Identifying Signatures of Selection in Genetic Time Series

Alison F. Feder,*[,1] Sergey Kryazhimskiy,†[,1] and Joshua B. Plotkin*[,2]
*Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, and †Department of Organismic and Evolutionary Biology and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138

ABSTRACT Both genetic drift and natural selection cause the frequencies of alleles in a population to vary over time. Discriminating between these two evolutionary forces, based on a time series of samples from a population, remains an outstanding problem with increasing relevance to modern data sets. Even in the idealized situation when the sampled locus is independent of all other loci, this problem is difficult to solve, especially when the size of the population from which the samples are drawn is unknown. A standard χ^2-based likelihood-ratio test was previously proposed to address this problem. Here we show that the χ^2-test of selection substantially underestimates the probability of type I error, leading to more false positives than indicated by its *P*-value, especially at stringent *P*-values. We introduce two methods to correct this bias. The empirical likelihood-ratio test (ELRT) rejects neutrality when the likelihood-ratio statistic falls in the tail of the empirical distribution obtained under the most likely neutral population size. The frequency increment test (FIT) rejects neutrality if the distribution of normalized allele-frequency increments exhibits a mean that deviates significantly from zero. We characterize the statistical power of these two tests for selection, and we apply them to three experimental data sets. We demonstrate that both ELRT and FIT have power to detect selection in practical parameter regimes, such as those encountered in microbial evolution experiments. Our analysis applies to a single diallelic locus, assumed independent of all other loci, which is most relevant to full-genome selection scans in sexual organisms, and also to evolution experiments in asexual organisms as long as clonal interference is weak. Different techniques will be required to detect selection in time series of cosegregating linked loci.

POPULATION geneticists typically seek to understand the forces responsible for patterns observed in contemporaneous samples of genetic data, such as the nucleotide differences fixed between species, polymorphisms within populations, and the structure of linkage disequilibrium. Recently, however, there has been a rapid increase in the availability of dynamic data, where the frequencies of segregating alleles in an evolving population are monitored through time, both in laboratory experiments (Hegreness *et al.* 2006; Bollback and Huelsenbeck 2007; Barrick *et al.* 2009; Lang *et al.* 2011; Orozco-terWengel *et al.* 2012; Lang *et al.* 2013) and in natural populations (Barrett *et al.* 2008; Reid *et al.* 2011; Denef and Banfield 2012; Winters *et al.* 2012; Daniels *et al.* 2013; Maldarelli *et al.* 2013; Pennings *et al.* 2013). One important question is whether the changes in allele frequencies observed in such data are the result of natural selection or are simply consequences of genetic drift or sampling noise. In principle, it seems that dynamic data should provide researchers with more power to detect and quantify selective forces while avoiding the assumptions of stationarity that are required for many inference techniques based on static samples (Sawyer and Hartl 1992; Boyko *et al.* 2008; Desai and Plotkin 2008). Nonetheless, the behavior and power of inference techniques based on time series data have not been thoroughly investigated.

There is a well-developed literature on inferring population sizes from genetic time-series data, assuming neutrality (Pollak 1983; Waples 1989; Williamson and Slatkin 1999; Wang 2001), and a rapidly growing literature on inferring natural selection from such time series (Bollback *et al.* 2008; Illingworth and Mustonen 2011; Illingworth *et al.* 2012; Malaspinas *et al.* 2012; Mathieson and McVean 2013). However, even the simplest case—the dynamics of two alternative alleles at a single genetic locus independent

doi: 10.1534/genetics.113.158220
Manuscript received October 2, 2013; accepted for publication November 28, 2013; published Early Online December 6, 2013.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.158220/-/DC1.

[1]These authors contributed equally to this work.

[2]Corresponding author: Department of Biology, University of Pennsylvania, Philadelphia, PA 19104. E-mail: jplotkin@sas.upenn.edu

of all other loci—presents a number of statistical challenges that have not been resolved. The main complication arises when the actual size of the population from which the serial samples are drawn is unknown. In this case, large changes in the frequency of an allele might indicate either that the allele is under selection or that the population size is small and genetic drift is strong. To favor one alternative over the other Bollback *et al.* (2008) proposed to fit two nested Wright–Fisher models to time-series data at a single locus (one model with selection and one without) and reject the neutral model, using the χ^2-distribution for the likelihood-ratio statistic. Such an approach is generally the most powerful and unbiased, at least for large data sets. Nonetheless, here we show that in practice the actual frequency of false positives under this approach can vastly exceed the nominal *P*-value obtained from the χ^2-distribution—and especially so at more stringent *P*-value cutoffs. Since the χ^2-distribution does not provide an accurate representation of the false positive rate, this approach cannot be used to draw sound statistical conclusions about selection from such time series. The underlying reason for this problem is that the likelihood-ratio statistic is χ^2-distributed only asymptotically, and convergence to this distribution is slow (Wilks 1938). In most practical applications, such as when sampling from natural populations (Reid *et al.* 2011; Denef and Banfield 2012; Winters *et al.* 2012; Daniels *et al.* 2013; Maldarelli *et al.* 2013; Pennings *et al.* 2013) or competing two microbial strains (Lenski *et al.* 1991; Bollback and Huelsenbeck 2007; Lang *et al.* 2013), the number of sampled time points is typically small (<10) and the distribution of the likelihood-ratio statistic is far from χ^2 under neutrality, leading to more false positives then expected.

We propose two solutions to fix this problem, providing unbiased tests for natural selection in time-series data sampled at a single genetic locus. First, we develop an algorithm for computing the exact distribution of the likelihood-ratio statistic under neutrality. Although feasible in many regimes, this direct approach suffers from several complications that we discuss below. We also propose an alternative, computationally efficient, albeit approximate, statistical method for rejecting the neutral model. Our approach builds directly on the work of Bollback *et al.* (2008), and it is likewise limited to studying time series of allele frequencies at a single locus under genic selection, assuming independence from all other loci. The more complicated problem of detecting selection from genomic time series of many linked loci has received attention elsewhere (Illingworth and Mustonen 2011; Illingworth *et al.* 2012), and the problems identified here likely apply to those situations as well.

We start our presentation by introducing a likelihood framework for time-series data at a single genetic locus. We then demonstrate that the *P*-value given by the χ^2-distribution for the likelihood-ratio statistic underestimates the actual false-discovery rate. Next, we introduce two methods to correct this bias, and we verify that they are virtually unbiased for large sample sizes and conservative for small sample sizes. We quantify the power of these two tests for selection in different parameter regimes, considering also noise in the measurements of allele frequencies. Finally, we apply our methods to three experimental data sets and demonstrate that the tests behave as expected in practical situations.

Materials and Methods

Approximate expression of the transition probability for the Moran process

Calculating the likelihood of an allele-frequency time series requires knowing the transition probability $P_s(x, t|x', t')$ that the frequency of the observed allele in the population at time t is x, given that it was x' at some previous time t'. The subscript s indicates that this probability depends on the selection coefficient of the allele. In general, it will also depend on the population size N and maybe on other parameters. Under most population-genetic models no exact analytical expressions for the transition probability $P_s(x, t|x', t')$ are available for arbitrary x, x', t, t', and s. The standard approximation to the discrete Wright–Fisher and Moran models is the diffusion approximation of Kimura and others (Ewens 2004). Although considerably simpler than the discrete models, the diffusion equation is still difficult to solve exactly and efficiently in a general case. Although some numerical methods are available (Kimura 1955a,b; Evans *et al.* 2007; Bollback *et al.* 2008; Song and Steinrücken 2012), they are often cumbersome to implement or computationally intensive.

Therefore, we use a Gaussian approximation to the Wright–Fisher process, which is less accurate than the diffusion approximation but allows us to obtain a simple analytical expression for the transition probability, which can be computed efficiently and is quite accurate provided the allele has not been lost or fixed during the period of observation. We emphasize that our two tests for selection proposed below do not intrinsically depend upon this Gaussian approximation (that is, they could in principle be implemented using the full Wright–Fisher model or the Kimura diffusion), but we nonetheless rely on this approximation for efficiency's sake. Moreover, as we discuss below, there is little additional power to be gained by considering time series that exhibit many sampled time points with fixed alleles, provided that sampling noise is small.

We describe the Gaussian approximation in detail in the *Appendix* and summarize it here. Briefly, if the timescale of observation is short compared to N in the case of a neutral allele or to $1/s$ in the case of a positively selected allele, *i.e.*, if absorption events can be neglected, the Moran process can be approximated by a sum of a deterministic process g and a Gaussian noise process Z (Pollett 1990). In the absence of genetic drift, *i.e.*, when $N \to \infty$, the allele frequency X behaves deterministically, $X \to g(t, x_0)$, where g satisfies the logistic equation

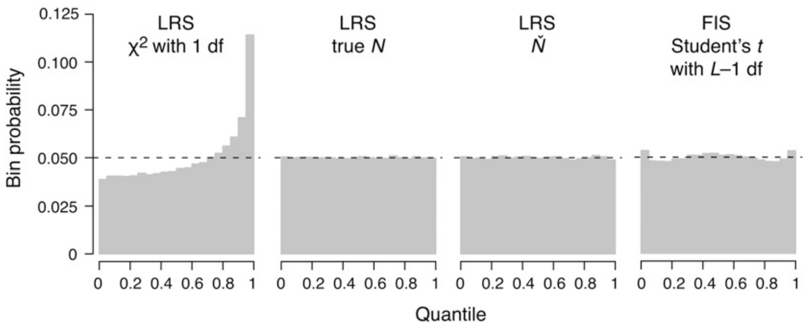

Figure 1 Distributions of test statistics under the neutral null hypothesis. Histograms show the probabilities that the value of a test statistic generated under the neutral null hypothesis falls within each vigintile (quantiles of size 0.05) of another, approximate, distribution. If the approximate distribution is close to the true distribution, the probability for each bin will approximately equal 0.05 (dashed line). The left three panels show the probability distributions for the likelihood-ratio statistic (LRS) to fall into the vigintiles of the χ^2-distribution with 1 d.f., the LRS distribution under the true N, and the empirical LRS distribution under $\check{N}$, respectively. The LRS falls in the top vigintiles of the χ^2-distribution more often than expected, indicating that the P-value given by the χ^2-distribution underestimates the probability of a type I error. The distribution of LRS under the true N is shown as a control case. The distribution of LRS under $\check{N}$ closely approximates the true LRS distribution. The rightmost panel shows the probabilities for the frequency increment statistic (FIS) to fall into each vigintile of Student's t-distribution with $L - 1$ d.f. Student's t is a good approximation for the true distribution of the FIS. Parameter values were $N = 10^3$, $T = 100$, $\Delta = 20$, $L = 5$, and $\nu_0 = 0.5$; the number of Wright–Fisher simulations was 3.5×10^5.

$$\dot{g} = sg(1-g), \tag{1}$$

$$g(0, x_0) = x_0, \tag{2}$$

whose solution is

$$g(t, x_0) = x_0\big(x_0 + (1-x_0)e^{-st}\big)^{-1}. \tag{3}$$

Here x_0 is the initial deterministic allele frequency. When $N < \infty$, genetic drift perturbs the allele frequency X from its deterministic value and so $X(t) = g(t, x_0) + Z(t)$, where $Z(t)$ is the noise process. Then for any two time points $t' \geq 0$ and $t > t'$, the transition probability is approximated by

$$P_s(x, t \mid x', t') \approx \sqrt{\frac{N}{2\pi\sigma^2(\Delta t, g')}} \times \exp\left\{-N\frac{(x-g-(x'-g')M(\Delta t, g'))^2}{2\sigma^2(\Delta t, g')}\right\}, \tag{4}$$

where

$$M(\Delta t, \xi) = e^{-s\Delta t}\Big(\xi + (1-\xi)e^{-s\Delta t}\Big)^{-2}, \tag{5}$$

$$\begin{aligned}\sigma^2(\Delta t, \xi) = M^2(\Delta t, \xi)(2+s)\xi(1-\xi)s^{-1} \\ \times [2\xi(1-\xi)s\Delta t + \xi^2 e^{s\Delta t} - (1-\xi)^2 e^{-s\Delta t} \\ + (1-\xi)^2 - \xi^2],\end{aligned} \tag{6}$$

and we used shorthands $g \equiv g(t, x_0)$, $g' \equiv g(t', x_0)$, $\Delta t = t - t'$. Under the neutral null hypothesis (*i.e.*, when $s = 0$), the transition probability simplifies to

$$P_0(x, t \mid x', t') \approx \sqrt{\frac{N}{2\pi\sigma_n^2(\Delta t, x_0)}}\exp\left\{-N\frac{(x-x')^2}{2\sigma_n^2(\Delta t, x_0)}\right\}, \tag{7}$$

with

$$\sigma_n^2(\Delta t, x_0) = 2x_0(1-x_0)\Delta t. \tag{8}$$

Note that functions (3)–(8) depend on parameters N and s and on the nuisance parameter x_0 that in principle can be estimated along with N and s. However, for the sake of reducing the number of fitted parameters, we fix x_0 to be equal to the observed allele frequency at time zero, $x_0 \equiv \nu_0$.

We assume here that time is measured in generations. If time is measured in physical units, Equations 3 and 5 still hold, with rescaled parameters $N \to N\tau$ and $s \to s/\tau$, where τ is the generation time; Equation 6 does not hold exactly because of the term $2 + s$, but holds approximately as long as $s \ll 1$, which is true in most cases. Thus, Equations 4 and 7 can still be used.

Implementation

In *Results and Discussion*, we obtain the expression for the likelihood $L(\text{Data}; N, s)$ of allele-frequency data as a function of two parameters, N and s. We estimate these parameters by maximizing this likelihood expression. First, consider the case when the allele frequency is measured at only two time points t_0 and t_1 with the corresponding frequencies being ν_0 and ν_1. Then the likelihood expression (9) with the Gaussian approximation (4) becomes

$$L(\text{Data}; N, s) = \sqrt{\frac{N}{2\pi\sigma^2(\Delta t, \nu_0)}}\exp\left\{-N\frac{(\nu_1 - g(t_1, \nu_0))^2}{2\sigma^2(\Delta t, \nu_0)}\right\},$$

which is maximized at $\hat{N} = \infty$ and

$$\hat{s} = \frac{1}{t_1 - t_0}\ln\left(\frac{\nu_1}{1-\nu_1}\frac{1-\nu_0}{\nu_0}\right).$$

In this case, the Gaussian likelihood function collapses to a δ-function centered at $\hat{s}$ so that $L(\text{Data}; \hat{N}, \hat{s}) = \infty$. In other words, with two data points there is enough information to estimate the selection coefficient but not the population size. Thus, the likelihood-ratio approach can be applied only to three or more sampled time points, in which case we find the maximum-likelihood parameter values using the Nelder–Mead simplex method (Nelder and Mead 1965) implemented in the Gnu Scientific Library (GSL) package. We limit the search to the interval $[-2, 2]$ for s (although in practice $|s| \ll 1$) and to the interval $[10^{-1}, 10^8]$ for N, and we allow a maximum of 3×10^4 function evaluations.

Table 1 Accuracy of the χ^2-based *P*-value in estimating the probability of type I error in the likelihood-ratio test

	Sampling parameters					α			
N	T	L	Δ	Absorption probability	$\check{N}/N$	0.05	0.01	0.001	0.0001
10^4	10	10	1	2.6×10^{-3}	1.3	1.4	1.7	2.3	3.3
10^4	100	10	10	7.9×10^{-4}	1.3	1.4	1.7	2.3	2.9
10^4	1000	10	100	1.2×10^{-3}	1.3	1.6	2.4	4.3	8.1
10^3	100	10	10	2.2×10^{-3}	1.3	1.7	2.5	4.5	8.3
10^4	1000	10	100	1.2×10^{-3}	1.3	1.6	2.4	4.3	8.1
10^5	10000	10	1000	1.3×10^{-2}	1.3	1.2	1.5	2.3	3.6
10^4	100	5	20	7.7×10^{-4}	1.7	2.0	2.9	5.2	8.5
10^4	100	10	10	7.9×10^{-4}	1.3	1.4	1.7	2.3	2.9
10^4	100	100	1	8.0×10^{-4}	1.0	1.1	1.1	1.2	1.4

Columns 1–4 show simulation and sampling parameters (see text for notations). Column 5 shows the probability that the allele fixes or goes extinct within the sampling period. Column 6 shows the ratio of the population size most likely under the neutral null hypothesis $\check{N}$ to the true population size N. Columns 7–10 show the ratio of the true fraction of false positives in the likelihood-ratio test to the fraction α expected under the assumption that the LRS is distributed as χ^2 with 1 d.f., across a range of α-values. We performed 10^6 neutral Wright–Fisher simulations with the initial allele frequency $\nu_0 = 0.5$. See Table S1 for results for other initial frequencies. Simulations with absorption events were excluded from the analysis.

Even though the frequency increment test described below does not rely on the calculation of $\hat{N}$ and $\hat{s}$, it too can be applied only when three or more sampled time points are available, for the same conceptual reason as described above. Mathematically, when only one frequency increment is observed, the variance of the distribution of increments cannot be estimated and the *t*-statistic cannot be computed.

Results and Discussion

We consider the problem of determining whether selection has played a role in shaping the fluctuations in the observed frequencies of an allele in a population sampled over time. Suppose that at each time point t_i $(0 < t_1 < \ldots < t_L)$ we sample a diallelic locus in n_i individuals from a given population of an unknown size N and observe that b_i individuals carry allele A and $1 - b_i$ individuals carry allele a. Thus, we observe sampled allele frequencies $\nu_0 = b_0/n_0$, $\nu_1 = b_1/n_1, \ldots, \nu_L = b_L/n_L$. We ask whether genetic drift and sampling noise alone are sufficient to explain the fluctuations in the sampled allele frequencies or whether these frequency changes implicate the action of natural selection at either the specified locus or another completely linked locus. Initially, we treat this problem while neglecting sampling noise. That is, we initially assume that $n_i \gg 1$, $1 \ll b_i \ll n_i$ for all i, so that the sampled allele frequencies ν_i accurately represent the actual frequencies in the entire population. We later investigate how sampling noise affects our conclusions.

We approach the problem using the standard likelihood-ratio test. Following Bollback *et al.* (2008), we consider a pair of nested hypotheses. Under the neutral null hypothesis, changes in the allele frequency are caused only by genetic drift; *i.e.*, the selection coefficient s of allele A is assumed to be zero. Under the alternative hypothesis there is no restriction on s. In both cases, allele a is not under selection. We calculate the likelihoods of the allele-frequency time series under each of these hypotheses, compute the likelihood-ratio statistic (LRS), and reject neutrality if the LRS falls in the tail of the χ^2-distribution with 1 d.f. Because the LRS need not be χ^2-distributed when the number of data points is small, we first report comparisons between the χ^2-distribution and the true distribution of the LRS, for a range of sample sizes. To do this, we simulate samples from the neutral Wright–Fisher process and report whether the probability of type I error in the χ^2-test is accurately predicted by the associated χ^2 *P*-value.

Likelihood of time-series data and the likelihood-ratio statistic

Under standard single-locus population-genetic models, the dynamics of an allele with selection coefficient s in a population are described by a Markov process that specifies the transition probability $P_s(x, t|x', t')$ that the allele frequency is x at time t, given that it was x' at some previous time t'. In addition to the selection coefficient s, this transition probability depends also on the population size N and possibly on other nuisance parameters (Ewens 2004). Ignoring sampling noise, the likelihood of observing allele frequencies $\nu_0, \nu_1, \ldots, \nu_L$ at times $0, t_1, \ldots, t_L$ is

$$L(\mathrm{Data}; N, s) = U(\nu_0) \prod_{i=1}^{L} P_s(\nu_i, t_i | \nu_{i-1}, t_{i-1}), \qquad (9)$$

and, under the neutral null hypothesis,

$$L(\mathrm{Data}; N, 0) = U(\nu_0) \prod_{i=1}^{L} P_0(\nu_i, t_i | \nu_{i-1}, t_{i-1}) \ . \qquad (10)$$

Here $U(x)$ denotes the probability of observing allele frequency x at time point 0, which for simplicity we set to be uniform on the interval (0, 1); *i.e.*, $U(x) \equiv 1$.

Computing likelihoods (9) and (10) is nontrivial even for the standard Wright–Fisher process, because no exact analytical expression for the transition probability $P_s(x, t|x', t')$ exists, and approximate numerical procedures, based on the diffusion equation (Kimura 1955a,b; Evans *et al.* 2007; Bollback *et al.* 2008; Song and Steinrücken 2012), are difficult to implement or computationally intensive. Since our investigation requires us to evaluate the likelihood function millions of times, we desire a fast algorithm for evaluating expressions (9) and (10). Therefore, we choose to compute these likelihoods, using analytical expressions obtained under the Gaussian approximation of the Wright–Fisher process, as described in *Materials and Methods* and in the *Appendix*. Because the Gaussian approximation is accurate only when the allele frequency is far from 0 or 1, our results are restricted to time-series data that lack absorption events.

Given an algorithm for computing expressions (9) and (10), we find the parameter values $\hat{N}$ and $\hat{s}$ that maximize the likelihood function (9) and the value $\check{N}$ that maximizes the likelihood function (10), and we compute the ratio

$$R(\text{Data}) = 2\,\log\left(\frac{L\left(\text{Data}; \hat{N}, \hat{s}\right)}{L\left(\text{Data}; \check{N}, 0\right)}\right). \tag{11}$$

Note that the likelihood-ratio statistic can be obtained only if the number of sampled time points is three or more, as explained in *Materials and Methods*.

If our null hypothesis were *simple*, *i.e.*, if the null distributions of the observed random variables did not depend on any free parameters, the Neyman–Pearson lemma would guarantee that the LRS defines the most powerful test of a given size for rejecting such a null hypothesis (Stuart *et al.* 2009, Chap. 20). In other words, the Neyman–Pearson lemma instructs us to reject the null hypothesis whenever $R(\text{Data}) > \varkappa_\alpha$, choosing $\varkappa_\alpha$ so that the probability of a type I error is α. This test is guaranteed to have the lowest probability of type II error among all tests that have the same probability of type I error, α.

In our case, however, the null hypothesis is *composite*; *i.e.*, the distributions of allele frequencies depend on a parameter, N, whose value is unknown. This implies that the distribution of the LRS under the null hypothesis is unspecified. Thus, not only is the likelihood-ratio test not guaranteed to be the most powerful, but also there is no general way of determining the critical regions for the LRS distribution. The standard way to circumvent the latter problem is to use the asymptotic distribution for the LRS. When the number of data points approaches infinity, the LRS distribution converges to the χ^2-distribution (in this case, with 1 d.f.), under appropriate regularity assumptions (Wilks 1938). This approach has been previously used in the context of allelic time series by Bollback *et al.* (2008). It is worth noting that, although the allele frequencies sampled at successive time points are not independent, the allele frequencies at successive time points *conditioned* on the frequencies at preceding time points are independent [this fact is reflected in expressions (9) and (10)], and so the classical convergence results for LRS still hold.

Although the LRS is guaranteed to be asymptotically χ^2-distributed, the rate of convergence to this distribution is $O(1/\sqrt{L})$, where L is the number of sampled time points (Wilks 1938). Therefore, we characterize how well the χ^2-distribution approximates the true distribution of the LRS when the number of data points is finite. This question is important because the use of an incorrect null distribution can result in a test that underestimates the fraction of type I errors and thus erroneously rejects the null hypothesis more often than indicated by its *P*-value.

Table 2 Accuracy of the *t*-distribution-based *P*-value in estimating the probability of type I error in the frequency increment test

	Sampling parameters			α			
N	T	L	Δ	0.05	0.01	0.001	0.0001
10^4	10	10	1	1.00	1.00	0.98	1.02
10^4	100	10	10	1.00	1.00	1.00	1.04
10^4	1,000	10	100	0.96	1.05	1.25	1.37
10^3	100	10	10	0.99	1.08	1.31	1.38
10^4	1,000	10	100	0.96	1.05	1.25	1.37
10^5	10,000	10	1,000	0.73	0.68	0.65	0.62
10^4	100	5	20	1.00	1.02	1.02	1.09
10^4	100	10	10	1.00	1.00	1.00	1.04
10^4	100	100	1	0.99	0.99	1.00	0.96

Columns 1–4 show simulation and sampling parameters (see text for notations). Columns 5–8 show the ratio of the true fraction of false positives in the FIT to the fraction α expected under the assumption that the frequency increment statistic is distributed according to Student's *t*-distribution with $L - 1$ d.f., across a range of α-values. We performed 10^6 neutral Wright–Fisher simulations with the initial allele frequency $\nu_0 = 0.5$. Unlike our implementations of the LRT and the ELRT, the FIT can formally be applied in cases when the observed allele is either fixed or lost at the last sampled time point. Thus, the fraction of simulations in which absorption events prevented us from applying the FIT was $<10^{-5}$.

Likelihood-ratio statistic is not χ^2-distributed for finite data

With this goal in mind, we simulated the neutral two-allele Wright–Fisher model with population size N, without mutation, with allele A initiated at 10%, 20%, 30%, 40%, or 50% of the population. We recorded the frequency of allele A every generation, for T generations. To ensure that absorption events are rare within the sampling period we set $T \leq N/10$. We then produced a data set consisting of these frequencies sampled every Δ generations. We sampled a total of $L + 1$ time points, so that $\Delta = T/L$. For each population size N, we simulated 10^6 allele-frequency trajectories, sampled allele frequencies from these trajectories using various combinations of T and L, and computed the LRS for each of the sampled time series. Thus, for each combination of N, L, and T we obtained the true distribution of the LRS under the neutral null hypothesis. We compared this distribution with the χ^2-distribution with 1 d.f. in two ways. First, we calculated the probabilities for the LRS to fall into each of the 20 vigintiles (quantiles of size 0.05) of the χ^2-distribution. Second, we computed the probability of type I error of the χ^2-based test for a range of nominal *P*-values α.

The results of these analyses are shown in Figure 1, Supporting Information, Figure S1, Table 1, and Table S1. Figure 1 and Figure S1 demonstrate that the χ^2-distribution is a poor approximation for the true distribution of the LRS

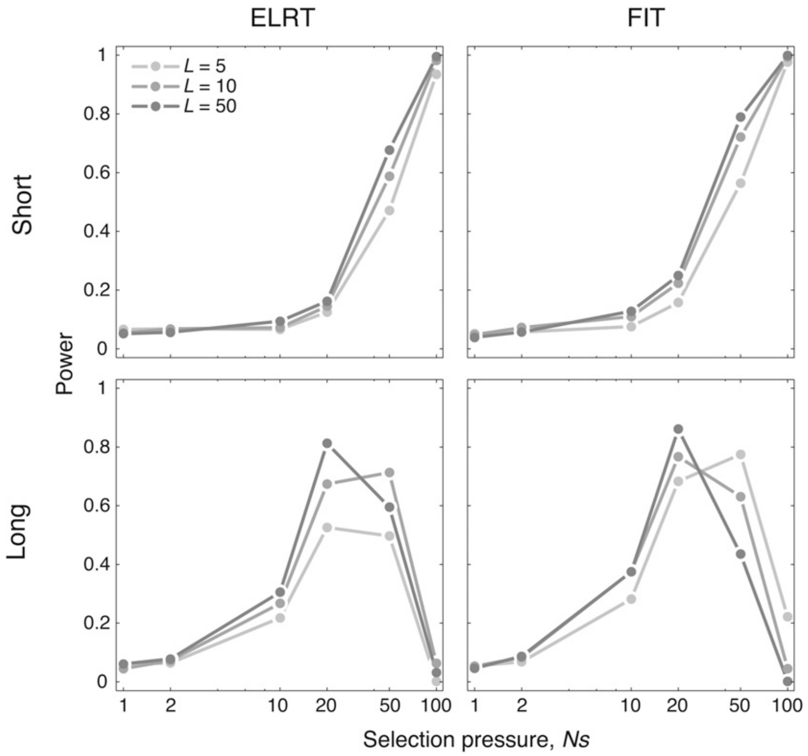

Figure 2 Power of the ELRT and the FIT to detect selection of different strength. Power is reported as the fraction of trial data sets generated by the Wright–Fisher model with selection for which the ELRT (left column) or the FIT (right column) rejects the neutral null hypothesis at *P*-value $\alpha = 0.05$ in short ($T = 0.01N$, top row) and "long" ($T = 0.1N$, bottom row) time series. Both tests gain power with increasing selection pressure, but in long time series they start to lose power when selection becomes very strong (see text for details). Power of both tests grows weakly with the number of sampled time points, *L*. We ran 10^3 trials with $N = 10^4$ and initial allele frequency $\nu_0 = 0.5$. Trials that produced absorption events within the sampling period were discarded.

under neutrality when the number of sampled time points is finite. If the LRS under the neutral null hypothesis followed the χ^2-distribution, then the probability for the LRS to fall into each vigintile of the χ^2-distribution would equal 0.05. Instead, the LRS more often falls in the top vigintiles of the χ^2-distribution and, correspondingly, less often in the bottom vigintiles. This fact is problematic because it implies that the *P*-values calculated from the χ^2-distribution will underestimate the probability of type I error. Table 1 and Table S1 show that this is indeed the case, even when as many as 100 time points are sampled. While the discrepancy between the actual probability of false positives and the χ^2-based *P*-value is moderate (less than a factor of 2) for relatively high nominal *P*-values (*e.g.*, >1%), the discrepancy becomes increasingly more severe for stringent *P*-values, so that in some regimes the χ^2-test rejects neutrality 50 times as often as it should (see Table S1).

The classical result of Wilks (1938) guarantees that the LRS distribution will converge to the χ^2-distribution as the number of data points increases. In our case, the LRS distribution should converge to the χ^2-distribution with 1 d.f. as the number of sampled points *L* increases (and Δ decreases), while the time-series length *T* remains constant. The χ^2-based *P*-value should likewise converge to the true probability of type I error. As expected, the values in columns 7–10 in Table 1 and in the corresponding columns in Table S1 approach 1 as *L* increases.

In addition to the deviation of the LRS distribution from the χ^2-distribution, the most likely population size under the null hypothesis, $\check{N}$, systematically overestimates the true population size, *N*, especially when the number of data points is small (see Table 1 and Table S1). This phenomenon is consistent with previous reports (Waples 1989; Williamson and Slatkin 1999; Wang 2001). The bias in the inferred population size decreases with increasing number of data points, almost independently of the true population size or the observation time (see Figure S2).

Two alternative tests of selection

The empirical likelihood-ratio test: We propose two approaches to fix the shortcomings of the χ^2-likelihood-ratio test for selection in time series data. The ideal approach would be to obtain the true distribution of the LRS by simulating the neutral Wright–Fisher model with the true population size, *N*. But since we are concerned with the case when *N* is unknown, we propose to use the estimated maximum-likelihood population size under neutrality, $\check{N}$ to obtain the null distribution of the LRS. We call this approach the empirical likelihood-ratio test (ELRT).

Figure 1 shows that the LRS distribution generated under $\check{N}$ is an excellent approximation to the true LRS distribution, even when the number of sampled time points is small. As a result, the *P*-values computed with the empirical LRS distribution provide an accurate description of the false-positive rate. Nevertheless, the ELRT approach suffers from two drawbacks, at least in its simplest implementation. First, the Gaussian approximation that we employed to calculate the likelihoods becomes problematic in cases when the observed allele-frequency changes are large (for example, if the allele is under very strong selection). Large changes in allele frequency lead to small $\check{N}$, which leads to a high probability of absorption events in neutral simulations, and the

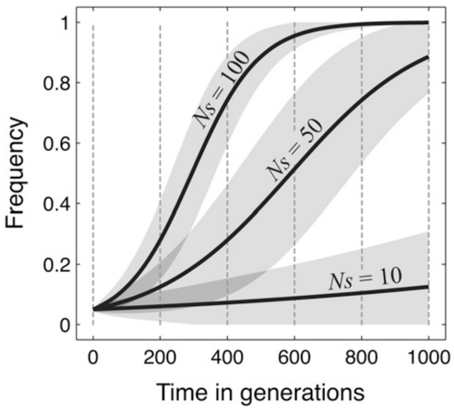

Figure 3 Schematic diagram describing the power of any test for selection in allele-frequency time-series data. Thick solid lines show the expected frequency dynamics (Equation 3) of alleles with selection coefficients s = 0.001, 0.005, 0.01, initiated at frequency $x_0 = 0.05$. Shaded areas denote $\pm\sqrt{\sigma^2(t, x_0)/N}$, where σ^2 is given by Equation 6 and $N = 10^4$, which illustrate the size of stochastic fluctuations around the expected frequency. Vertical dashed shaded lines show hypothetical sampling time points. When the selection coefficient is low (Ns = 10), stochastic fluctuations dominate, and tests of selection have low power. When the selection coefficient is high (Ns = 100), fixation events occur within the sampling interval and some sampling points (at 800 and 1000 generations) become uninformative, which also leads to loss of power. For a given sampling interval T power is maximized for intermediate selection coefficients (Ns = 50).

Gaussian approximation becomes inaccurate. Thus, somewhat paradoxically, we expect the ELRT based on the Gaussian approximation to lose power when the data come from populations under very strong selection. This problem is not intrinsic to the ELRT method, and indeed it could be remedied by calculating likelihoods using the (computationally intensive) diffusion approximation. The second drawback of the ELRT is that it is computationally intensive, even when using the fast Gaussian approximation for likelihoods. In particular, to obtain the approximate empirical LRS distribution, many Wright–Fisher simulations must be performed, each accompanied by the calculation of the LRS.

In the next section we propose another alternative to the χ^2-likelihood-ratio test that is computationally inexpensive, but somewhat less accurate than the ELRT.

The frequency increment test: We define the rescaled allele-frequency increments as

$$Y_i = \frac{\nu_i - \nu_{i-1}}{\sqrt{2\nu_{i-1}(1-\nu_{i-1})(t_i - t_{i-1})}}, \quad i = 1, 2, \ldots, L.$$

Since under the neutral null hypothesis the allele frequency ν behaves, away from the boundaries 0 or 1, approximately as Brownian motion (see Ewens 2004 and the *Appendix*), the random variables Y_i are independent and approximately normally distributed with mean 0 and variance $1/N$ (see Equations 7 and 8). Under the alternative hypothesis, Y_i are also independent and approximately normally distributed, but with a nonzero mean and a different variance (see Equations 4–6). Thus, the problem of testing whether serial data come from a neutral population reduces to the problem of testing whether the rescaled allele-frequency increments come from a normal distribution with mean zero (and unknown variance). The latter problem is one the most classical problems in statistics, and it has a well-known and elegant solution: the t-test. The frequency increment statistic (FIS), defined as

$$t_{\mathrm{FI}}(\mathrm{Data}) = \frac{\overline{Y}}{\sqrt{S^2/L}}, \tag{12}$$

where $\overline{Y}$ and S are the sample mean and the sample variance

$$\overline{Y} = \frac{1}{L}\sum_{i=1}^{L} Y_i \quad \text{and} \quad S^2 = \frac{1}{L-1}\sum_{i=1}^{L}(Y_i - \overline{Y})^2,$$

is distributed according to Student's t-distribution with $L - 1$ d.f., under the neutral null hypothesis. Note that the unknown nuisance parameter, N, in the population-genetic problem corresponds to the unknown variance in the t-test. We call this test the frequency increment test (FIT). In addition to being simple and computationally trivial, this test is also *the most powerful similar test* (see Stuart *et al.* 2009, Chap. 21) of the selection hypothesis against the neutral null hypothesis, provided frequencies are far from the boundaries 0 and 1.

Figure 1 and Table 2 show that the FIT substantially outperforms the χ^2-likelihood-ratio test, in the sense that the nominal FIT P-value represents the probability of a type I error more accurately than does the χ^2-based P-value. Nevertheless, the FIT P-value is not exact. Under most parameter regimes where the probability of type I error deviates from the P-value reported by the t-distribution, the FIT appears to be overly conservative (*i.e.*, the P-value overestimates the probability of type I error), but how precisely this depends on N, L, and Δ is complicated (Table 2). In any case, the inaccuracies in the probability of type I error under the FIT are an order of magnitude smaller than those under the χ^2-LRT, in all parameter regimes tested.

Power of the ELRT and the FIT to detect selection

Next we determined the power of the ELRT and the FIT to detect selection in allele-frequency data, in terms of the strength of selection, time-series length, and sampling frequency. To this end, we ran Wright–Fisher simulations with population size $N = 10^4$ as described above, but now with allele A possessing selective advantage s. For each value of the scaled selection coefficient Ns ranging from 1 to 100 we simulated 10^4 allele-frequency trajectories and sampled from them in the first $T = N/100 = 100$ or in the first $T = N/10 = 1000$ generations. These two sampling schemes gave rise to the "short" and "long" allele-frequency time series. For each time series, we sampled the frequencies of the selected allele at $L + 1$ time points equally spaced Δ generations apart, with L taking values 5, 10, and 50. For

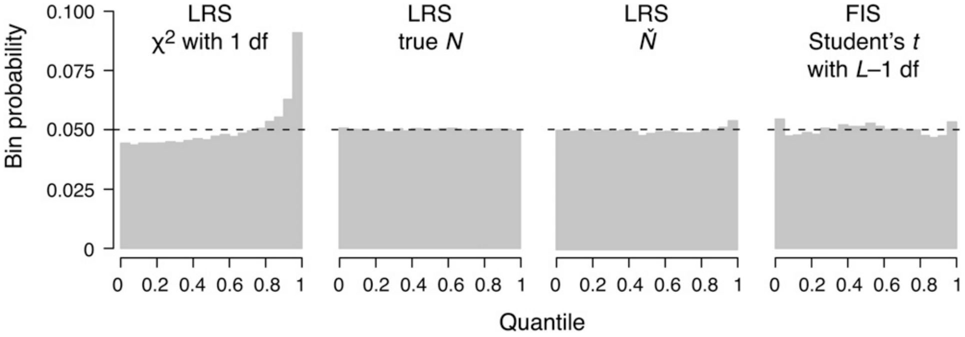

Figure 4 Distributions of test statistics under the neutral null hypothesis, when allele frequencies are sampled with noise. Histograms show the probabilities that the value of a test statistic generated under the neutral null hypothesis falls within each of the vigintiles (quantiles of size 0.05) of another, approximate, distribution. Notations are as in Figure 1. Parameter values were $N = 10^3$, $T = 100$, $\Delta = 20$, $L = 5$, $\nu_0 = 0.5$, and $n = 500$; the number of Wright–Fisher simulations was 2×10^5.

each combination of s, T, and L, we performed the ELRT and the FIT, and we calculated the frequency with which they reject neutrality at P-value $= 0.05$. When computing the null distribution of the LRS in the ELRT, we encountered some neutral Wright–Fisher trials that exhibited an absorption event during the observation period T. Instead of discarding such trials, we include them into the estimation of the empirical LRS distribution by conservatively assigning them to the maximum LRS value of the neutral trials.

Figure 2 shows that both tests possess substantial power to detect moderate to strong selection ($Ns > 10$), but they lose power when selection is very strong. As illustrated in Figure 3, such behavior is expected for any test of selection from time-series data. Consider a fixed sampling duration T. Clearly, if selection is very weak, it will not be able to change the allele frequency substantially during this time interval, and so the observed allele-frequency changes will be dominated by noise. On the other hand, when selection is very strong, the allele will go to fixation within the interval T, and so some of the samples in the later part of the interval will carry no information about the allele dynamics. For example, in Figure 3, an allele with selection coefficient $Ns = 100$ typically fixes in <800 generations, and so samples taken after generation 800 are uninformative. In the extreme case of very strong selection, the allele will fix between the first and second sampling time points. In this case, without knowledge of the population size, we could not determine whether the time series was caused by strong selection or strong genetic drift. Thus, any test of selection based on time series data will lose power for either very weak or very strong selection pressures.

The intuition outlined above suggests that a given sampling interval T sets the scale for selection coefficients that we have power to detect, $s_{\text{power}}(T)$. We can estimate $s_{\text{power}}(T)$ by inverting the logic of this intuition: for selection strength s, there is an optimal sampling interval that maximizes the power of tests to detect this selection. Such a sampling interval should be long enough for selection to substantially change the allele frequency but short enough to avoid fixation. From Equation 3, the expected time $t(x_f, x_0; s)$ it takes for an allele with selection coefficient s to reach frequency x_f from the initial frequency x_0 is approximately given by

$$t(x_f, x_0; s) = \frac{1}{s}\ln\left(\frac{x_f}{1-x_f}\frac{1-x_0}{x_0}\right).$$

Setting $t(x_f, x_0; s_{\text{power}}) = T$ with some arbitrary x_f close to 1, we predict that tests of selection in a time series of length T will have the maximal power to detect selection coefficients on the order of

$$s_{\text{power}}(T) = \frac{1}{T}\ln\left(\frac{x_f}{1-x_f}\frac{1-x_0}{x_0}\right).$$

Setting $x_0 = 0.5$ as in our simulations and $x_f = 0.95$ (this choice is arbitrary and not critical for determining the order of magnitude of s_{power}), we predict that tests of selection will have maximal power to detect selection of strength $s_{\text{power}} = 0.029$ in time series of length $T = 100$ generations and $s_{\text{power}} = 0.0029$ in time series of length $T = 1000$ generations. For a population of size $N = 10^4$, this translates into $Ns_{\text{power}} = 290$ and $Ns_{\text{power}} = 29$, respectively, which is consistent with our numerical results (Figure 2). These power calculations are generic properties of any test of selection in time-series data.

The ELRT and the FIT have an additional complication in that they cannot be applied to data points after an absorption event. In plotting Figure 2, we discarded all trials in which an absorption event occurred within the sampling period, even though some of these trials likely had a detectable signature of selection prior to the absorption event. Thus, Figure 2 shows the lower bound on the power of our tests.

Aside from these gross properties of power, we found that the FIT has slightly more power than the ELRT and that power of both tests increases weakly with the number of sampled time points L, with all other parameters being equal.

The effects of noisy sampling

So far we have studied tests of selection, assuming that allele frequencies are measured with perfect accuracy in successive time points. In this section, we investigate the behavior of the FIT and the ELRT in a more realistic situation—when allele frequencies are estimated, at each time point, by sampling a limited number of individuals from the population and typing them with respect to the focal locus. To study this, we used the same simulated

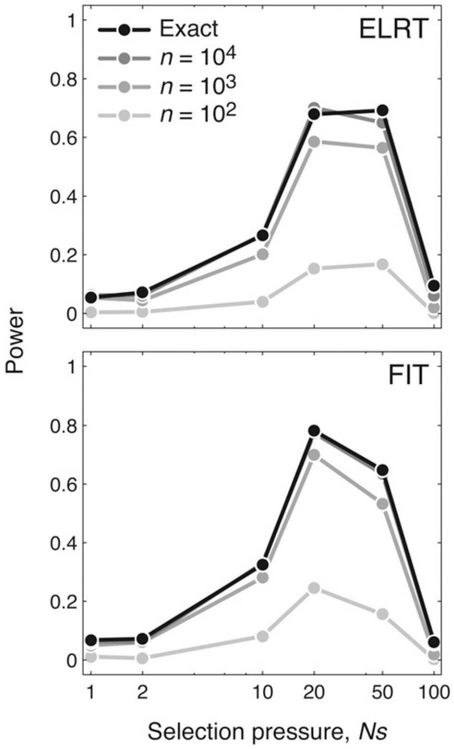

Figure 5 Power of the ELRT and the FIT to detect selection of different strengths, under various sampling regimes. Parameter values were $N = 10^4$, $T = 1000$, $\Delta = 100$, $L = 10$, and $\nu_0 = 0.5$; the number of Wright–Fisher simulations was 10^3.

time-series trajectories as in previous sections, but instead of analyzing the true allele frequencies, x, we drew binomial random variables with sample size n and success probability x to obtain the sampled allele frequencies ν. We then analyzed the test size and power, treating the sampled allele frequencies ν as the data.

As shown in Figure 4, when sample sizes are sufficiently large ($n = 500$), the *P*-values produced by the ELRT and the FIT remain accurate representations of the true type I error probability. When the sample sizes become too small ($n \leq 100$), both tests become overly conservative; *i.e.*, the *P*-values produced by the ELRT and the FIT overestimate the probability of type I error (see Figure S3). Note that the LRT also becomes overly conservative in this regime, even if the χ^2-distribution or the distribution of the LRS under true N is used (Figure S3). This in itself is not problematic and it simply implies that the *P*-values from such tests should be viewed as upper bounds on the actual probability of type I error. More problematic is the associated decline in power of both tests as samples size n decreases (Figure 5). The dependence of power on the strength of selection in the presence of sampling noise remains the same as in the absence of sampling noise, with the power curves shifted downward (Figure 5).

Applications to empirical data

In this section, we apply our tests of selection to allele-frequency time series from three previously published experimental data sets, as well as some additional new experimental data.

Bacteriophage evolved at high temperature: The first data set is from an experiment described by Bollback and Huelsenbeck (2007). Bollback and Huelsenbeck (2007) evolved three lines of bacteriophage MS2, which infects *Escherichia coli*, at increasingly high temperatures, from 39° to 43°. After 50 passages, each corresponding to approximately three bursts, they identified mutations that were segregating in the populations and determined the frequencies of these mutations at the previous time points. From this data set we selected allele-frequency trajectories that remained at intermediate frequencies between 0 and 1 for at least two consecutive time points and applied the FIT, but not the ELRT (Table 3). We could not apply the ELRT to these data for two reasons. First, some time series had only two time points at which the mutant allele was at intermediate frequencies. The maximum-likelihood approaches cannot estimate both N and s in such cases (see *Materials and Methods*). Second, the frequencies of the remaining alleles changed so fast (*e.g.*, from 30% to 90% in 10 passages) that the maximum-likelihood (ML)-estimated population sizes under neutrality, $\check{N}$, were very small (see Table 3), and so neutral simulations were dominated by absorption events.

When we applied the FIT to these data, we found that only one time series produced a significant *P*-value (mutation C3224U in line 3), despite the fact that most of the identified mutations are likely to be beneficial. The poor performance of our tests on these data are expected for two reasons. First, the sample sizes in these data set are very small ($n \leq 10$), and we expect our tests to have very low power. Second, even though all mutations are probably beneficial, not all frequency trajectories are monotonically increasing, and some of them are even decreasing (*e.g.*, mutation C1549U/A in line 3), presumably due to clonal interference (Gerrish and Lenski 1998), which further reduces the power of our test.

Deep population sequencing of adapting yeast populations: The second data set we analyzed is from an experiment in which Lang *et al.* (2011) evolved 592 populations of the yeast *Saccharomyces cerevisiae* in rich medium for 1000 generations. The original experiment tracked the appearance and fate of sterile mutations that are known to be beneficial under the chosen experimental conditions (Lang *et al.* 2011). Subsequently, some of these populations were deep sequenced, and many other adaptive mutations were identified (Lang *et al.* 2013). From this large data set, we selected three allele-frequency trajectories of mutations in genes *STE11*, *IRA1*, and *IRA2* that arose in three different populations (Figure 6, Table S2). Applying the ELRT and the FIT to these time series, we found that our tests return best results when used on subsets of each time series (Figure 6, Table S2). Based on these truncated time series, both the ELRT and the FIT identified that the trajectories of the mutant *STE11* and *IRA1* alleles, but not that of the mutant *IRA2* allele, were positively

Table 3 Mutant allele frequencies in bacteriophage data from Bollback *et al.* (2007) and the application of the ELRT and the FIT

		Passage[a]								ELRT[b]			FIT		
Line	Mutation	0	10	15	20	25	30	40	50	$\check{N}$	$\hat{N}$	$\hat{s}$	t_{FI}	d.f.	*P*
1	C206U	0/10	0/10	NA	**1/10**	NA	**3/10**	**10/10**	10/10	NA	NA	NA	2.55	1	0.119
	C1549U/A	0/10	0/10	NA	0/10	NA	**6/10**	**5/10**	**9/10**	56.5	9.7	0.20	0.59	1	0.330
	G1551A	0/10	0/10	NA	0/10	NA	**1/10**	**5/10**	**1/10**	11.3	4.2	−0.01	0.25	1	0.422
2	C206U	0/10	0/9	**1/9**	**1/10**	NA	**3/10**	**9/10**	**9/10**	19.7	99.2	0.16	1.51	3	0.114
	C1549U/A	0/10	0/9	0/10	**9/10**	NA	**6/9**	**10/10**	8/10	NA	NA	NA	−0.10	1	0.532
	G1551A	0/10	0/9	0/10	**1/10**	NA	**2/9**	**0/10**	2/10	NA	NA	NA	−0.14	1	0.545
	U466C	0/10	**2/9**	**3/9**	**10/10**	NA	10/10	10/10	10/10	NA	NA	NA	1.29	1	0.210
3	C206U	0/10	0/10	NA	**1/10**	**5/10**	**4/10**	**7/10**	**10/10**	12.6	30.3	0.21	1.61	3	0.103
	C1549U/A	0/10	0/10	NA	0/10	**5/10**	**5/10**	**3/10**	**1/10**	187.5	405.4	−0.09	−1.99	2	0.908
	C3224U	0/10	0/10	NA	0/10	0/10	**3/10**	**7/10**	**10/10**	NA	NA	NA	7.00	1	0.045
	C3220U	0/10	0/10	NA	0/10	0/10	**3/10**	**8/10**	**10/10**	NA	NA	NA	2.69	1	0.113
	G1551A	0/10	0/10	NA	0/10	0/10	**2/10**	**7/10**	**9/10**	22.1	652.4	0.19	2.07	1	0.143
	U466C	0/10	0/10	NA	**4/10**	**5/10**	**9/10**	**10/10**	10/10	28.2	39.2	0.33	2.13	2	0.083

[a] Estimated mutant allele frequencies at different time points (passages) are shown. Data points used for the ELRT and the FIT are in boldface type (if the last of these data points was at frequency 0 or 1, it was not used for the ELRT).

[b] ML parameter values are shown. $\hat{N}$ and $\hat{s}$ maximize likelihood (9) and $\check{N}$ maximizes likelihood (10) under the Gaussian approximation. Note that since all $\check{N}$ are very small, the ELRT *P*-values cannot be obtained due to frequent absorption events in neutral Wright–Fisher simulations.

selected. Given the knowledge that the experimental population sizes exceed 10^4 individuals and the fact that mutations in genes *STE11*, *IRA1*, and *IRA2* independently arose and spread in several parallel lines, it is likely that all three mutations are in fact beneficial (Lang *et al.* 2013). Our tests do not take these two critical pieces of information into account, but they are still able to identify the action of positive selection in two of three cases, based solely on allele frequencies estimated from samples of size $n \leq 150$.

Yeast populations evolved at different population sizes: The third data set we analyzed is from an experiment performed by one of us (S. Kryazhimskiy) and described in Kryazhimskiy *et al.* (2012). In this experiment, 1008 populations of the yeast *S. cerevisiae* were evolved under conditions similar to those in the experiment by Lang *et al.* (2011), but under various population sizes and migration regimes. After 500 generations of evolution, fitnesses of these populations were measured in a competition experiment. Fitness data for 976 of these populations were previously described in Kryazhimskiy *et al.* (2012). Here, we analyzed the published competition assays from 736 well-mixed (WM) populations, referred to as "No", "Small WM", and "Large WM", as well as unpublished data from an additional 32 well-mixed populations of intermediate size referred to as "Medium WM". All these populations were evolved in exactly identical conditions, except for the serial transfer bottleneck size. In particular, the bottleneck size was $\sim 10^3$ individuals in the No populations and 5, 10, and 20 times larger than that in Small WM, Medium WM, and Large WM, respectively. The fitnesses of all populations were measured in competition assays with at least threefold replication. As described in Kryazhimskiy *et al.* (2012), each competition assay consists of measuring the frequency of the evolved population relative to a fluorescently labeled reference strain at two time points. The raw flow cytometry counts for all populations used here (including those published previously) are reported in Table S4.

As mentioned in *Materials and Methods*, when the time series contains only two time points, there is not enough information to estimate the population size (or, equivalently, the variance of the distribution of frequency increments).

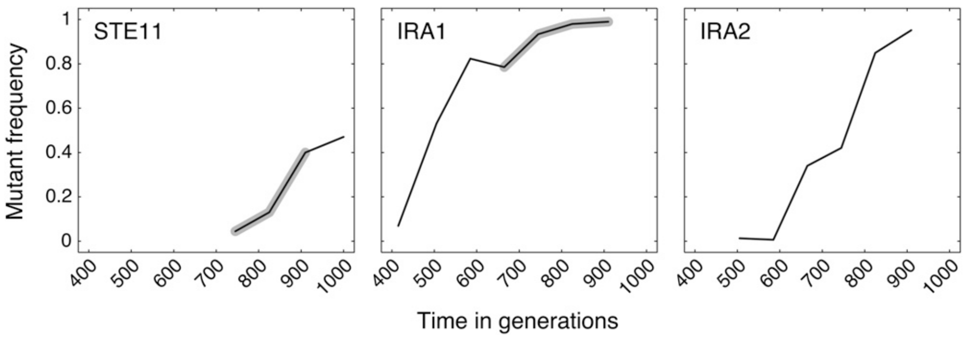

Figure 6 Application of the ELRT and the FIT to allele-frequency time series from Lang *et al.* (2011, 2013). Each panel shows the estimated frequency of a mutant allele in the long-term evolution lines described in Lang *et al.* (2011, 2013): the left panel shows the frequency of mutation D579Y in gene *STE11* in population RMB2-F01; the center panel shows the frequency of mutation Y822* in gene *IRA1* in population RMS1-D12; and the right panel shows the frequency of mutation A2698T in gene *IRA2* in population BYS2-D06. Shading highlights the data points for which the FIT and the ELRT identify selection.

Table 4 Application of the FIT to yeast data from Kryazhimskiy *et al.* (2012)

			α = 0.05		α = 0.01		α = 0.001	
Treatment	Bottleneck size	No. populations	Obs[a]	Exp[b]	Obs[a]	Exp[b]	Obs[a]	Exp[b]
No	10^3	639	218	32.0	74	6.4	7	0.64
Small WM	5×10^3	64	61	3.2	53	0.6	34	0.06
Medium WM	10^4	32	32	1.6	32	0.3	32	0.03
Large WM	2×10^4	32	32	1.6	32	0.3	32	0.03

[a] Number of populations for which the FIT rejects the neutral null hypothesis at the given *P*-value threshold α.

[b] Expected number of false positives at the given *P*-value threshold α.

However, the FIT can be easily applied to frequency increments pooled across replicate fitness measurements. In particular, if ν_{ki} is the frequency of the evolved population at time point i (with $t_0 = 0$ and $t_1 = 20$) in replicate assay k (with $k = 1, \ldots, K$), then we define the frequency increment in replicate k as

$$Y_k = \frac{\nu_{k1} - \nu_{k0}}{\sqrt{2\nu_{k0}(1 - \nu_{k0})(t_1 - t_0)}}$$

and calculate the frequency increment statistic according to Equation 12, with L replaced by the number of replicates K.

The results of the FIT applied to these data are reported in Table S3 and summarized in Table 4. We find that the FIT rejects the neutral null hypothesis at various stringency cutoffs for all Medium WM and Large WM populations and for the majority of Small WM populations. At the same time, the FIT rejects neutrality for only ~34% of the No populations at the *P*-value cutoff of 0.05 and only ~1% of the No populations at the *P*-value cutoff of 0.001. In both of these cases the observed numbers of positives significantly exceed the numbers of false positives expected due to multiple testing. These results demonstrate that the FIT reliably detects the action of natural selection in data from microbial evolution experiments. Moreover, since we do not know which populations truly adapted in this experiment, these results inform us that, when the bottleneck size exceeds 5000 individuals, nearly all populations undergo significant adaption during 500 generations of evolution, but when the bottleneck size is 1000, only ~34% of populations do so. These results are consistent with the expectation that larger populations adapt faster and suffer less from the accumulation of deleterious mutations, compared to small populations.

Conclusions

We have shown that the standard χ^2-based test for selection in time series of allele frequencies (Bollback *et al.* 2008) is subject to a greatly elevated false discovery rate in the practical regime of relatively few sampled time points. As a result of this bias, the χ^2-LRT is not a reliable test for selection in many practical time series, because its *P*-value underestimates the rate of false positives, especially when the allele frequencies are measured accurately. We proposed two new tests to address this problem, and we showed that both of them accurately estimate the probability of type I error and have power to detect selection in parameter regimes that are reasonable for many evolution experiments and natural populations.

Our tests were initially developed under the assumption that sampling noise is negligible and that the estimated allele frequencies can be treated as exact. In many situations, such as microbial laboratory experiments, this assumption is not restrictive. Indeed, when allele frequencies are measured with high-throughput methods such as flow cytometry (Lang *et al.* 2011; Kryazhimskiy *et al.* 2012) or deep population sequencing (Smith *et al.* 2011; Lang *et al.* 2013), the sample sizes often exceeds the population size. On the other hand, when samples are derived from natural populations, this assumption is likely to be violated. In this case, our tests remain conservative, but lose power to detect selection, especially when selection is weak. This is expected because when sampling noise dominates demographic stochasticity, the information about the population size that is contained in allele-frequency fluctuations is lost. In principle, the population size can be inferred even in the presence of high sampling noise, if the time series is long enough. Indeed, if large frequency fluctuations are caused by small population size, time to absorption will be short, but if they are caused by sampling noise, time to absorption will be long. Moreover, incorporating time to absorption into tests of selection in time-series data would alleviate the ascertainment bias that arises when, for example, only those alleles are analyzed that reach sufficiently high frequencies in the population.

The methods proposed here, just as the earlier χ^2-based test, are limited to the regime in which the frequencies of alleles observed at a locus are not influenced by mutations that may arise elsewhere in the genome during the time of observation. Thus, our tests are perhaps most readily applicable to selection scans in full-genome time-series data like those now actively generated in evolution experiments in *Drosophila* (Burke *et al.* 2010; Orozco-terWengel *et al.* 2012). They will also be applicable to asexual organisms when clonal interference is absent or weak, for example in competitive fitness assays (Lenski *et al.* 1991; Gallet *et al.* 2012; Kryazhimskiy *et al.* 2012) or in tracking known polymorphisms in natural populations for a relatively short time (Barrett *et al.* 2008; Winters *et al.* 2012; Pennings *et al.* 2013). By contrast, inferring selection coefficients when allele dynamics are influenced by multiple linked sites is a substantially more difficult problem, which has begun to be addressed elsewhere (Illingworth and Mustonen 2011; Illingworth *et al.* 2012), although not within the same rigorous population-genetic framework that treats all genotypic dynamics stochastically.

Acknowledgments

The authors thank Todd Parsons for help with the Gaussian approximation, Daniel P. Rice for providing experimental data, and two anonymous reviewers for thoughtful comments. S.K. was supported by the Burroughs Wellcome Fund Career Award at Scientific Interface. J.B.P. acknowledges support from the David and Lucile Packard Foundation, the Burroughs Wellcome Fund, the Alfred P. Sloan Foundation, the U S. Army Research Office (W911NF-12-1-0552), and the Foundational Questions in Evolutionary Biology Fund (RFP-12-16).

Literature Cited

Barrett, R. D. H., S. M. Rogers, and D. Schluter, 2008 Natural selection on a major armor gene in threespine stickleback. Science 322: 255–257.

Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh *et al.*, 2009 Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243–1247.

Bollback, J. P., and J. P. Huelsenbeck, 2007 Clonal interference is alleviated by high mutation rates in large populations. Mol. Biol. Evol. 24: 1397–1406.

Bollback, J. P., T. L. York, and R. Nielsen, 2008 Estimation of $2N_es$ from temporal allele frequency data. Genetics 179: 497–502.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083.

Burke, M. K., J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose *et al.*, 2010 Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature 467: 587–590.

Daniels, R., H.-H. Chang, P. D. Séne, D. C. Park, D. E. Neafsey *et al.*, 2013 Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. PLoS ONE 8: e60780.

Denef, V. J., and J. F. Banfield, 2012 In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science 336: 462–466.

Desai, M. M., and J. B. Plotkin, 2008 The polymorphism frequency spectrum of finitely many sites under selection. Genetics 180: 2175–2191.

Evans, S. N., Y. Shvets, and M. Slatkin, 2007 Non-equilibrium theory of the allele frequency spectrum. Theor. Popul. Biol. 71: 109–119.

Ewens, W. J., 2004 *Mathematical Population Genetics*. Springer Science + Business Media, New York.

Gallet, R., T. F. Cooper, S. F. Elena, and T. Lenormand, 2012 Measuring selection coefficients below 10^{-3}: method, questions, and prospects. Genetics 190: 175–186.

Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. Genetica 102/103: 127–144.

Hegreness, M., N. Shoresh, D. Hartl, and R. Kishony, 2006 An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311: 1615–1617.

Illingworth, C. J., L. Parts, S. Schiffels, G. Liti, and V. Mustonen, 2012 Quantifying selection acting on a complex trait using allele frequency time series data. Mol. Biol. Evol. 29: 1187–1197.

Illingworth, C. J. R., and V. Mustonen, 2011 Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. Genetics 189: 989–1000.

Kimura, M., 1955a Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41: 144–150.

Kimura, M., 1955b Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harb. Symp. Quant. Biol. 20: 33–53.

Kryazhimskiy, S., D. P. Rice, and M. M. Desai, 2012 Population subdivision and adaptation in asexual populations of *Saccharomyces cerevisiae*. Evolution 66: 1931–1941.

Kurtz, T. G., 1970 Solutions of ordinary differential equations as limits of pure jump Markov processes. J. Appl. Probab. 7: 49–58.

Kurtz, T. G., 1971 Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. J. Appl. Probab. 8: 344–356.

Lang, G. I., D. Botstein, and M. M. Desai, 2011 Genetic variation and the fate of beneficial mutations in asexual populations. Genetics 188: 647–661.

Lang, G. I., D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock *et al.*, 2013 Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500: 574.

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler, 1991 Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am. Nat. 138: 1315–1341.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating allele age and selection coefficient from time-serial data. Genetics 192: 599–607.

Maldarelli, F., M. Kearney, S. Palmer, R. Stephens, J. Mican *et al.*, 2013 HIV populations are large and accumulate high genetic diversity in nonlinear fashion. J. Virol. 87: 10313–10323.

Mathieson, I., and G. McVean, 2013 Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 193: 973–983.

Nagylaki, T., 1990 Models and approximations for random genetic drift. Theor. Popul. Biol. 37: 192–212.

Nelder, J. A., and R. Mead, 1965 A simplex method for function minimization. Comput. J. 7: 308–313.

Orozco-terWengel, P., M. Kapun, V. Nolte, R. Kofler, T. Flatt *et al.*, 2012 Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Mol. Ecol. 21: 4931–4941.

Pennings, P., S. Kryazhimskiy, and J. Wakeley, 2014 Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. (in press).

Pollak, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics 104: 531–548.

Pollett, P. K., 1990 On a model for interference between searching insect parasites. J. Austral. Math. Soc. Ser. B 32: 133–150.

Reid, B. J., R. Kostadinov, and C. C. Maley, 2011 New strategies in Barrett's esophagus: integrating clonal evolutionary theory with clinical management. Clin. Cancer Res. 17: 3512–3519.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Smith, A. M., T. Durbic, J. Oh, M. Urbanus, M. Proctor *et al.*, 2011 Competitive genomic screens of barcoded yeast libraries. J. Vis. Exp. 54: e2864.

Song, Y. S., and M. Steinrücken, 2012 A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. Genetics 190: 1117–1129.

Stuart, A., K. Ord, and S. Arnold, 2009 *Kendall's Advanced Theory of Statistics*, *Classical Inference and the Linear Model*. John Wiley & Sons, New York.

Wang, J., 2001 A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet. Res. 78: 243–257.

Waples, R. S., 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121: 379–391.

Wilks, S. S., 1938 The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9: 60–62.

Williamson, E. G., and M. Slatkin, 1999 Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics 152: 755–761.

Winters, M. A., R. M. Lloyd, R. W. Shafer, M. J. Kozal, M. D. Miller *et al.*, 2012 Development of elvitegravir resistance and linkage of integrase inhibitor mutations with protease and reverse transcriptase resistance mutations. PLoS ONE 7: e40514.

Communicating editor: Y. S. Song

Appendix

Gaussian Approximation to the Moran Process

We approximate the continuous-time Moran processes with a combination of a deterministic process and a Gaussian noise process. We follow here the procedure outlined by Pollett (1990), which is based on the results by Kurtz (1970, 1971). The Gaussian approximation used here is slightly different from that described by Nagylaki (1990) in that it (a) does not assume that selection is weak and (b) allows for the values of the original and limiting processes at the initial time point to be different.

Moran's stochastic process describes the number $n^{(N)}(t)$ of mutants in a population of constant size N at time t. This number can increase by one from i to $i + 1$ with rate

$$r^{(N)}(i,i+1)=\mu_{\mathrm{m}}i\frac{\lambda_{\mathrm{w}}(N-i)}{\lambda_{\mathrm{w}}(N-i)+\lambda_{\mathrm{m}}i}$$

and decrease by one with rate

$$r^{(N)}(i,i-1)=\mu_{\mathrm{w}}(N-i)\frac{\lambda_{\mathrm{m}}i}{\lambda_{\mathrm{w}}(N-i)+\lambda_{\mathrm{m}}i}.$$

Here, μ_{w} and λ_{w} are the birth and death rates of the wild type, and μ_{m} and λ_{m} are the birth and death rates of the mutant type, respectively. We assume $\lambda_{\mathrm{w}} = \lambda_{\mathrm{m}}$, $\mu_{\mathrm{w}} = 1$, and let $\mu_{\mathrm{m}} = (1 + s)\mu_{\mathrm{w}} = 1 + s$. Then

$$r^{(N)}(i,i+1)=Nf_{+1}\left(\frac{i}{N}\right),\quad r^{(N)}(i,i-1)=Nf_{-1}\left(\frac{i}{N}\right)\tag{A1}$$

with

$$f_{+1}(x)=(1+s)x(1-x),\quad f_{-1}(x)=x(1-x).$$

Define

$$F(x)=\sum_{\delta\in\{-1,+1\}}\delta f_{\delta}(x)=sx(1-x)$$

$$G(x)=\sum_{\delta\in\{-1,+1\}}\delta^{2}f_{\delta}(x)=(2+s)x(1-x).$$

Let $X^{(N)}(t) = n^{(N)}(t)/N$ be the frequency of the mutant in the population at time t. The limit of $X^{(N)}$, $g(t, x_0) = \lim_{N\to\infty}X^{(N)}(t)$, is a deterministic function that, under certain regularity conditions, satisfies Equations 1 and 2 with $x_0 = \lim_{N\to\infty}X^{(N)}(0)$ and the solution given by (3).

Now let

$$Z(t)=\lim_{N\to\infty}\sqrt{N}\left(X^{(N)}(t)-g(t,x_0)\right)\tag{A2}$$

be the asymptotic process that describes the noise around the deterministic trajectory. If we knew the distribution of $Z(t)$, we could approximate the frequency $X^{(N)}$ at a finite N by

$$X^{(N)}(t)\approx g(t,x_0)+\frac{1}{\sqrt{N}}Z(t).\tag{A3}$$

The asymptotic noise process is in general a diffusion process, but, as long as it remains far from absorbing boundaries, it can be approximated by a Gaussian process with the corresponding first two moments. The advantage of this approach is that the first two moments of the diffusion process can be computed analytically, resulting in an expression for the probability distribution of the allele frequency at time t.

If $z_0=\lim_{N\to\infty}\sqrt{N}(X^{(N)}(0)-x_0)$ is the initial value of the limiting noise process, then the mean and variance of the noise process at time $t \geq 0$ are $\mathbb{E}Z(t) = M(t, x_0)z_0$ and Var $Z(t) = \sigma^2(t, x_0)$ respectively, where $M(t, x_0)$ satisfies the equations

$$\frac{dM}{dt}=F'(g(t,x_0))M=s\frac{(1-x_0)e^{-st}-x_0}{(1-x_0)e^{-st}+x_0}M\tag{A4}$$

$$M(0,x_0)=1\tag{A5}$$

and $\sigma^2(t,\, x_0)$ satisfies the equations

$$\begin{aligned}\tfrac{d\sigma^2}{dt} &= 2F'(g(t,x_0))\sigma^2 + G(g(t,x_0)) \\ &= \tfrac{(1-x_0)e^{-st}-x_0}{(1-x_0)e^{-st}+x_0}2s\sigma^2 + \tfrac{(2+s)x_0(1-x_0)e^{-st}}{((1-x_0)e^{-st}+x_0)^{-2}}\end{aligned} \tag{A6}$$

$$\sigma^2(0,x_0) = 0. \tag{A7}$$

The solution to Equations A4 and A5 is given by

$$M(t,x_0) = \exp\left\{ \int_0^t F'(g(\tau,x_0))d\tau \right\},$$

which, after substituting F' and g, yields

$$M(t,x_0) = e^{-st}\big(x_0 + (1-x_0)e^{-st}\big)^{-2}.$$

The solution to Equations A6 and A7 is given by

$$\sigma^2(t,x_0) = M^2(t,x_0)\int_0^t M^{-2}(\tau,x_0)G(g(\tau,x_0))d\tau,$$

which, after substituting G and g, yields

$$\begin{aligned}\sigma^2(t,x_0) = {} & M^2(t,x_0)(2+s)x_0(1-x_0)s^{-1} \\ & \times \Big[2x_0(1-x_0)st + x_0^2e^{st} - (1-x_0)^2e^{-st} \\ & + (1-x_0)^2 - x_0^2\Big].\end{aligned}$$

If the true state of the stochastic process $X^{(N)}$ is known to be $X^{(N)}(0)$ at time point 0, we can approximate the initial value of the limiting noise process as $z_0 \approx \sqrt{N}(X^{(N)}(0) - x_0)$. Then from (A3) we have

$$\mathbb{E}X^{(N)}(t) \approx g(t,x_0) + M(t,x_0)\Big(X^{(N)}(0) - x_0\Big),$$

$$\mathrm{Var}\, X^{(N)}(t) \approx \frac{1}{N}\sigma^2(t,x_0).$$

Analogously, if the value of the process $X^{(N)}$ is known to be $X^{(N)}(t')$ at a later time $t' \geq x_0$, then at time $t \geq t'$ we have

$$\mathbb{E}_{t'}X^{(N)}(t) \approx g(t,x_0) + M(\Delta t, g(t',x_0))\Big(X^{(N)}(t') - g(t',x_0)\Big), \tag{A8}$$

$$\mathrm{Var}_{t'}X^{(N)}(t) \approx \frac{1}{N}\sigma^2(\Delta t, g(t',x_0)), \tag{A9}$$

where $\Delta t = t - t'$, and $\mathbb{E}_{t'}$ and $\mathrm{Var}_{t'}$ denote conditional expectation and variance given the state of the process at time t'. Thus, the conditional distribution of the allele frequency $X^{(N)}$ at time t given its value at time $t' \leq t$ can be approximated by a Gaussian distribution with mean given by (A8) and variance given by (A9). We apply this approximation to every observation interval $(t_{i-1},\, t_i)$, $i = 1, \ldots, L$. As noted above, the initial value of the deterministic process, x_0, is a free parameter that can be fitted along with N and s. However, we set x_0 to be equal to the observed allele frequency ν_0 at time 0 to reduce the number of fitted parameters.

Note that the approximations described here work for the Moran process that is density dependent as can be seen from Equations A1. The Wright–Fisher process is not density dependent and, strictly speaking, the approximations described here are not valid, although in practice they work well.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.158220/-/DC1

Identifying Signatures of Selection in Genetic Time Series

Alison F. Feder, Sergey Kryazhimskiy, and Joshua B. Plotkin

DOI: 10.1534/genetics.113.158220

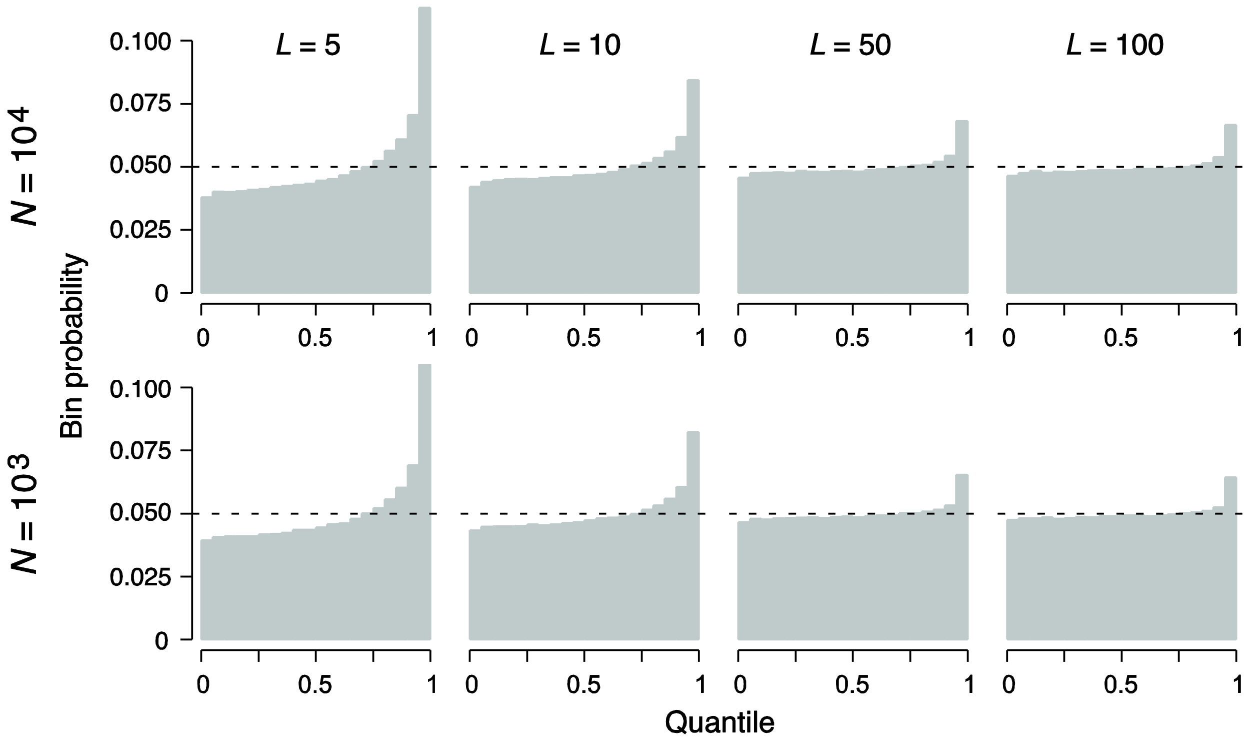

Figure S1 Comparison of the true LRS distribution to the χ^2 distribution with 1 df. Panels show comparisons for different values of the population size *N* and the number of sampled time points *L*, as indicated on the left and on top. Notations are as in Figure 1. Parameter values: *T* = 100, Δ = 20, ν_0 = 0.5; the number of Wright-Fisher simulations was 10^6.

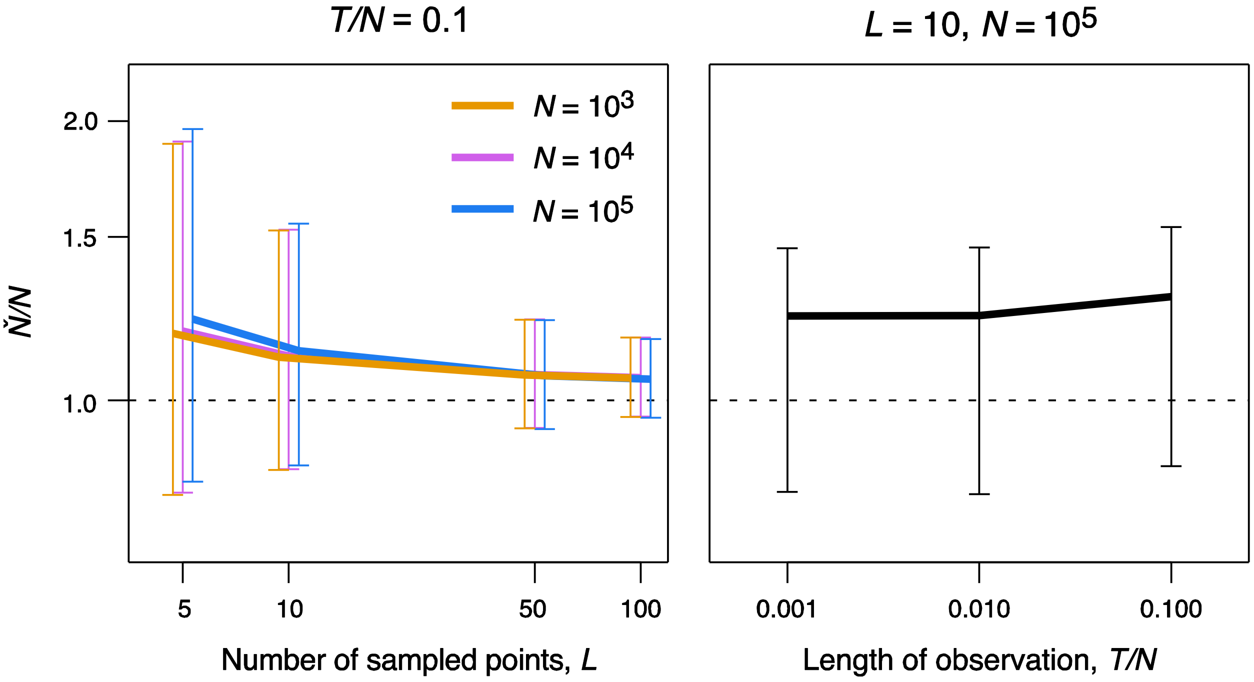

Figure S2 Bias in the maximum-likelihood estimate of population size under neutrality. The figure shows the ratio of the most-likely population size under neutrality, $\check{N}$, to the true population size, N, as a function of the number of sampled points L (left panel) and as a function of the length of the observed time series T (right panel). Whiskers indicate quartiles of the distribution of $\check{N}/N$. In the right panel, curves for different population sizes are slightly shifted along the x-axis for clarity. Bias in $\check{N}$ decreases as the number of sampled time points increases. The bias is nearly independent of N and of the length of the sampling period. The number of Wright-Fisher simulations was 10^5.

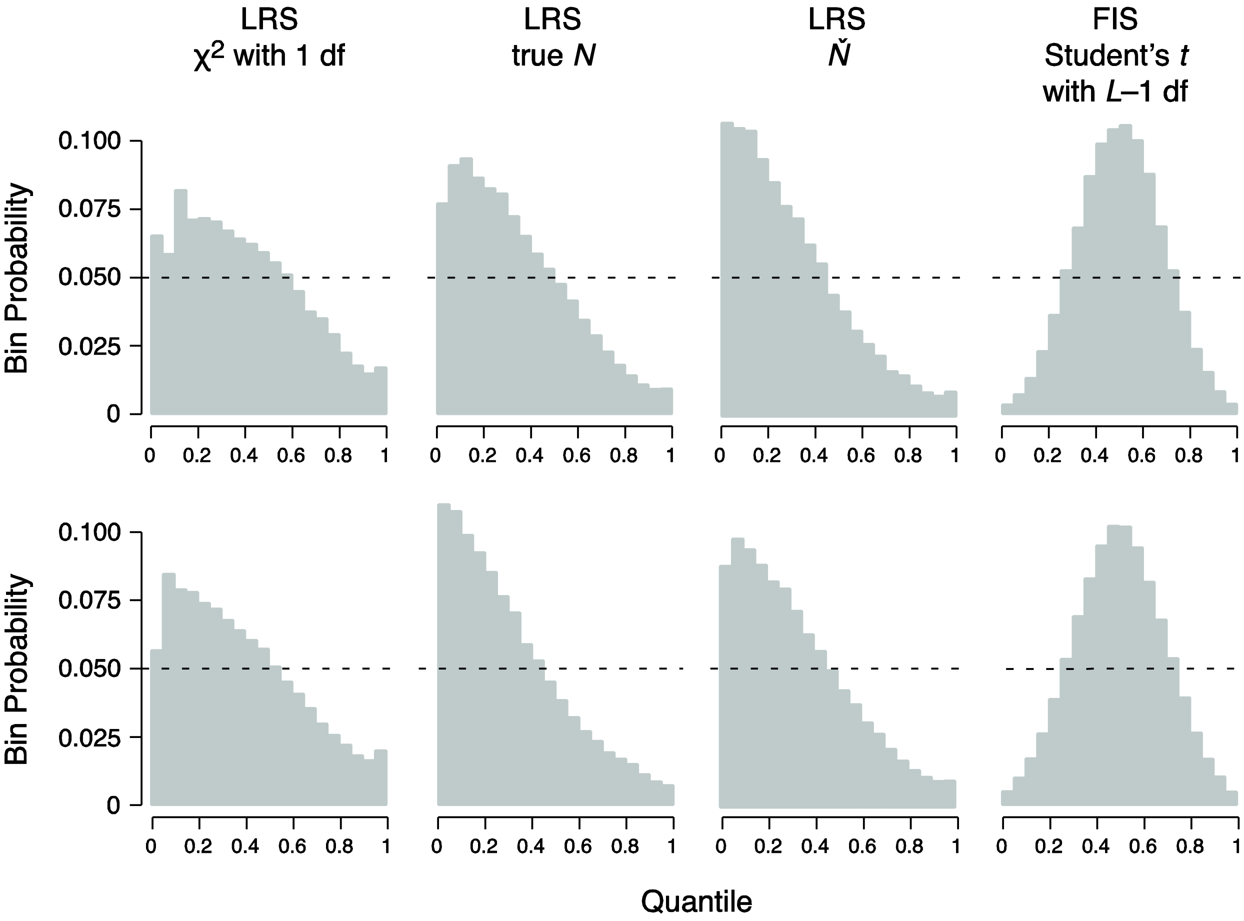

Figure S3 Distributions of various test statistics under the neutral null hypothesis, when allele frequencies are sampled with noise. Top row, n = 50. Bottom row, n = 100. Notations as in Figure 1. Parameter values: $N = 10^3$, T = 10, Δ = 2, L = 5, ν_0 = 0.5; the number of Wright-Fisher simulations was 10^5.

Table S1 Deviation of the distribution of LRS from the χ^2 distribution for different initial allele frequencies

Simulation parameters		Sampling parameters					α			
v_0	*N*	*T*	*L*	Δ	Absorption probability	$\check{N}/N$	0.05	0.01	0.001	0.0001
0.1	10^4	10	10	1	8×10^{-5}	1.3	1.4	1.8	2.3	3.7
	10^4	100	10	10	2.5×10^{-4}	1.3	1.6	2.2	3.6	5.9
	10^4	1000	10	100	0.067	1.5	2.2	4.4	14.3	51.8
	10^3	100	10	10	0.123	1.5	1.9	3.2	7.9	20.9
	10^4	1000	10	100	0.067	1.5	2.2	4.4	14.3	51.8
	10^5	10000	10	1000	6.6×10^{-4}	1.2	0.2	0.4	0.9	1.9
	10^4	100	5	20	2.3×10^{-4}	1.7	2.2	3.4	6.3	11.5
	10^4	100	10	10	2.5×10^{-4}	1.3	1.6	2.2	3.6	5.9
	10^4	100	100	1	2.4×10^{-4}	1.1	1.2	1.6	2.5	4.9
0.2	10^4	10	10	1	9.8×10^{-5}	1.3	1.4	1.7	2.4	3.4
	10^4	100	10	10	3.4×10^{-4}	1.3	1.5	1.9	2.7	3.8
	10^4	1000	10	100	6.9×10^{-3}	1.4	2.2	4.0	11.4	35.7
	10^3	100	10	10	0.014	1.4	2.2	4.0	10.5	29.8
	10^4	1000	10	100	6.9×10^{-3}	1.4	2.2	4.0	11.4	35.7
	10^5	10000	10	1000	1.3×10^{-3}	1.3	0.5	0.7	1.6	3.5
	10^4	100	5	20	3.2×10^{-4}	1.7	2.0	3.1	5.5	9.5
	10^4	100	10	10	3.4×10^{-4}	1.3	1.5	1.9	2.7	3.8
	10^4	100	100	1	3.5×10^{-4}	1.0	1.1	1.2	1.6	1.8
0.3	10^4	10	10	1	1.0×10^{-4}	1.3	1.4	1.7	2.3	3.2
	10^4	100	10	10	4.3×10^{-4}	1.3	1.4	1.8	2.5	3.5
	10^4	1000	10	100	1.6×10^{-3}	1.3	1.9	3.1	7.0	17.4
	10^3	100	10	10	1.3×10^{-3}	1.3	2.0	3.2	7.5	18.7
	10^4	1000	10	100	1.6×10^{-3}	1.3	1.9	3.1	7.0	17.4
	10^5	10000	10	1000	0.012	1.3	1.3	1.7	2.7	4.5
	10^4	100	5	20	3.5×10^{-4}	1.7	2.0	3.0	5.2	8.6
	10^4	100	10	10	4.3×10^{-4}	1.3	1.4	1.8	2.5	3.5
	10^4	100	100	1	3.7×10^{-4}	1.0	1.1	1.1	1.2	1.5
0.4	10^4	10	10	1	1.1×10^{-4}	1.3	1.4	1.7	2.4	3.1
	10^4	100	10	10	4.1×10^{-4}	1.3	1.4	1.8	2.3	3.0
	10^4	1000	10	100	1.3×10^{-3}	1.3	1.7	2.5	4.9	9.8
	10^3	100	10	10	2.6×10^{-4}	1.3	1.8	2.6	5.2	10.0
	10^4	1000	10	100	1.3×10^{-3}	1.3	1.7	2.5	4.9	9.8
	10^5	10000	10	1000	0.013	1.3	1.2	1.6	2.4	3.9
	10^4	100	5	20	6.2×10^{-4}	1.7	2.0	3.0	5.3	8.5
	10^4	100	10	10	4.1×10^{-4}	1.3	1.4	1.8	2.3	3.0
	10^4	100	100	1	4.1×10^{-4}	1.0	1.1	1.1	1.1	1.2
0.5	10^4	10	10	1	2.6×10^{-3}	1.3	1.4	1.7	2.3	3.3
	10^4	100	10	10	7.9×10^{-4}	1.3	1.4	1.7	2.3	2.9
	10^4	1000	10	100	1.2×10^{-3}	1.3	1.6	2.4	4.3	8.1
	10^3	100	10	10	2.2×10^{-3}	1.3	1.7	2.5	4.5	8.3
	10^4	1000	10	100	1.2×10^{-3}	1.3	1.6	2.4	4.3	8.1
	10^5	10000	10	1000	1.3×10^{-2}	1.3	1.2	1.5	2.3	3.6
	10^4	100	5	20	7.7×10^{-4}	1.7	2.0	2.9	5.2	8.5
	10^4	100	10	10	7.9×10^{-4}	1.3	1.4	1.7	2.3	2.9
	10^4	100	100	1	8.0×10^{-4}	1.0	1.1	1.1	1.2	1.4

Table S2 Mutant allele frequencies in yeast data from Lang et al, 2013 and the application of ELRT and FIT

			Time in generations[a]								ELRT[b]						FIT		
Line	Gene	Amino acid change	415	505	585	665	745	825	910	1000	$\check{N}$	$\hat{N}$	$\hat{s}$,%	P_{abs}	*P-val* bound	Cond *P*-val	t_{FI}	*df*	*P*-val
RMB2-F01	*STE11*	D579Y	n/a	n/a	n/a	n/a	**3**	**9**	**24**	32	181	1.2×10^5	1.7	0.977	0.983	0.261	18.1	1	0.018
			n/a	n/a	n/a	n/a	**67**	**69**	**60**	68									
RMS1-D12	*IRA1*	Y822*	8	34	84	**73**	**98**	**97**	**97**	n/a	3338	6.7×10^4	1.6	0.048	0.049	0.001	5.9	2	0.014
			115	64	102	**93**	**105**	**99**	**98**	n/a									
BYS2-D06	*IRA2*	A2698T	n/a	2	1	**51**	**64**	**79**	**119**	n/a	539	850	2.0	0.440	0.557	0.209	2.3	2	0.074
			n/a	154	147	**150**	**152**	**93**	**125**	n/a									

[a]For each mutation, top row shows the number of reads with the mutant allele and the bottom row shows the total coverage at that site. Data points that give the lowest FIT *P*-value for each mutation are in bold.

[b]For ELRT, we show the parameter values that maximize likelihood functions (9) and (10) under the Gaussian approximation, the probability P_{abs} of an absorption event during the sampling period in the neutral Wright-Fisher trials with $\check{N}$, an upper bound on the ELRT *P*-value given by the fraction of trials that have a higher than observed LRS among all trials, and the conditional *P*-value given by the fraction of trials that have a higher than observed LRS among trials without an absorption event.

Tables S3 and S4 are available for download as Excel files at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.158220/-/DC1.

Table S3 Results of FIT applied to data from yeast populations evolved at different population sizes

Table S4 Raw flow cytometry counts for yeast populations evolved at different population sizes