Molecular Ecology (2015) doi: 10.1111/mec.13478

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

Inferring the age of a fixed beneficial allele

LOUISE ORMOND,*† MATTHIEU FOLL,*† ‡ GREGORY B. EWING,*† SUSANNE P. PFEIFER*† and JEFFREY D. JENSEN*†

**School of Life Sciences, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, †Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, ‡International Agency for Research on Cancer (IARC), Lyon, France*

Abstract

Estimating the age and strength of beneficial alleles is central to understanding how adaptation proceeds in response to changing environmental conditions. Several haplotype-based estimators exist for inferring the age of segregating beneficial mutations. Here, we develop an approximate Bayesian-based approach that rather estimates these parameters for fixed beneficial mutations in single populations. We integrate a range of existing diversity, site frequency spectrum, haplotype- and linkage disequilibrium-based summary statistics. We show that for strong selective sweeps on de novo mutations the method can estimate allele age and selection strength even in nonequilibrium demographic scenarios. We extend our approach to models of selection on standing variation, and co-infer the frequency at which selection began to act upon the mutation. Finally, we apply our method to estimate the age and selection strength of a previously identified mutation underpinning cryptic colour adaptation in a wild deer mouse population, and compare our findings with previously published estimates as well as with geological data pertaining to the presumed shift in selective pressure.

Keywords: adaptation, ecological genetics, population genetics – empirical, population genetics – theoretical

Received 8 May 2015; revision received 14 October 2015; accepted 9 November 2015

Introduction

Selective sweeps are believed to have played a role in shaping genomic patterns of variation across a wide range of species. Estimating the parameters underlying this process, including the beneficial allele age and associated selection strength, can provide deeper insights into the mode and tempo of adaptation. With regard to allele age in particular, one question that has remained of particular focus is whether specifically identified beneficial mutations correspond with the timing of an environmental change experienced by the population in question – be it the colonization of a novel habitat or a sudden geological event. This question is often posed in the context of whether adaptive events more commonly draw on new or standing genetic variation – and indeed, significant debate remains around this topic (Jensen 2014). Adaptation from new mutations may be said to be 'mutation limited', in that the appropriate mutation would need to occur after the shift in selective pressure. Thus, the ability to accurately infer the age and the starting frequency at the onset of selection of identified beneficial mutations relative to known environmental shifts will be key for advancing this debate.

Many tests have been designed to identify the action of selection in the genome from patterns of polymorphism (see review of Thornton *et al.* (2007); Bank *et al.* (2014)). These rely on frequency changes in linked neutral sites induced by a selective sweep, a process known as 'genetic hitchhiking' (Kaplan *et al.* 1989). Polymorphism-based signals are relatively fleeting and are typically visible only on a timescale of 0.1 N_e generations or less, for an effective population size N_e (Przeworski 2003). Yet the majority of approaches are intended to only identify beneficial fixations, and comparatively few approaches exist for inferring the age of these variants. Over the last few years, method development has largely focused on time-sampled data sets, and much progress has been made in this area (e.g. McVean 2002;

Correspondence: Jeffrey D. Jensen,
E-mail: jeffrey.jensen@epfl.ch

Malaspinas *et al.* 2012; Mathieson & McVean 2013; Foll *et al.* 2014; Steinrücken *et al.* 2014). However, apart from experimentally evolved or clinical populations, or the handful of ancient genomes, the great majority of available data is collected at a single time point (i.e. present), and there is thus a compelling incentive to improve single time point methods.

Despite the fast transit time characterizing beneficial fixations, the majority of single time point methods to date have aimed to estimate these parameters for segregating, rather than fixed, beneficial mutations using haplotype structure (e.g. Slatkin 2008; Peter *et al.* 2012; Chen & Slatkin 2013; Chen *et al.* 2015). Most recently, Chen *et al.* (2015) used a hidden Markov model to explore haplotype structure and developed a likelihood estimation approach assuming strong selection (and thus a deterministic allele trajectory) for currently segregating beneficial mutations. For fixed mutations, the state-of-the-art approach was proposed by Przeworski (2003) to estimate the age of a fixed beneficial mutation in an approximate Bayesian (ABC) framework based on a combination of diversity, site frequency spectrum (SFS) and haplotype statistics. We continue this focus to develop an improved estimator for the age of fixed beneficial mutations using the past decade of statistical method development, and utilize the Przeworski (2003) estimator as a performance benchmark.

Most notably, the characteristic pattern of linkage disequilibrium (LD) generated by a complete selective sweep suggests the opportunity to utilize this in an ABC framework. Simulation and theoretical studies (e.g. Stephan *et al.* 2006; Jensen *et al.* 2007; McVean 2007; Pavlidis *et al.* 2010) have described strong LD at linked sites on either side of the beneficial fixation, but not spanning the selected site. In addition, there is a reduction in LD across the target of selection. Kim & Nielsen (2004) designed a statistic ω_{max} that captures this complex pattern, with Jensen *et al.* (2007) subsequently demonstrating that ω_{max} exhibits different density distributions under selective sweep models in both equilibrium and nonequilibrium populations.

Here, we explore the combination of frequency spectrum- and linkage disequilibrium-based expectations as an approach to improve our ability to estimate the age of a fixed beneficial mutation based on observed patterns of polymorphism. We develop an ABC-based method that is demonstrated to outperform existing approaches. This approach is not intended to identify loci under selection from genomewide scans: rather, it is applicable to previously identified loci. We extend this approach to co-estimate allele age and selection strength assuming that selection acts on a de novo mutation. Next, we relax the assumption of selection on a de novo mutation to co-estimate the starting frequency of the segregating allele with allele age and selection strength. Finally, we apply these developed methodologies to explore the selective history of cryptic coloration in a wild deer mouse population, and compare our newly developed estimates with previous published inference.

Methods

We present three sets of methods. First, we infer allele age T (the time since the allele fixed) alone assuming that the selection coefficient s is known and that a model of selection on de novo mutations applies. Secondly, s and T are co-estimated while continuing to assume a model of selection from de novo mutation. Thirdly, we co-infer the starting frequency f at which the previously neutral allele was segregating in the population at the onset of selection s and the age at which selection begins T_s. Although the underlying assumption is that a test of selection has been applied using other tools, we demonstrate that this approach has power to correctly infer neutrality as well.

Approximate Bayesian Computation (ABC)

A standard ABC approach was applied following Tavaré *et al.* (1997) and Beaumont *et al.* (2002). We used the R package *abc* (Csillery *et al.* 2012) and implemented the method in the following series of steps:

Simulations. For each scenario considered, 5×10^5 simulations were generated using the program MSMS (Ewing & Hermisson 2010). Briefly, neutral genealogies are traced backwards in time for a random sample of alleles using standard coalescent theory, incorporating recombination and demographic changes where applicable. Selection is modelled at a single predetermined locus by applying forward simulations. In this study, selection is assumed to be additive such that genotypes that are homozygous and heterozygous for the selected derived allele have fitness $1 + s$ and $1 + s/2$ respectively, whereas genotypes that are homozygous for the ancestral allele have fitness 1. For a sample of n chromosomes of length L, and assuming an effective diploid population size $N = 10\,000$, a coalescent history was constructed assuming a population-scaled mutation rate $\theta = 4N_eL\mu$ with mutation rate $\mu = 10^{-7}$ per base pair per generation, and a population-scaled recombination rate $\rho = 4N_eLr$ with recombination rate $r = 10^{-7}$ per base pair per generation. Unless a model of selection from standing variation is stipulated, simulations are designed to model selective sweeps from de novo mutations arising on a single chromosome in the population, which have ultimately fixed. The strength of selective sweeps is determined using the population-scaled parameter $\alpha = 2N_es$. Unless specified otherwise, simulations were run using $L = 20$ kb for the inference of T alone and using $L = 10$ kb for the co-estimation of s and T, and the selected mutation is positioned in the centre of the region. These lengths were chosen to capture the full signature of the selective sweep for the parameters used, based on theoretical results demonstrating an effect over $L = 0.01 \times s/r = 10$ kb for a selective sweep of coefficient $s = 0.1$ and recombination rate $r = 10^{-7}$ crossovers per base pair per generation (Kaplan *et al*. 1989).

Equilibrium populations are modelled as panmictic diploid populations of constant size $N_e = 10\,000$. Allele age T is taken to be the time since the allele fixed, using the –SF option in MSMS simulations. For equilibrium demographic scenarios, the prior distributions for s and T were $\log_{10}(s) \sim U(-4, -0.5)$ and $\log_{10}(T) \sim U(-4, -0.5)$ where U is a uniform distribution. T is reported in units of $4N_e$ generations in keeping with standard coalescent theory. These distributions were chosen in order to span different orders of magnitude from neutrality (where $N_es \leq 1$) to strong selection ($s = 0.3$), and from very recent to distant ages of the selected allele. Przeworski (2002) have shown that $T = 0.1 \times 4N_e$ is approximately the upper limit for detecting selective sweeps, after which the signature in polymorphism data becomes rapidly obscured by subsequent mutation, recombination, and genetic drift.

Choice of summary statistics. We used the program MSSTATS (Thornton 2003) to calculate a panel of 21 frequently used summary statistics (see Table S1, Supporting information for details of statistics) from the standard MSMS single nucleotide polymorphism (SNP) output simulated in step 1 above. Figure S1 (Supporting information) shows the correlation of a range of informative diversity, SFS- and LD-based statistics with s for recent sweeps ($T = 0.01 \times 4N_e$ generations). For older sweeps ($T = 0.1 \times 4N_e$ generations), we find that the signature of selection becomes rapidly obscured (data not shown). Following Wegmann *et al.* (2009), we employ a partial least squares method (PLS) to incorporate the most informative statistics into our method. PLS is similar to principal component analysis, but determines orthogonal components from a high dimensional set of statistics by maximizing the covariance between the statistics and the variables. Applying PLS has been shown to improve the performance of ABC methods, partly by reducing the dimensionality of the set of summary statistics and partly by removing noise from uninformative statistics (Joyce & Marjoram 2008). Wegmann *et al.* (2009) have shown that incorporating a large number of noninformative summary statistics may bias the resulting posteriors. The *pls* package in R (Bjorn-Helge & Wehrens 2007) was used to calculate PLS components based on a subset of size 10^4 out of the total 5×10^5 simulations. Prior to implementing PLS, we apply a Box–Cox transformation (Box & Cox 1964) to normalize the statistics. We adapted a script available through ABC TOOLBOX for this purpose (Wegmann *et al.* 2010). Incorporating PLS into our ABC method was shown to reduce relative bias and root mean square error (RMSE), and was therefore used in all ABC calculations.

ABC inference of s and T. To evaluate the performance of our ABC method, we selected values of T only, or of s and T over different orders of magnitude, and ran 100 simulations for each selected pair of values that we considered as pseudo-observables. Summary statistics were calculated from the SNP output using *msstats* and transformed into PLS components using the same loadings as in step 2 above.

Posterior distributions for the parameters were generated using an ABC rejection algorithm and a tolerance level of 0.005, which was found to be optimal. A total of 2500 simulations were therefore retained out of the total number of simulations of 5×10^5. Using local linear or ridge regression ABC methods did not significantly improve results (data not shown).

Point estimates for s and T were calculated from the mode of the joint density posterior distribution using the two-dimensional kernel density function in the MASS package in R (Venables & Ripley 2002). For estimating allele age alone, the mode of the posterior distribution for T was calculated to give a point estimate.

Relative bias and RMSEs were calculated between these predicted values and the true pseudo-observable values for s and T (Tables S2 and S3, Supporting information). Relative bias is defined as the mean difference between the predicted value y and the true pseudo-observable y_t divided by the value of the true pseudo-observable y_t. RMSE is defined as the square root of the squared difference between the predicted value y and the true pseudo-observable y_t divided by the number of observations n:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_1^n (y - y_t)^2}{n}}.$$

Nonequilibrium demographic scenarios

For nonequilibrium populations, allele age is taken to be the time since the onset of selection (T_s) by applying the –SI option in MSMS. It is not possible in the current version of MSMS (or in other simulation programs) to model the time since the allele fixed T under changing demographic parameters, but only to model the time T_s since the onset of selection. To ensure that the selected allele fixes in simulations, the –SFC option is used to prevent loss owing to genetic drift, and the –oTrace switch is applied to track the frequency of the selected allele in the population through time using a python script. Only simulations where the frequency of the selected mutation is above 0.99 at the time of sampling are retained.

The demographic models are assumed to have been inferred using other methods (e.g. δaδi (Gutenkunst *et al.* 2009), fastsimcoal (Excoffier *et al.* 2013)), and are incorporated in the simulations in step 1 above to run 5×10^5 simulations. The selection coefficient s is drawn from a log uniform prior as for equilibrium scenarios: $\log_{10}(s) \sim \mathrm{U}(-4, -0.5)$, and the prior for allele age T_s is adjusted to account for the allele's sojourn time and to ensure that the selected mutation has sufficient time to fix. Based on the analytical derivation of the sojourn time T_{soj} provided by Stephan *et al.* (1992)

$$T_{soj} = \frac{2\ln(2N_e)}{s},$$

we adjust the prior for T_s to $\log_{10}(T) \sim \mathrm{U}(\log_{10}(T_{soj}), \log_{10}(0.3 + T_{soj}))$. T_{soj} as calculated here represents the expected sojourn time under equilibrium demography and is therefore an approximation of the sojourn time under nonequilibrium demography.

Two scenarios were chosen to model size-change events. In both cases, bottlenecks are assumed to occur relatively recently at $0.01 \times 4N_e$ in the past. First, we model a shallow and long bottleneck of length $0.02 \times 4N_e$ with a 95% reduction in population size, and second, we model a narrow and severe bottleneck of length $0.002 \times 4N_e$ with a 99.8% reduction in population size. These parameters were chosen to be consistent with other studies (Pavlidis *et al.* 2010). In addition, a growth scenario was modelled assuming exponential growth following a bottleneck at $0.01 \times 4N_e$ pastward which reduced the population size to 1% of its current size, with a calculated $\alpha = 460.5$. This last scenario was chosen for its similarity to the demographic parameters inferred for *Peromyscus maniculatus* deer mice in the Nebraska Sand Hills for the data application presented here (Linnen *et al.* 2013).

Co-estimating allele starting frequency f

The previous sections assume a model of selection acting on a de novo mutation. In the third part of our method, we relax this assumption and extend our approach to co-infer the allele frequency f when selection begins, along with T_s and s. The same steps as for the joint inference of s and T_s in nonequilibrium scenarios described above were applied, but with the additional specification of f. The software *msms* allows for f to be input using the –SI switch. In simulated samples, s is drawn from a log uniform prior $\log_{10}(s) \sim \mathrm{U}(-4, -0.5)$, and the prior for T_s is adjusted to take account of sojourn time, to $\log_{10}(T_s) \sim \mathrm{U}(\log_{10}(T_{soj}), \log_{10}(0.3 + T_{soj}))$, to give the selected allele sufficient time to fix in the population, as before. The starting frequency f is drawn from a log uniform prior $\log_{10}(f) \sim \mathrm{U}(-4, -0.5)$ spanning the case of selection on a de novo mutation (with $N_e = 10^4$) to selection on a previously neutral segregating mutation with a frequency of 30%. Point estimates for s, T_s and f were calculated using the three-dimensional kernel density estimate of the joint posterior mode in the MISC3D package (Feng & Tierney 2015).

ω_{max}-ABC methodology

In addition to the MSSTATS-based ABC methodology described above, we also derived a methodology to incorporate the statistic ω_{max}. The same steps as for MSSTATS-ABC were implemented with the adjustments detailed in this section.

As described in the introduction, ω_{max} was designed by Kim & Nielsen (2004) to capture the specific LD pattern associated with selective sweeps, and in particular, the reduction in LD that occurs across the selected site after a sweep. The statistic ω is defined as

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1} \left(\sum_{i,j\epsilon L} r_{ij}^2 + \sum_{i,j\epsilon R} r_{ij}^2\right)}{(1/l(S-l)) \sum_{i\epsilon L, j\epsilon R} r_{ij}^2}.$$

At each site l of S polymorphic sites, the statistic splits sites into two groups, from the first to the l^{th} polymorphic site to the left, and from $(l + 1)^{\text{th}}$ to S polymorphic sites to the right. Within each group, singletons are excluded and the correlation coefficient r^2_{ij} is calculated between the i^{th} and j^{th} sites. The value of l that maximizes ω (ω_{max}) can also be obtained.

Under equilibrium demography, and assuming $T = 0.01 \times 4N_e$, simulations for different selection coefficients generate limited differences in distributions of ω_{max} overall, but do produce a skewed distribution for the top 5% values in selection scenarios compared to neutral simulations (Fig. S2, Supporting information). This observation holds over different sequence lengths ($L = 10^4$, 5×10^4 and 10^5 bps). This result is consistent with the findings of Jensen *et al.* (2007), who demonstrated via simulation that for large sample sizes ($n = 50$, as in our simulation study) in equilibrium populations, ω_{max} distributions are characterized by a tail of large values in selection scenarios, which increases with the size of selection coefficients.

To incorporate ω_{max} into an ABC framework, 100 simulations were generated in *msms* for each pair of values of s and T drawn from the priors, but only the top 5% by value of ω_{max} were retained; these were combined with the MSSTATS statistics calculated for those simulations. Taking the top 5% of simulations by value of ω_{max} for both the prior and for pseudo-observables replicates the ascertainment process (i.e. significant P-values). Not correcting for such ascertainment in multilocus genome scans has been shown to generate a high rate of false positives (Thornton & Jensen 2007). The approach of retaining the top 5% simulations by value of ω_{max} is consistent with the idea of an outlier approach where 100 loci are scanned, as done here, and only extreme values in the tails of distributions are retained as possible candidates for sites under selection.

Values of ω and ω_{max} for each simulation were calculated using OMEGAPLUS (Pavlidis *et al.* 2010).

Application to data on cryptic colour adaptation in deer mice

Data on 91 *Peromyscus maniculatus* deer mice were obtained from a previous study by Linnen *et al.* (2013). Briefly, mutations associated with traits underpinning cryptic colour adaptation to a light phenotype have been identified in mice living in the Nebraska Sand Hills. A serine deletion at position 128150 on exon 2 has been shown to be associated with several potentially adaptive traits, with a previously estimated selection coefficient of 0.126. Enrichment, sequencing and genotyping are described in Linnen *et al.* (2013). The sequence data were partitioned according to phenotype, and alleles with the serine deletion were extracted from the data set. The data were adjusted to cover a region of 20 kb on either side of the deletion. Of the 100 alleles with the serine deletion, 36 were discarded based on a threshold of more than 15% unknown sites. Of the remaining 64 alleles, all cases where the site was unknown for at least one individual were removed. The filtered data contained 418 segregating sites in the 40-kb region surrounding the deletion. We explored the impact of changing the filtering to 25% or 10% of individuals with more than 25% unknown sites, but this did not markedly change the results. We applied the ABC method described above for estimating s and T conditioning on the number of segregating sites S as well as θ and ρ. PLS was used to generate components to drive the inference procedure from the MSSTATS statistics after excluding S and any invariant statistics. We verified the accuracy of our ABC estimator using simulations with the mouse parameters. Point estimates for T and s were calculated from the mode of the joint density posterior distribution as before. A point estimate for T alone (assuming the previously published estimate of $s = 0.126$) was also derived for comparison purposes. We then explicitly incorporate into our simulations the previously inferred demographic scenario – a bottleneck 2900 years ago that reduced the population to 0.004 of its original size, followed by an exponential recovery to 65% of its original size (Linnen *et al.* 2013) and estimate s and the time of the onset of selection T_s. Finally, we co-estimate the starting frequency f of the serine deletion with s and T_s. We analysed the data over one additional length, 80 kb, with the selected mutation positioned centrally. We obtained a slightly lower sample sizes after applying the filtering process described above of 48 alleles for the 80 kb region. Simulations for the ABC calculation were run assuming $N_e = 53\,080$, a mutation rate $\mu = 3.62 \times 10^{-8}$ and a recombination rate $r = 0.62 \times 10^{-8}$ per base pair per generation (all assumptions are from Linnen *et al.* (2013).

Code for implementing the method is available through http://jensenlab.epfl.ch/.

Results

Inference of allele age (T) alone

Initially, we fixed the selection coefficient s to be 0.1 (strong selection), 0.01 (moderately strong selection) or 0.001 (weak selection) and we replicated previous results for the inference of allele age only using three statistics (the number of segregating sites S, Tajima's D (Tajima 1989) and the number of haplotypes), following Przeworski (2003). We sought to improve on these using the statistics available through MSSTATS (MSSTATS-ABC) and by incorporating the ω_{max} statistic (ω_{max} -ABC). All ages T are in units of $4Ne$ generations. The choice of whether ω_{max}–ABC or MSSTATS-ABC is used will depend on how the location of the mutation has been established, and therefore whether a method that corrects for ascertainment (ω_{max}–ABC) or one that does not correct for ascertainment (*msstats*-ABC) is appropriate (see Discussion). Figure 1A shows the results of inferring allele age T for 6 cases ($T = 0.001$, 0.01, 0.05, 0.1, 0.2 and 0.3) assuming strong selection ($s = 0.1$) and a sequence of 20 kb. Boxplots represent the distribution of point estimates, which are the modes of posterior distributions. Both MSSTATS-ABC and ω_{max}-ABC differentiate age well for 3 orders of magnitude, for $T = 0.001$, $T = 0.01$ and $T = 0.1$, and outperform the previously implemented summary statistics in (Przeworski 2003). Above $T = 0.05$, the age of sweeps is inferred with high accuracy. Relative bias and RMSE estimates support this conclusion (Table S2, Supporting information). The age of very young sweeps ($T = 0.001$) is underestimated, presumably because the signature of the selective sweeps is not yet apparent in all of the statistics utilized.

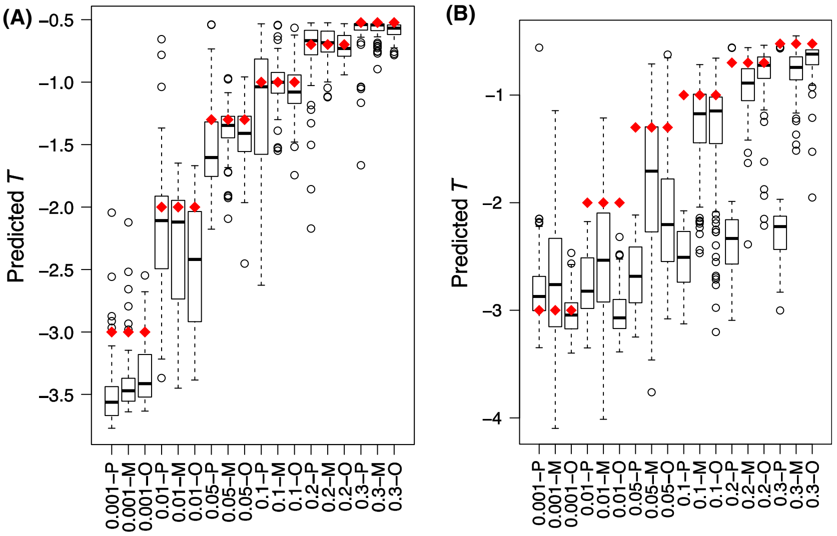

Fig. 1 Inference of allele age *T* alone. Boxplots compare results from msstats-ABC (marked M) and ω_{max} –ABC (marked O) with the Przeworski 2003 ABC method (marked P). Boxplots represent the modes of posterior distributions for inferring *T* alone for 100 pseudo-observables. The value of *s* is assumed to be known: (A) $s = 0.1$ ($\alpha = 2N_e s = 2000$) and (B) $s = 0.01$ ($\alpha = 2N_e s = 200$). *T* is drawn from a log uniform prior: $\log_{10}(T) \sim U(-4, -0.5)$. Other parameters are as described in methods, with $L = 20$ kb. Red diamonds indicate the true values for each case ($T = 0.001, 0.05, 0.01, 0.1, 0.2, 0.3$).

For moderate selection ($s = 0.01$), both *msstats*-ABC and ω_{max}-ABC differentiate *T* well between two orders of magnitude rather than three, effectively separating old sweeps ($T \geq 0.1$) from young sweeps ($T < 0.01$) (Fig. 1B). In this case, the estimators significantly improve performance over the statistics employed by Przeworski (2003). In contrast, for weak selection ($s = 0.001$), we find that the estimators perform poorly and identify all sweeps as very young (Fig. S3, Supporting information). Thus, the estimator for *T* alone works well only for strong and moderately strong selection.

We additionally explored the impact of choosing different window sizes surrounding a selected mutation ($L = 20$, 40 and 80 kb) (Fig. S4, Supporting information). We find that window sizes of 10 kb or 20 kb provide the best estimates for the parameter ranges investigated here, and that larger window sizes slightly underestimate allele age, due to a dilution of the statistics. This result is in line with theoretical results estimating the size of a swept region subject to a reduction of diversity as $L = 0.01 \times s/r$ (Kaplan *et al.* 1989).

Joint inference of s and T under equilibrium demography

Here, we extended our approach to jointly infer *s* and *T* for fixed mutations using a simple and computationally efficient approach. Simulations demonstrate that for young sweeps ($T = 0.01$) and old sweeps ($T = 0.1$), neutral scenarios (where $N_e s \leq 1$, i.e. $s = 0.0001$ and $s = 0$) can be readily differentiated from selection scenarios, for both young and old sweeps, using either MSSTATS-ABC (Figs 2C,D and S5, Supporting information) or ω_{max}–ABC (Figs S6 and S7, Supporting information). Additionally, we can infer strong and moderately strong selection ($s = 0.1$ and $s = 0.01$) well (Figs 2A,B and S5–S9, Supporting information) using either methodology. One of the weaknesses of both methods is that weak selection ($s = 0.001$), whether for old or young sweeps, can be misinferred as stronger, older selection ($s = 0.01$ or $s = 0.1$) (Figs 2C and S5C, Supporting information). This limitation owes to the fact that the patterns of polymorphism for weak sweeps resemble that of older, stronger sweeps. We find that ω_{max}–ABC is a more accurate estimator of weak selection than *msstats*-ABC (Table S3 and Fig. S9, Supporting information).

With regard to allele age, both methods are able to differentiate old sweeps ($T = 0.1$) from young sweeps ($T = 0.01$). We find that the methods do not have the power to accurately infer the age of young sweeps but only to establish whether sweeps are either $T = 0.01$ or younger. Results of inference for very young sweeps ($T = 0.001$) are similar to the results of inference for moderately young sweeps ($T = 0.01$) (data not shown). In contrast, for older sweeps, the additional time may enable different statistics to be impacted at different rates, and for a subset of these statistics to return towards equilibrium. Simulation studies have shown that statistics reliant on intermediate frequency alleles such as Fay and Wu's *H* (Fay & Wu 2000) decay rapidly after a selective sweep and retain very little signal at $0.1 \times 4N_e$, whereas statistics reliant on singletons such as Tajima's *D* retain a signal longer (Przeworski 2002). The decay of both of these types of statistics at different rates most likely underpins the accuracy of both estimators to infer *T* and *s* for sweeps of $T = 0.1$.

If a model of de novo mutation is assumed ($f = 1/2N_e$), a pseudo-observable of selection from a rare mutation ($f \leq 0.01$) will have an inferred allele age as the time at which selection starts to act on that mutation. In contrast, inference from a pseudo-observable of selection from high levels of standing variation ($f > 0.01$) will be erro-

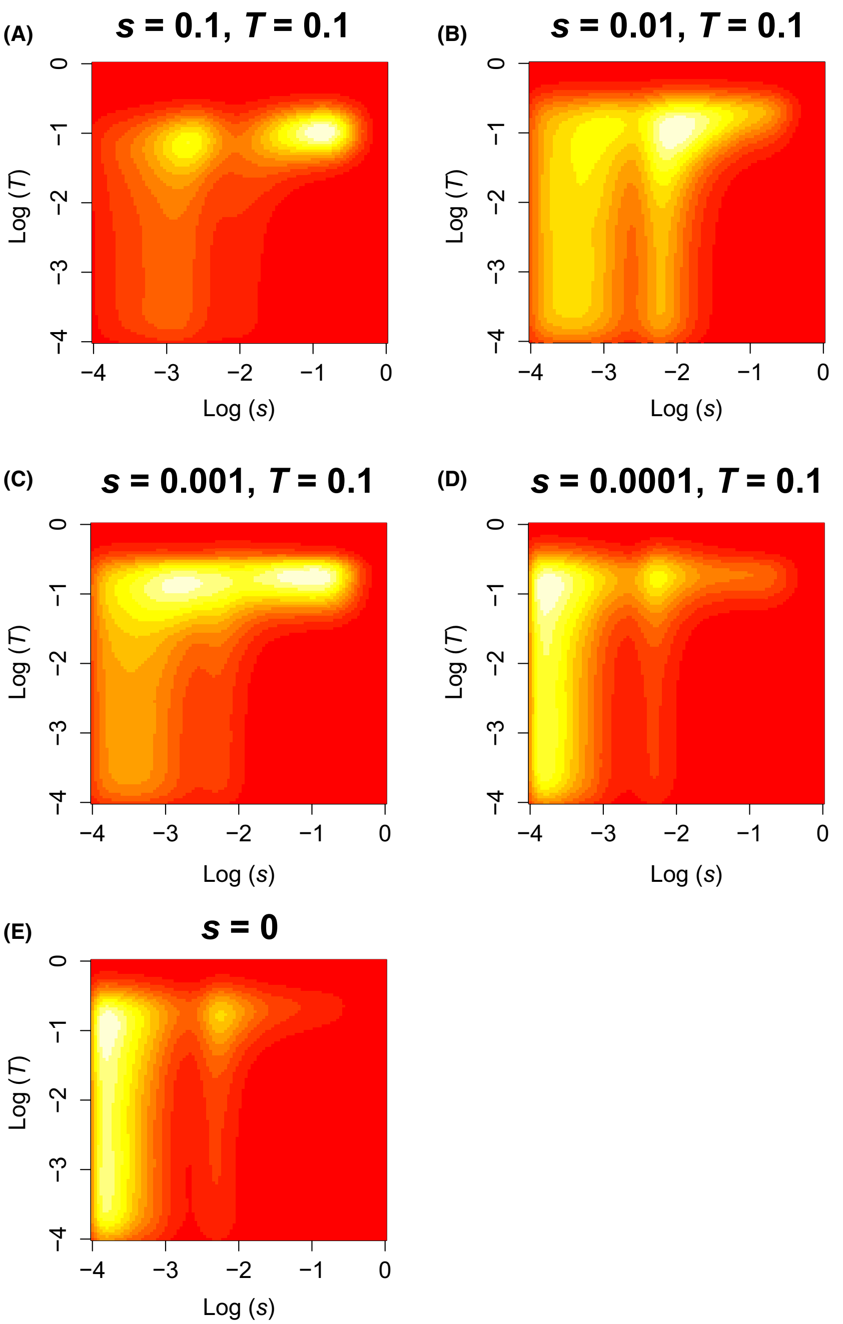

Fig. 2 Joint inference of s and T in equilibrium populations for old sweeps (T = 0.1) (MSSTATS-ABC). Figures show the cumulative joint posterior density plots for 100 pseudo-observable simulations over different orders of magnitude of the selection coefficient s, for old sweeps (T = 0.1) and (A) s = 0.1 ($\alpha = N_es = 10^3$); (B) s = 0.01($\alpha = N_es = 10^2$); (C) s = 0.001 ($\alpha = N_es$ = 10). The bottom two panels represent neutral scenarios with (D) s = 0.0001 ($\alpha = N_es$ = 1); and (E) s = 0. The white, yellow and red colours mark areas of high, moderate and low joint density respectively. Black crosses indicate the true values of pseudo-observables. s and T are drawn from log uniform priors: $\log_{10}(s) \sim U(-4, -0.5)$ and $\log_{10}(T) \sim U(-4, -0.5)$. Other parameters are as described in Methods.

neous and usually indicate an older age (data not shown).

Co-estimating allele starting frequency f under equilibrium demography

In this section, we relax the assumption that selection proceeds from de novo mutation, and allow the frequency of the selected allele to be co-inferred along with the time at which selection starts T_s (in contrast to the previous section, in which the time T since fixation is inferred), and the selection coefficient s. Analytical derivations by Stephan *et al.* (1992) predict that if $f < 1/2N_es$ and selection is strong, the reduction in linked neutral diversity associated with selection from rare mutations should resemble that from selection on de novo mutations. Subsequent analysis has shown that selection from either a de novo mutation or from a rare mutation results in a classical 'hard sweep' pattern where a single copy of the mutation is swept to fixation (Orr & Betancourt 2001; Hermisson & Pennings 2005). In contrast, selection from high levels of standing variation ($f \gg 1/2N_es$) results in the fixation of multiple haplotypes in a 'soft sweep' pattern (the other common definition of a soft sweep, in which haplotype diversity is the result of multiple beneficial, is not considered here). In line with theoretical expectations, simulations by Przeworski *et al.* (2005) showed similar patterns of reduction in diversity for $f = 1/2N_e$, $f = 0.001$ and $f = 0.01$ (where $N_e = 10^4$ and $s = 0.05$), but almost no reduction in diversity for selection from high levels of standing variation ($f = 0.05$ and $f = 0.20$). Here, we show results that are consistent with these previous findings. For strong and moderately strong selection ($s = 0.1$ and $s = 0.01$) in equilibrium populations, we find that the ABC estimator performs well for inferring f, s and T_s as long as the pseudo-observable satisfies the condition $f < 1/2N_es$, (i.e. the cases where $f = 1/2N_e$, $f = 0.001$, $f = 0.01$) (Figs 3 and S10, Supporting information). We find marginally better inference for $s = 0.1$ than for $s = 0.01$ right up to $f = 0.01$, which appears to be the cut-off for accurate inference. For weak selection ($s = 0.001$), T_s and s are well inferred but f is not (Fig. S10D, Supporting information).

In contrast, when the pseudo-observable sweep is characterized by $f \gg 1/N_es$, we generally identify that $f \gg 1/2N_es$ – but with the drawback that s and T_s are poorly co-estimated. As levels of standing variation increase, rising haplotype diversity means the method infers older, weaker sweeps than are the case for the pseudo-observables.

Robustness to nonequilibrium demography

Nonequilibrium demography can mimic signatures of selection (e.g. Przeworski 2002; Jensen *et al.* 2005) and compromise inference. However, an advantage of the ABC approach is that demographic parameters can be explicitly modelled in simulations, and therefore, demography can be taken into account in the inference method. We explored the robustness of the method for inferring first, s and T_s, and second s, T_s and f under three nonequilibrium scenarios where demographic parameters are explicitly known and modelled: (i) a shallow and long bottleneck of length 0.02 × $4N_e$ with a 95% reduction in population size, (ii) a narrow and severe bottleneck of length 0.002 with a 99.8% reduction in population size, and (iii) an exponential growth scenario following a sharp 99% reduction in population size with $\alpha = 460.5$. The bottlenecks are modelled to occur at $T = 0.01 \times 4N_e$. Our results are described for MSSTATS-ABC, but similar results were obtained for ω_{max}–ABC.

In the case of inferring s and T_s, Figs 4 and S13 (Supporting information) show the results for the third scenario of a bottleneck followed by exponential growth. We find that for old (Figs 4A and S11A, Supporting information), young (Figs 4B and S11B, Supporting information) and very young (Figs 4C and S11C, Supporting information) sweeps, both T_s and s are well inferred. We note that it is difficult to distinguish weak sweeps, where $s = 0.001$ (Fig. S11D,E, Supporting information), from neutral scenarios (Figs 4D

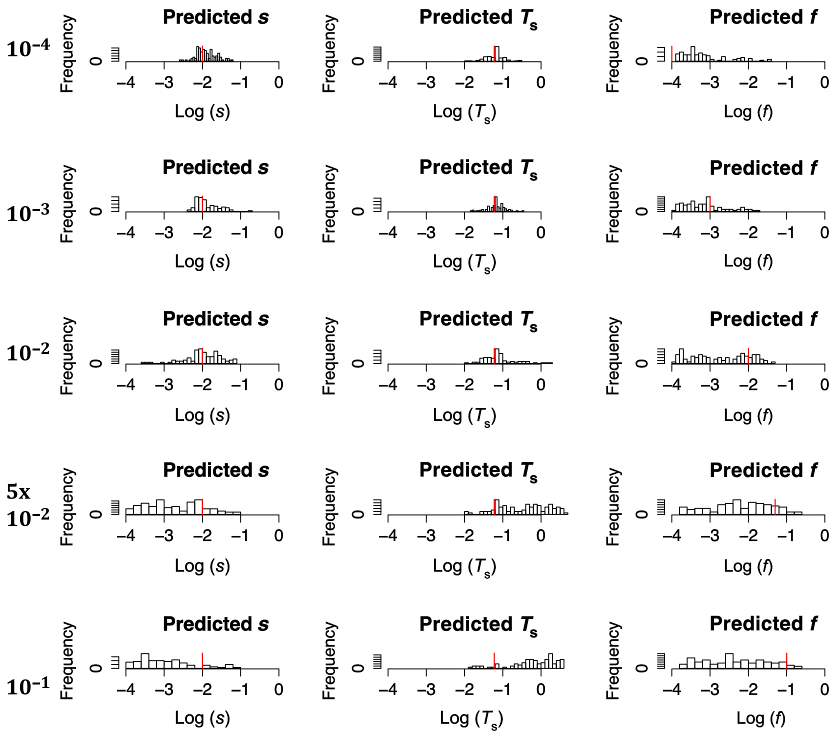

Fig. 3 Joint inference of s, T_s and f in equilibrium populations. Figures show the predicted values for 100 pseudo-observables for the example of $s = 0.01$, T_s=0.060 and $f = 0.0001$, 0.001, 0.01, 0.05, 0.1. Estimates of s, T_s and f were obtained from the mode of the joint posterior density. Red lines indicate the known values of the pseudo-observables. T_s represents the time since selection began acting on the allele (calculated using T_s = time to fixation ($T = 0.01$) + sojourn time T_{soj}).

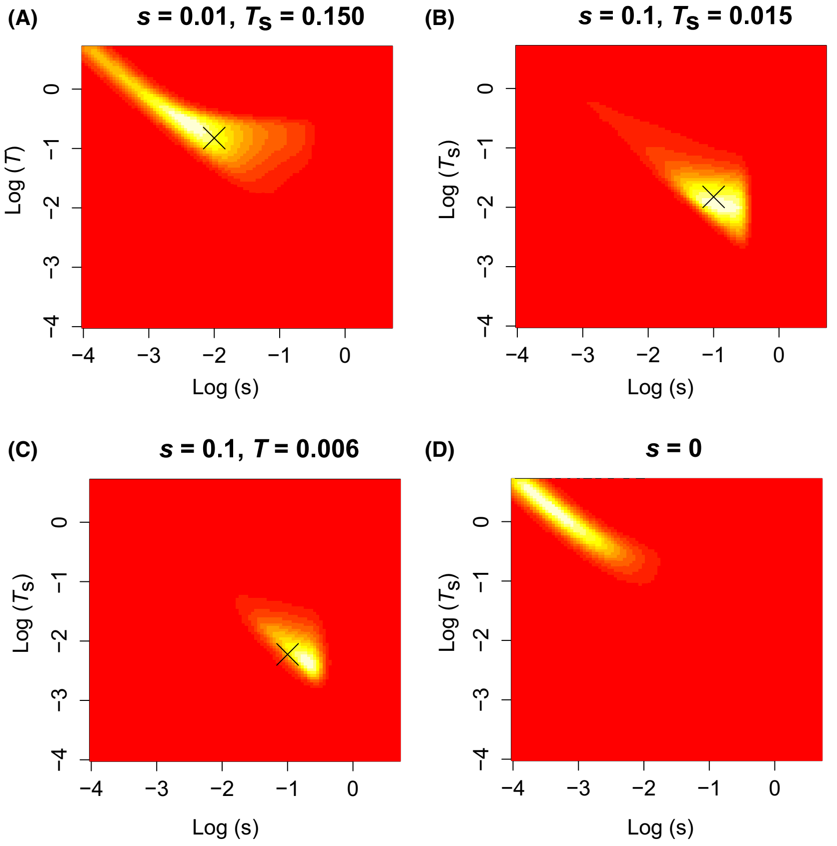

Fig. 4 Joint inference of s and T_s in demographic model for a strong bottleneck followed by exponential growth (demographic model 3) using MSSTATS-ABC. Figures show the cumulative joint posterior density plots for 100 pseudo-observable simulations. s is drawn from a log uniform prior: $\log_{10}(s) \sim U(-4, -0.5)$ and T_s from an adjusted log uniform prior: $\log_{10}(T_s) \sim U(\log_{10}(T_{soj}), \log_{10}(0.3 + T_{soj}))$. For the pseudo-observables, T_s is calculated from $T_s = T + T_{soj}$ where T is the time since fixation and the sojourn time $T_{soj} = (2\ln(2N_e)/s)/4N_e$. The white, yellow and red colours mark areas of high, moderate and low joint density respectively. Black crosses indicate the true values of pseudo-observables. (A) Inference for a moderately strong, old sweep with pseudo-observable values $s = 0.01$ and T_s=0.15 (calculated from $T_s = T + T_{soj}$ where $T = 0.1$). (B) Inference of a strong, very recent sweep with pseudo-observable values $s = 0.01$, $T_s = 0.006$ (calculated from $T_s = T + T_{soj}$ where $T = 0.001$) (C) Inference of a strong, very recent sweep with pseudo-observable values $s = 0.01$, $T_s = 0.006$ (calculated from $T_s = T + T_{soj}$ where $T = 0.001$). (D) Results of inference where no selected mutation was included in simulations.

and S11F, Supporting information). Similar results were obtained for the two bottlenecks scenarios (data not shown). One of the reasons for the estimator's strong results in nonequilibrium populations – and indeed its limitation – is that T_s rather than T is inferred. In our methodology, the prior for T_s is set as a function of s using an estimate of sojourn time T_{soj} in equilibrium populations. This analytical derivation most likely overestimates T_{soj} for alleles fixing in bottlenecked populations, and therefore, the full potential parameter space for T_s is not covered by our adjusted prior. This shortcoming could be corrected using another set of simulations, rather than the analytical derivation of Stephan *et al.* 1992, to estimate the minimum T_{soj} under a specific demographic scenario for a mutation of strength s. This would give a broader and more accurate prior from which to draw T_s.

When co-inferring f with T_s and s, we find the performance of the estimator deteriorates under nonequilibrium demography (Fig. S12, Supporting information). For both young and old sweeps, inference is only reliable if the method correctly identifies a de novo or very rare mutation ($f \leq 0.001$). It is difficult to correctly infer f and therefore to establish whether co-estimates of T_s and s are robust. As for equilibrium populations, a high level of standing variation is inferred as an older, weaker sweep than is the case for the pseudo-observable, due to high levels of diversity.

Data application: mouse coat colour evolution

We applied our methods to a previously published data set for 91 *P. maniculatus* deer mice living in the recently formed Nebraska Sand Hills (estimated age

8000 years) (Linnen *et al.* 2013). The data set was adjusted to cover SNPs over 20 kb on either side of a serine deletion on exon 2 which has been implicated in several traits associated with cryptic colour adaptation to a light phenotype for predator avoidance. After filtering for the serine deletion and genotyping quality, we retain SNP data from 64 alleles for analysis, and remove any further unknown sites. First, our aim was to co-estimate s and T assuming an equilibrium population of 53 080. Second, we estimate allele age T alone assuming a previously published estimate of $s = 0.126$ (Linnen *et al.* 2013). Third, we explicitly model the demographic scenario that had been previously inferred in our simulations (of a bottleneck 2900 years ago which reduced the population to 0.04% followed by an exponential recovery to 0.65% of the original population size). Lastly, we co-estimate f with s and T_s. We used msstats-ABC as this is consistent with the initial identification of the selected site described in Linnen *et al.* (2013). We simulated pseudo-observables with the specific mouse parameters to establish how well our methods work before applying these to the data set.

Assuming an equilibrium population, the joint inference of s and T showed a young, moderately strong selective sweep, with an inferred s of 8.7×10^{-3} (1.1×10^{-4} – 3.3×10^{-2}) and $T \leq 0.01$ (Fig. 5A), using a window size of 80 kb to ensure that diversity patterns are fully captured. Applying a window of 40 kb reduced the signal of the sweep (Fig. S13A, Supporting information). If s is assumed to be 0.126, as estimated in Linnen *et al.* (2013), the inference of allele age alone gives the same result of a young or very young sweep with $T \leq 0.01$ (Fig. S14, Supporting information).

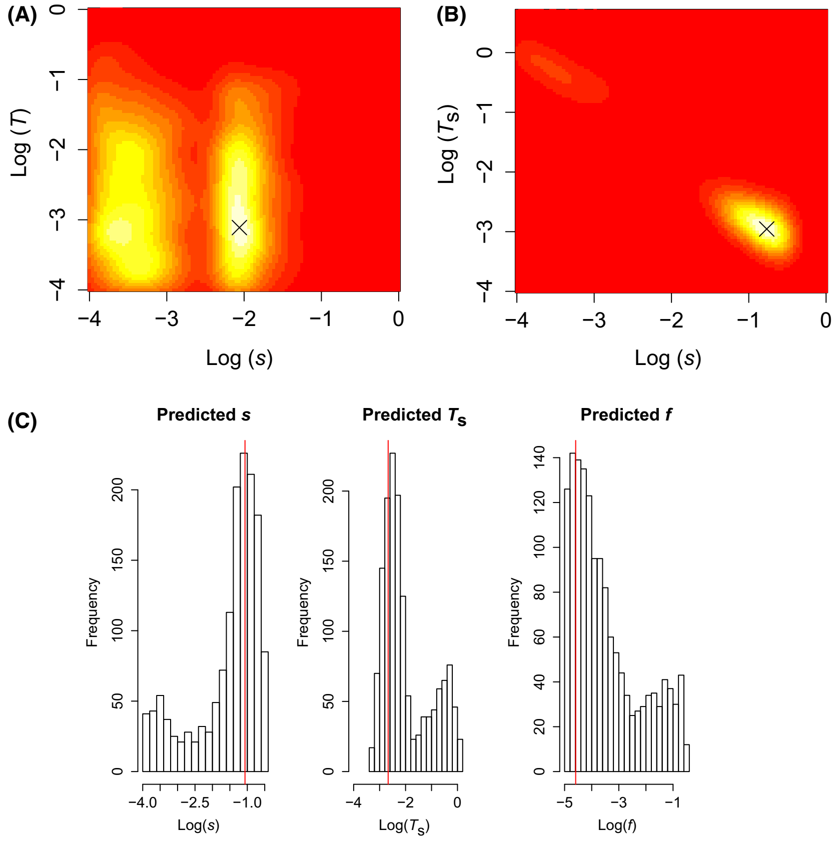

Fig. 5 Joint inference of allele age, selection coefficient and starting frequency for *Peromyscus maniculatus.* The joint density plots in A and B represent the results of the joint inference for the serine deletion at position 128,150 on exon 2, in (A) for s and T assuming an equilibrium population with $N_e = 53{,}080$ and in (B) for s and T_s with the demographic scenario inferred in (Linnen *et al.* 2013) explicitly included in simulations. In (C) f is co-inferred with T_s and s also assuming the demographic scenario inferred in (Linnen *et al.* 2013); histograms represent the posterior distributions from the ABC inference, with the red lines indicating the mode of the joint posterior density for the three parameters. The density plots are shown for $L = 80$ kb, with the mutation positioned centrally ($x = 0.5$). Other parameters for deer mice simulations are as described in methods. For (A), the mode of the joint density occurs at $s = 8.7 \times 10^{-3}(1.1 \times 10^{-4} - 3.4 \times 10^{-2})$ and $T = 7.7 \times 10^{-4}$ $(1.2 \times 10^{-4} - 1.0 \times 10^{-1})$. For (B), the mode occurs at $s = 1.7 \times 10^{-1}(1.5 \times 10^{-4} - 3.0 \times 10^{-1})$ and $T_s = 1.1 \times 10^{-3}(5.8 \times 10^{-4} - 8.8 \times 10^{-1})$. For (C), the mode occurs at $s = 8.6 \times 10^{-2}(1.5 \times 10^{-4} - 2.9 \times 10^{-1})$, T_s=$2.1 \times 10^{-3}(7.5 \times 10^{-4} - 9.0 \times 10^{-1})$ and $f = 2.6 \times 10^{-5}(1.1 \times 10^{-5} - 1.9 \times 10^{-1})$.

If the demography inferred in Linnen *et al.* (2013) is explicitly included in the simulations for the ABC calculation, simulations using pseudo-observables show that signals from selective sweeps with T_s coincident with or older than the bottleneck are usually quenched, leading to the inference of neutral scenarios (data not shown). In contrast, s and T_s for strong sweeps that are younger than the bottleneck (of the order of $T_s = 0.005$) are accurately inferred. This result illustrates the importance of using simulations to establish the limits of inference for specific scenarios. In applying our method to the mouse data, we infer a strong, recent sweep, with $s = 1.7 \times 10^{-1}$ $(1.5 \times 10^{-4} - 3.0 \times 10^{-1})$ and $T_s = 1.1 \times 10^{-3}$ $(5.8 \times 10^{-4} - 8.8 \times 10^{-1})$ (Fig. 5B). These results are consistent with those obtained under equilibrium demography but with a stronger estimate of s. Using a length of 40 kb, we find the qualitatively similar result of a strong recent sweep (Fig. S13B, Supporting information). We also find that simulated pseudo-observable sweeps that are either coincident or older than the mouse bottleneck are sometimes correctly inferred over this length, which is an improvement over the 80 kb length (data not shown), but are mostly inferred as neutral.

Our method is subject to the limitation that an estimate of sojourn time under equilibrium demography is used to set the prior for T_s, under the assumption that the mutation fixes. In our simulations, very recent values of T_s are therefore only associated with strong s. Here, we have checked with simulations that sojourn time is longer than under the equilibrium scenario, and therefore that the prior for T_s is broader than required, to reduce this source of error.

Co-inferring f, s and T_s jointly supports a recent, strong sweep acting on a de novo or rare mutation ($s = 8.6 \times 10^{-2}(1.5 \times 10^{-4} - 2.9 \times 10^{-1})$, $T_s = 2.1 \times 10^{-3}(7.5 \times 10^{-4} - 9.0 \times 10^{-1})$ and $f = 2.6 \times 10^{-5}$ $(1.1 \times 10^{-5} - 1.9 \times 10^{-1})$) (Fig. 5C). Simulations underpinning this estimate incorporate the demographic scenario from Linnen *et al.* (2013). In comparison with the age of the Sand Hills (i.e. 0.075 in units of $4Ne$ generations, assuming one generation every 6 months), these results support previous claims of selection acting on a young de novo mutation subsequent to the environmental change.

Discussion

We present ABC methods that estimate allele age, selection strength and starting frequency for fixed mutations using single population, single time point data sets. We demonstrate that it is possible to distinguish between different orders of magnitude of the selection coefficient s, between old and young sweeps, and between de novo/rare and common starting frequencies. There are significant differences between the ABC method that integrates ω_{max} and the MSSTATS-ABC, which undermine a direct comparison between the two methods. Namely, one takes account of ascertainment bias while the other does not. The ω_{max} approach was designed to be consistent with an approach for identifying sites under selection using the top ω_{max} values. Our simulations show that ω_{max} marginally outperforms a simple MSSTATS-ABC approach, particularly in estimating parameters for weak sweeps, as it is able to leverage a statistic that captures the specific LD pattern existing immediately after a selective fixation, but this is conditional on it being the appropriate method for the data analysed.

One of the major advantages of an ABC approach is that demography can be explicitly accounted for in simulations, which removes a source of error in estimating the strength of selective sweeps. Here, we illustrate this by explicitly including the previously estimated demographic model for our ABC estimation of s and T in deer mice. We find results that are consistent with those obtained under the assumption of an equilibrium population, but with slightly stronger estimates of selection. We also find that we can distinguish cases of selection on de novo and rare mutations from selection on common standing variation resulting in soft sweeps, and in the first case, we are able to co-infer T_s and s to within an order of magnitude, assuming equilibrium demography. Here, we find the most likely model to be one of selection on de novo or rare mutation. This is consistent with our estimates of allele age and provides support for the previously published notion of mutation-limited adaptation underpinning cryptic coloration in deer mice (Linnen *et al.* 2009, 2013; Poh *et al.* 2014).

Many haplotype methods such as iHS rely on a comparison between haplotype lengths for ancestral and derived alleles, and therefore have power to detect selected mutations at low or intermediate frequencies (Voight *et al.* 2006). Beyond this frequency level, power declines because these methods depend on a comparison with alternative allele haplotype structure. For example, the method published by Chen *et al.* (2015) applies to alleles under strong selection that are not yet fixed. Peter *et al.* (2012) use a range of haplotype- and SFS-based statistics including EHH and iHS to estimate allele age and selection coefficients for segregating mutations in models of de novo mutation and standing variation. The importance sampling method developed by Slatkin (2008) is specifically designed to identify s and T for low frequency alleles such as the A-allele of G6PD in Africa. In contrast to these methods predicated mainly on haplotype structure, our methods use SFS-based statistics that are sensitive to different parts of the SFS, as well as LD- and haplotype-based statistics that recover to equilibrium at different rates. Our methods thus fit an important niche, and may be utilized to infer the relative age, strength and frequency of fixed beneficial mutations relative to the timing of environmental shifts – in order to quantify, for example, the age of variants conferring cryptic coloration following the last ice age, as seen here in the mouse example and in the Laurent, Pfeifer *et al.* example in lizard populations also appearing in this issue.

Acknowledgements

This work was supported by grants from the Swiss National Science Foundation, and a European Research Council (ERC) Starting Grant, to JDJ. We thank Kristen Irwin for a careful reading of the manuscript, and Catherine Linnen, Yu-Ping Poh and Stefan Laurent for assistance with the mouse analysis.

References

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, **30**, 540–546.

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

Bjorn-Helge M, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, **18**, 1–23.

Box G, Cox D (1964) An analysis of transformations. *Journal of the Royal Statistical Society*, **26**, 211–243.

Chen H, Slatkin M (2013) Inferring selection intensity and allele age from multilocus haplotype structure. *G3 (Bethesda)*, **3**, 1429–1442.

Chen H, Hey J, Slatkin M (2015) A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Biology*, **99**, 18–30.

Csillery K, Francois O, Blum M (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

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, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

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

Feng D, Tierney L (2015) Package "misc3d", CRAN repository.

Foll M, Poh YP, Renzette N *et al.* (2014) Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genetics*, **10**, e1004185.

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.

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

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

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.

Joyce P, Marjoram P (2008) Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article26.

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

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

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science*, **325**, 1095–1098.

Linnen CR, Poh YP, Peterson BK *et al.* (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science*, **339**, 1312–1316.

Malaspinas AS, Malaspinas O, Evans SN, Slatkin M (2012) Estimating allele age and selection coefficient from time-serial data. *Genetics*, **192**, 599–607.

Mathieson I, McVean G (2013) Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, **193**, 973–984.

McVean GA (2002) A genealogical interpretation of linkage disequilibrium. *Genetics*, **162**, 987–991.

McVean G (2007) The structure of linkage disequilibrium around a selective sweep. *Genetics*, **175**, 1395–1406.

Orr HA, Betancourt AJ (2001) Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, **157**, 875–884.

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.

Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genetics*, **8**, e1003011.

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

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

Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.

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

Slatkin M (2008) A Bayesian method for jointly estimating allele age and selection intensity. *Genetics Research (Cambridge)*, **90**, 129–137.

Steinrücken M, Bhaskar A, Song YS (2014) A novel spectral method for inferring general diploid selection from time series genetic data. *The Annals of Applied Statistics*, **8**, 2203–2222.

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

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

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

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

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

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

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

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York.

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

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

All authors designed and performed, L.O. and J.D.J. wrote.

Data accessibility

All code described in the text can be found on the Software page of the Jensen Lab website: http://jensenlab.epfl.ch and on the Dryad Digital Repository: http://datadryad.org/(doi: 10.5061/dryad.qb6b5). All data described in the text is publicly available (see Linnen *et al*. 2013) and has been deposited in the NCBI Short Read Archive (Accession no. SRP017939).

Supporting information

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

Table S1 Summary statistics calculated in msstats

Table S2 Relative bias and RMSE estimates for inference of *T* alone

Table S3 Relative bias and RMSE estimates for joint inference of *s* and *T* for young and old sweeps (L = 10 kb x = 0.5)

Fig. S1 Variation of selected summary statistics with population scaled selection coefficient $N_e s$.

Fig. S2 Variation of ω_{max} with selection.

Fig. S3 Boxplot of the modes of posterior distributions of allele age *T* under weak Selection (s = 0.001).

Fig. S4 Boxplot comparing the impact of different window lengths on Inference power for allele age T.

Fig. S5 Joint inference of *s* and *T* in equilibrium populations for young sweeps (T = 0.01) (msstats---ABC).

Fig. S6 Joint inference of *s* and *T* in equilibrium populations for old sweeps (T = 0.01) (ω_{max} –ABC).

Fig. S7 Joint inference of *s* and *T* in equilibrium populations for young sweeps (T = 0.01) (ω_{max} –ABC).

Fig. S8 Boxplots of predicted values from joint inference of *s* and *T* in equilibrium populations (msstats---ABC).

Fig. S9 Boxplots of predicted values from joint inference of *s* and *T* in equilibrium populations (*w_max*-ABC).

Fig. S10 Joint inference of *s*, *T*s and f in equilibrium populations.

Fig. S11 Joint inference of *s* and *T*s under the demographic scenario of exponential growth following a sharp bottleneck using msstats–ABC.

Fig. S12 Joint inference of *s*, *T*s and *f* in a population undergoing exponential growth after a sharp bottleneck (demographic scenario 3).

Fig. S13 Joint inference of allele age and selection coefficient for *P. maniculatus*.

Fig. S14 Inference of allele age *T* alone for *P. maniculatus*.