MOLECULAR ECOLOGY

Molecular Ecology (2015) doi: 10.1111/mec.13390

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

The consequences of not accounting for background selection in demographic inference

GREGORY B. EWING*† and JEFFREY D. JENSEN*†

*Ecole Polytechnique Fédérale de Lausanne (EPFL), EPFL SV IBI-SV UPJENSEN, AAB 0 46, Station 15, CH 1015, Lausanne, Switzerland, †Swiss Institute of Bioinformatics (SIB), EPFL SV IBI-SV UPJENSEN, AAB 0 46, Station 15, CH 1015, Lausanne, Switzerland

Abstract

Recently, there has been increased awareness of the role of background selection (BGS) in both data analysis and modelling advances. However, BGS is still difficult to take into account because of tractability issues with simulations and difficulty with nonequilibrium demographic models. Often, simple rescaling adjustments of effective population size are used. However, there has been neither a proper characterization of how BGS could bias or shift inference when not properly taken into account, nor a thorough analysis of whether rescaling is a sufficient solution. Here, we carry out extensive simulations with BGS to determine biases and behaviour of demographic inference using an approximate Bayesian approach. We find that results can be positively misleading with significant bias, and describe the parameter space in which BGS models replicate observed neutral nonequilibrium expectations.

Keywords: evolutionary theory, natural selection and contemporary evolution, population dynamics, population genetics—theoretical

Received 7 May 2015; revision received 5 August 2015; accepted 25 August 2015

Introduction

With the wide availability of large, extensive data sets, it is common to estimate parameters from complex nonequilibrium demographic scenarios (Li & Stephan 2006; Gutenkunst *et al.* 2009; Excoffier & Foll 2011; Excoffier *et al.* 2013). Demographic parameters are often of direct interest in the study concerned; however, they may also act as nuisance parameters when required for accurate tests of positive selection or in the detection of selective sweeps with acceptable false discovery rates (Jensen *et al.* 2005, 2007; Thornton & Jensen 2007). Such requirements exist because expected patterns of diversity can be similar for selection and demographic models, the canonical example being a selective sweep and population bottleneck (Nielsen *et al.* 2005; Prezeworski *et al.* 2005). Both models are characterized by similar reduction in diversity and shifts in the site frequency spectrum (SFS), and distinguishing the presence of either has received significant attention (Thornton & Andolfatto 2006).

Background selection (BGS) (Charlesworth *et al.* 1993; Charlesworth 1994) is also widely accepted to have a significant effect on patterns of genetic diversity and the efficiency of selection via linkage; despite this, it is rarely taken into account in current studies due to the practical difficulties of doing so. Intermediate levels of BGS pose a difficult problem in modelling, at least with coalescent simulators, despite the relative ease of simulating BGS in a forward simulation framework (Hernandez 2008; Messer 2013). While strongly deleterious mutations are purged from the population immediately, and very slightly deleterious mutations behave neutrally, intermediately deleterious mutations accumulate according to non-neutral dynamics that are difficult to characterize in nonequilibrium demographic models.

Some studies have corrected for BGS with a rescaling of effective population size (N_e) (e.g. Prüfer *et al.* 2012). This is appealing because it is easy to apply

Correspondence: Gregory B. Ewing, E-mail: gregory.ewing@epfl.ch

without adjusting existing methods and is theoretically motivated. And it is likely valid when the effects of BGS are strong, as every individual that receives a deleterious mutation is effectively removed from the population. Thus, in each generation, an approximately fixed number of individuals are removed from the population resulting in an effective reduction in N_e (Charlesworth *et al.* 1993). However, recent studies have shown that BGS may often be more intermediate in magnitude (Bank *et al.* 2014; Comeron 2014).

Therefore, it is currently unknown whether BGS causes significant problems in inference and if so, for which parameters and at what magnitudes. In this study, we consider the consequences of ignoring BGS in demographic inference under nontrivial demographic models. Briefly, forward simulations are used to generate data, which, while quite slow, do not require many iterations. Inference is then carried out with fast coalescent simulations in a simple ABC setting. Although we consider ABC primarily for the point of illustration, we expect these results to be relevant to all inference methods based on the site frequency spectrum.

Methods

All simulations and inference were carried out with a two-population model with a single founding ancestral population as shown in Fig. 1. In all cases, we normalize the population size to the ancestral population size and set $N_e = 1000$, and 20 chromosomes are sampled from each population. We assume symmetric migration and distinct population sizes and growth parameters. In each data set, we simulate 1000 independent loci, using a within-locus recombination rate of $2N_e\rho = 10$ and $4N_e\mu = 100$. Here, we define a locus as an independent observation of the coalescent process; that is, we assume no linkage between loci, although linkage and recombination are included within a locus. BGS is included as a simple single selection coefficient and a probability parameter that any mutation results in a selected mutation. In all cases, we assume that BGS is negatively selected and that selection is additive with multiple mutations. True parameters of all simulated data are shown in Table 1; however, only representative results are shown here for conciseness. All forward simulations were carried out with SFSCODE (Hernandez 2008) and SLIM (Messer 2013) with $10N_e$ generations to reach equilibrium. Neutral data sets were also compared to coalescent simulations as a check (ms and msms) (Hudson 2002; Ewing & Hermisson 2010).

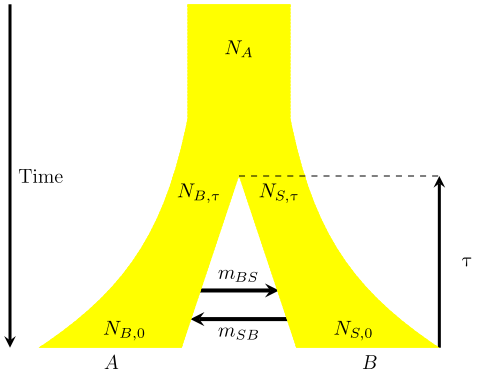

Fig. 1 Basic population model, including all estimated parameters. All simulated data use symmetric migration, and all population sizes are relative to the ancestral size. Growth is parameterized based on start and end population sizes. τ is fixed at $2N_e$ generations.

All inference was carried out with pure rejection ABC using binned joint site frequency spectrum (JSFS) as the summary statistics. The binning has previously shown to be informative (Naduvilezhath *et al.* 2011) where the mean and standard deviation across the 1000 loci are used for each bin. These statistics simplify scaling, and such classic population genetic statistics such as F_{st} are indeed summary statistics of the JSFS. In cases where we used a Poisson composite likelihood metric, the standard deviation was ignored. The tolerance of the ABC was set such that we kept the best 5000 simulations of 10 million using either a Euclidean distance metric or a Poisson composite likelihood metric (Excoffier *et al.* 2013). Priors are flat and reported in Table 1. Although other inference methods are available, such as *dadi* or *fastsimcoal2*, pure rejection ABC permits much faster inference once the initial set of simulations for ABC is complete, and only the rejection step needs to be repeated for each data set. This also facilitates simpler interpretation of the results in the context of comparison; it is noted that we are not demonstrating inference methods, but rather the general misinference that is possible regardless of methodology. Nevertheless, inference with *fastsimcoal2* was carried out for a few example data sets for comparison, and the same general trends were observed. Furthermore, we carried out regression post-processing (Beaumont *et al.* 2002) for some ABC results and found this made no significant qualitative difference and is therefore not presented here.

In all inference, we condition on the number of SNPs we observe in the data and do not directly estimate $\theta \propto N_e\mu$, but rather estimate parameters relative to the ancestral population. This avoids needless simulations with widely different SNP counts that would always be rejected, and reflects practical inference for real data more accurately. A further feature that is used in inference while conditioning on the number of SNPs is weighted mutations. In the coalescent simulations for

Table 1 All parameter values used for simulations, and priors where applicable for ABC estimation

Parameter	Values							Prior min	Prior max
$N_{A,0} = N_{A,\tau}$	0.5	1	2	5	10			0.1	100
$N_{B,0} = N_{B,\tau}$	0.5	1	2	5	10			0.1	100
$N_{A,\tau} = N_{B,\tau}$	0.5							0.01	100
$m_{A\to B} = m_{B\to A}$	0	0.5	1	10	100			0	500
γ	0	1	2	5	10	20	100		
p_d	0.01	0.05	0.1	0.2					
τ	2								

ABC, rather than generating a Poisson number of neutral SNPs on each branch of a coalescent tree, we can report the expected mean mutation count that can be directly included in the JSFS. This avoids simulation noise arising from the Poisson process, while the coalescent noise is still present (that is why we sample 1000 such trees with 1000 loci). Such a feature is included in *fastsimcoal2* when carrying out inference. *Dadi*, using the diffusion approximation, also avoids this noise in deriving the expected JSFS for a given model and parameters. We also carried out, as a check, ABC with normal mutation processes and note only slightly wider posterior confidence intervals.

Results

Figure 2 shows the inferred marginal posterior distribution for population size estimates of A and B, when population A's true size is 1 and population B's is 5, for different $N_e s$ parameters. The posterior densities are quite wide compared to the priors in this inference despite such a large simulation size. However, as we are comparing inferences with model mis-specification, this does not affect our conclusions. Using the mode as an estimate, when Ns is small we see no deviation from the neutral case for estimates of both populations A and B, as expected. However, with intermediate levels of BGS, there is a significant departure from the true value. With high values of BGS ($Ns > 100$), we again estimate population size accurately. This is expected as with strong BGS all mutations are effectively lethal and result in a reduction in effective population size, that is a thinning process. Recall that population size estimates are relative to the ancestral population size and both sizes are reduced equally by BGS.

Other parameters show similar trends of misinference with intermediate BGS values. Figure 3 illustrates the level of misinference with varying strength of selection ($N_e s$) and the probability of such mutations. Low selection coefficients combined with a high probability of such selected mutations show the same general trends as intermediate selection coefficients. This is due to multiple, linked, negatively selected sites on the same locus and is equivalent to a higher combined effect of BGS.

It should be stressed that we are not observing a simple rescaling of various parameters, as we are estimating parameters relative to the ancestral population; the ancestral population size is always some effective $\hat{N}_e$ that is not directly estimated. Also note that the ratio between populations A and B is also misinferred, providing further evidence that BGS effects are not a simple rescaling of N_e.

In Fig. 4, we compare the global site frequency spectrum between the estimated parameters and the true parameters. The global SFS is presented rather than the JSFS for conciseness, illustrating the same results. The global SFS was generated with independent simulations from the inferred parameters (mode estimates) and is not the posterior SFS. Close agreement is found between the means for the intermediate-BGS case ($Ns = 10$) despite the large bias in inferred parameters. Further, comparison with the Poisson CL metric shows similar fit. However, the standard deviations appear to mismatch somewhat and typically are higher in the BGS data compared to that simulated without BGS. To further investigate, we carried out ABC inference using a Poisson composite likelihood metric and ignored the standard deviations. The results are shown in Fig. 4; the resulting SFS is similar, and parameters are estimated with similar bias and we conclude that the results are insensitive to the choice of metric. This shows that we can have a case of positively misleading inference where we infer parameters that are not a good reflection of actuality, yet appear to explain our data well.

An interesting observation is that using a Poisson CL metric does not affect the quality of the fit nor the bias. The good SFS fit is not surprising as we are now directly attempting to match the expected JSFS with the observed JSFS, yet we note the mismatch in variance; it is perhaps surprising that the Euclidean metrics, with twice as many statistics, do not perform better. This could be due to the curse of dimensionality: as more summary statistics are used, it is less likely that a random sampling of space will be close to the data, giving wider confidence intervals and larger bias with more

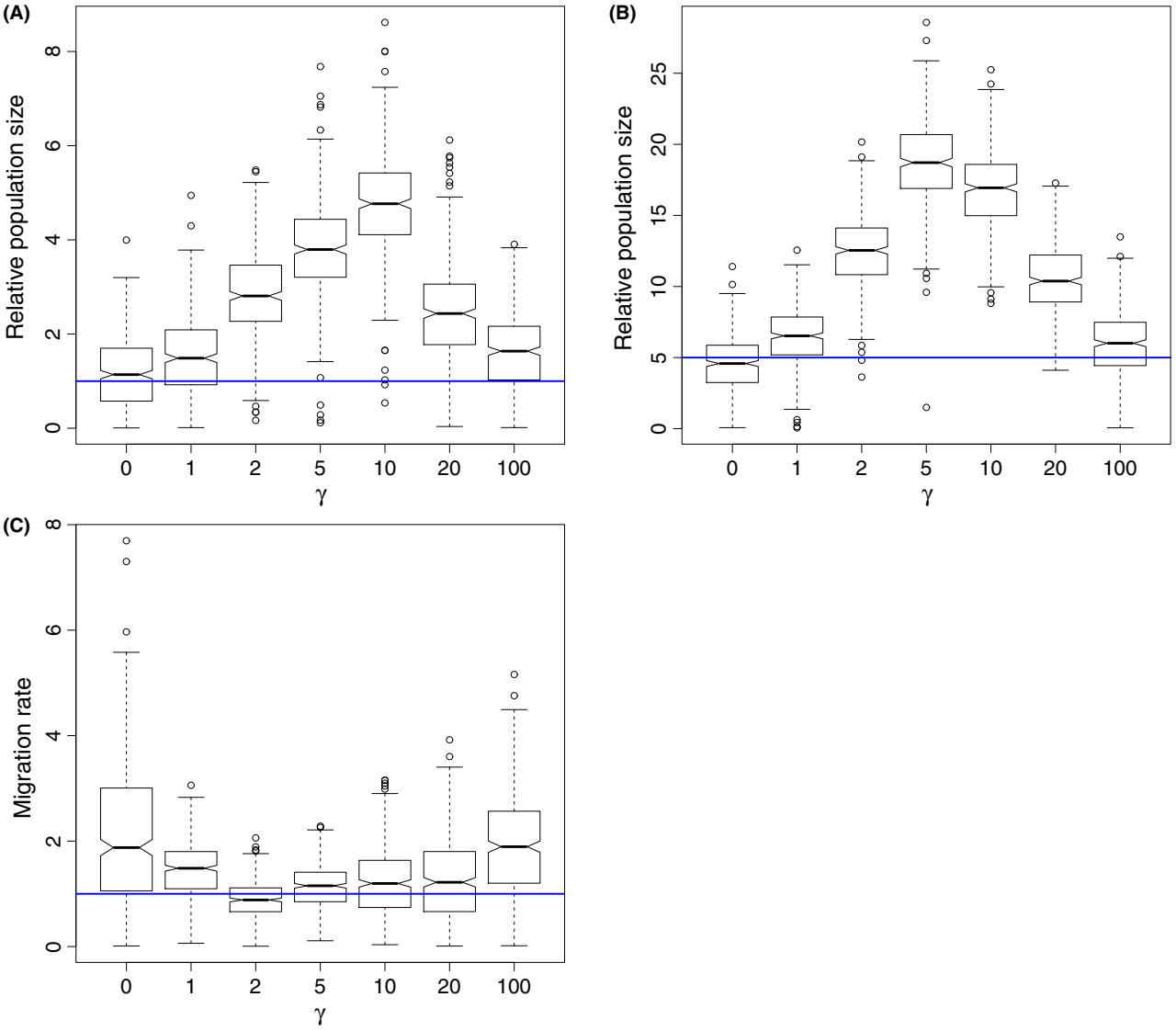

Fig. 2 Bias in inference with different levels of background selection (BGS). (A) Population A, (B) population B and (C) migration. The solid blue line indicates the true value, while the box plots show the posterior density from the ABC inference. As shown, for intermediate levels of background selection, growth models are strongly overestimated. We note that migration estimation appears to improve with increasing BGS.

statistics. Alternatively, the Poisson nature of the metric may give more weight to more important statistics. Finally, as noted above, the standard deviations did not match well, so including them may well be futile. The simulated data, while excluding BGS, simply cannot fit as well when considering standard deviations in the summary statistics; including both means and standard deviations confers a larger bias.

The larger expected standard deviation may be understood when we consider the distribution of ancestral recombination graphs (ARG). The observed negatively selected mutations are more likely to be young than old, and thus, internal branches will tend to be shorter relative to the neutral case. Thus, as we go back through time in a linage, we are less likely to have a parent with a deleterious mutation the further back we go, as such a linage has a low probability of surviving until sampling time. Because the pool of likely parents is thus effectively reduced further back in time, the coalescent rate increases, or in other words, N_e is reduced the further back in time we proceed. Note the qualitative agreement in Fig. 1 where the derived population size estimates are increased relative to the ancestral population size and that exponential growth is always overestimated with BGS. However, unlike with population growth, or other demographic effects, this process is driven by the

random Poisson distribution of mutations in addition to the stochastic coalescent process, thereby increasing observed variance compared to the purely demographic case.

Discussion

We have shown that with intermediate levels of BGS present, positively misleading demographic inference is possible when BGS is not taken into account, while with both strong and weak BGS, accurate inference is possible. This result is noteworthy, given the common assumption in demographic inference that all SNPs are selectively neutral and unaffected by linkage to a selected site. Examining a wide range of parameters, we also demonstrate that in many cases, the expected SFS between neutral demographic models and BGS models is confounded. The direction of bias is well explained as a reduction in N_e further past-ward, resulting in overestimation of population growth and size in the present relative to the past. However, qualitative expectations are difficult to derive in nonequilibrium populations.

The most notable difference between BGS and non-BGS models identified here was the expected standard deviation of site frequency spectra. In all cases, BGS increased the standard deviation of SFS means across loci, which can be understood when considering the additional Poisson process that now affects the resulting ARG. This also suggests possible ways to investigate whether a given data set has important levels of BGS, as we expect smaller effective N_e past-ward and a larger standard deviation of expected site frequency spectra. Unfortunately, both increasing population size and BGS are reasonable expectations for many data sets, and thus, information on population size history would be insufficient. It is tempting to consider the standard deviation, but for real data many processes such as recombination hot spots, selective sweeps, mutation variation and ascertainment bias can all be expected to increase the SFS variance compared to simulated demographic models. Indeed, the SFS fits shown in Fig. 4 are considerably better than those typically associated with real data sets.

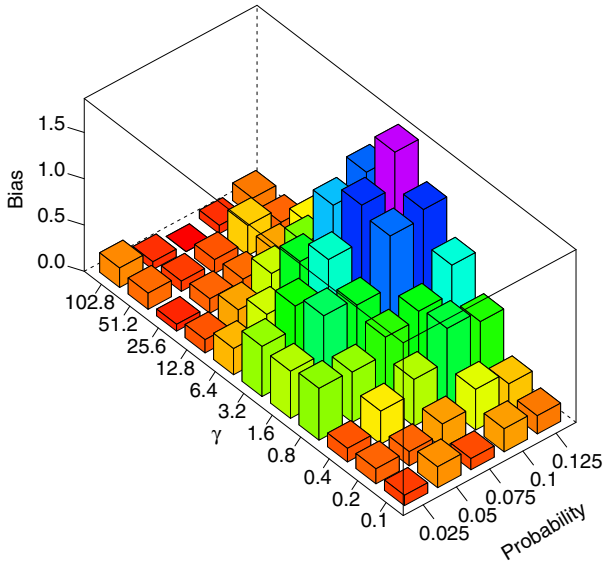

Fig. 3 Median estimator bias of relative population size estimation with different background selection (BGS) parameters. The true population size is 5. The *Z*-axis is the absolute mean bias, the *X*-axis is the probability that a mutation is negatively selected, and the *Y*-axis is the strength of selection in units of $2N_es$. Bar colours denote bias magnitude. Intermediate BGS has a large bias, while both strong and weak BGS do not.

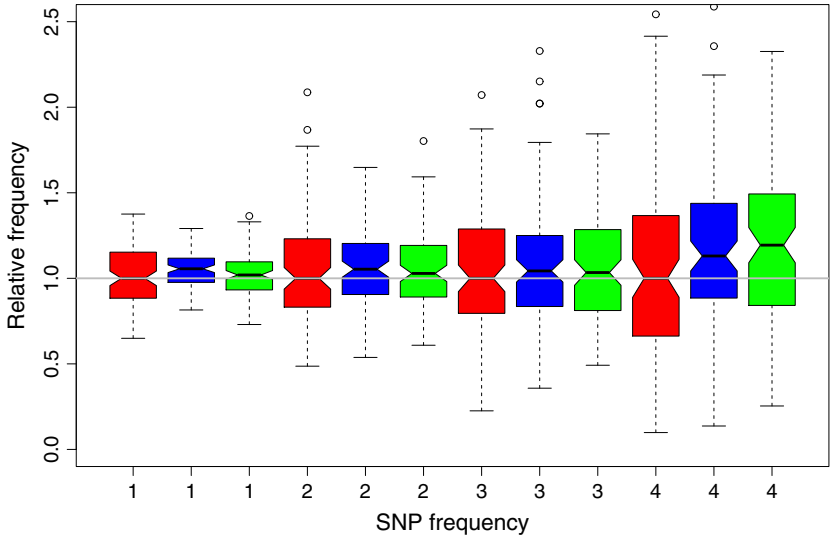

Fig. 4 Site frequency spectrum of background selection (BGS) data compared with a fitted demographic model. We normalize all frequencies to the BGS medians for each category. Red: BGS data. Blue: estimated data using the Euclidean metric. Green: estimated data using the Poisson CL metric. In both the cases, the best simulation was used to regenerate data rather than the posterior. We note the apparently larger variance under the BGS model. In all cases, $\gamma = 10$ and the probability of a selected mutation is 0.1.

The results presented here indicate that previously analysed data sets may in fact be misinferred. Given the wide range of previously estimated selection parameters and recent BGS estimates (Charlesworth 2012; Comeron 2014), it is reasonable to assume that intermediate rates of BGS do occur, leading to important deviations from neutrally demographic model assumptions. Recent experimental results regarding the distribution of fitness effects further support the notion that some mutations will be within the ranges tested here (e.g. Bank *et al.* 2014).

Separating BGS from demography therefore remains a challenging problem. While forward simulations of BGS are straightforward and suggest the possibility of co-estimation of BGS and demography within an ABC framework as a possible solution, problems remain. Although available computing resources have increased dramatically, forward simulators remain slow for general inference methods. For example, the ABC simulations shown here took only days to weeks to complete with a coalescent simulator and available cluster resources, compared to an estimated computation time of almost 1 year with SFSCode. Progress in this area is ongoing. For example, SLiM computes much faster than SFSCode by only tracking segregating sites rather than entire genomes. For rough approximations, coalescent simulators have been implemented (Zeng & Charlesworth 2011; Zeng 2013), and work in this area continues towards relaxing some of these approximations. Further development is also needed to identify sufficient statistics for the co-estimation of BGS and demography.

Acknowledgements

This work was funded by grants from the Swiss National Science Foundation, and a European Research Council (ERC) Starting Grant to JDJ. We would like to thank Vital-IT for providing cluster resources, and Kristen Irwin for a careful reading of the manuscript.

References

Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD (2014) A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. *Genetics*, **196**, 841–852.

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

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, **63**, 213–227.

Charlesworth B (2012) The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila × chromosome. *Genetics*, **191**, 233–246.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Comeron JM (2014) Background selection as baseline for nucleotide variation across the Drosophila genome. *PLoS Genetics*, **10**, e1004434.

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 (2011) Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, **27**, 1332–1334.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, doi:10.1371/journal.pgen.1003905.

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.

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

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

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

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

Messer PW (2013) SLiM: simulating evolution with selection and linkage. *Genetics*, **194**, 1037–1039.

Naduvilezhath L, Rose LE, Metzler D (2011) Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology*, **20**, 2709–2723.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Prezeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. *Evolution*, **59**, 2312–2323.

Prüfer K, Munch K, Hellmann I *et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature*, **486**, 527–531.

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.

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

Zeng K (2013) A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity*, **110**, 363–371.

Zeng K, Charlesworth B (2011) The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, **189**, 251–266.

J.D.J. and G.B.E. conceived of the study and authored the manuscript; G.B.E. carried all the simulations and wrote the required code; J.D.J. and G.B.E. analyzed and interpreted the results.

Data accessibility

All required scripts and protocols, together with example ABC data: Dryad doi:10.5061/dryad.37hc1.