MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 28–44 doi: 10.1111/j.1365-294X.2011.05308.x

INVITED REVIEW

Joint analysis of demography and selection in population genetics: where do we stand and where could we go?

JUNRUI LI,*† HAIPENG LI,* MATTIAS JAKOBSSON,‡ SEN LI,‡ PER SJÖDIN‡ and MARTIN LASCOUX*§

*Laboratory of Evolutionary Genomics, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China, †Graduate School of the Chinese Academy of Sciences, Beijing 100039, China, ‡Department of Evolutionary Biology, Evolutionary Biology Centre, and Science for Life Laboratory, Uppsala University, 752 36 Uppsala, Sweden, §Program in Plant Ecology and Evolution, Department of Ecology and Genetics, Uppsala University, 752 36 Uppsala, Sweden

Abstract

Teasing apart the effects of selection and demography on genetic polymorphism remains one of the major challenges in the analysis of population genomic data. The traditional approach has been to assume that demography would leave a genome-wide signature, whereas the effect of selection would be local. In the light of recent genomic surveys of sequence polymorphism, several authors have argued that this approach is questionable based on the evidence of the pervasive role of positive selection and that new approaches are needed. In the first part of this review, we give a few empirical and theoretical examples illustrating the difficulty in teasing apart the effects of selection and demography on genomic polymorphism patterns. In the second part, we review recent efforts to detect recent positive selection. Most available methods still rely on an a priori classification of sites in the genome but there are many promising new approaches. These new methods make use of the latest developments in statistics, explore aspects of the data that had been neglected hitherto or take advantage of the emerging population genomic data. A current and promising approach is based on first estimating demographic and genetic parameters, using, e.g., a likelihood or approximate Bayesian computation framework, focusing on extreme outlier regions, and then using an independent method to confirm these. Finally, especially for species where evidence of natural selection has been limited, more experimental and versatile approaches that contrast populations under varied environmental constraints might be more successful compared with species-wide genome scans in search of specific signatures.

Keywords: contemporary evolution, ecological genetics, population genetics—theoretical, population genetics—empirical

Received 14 June 2011; revision received 30 August 2011; accepted 7 September 2011

Introduction

Two major goals of population genetics are to reconstruct the demographic history of populations and species and to identify the parts of the genome that are, or have been, under natural selection. Unfortunately, selection and demographic events can leave very similar signatures in the genome, and one of the remaining challenges of population genetics is developing approaches to disentangle selection from demography. As we shall argue below, this problem is compounded by the fact that selection events are often associated with demographic changes. In this review, we will focus on positive selection and, in most cases, on recent selection (say, <0.1 N_e generations ago where N_e is the effective population size).

Correspondence: Martin Lascoux, Fax: +46 (18) 4716457;
E-mail: martin.lascoux@ebc.uu.se

When analysing sequence data, the basic approach to disentangling the effect of recent selection and demography has been to assume that all loci are affected in the same way by demography, whereas selection affects a particular genomic region. Hence, the genome-wide pattern of genetic variation will capture the effect of demography and the extreme tails of the distribution (the loci departing from the genome-wide pattern) will be suggestive of regions under selection. Based on recent genomic resequencing data in *Drosophila* (Begun *et al.* 2007) and based on theoretical results, in particular John Gillespie's work on genetic draft (Gillespie 2000, 2001), Hahn (2008) argued that this 'outlier' approach could be badly misleading if selection is extremely common in the genome (which seems to be the case in Drosophila). The patterns in the genome caused by genetic hitchhiking would then be incorrectly interpreted as reflecting demographic history. Ten years ago, such a claim might not have received much attention, but a series of recent studies has given more credence to a pervasive role of positive selection in shaping sequence polymorphism in genomes (Sella *et al.* 2009; Siol *et al.* 2010; Stephan 2010).

For example, some Drosophila species have very large effective population sizes and positive selection seems indeed important in those species (e.g. Kern *et al.* 2002; Li & Stephan 2006; Begun *et al.* 2007; Jensen *et al.* 2008; Andolfatto *et al.* 2011; Jensen & Bachtrog 2011; Sattah *et al.* 2011). However, the difficulty in distinguishing the effects of selection and demography is not restricted to species with very large effective population sizes. An illustrative example of the difficulties one can face when attempting to separate the effect of selection and demography is provided by a series of recent studies in humans—a species with a small effective population size—focusing on the ratio between the effective sizes of the X chromosomes, N_eX, and the effective size of the autosomes, N_eA, (Emery *et al.* 2010; Gottipati *et al.* 2011, and references therein). If males and females are present in equal numbers, the ratio $Q = N_eX/N_eA$ is expected to be 0.75. Many processes such as dispersal or mating are potentially sex-biased: females might tend to stay in their natal territory and males commonly have multiple female mates. If there is a male bias so that the variance of reproductive success is smaller for males than for females, Q will be <0.75, while if there is a female bias, it will be larger than 0.75. Two recent studies estimated Q from nucleotide diversity data and reached seemingly contradictory results. Keinan *et al.* (2009) using extensive single-nucleotide polymorphisms (SNPs) data found evidence for male bias, whereas Hammer and colleagues (Hammer *et al.* 2010) found evidence of female bias. Two (nonmutually exclusive) explanations have been put forward. Emery *et al.* (2010) first noted that the two studies used different summary statistics to estimate the Q ratio: Hammer *et al.* used nucleotide diversity, π, while Keinan *et al.* used Wright's fixation index F_{ST}. Emery *et al.* (2010) concluded that the two estimates detect sex-biased events on different timescales, which could explain the discrepancy. The estimate of Q based on π, Q_π, will not be well suited to detect recent sex biases and better reflects more ancient biases. The opposite is true for the estimate of Q based on F_{ST}, Q_{FST}. Finally, using coalescent simulations, Emery *et al.* (2010) showed that the data are consistent with a recent male bias and a more ancient female bias, an explanation that does not invoke selection. In contrast, Hammer *et al.* (2010) attempted to explain the discrepancy by the action of selection. That selection could be an important force moulding genetic variation on the X chromosome has been suggested by two other recent studies (Casto *et al.* 2010; Lambert *et al.* 2010) as well. Hammer *et al.* (2010) found that regions with a short genetic distance to genes have a deficit of diversity on the X resulting in a $Q_\pi < 0.75$, and regions further from genes have an excess of diversity ($Q_\pi > 0.75$) (see also Luca *et al.* 2011 for the same observation on autosomes). Hammer *et al.* (2010) concluded that in the case of the X chromosome, 'if we wish to disentangle the history of selection, recombination and demography, a targeted set of carefully chosen regions at sufficient genetic distance from functional element is needed. Intriguingly, at least for the human X chromosome, the signature left solely by demographic history may be hidden in the small fraction of selectively neutral polymorphisms that resides far from genes'. As pointed out by Emery *et al.* (2010), the distance from genes might suffice to explain the contrasting results obtained by Keinan *et al.* and Hammer *et al.* but not the discrepancy between the estimate of Q based on F_{ST} and the estimate of Q based on π. Using extensive genomic data, Gottipati *et al.* (2011) confirmed the increase in the ratio with distance from genes across populations, but they also showed that the ratio is lower in West Africans than in Europeans, something that is most easily explained by the contrasted demographic histories of the two populations. This last study also nicely illustrates how much can be gained by considering different populations when trying to disentangle the effects of selection and demography on genetic polymorphisms.

If Hammer *et al.* are correct and all species were to be similar to Drosophila or the X chromosome in humans, these results do not augur well for the separation of the effects of selection and demographic events. However, it is unclear how much can be extrapolated from these studies. Drosophila species often have extremely large effective population sizes, and positive

selection is therefore particularly efficient. This is vindicated by the fact that evidence for positive selection is overwhelming in Drosophila compared with humans and baker's yeast, which both have much smaller effective population sizes (e.g. Li *et al.* 2008; Hernandez *et al.* 2011; Sattah *et al.* 2011), and that within *Drosophila*, species with larger current effective population sizes also have higher rates of recent adaptive evolution (Jensen & Bachtrog 2011). Furthermore, the X chromosome has a lower recombination rate than the autosomes, which will enhance the effect of selection on diversity patterns across the chromosome because of linkage. Another empirical reason to have doubts about the generality of the conclusions drawn from Drosophila, or the X chromosome, is the fact that inferences of selection or demographics have, in some cases, been corroborated by independent data. For instance, Wright *et al.* (2005) used a maximum-likelihood approach to identify genes under selection since the domestication of modern maize from its ancestor teosinte. Many of the loci picked by this approach colocalize with QTLs and/or have been independently shown to be involved in morphological changes that led from teosinte to maize (e.g. *tb1*). Interestingly too, Wright *et al.* (2005) concluded that around 4–10% of the genome has been under selection since the domestication event, something that does not fit too well with the idea of selection affecting the whole genome. This discrepancy might have different causes: Wright *et al.* (2005) consider a short period of time and they look at domestication, where selection is different from natural selection. In the case of natural populations, the implication of the view that the entire genome has been, directly or indirectly, under selection implies that a very large number of loci are, or have been during the considered period, associated with fitness, where fitness is the ability of organisms to survive and reproduce in the environment where the organisms find themselves (Orr 2009). That large parts of the genome are affected by selection may not be too surprising if very long periods of time are considered (as when criteria such as d*N*/d*S* are used to quantify the importance of positive selection on individual loci), but it appears more paradoxical when shorter periods of time are considered. In the latter, one might expect that a more limited number of loci in the genome are related to fitness. Also, Wright *et al.* (2005) are dealing with complex traits for which models of hard sweeps—selection acting on new mutations—are less relevant than soft selection—selection from standing variation. At any rate, the latter seems to be underlying selective changes observed in artificial experiments (Burke *et al.* 2010; Johansson *et al.* 2010) and will leave a much weaker signature of selection than selective sweeps, implying that classical tests for selection will miss those. There are good reasons to believe that soft selection may be more important than hard sweeps also in natural populations (Colosimo *et al.* 2004, 2005; Pritchard *et al.* 2010).

In comparison with selection, it is generally more difficult to obtain independent support for demographic inferences. Yet, in some cases, demographic inferences based on genetic variation generate results that are supported by the fossil or climatic record. For example, in humans, estimates of the out-of-Africa migration event are reasonably similar based on genetic data and fossil data (Voight *et al.* 2005; Mellars 2006).

Hahn's criticism of the traditional approach may have been a bit excessive, but it, nonetheless, captures the major point, namely the intrinsic association between selection and demography. In this respect, the presence of a correlation between effective population size, viewed as a measure of random genetic drift, and measures of positive selection, such as mean d*N*/d*S* (e.g. Ellegren 2009), does not invalidate Hahn's criticism. Some 50 years ago, Barker and coworkers (Frankham *et al.* 1968a,b; Jones *et al.* 1968) demonstrated experimentally that random drift affects the efficacy of selection, at least in small populations, but the question that remains to this day is whether the effects of demography and selection can jointly be estimated from polymorphism data, without resorting to an a priori classification of the polymorphic sites or loci. Another aspect of the interplay between demography and selection recently pointed out by Pavlidis *et al.* (2010) is that 'in natural populations positive selection may occur simultaneously with demographic changes'. For instance, a species colonizing a new habitat is likely to simultaneously undergo a population bottleneck and experience intense selection as it adapts to its new environment. In a recent paper, Kim & Gulisija (2010) used computer simulations and analytical derivations to show that the details of the process of colonization of a new habitat do matter and can strongly affect our ability to detect selection. Kim & Gulisija (2010) compared a simple population split (PS) model with a more gradual and complicated adaptive niche expansion model in which a small number of migrants carrying specific alleles initiate a logistic growth in a new habitat [genotype-dependent colonization and introgression, (GDCI)]. Here, the first model can be viewed as a simplification of the second (Fig. 1). The resulting pattern of genetic variation at both linked and unlinked neutral loci was strikingly different. In particular, the GDCI model produces a greater excess of high-frequency-derived alleles than the PS model but a similar reduction in expected heterozygosity. So, if one assumes a PS model, but the population actually follows a GDCI model, one may greatly overestimate the importance of selection.

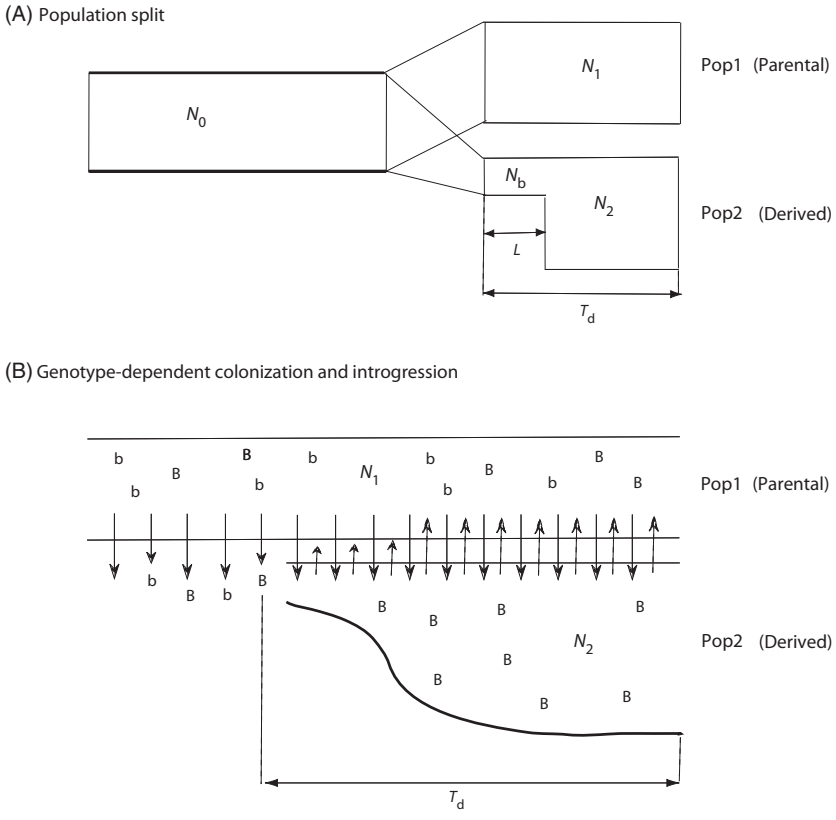


Fig. 1 Two demographic models approximating the same process of adaptive niche expansion but leading to different site frequency spectrum at neutral loci linked to loci under selection. The genotype-dependent colonization and introgression model produces a greater excess of high-frequency-derived alleles than the population split (PS) model but a similar reduction in expected heterozygosity. In both models, T_d is the time of PS. In B, arrows indicate the direction of migration. Migration is primarily from Pop1 to Pop2, but migration in the opposite direction is also possible. B is the beneficial allele in the new environment (adapted from Kim & Gulisija 2010).

Similarly, population expansions can create patterns in spatial distribution of allele frequencies similar to those expected from selection (Excoffier *et al.* 2009). Related models have also been developed in relation to speciation (Gavrilets 2004), although they have often been general and have not aimed at the estimation of specific parameters. Also, as emphasized by Zeng & Charlesworth (2010), inferring demographics based on an inadequate selection model may lead to unreliable results and *vice versa*. Zeng and Charlesworth's study focuses on codon usage bias and is based on introns and synonymous variation from a single population of *D. melanogaster* from Zimbabwe, which is presumably at equilibrium. Both autosomal and X chromosome data were considered and comparisons between autosomal and X-linked genes were restricted to genes falling within overlapping ranges of their evolutionarily effective recombination rates. Different demographic and selection models were fitted to the data, and for the overall sequence data, there was no support for recent demographic changes for the autosomes but evidence of population growth for the X chromosome. This is not expected as demography is supposed to affect all parts of the genome in the same way, although perhaps not to the same degree. However, when only synonymous sites were considered, there was no evidence of growth in either X or autosomes. In contrast, introns in the autosomal data showed weak evidence for population expansion, whereas X-linked introns indicated a fivefold increase in population size. For both chromosome sets, a model including population expansion *and* selective difference between AT and GC sites fits the intron data better than models that would include either population expansion *or* selective difference between AT and GC sites. Interestingly, a model including a recent change in mutational parameters as well as a model of an equilibrium population with two classes of selected sites fits the data equally well as the model with population expansion combined with a selective difference between AT and GC sites. Had one therefore assumed a constant population size, one would have inferred the presence of selection, but if one had focused on selective differences between AT and GC sites, one would have instead inferred some population expansion.

In summary, and as expressed by Wakeley (2010), it might be timely to start asking 'what is to become of the sophisticated coalescent machinery for making inferences about the demographic history of populations? To what extent will we be able to rely on genetic markers containing information about changes in population size over time or patterns of population

structure?' Should we really reconsider the standard approach used so far, as suggested by Hahn (2008)?

In the following, we will first briefly review what is known about the effect of selection on gene genealogies. Second, we will illustrate some effects of selection on demographic inferences, in particular in the case where selection and demographic change occur simultaneously. Finally, we will describe some possible directions to disentangle the signatures of selection and demographic events in the genome. Our aim is not to reiterate the point made by Hahn (2008) and by Sella *et al.* (2009) about the pervasive effect of selection in the genome but rather to explicitly consider which approaches are currently available when attempting to disentangle the effects of selection and demography and sketch out new directions.

Selection and gene genealogies: theoretical insights

The effect of selection on gene genealogies, in particular the effect of selective sweeps, has recently been reviewed by Wakeley (2010), and additional information is also available in Wakeley (2008). We refer the reader to those two sources for a more extensive coverage of the topic. Here, we will simply note the most important features that are relevant for this review. One early and surprising result of coalescent theory was that weak purifying selection has a limited effect on the genealogy of a sample (Golding 1997; Krone & Neuhauser 1997). However, as shown by Przeworski *et al.* (1999), while this is indeed true for the sites under selection, linked neutral variants would be affected and therefore there would still be a way to test for weak purifying selection along the genome. This observation, together with the fact that there are probably only a few strongly selected adaptive substitutions and that most signs of selection will be carried by linked sites, is the basis for the importance of hitchhiking models. There are a few different approaches to model gene genealogies at sites linked to loci under selection. The structured coalescent approach (Hudson & Kaplan 1986, Kaplan *et al.* 1988) is arguably one of those that give the most intuition about the relationship between selection and demography. The structured coalescent is also at the heart of simulation programmes modelling the coalescent with selection (Spencer & Coop 2004; Teshima & Innan 2009; Ewing & Hermisson 2010). The structured coalescent is based on the realization that the effect of selection is similar to that of population structure and it assumes that the effect of selection is strong enough relative to the effect of drift, so that random fluctuations in allele frequencies around the trajectory of the selected allele can be ignored. Under such an assumption, and if selection operates at a locus with two alleles A1 and A2, the rate of coalescence within an allelic class in a sample depends inversely on the allele frequency of the allele favoured by selection in that class. If there are i alleles in the class and the frequency of the allele is $x(t)$ at time t in the past, the coalescent rate is $i(i-1)/2x(t)$. Hence, the rate of coalescence is equal to the neutral rate of coalescence if $x(t) = 1$ but is greater than the rate under neutrality if $x(t) < 1$. Thinking forward in time, most branching events will take place at the beginning of the selective process and the time to fixation will be very short on a coalescent timescale. Selective sweeps are a bit more complicated as escape from selection through recombination needs to be considered but the dynamics of the processes are the same. Because selective sweeps occur so quickly, especially when selection intensity is high, selective sweeps will only be possible to detect in a small window of time.

The dynamics of selective sweeps in the presence of demographic changes are illustrated through their effect on the site frequency spectrum (SFS) in Box 1. Box 2 illustrates the joint effect of selection and population growth on the time to the most recent common ancestor. Box 3 illustrates the effect of fluctuating population size on the probability of fixation of an advantageous allele. It demonstrates that the probability of fixation for an allele with selective advantage s can be very different even between models with the same (well-defined and time-invariant) N_e.

Attempts to jointly estimate demography and selection or to avoid the problem

An approach estimating demography and selection in a single analysis is yet to be devised but there have been many attempts to define test statistics sensitive only to selection or to account for demography when estimating selection parameters.

Combining summary statistics

Various efforts have been made to combine existing summary statistics in order to capture different signals left by selection along the genome. Early attempts focused on the SFS (Zeng *et al.* 2006). The basic idea is to combine existing test statistics, for example Tajima's D and Fay and Wu's H. As D and H were designed to capture departures from the expected SFS in the intermediate- and high-frequency variant ranges, respectively, the new combined test statistics should be more powerful. The authors also conjectured that because D and H are sensitive to different demographic factors, the joint test should be less sensitive to other forces. For

Box 1 Interaction of demography and selection

Model description

To illustrate the interaction between demography and selection, two different demographic models with or without selection were constructed (Fig. 2). We considered two demographic models, a constant population size model and a population bottleneck model. In the constant population size model, effective population size is set to N_0. In the bottleneck model, a bottleneck happens $0.1T_0$ generations in the past (T_0 is equal to $4N_0$) and lasted for $0.05T_0$ generations during which population size was reduced to $0.1N_0$. For these demographic models, we investigated the case without selection or the case of positive selection. The beneficial mutation arose at two different time points: one is $0.007T_0$ (after bottleneck) and the other is $0.07T_0$ (in bottleneck). We also considered a more ancient introduction of the mutation, but in such cases, there was no effect

Simulations

The simulated DNA sequence is composed of 10 000 segments, each of length 1000 bp. Each segment is simulated independently and can be neutral or under selection (i.e. contain a beneficial allele right in the middle). The proportion of DNA segments under selection is 0% (neutral), 30%, 70% and 100%

The simulated data were generated using the program *mbs* (Teshima & Innan 2009, Table S1, Supporting information), which can incorporate any demographic history and any model of selection. Briefly, *mbs* runs coalescent simulations and output-relevant patterns of single-nucleotide polymorphism data, conditional on the histories (trajectories) of allele frequency and population size. In this simulation, the parameters are set as shown in Fig. 2: effective population size $N_0 = 10^6$, recombination rate $r = 0.05$, selective strength $s = 0.001$ if segment is under selection otherwise $s = 0$ (neutral). For each run, 100 trajectories are generated and 100 replicates are generated conditional on each trajectory

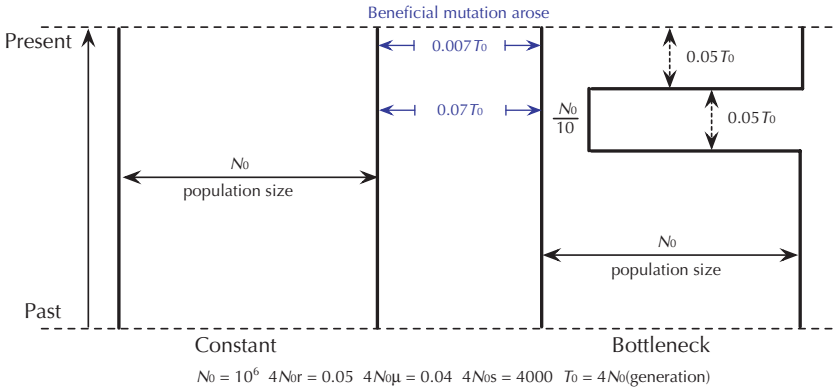


Fig. 2 Two different demographic models. The time at which selection occurs is indicated by the blue arrows.

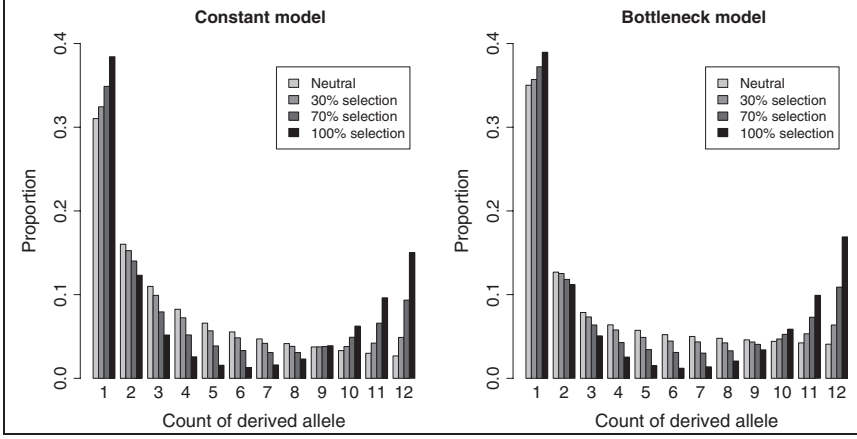


Fig. 3 Site frequency spectrum for the model in which beneficial mutation arose at $0.007T_0$ generations towards the past.

Box 1 *Continued*

Results

The site frequency spectra for DNA sequences generated under the different models were then calculated. Fig. 3 shows the site frequency spectrum (SFS) for the model in which beneficial mutation arose at $0.007T_0$ generations in the past. Compared with the constant neutral model, there is an excess of low-frequency and high-frequency alleles in the constant population model because of positive selection and also an excess of singletons and high-frequency allele in the bottleneck neutral model. However, the site frequency spectra are almost the same under both demographic models. This is mainly because selection happened after the bottleneck and it is so strong that the effect of the bottleneck can be ignored

Figure 4 shows the SFS for the model in which the beneficial mutation arose at $0.07T_0$ generations towards the past. Comparing the constant population models in Figs 3 and 4, we can see that the effect of selection is weakened because of a long time of neutral evolution after fixation of the beneficial allele. Furthermore, the bottleneck nearly wiped out the effect of positive selection

There is an interesting phenomenon in Fig. 4; both selection and bottleneck tend to decrease the proportion of doubletons, and however, their interaction leads to a weaker rather than a stronger reduction

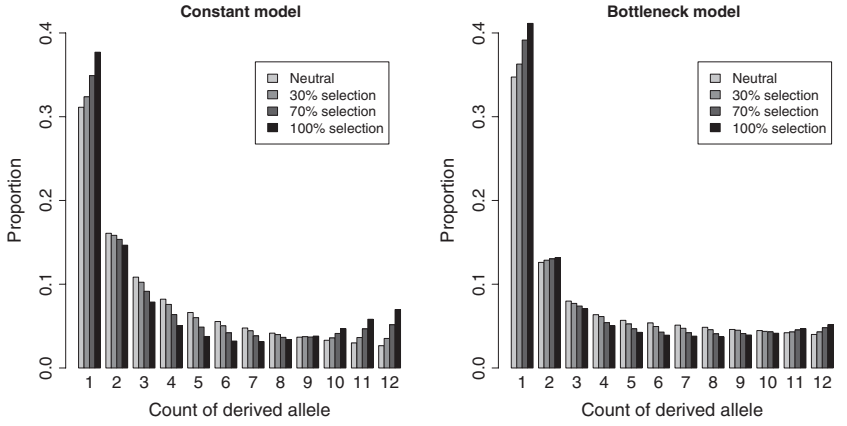


Fig. 4 Site frequency spectrum for the model in which beneficial mutation arose at $0.07T_0$ generation towards the past.

example, the sensitivity of D to population expansion may be counterbalanced by the insensitivity of H to the same factor in the joint DH test. However, results from simulations are not particularly conclusive and later simulation studies (Li 2011) show that most tests based on the SFS tend to be sensitive to both demography and selection. While DH test proposed by the Zeng *et al.* (2006) focused exclusively on the SFS, the composite of multiple signals (CMS) test proposed by Grossman *et al.* (2010) relies on three different signatures of positive selection, namely (i) long haplotypes, (ii) high-frequency-derived alleles and (iii) high differentiation among populations at selected loci. Combining these three signals, apart from reducing the dependence on demography, will extend the temporal range captured by CMS with long haplotypes confined to recent selective sweeps, population differences persisting over an intermediate-range and high-frequency-derived alleles being able to capture more ancient events (Sabeti *et al.* 2006). In contrast to Zeng's DH test that was rooted in 'classical' single-locus population genetics—even if, in practice, many loci may be needed to carry out the test in a meaningful way—the CMS implicitly assumes that genomic data are available. In coalescent simulations implementing demographic scenarios inspired by human history, the new composite test proved rather robust with only a slight increase in false positive for the most extreme bottleneck. Pavlidis *et al.* (2010) also assessed the power of a combination of tests statistics based on the SFS and linkage disequilibrium to detect positive selection in equilibrium and nonequilibrium populations. For severe bottlenecks, as those estimated in *Drosophila melanogaster* (Li & Stephan 2006; Thornton & Andolfatto 2006), the separate statistics fail to identify the target of selection. Pavlidis *et al.* (2010) also explicitly modelled a case where selection occurred during the bottleneck period, a scenario that is plausible, for example, in a case of colonization of a new habitat. When selection occurred at the end or the beginning of the bottleneck, the effects of selection cannot be discriminated from the neutral spectrum of variation, regardless of a shallow or a deep bottleneck. However, when

Box 2 Selection and growth affecting the time to the most recent common ancestor

Positive selection can have a strong effect on gene genealogies causing shorter expected time to the most recent common ancestor (T_{MRCA}) in affected genome regions than in neutral regions (e.g. Nordborg 2001). Several demographic changes, such as population growth, will also shorten the expected T_{MRCA} (e.g. Griffiths & Tavaré 1994), which can make signals of selection difficult to detect for many species and populations

We used the software *msms* (Ewing & Hermisson 2010) to simulate gene genealogies for a model with positive selection where the population either has a constant size or grows exponentially (with a growth rate α). Like the program *mbs* that was used in Box 1, the program *msms* combines forward simulation and coalescent (backward) simulations. We simulate a partial sweep that starts $0.1 \times 4N_e = 4000$ generations before present (using the 'SI' switch). The frequency of the advantageous variant was set to zero $0.1 \times 4N_e = 4000$ generations ago, and the forward mutation rate ($4N_e\mu$') from the neutral variant to the advantageous variant was set to 1 (using the 'Smu' switch), so that, on average, one new advantageous variant arises every generation. The selection strength ($2N_es$) varied from 0 to 500 (using the 'Sc' switch) for one copy of the advantageous variant, and the selection intensity for the advantageous homozygote was assumed to be twice the intensity of the heterozygote (fitness values: AA: $1 + 2s$, Aa: $1 + s$, aa: 1). We simulate gene genealogies for 100 haploid individuals and set the effective population size to $N_e = 10\,000$ and the growth rate to $4N_e\alpha = 100$ (except for the case of constant size). In the cases of population growth, the start of the growth was set to $0.02 \times 4N_e = 800$ generations, $0.06 \times 4N_e = 2400$ generations and $0.1 \times 4N_e = 4000$ generations before present. We simulated 1 million replicates with different values of the selection strength (including no selection) and plotted the distributions of the T_{MRCA} in Fig. 5

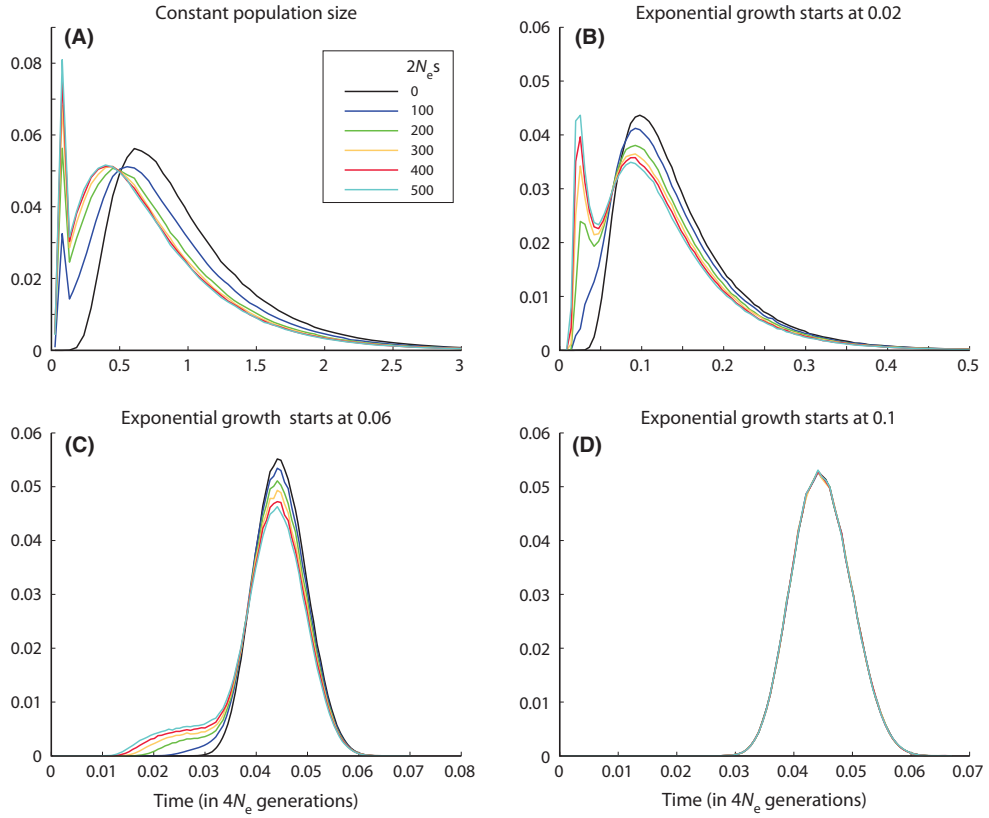


Fig. 5 Distributions of T_{MRCA} for different selection intensities in a population with (A) constant population size, (B) exponential growth starting $0.02 \times 4N_e = 800$ generations ago, (C) exponential growth starting $0.06 \times 4N_e = 2400$ generations ago and (D) exponential growth starting $0.1 \times 4N_e = 4000$ generations ago. Note that in all cases, the selection for the advantageous variant starts $0.1 \times 4N_e = 4000$ generations ago.

Box 2 *Continued*

For the case of a population with constant size (Fig. 5A), the distribution of T_{MRCA} has a peak around the time of the start of selection (0.1), which is not seen in the neutral case. Generally, the greater the selection intensity, the more pronounced becomes the peak around 0.1 shifting the T_{MRCA} away from the neutral expectation. For the cases of population growth starting $0.02 \times 4N_e$ generations before present (Fig. 5B), the T_{MRCA} are generally much shorter than for the constant population case. For the cases with strong selection, a mode appears around the expansion starting time. When the expansion starting time is closer to the selection start time (Fig. 5C), the distribution of T_{MRCA} is becoming unimodal with the mode smaller than both the expansion starting time and the selection starting time. However, the effect of selection can still be observed in very short T_{MRCA}:s for some genealogies. When the selection and the population expansion start simultaneously at 0.1 (Fig. 5D), the effect of population expansion and the effect of selection are confounded and the distributions of T_{MRCA} are indistinguishable

The combined effect of selection and population growth produces some irregular patterns in the distributions of T_{MRCA} that depend on the onset of both selection and growth. If the population starts growing much later than the start of the selection, genes under selection have, on average, much shorter T_{MRCA}, and the distribution of T_{MRCA} is quite different from the neutral case (e.g. Fig. 5B). However, if the selection and the population expansion start relatively close in time, it is difficult or even impossible to separate distributions of T_{MRCA} for the neutral and the selection cases, and in these cases, it will be difficult to detect signals of selection

Box 3 Probability of fixation in a fluctuating environment

A classical result in population genetics is that the probability of fixation of an allele with frequency p in the population is

$$\frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}}$$

where s is the selective advantage of the allele and N_e is the effective population size (Kimura 1957, 1962). As a new mutation is initially at frequency $1/2N$, where N is the census size, the probability of fixation for a new mutation is

$$\frac{1 - e^{-2N_e s/N}}{1 - e^{-4N_e s}} \approx 2s\frac{N_e}{N}$$

which is equal to $2s$ if $N = N_e$ (that the probability of fixation is approximately $2s$ in the case of a constant population dates back to Haldane (1927)). Imagine a population that has population size N_1 every odd generation and $N_2 = bN_1$ every even generation, so that the population size oscillates back and forth between N_1 and N_2 (this is a simplification introduced by Ewens 1967). Population size changes every generation which ensures that the effective population size is well defined if N_1 and N_2 are not too small (Krone & Nordborg 2002; Sjödin *et al.* 2005) and equal to the harmonic mean $N_e = 2N_1 b/(1 + b)$. The probability that a mutation will arise in an odd generation is then $N_1/(N_1 + N_2) = 1/(1 + b)$ and in an even generation $b/(1 + b)$. The probability that a mutation with selective advantage s is fixed given this demographic model is then

P(mutation is fixed/mutation during odd generation) $\times$ P(mutation during odd generation)
+ P(mutation is fixed/mutation during even generation) $\times$ P(mutation during even generation)
or

$$2s\frac{N_e}{N_1}\frac{1}{1+b} + 2s\frac{N_e}{N_2}\frac{b}{1+b} = 2s\frac{4b}{(1+b)^2}$$

This expression is maximized for $b = 1$ ($N_1 = N_2 = N_e$) and tends to zero for both $b \to 0$ and $b \to \infty$ (Fig. 6). The probability of fixation is thus lowered by a factor $4b/(1 + b)^2$ relative to $2s$—the fixation probability when the population size is constant. In other words, s and N_e do not provide sufficient information to calculate the fixation probability even in cases where N_e is well defined and constant over time. This point is well illustrated by a recent

Box 3 *Continued*

empirical study by Karasov *et al.* (2010) where they study a gene in which they find several mutations that entail insecticide resistance in *Drosophila melanogaster* on the same genetic background. They show that such a fast increase in frequency—and on the same genetic background—would be impossible given the time that the insecticide has been used and the well-accepted estimate of $4N_e\mu = 0.001$ under standard population genetic models (as well as some more complicated models). They point out that what is important for the birth and fixation of new mutations is not the effective population size (at least not the effective population size measured over long periods of time) but instead the census population sizes during recent times

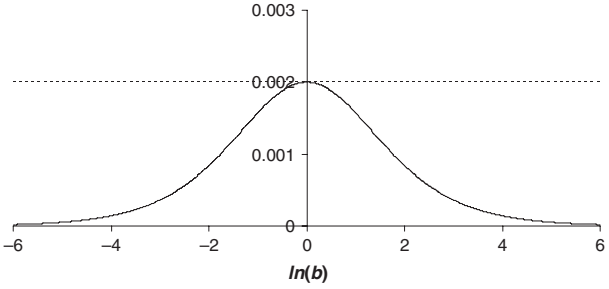


Fig. 6 Probability of fixation for a mutation with selective advantage $s = 0.001$. The dashed line represents the probability with a constant population size, while the continuous line represents the model used in the text. In this model, the population size is changing every generation between two population sizes, N_1 and N_2, with $N_2 = bN_1$.

the sweep occurs in the middle of the bottleneck, the discrimination power is greater. However, for every investigated case, the tests failed to locate the true target of selection.

Machine-learning algorithms

Machine learning focuses on automatic techniques for learning to make accurate predictions based on past observations (Schapire 2003). The machine-learning process can be started from either available empirical data or simulated data. For instance, if the aim is to learn how to discriminate the signatures of selection from the signatures of a bottleneck, one can use two sets of training samples generated under a selection and a bottleneck model, respectively. Pavlidis *et al.* (2010) and Lin *et al.* (2011) used machine-learning algorithms to maximize the predictive performance of combinations of summary statistics. For example, Lin *et al.* (2011) used a recent machine-learning approach called boosting (see Schapire 2003; Bühlmann & Hothorn 2007) to classify population genetics models based on a set of summary statistics of the SFS (Watterson's θ_W, Tajima's θ_π and Fay and Wu's θ_H and two combinations of those; Tajima's D and Fay and Wu's H) and summary statistics of linkage disequilibrium (integrated extended haplotype homozygosity, iHH, Voight *et al.* 2006). Simulations showed that boosting increases the accuracy of the prediction and decreases the number of false positives when the training and testing parameters are similar but classification was more difficult when they differed. Summary statistics reflecting linkage disequilibrium were informative to detect recent sweeps, whereas θ_π was more useful for detecting older sweeps. Overall, the most informative summary statistics to distinguish between bottlenecks and selection was Watterson's θ_W. Thus, boosting might be useful to find the best combination of summary statistics for approximate Bayesian computation (ABC) analyses. In Lin *et al.*'s (2011), boosting was employed for classification but a similar approach could also be used in a regression framework to estimate the intensity of selection.

Likelihood models

Various likelihood models have been developed that aim to infer selection and population demography simultaneously. Some likelihood methods consider different classes of site variants with the assumption that variants in some classes are a priori more likely to be selectively neutral than variants in other classes. For instance, one can contrast polymorphisms at synonymous and nonsynonymous sites. Williamson *et al.* (2005) developed a maximum-likelihood framework for inferring selection and demography based on this principle. The method assumes that noncoding SNPs are selectively neutral and uses these polymorphisms to test for demographic departures from the standard neutral model. The estimated parameters are used in the next step to correct for the effects of demography when testing for selection at synonymous, nonsynonymous

and insertion/deletion polymorphisms. The method is based on the SFS and is an extension of the Poisson random field (PRF) approach first introduced by Sawyer & Hartl (1992) to the case where the population changes size. Being based on the PRF, the method assumes no linkage among sites, no interference among mutations and an infinitely-many sites mutation model. The method further requires inferring the ancestral state of each site, something that can be problematic in many cases. Simulations under a variety of demographic models suggest that this two-step approach of estimating selection is robust to many violations of the demographic assumptions. As pointed by the authors, the core of the approach is not so much the estimation of demography; rather, the core is the comparison of the site frequency spectra of the two functional classes in the context of a reasonable demographic model.

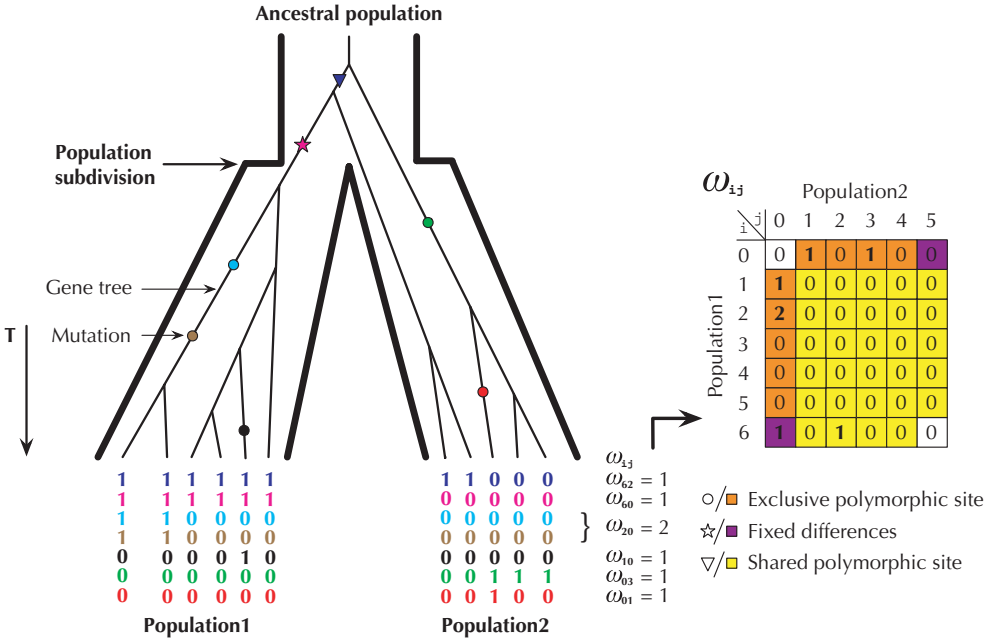
The analysis of demography and selection proposed by Li & Stephan (2006) is also a two-step likelihood approach. However, it builds on the premise that demographic changes affect the genome-wide polymorphism pattern. Thus, in contrast to the previous method that a priori distinguishes neutral and putatively selected sites, demography is here estimated from the whole data set and all SNPs are initially assumed to evolve neutrally. The model considers multiple populations and is based on the joint mutation frequency spectrum (see Box 4). The data analysed in the study by Li & Stephan (2006) consist of two populations of *Drosophila melanogaster*, one African and one European. Considering the joint frequency spectrum increases the power of the demographic analysis. Once the best parameters have been estimated for the chosen demographic model, a likelihood ratio test is used to compare a model incorporating demography and selection (alternative hypothesis)—in this case a hitchhiking model—with the model incorporating demography only (null hypothesis). A sliding window analysis is then carried out to identify the chromosomal regions affected by hitchhiking. Assuming that all windows are independent of one another, a rejection sampling procedure is used to estimate the rate of adaptive substitution (δ) and the distribution of substitutions with selection coefficient s, $f(s)$. In this last step, the fact that the European population is derived from the African population is taken into account, namely that sweeps can have occurred either before or after the PS. The method developed by Nielsen *et al.* (2009) differs in some details but is similar in spirit to the method of Li & Stephan (2006). There too, the whole data are first used to estimate demographic parameters and neutrality is tested as a departure from that demographic scenario. The main innovations are the consideration of population admixture, the use of a new approach to infer the joint demographic history of multiple populations based on the joint frequency spectrum (Gutenkunst *et al.* 2009) and the development of summary statistics of the joint frequency spectrum (called the 2D-SFS as two human populations European and African Americans are considered in the study). Nielsen *et al.* (2009)'s G2D test measures the fit of a SNP frequency spectrum and number of fixed differences between populations of individual loci compared with the overall pattern in the genome. It should therefore pick any deviation from the pattern generated by demography and not only those because of selection.

Approximate Bayesian computation

Approximate Bayesian computation (Tavaré *et al.* 1997; Pritchard *et al.* 1999, Beaumont *et al.* 2002) has been successfully used to infer a variety of demographic parameters for a wide range of species (see e.g. Bertorelle *et al.* 2010 and Csilléry *et al.* 2010). Briefly, the ABC framework uses Bayes' theorem and generates a posterior distribution of some parameter of interest from combining prior belief about the parameter and information from some (potentially genetic) data. The ABC approach relies on three basic steps: drawing parameter values from a prior distribution, simulating data under an explicit model and retaining the parameter values that generated simulated data sufficiently similar to the observed data (using a rejection algorithm). For most population genetic problems, the full data are too large or too complex to be efficiently evaluated using Monte Carlo simulation. Because ABC relies on summaries of the data instead of considering the full data, it is well adapted to make inferences from complex biological and population genetic data. However, ABC has rarely been used to infer properties of selection (but see Jensen *et al.* 2008), despite that ABC is likely to be a powerful approach for addressing many features of selection at the same time as demographic events can be modelled and thereby incorporated in the estimation procedure. In order to apply ABC to infer some properties of selection, such as the rate of selection, the strength of selection, finding genes targeted by selection or determining the fraction of the genome affected by selection, we need summary statistics that are particularly sensitive to the genetic patterns created by, for example, selective sweeps. Furthermore, we need flexible modelling tools that can simulate both demographic change and selection, and these methods and software are now becoming available (SFS_code, Hernandez 2008; msms, Ewing & Hermisson 2010; mbs, Teshima & Innan 2009; Jensen *et al.* 2007). Jensen *et al.* (2008) used simulations to investigate the rate and strength of selective sweeps for realistic choices of

Box 4 Joint mutation frequency spectrum

The joint mutation frequency spectrum (joint MFS) describes the joint distribution of polymorphisms across populations and is the equivalent of the site frequency spectrum (SFS) for just one population and is playing an important part in the development of new methods to isolate signatures of selection from those of demographic processes in structured populations (Li & Stephan 2006; Gutenkunst *et al.* 2009; Lukić *et al.* 2011). The figure below shows the joint MFS for two populations derived from the same ancestral population. Thick lines represent population boundaries, and thin lines trace the ancestral lineages of samples. The coloured circles, star and triangle stand for mutations. Here, seven mutations are considered, and the 0s and 1s under the genealogy represent ancestral and derived alleles, respectively



ω_{ij}	Population2					
Population1	0	1	2	3	4	5
0	0	**1**	0	**1**	0	0
1	**1**	0	0	0	0	0
2	**2**	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	**1**	0	**1**	0	0	0

Generally, the matrix describing the JMFS, ω_{ij}, is of dimension $n_1 \times n_2$, where n_1 and n_2 are the sample sizes of populations 1 and 2, respectively. ω_{ij} is the number of derived mutations carried by i sampled chromosomes in the sample from population 1 and by j sampled chromosomes in the sample from population 2. Taking the cyan mutation as an example, two chromosomes in population 1 carry this mutation, but none in population 2. There are two mutations (the other is brown) exhibiting such a configuration, so we note $\omega_{20} = 2$. The values of ω_{00} and ω_{n1n2} (denoting the numbers of mutations that are not present and fixed in the sample, respectively) are not considered in the matrix

This matrix can be summarized by statistics introduced by Wakeley & Hey (1997): the number of exclusive polymorphic site which means polymorphisms specific to the samples from populations 1 and 2 (orange region); the number of sites fixed in either sample (purple region); and the number of shared sites between two samples (yellow region). The definition used here is slightly different from the initial definition, but similar to that used by Becquet & Przeworski (2007). Here, we consider the mutation represented by a triangle shape (the highest one in the tree) as a shared polymorphic site rather than an exclusive polymorphic site because we assume that we know the ancestral allele state. These summary statistics can be used to estimate ancestral population parameters, such as population size and the time of split (Wakeley & Hey 1997)

mutation rates and recombination rates (from both Drosophila and humans). They further developed an ABC estimation procedure to infer both the rate and the strength of selective sweeps. In short, their ABC procedure built on using prior distributions for the rate and the strength of selective sweeps, generating simulated data under a recurrent selective sweep model (Jensen *et al.* 2007) and choosing a set of summary statistics

that were informative about selective sweeps. Another use of ABC that has yet to be developed is for searching the genome for signs of positive selection. In principle, ABC allows a flexible choice of demographic models, and using summary statistics that are informative about genomic patterns of selective sweeps, it may be possible to find outlier genes that deviate from the spectrum of neutral regions. For example, since (strong) selection (both directional and balancing), impact the properties of genealogies (Kaplan *et al.* 1988; Hudson & Kaplan 1988; see also Box 2), potentially even more than different demographic scenarios (see e.g. Blum & Jakobsson 2011, for an example from humans), summary statistics that captures information about the genealogy might be useful for detecting selection in an ABC framework.

Unbalanced trees and sample size

Recently, a new statistical test based on the maximum frequency of derived mutations (MFDM) has been proven analytically and empirically free from the confounding impact of varying population size, and the MFDM has also been shown to have high power to detect recent positive selection (Li 2011). Compared with the commonly used 'neutrality' tests (Tajima 1989; Fu & Li 1993; Fay & Wu 2000; Zeng *et al.* 2006), which are based on the SFS, the MFDM test uses an approach based on the tree topology that is not affected by varying population size in a random mating population (Hudson 1990). The principle of this test is to use the maximum frequency of derived mutation in the sample to detect the presence of an unbalanced tree (Fig. 7). Once the unbalanced tree is inferred at the locus, it implies that a nearby locus may have experienced recent positive selection. The MFDM test is not only free from the confounding effects of varying population size, including bottlenecks and population expansion, but also robust to population subdivision and admixture after excluding migration by a phylogenetic method and a simple sampling scheme (Li 2011). The MFDM test has a strikingly low false-positive rate, while its power to detect positive selection is still reasonably high. Its power can, specifically, be much higher than the Fay and Wu's H tests (Fay & Wu 2000) in an expanding population (Li 2011) such as humans. Notably, this has been achieved by analysing the DNA polymorphism of a single locus (i.e. a short piece of DNA fragment, say a few hundred base pairs). However, to achieve these properties, the MFDM test needs a relatively large sample size (the minimum sample size required is 41 chromosomes or 21 diploid individuals). The MFDM test is a single-locus test, but a multilocus extension could, in principle, be developed given that a reliable map of recombination events along the genome is available.

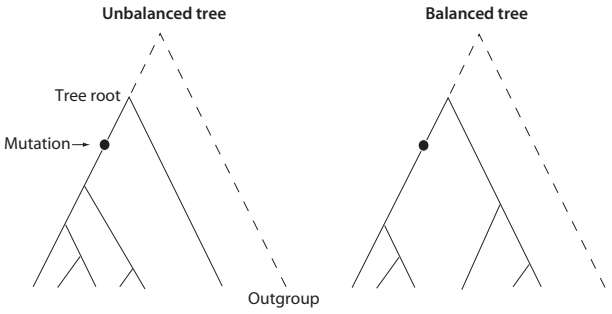


Fig. 7 Unbalanced and balanced trees. According to coalescent theory (Wakeley 2008), varying population size does not affect tree topology in a single Wright–Fisher population. The probability of an unbalanced tree is usually very small under neutrality, as the number of possible trees grows very quickly. However, under a hitchhiking model, this probability will increase substantially when the neutral locus is partially linked to the selected locus.

Discussion

We have reviewed recent efforts to obtain joint estimates of selection and demography from sequence polymorphism data (Table 1). Our interest in the question was initially triggered by Hahn's article 'Toward a selection theory of molecular evolution' (2008) and, in particular, by his challenge to the widely used two-step approach to detect selection where the average pattern of variation across the genome is assumed to reflect demography and where the parts of the genome under selection are those departing from this average pattern. Hahn's rationale was that if hitchhiking is all pervading, as seems to be the case in species with very large effective population sizes such as Drosophila, then no site can safely be assumed to evolve solely under drift and mutation. It remains to be seen whether this is the case in species with smaller effective population sizes, but even so, what our review suggests is that in the search for genome regions under selection, there are many reasons to worry about the confounding effect of demographics, especially if genome scans are used. One such reason is that selective and demographic events are likely to be associated throughout the evolution of species and will also leave similar signatures across the genome. Another reason is that the details of the particular demographic model that is assumed have a large impact on the inference of genes targeted by selection (Box 1; Kim & Gulisija 2010). If an incorrect model is assumed or inferred, estimates of selection are in turn going to be wrong. What can be done?

A few major conclusions emerge from our survey of the current literature. First, the 'two-step approach' still remains the main strategy to estimate demography and

Table 1 Summary of the methods presented in the paper whose aim is to jointly estimate selection and demography or estimate selection while controlling for demographic effects

Methods	Strength	Weakness	References
Combining summary statistics	Ease of use	Sensitive to both demography and selection	Grossman *et al.* (2010)
Machine-learning algorithms	Decrease in the number of false positive	Same as above	Pavlidis *et al.* (2010) Lin *et al.* (2011)
Likelihood models	Optimal use of the data. Closest approach to a true joint analysis of demography and selection	Limited to simple models	Williamson *et al.* (2005) Li & Stephan (2006) Nielsen *et al.* (2009)
Approximate Bayesian computation	Easy to implement and can consider realistic models	Approximate method	Tavaré *et al.* (1997) Pritchard *et al.* (1999) Beaumont *et al.* (2002)
Unbalanced tree	Low sensitivity to demography	So far limited to completed sweeps and selection on standing variation with lowfrequency	Li (2011)

selection, although new methods are emerging that may ultimately lead to a simultaneous estimation of demography and selection. These new methods make use of the latest developments in statistics (Pavlidis *et al.* 2010; Lin *et al.* 2011), explore aspects of the data that had been neglected hitherto (Li 2011) or take advantage of the emergence of population genomic data (Gottipati *et al.* 2011; Hernandez *et al.* 2011; Sattah *et al.* 2011). None of these methods truly constitutes a joint estimation method of demography and selection. In the meantime, the most realistic path of action seems to first use a likelihood or ABC approach to estimate the distribution of demographic and genetic parameters and then look for outliers. At that stage, one may also try to obtain an independent estimate of selection on individual loci, for instance, by using the MFDM test. If the number of loci detected by both approaches is large and recombination is limited, one knows at least that the demographic estimates might be questionable. If, on the other hand, the number of loci detected by both approaches is limited, one may reasonably assume that selection either directly or through hitchhiking has not been powerful enough to affect estimates of demographic parameters. Also, in species where recent selection is not too strong, it will help to consider sites distant from genic regions when assessing demography as a consistent decrease of polymorphism is observed in the latter (Gottipati *et al.* 2011; Luca *et al.* 2011).

Second, there is much to gain by including as much information as possible on the biology of the species, including its past history. As we have argued earlier, selection events are likely to be associated with major demographic transitions. The Hernandez *et al.*'s (2011) is particularly instructive in this respect as it shows that even with extensive genomic data at hand, detecting the signature of selection remains a difficult task, at least in organisms such as humans with a limited effective population size. One of the main conclusions of this study is that 'in search for targets of human adaptation, a change of focus is warranted'. Until recently, most studies of selection in the human genomes have focused on selective sweeps. However, those appear rare and there are also good biological reasons to expect them to be so (Pritchard *et al.* 2010). Instead, Hernandez *et al.* (2011) believe that what is needed are 'new tests to detect other modes of selection, such as comparisons between closely related populations that have adapted to drastically different environments or methods that consider loci that contribute to the same phenotype jointly'. We certainly agree that an approach more rooted in the ecology of the species under study than current genome scans would be the way forward when searching for genes under selection, as well as an explicit consideration of the factors that led to the selection pressure (the environments) and its target (the phenotype). It remains to be seen whether this will lead to the discovery of large amount of positive selection or whether natural selection has actually played a limited role in the recent evolution of species like humans, but it definitely offers a promising alternative to available methods.

Recombination, which we have avoided discussing at any length, is also an important factor when attempting to separate genetic signatures of demography from selection. The underlying rationale for outliers in genomic scans representing selected sites is that demography will affect the whole genome—equally—while selection only affects parts of it, but this is only true if there is recombination. Recombination is in fact a key parameter in many of the methods for detecting selection. Some approaches deal with recombination directly

by explicitly incorporating prior estimates of the recombination landscape using genetic maps, while others avoid the influence of recombination by picking sufficiently spaced markers (e.g. in the PRF approach, Sawyer & Hartl 1992). Finally, the recombination process in itself can mimic selection by biased gene conversion, a process where certain aspects of a sequence influence its probability of being chosen as the template in a gene conversion tract (e.g. Nagylaki 1983; Berglund *et al.* 2009).

Conclusion

Our review of the literature suggests that there is not yet any obvious answer to Wakeley's (2010) question about the fate of the sophisticated coalescent machinery for making inferences about the demographic history of populations. Inferential tools robust to the presence of natural selection, especially for species in which selection is a dominant force, still seem in their infancy, and the contribution of positive selection to shaping species still remains a controversial issue. Similarly, the hope that population genomic data would greatly facilitate the resolution of the neutralist–selectionist dispute has not yet materialized and the first attempts of using genomic data alone have not done the trick, in the same way as they did not allow us to overcome Fisher's 'infinitesimal curse' for quantitative traits (Visscher *et al.* 2010). At least, we should be pardoned to think that we are now better equipped than ever to move away from silliness (Gillespie 2004) and start grappling with those fundamental evolutionary issues in all their complexity.

Acknowledgements

We thank Matthew Hahn, Sylvain Glémin and three anonymous reviewers for comments on the manuscript. JL and HL are supported by the 973 project (No.2012CB316505) and the National Natural Science Foundation of China (No.31172073). ML would like to acknowledge support from the Chinese Academy of Sciences (visiting professorship), the EU (Noveltree project) and the Swedish Research Council. MJ is supported by the Swedish Research Council Formas.

References

Andolfatto P, Wong KM, Bachtrog D (2011) Effective population size and the efficacy of selection on the X Chromosomes of two closely related Drosophila species. *Genome Biology and Evolution*, **3**, 114–128.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in population genetics. *Genetics*, **162**, 2025–2035.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Begun DJ, Holloway AK, Stephens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biology*, **7**, e26.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Blum MGB, Jakobsson M (2011) Deep divergences of human gene trees and models of human origins. *Molecular Biology and Evolution*, **28**, 889–898.

Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, **467**, 587–590.

Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW (2010) Characterization of X-linked SNP genotypic variation in globally distributed human populations. *Genome Biology*, **11**, R10.

Colosimo PF, Peichel CL, Nereng K *et al.* (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology*, **2**, e0635.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, **25**, 410–418.

Ellegren H (2009) A selection model of molecular evolution incorporating the effective population size. *Evolution*, **63**, 301–305.

Emery LS, Felsenstein J, Akey JM (2010) Estimators of the human effective sex ratio detect sex biases on different timescales. *American Journal of Human Genetics*, **87**, 848–856.

Ewens WJ (1967) The probability of survival of a new mutant in a fluctuating environment. *Heredity*, **22**, 438–443.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review in Ecology, Evolution, and Systematics*, **40**, 481–501.

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

Frankham R, Jones LP, Barker JSF (1968a) Effects of population size and selection for a quantitative character in Drosophila. 1. Short-term response to selection. *Genetical Research*, **12**, 237–248.

Frankham R, Jones LP, Barker JSF (1968b) Effects of population size and selection for a quantitative character in Drosophila. 3. Analysis of lines. *Genetical Research*, **12**, 267–277.

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

Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*, Vol. **41**, 432 pp, Monographs in Population Biology. Princeton University Press, Princeton, NJ.

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

Gillespie JH (2001) Is the population size of a species relevant to its evolution? *Evolution*, **55**, 2161–2169.

Gillespie JH (2004) Why k=4Nus is silly. In: *The Evolution of Population Biology* (eds Singh RS, Uyenoyama MK), pp. 178–192. Cambridge University Press, Cambridge UK.

Golding GB (1997) The effect of purifying selection on genealogies. In: *Progress in Population Genetics and Human Evolution*, **Vol. 87**. IMA volumes in mathematics and its applications (eds Donnelly P, Tavare S), pp. 271–285. Springer Verlag, New York.

Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature Genetics*, **43**, 741–743.

Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **344**, 403–410.

Grossman SR, Shylakhter I, Karlsson EK *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

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

Haldane JBS (1927) A mathematical theory of natural and artificial selection. V. Selection and mutation. *Proceedings of the Cambridge Philosophical Society*, **23**, 838–844.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics*, **42**, 830–831.

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **23**, 2786–2787.

Hernandez RD, Kelley JL, Elyashiv E *et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science*, **331**, 920–924.

Hudson RR (1990) Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds Futuyma D, Antonovics J), pp. 1–44. Oxford University Press, New York.

Hudson RR, Kaplan NL (1986) On the divergence of alleles in nested subsamples from finite populations. *Genetics*, **113**(4), 1057–1076.

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Jensen JD, Bachtrog D (2011) Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin Domain. *Genome Biology and Evolution*, **3**, 687–701.

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

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

Johansson AM, Petersson ME, Siegel PB, Carlborg Ö (2010) Genome-wide effects of long-term divergent selection. *PLoS Genetics*, **6**, e1001188.

Jones LP, Frankham R, Barker JSF (1968) Effects of population size and selection for a quantitative character in Drosophila. 2. Long-term response to selection. *Genetical Research*, **12**, 249–266.

Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. *Genetics*, **120**, 819–829.

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

Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genetics*, **41**, 66–70.

Kern AD, Jones CD, Begun DJ (2002) Genomic effects of nucleotide substitutions in *Drosophila simulans*. *Genetics*, **162**, 1753–1761.

Kim Y, Gulisija D (2010) Signatures of recent directional selection under different models of population expansion during colonization of new selective environments. *Genetics*, **184**, 571–585.

Kimura M (1957) Some problems of stochastic processes in genetics. *Annals of Mathematical Statistics*, **28**, 882–901.

Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–719.

Krone SK, Neuhauser C (1997) Ancestral processes with selection. *Theoretical Population Biology*, **51**, 210–237.

Krone S, Nordborg M (2002) Separation of time scales and convergence to the coalescent in structured populations. In: *Modern Developments in Theoretical Population Genetics* (eds Slatkin M, Veuille M), pp. 194–232. Oxford University Press, Oxford, UK.

Lambert CH, Connelly CF, Macdeoy J, Qiu R, Olson MV, Akey JM (2010) Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. *American Journal of Human Genetics*, **86**, 34–44.

Li H (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution*, **28**, 365–375.

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

Li YF, Costello JC, Holloway AK, Hahn MW (2008) ''Reverse ecology'' and the power of population genomics. *Evolution*, **62**, 2984–2994.

Lin K, Li H, Schlötterer C, Futschik A (2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, **187**, 229–244.

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and application to human evolution. *Genome Research*, **21**, 1087–1098.

Lukić S, Hey J, Chen K (2011) Non-equilibrium allele frequency spectra via spectral methods. *Theoretical Population Biology*, **79**, 203–219.

Mellars P (2006) Why did modern humans populations disperse from Africa ca. 60,000 years ago? *Proceedings of the National Academy of Sciences, USA*, **103**, 9381–9386.

Nagylaki T (1983) Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences, USA*, **80**, 6278–6281.

Nielsen R, Hubisz MJ, Hellmann I *et al.* (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, **19**, 838–849.

Nordborg M (2001) Coalescent theory. In: *Handbook of Statistical Genetics* (eds Balding DJ, Bishop MJ, Cannings C), pp. 179–212. John Wiley & Sons, Inc., Chichester, UK.

Orr HA (2009) Fitness and its role in evolutionary genetics. *Nature Reviews in Genetics*, **10**, 531–539.

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

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

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

Przeworski M, Charlesworth B, Wall J (1999) Genealogies and weak purifying selection. *Molecular Biology and Evolution*, **16**(2), 246–252.

Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.

Sattah S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genetics*, **7**, e1001302.

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

Schapire RE (2003) The boosting approach to machine learning: an overview. In: *Nonlinear Estimation and Classification* (eds Denison DD, Hansen MH, Holmes C, Mallick B, Yu B), pp. 149–172. Springer, New York.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? *PLoS Genetics*, **5**, e1000495.

Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist*, **188**, 313–332.

Sjödin P, Krone S, Kaj I, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. *Genetics*, **169**, 1061–1070.

Spencer CCA, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, **20**, 3673–3675.

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

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

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Teshima K, Innan H (2009) mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, **10**, 166.

Thornton KR, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics*, **172**, 1607–1619.

Visscher PM, McEvoy B, Yang J (2010) From Galton to GWAS: quantitative genetics o human height. *Genetics Research, Camb*, **92**, 371–379.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences, USA*, **102**, 18508–18513.

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

Wakeley J (2008) *Coalescent Theory: An Introduction*. Roberts and Company, Greenwood Village, Colorado.

Wakeley J (2010) Natural selection and coalescent theory. In: *Evolution since Darwin: The First 150 Years* (eds Bell MA, Futuyma DJ, Eanes WF, Levinton JS), pp. 119–149. Sinauer and Associates, Sunderland, Massachusetts.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Williamson SH, Hernandez R, Fiedel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences, USA*, **102**, 7882–7887.

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

Zeng K, Charlesworth B (2010) Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *Journal of Molecular Evolution*, **70**, 116–128.

Zeng K, Fu Y-X, Shi S, Wu C-I (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.

The authors of the paper are broadly interested in population genetics and evolution theory. Their experimental work covers a large range of organisms, including humans, Drosophila, forest trees and model weeds. Some are primarily interested in demographic inferences while others are more eager to find signatures of selection in the genome. All are faced with the fact that both selection and demography affect genetic variation.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Simulation programs that include demography and selection.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.