al.* (1992) and Barton (1998), as restriction fragment length polymorphism (RFLP) surveys in two *Drosophila* species revealed that in regions of low recombination, variation is drastically reduced (Aguadé *et al.* 1989; Stephan & Langley 1989). Since divergence to closely related species was relatively unaffected by recombination, mutagenic effects of recombination could be excluded, such that a neutral interpretation of the data became untenable (Begun & Aquadro 1992). Thus, until the year 1993 when the paper on BGS by Charlesworth and co-workers appeared, the observation of reduced nucleotide variation in regions of restricted recombination was generally interpreted as evidence for genetic hitchhiking (Aguadé *et al.* 1989; Stephan & Langley 1989; Lange *et al.* 1990; Begun & Aquadro 1992; Martín-Campos *et al.* 1992; Stephan & Mitchell 1992; Langley *et al.* 1993).

The discovery of reduced levels of variation in genomic regions of restricted crossing-over and the ensuing controversy over its interpretation initiated an important phase in the history of molecular population genetics (i.e. population genetics that uses molecular data such as RFLP, microsatellite or DNA sequence data). Since the phenomenon of reduced variation in regions of restricted recombination has also been found in organisms other than *Drosophila* (for instance, in humans (Nachman *et al.* 1998; Hellmann *et al.* 2003) and several plant species such as wild tomatoes (Stephan & Langley 1998; Roselius *et al.* 2005)), it has provoked extensive modelling and analysis efforts. The development of methods for distinguishing BGS and hitchhiking was a major activity in those years (until about 2000). An important question has been whether BGS alone can account for the patterns of variation we observe, or whether positive selection also needs to be invoked.

Although numerous attempts have been made, the controversy on the relative importance of BGS and genetic hitchhiking could not be resolved using data from genomic regions of low recombination. However, this situation changed around the year 2000 with the advent of genomics when—owing to technical advances—huge amounts of DNA sequence data (mostly from genomic regions of normal recombination) became available. The theoretical methods developed for disentangling positive and negative selection could then be readily adapted to these new data. Therefore, our efforts to distinguish BGS and hitchhiking have

*stephan@bio.lmu.de

One contribution of 16 to a Theme Issue 'The population genetics of mutations: good, bad and indifferent' dedicated to Brian Charlesworth on his 65th birthday.

had a significant impact on the analyses of DNA sequence evolution in the years to come. In the following I will describe both the controversy surrounding BGS and hitchhiking, and its major implications for population genetics and neighbouring fields.

2. PROPERTIES AND PREDICTIONS OF THE HITCHHIKING AND BACKGROUND SELECTION MODELS

The hitchhiking model as proposed by Maynard Smith & Haigh (1974) assumes that positive directional selection operates at a single locus that is partially linked to an existing neutral polymorphism. Thus it describes the reduction of nucleotide heterozygosity at a neutral site owing to a single hitchhiking (SHH) event caused by the fixation of a new beneficial allele. This reduction of heterozygosity at the time of fixation of the selected allele essentially depends on the ratio s/r, where—in the simplest case of no dominance—s is the selective advantage of the beneficial mutation and r the recombination fraction between the selected and neutral loci.

Further analysis has integrated finite population size into the SHH model using coalescent theory (Kaplan *et al.* 1989) and diffusion theory (Stephan *et al.* 1992). Both approaches showed that the results of single hitchhiking events on expected heterozygosity are relatively independent of effective population size N_e or the selection intensity $\alpha = 2N_e s$, if α is very large (>100). The same authors have also extended the SHH model that considers only single hitchhiking events to recurrent hitchhiking (RHH), where hitchhiking events occur randomly along a chromosome according to a time-homogeneous Poisson process (Kaplan *et al.* 1989; Wiehe & Stephan 1993). Apart from a reduction of expected heterozygosity, coalescent simulations have revealed that both hitchhiking models predict other important patterns in polymorphism data: (i) a skew in the site frequency spectrum towards rare variants (Braverman *et al.* 1995); and (ii) an excess of high frequency-derived alleles in recombining regions (Fay & Wu 2000). In particular, this second property has later become important in identifying hitchhiking events in chromosomal regions of normal recombination rates (Kim & Stephan 2002; see below).

While a hitchhiking effect may be caused by the fixation of a single advantageous mutation, the BGS model imagines a chromosome with many partially linked sites at which strongly deleterious alleles are maintained in a mutation–selection balance (Charlesworth *et al.* 1993). Furthermore, the BGS model in its original form assumes multiplicity of selective effects at the selected loci, so that with selection coefficient s and dominance coefficient h the fitness of the genotype heterozygous for i and homozygous for j mutant alleles is $(1 - hs)^i(1 - s)^j$. Recombination frequency between adjacent selected loci is r, and the deleterious mutation rate per diploid genome U. Typically, a large number of selected loci were used (Charlesworth *et al.* 1993).

The effect of BGS on a focal site at which a neutral (or weakly selected) polymorphism occurred was calculated analytically and by extensive computer simulations (Charlesworth *et al.* 1993; Hudson & Kaplan 1995; Nordborg *et al.* 1996; Stephan *et al.* 1999). This work showed that diversity at the focal locus is reduced by BGS owing to recurrent strongly deleterious mutations as though effective population size N_e is reduced by a factor that depends on U, the selection coefficients and the recombination rates between the selected sites and the focal locus. As alleles that carry deleterious mutations are eliminated within a relatively short time, a skew in the frequency spectrum of polymorphic sites was not predicted by this model, contrary to predictions of the SHH and RHH models.

3. EFFORTS TO DISTINGUISH THE HITCHHIKING AND BGS MODELS

In the early- to mid-1990s, the first attempts were made to explain the observed broad ('genome-wide') patterns of genetic variation in *D. melanogaster*, using the BGS (Charlesworth 1996, 2009) and the RHH models (Wiehe & Stephan 1993; Stephan 1995).

A general goal of the population genetics community at that time has been to distinguish the hitchhiking and BGS models. Since the SHH and RHH models predict a shift towards low-frequency variants, while the spectrum of the BGS model is essentially neutral, the standard approach for distinguishing these models has been based on analyses of the site frequency spectrum. The polymorphism data were mostly summarized in statistics such as Tajima's (1989) D, Fu & Li's (1993) D or Fay & Wu's (2000) H. These statistics measure distinct features of the frequency spectrum. In most cases, however, clear conclusions could not be reached. Deviations from the standard neutral model (of a panmictic population of constant size) were relatively infrequent, owing to small sample sizes and a focus on individual genes (instead of genomic approaches). A more conceptual problem of these analyses has been that these summary statistics were oriented towards the standard neutral model, but did not take deviations such as demography and population structure into account. Efforts alleviating this problem by—for instance—calculating Tajima's D for non-equilibrium populations of varying size are relatively recent (Innan & Stephan 2000; Zivkovic & Wiehe 2008), but have not yet been fully exploited in data analyses.

Another test of the BGS hypothesis for subdivided populations was proposed by Stephan *et al.* (1998) and further developed by Chen *et al.* (2000) and Baines *et al.* (2004). They simulated BGS for genes in regions of low recombination in a finite island model assuming that the migration rates at these loci are identical to those estimated for neutrally evolving reference genes (located in regions of normal recombination). Application to multi-locus *Drosophila ananassae* single nucleotide polymorphism (SNP) data showed that F_{ST} in regions of normal recombination is relatively homogeneous among all pairs of populations. However, subsets of northern (and also southern) populations exhibited significantly reduced

F_{ST} values at two loci located in regions of strongly restricted recombination (*furrowed* and *vermilion*) near the centromere of the X chromosome. The observed homogeneity among the northern and the southern populations deviates from predictions of the BGS model. Instead, this homogeneity may have been caused by recent local hitchhiking events that were limited to the northern (southern) species range of *D. ananassae*.

4. INTERFERENCE OF MUTATIONS IN REGIONS OF REDUCED RECOMBINATION

Following the insight that close linkage between selected variants in regions of reduced recombination may cause Hill–Robertson interference (Hill & Robertson 1966), the original BGS and RHH models that did not incorporate interference among selected alleles have been modified in recent years. Previous studies of RHH have assumed that at most one beneficial mutation is on the way to fixation at a given time (Kaplan *et al*. 1989; Stephan *et al*. 1992; Wiehe & Stephan 1993). However, for a high rate of selected substitutions and low recombination rate, this assumption can be easily violated. Kim & Stephan (2003) have investigated this problem using forward simulations and analytical approximations. They found that interference between linked beneficial alleles causes a reduction of their fixation probability. The hitchhiking effect on linked neutral variation for a given substitution also slightly decreases owing to interference, so that the strength of RHH is weakened. However, this effect is significant only in chromosomal regions of relatively low recombination rates where the level of variation is greatly reduced. Analytical approximations were obtained for the case of complete linkage.

These results suggest that the formula for the reduction of variation owing to RHH that was derived under the assumption that at most one beneficial mutation is on the way to fixation (Wiehe & Stephan 1993),

$$\pi = \theta\rho/(\rho + k\alpha\nu) \tag{4.1}$$

is also valid for relatively small recombination rates. Here, ρ is the local recombination rate per nucleotide site, α the selection intensity, ν the rate of adaptive substitution per site and $k = 0.075$. π is nucleotide diversity and $\theta = 4N_e\mu$, where μ is the nucleotide mutation rate. The function on the right-hand side of equation (4.1) is convex and thus different from the corresponding behaviour of the BGS model. To apply this formula to multi-locus polymorphism data, Innan & Stephan (2003) analysed the correlation between nucleotide diversity θ per locus and θ/ρ. For the *D. melanogaster* data of Andolfatto & Przeworski (2001) that were collected in regions of low recombination, they found that the correlation coefficient deviates significantly from that predicted by the BGS model. This suggests that positive selection needs to be invoked to explain the data.

Interference has recently also been taken into account for strongly deleterious mutations (Kaiser & Charlesworth 2009). Their model deviates from the original BGS model in that the effects of deleterious mutations present at multiple sites were no longer studied at a single focal neutral site, but pairs of adjacent selected sites followed by a neutral site were distributed along the chromosome. Thus, a piece of a chromosome was simulated mimicking gene regions of different length. The authors have carried out simulations of sequence evolution in genomic regions with reduced rates of recombination, which show that Hill–Robertson interference among strongly selected variants occurs when the region of restricted recombination is large. This reduces the effective strength of selection on strongly selected sites. As a consequence, the reduction of variation owing to BGS is less severe. This result agrees well with data from the *D. melanogaster* fourth chromosome and the neo-Y chromosome of *D. miranda*, in contrast to the results of previous modelling efforts based on the original BGS model. Using recent estimates of selection effects on deleterious mutations, these attempts had predicted near-zero N_e for chromosomal regions that lack recombination (Loewe & Charlesworth 2007), which, however, is not observed.

5. JOINT ACTION OF BGS AND RHH

One problem in estimating the parameter $\alpha\nu$ in formula (4.1) is that BGS has not been considered in the derivation of this equation (Stephan 1995). To remedy this problem, Kim & Stephan (2000) showed that the joint effects of BGS and RHH on neutral variation can be approximated by

$$\pi = \theta f\rho/(\rho + kf\alpha\nu), \tag{5.1}$$

where f describes the reduction of effective population size owing to BGS. Due to BGS, both f and ν depend on the recombination rate ρ. For small but not too low recombination rates, equation (5.1) can be approximated by

$$\pi = \theta\rho/(k\alpha\nu_0), \tag{5.2}$$

where ν_0 is the rate of beneficial substitution at zero recombination (provided f does not converge to zero for $\rho \to 0$; see Kaiser & Charlesworth (2009) for an analysis of this problem). This rate can be estimated for the *D. melanogaster* non-recombining fourth chromosome as follows. The time back to the last hitchhiking event in *D. melanogaster* is about $0.1N_e$ generations on average (Perlitz & Stephan 1997). Based on the genomic features of the fourth chromosome (its size is about 1% of the haploid genome), this corresponds to $\nu_0 = 5.9 \times 10^{-6} N_e^{-1}$. Using this value, the average selection coefficient can be estimated. For example, for an RFLP dataset of 15 loci from the third chromosome of an American *D. melanogaster* population by Aquadro *et al*. (1994), one obtains $\alpha\nu_0 = 4.6 \times 10^{-8}$. This leads to an average selection coefficient s of approximately 3.9×10^{-3}. However, the value of ν_0 in regions of restricted recombination on the third chromosome may be higher than on the fourth, since the density of potential target sites of strong selection on the third chromosome is likely to be higher. Thus, applying the ν_0 estimate of the fourth chromosome to the third chromosome probably

leads to an overestimate of s (for a given estimate of αv_0). Indeed, for African SNP datasets, Li & Stephan (2006) and Jensen *et al.* (2008) have estimated the average selection coefficient as approximately 2×10^{-3}, using more advanced maximum likelihood and approximate Bayesian methods, respectively.

6. BGS IN REGIONS OF NORMAL RECOMBINATION

While most work has examined BGS in large genomic regions of low recombination, some recent analysis focused on the level of a single gene or a small group of genes located in regions of normal recombination. Loewe & Charlesworth (2007) investigated how the effects of BGS caused by non-synonymous mutations are influenced by the length of coding regions, the number and length of introns, and intergenic distances. To calculate the impact of BGS on sequence variation, they used estimates of the distribution of fitness effects of non-synonymous mutations obtained from *D. melanogaster*. Results for genes in regions with normal recombination frequencies suggest that BGS may reduce the effective population size of genes somewhat, but not nearly to the same extent as in regions of low recombination. On the other hand, BGS may influence N_e of different regions of the same gene, consistent with observed differences in codon usage bias along genes. That is, BGS may influence phenomena that are usually included in the realm of weak selection at the level of genes, but strong reductions of diversity at individual genes in regions of normal recombination can usually (i.e. in the absence of heterogeneities in the mutation rate) not be attributed to the effects of BGS as the number of linked deleterious mutations is too small.

7. SELECTIVE SWEEPS IN REGIONS OF NORMAL RECOMBINATION

Since around the year 2000, research on positive directional selection and its effects on patterns of genetic variation in natural populations focused on regions of normal recombination. This shift in focus became possible through the availability of whole-genome sequence data from relevant species, in particular *Drosophila* and humans. It has allowed population geneticists to collect genome-wide data on DNA sequence variation and search for evidence for positive selection in these data.

These developments have had important consequences for the controversy between the BGS and hitchhiking models. The hallmark of SHH is an elimination or severe reduction of nucleotide diversity caused by a 'selective sweep', as SHH is now often referred to. As outlined above, in regions of normal recombination (and in the absence of heterogeneities in the mutation rate) this reduction cannot be explained by BGS, in contrast to the situation in regions of reduced recombination. This means that the—up to that time—joint model development of background selection and hitchhiking began to diverge.

(a) *New statistical tests for selection*

In 2002, Kim and Stephan showed that in regions of normal recombination rates, the level of genetic variation is greatly reduced at the site of strong directional selection and increases as the recombinational distance from the site of selection increases. Their main conclusion was that the characteristic footprint of SHH found in regions of restricted recombination should also be detectable in regions of normal recombination rates, if the distribution of SNPs along the genome is sufficiently dense (valleys of reduced variation are expected to occur on a smaller scale of typically only a few kilobases). Based on these results they proposed a statistical method (a composite likelihood ratio (*CLR*) test) to examine the significance of a local reduction of variation and a skew in the site frequency spectrum caused by a selective sweep. In this test, the presence of high frequency-derived variants (as proposed by Fay & Wu 2000) has been implemented as a critical property of a selective sweep in regions of normal recombination. This method also allowed them to estimate the strength and location of positive selection from DNA sequence data.

Since this time, searching for strong positive selection in the genome within a natural population has been the focus of a multitude of studies (Harr *et al.* 2002; Glinka *et al.* 2003; Akey *et al.* 2004; Orengo & Aguadé 2004, and so forth). 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. Some studies analysed SNP by directly sequencing small fragments of DNA at multiple loci. Other ones used microsatellite markers to measure polymorphism and looked for regions of depleted variability as an indicator of a selective sweep. 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. Ometto *et al.* 2005). Only very recently more sophisticated methods have been developed for analysing genome-wide polymorphism data, including tests based on the background site frequency spectrum (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 Tajima (1989) and Fu & Li (1993)). In addition, however, specific tests for positive selection to these sub-genomic regions such as a combination of the *CLR* and goodness-of-fit (*GOF*) tests (Kim & Stephan 2002; Jensen *et al.* 2005) and SweepFinder (Nielsen *et al.* 2005) are used; these latter tests have the advantage that they can also estimate the strength of selection and the approximate location of the beneficial mutation within the region (see Pavlidis *et al.* 2008).

Most recently, for a few model species (humans and *D. melanogaster*) continuous SNP data have become available along the entire genome. However, most of these data have not yet been analysed.

(b) *Footprints of recent positive selection in genome-wide and sub-genomic data*

In the following, I describe how these new tests have been used successfully to identify the targets of recent, strongly positive selection in model species.

(i) Drosophila melanogaster

The fruitfly 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. *Drosophila melanogaster* was originally a tropical species from sub-Saharan Africa. After the last glaciations it moved to the more temperate zones of Eurasia approximately 15 000 years ago (David & Capy 1988). As a human commensal, *D. melanogaster* is nowadays found on all continents. This provides an 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. Additionally, the study included the scan of a 578-kb large putatively neutrally evolving autosomal region, which served as a control.

The first genome scan with SNP markers has been performed by Glinka *et al.* (2003). Polymorphism was analysed in an African and a European population by sequencing short fragments of non-coding 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 dataset 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 and colleagues 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. Recently, SNP variability was also analysed at a large number of loci on chromosome 3 from the aforementioned two populations (Hutter *et al.* 2007).

The resulting list of sub-genomic regions has been 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 (explained above).

One of these regions encompassed the gene *polyhomeotic-proximal* (*ph-p*). 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) in the African population (Beisswanger *et al.* 2006; Beisswanger & Stephan 2008). It is relatively old, dating to about 50 000 years ago; i.e. to a time before *D. melanogaster* migrated out of Africa. The gene *ph-p* and its duplicate *ph-d* code for proteins of the polycomb group. Although the *ph* duplication is at least 25–30 Myr 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 had been driving the functional divergence of these gene duplicates. To my knowledge, this is the first case of neofunctionalization driven by positive selection in the presence of gene conversion. The Ph proteins are chromatin regulators that may have an important function in temperature adaptation (Schwartz & Pirrotta 2007).

Another successful application of the *CLR–GOF* tests was reported by Glinka *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).

Most recently Svetec *et al.* (2009) have reported evidence for a sweep in the 3′ end of the *HDAC6* gene in the African population of *D. melanogaster*. This gene codes for an unusual histone deacetylase that is localized in the cytoplasm and thought to be a key regulator of cytotoxic stress resistance. In this case, the target of selection was delimited to a region of 2.7 kb.

A further study investigating patterns of non-coding SNPs has been published by Orengo & Aguadé (2004). The authors concentrated on the X chromosome and analysed 109 short non-coding DNA fragments with an average distance of approximately 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 re-sequenced 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 with 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. Two more recent surveys have expanded the microsatellite dataset 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 two latter 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 datasets.

Re-sequencing 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 *CG18 508* and *Fcp3C* (*Follicle cell protein 3C*) in non-African populations (from the United States and China). However, while the *CLR* test produced a highly significant result, the *p*-value of the *GOF* test was relatively low (0.114), which indicates that demography rather than selection explains the data. 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 owing to positive selection (reviewed in Pavlidis *et al*. 2008).

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).

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 and the observed clinal pattern of variation of this trait (Pavlidis *et al*. 2008).

(ii) *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.

In plants, the genomic approach to adaptation and natural selection that is highlighted in this review is most advanced for *Arabidopsis lyrata*. This species has become a model system for plant molecular population genetics, in part, because it has large population sizes, thereby facilitating the detection of natural selection. In a systematic search for adaptive evolution, Ross-Ibarra *et al*. (2008) obtained DNA sequence polymorphism for 71 *A. lyrata* plants coming from six different natural populations. 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, which then served as a null model for tests using F_{ST}. Eight genes showed a signal of significantly elevated F_{ST} in at least one pairwise population comparison, indicating the action of adaptive differentiation. After correction for multiple testing, only the flowering time gene *FCA* remained statistically significant.

Domesticated crop species are also 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 (Vigoroux *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 a significantly increased F_{ST} when compared with 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. Fifteen 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. The DNA polymorphism data revealed lower levels of diversity than expected, 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 dataset. Then, a bottleneck of 10-fold severity 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 severity and therefore seems to additionally have been under positive selection. Overall, 2–4 per cent 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 contain many genes with putative functions in plant growth and genes involved in amino acid biosynthesis.

(iii) *Functional categories of the genes under positive selection*

The genes that have been identified by selection mapping in natural populations of *D. melanogaster* and plants appear to fall into four functional categories: genes in sensory pathways, genes determining body size and temperature adaptation, 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.

For the genes that experienced positive selection in humans, further categories (or sub-categories) can be defined (Voight *et al*. 2006). 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 the lactase gene, *LCT*). Furthermore, olfactory and pigmentation genes are important sub-categories of the genes involved in sensory perception.

8. CONCLUSIONS

The controversy surrounding the relative significance of strongly beneficial and deleterious mutations in evolution has stimulated research activities for almost a decade. Despite efforts from many theorists and empiricists, fundamental questions are still open, in particular for the population genetics of regions of reduced recombination. On the other hand, the development of the SHH/RHH and BGS models and the long struggle for distinguishing them led to quite practical advances that are useful for the identification of genes involved in adaptation and domestication, and of genes possibly important for biomedicine.

This paper is dedicated to Brian Charlesworth who has always been a friend and one of my most important academic teachers. I also want to commemorate a legendary day in May 2008 with friends (Adelgunde, Eve, Werner), an Alpine setting to include everything Bavarian (*Alphörner*, *Lederhosen*, *Dirndl*, etc.), and Brian's comments. This research was supported by the grants Ste 325/7-1 and 12-1 (Research Unit 1078) from the German Research Foundation.

REFERENCES

Adams, M. D. *et al.* 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195. (doi:10.1126/science.287.5461.2185)

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

Akey, J. M. *et al.* 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286. (doi:10.1371/journal.pbio.0020286)

Andolfatto, P. & Przeworski, M. 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158**, 657–665.

Aquadro, C. F., Begun, D. J. & Kindahl, E. C. 1994 Selection, recombination, and DNA polymorphism in *Drosophila*. In *Non-neutral evolution: theories and molecular data* (ed. B. Golding), pp. 46–56. New York, NY: Chapman & Hall.

Baines, J. F., Das, A., Mousset, S. & Stephan, W. 2004 The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. *Genetics* **168**, 1987–1998. (doi:10.1534/genetics.104.027482)

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

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

Beaumont, M. A. & Balding, D. J. 2004 Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13**, 969–980. (doi:10.1111/j.1365-294X.2004.02125.x)

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

Beisswanger, S. & Stephan, W. 2008 Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 5447–5452. (doi:10.1073/pnas.0710892105)

Beisswanger, S., Stephan, W. & De Lorenzo, D. 2006 Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* **172**, 265–274. (doi:10.1534/genetics.105.049346)

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

Charlesworth, B. 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* **68**, 131–149. (doi:10.1017/S0016672300034029)

Charlesworth, B. 2009 Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205. (doi:10.1038/nrg2526)

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

Chen, Y., Marsh, B. J. & Stephan, W. 2000 Joint effects of natural selection and recombination on gene flow between *Drosophila ananassae* populations. *Genetics* **155**, 1185–1194.

David, J. R. & Capy, P. 1988 Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet.* **4**, 106–111. (doi:10.1016/0168-9525(88)90098-4)

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

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

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.

Glinka, S., De Lorenzo, D. & Stephan, W. 2006 Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Mol. Biol. Evol.* **23**, 1869–1878. (doi:10.1093/molbev/msl069)

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

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

Hill, W. G. & Robertson, A. 1966 The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294. (doi:10.1017/S0016672300010156)

Hudson, R. R. & Kaplan, N. L. 1995 Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.

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 chromosome-wide single nucleotide polymorphism data. *Genetics* **177**, 469–480. (doi:10.1534/genetics.107.074922)

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

Innan, H. & Stephan, W. 2003 Distinguishing the hitchhiking and background selection models. *Genetics* **165**, 2307–2312.

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

Jensen, J. D., Bauer DuMont, V., Ashmore, A. B., Gutierrez, A. & Aquadro, C. F. 2007 Patterns of sequence variability and divergence at the *diminuitive* gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics* **177**, 1071–1085. (doi:10.1534/genetics.106.069468)

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

Kaiser, V. B. & Charlesworth, B. 2009 The effects of deleterious mutations on evolution in non-recombining chromosomes. *Trends Genet.* **25**, 9–12. (doi:10.1016/j.tig.2008.10.009)

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

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.

Lange, B. W., Langley, C. H. & Stephan, W. 1990 Molecular evolution of *Drosophila* metallothionein genes. *Genetics* **126**, 921–932.

Langley, C. H., MacDonald, J., Miyashita, N. & Aguadé, M. 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism in the *suppressor of forked* region in *Drosophila melanogaster* and *D. simulans*. *Proc. Natl Acad. Sci. USA* **90**, 1800–1803. (doi:10.1073/pnas.90.5.1800)

Li, H. & Stephan, W. 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* **2**, e166. (doi:10.1371/journal.pgen.0020166)

Loewe, L. & Charlesworth, B. 2007 Background selection in single genes may explain patterns of codon bias. *Genetics* **175**, 1381–1393. (doi:10.1534/genetics.106.065557)

Martín-Campos, J. M., Comerón, J. M., Miyashita, N. & Aguadé, M. 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *D. melanogaster*. *Genetics* **130**, 805–816.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.

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

Nielsen, R. *et al.* 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575. (doi:10.1101/gr.4252305)

Nordborg, M., Charlesworth, B. & Charlesworth, D. 1996 The effect of recombination on background selection. *Genet. Res.* **67**, 159–174. (doi:10.1017/S0016672300033619)

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. *Mol. Biol. Evol.* **22**, 2119–2130. (doi:10.1093/molbev/msi207)

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

Orengo, D. J. & 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*. *Mol. Biol. Evol.* **24**, 1122–1129. (doi:10.1093/molbev/msm032)

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

Perlitz, M. & Stephan, W. 1997 The mean and variance of the number of segregating sites since the last hitchhiking event. *J. Math. Biol.* **36**, 1–23. (doi:10.1007/s002850050087)

Pool, J. E., Bauer DuMont, V., Mueller, J. L. & Aquadro, C. F. 2006 A scan of molecular variation leads to a narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. *Genetics* **172**, 1093–1105. (doi:10.1534/genetics.105.049973)

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

Roselius, K., Stephan, W. & Städler, T. 2005 The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753–763. (doi:10.1534/genetics.105.043877)

Ross-Ibarra, J. *et al.* 2008 Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS One* **3**, e2411. (doi:10.1371/journal.pone.0002411)

Schwartz, Y. B. & Pirrotta, V. 2007 Polycomb silencing mechanisms and the management of genomic programmes. *Nat. Rev. Genet.* **8**, 9–22. (doi:10.1038/nrg1981)

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

Stephan, W. & Langley, C. H. 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts

between the *vermilion* and *forked* loci. *Genetics* **121**, 89–99.

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

Stephan, W. & Mitchell, S. J. 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. *Genetics* **132**, 1039–1045.

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

Stephan, W., Xing, L., Kirby, D. A. & Braverman, J. M. 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. *Proc. Natl Acad. Sci. USA* **95**, 5649–5654. (doi:10.1073/pnas.95.10.5649)

Stephan, W., Charlesworth, B. & McVean, G. 1999 The effect of background selection at a single locus on weakly selected, partially linked variants. *Genet. Res.* **73**, 133–146. (doi:10.1017/S0016672399003705)

Svetec, N., Pavlidis, P. & Stephan, W. 2009 Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. *Mol. Biol. Evol.* **26**, 1549–1556. (doi:10.1093/molbev/msp065)

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

Vigoroux, Y. *et al.* 2002 Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19**, 1251–1260.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. 2006 A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72. (doi:10.1371/journal.pbio.0040072)

Wiehe, T. H. E. & Stephan, W. 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**, 842–855.

Wright, S. I. & Gaut, B. S. 2005 Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* **22**, 506–519. (doi:10.1093/molbev/msi035)

Wright, S. I. *et al.* 2005 The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314. (doi:10.1126/science.1107891)

Zivkovic, D. & Wiehe, T. 2008 Second order moments of segregating sites under variable population size. *Genetics* **180**, 341–357. (doi:10.1534/genetics.108.091231)