An ABC Method for Estimating the Rate and Distribution of Effects of Beneficial Mutations

Jorge A. Moura de Sousa[1], Paulo R.A. Campos[2], and Isabel Gordo[1,*]

[1]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, Oeiras, Portugal

[2]Departamento de Física, Universidade Federal de Pernambuco, Cidade Universitaria, Recife PE, Brazil

*Corresponding author: E-mail: igordo@igc.gulbenkian.pt.

Accepted: March 23, 2013

Abstract

Determining the distribution of adaptive mutations available to natural selection is a difficult task. These are rare events and most of them are lost by chance. Some theoretical works propose that the distribution of newly arising beneficial mutations should be close to exponential. Empirical data are scarce and do not always support an exponential distribution. Analysis of the dynamics of adaptation in asexual populations of microorganisms has revealed that these can be summarized by two effective parameters, the effective mutation rate, U_e, and the effective selection coefficient of a beneficial mutation, S_e. Here, we show that these effective parameters will not always reflect the rate and mean effect of beneficial mutations, especially when the distribution of arising mutations has high variance, and the mutation rate is high. We propose a method to estimate the distribution of arising beneficial mutations, which is motivated by a common experimental setup. The method, which we call One Biallelic Marker Approximate Bayesian Computation, makes use of experimental data consisting of periodic measures of neutral marker frequencies and mean population fitness. Using simulations, we find that this method allows the discrimination of the shape of the distribution of arising mutations and that it provides reasonable estimates of their rates and mean effects in ranges of the parameter space that may be of biological relevance.

Key words: experimental evolution, mutation rate, distribution of fitness effects, parameter estimation.

Introduction

At what rate do beneficial mutations arise and what are their fitness effects? These are two of the most important questions regarding adaptation of organisms to novel environments (Kimura and Ohta 1974; Lang et al. 2011). Reflecting its importance, estimating genomic mutation rates of new beneficial alleles (U) and uncovering the mean effects of those beneficial mutations ($E(S)$) have been the subject of many studies (Rozen et al. 2002; Perfeito et al. 2007; Sawyer et al. 2007; MacLean and Buckling 2009; Bataillon et al. 2011; Estes et al. 2011; Sousa et al. 2012). Experimental evolution in clonal populations presents some advantages in determining these parameters, but some difficulties still arise, even in these controlled and relatively simple environments. One of these difficulties is being able to assay all the beneficial mutations. Different distributions of fitness effects are important to the adaptive process: the distribution of newly arising mutations, the distribution of contending mutations, which escape initial stochastic loss, and the distribution of mutations that survive competition with other mutations (clonal interference) and are able to actually fix, contributing to long-term adaptation (see Gordo et al. [2011] for a review). The greatest difficulty is to uncover the distribution of arising mutations, because they may easily be lost before reaching detectable frequencies. Despite this difficulty, determining the distribution that characterizes arising mutations, $f(S)$, is important, because it is this distribution that determines the nature of adaptation (Rozen et al. 2002; Perfeito et al. 2007; Orr 2010; Sousa et al. 2012). For this reason, some studies have tried to determine this distribution in viruses (Sanjuán et al. 2004; Rokyta et al. 2008), in bacteria (Kassen and Bataillon 2006; Stevens and Sebert 2011), and in other organisms (Desai et al. 2007; Schoustra et al. 2009; Burke et al. 2010; Orozco-terWengel et al. 2012). Experimental support for an exponential distribution of arising beneficial mutations has been obtained (Kassen and Bataillon 2006; MacLean and Buckling 2009), but this has not always been the case in all organisms and environments (Barrett et al. 2006; Rokyta et al. 2008; Bataillon et al. 2011; Gordo et al. 2011; Mcdonald et al. 2011). From the mutations that arise, those that end up outcompeting other beneficial mutations will drive long-term adaptation (Gerrish and Lenski 1998;

Good et al. 2012). The difference between the distributions of arising, contending, and fixed mutations is expected to depend on the effective population size (Crow and Kimura 1970), the mutation rate (Charlesworth et al. 1993), and the level of maladaptation, with increasingly adapted organisms having access to increasingly lower amount of beneficial mutations (Fisher 1930).

The biggest challenge in determining $f(S)$ lies in the rarity of beneficial mutations. In principle, this distribution can be determined directly by measuring fitness effects of extremely large samples of mutants (Lind et al. 2010; Hietpas et al. 2011). It can also be inferred from sequence data collected from natural populations (Nielsen 2005; Eyre-Walker and Keightley 2007; Jensen, Thornton, Andolfatto 2008; Jensen, Thornton, Aquadro 2008; Schneider et al. 2011). Indeed, scans for signatures of positive selection across the genome of different species, including our own, have been performed (Biswas and Akey 2006; Hancock and Di Rienzo 2008; Cutter and Choi 2010; Enard et al. 2010; Grossman et al. 2013). Disentangling the signature of selection from that caused by a complex demography is difficult (Grossman et al. 2010; Sinha et al. 2011), and checking the performance of different methods under departures from model assumptions is therefore an important task (Keightley and Eyre-Walker 2010). Recent advances have been made in developing methods for estimating selection coefficients from time series data of allele frequencies (Bollback et al. 2008; Malaspinas et al. 2012; Mathieson and McVean 2013) and also in disentangling alleles under positive selection from passenger mutations (Illingworth and Mustonen 2011). In the context of experimentally evolved populations, where typically the experimenter imposes a particular demographic regime, one method that has been used proposes to study beneficial mutations through assaying the evolutionary dynamics of neutral markers in asexual populations (Imhof and Schlotterer 2001; Hegreness et al. 2006). The basic principle underlying this method relies on the "hitchhiking effect" of a neutral allele with mutations that give an advantage to the organism (Maynard-Smith and Haigh 1974). This same principle is at the heart of methods to detect positive selection across the genome of sexually reproducing organisms (Thornton et al. 2007). In experimentally evolved populations, the frequency of a neutral allele can be easily measured (e.g., by using neutral fluorescent markers), and inferring evolutionary parameters from neutral marker dynamics can thus be performed under certain theoretical assumptions (Dykhuizen and Hartl 1983; Hegreness et al. 2006; Barrick et al. 2010; Illingworth and Mustonen 2012). A simple and quite elegant method was proposed by Hegreness et al. (2006): Using simulations, they showed that a simple population genetics model, where all beneficial mutations have the same effect, is able to reproduce the dynamics of a commonly used marker system involving one locus with two neutral markers. The dynamics can therefore be summarized by two parameters that theoretically represent the evolutionary process: the effective mutation rate (U_e) and the effective selection coefficient (S_e). Barrick et al. extended this method and determined the values of U_e and S_e in different strains of *Escherichia coli* (Barrick et al. 2010; Woods et al. 2011). Although it may be useful to be able to summarize the process under a single mutational effect, far more realistic distributions of fitness effects can also explain the data. Recently, Illingworth and Mustonen (2012) proposed a new method to estimate the distribution of haplotype fitnesses in experimentally evolving populations. When tested against simulated data under the assumption of an exponential distribution of arising beneficial mutations, the method is able to retrieve the correct distribution of haplotype fitnesses for values U below 10^{-6}. It is, however, not known how the method performs for other distributions of arising beneficial mutations and for larger values of U. Moreover, this method estimates the distribution of haplotype fitnesses segregating in populations and not the distribution of beneficial arising mutations. Here, we ask two questions: how do the effective parameters compare with the more biologically meaningful parameters U and $E(S)$? and, because frequency dynamics appear insufficient to distinguish between different distributions (Hegreness et al. 2006), is there a reasonable set of data that can be obtained, which allows the determination of the distribution of arising beneficial mutations?

We address both these questions from a theoretical perspective, taking a commonly used experimental setup to study the adaptation of asexual populations in controlled environments as a motivation. This setup simply involves tracking a marker locus with two neutral alleles. We show that the effective evolutionary parameters can provide good estimates of U and of the mean effect of beneficial mutations only when the distribution of effects of arising mutations has limited variance. However, when the variance is increased (e.g., if arising mutations follow an exponential distribution), we find that U_e can underestimate the true value U, whereas S_e can overestimate the true value of $E(S)$. We propose a new method based upon measurements of both the frequency of neutral markers and mean population fitness, at periodic time intervals. This method, which was motivated by typical experimental setups easily applied to experimental evolution, is theoretically expected to estimate the mutation rate reasonably well and allows distributions of arising beneficial mutations with different shapes to be distinguished.

Materials and Methods

Model of Adaptation to Simulate Evolutionary Dynamics

We assume a clonal population reproducing according to the Wright–Fisher model, where periodic bottlenecks occur (with a period of T_{bot}). The population is initially isogenic, with the exception of a neutral marker, which is biallelic and has a frequency $f_0 = 0.5$ for one of the alleles. The initial population

size is N_0. Generations are discrete and the population doubles each generation for $t < T_{bot}$. With period T_{bot}, the population size is reduced by random sampling to N_0. This assumed demography, involving periodic bottlenecks where the number of individuals is fixed, is typical of most experimental setups, where daily passages of a sample of the population are performed, and population numbers are experimentally controlled. At each generation, mutations occur at a rate U per genome, following a Poisson distribution. All mutations are beneficial, and the effects of each mutation (S) are drawn from a continuous distribution $f(S)$. We allow for variation of the selective effects of arising mutations, assuming a Gamma distribution, with shape and scale parameters, α and β, respectively, implying a mean $E(S) = \alpha\beta$. Similar to other studies that previously proposed to estimate the distribution of arising deleterious mutations (Keightley 1998; Eyre-Walker and Keightley 2007), we have assumed a Gamma distribution because it can have a wide range of shapes. Multiplicative fitness is assumed, so that the effects of mutations do not depend on the genetic background where they arise. This is obviously an oversimplification, because the distribution can change along the adaptive walk (Martin and Lenormand 2006; Sousa et al. 2012), but we consider a short-term evolution scenario where U and $f(S)$ may be assumed constant. Genetic drift is modeled by sampling, from a multinomial distribution, classes of individuals with the same fitness. The frequency dynamics of the neutral marker ($f(t)$), as well as the mean population fitness ($w(t)$), are followed. This model of adaptation is used to produce a set of simulated evolutionary dynamics, from which evolutionary parameters are estimated using different methods: a method developed by Hegreness et al. (and extended by Barrick et al.) and a new method that we propose here that simultaneously tries to estimate U and $f(S)$ (see later).

The range of parameters chosen to produce simulated data with the described model was made in accordance with current estimates in different systems but mostly in microorganisms. U is currently estimated to achieve values between 10^{-4} and 10^{-9}, depending on the environment and genetic background (Drake et al. 1998; Perfeito et al. 2007; Lang et al. 2011; Denver et al. 2012). An effective population size $N_e = 10^5$ was assumed (corresponding to bottlenecks with a period $t_{bot} = 5$ generations) for all simulations except when indicated differently.

Generating Pseudo-Observed Data

Pseudo-observed data sets were generated under the model of adaptation described earlier, with a specific value of the mutation rate U and a specific Gamma distribution (with parameters α and β) with mean $E(S)$. These data sets represent biological data that can be acquired in an experiment. The new method proposed here with the goal of estimating $f(S)$ and U requires the periodic measure of the frequency of the neutral markers and the mean population fitness (every 50 generations for a 300-generation experiment). These appear reasonable to obtain experimentally and require experimental work that is typical in evolution experiments performed in controlled environments: In addition to assaying the frequency of the markers (as already is typically done [Woods et al. 2011]), fitness has to be measured by performing either a direct competitive fitness assay against the ancestral strain or a measurement of the population growth rate at different times along the experiment (Gordo et al. 2011). Furthermore, the choice of studying 100 replicate populations reflects the 100- or 96-well plate experimental setup that is commonly used (Lemonnier et al. 2008; Kvitek and Sherlock 2011). These plates are affordable by most laboratories, and, with a multichannel pipette, several passages can be performed in little time, space, and at low cost, particularly when studying microbial populations. Regarding the markers, many strains expressing different fluorescent alleles are available, which makes the acquisition of frequency data a relatively easy task. This can be performed using flow cytometry or another fluorescence reader. Competitive fitness measurements (Elena and Lenski 2003) can also be easily performed using a similar setup.

The pseudo-observed data therefore consist of the marker frequencies and the fitnesses at periodic time points of the experiments (ti) for n independently evolved populations. Different pseudo-observed data sets assuming different distributions of S were generated to test the two different methods: Barrick et al. method (which requires the marker frequencies only), to compare U_e with U and S_e with $E(S)$, and the One Biallelic Marker Approximate Bayesian Computation (ABC, which requires both the marker frequencies and the fitnesses) to assess its ability to estimate U, α, and β (see later).

Estimation of U and $E(S)$ by U_e and S_e Based upon the Dynamics of the First Significant Deviation of $f(t)$

For a given set of pseudo-observed data, we obtained the effective parameters U_e and S_e and compared them with the biologically meaningful values of U and $E(S)$. To obtain U_e and S_e, we followed Barrick et al. A large set of simulated evolutionary dynamics under the assumption that all beneficial mutations have the same value of S was generated. This simulated data consist of sets of 100 replicate populations evolved under different parameter combinations of U and S. The range of $\log_{10}(U)$ was $[-8; -3.95]$, with increments of 0.15, and the range of S was [0.01; 0.18], with increments of 0.01. This simulated data are the input of Barrick et al. (2010) method to obtain U_e and S_e. For each simulation, it consists of the logarithm of the ratio of the two subpopulation frequencies $(\mathrm{Rf} = f(\mathrm{ti})/(1 - f(\mathrm{ti}))$ at several time points, ti, $(\ln(\mathrm{Rf}(\mathrm{ti})))$, where $\mathrm{ti} = 5 \times i$. We then use this input in the program marker_divergence_fit.pl, whose output is fed into the program marker_divergence_significance.pl, both

available at http://barricklab.org/twiki/bin/view/Lab/ToolsMarkerDivergence, to obtain U_e and S_e. The first program summarizes the evolutionary dynamics (both in the simulated data sets and in the pseudo-observed data) in two statistics: τ_e and α_e. The first is the time, τ_e, where a significant deviation of $\ln(Rf(\tau_e))$ from $\ln(Rf(t=0))$ occurs. The second is the rate of change of $\ln(Rf(t))$ with time, that is, τ_e sets the time of divergence of marker frequency and α_e the rate of divergence. Each replicate population is summarized by a single value of τ_e and α_e, and the n replicate populations (characterized by a given combination (U, S)) result in a distribution of $T(\tau_e)$ and $A(\alpha_e)$. These distributions are then compared, using the second program, to the distributions of τ_e and α_e that summarize the pseudo-observed data $To(\tau_e)$ and $Ao(\alpha_e)$ using a two-dimensional Kolmogorov–Smirnov to test the fit between the simulated data and the pseudo-observed data. The combination (U, S) that gives rise to the highest P value is taken as U_e and S_e, even when the hypothesis that the distributions are different cannot be rejected.

This procedure was done to obtain the results in figure 1 and supplementary figures S1 and S4, Supplementary Material online, where 20 independent replicates of each pseudo-observed data set (under the same U, α, and β) were performed, and the average of U_e and S_e obtained for each pseudo-observed data set is presented.

New Estimation Method Based upon the Dynamics of Frequency and Fitness

We propose a new method, the One Biallelic Marker ABC (fig. 2), which aims to infer the distribution of arising mutations. The pseudo-observed data used to infer the performance of the method are generated under the model of adaptation explained before, but the method now analyses the distributions, along time intervals (ti), of both marker frequency (f(ti)) and fitness (w(ti)), where $ti = i \times 50$ generations ($i = 0$–6) are measured for 100 replicate populations evolving under a given U, α, and β. A large data set with 1 million simulated evolutionary dynamics is produced, each with a set of 100 replicate populations evolving under a specific combination of parameters U, α, and β. For each of the simulations, each parameter is randomly chosen from the following distributions: $\log_{10}(U) \sim$ Uniform [−9; −4]; $\alpha \sim$ Uniform [0.5; 15]; and $\log_{10}(\beta) \sim$ Uniform [−4; −0.08]. Both the pseudo-observed data and the simulated data are summarized as the distribution of the values of $|0.5 - f(ti)|$ represented as a histogram, with five binned classes, for the marker frequency at different time points (ti), and of the distribution of fitness effects at the same time points, w(ti), represented as a histogram with six binned classes. This results in 11 summary statistic values for each of the six time points used in the analysis (table in fig. 2).

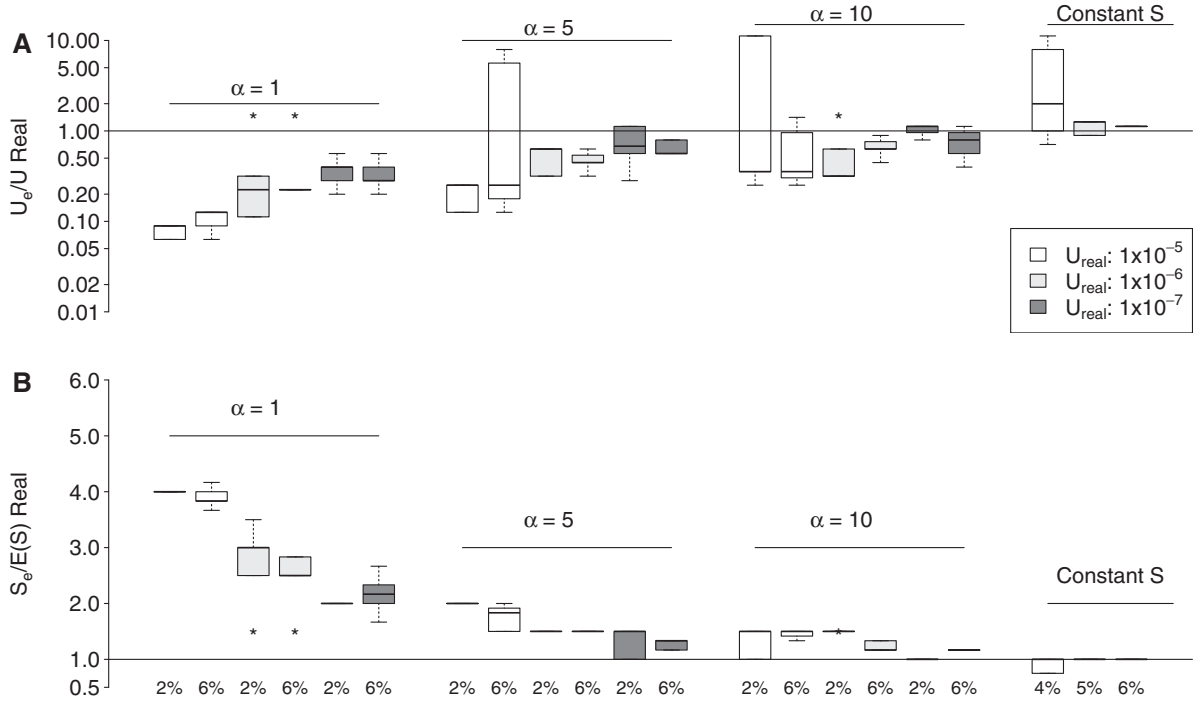

Fig. 1.—Performance of the single S model, according to the highest scoring estimates for pseudo-observed data with $f(S)$ as Gamma distributions of different variances. (*A*) Ratios of estimates of U over real parameter U. (*B*) Ratios of estimates of S over the mean effect of S. The box plots of 20 independent estimation processes are shown, with the median indicated as a bar. Asterisks indicate cases where none of the 20 independent replicates was fitted significantly.

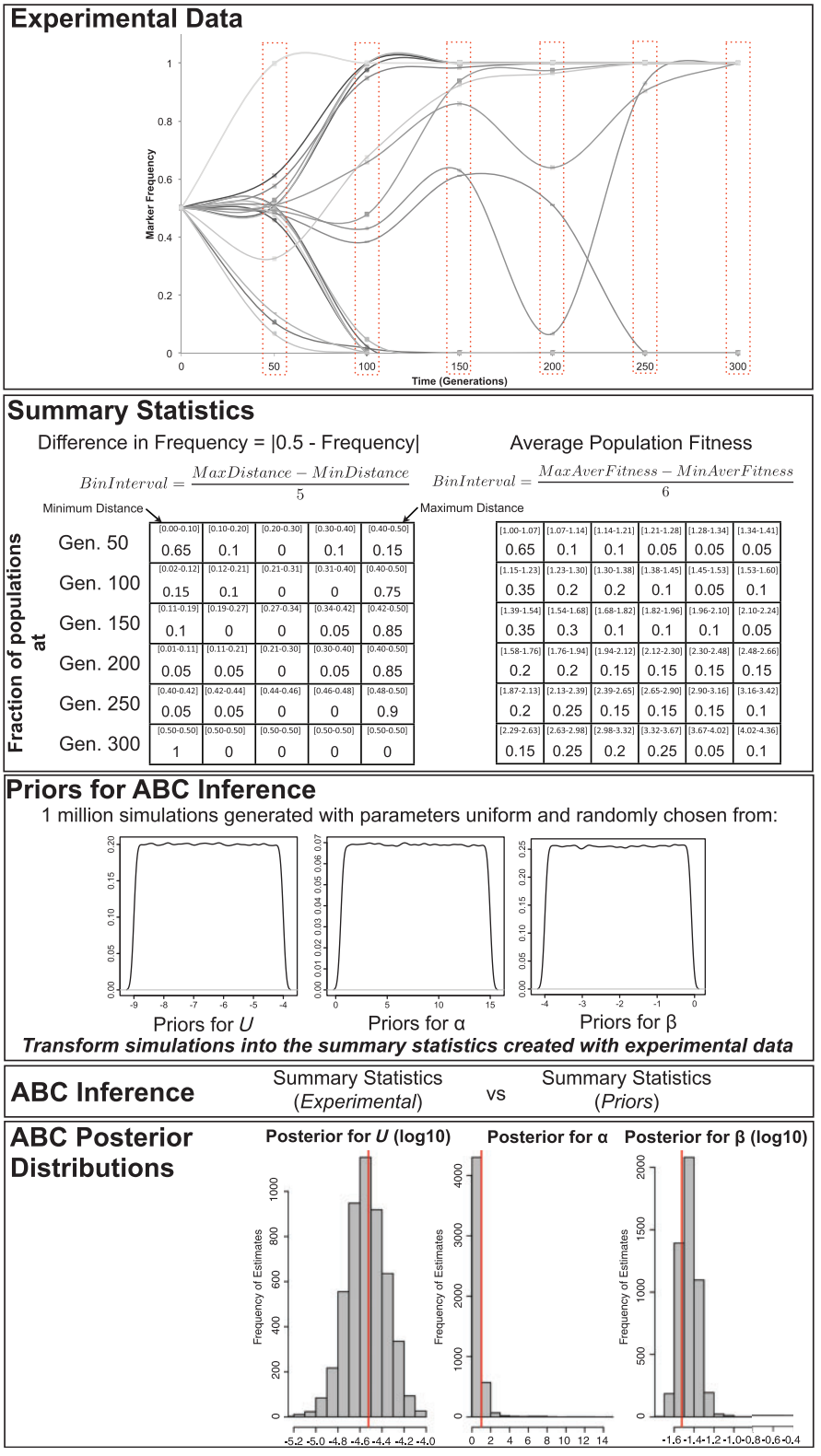

Fig. 2.—Schematic description of the One Biallelic Marker ABC. Data are obtained from an evolution experiment (here called pseudo-observed data), at specific time points, involving replicate adaptations to a common environment (an example of 20 replicate populations is shown). For each time point, the data are condensed to summary statistics, for marker frequency and mean population fitness, which are histograms with the frequency of populations that fall in different bins (5 for frequency data and 6 for fitness data) at every 50 generations (Gen.). The choice of the bin for the frequency statistics is dictated by

(continued)

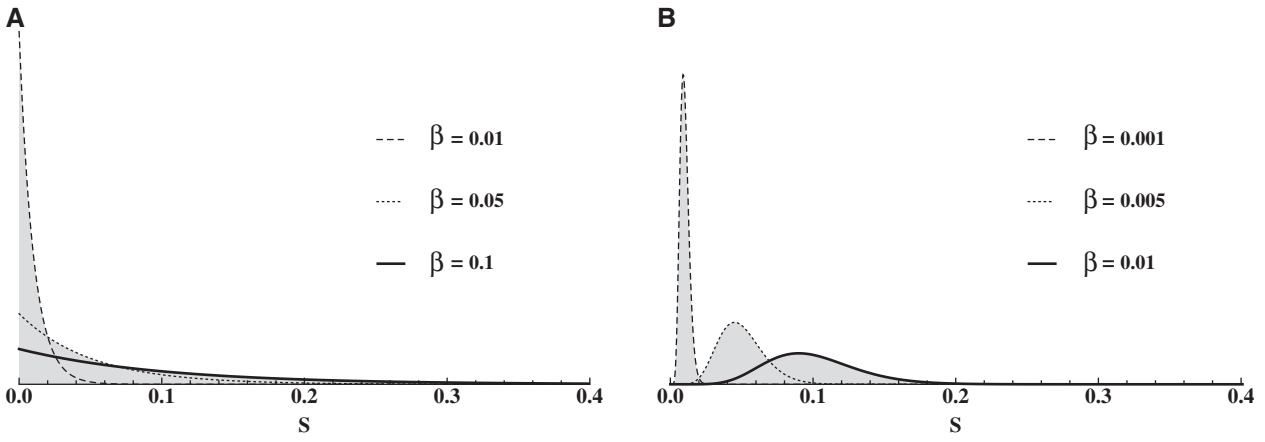

Fig. 3.—Theoretical distributions of beneficial selective coefficients assumed to produce pseudo-observed data. (*A*) Gamma distribution with shape parameter $\alpha = 1$ (exponential distribution), for different scale (β) parameters. (*B*) Gamma distribution with shape parameter $\alpha = 10$, for different scale (β) parameters.

Summary statistics from pseudo-observed and simulated data are compared using an ABC method (Beaumont et al. 2002) implemented in R (Csilléry et al. 2012) (package downloaded from and available at http://cran.r-project.org/web/packages/abc/index.html). ABC approaches have previously been used, for example, to determine rates of selective sweeps using sequence data from populations of *Drosophila melanogaster* (Jensen, Thornton, Andolfatto 2008). The inputs of the ABC method are the summary statistics of the 100 replicate populations that compose the pseudo-observed data ($S(y_0)$) and the previously described 1 million simulations ($S(y_i)$). The ABC method computes the posterior probability distribution of a multivariate parameter, θ (composed of a combination of *U*, α, and β). A value for this parameter, θ_i, is sampled from the prior distributions, and the summary statistics computed from simulated data $S(y_i)$ are compared with those of the pseudo-observed data $S(y_0)$ using the Euclidian distance *d*. If *d* is below a given threshold, the parameter value θ_i is accepted. The threshold (tolerance) chosen was 0.5%, which corresponds to the proportion of accepted simulations. The estimation of the posterior probability distribution for θ can be improved by different regression-based methods available in the ABC R package (Csilléry et al. 2012): local linear regression and neural networks. We used the neural network method, which performs a dimensionality reduction in the summary statistics, and is suggested to be appropriate for use with high dimensionality (Csilléry et al. 2012). Through this procedure, we obtain estimates for *U*, α, and β, outputted as posterior distributions for each parameter. For each combination of parameters (*U*, α, and β), 20 independent pseudo-observed data sets were considered to produce the statistics presented in the results. A scheme with the different steps described here is represented in figure 2.

Effect of Variation in Initial Frequency of Marker and Presence of Deleterious Mutations

We tested the effect of small variations in the initial frequency of each of the initial subpopulations (supplementary fig. S3, Supplementary Material online) as they may occur in any experimental setup. We also tested how the estimates would be affected by the occurrence of deleterious mutations (supplementary fig. S2, Supplementary Material online). For the first scenario, we generated pseudo-observed data sets under the same assumptions of the adaptation model described earlier except that the initial frequency of the neutral marker $f(t=0) = 0.5 + \varepsilon$, where ε is drawn from a Uniform distribution, $\varepsilon \sim$ Uniform [−0.03; 0.03] (supplementary fig. S3, Supplementary Material online). For testing the effect of deleterious mutations, we generated pseudo-observed data sets assuming that, in addition to beneficial mutations, deleterious mutations can also occur at a rate of 10^{-3} and each having a selection coefficient $S_{del} = 2\%$. Multiplicative fitness

Fig. 2.—Continued
the module of the difference between the initial and current frequency of the subpopulations, so that this value is, at most, 0.5 (for marker frequencies of 1 or 0). A large simulated data set is built against which the experimental data are compared. The priors chosen to produce the simulated data set, which consist in 1 million simulations, are shown. Each simulated data (obtained with a given value of *U*, α, and β) are then classified according to the same summary statistics as calculated for the observed data—called Summary stats (Priors) and Summary stats (Experimental), respectively. Using ABC inference, these summary statistics are compared and the ones closest to the experimental data chosen. The 5,000 top-ranked values (0.5%) of each of the parameters are shown as the posterior distribution where the median value is highlighted in red.

was also assumed (supplementary fig. S2, Supplementary Material online). All other assumptions were kept the same. Pseudo-observed data sets for both cases were generated with a Gamma distribution of fitness effects with $\alpha = 1$.

Results

Comparison of Effective Parameters U_e and S_e with U and the Average Effect of Beneficial Mutations

To determine whether the effective parameters U_e and S_e are good estimates of U and S, we generated pseudo-observed data for a given value of U and with fitness effects drawn from a Gamma distribution with different shape and scale parameters (fig. 3). For populations with $N_e = 10^5$, the estimated values of the effective parameters are shown in figure 1, where we also have included results for the case where pseudo-observed data were generated under a model where all beneficial mutations have the same effect, because this is the case where the estimates are expected to perform best. Figure 1*A* shows that U_e provides a good estimate of U for Gamma distributions with shape parameter bigger than 1. This is observed in the cases where U is low ($<10^{-6}$), but when $U = 10^{-5}$ and $E(S) = 0.02$, U_e underestimates U by a quarter of its real value. Larger biases can be seen for the exponential distribution ($\alpha = 1$), particularly under high mutation rates ($>10^{-7}$), where clonal interference may be more pronounced. We find that when the distribution of S is exponential with mean 2% and the mutation rate is 10^{-5} (parameters that have been estimated in some bacteria evolution experiments [Perfeito et al. 2007]), U_e considerably underestimates the real value of U by an order of magnitude. This also happens when the mean effect of beneficial mutations is 6%. The underestimation becomes smaller when either the mutation rate or the variance in S decreases. As expected, U_e provides an accurate estimate of U when S is constant (except for the case where a high value of the mutation rate is considered). Figure 1*B* shows that S_e overestimates $E(S)$ two to four times for a mutation rate higher 10^{-6}, with $\alpha = 1$. For $\alpha = 10$, this overestimation is small (<1.5-fold). Importantly, however, most of these values for S_e seem to provide an estimate of the order of magnitude of the mean effect of beneficial mutations. To test whether the bias in U_e and S_e increases with clonal interference, we also studied populations with increased effective population size ($N_e = 10^6$). Indeed, we find that both U_e (supplementary fig. S1*A*, Supplementary Material online) and S_e (albeit to a lesser extent) (supplementary fig. S1*B*, Supplementary Material online) show larger deviations from U and $E(S)$, which can be up to a 50-fold underestimates of U and a 6-fold overestimates of $E(S)$. In sum, higher levels of clonal interference (more pronounced in larger populations and with higher values of the mutation rate) lead to larger biases in U_e and S_e. These biases are dependent upon the underlying distribution of beneficial mutations.

Estimation of the Distribution of Arising Beneficial Effects

To go beyond the mean effect of beneficial mutations and to try to estimate the distribution of arising beneficial mutations, we developed a new method, which we call One Biallelic Marker ABC. To test its performance in retrieving the evolutionary parameters U, α, and β, we explored different sets of pseudo-observed data with combinations of parameter values that seem reasonable given the current literature (Perfeito et al. 2007; Lang et al. 2011; Denver et al. 2012).

In figure 4, we show the ability of the One Biallelic Marker ABC method to estimate U, α, and β, when the distribution of arising mutations is exponential. This is the most commonly assumed distribution in theoretical studies of the adaptive process (Betancourt and Bollback 2006; Orr 2010). Figure 4*A* shows that the One Biallelic Marker ABC method provides estimates of U within an order of magnitude, for all cases tested. The worst performance lies in retrieving U for both high values of $E(S)$ (5% and 10%) and high U (3×10^{-5}), but even in these cases, the estimated value allows for a correct estimate of the order of magnitude of U. For the intermediate value of the mutation rate studied ($U = 3 \times 10^{-6}$), the method provides an accurate estimate of U. Figure 4*B* and *C* provide the results for the estimates of the shape and scale parameters of $f(S)$. As shown in figure 4*B*, the estimate of the shape parameter α is close to 1 or 2, for the majority of the cases considered. Exceptions occur for the high mutation rate and the larger β values, which have a very high variance. Estimation of β, shown in figure 4*C*, is remarkably good, across the parameter range studied, being always below 2-fold the real value of β.

We also studied the case where the distribution of arising beneficial mutations has a different shape, specifically $\alpha = 10$ (fig. 3). As shown in figure 5*A*, the estimated values of U are very close to the real ones in this case, rarely exceeding two times the real U values, although it can be either over or underestimated, depending upon the average selective effect. The two parameters characterizing the distribution of arising mutations are also remarkably close to the real values. Figure 5*B* shows that α is always estimated to be close to its true value (between 7 and 12), irrespectively of the value of U. Importantly, this estimate of α allows us to detect that the distribution of arising mutations is not exponential. The method, therefore, has power to reliably distinguish between distributions of effects with distinct shapes. In figure 5*C*, the performance of the estimates regarding the β parameter of the distribution of effects is shown. β is well estimated, never exceeding twice the real value.

To further assess the power of the method in distinguishing distributions with different shapes, we studied intermediate values of α, between 0.75 and 10. In figure 6, we show that the One Biallelic Marker ABC method is able to discriminate not only between the two limiting cases in our simulations but also between intermediate α values. The method fails to

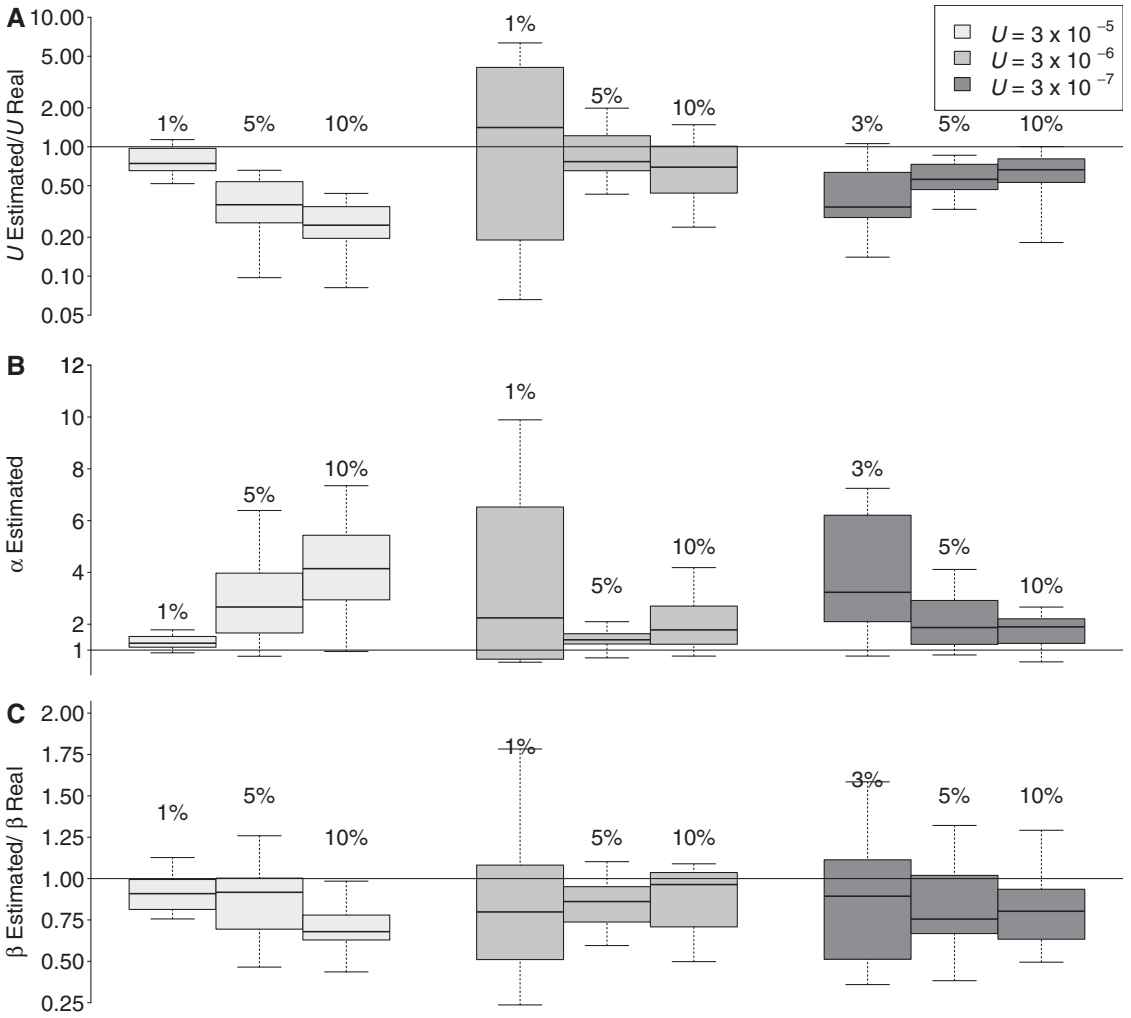

Fig. 4.—Performance of the One Biallelic Marker ABC method, for pseudo-observed data generated with $f(S)$ as a Gamma distribution with $\alpha = 1$ and different scales (β). Labels above the bars show the mean selection coefficient for each case, and x axis shows the value assumed for the mutation rate. (*A*) Ratios of estimates of U over real parameter U. (*B*) Estimates of the shape (α) parameter. (*C*) Ratios of estimates of the scale (β) parameter over the real scale parameter. The box plots of 20 independent estimation processes are shown, with the median indicated as a bar.

distinguish the shape of the distribution of arising mutations when $0.75 < \alpha < 2$, especially when U is large. In these cases, α is overestimated (by about 2-fold). When $\alpha > 2$, estimates of α consistent with the true value are obtained. When $\alpha = 4$, rejection of an exponential distribution is obtained. Overall, the method provides a reliable distinction between different shapes of the distribution of arising mutations, although distinguishing between α values lower than 2 remains difficult.

Discussion

To estimate the parameters that describe the dynamics of adaptation, we need powerful methods. Beneficial mutations are essential in driving adaptation and their statistical properties remain an open question (Orr 2010). Although methods developed to tackle this subject may never perfectly capture the complete nature of the evolutionary process, they can provide reasonable estimates regarding the strength of the forces involved in the process (Thornton et al. 2007; Keightley and Eyre-Walker 2010).

A simple theoretical approach assumes that all mutations have the same fitness effect and has been shown to have predictive power in explaining certain patterns of data obtained in experimentally evolved populations (Hegreness et al. 2006). Notwithstanding, several direct measurements of mutation effects point to the existence of considerable variation (Kassen and Bataillon 2006), which motivates the development of new methods that try to infer the underlying distribution of arising beneficial effects.

Regarding the estimated effective evolutionary parameters studied here, it seems clear that the relation with the real parameters is dependent on the actual distribution of effects

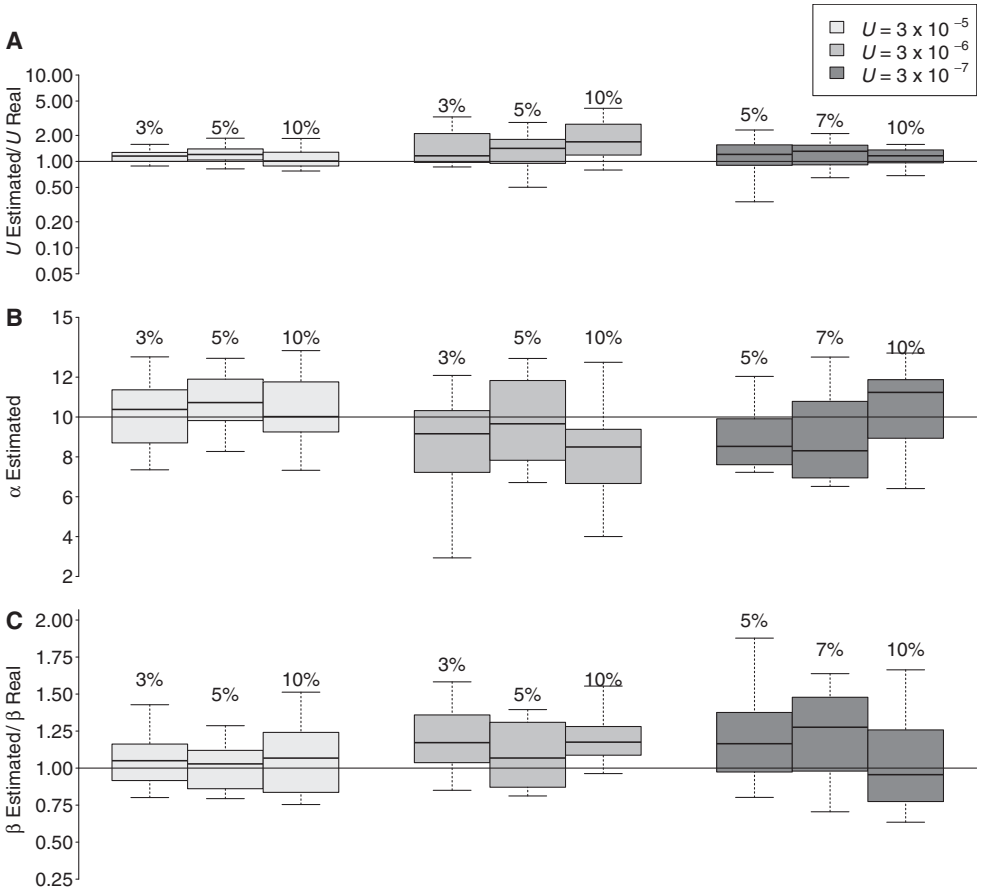

Fig. 5.—Performance of the One Biallelic Marker ABC method, for pseudo-observed data generated with $f(S)$ as a Gamma distribution with $\alpha = 10$ and different scales (β). Labels and symbols are as in figure 4.

of arising mutations: exponential-like distributions of beneficial effects result in values of U_e below the true mutation rate and values of S_e above the true mean effect of mutations, with the difference being reduced when the distribution of effects decrease in variance. Nevertheless, assuming a fixed value for S has been a commonly used method to infer the evolutionary parameters from experimental data, for example, in studies that address how evolvability is dependent upon the genetic background. In one such study, Barrick et al. (2010) isolated eight clones of *E. coli* with different mutations in the *rpoB* gene, encoding the β subunit of RNA polymerase. As these mutations are generally deleterious in environments without antibiotics, and they can cause a wide range of fitness defects (Trindade et al. 2010), the authors estimated U_e and S_e to determine the evolvability of different (but related) genotypes. The two neutral markers dynamics were used to estimate the evolutionary parameters, and, from these dynamics, it was inferred that mutants with a higher fitness defect had a higher evolvability caused by a stronger selective effect of beneficial mutations. Interestingly, the inferred mutation rate (through U_e) appears to be independent of the genetic

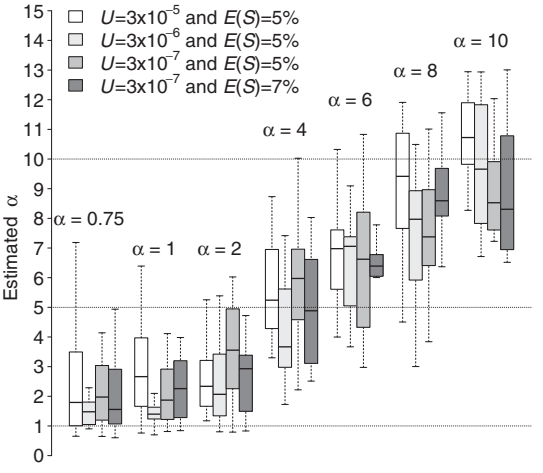

Fig. 6.—Estimates for the shape (α) parameter, for pseudo-observed data generated with $f(S)$ as a Gamma distribution with varying α and scales (β), in order for $E(S)$ to be constant within a mutation rate. Labels show the α parameters used as pseudo-observed data for each case. The box plots of 20 independent estimation processes are shown, with the median indicated as a bar.

background. Because we show here that U_e may be below U and S_e above $E(S)$, some caution is to be taken when drawing conclusions regarding the relation between evolvability and fitness effects of such mutations. Similar caveats apply in the study of Woods et al. (2011). That study involved a long-term evolution experiment, running for more than 50,000 generations, where clones sampled at generation 500 were found to carry mutations in *topA* and *rbs*. These were shown to be beneficial and fixed after generation 1,500, and their carriers were called "eventual winners." Other contemporaneous genotypes (with other mutations) were deemed "eventual losers." Even though both sets of clones had increased fitness related to the ancestral, the "eventual losers" also had, counter intuitively, increased fitness relative to the "eventual winners." To understand why the "eventual winners" ultimately won the competition, their evolvability was studied, and U and $E(S)$ were inferred (through U_e and S_e) by assaying neutral marker dynamics. The authors found that "eventual winners" had, indeed, the ability to generate beneficial mutations with stronger effects, compared with the "eventual losers."

The approach used in both studies to determine evolvability may provide an overestimate of the mean selective coefficient in the order of two to three times the real values if the mutation rates are in the order estimated by the authors, or even more, if the mutation rates are underestimated (fig. 1*B*). As a consequence, this could imply that the actual mean selective coefficients are lower than the one estimated, and small differences in evolvability may be difficult to detect.

In general, inferring evolutionary parameters and, more specifically, the distribution of arising mutations, from data of evolving populations is a difficult task. Experimentally, one way to gain further insight into the distribution of effects is to use more than two neutral markers, which can bring more power (Perfeito et al. 2007). Theoretically, we can expect that new and improved methods are likely to emerge. Recently, Zhang et al. (2012) extended the previous model by Hegreness et al. to incorporate a continuous initial growth phase, dividing it in 50 time intervals, and developing an analytical model to find the distributions of estimators for U and S. Similar to the previous work, however, only the initial dynamics are considered (the first significant deviation), and the method does not consider the occurrence of clonal interference. Illingworth and Mustonen (2012), on the other hand, developed a maximum likelihood method where the marker dynamics over the total amount of time followed is used. The method determines the minimum number of mutations that best describe the dynamics and allows inferring the distribution of haplotype fitnesses that are segregating. Although the performance of the method is quite good under certain conditions, it is not clear how it will perform under a wide range of mutation rates.

Here, we propose a new theoretical approach that is expected to contribute to improved insight regarding the distribution of arising beneficial mutation effects. Using ABC, we propose a set of summary statistics to be used under a simple experimental setup, where distributions of marker frequencies and the mean fitness of the population are recorded at periodic time intervals. These statistics allow a reasonable estimation of the distribution of arising mutations and of the mutation rate, provided that we accept that such a distribution may be well approximated by a Gamma. Combining the parameters of the Gamma distribution (α and β, it is also possible to estimate the mean effect of arising beneficial mutations ($E(S)$). Figure 7 shows the estimates of $E(S)$ given by the method when $\alpha = 1$ or $\alpha = 10$. Under an exponential distribution of fitness effects (fig. 7*A*), which is commonly assumed, the mean effect can be overestimated up to 5- or 6-fold, for large values of the mutation rate. For $\alpha = 10$, the estimate $E(S)$ is very accurate, reflecting its real value for every condition tested (fig. 7*B*).

The One Biallelic ABC method seems to allow distinguishing between distributions with different shapes and scales. The underlying model used makes several assumptions, which could be violated in a real experiment. In particular, it assumes that the initial population is composed of two equally sized subpopulations, each with a different marker, and it also assumes that no deleterious mutations occur. To test the robustness of the approach in the face of these assumptions, we performed new simulations where pseudo-observed data were generated. In one case, the initial marker frequency was allowed to deviate from its expectation of 0.5 (supplementary fig. S3, Supplementary Material online). In the other case, deleterious mutations were allowed to occur with rates and effects typical of those inferred from mutation accumulation experiments with bacteria (Kibota and Lynch 1996; Trindade et al. 2010) (supplementary fig. S2, Supplementary Material online). In both these cases, the inference of the values of U, α, and β was similar to those obtained before.

We performed the analysis of a method, which assumes a common experimental setup with only one neutral locus with two alleles and fitness measurements at periodic time intervals. In principle, this setup can be extended to follow variation of one locus with more alleles or neutral variation at more loci. The method could then be extended, and a thorough study of the best summary statistics would be needed to ask what would be the minimal set of data required to reasonably estimate the rate and distribution of arising beneficial mutations.

We have also tested the effect of considering a smaller number of populations to determine whether the approaches can provide reasonable estimates when applied to data that have been obtained in studies involving experimental evolution with fewer replicates. Supplementary figure S4, Supplementary Material online, shows the comparison of U_e with U and S_e with $E(S)$ when the number of replicate populations is 10, which corresponds to the approximate size of previously published experiments (Hegreness et al. 2006; Barrick et al. 2010; Woods et al. 2011). We observed similar biases to those found when considering 100 replicate evolved populations.

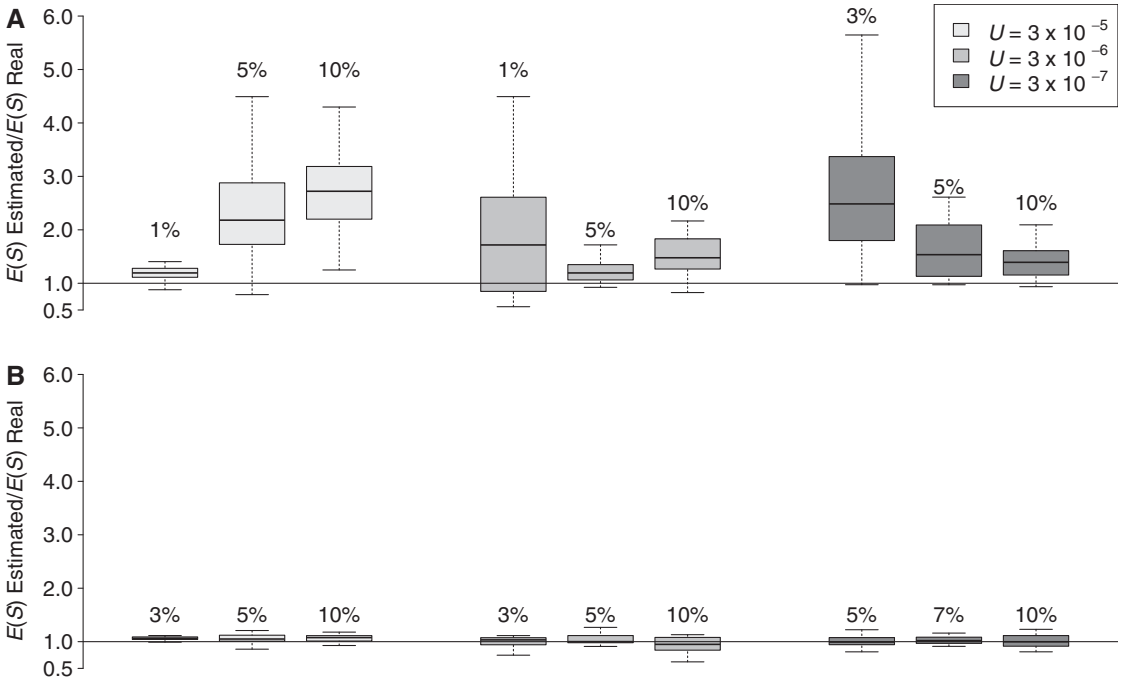

Fig. 7.—Ratios of estimated average beneficial selective coefficient over the real average beneficial selective coefficient (both calculated as $\alpha\beta$). Labels above the bars show the mean selection coefficient for each case, and *x* axis shows the value assumed for the mutation rate. Pseudo-observed data were generated with $f(S)$ as a Gamma distribution with $\alpha = 1$ (exponential distribution of $f(S)$). The box plots of 20 independent estimation processes are shown, with the median indicated as a bar.

Regarding the One Biallelic Marker ABC approach, we can observe that even with these reduced number of populations, reasonable estimates of U can be obtained; the estimated values of α tend to produce an overestimation, which can be up to 15-fold, whereas the estimates of β are close to the real ones (supplementary fig. S5, Supplementary Material online).

The One Biallelic Marker ABC method, as the alternatives discussed earlier, displays certain limitations in its performance, which are particularly apparent when dealing with very intense clonal interference, for which a system with more markers would be desirable. It is a method that tries to estimate the distribution without limiting the number of mutations in a given genetic background and taking into account the dynamics of the entire process of adaptation. For a wide spectrum of mutation rates, we are able to estimate the parameters of the underlying distribution of beneficial mutations. The One Biallelic Marker ABC method was tested over a range of distributions of beneficial selective coefficients and beneficial mutation rates, including high mutation rates, which are typically not studied in the analysis of other methods. This gives us a fairly good degree of confidence that, in applying the method to real biological data from adaptation experiments of clonal populations using the two-marker methodology, we are able to gain information on the distribution of beneficial arising mutations.

Supplementary Material

Supplementary figures S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Lilia Perfeito, Ana-Hermina Ghenu, Lindi Wahl, the Gordo's Laboratory members, two anonymous referees, and the editor for their comments and suggestions. This work was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 260421 – ECOADAPT; the scholarship provided by Fundação Calouste Gulbenkian (FCG) and Fundação para a Ciência e Tecnologia (FCT) to J.A.M.S.; the salary support of LAO/ITQB and FCT to I.G.; and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), and program PRONEX/MCT-CNPq- FACEPE to P.C.

Literature Cited

Barrett RDH, MacLean RC, Bell G. 2006. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. Biol Lett. 2:236–238.

Barrick JE, Kauth MR, Strelioff CC, Lenski RE. 2010. *Escherichia coli rpoB* mutants have increased evolvability in proportion to their fitness defects. Mol Biol Evol. 27:1338–1347.

Bataillon T, Zhang T, Kassen R. 2011. Cost of adaptation and fitness effects of beneficial mutations in *Pseudomonas fluorescens*. Genetics 189: 939–949.
Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.
Betancourt AJ, Bollback JP. 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev. 16:618–623.
Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. 22:437–446.
Bollback JP, York TL, Nielsen R. 2008. Estimation of 2Nes from temporal allele frequency data. Genetics 179:497–502.
Burke MK, et al. 2010. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature 467:587–590.
Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.
Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper & Row.
Csilléry K, François O, Blum MGB. 2012. ABC: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 3:475–479.
Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. Genome Res. 20:1103–1111.
Denver DR, et al. 2012. Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. Genome Biol Evol. 4:513–522.
Desai MM, Fisher DS, Murray AW. 2007. The speed of evolution and maintenance of variation in asexual populations. Curr Biol. 17: 385–394.
Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667–1686.
Dykhuizen DE, Hartl DL. 1983. Selection in chemostats. Microbiol Rev. 47: 150–168.
Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet. 4:457–469.
Enard D, Depaulis F, Roest Crollius H. 2010. Human and non-human primate genomes share hotspots of positive selection. PLoS Genet. 6: e1000840.
Estes S, Phillips PC, Denver DR. 2011. Fitness recovery and compensatory evolution in natural mutant lines of *C. elegans*. Evolution 65: 2335–2344.
Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610–618.
Fisher RA. 1930. The genetical theory of natural selection. Oxford: The Clarendon Press.
Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. Genetica 102–103:127–144.
Good BH, et al. 2012. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci U S A. 109:4950–4955.
Gordo I, Perfeito L, Sousa A. 2011. Fitness effects of mutations in bacteria. J Mol Microbiol Biotechnol. 21:20–35.
Grossman SR, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327: 883–886.
Grossman SR, et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703–713.
Hancock AM, Di Rienzo A. 2008. Detecting the genetic signature of natural selection in human populations: models, methods, and data. Annu Rev Anthropol. 37:197–217.
Hegreness M, Shoresh N, Hartl D, Kishony R. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311:1615–1617.
Hietpas RT, Jensen JD, Bolon DNA. 2011. From the cover: experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A. 108: 7896–7901.
Illingworth CJ, Mustonen V. 2011. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. Genetics 189:989–1000.
Illingworth CJ, Mustonen V. 2012. A method to infer positive selection from marker dynamics in an asexual population. Bioinformatics 28: 831–837.
Imhof M, Schlotterer C. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. Proc Natl Acad Sci U S A. 98: 1113–1117.
Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet. 4:e1000198.
Jensen JD, Thornton KR, Aquadro CF. 2008. Inferring selection in partially sequenced regions. Mol Biol Evol. 25:438–446.
Kassen R, Bataillon T. 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet. 38:484–488.
Keightley PD. 1998. Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: a simulation study. Genetics 150:1283–1293.
Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos Trans R Soc Lond B Biol Sci. 365:1187–1193.
Kibota TT, Lynch M. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. Nature 381:694–696.
Kimura M, Ohta T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A. 71:2848–2852.
Kvitek DJ, Sherlock G. 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet. 7:e1002056.
Lang GI, Botstein D, Desai MM. 2011. Genetic variation and the fate of beneficial mutations in asexual populations. Genetics 188:647–661.
Lemonnier M, et al. 2008. The evolution of contact-dependent inhibition in non-growing populations of *Escherichia coli*. Proc Biol Sci. 275:3–10.
Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. Science 330:825–827.
MacLean RC, Buckling A. 2009. The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. PLoS Genet. 5: e1000406.
Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. 2012. Estimating allele age and selection coefficient from time-serial data. Genetics 192: 599–607.
Martin G, Lenormand T. 2006. The fitness effect of mutations across environments: a survey in light of fitness landscape models. Evolution 60: 2413–2427.
Mathieson I, McVean G. 2013. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 193:973–984.
Maynard-Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.
Mcdonald MJ, Cooper TF, Beaumont HJE, Rainey PB. 2011. The distribution of fitness effects of new beneficial mutations in *Pseudomonas fluorescens*. Biol Lett. 7:98–100.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.
Orozco-terWengel P, et al. 2012. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Mol Ecol. 21:4931–4941.
Orr HA. 2010. The population genetics of beneficial mutations. Philos Trans R Soc Lond B Biol Sci. 365:1195–1201.

Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: high rate and small effects. Science 317:813–815.
Rokyta DR, et al. 2008. Beneficial fitness effects are not exponential for two viruses. J Mol Evol. 67:368–376.
Rozen DE, de Visser JAGM, Gerrish PJ. 2002. Fitness effects of fixed beneficial mutations in microbial populations. Curr Biol. 12: 1040–1045.
Sanjuán R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A. 101:8396–8401.
Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci U S A. 104:6504–6510.
Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437.
Schoustra SE, Bataillon T, Gifford DR, Kassen R. 2009. The properties of adaptive walks in evolving populations of fungus. PLoS Biol. 7: e1000250.
Sinha P, et al. 2011. On detecting selective sweeps using single genomes. Front Genet. 2:1–5.
Sousa A, Magalhães S, Gordo I. 2012. Cost of antibiotic resistance and the geometry of adaptation. Mol Biol Evol. 29:1417–1428.
Stevens KE, Sebert ME. 2011. Frequent beneficial mutations during single-colony serial transfer of *Streptococcus pneumoniae*. PLoS Genet. 7: e1002232.
Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity (Edinb) 98:340–348.
Trindade S, Perfeito L, Gordo I. 2010. Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. Philos Trans R Soc Lond B Biol Sci. 365:1177–1186.
Woods RJ, et al. 2011. Second-order selection for evolvability in a large *Escherichia coli* population. Science 331:1433–1436.
Zhang W, et al. 2012. Estimation of the rate and effect of new beneficial mutations in asexual populations. Theor Popul Biol. 81:168–178.

Associate editor: Cécile Ané