# The Limits to Estimating Population-Genetic Parameters with Temporal Data

Michael Lynch* and Wei-Chin Ho

Biodesign Center for Mechanisms of Evolution, Arizona State University

*Corresponding author: E-mail: mlynch11@asu.edu.

**Accepted:** March 13, 2020

## Abstract

The ability to obtain genome-wide sequences of very large numbers of individuals from natural populations raises questions about optimal sampling designs and the limits to extracting information on key population-genetic parameters from temporal-survey data. Methods are introduced for evaluating whether observed temporal fluctuations in allele frequencies are consistent with the hypothesis of random genetic drift, and expressions for the expected sampling variances for the relevant statistics are given in terms of sample sizes and numbers. Estimation methods and aspects of statistical reliability are also presented for the mean and temporal variance of selection coefficients. For nucleotide sites that pass the test of neutrality, the current effective population size can be estimated by a method of moments, and expressions for its sampling variance provide insight into the degree to which such methodology can yield meaningful results under alternative sampling schemes. Finally, some caveats are raised regarding the use of the temporal covariance of allele-frequency change to infer selection. Taken together, these results provide a statistical view of the limits to population-genetic inference in even the simplest case of a closed population.

**Key words:** population genomics, effective population size, selection coefficient, fluctuating selection.

## Introduction

Most evolutionary features ultimately depend on the magnitudes of a few key population-genetic parameters—rates of mutation, recombination, and migration, the strength and pattern of selection, and the stochastic noise caused by random genetic drift. As the magnitudes of these forces are typically quite small at the molecular level, most estimates derive from observations on standing patterns of variation in samples of multiple individuals acquired at a single time point, which are typically assumed to reflect cumulative, equilibrium effects. A diversity of methods relies on such strategies, including those focused on variation at genomic sites or on the magnitude of linkage disequilibrium among sites (Walsh and Lynch 2018, chapters 2–4).

One limitation of single-sample approaches is that most patterns of variation are functions of at least two evolutionary forces. If one desires an estimate of one particular population-genetic parameter, this then requires a preexisting estimate of another key parameter. For example, a common procedure uses silent-site diversity ($\pi_s$, obtained as an average from a sample of multiple individuals and nucleotide sites) as an estimator of the equilibrium expectation $4N_eu$, which is equivalent to the ratio of the power of mutation to that of drift, where $N_e$ is the effective population size (the inverse of which governs the stochasticity of allele-frequency change) and $u$ is the mutation rate per nucleotide site per generation. Extrapolation of an estimate of $N_e$ from $\pi_s$ requires an estimate of $u$. In the absence of such information, it is often simply assumed that $\pi_s$ will scale positively with $N_e$, but this ignores the fact that the mutation rate varies by at least three orders of magnitude across the Tree of Life, scaling nearly inversely with $N_e$ (Lynch et al. 2016; Long et al. 2018). Additional uncertainties associated with single-sample methods include the assumptions of drift-mutation equilibrium and neutrality of the observed polymorphisms.

Similar caveats arise when measures of linkage disequilibrium are used to infer the scaled recombination rate $4N_ec$, where $c$ is the recombination rate between sites, and when measures of variation at putatively selected sites are used to estimate $4N_es$, where $s$ is the strength of selection operating on a nucleotide site. Although some single-sample methods have been derived for reconstructing historical patterns of $N_e$ from the information in site-frequency spectra (e.g., Liu and Fu 2015) or patterns of linkage disequilibrium (Li and Durbin 2011), thereby providing insight into population-size stability, these are unable to estimate $N_e$ within the past several hundred generations, simply because the pool of very young (and rare) alleles is essentially unobservable unless sample sizes are enormous. Moreover, such methods typically retain the assumption of a closed-population structure.

With the field of population genomics now well established, the goal here is to evaluate the limits to the information that can be extracted from temporal sequences of samples of the genomes of multiple individuals, where $10^6$ or so polymorphisms in a sample is in the realm of possibility. An advantage of sequential-sample estimators of $N_e$ and $s$ is that they often do not require estimates of mutation or recombination rates (provided these forces are weak relative to those being estimated). In principle, for example, it ought to be possible to estimate the current effective size of a population from observed temporal fluctuations in allele frequencies, provided sample sizes are large enough that the true signal of drift is not overwhelmed by sampling error (Krimbas and Tsakas 1971; Nei and Tajima 1981; Pollak 1983; Waples 1989). However, because such $N_e$ estimators still rely on the assumption of neutrality and a closed-population structure, there is a need for methodology to verify in advance that the molecular markers under consideration are actually consistent with such conditions.

Here, we explore three general issues relevant to the estimation of population-genetic parameters using temporal series of data. First, we develop a simple test for the consistency of the data with neutrality and closed-population structure, as these conditions are ultimately required if temporal fluctuations in allele frequencies are to be reliable indicators of $N_e$. Second, we explore ways to further evaluate whether polymorphisms at individual nucleotide sites are experiencing significant directional selection, and if not, whether significant fluctuating selection, for example, quasi-neutrality in the sense of Wright (1948) and Kimura (1954) is occurring. Finally, for situations in which the data appear to be consistent with the assumptions of the model, we evaluate the challenges of estimating $N_e$ in large populations by the sequential sampling method.

As the goal is to determine the limits to the ability to estimate current-day population-genetic parameters, it will be assumed throughout that accurate estimates of individual genotypes have been obtained (assured, e.g., with adequate depth of sequencing coverage per site and application of suitable base-call quality screening). All sampling error is therefore associated with the number of surveyed individuals and nucleotide sites. Polymorphic sites will be assumed to be biallelic, as is almost always observed in population samples, although pooling of alleles can be performed with tri- or tetra-allelic sites. As the focus is on populations of relatively large size, it will be assumed that the sampling procedure has minimal effect on the change in population allele frequencies over time. For purposes of presentation, an even temporal distribution of samples will be assumed throughout. The derived expressions for the expected sampling variances of estimates are those for single surveys, but if one had the luxury of performing $\eta$ experimental replications, then the expected sampling variances would be $1/\eta$ times the reported expressions, and the standard errors should be multiplied by $1/\sqrt{\eta}$.

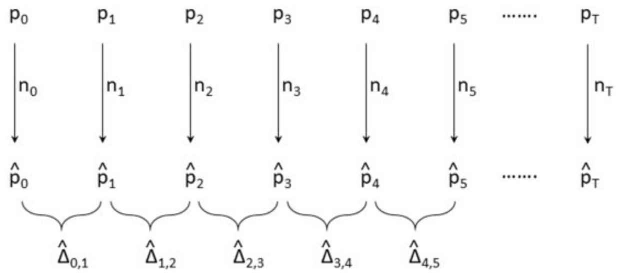

FIG. 1.—Sampling scheme for allele frequencies, with $p_i$ denoting the parametric frequency in the population, $\widehat{p}_i$ denoting the estimated frequency after sampling $n_i$ individuals, and $\widehat{\Delta}_{ij}$ denoting the interval-specific estimated change in frequency.

## Validating the Primacy of Random Genetic Drift

A critical assumption in the application of temporal methods for estimating $N_e$ is that the series of sequential changes in allele frequencies across sampling points are consistent with a random walk. Consider a temporal series of allele frequencies with population parametric values at nucleotide site $i$ equal to $p_{i,0}, p_{i,1}, \ldots, p_{i,T}$ over $T$ equally spaced intervals (e.g., generations) (fig. 1). At each sampling point $t = 0, \ldots, T$, the genotypes are determined for $n_{i,t}$ individuals (here assumed to be diploid), yielding a series of allele-frequency estimates $\widehat{p}_{i,0}, \widehat{p}_{i,1}, \ldots, \widehat{p}_{i,T}$. In the following, we will drop the subscript for nucleotide sites for simplicity unless otherwise noted, and we will assume a constant sample size $n$. From these data, a series of allele-frequency change estimates can be obtained between adjacent samples: $\widehat{\Delta}_{0,1}, \widehat{\Delta}_{1,2}, \ldots, \widehat{\Delta}_{T-1,T}$. More generally, allowing for any number of generations separating samples $\widehat{\Delta}_{j,k} = \widehat{p}_k - \widehat{p}_j$.

If random genetic drift is the predominant evolutionary force, because the process has no memory, there should be no correlation between the parametric allele-frequency changes in different intervals, and the interval-specific changes should be random values with expectations equal to zero. The expected variance of change between generation $j$ and $j+1$ is equal to $\sigma^2(\Delta_{j,j+1}) = p_j(1-p_j)/(2N_e)$, (assuming constant $N_e$ across generations), although as discussed below, additional variance is introduced by sampling of individuals each generation.

There are a number of ways to test for randomness in allele-frequency change. For example, one might evaluate whether the changes in allele frequencies in intervals spaced by a specific number of generations are correlated, which would not be expected if random drift were the predominant evolutionary force (Buffalo and Coop 2019). Even here, however, a number of subtleties can lead to misleading interpretations. First, even under a pure drift model, changes in adjacent intervals will be negatively correlated, owing to the sharing of an intermediate sampling point—if $p_t$ is overestimated by investigator sampling error, then $\widehat{\Delta}_{t-1,t}$ will be overestimated and $\widehat{\Delta}_{t,t+1}$ will be underestimated, and vice versa if $p_t$ is underestimated. As this leads to a substantial expected covariance equal to $-p_t(1-p_t)/(2n)$, a comparison of allele-frequency changes in adjacent intervals must be avoided. One might imagine an efficient use involving the pairing of changes separated by single intervals, that is, $\widehat{\Delta}_{0,1}$ with $\widehat{\Delta}_{2,3}$, $\widehat{\Delta}_{3,4}$ with $\widehat{\Delta}_{5,6}$, etc. However, even in this case, owing to the sharing of time points 3, 6, and so on, the expected covariance of pairs of changes has a positive bias $\simeq p(1-p)/(2nm)$, where $m = T/3$ is the total number of paired intervals. Thus, we now focus on three sampling schemes with minimal to zero bias in the expected values for the covariance in allele-frequency changes under neutrality, the first two focused on the specific behavior of individual sites, and the third involving an aggregate measure over multiple sites.

## Sampling Scheme 1

Given the amount of effort invested in a temporal survey, one will want to use as much of the data as possible, and avoiding the types of adjacent pairings noted above, this might still be accomplished by evaluating the covariance of allele-frequency changes for all pairs of observations separated by single intervals. This can be accomplished by pairing $\Delta_{0,1}$ with $\Delta_{2,3}$, $\Delta_{1,2}$ with $\Delta_{3,4}$, $\Delta_{2,3}$ with $\Delta_{4,5}$, up to $\Delta_{T-3,T-2}$ with $\Delta_{T-1,T}$. One issue with this approach is that some of the $\Delta$ values are used twice (e.g., $\Delta_{2,3}$), and most still share a sampling point (e.g., $\Delta_{0,1}$ and $\Delta_{1,2}$), that is, the terms involved in the analysis are still not completely independent. The net result is that the expected covariance is slightly negative, rather than zero.

To see this, note that the unbiased sample covariance estimator for $m$ pairs of variables, $x$ and $y$, is

$$\mathrm{Cov}(x,y) = \frac{m(\overline{xy} - \bar{x}\cdot\bar{y})}{(m-1)}. \tag{1}$$

For this particular sampling scheme, $m = T - 2$, so that when $T = 5$ (six samples in total), $m = 3$, and the covariance for the site-specific changes is obtained by letting

$$\begin{aligned}
\overline{xy} &= [(\Delta_{0,1}\cdot\Delta_{2,3}) + (\Delta_{1,2}\cdot\Delta_{3,4}) + (\Delta_{2,3}\cdot\Delta_{4,5})]/3 \\
\bar{x} &= (\Delta_{0,1} + \Delta_{1,2} + \Delta_{2,3})/3 \\
\bar{y} &= (\Delta_{2,3} + \Delta_{3,4} + \Delta_{4,5})/3.
\end{aligned}$$

Generalizing to arbitrary $T$, taking expectations of all terms, and noting that all expected cross-products of terms not sharing sampling points are equal to zero (as they share no drift or sampling deviations), the expected covariance for this sampling scheme under neutrality becomes

$$E[\mathrm{Cov}_1] = -\frac{E(\delta^2_{d,2} + \cdots + \delta^2_{d,T-3})}{(T-3)(T-2)}, \tag{2a}$$

where $\delta^2_{d,t} = p_t(1-p_t)/(2N_e)$ is the expected squared deviation of allele-frequency change over interval $(t, t+1)$ resulting from genetic drift. Because the expected heterozygosity declines by a factor of $1/(2N_e)$ per generation, the previous expression reduces to

$$E[\mathrm{Cov}_1] = -\frac{p_0(1-p_0)}{2(T-3)(T-2)N_e}\cdot\sum_{i=2}^{T-3}\lambda^i, \tag{2b}$$

where $\lambda = 1 - [1/(2N_e)]$. However, for populations with $N_e \gg T$ (most practical applications), $\lambda \simeq 1.0$, and

$$E[\mathrm{Cov}_1] \simeq -\frac{p_0(1-p_0)(T-4)}{2(T-3)(T-2)N_e}. \tag{2c}$$

This shows that the expected sampling covariance of allele-frequency changes separated by an interval is negative if all intervals are used. Although the absolute bias is expected to be very small in the case of very large populations, the dependency on $N_e$ remains a concern, as this is the parameter that one wishes to eventually estimate.

## Sampling Scheme 2

Now consider the situation in which there is no overlap at all in the use of time points in different pairings, that is, contrasting $\Delta_{0,1}$ with $\Delta_{2,3}$, $\Delta_{4,5}$ with $\Delta_{6,7}$, $\Delta_{8,9}$ with $\Delta_{10,11}$, up to $\Delta_{T-3,T-2}$ with $\Delta_{T-1,T}$. The expected covariance under neutrality is zero with this sampling scheme, although the cost is a substantially longer experiment. For example, with scheme 1, three sets of comparisons can be obtained with $T = 5$, but this requires $T = 11$ with scheme 2. A simple expression for the expected sampling variance of the covariance under scheme 2 can be obtained by noting from equation (1),

$$\sigma^2[\mathrm{Cov}(x,y)] = \left(\frac{m}{m-1}\right)^2\{\sigma^2(\overline{xy}) - 2\sigma[\overline{xy}, (\bar{x}\cdot\bar{y})] + \sigma^2(\bar{x}\cdot\bar{y})\}. \tag{3a}$$

Expanding the terms for $\overline{xy}$, $\bar{x}$, and $\bar{y}$, noting that $\sigma^2(xy) = \sigma^2_x \cdot \sigma^2_y$ for two uncorrelated variables each with expectations zero, and that in this case, $\sigma^2(\widehat{\Delta}) = \sigma^2_x = \sigma^2_y \simeq p(1-p)/n$ (because $\widehat{\Delta}$, where $\widehat{}$ denotes an estimate, is derived from two samples containing $2n$ genes), leads to the expected sampling standard error for the covariance of changes at a particular site

$$\sigma[\mathrm{Cov}_2] = \frac{p_0(1-p_0)}{n\sqrt{m-1}}, \tag{3b}$$

where $m$ is the number of paired $\Delta$ comparisons. Averaging over $L$ independent sites with different allele frequencies,

$$\sigma[\overline{\mathrm{Cov}_2}] = \frac{\overline{p_0(1-p_0)}}{n\sqrt{L(m-1)}}. \tag{3c}$$

In contrast, for sampling scheme 1, owing to the nonindependence of data and the associated reduction in degrees of freedom,

$$\sigma[\overline{\mathrm{Cov}_1}] \simeq \frac{\overline{p_0(1-p_0)}}{n\sqrt{L(m-1)/2}}. \tag{4}$$

Assuming $T+1$ consecutive samples, $m=(T+1)/4$ for scheme 2, and equation (3c) reduces to

$$\sigma[\overline{\mathrm{Cov}_2}] = \frac{2 \cdot \overline{p_0(1-p_0)}}{n\sqrt{L(T-3)}}, \tag{5}$$

and because $m-1=T-3$ with scheme 1, for equivalent $T$, the sampling error associated with scheme 2 is $2/\sqrt{2} \simeq 1.4\times$ that for scheme 1. Thus, aside from the slight bias, scheme 1 may be viewed as yielding more efficient use of the data. In either case, the preceding expressions indicate that the expected standard error of the test statistic is inversely proportional to the number of individuals sampled per time point ($n$), but only with the approximate square root of the total number of nucleotide sites sampled ($\simeq LT$ for large $T$).

Note that here and below it is assumed that when estimators are averaged over sites, the vast majority of sites are effectively in linkage equilibrium, as the $L$ deployed in the sampling-variance estimators is equivalent to the number of degrees of freedom. This will generally not be a problem if unlinked sites are relied upon. However, studies involving specialized base-population crosses, such as multiparent populations that initiate with high levels of linkage disequilibrium (King and Long 2017), could be especially challenging here. Although the actual degrees of freedom might be approximated through computational comparison of observed sampling variances with those expected with $L$ degrees of freedom, this would require prior information on recombination rates. The key point is that, when applied to composite estimators based on multiple loci, all expressions derived herein yield lower-bound estimates of the sampling variance if $L$ is used as the number of degrees of freedom.

## Genome-Wide Survey

Whereas the two preceding sampling schemes consider the covariance of change within particular loci over multiple intervals, such approaches are strongly limited by the duration of sampling. An alternative route is to evaluate the average genome-wide covariance of change across $L$ sites with a shorter temporal-series duration,

$$\mathrm{Cov}_3 = \frac{L \cdot (\overline{\Delta_{ij}\Delta_{ik}} - \overline{\Delta_{ij}} \cdot \overline{\Delta_{ik}})}{L-1}, \tag{6}$$

where $j$ and $k$ denote two time intervals, and the means in the numerator are over all $i=1,\ldots,L$ sites. Under neutrality, assuming the time intervals share no samples, $\mathrm{Cov}_3$ has expected value zero, and assuming $L \gg 1$, sampling standard error

$$\sigma[\mathrm{Cov}_3] = \frac{\overline{p_0(1-p_0)}}{n\sqrt{L}}, \tag{7}$$

identical to equation (3c) with $m=2$. Note that this sort of analysis can be performed on any pair of intervals over the sampling period, provided there is no overlap in the sampling points. One might, for example, consider the covariance of change between interval (0,1) and interval (2,3), interval (3,4), etc.

For hypothesis testing, we take advantage of work by Pearson et al. (1929), where the sampling distribution of a covariance for a bivariate normal sampling distribution was derived. Letting $\hat{\sigma}_{x,y}$ denote a sample covariance, and $\sigma_x$, $\sigma_y$, and $\rho$ be the parametric standard deviations of $x$ and $y$, and their correlation coefficient, and using the standardized variable $v = L\hat{\sigma}_{x,y}/[(1-\rho^2)\sigma_x\sigma_y]$,

$$\mathrm{pdf}(v) = \frac{(1-\rho^2)^{(L-1)/2}}{\sqrt{\pi} \cdot 2^{(L/2)-1} \cdot \Gamma[(L-1)/2]} \cdot \exp(\rho v) \cdot v^{(L/2)-1} \cdot K_{(L/2)-1}(v), \tag{8a}$$

where $\Gamma(x)$ is the gamma function, and $K_{(L/2)-1}(v)$ denotes a modified Bessel function of the second kind. For the special case in which $\rho=0$ (our null hypothesis), Pearson et al. (1929) found that the preceding expression can be closely approximated (perfectly to the first four moments) with

$$\mathrm{pdf}(v) = \frac{\Gamma[(L+4)/2]}{\Gamma[(L+3)/2]\sqrt{\pi(L^2-1)}} \left(1 + \frac{v^2}{L^2-1}\right)^{-(L+4)/2}. \tag{8b}$$

In our case, $v = Ln\hat{\sigma}_{x,y}/[p_0(1-p_0)]$, and with $L \gg 1$, equation (8b) simplifies further to

$$\mathrm{pdf}(v) \simeq \frac{1}{\sqrt{2\pi L}} \cdot \exp\left(-\frac{v^2}{2L}\right), \tag{8c}$$

showing that the standardized covariance measure $v$ is normally distributed with variance $L$.

This suggests a relatively simple test for random allele-frequency changes across different intervals. Let $\hat{\sigma}(\Delta')$ be the covariance of standardized allele-frequency changes for pairs of intervals, with the change at each site being normalized by the expected standard deviation of changes, that is, using $\widehat{\Delta}'_{j_0,j_1} = (\widehat{p}_{j_1} - \widehat{p}_{j_0})/\sqrt{\widehat{p}(1-\widehat{p})/n}$ as the standardized change between the time points $j_0$ and $j_1$ for interval $j$, with $\widehat{p}$ being the average allele frequency over the two time samples at the site, with a similar definition for interval $k$. Because $\widehat{\sigma}(\Delta') = v/L$, it follows that

$$\mathrm{pdf}[\widehat{\sigma}(\Delta')] \simeq \frac{1}{\sqrt{2\pi/L}} \cdot \exp\{-\frac{[\widehat{\sigma}(\Delta')]^2}{2/L}\}, \qquad (8d)$$

that is, under the null hypothesis, the sample covariance of standardized allele-frequency changes has expectation zero and variance $1/L$. (Again, as noted above, if the assayed loci are linked, $L$ needs to be replaced with the appropriate degrees of freedom.) A test for random change is then achieved by comparing the absolute value of $\widehat{\sigma}(\Delta')$ with the critical two-tail cutoff points of the normal distribution. If the data are consistent with a model of pure drift, for large $L$, there is a 5% probability that the absolute value of $\widehat{\sigma}(\Delta')$ will exceed $1.96/\sqrt{L}$ by chance, and a 1% probability that it will exceed $2.58/\sqrt{L}$ by chance. Observed values of $|\widehat{\sigma}(\Delta')|$ beyond these critical points imply that sampling heterogeneity problems, migration, and/or selection have contributed to the average observed pattern of change.

## Estimation of the Mean and Variance of Selection Intensity

Results from hundreds of single-sample studies in molecular population genetics suggest that the intensity of directional selection operating at the single-nucleotide level is often on the order of the reciprocal of $N_e$ or a factor several-fold larger. Selection coefficients at the nucleotide level >0.01 are exceedingly rare in studies of natural populations, and as these only induce an ~1% change in allele frequency per generation, the challenges in estimating selection at the DNA level with temporal data are clear. An additional issue (aside from possible contributions from nonselective forces) is that temporal changes in allele frequencies may result from direct selection on the nucleotide site of interest or indirectly from selection operating on adjacent sites in linkage disequilibrium. Thus, the best that we can hope to achieve with a temporal survey is a measure of the net strength of selection operating on a site.

One way forward is to note that in the absence of frequency-dependent selection, and assuming additive fitness effects, the expected change in allele frequency over an interval of $t$ time units is

$$\zeta_t = \zeta_0 - \mu_s t, \qquad (9a)$$

where

$$\zeta_t = \ln\left(\frac{1-p_t}{p_t}\right), \qquad (9b)$$

and $\mu_s$ is the average interval-specific selection coefficient over the entire sampling period (Crow and Kimura 1970). It follows that the slope of a regression of $\zeta_t$ on time provides an estimate of $\mu_s$ for the allele designated by frequency $p$. This estimator is only defined when the allele-frequency estimate is $0 < \widehat{p}_t < 1$ at all sample points. In principle, the selection coefficient may also fluctuate in time, one case of special interest being quasi-neutrality, wherein the long-term average $s$ is equal to zero with the interval-specific magnitudes of selection wandering randomly around this mean with temporal variance $\sigma_s^2$ (Wright 1948; Kimura 1954).

### Estimation of Mean Selection Coefficients

An efficient means of estimating $\mu_s$ for a nucleotide site is to perform a least-squares regression of $\zeta_t$ on time. Allowing for both selection and drift in a Wright–Fisher framework, followed by random sampling, computer simulations indicate that the regression coefficients provide unbiased estimates of $\mu_s$ over reasonable sample sizes and allele frequencies, so long as selection is strong enough to dominate random genetic drift (fig. 2, left). Negative bias occurs, independent of the experimental duration and sample size, when $N_e s p_0 < 1$, consistent with the view that selection operates in nearly deterministic fashion only after an allele frequency exceeds $1/(N_e s)$ (Walsh and Lynch 2018, chapter 7), as assumed in equation (9b). In principle, a more elaborate expression for allele-frequency change that allows for the influence of drift might be developed, but this would require an estimate of $N_e$.

An expression for the expected sampling variance of the regression coefficient (the sample estimate $\widehat{s}$ of the population parameter $\mu_s$), the steps of which can be found in the Appendix, is

$$\sigma^2(\widehat{s}) \simeq \left[\frac{6}{(T-1)(T+1)}\right]\left[\frac{1}{Tnp_0(1-p_0)} + \sigma_s^2\right], \qquad (10)$$

where $\sigma_s^2$ is the among-generation variance in $s$. This expression provides an essentially unbiased estimate of the sampling variance obtained in computer simulations, for a wide range of experimental durations ($T$ = 5–90), sample sizes ($n$ = 100–1,000), full range of allele frequencies ($p_0$ = 0.05–0.50), and a range of selection coefficients ($s = 10^{-5}$ to $10^{-2}$), provided $N_e > 10^5$ (fig. 2, right). For weak selection and long experimental durations, equation (10) somewhat underestimates the sampling variance because it does not account for the cumulative effects of random genetic drift.

In practical applications, one would ordinarily accept the estimate of the sampling variance of the regression coefficient from direct statistical analysis, but the expectation given by equation (10) provides insight into the optimal design of sampling schemes for estimating $\mu_s$. Regardless of the average strength of selection, provided $T\sigma_s^2$ is small relative to the

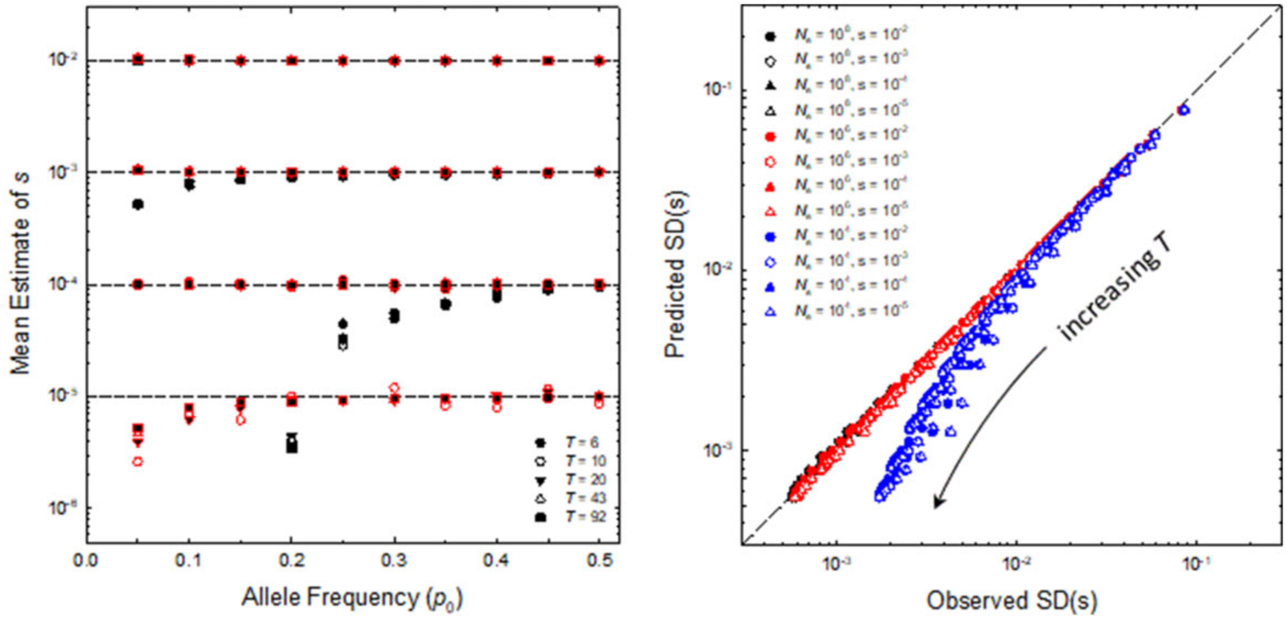

Fig. 2.—(Left) Mean estimates of the selection coefficient $s$ obtained from the least-squares regression approach. Each point is the average of the results from $10^7$ simulations based on Wright–Fisher allele-frequency dynamics incorporating selection and drift, followed by random sampling of $n = 100$ diploid individuals at each sampling point. Black symbols are for effective population size $N_e = 10^4$, and red for $N_e = 10^6$, and results are reported for a range of starting allele frequencies, $p_0$. The horizontal dashed lines denote the expectations for four evaluated selection coefficients (with temporal variance, $\sigma_s^2$, equal to zero), and the different symbols denote experiments of different durations ($T$). (Right) Sampling standard deviations for estimates of $s$ for the case of $\sigma_s^2 = 0$, from simulations as noted above for three values of $N_e$, four of $s$, and a sample size of 100, compared with the theoretical expectation, equation (10). The diagonal dashed line denotes points of perfect agreement, and many symbols cannot be seen as they overlie each other on this line.

sampling variance of $\zeta$, for $T > 10$ or so, the sampling variance of $\widehat{s}$ is inversely related to the product of the sample size and the cube of the number of temporal samples. Thus, for a fixed investment in the total amount of genotyping that can be done, which is proportional to $Tn$, there is a very strong premium on extending the experiment in time, as the expected standard error of $\widehat{s}$ will be inversely proportional to $1/T$.

One can go further and consider the overall design necessary to detect a nucleotide with mean selection coefficient $\mu_s$. Assuming $\sigma_s^2$ is small relative to the sampling-error term in equation (10), which seems likely for most reasonable scenarios, the minimum sampling variance reduces to $\simeq 6/[T^3 np_0 (1 - p_0)]$. To detect a selection coefficient at the 5% significance level, one then requires $24/[T^3 np_0(1 - p_0)] < \mu_s^2$. The greatest power is achieved with high allele frequencies, so letting $p_0 = 0.5$, the critical value for detection in this case is $T^3 n = 96/\mu_s^2$, which implies $T^3 n > 10^6$ for $\mu_s = 0.01$, and $> 10^8$ for $\mu_s = 0.001$. Assuming a moderate sample size of $n = 100$, the critical experimental durations in these two cases become 21 and 100 consecutive generations of allele-frequency estimation. For a rarer allele with frequency $p_0 = 0.1$, these critical values become 2.8× larger.

The key point here is that when selection is weak, as is generally the case at the nucleotide level, its detection using temporal series of data demands very long surveys. Increasing the sample size helps, but in expanding $n$ to 1,000, the above critical $T$ values decline by only ~50%, and temporal variance in the selection coefficient will make such an enterprise more demanding. If one simply desires an estimate of the average absolute value of $\mu_s$ over a large sample of sites (e.g., particular sites within codons at particular frequencies), the sampling variance of the mean estimate is given by equation (10) divided by the number of sites jointly evaluated.

How much accuracy in estimation would be lost if one simply relied upon a two-point estimate of $\widehat{s}$,

$$\widehat{s}_2 = \frac{\zeta_0 - \zeta_T}{T}, \tag{11}$$

rather than using allele-frequency estimates at each time point? From Lynch (1987), the expected sampling variance of the two-point estimate is

$$\sigma^2(\widehat{s}_2) = \frac{1}{T^2 n}\left[\frac{1}{p_0(1 - p_0)} + \frac{1}{p_T(1 - p_T)}\right]. \tag{12}$$

Assuming $\sigma_s^2 \simeq 0$ and $T \gg 1$, the ratio of the sampling variance in this case relative to that with a full survey is

$$\frac{\sigma^2(\widehat{s}_2)}{\sigma^2(\widehat{s}_T)} \simeq \frac{T}{6}\left[1 + \frac{p_0(1 - p_0)}{p_T(1 - p_T)}\right]. \tag{13}$$

The minimum improvement gained by the full survey is therefore a reduction in the standard error of the estimate $\widehat{s}$ by a factor of $(T/6)^{1/2}$, that is, 2× with $T = 24$, and 4× with $T = 96$. In the limit of weak selection and/or short survey duration, such that $p_0(1-p_0) \simeq p_T(1-p_T)$, the inflation in sampling variance with the simpler method is a factor of $\simeq T/3$, whereas as the allele frequency approaches loss or fixation, that is, $p_T(1-p_T) \rightarrow 0$, the inflation factor can exceed $T$.

Equation (10) can also be used to evaluate the consequences of more intermediate sampling schemes. Rather than sampling each of $(T+1)$ consecutive generations, one could skip various generations, so that the duration of each sampling interval is $D$ (rather than 1 or $T$) generations. The expected sampling variance of $\widehat{s}$ is then obtained by dividing equation (10) by $D$ and substituting the number of multigenerational time intervals, $T'$, for $T$. For $T$ divisible by $D$, the inflation in the sampling standard error is $\simeq \sqrt{D}$. As an example, for a full survey with $T = 49$ and $D = 1$, from equation (10), the expected sampling variance is $\simeq 0.000050/[np_0(1-p_0)]$. Keeping $np_0(1-p_0)$ constant, and reducing the overall effort by half by skipping single generations, $T' = 24$ and $D = 2$, and the expected inflation of the standard error of $\widehat{s}$ is 1.5×. With $T' = 12$ and $D = 4$ (skipping periods of three generations), the expected inflation is 2.1×, and with $T' = 6$ and $D = 8$, the expected inflation is 2.6×. From equation (13), the expected inflation in the extreme case of sampling at just the starting and ending points (equivalent to a 25-fold reduction in effort) is $\sim\sqrt{T/3} = 2.9$. The key point here is that, for a given total survey duration, the improvement in the accuracy of estimation of $\mu_s$ with increased frequency of sampling is relatively small compared with the increase in effort.

## Estimator for the Variance of Selection Coefficients

To estimate the variance in true selection coefficients among generations, $\sigma_s^2$, we note that the estimated selection coefficient in interval $i$ can be partitioned as

$$\widehat{s}_i = s_i + e_i, \tag{14}$$

where $s_i$ is the true selection coefficient operating on the site in generation $i$, and $e_i$ is the estimation error resulting from finite sample size. From this, it follows that an estimator of the variance in $s$ among intervals is

$$\widehat{\sigma}_s^2 = \mathrm{Var}(\widehat{s}) - \mathrm{Var}(e), \tag{15}$$

where $\mathrm{Var}(e)$ is the average sampling variance of the $s_i$. Here, we focus on the most fine-grained sampling scheme of $T$ intervals of single-generation duration, as this will yield estimates of $\sigma_s^2$ with the highest degree of accuracy. To obtain an estimate of $\mathrm{Var}(\widehat{s})$ unfettered by nonindependence problems, we focus on the variance of $s_i$ estimates obtained for nonoverlapping time intervals, for example, $(0,1), (2,3), \ldots, (T-2, T-1)$. For even $T$, this yields $\tau = T/2$ estimates, with

$$\mathrm{Var}(\widehat{s}) = \frac{1}{\tau - 1}\sum_{i=1}^{\tau}(\widehat{s}_i - \bar{s})^2 \tag{16}$$

being the estimated variance among interval-specific sample estimates of $s_i$. From equation (11) with $T = 1$, estimates of $s_i$ based on adjacent generations are simply equal to the difference in estimates of $\zeta$ across the interval.

To estimate the average sampling variance, we utilize all of the data but allow for nonindependence of adjacent $s_i$ estimates (owing to allele frequencies at shared time points),

$$\mathrm{Var}(e) = \frac{1}{T}\sum_{i=1}^{T}\sigma^2(\widehat{s}_i) - \frac{2}{T(T-1)}\sum_{i<j}^{T}\sigma(\widehat{s}_i, \widehat{s}_j) \tag{17}$$

(from Lynch and Walsh 1998, p. 845). Note that the second term, which accounts for nonindependent estimates of the selection coefficient, has nonzero entries only for pairs of estimates in adjacent intervals. From Lynch (1987, eq. 12),

$$\begin{aligned} \sigma^2(\widehat{s}_i) &= \phi_{i-1} + \phi_i \\ \sigma(\widehat{s}_i, \widehat{s}_{i+1}) &= -\phi_i \end{aligned} \tag{18}$$

with $\sigma(\widehat{s}_i, \widehat{s}_j) = 0$ for $j \neq i+1$, and $\phi_i = 1/[2n_i\widehat{p}_i(1-\widehat{p}_i)]$, where $n_i$ is the number of diploid individuals in the $i$th sample (the two being removed in the case of haploidy). Substituting into equation (17), for the situation in which a string of $T$ consecutive estimates of $s_i$ is available,

$$\mathrm{Var}(e) = \frac{1}{T}\left(\phi_0 + \phi_T + \frac{2T}{T-1}\sum_{i=2}^{T}\phi_{i-1}\right). \tag{19}$$

Solving equations (16 and 19), and applying to equation (15) then provides an estimate of the variance in the selection coefficient for a nucleotide site.

Computer simulations incorporating generational episodes of selection and random genetic drift, with $\mu_s = 0$, were used to determine the bias and sampling error associated with this estimator of $\sigma_s^2$ (fig. 3). Two points are immediately apparent. First, the estimates for $\sigma_s^2$ tend to be downwardly biased, particularly when initial allele frequencies are low and sample sizes are on the order of 100 or smaller. This bias becomes negligible when sample sizes are as large as 1,000. However, even in the latter case, and even for the long experimental durations illustrated, an unbiased estimate of $\sigma_s^2$ cannot be achieved if $\sigma_s^2 < 10^{-4}$. Given that the latter implies a standard deviation of $s$ of 0.01, which may be beyond what operates at most nucleotide sites, the implication is that achieving accurate estimates of $\sigma_s^2$ at single-nucleotide sites is nearly unattainable without enormous sample sizes and survey durations.

Second, of even greater concern is the coefficient of variation (CV) of estimates of $\sigma_s^2$, which is virtually always >1.0 and often as high as 500. With a sampling CV of 1.0, if one

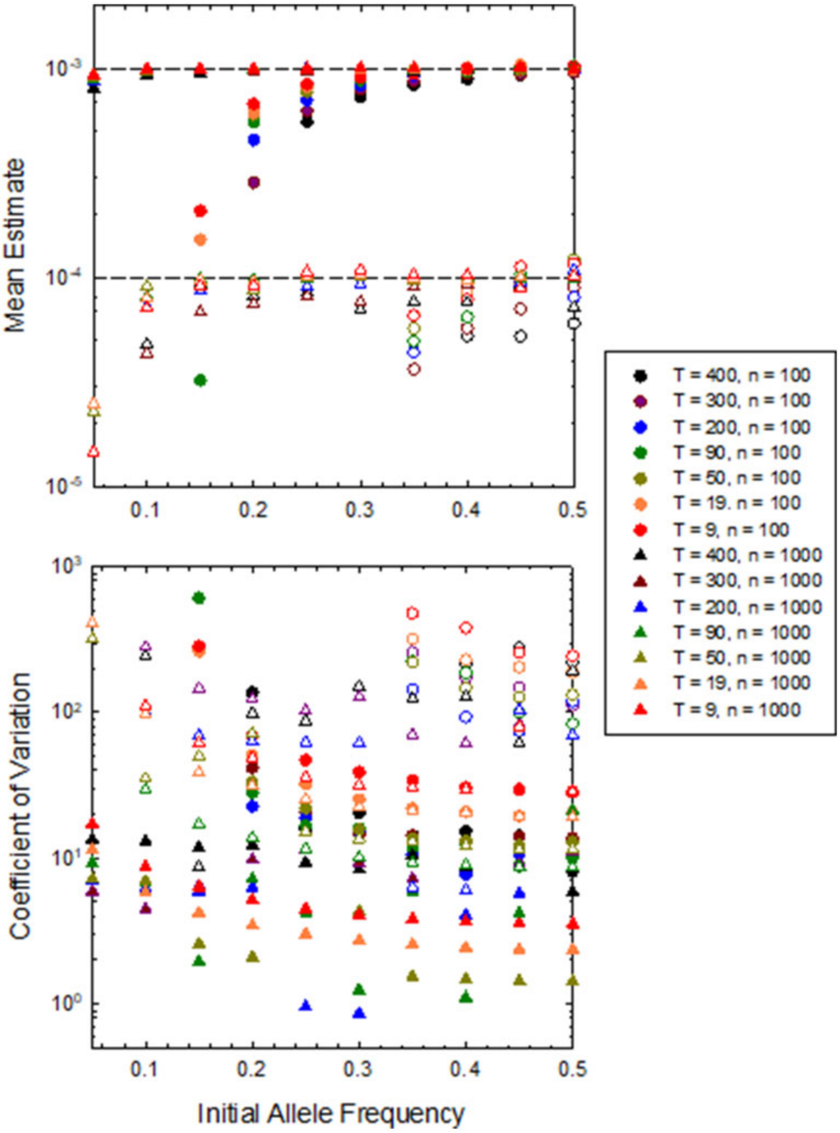

Fig. 3.—Mean and CV of estimates of $\sigma_s^2$ for series of samples taken at $T+1$ consecutive time points, each involving sample sizes of $n = 100$ or 1,000 diploid genomes. Results are given for a range of initial allele frequencies, each based on $10^6$ simulations with an effective population size of $10^8$ individuals, ensuring essentially no genetic drift on the time scale of the analyses, and mean selection coefficient $\mu_s = 0.0$. Closed points refer to situations in which $\sigma_s^2 = 10^{-3}$, whereas open points are for $\sigma_s^2 = 10^{-4}$. Data points are excluded for some cases at low allele frequencies where the mean estimates of $\sigma_s^2$ were negative.

wanted an average estimate of $\sigma_s^2$ pooled over sites to have a standard error <0.1 of the mean, 100 sites would need to be pooled, and with a per-site sampling CV of 500, this same level of accuracy would require the pooling of 25,000,000 sites.

## Estimation of Short-Term $N_e$

The effective size of a population ($N_e$) governs a wide range of evolutionary features, including the probability of fixation, levels of standing molecular variation, patterns of linkage disequilibrium, and the range of selection coefficients of mutant alleles that are effectively neutral (Charlesworth 2009; Walsh and Lynch 2018). Unfortunately, $N_e$ is also an exceptionally difficult population-genetic parameter to estimate directly. Almost all direct approaches rely on estimates of allele-frequency change, which invariably have very high sampling variances. Nonetheless, as studies of temporal changes in allele frequencies have become increasingly common, there has been a resurgence of interest in using such data to estimate $N_e$.

The motivation for the general approach is that random genetic drift generates stochastic changes in neutral allele frequencies to a degree determined by $1/(2N_e)$ in diploids (and $1/N_e$ in haploids). With suitable correction for the additional variation in allele-frequency estimates induced by sampling, observed divergence can then be used to infer $N_e$ provided the variation evaluated is neutral (Krimbas and Tsakas 1971). Often, a method of moments is used to infer $N_e$ from the observed allele-frequency changes and the sample sizes (Nei and Tajima 1981; Pollak 1983; Waples 1989), although likelihood-based procedures have also been suggested (Wang 2001; Bollback et al. 2008; Malaspinas et al. 2012; Hui and Burt 2015; Ferrer-Admetlla et al. 2016). Almost all applications have been confined to small populations, with $N_e$ typically well below 5,000. In this case, with suitable numbers of markers and time spans between samples, reasonable $N_e$ estimates (often within 10% of the expected values) are obtainable.

Because most natural populations are thought to have $N_e$ in the range of $10^4$ to $10^9$ (Lynch 2007), it is desirable to know in advance the joint influence of the numbers of sampled nucleotide sites, individuals, and generations separating samples on the ability to achieve a perceptible signal from drift. Is there an important tradeoff between the size of samples and the duration of sampling? Are there significant advantages of incremental sampling, as opposed to simply sampling at the starting and ending points of a survey? Do the frequencies of alleles at the sampled nucleotide sites have a substantial degree of influence?

To answer these questions, we proceed under the assumption that the investigator has settled on a set of molecular markers that suitably fulfill the expectations under neutrality based, for example, on the types of tests suggested above. The key indicator variable for estimating $N_e$ is then

$$\phi_{ijk} = \widehat{F}_{ijk} - \frac{1}{2n_{ij}} - \frac{1}{2n_{ik}}, \tag{20}$$

where $i$ denotes the nucleotide site, $j$ and $k$ denote two sampling times, and $n_{ij}$ and $n_{ik}$ are the associated numbers of sampled individuals (here assumed to be diploid). $\widehat{F}_{ijk}$ is a normalized measure of the observed allele-frequency change between the two time points,

$$\widehat{F}_{ijk} = \frac{(\widehat{p}_{ij} - \widehat{p}_{ik})^2}{\overline{p}_i(1-\overline{p}_i)}, \tag{21}$$

where $\widehat{p}_{it}$ is the estimated allele frequency at time $t$, and $\overline{p}_i = (\widehat{p}_{ij} + \widehat{p}_{ik})/2$. In essentially all of the following analyses, simulations show that the ratio of the variance of individual $\widehat{F}_{ijk}$ estimates to the squared average value is in the range of 1.90–2.00, as expected for a $\chi^2$-distributed variable (Nei and Tajima 1981). In the following, it will be assumed that the sample sizes are constant across time periods, so that $n = n_{ij} = n_{ik}$.

The logic underlying the use of $\widehat{F}_{ijk}$ is that the expected value of $\phi_{ijk}$ is very close to $t/(2N_e)$ provided the number of generations separating the two samples ($t$) is much smaller than $2N_e$ (Nei and Tajima 1981; Pollak 1983; Waples 1989), which will generally be the case for a study of natural populations. The resultant method-of-moments estimator of $N_e$ from two temporal samples is then

$$\widehat{N}_e = \frac{t}{2\overline{\phi}_i}, \tag{22}$$

where $\overline{\phi}_i$ is the average value of $\phi_i$ over all sites. (This averaging of the denominator term to obtain a single estimate of $N_e$ rather than obtaining an estimate for each site and then averaging minimizes the spurious sampling variation that can occur with individual ratios.)

Pollak (1983; see also Waples 1989) has usefully found that the expected sampling variance of $\widehat{N}_e$ is

$$\sigma^2(\widehat{N}_e) \simeq \left(\frac{8N_e^4}{L}\right)\left[\frac{1}{(2N_e)^2} + \frac{1}{tN\tilde{n}} + \frac{1}{(t\tilde{n})^2}\right], \tag{23a}$$

where $L$ is the number of sampled polymorphic sites (assumed to be independent), $\tilde{n}$ is the harmonic mean of the sample sizes (equal to $n$ in all that follows), and $N$ is the actual number of breeding adults (typically much larger than $N_e$; Walsh and Lynch 2018). Although an approximation, this formula provides immediate insight into the determinants of the precision of $N_e$ estimates by the temporal method. Because we expect $t\tilde{n} \ll 2N_e < N$, equation (23a) reduces to

$$\sigma^2(\widehat{N}_e) \simeq \left(\frac{2}{L}\right)\left(\frac{2N_e^2}{t\tilde{n}}\right)^2. \tag{23b}$$

The coefficient of sampling variation of $\widehat{N}_e$ (the ratio of the standard deviation of estimates to the expected value) is then

$$\mathrm{CV}(\widehat{N}_e) \simeq \frac{2N_e\sqrt{2/L}}{t\tilde{n}}. \tag{24}$$

The remaining issue is whether the preceding formulae, all of which involve approximations, do indeed yield reasonable estimates of $N_e$ and its sampling variance. This was evaluated by simulating three sources of binomial sampling of allele frequencies: random genetic drift between points $j$ and $k$ and individual sampling at the two time points. For any particular set of conditions, starting with a known allele frequency $p_j$, drift involved the random sampling of $N_e = N$ individuals to generate $p_k$, and then from these two true population values, the estimated frequencies $\widehat{p}_j$ and $\widehat{p}_k$ were obtained. For any particular set of $N_e$, $L$, $t$, and $n$, the mean and variance of $\widehat{N}_e$ (estimated by eq. 22) was obtained over a full range of allele frequencies. Over a range of $N_e = 10^3$ to $10^6$, the preceding theory is highly consistent with the results from simulations—equations (22 and 23b) give essentially unbiased estimates of the mean and variance of $N_e$ estimates, provided $tn\sqrt{L} \gg 5N_e$.

Equation (24) indicates that changing the survey duration $t$ or the sample size $n$ by the same factor has an equivalent effect on the accuracy of $N_e$ estimates by the temporal method. Such a scaling provides a logical basis for the optimal sampling design constrained by economic considerations and practicalities with respect to long-term sampling. The CV scales with $N_e/t$ because the variance of allele-frequency change resulting from drift (the signal that we wish to detect) depends on the reciprocal of this quantity. On the other hand, the CV scales only with the inverse of the square root of the number of loci sampled, which assuming complete genome sequencing is an absolute limit that cannot be modified by sampling.

This sampling theory highlights the difficulties in estimating $N_e$ in large populations by direct observations of allele-frequency changes. Nonsensical estimates of $N_e$ can be expected if $tn\sqrt{L} < 5N_e$, as then $\mathrm{CV}(\widehat{N}_e)$ exceeds 0.5, and even negative estimates will be likely. Suppose one desires to reduce the CV of an $N_e$ estimate to 0.1. According to equation (24), this requires $0.035tn\sqrt{L} > N_e$. Further supposing the luxury of $L = 10^6$ sites, the critical point reduces to $35tn > N_e$, which means that with $N_e = 10^6$, $tn$ must exceed 28,600, a steep task—sample sizes of nearly 3,000 if one were to rely on a temporal duration of 10 generations. On the other hand, with $N_e = 10^4$, $tn$ need only exceed 290, an achievable situation with today's sequencing costs (e.g., two samples of size 60, five generations apart).

The preceding estimator is based on a simple two-point comparison, and one might imagine that more resolution would be achievable by sampling at multiple intervals during time span $t$. There is, however, reason to expect a minimum payoff of such additional monitoring, the primary being that there is such a strong covariance in allele frequencies across generations that most of the information is contained within the contrast between starting and ending points (Hill 1972; Lynch 1987). Indeed, an analysis of a multigenerational maximum-likelihood (ML) procedure for estimating $N_e$ from temporal data (Hui and Burt 2015, table 1) shows that for the two-sample situation, the method-of-moments and ML procedures have almost identical sampling variance, and whereas the method of moments yields higher sampling variance when extended to three samples (over an equal total time span), the sampling error with the ML method is essentially identical whether two or three samples are applied. Such behavior can be understood from the structure of equation (23b), which shows that the sampling variance of $N_e$ scales inversely with $t^2$. Subdivision of $t$ into $x$ smaller episodes of equal length $\tau$ ($t = x\tau$) yields more independent estimates of $N_e$, which reduces the sampling variance linearly in $x$, but the sampling variance for each episode increases quadratically, the end result being that the sampling variance scales as $\sim 1/(x\tau^2) = x/t^2$.

To evaluate this matter more formally, allowing for larger $N_e$, the above kinds of analyses were performed three ways with six samples (providing five intervals of data): 1) using the simple starting and ending points, 2) performing two two-generation analyses (intervals 0–2, and 3–5), and 3) performing three single-generation analyses (intervals 0–1, 2–3, and 4–5). In the limit of $N_e \gg tn$, the preceding argument suggests that the sampling standard deviation of $N_e$ for these three schemes will scale approximately as 1.0:1.4:1.7. The simulation results indicate that the increase in the sampling standard deviation is reasonably consistent with the scaling suggested above, but that the use of multiple estimates also leads to downward bias in the mean estimate. Both features may be a consequence of the positive covariance between seemingly independent samples. An allele-frequency change resulting from drift causes a correlated change in the denominator of consecutive applications of equation (21) via the effects on heterozygosity at the site. In principle, computationally intensive ML techniques might be constructed to alleviate this problem, but taken together, these and prior analyses suggest that there are minimal advantages to doing so.

## Discussion

To ease the presentation, the formulations developed above assumed a setting in which populations are sampled on a timescale equal to generation length but are readily extended to longer intervals. For example, if the sampling interval is equivalent to $D$ generations, to account for the temporal accrual of selection effects, the estimated $\bar{s}$ should be divided by $D$, and the estimated variance of $s$ should be divided by $D^2$. Likewise, the estimated $N_e$ would need to be multiplied by $D$, as the effects of drift accumulate linearly with the number of generations. The expressions for the standard errors of these estimates would all have to be scaled in the same way, leaving the coefficients of sampling variance unchanged. There are, of course, alternative sampling schemes such as the fortuitous acquisition of DNA samples from single individuals at uneven intervals in the past. However, these do not easily lend themselves to the sampling theory outlined above, as there is additional uncertainty on the number of generations between intervals and even on lines of descent.

Our results demonstrate that although population genomics offers opportunities for obtaining relatively accurate estimates of previously elusive but key population-genetic parameters, the sampling requirements for achieving such ends remain substantial. As one example, Karasov et al. (2010) have suggested that the current effective population size of *Drosophila melanogaster* exceeds $10^8$, more than an order of magnitude larger than prior estimates. If one wished to show by direct estimation that $N_e$ is significantly $>10^7$ by an order of magnitude, a standard error of the estimate of $N_e$ of about 45,000,000 would be required. From equation (20b), generously assuming $10^6$ informative neutral sites, this requires that the product of the duration of the survey

($t$, in generations) and the sample size at each point ($n$) exceed $6 \times 10^5$, which is equivalent to fully genotyping 60,000 individuals at two time points separated by 10 generations. Although such an extreme might be possible in the not too distant future by pooled population sequencing to $\sim 100,000$ $\times$ depth of coverage, pooled sequencing is compromised by sequencing errors that obscure small frequency changes, as well as by likely variation in the contributions of individuals of different sizes to the pooled DNA. Thus, estimation of large $N_e$ almost certainly requires information on individual genotypes, necessary for factoring out contributions from sequencing errors (Lynch et al. 2014; Maruki and Lynch 2015).

Although the methods outlined above involving temporal covariance of allele-frequency changes provide a formal basis for testing whether temporal series of data are consistent with a random-walk model, it should be noted that violations of the null model need not imply the action of selection. For example, in nonequilibrium situations, persistent migration into a sampling site can lead to consistent directional changes in allele frequencies across intervals. One of the most likely problems, perhaps nearly unavoidable in natural settings, is the presence of spatial heterogeneity within populations. Owing to physical limitations on all organisms, nonrandom dispersal can be expected to be the rule even in populations uninfluenced by migration. Even if neutral, microlocale (spatial) heterogeneity in allele frequencies will manifest as temporal heterogeneity in a series of samples that do not fully account for such structure. In a test tube of microbes, such microheterogeneity might be eliminated by thoroughly mixing the sample before sampling, but this is essentially impossible for natural populations. Given that the shifts in allele frequency generated by random genetic drift are very small when $N_e$ is large, even a small amount of any of these kinds of biases can be problematical.

Assuming an absence of these kinds of problems, Buffalo and Coop (2019) suggested that genome-wide patterns of selection can be inferred indirectly by estimating the covariance in temporal allele-frequency change of neutral markers. Their proposed measure is closely related to the genome-wide covariance method outlined above, except that they normalize the measures of frequency change by dividing by $p(1-p)$. The motivation of this method is the idea that the expected covariance of allele-frequency change is zero in the absence of linked selection, and that when the latter prevails, alleles with larger frequency changes consistently exhibit such behavior across intervals because they happen to be stochastically associated with linked variants with greater fitness effects. The magnitude of such associations depends on the recombination rate among sites and on the time interval between surveys.

It is difficult to project the likely values of the scaled covariance of frequency change with linked selection in natural populations, but preliminary calculations by Buffalo and Coop (2019) suggest that values well below $10^{-3}$ may be common. This could present substantial challenges. Rearranging equation (7) indicates that the expected standard error of the temporal covariance of change in the null situation of no linked selection is $\simeq 1/(n\sqrt{L})$, where $L$ is the number of polymorphic sites in the analysis. Assuming a sample size of $n = 100$, then if an observed measure of standardized covariance is $10^{-3}$, rejection of the null hypothesis at the 5% level would require $L > 400$, and for covariances of $10^{-4}$ and $10^{-5}$, the minimum number of sites grows to 40,000 and 4,000,000, respectively. These constraints are generous in that they assume unlinked loci. If analyses were restricted to bins of strongly linked region of markers, as might be the case in searches for chromosomal regions under various levels of selection, to a first-order approximation, $L \simeq 1$, and the expected standard error is $\simeq 1/n$. The minimum sample size for rejecting the null hypothesis is then on the order of twice the reciprocal of the observed measure of covariance of change. Thus, aside from the physical limitations of obtaining completely random samples, the method of Buffalo and Coop (2019) appears to be extremely demanding with respect to required sampling effort.

There is also an issue with the view that linked selection will always generate positive covariance of allele-frequency change across generations. Even assuming that the selection coefficients of all polymorphisms remain constant in time, it can be seen from equations (9a and 9b) that for any strength of selection, there is an inflection point in the rate of allele-frequency change at $p = 0.5$ such that on opposite sides of this point the expected rate of change (although always of the same sign) will go from an acceleration to a deceleration phase. This behavior, which is a simple consequence of the frequency dynamics of selected alleles being proportional to $p(1-p)$, means that even with constant selection, marker alleles in different frequency classes are expected to yield different covariances of frequency change ranging from positive to near-zero to negative. The extent to which such complications will obscure the expected signature of linked selection will depend on the site-frequency spectrum of selected alleles and the relative incidence and magnitude of positive and purifying selection. This raises questions about the interpretation of the covariance of allele-frequency change, including whether the neutral expectation of zero overall covariance of change is also likely to arise emerge under some conditions of selection.

Computationally intensive methods involving likelihood and Bayesian frameworks are becoming increasingly popular routes to estimating population-genetic parameters such as $N_e$ and $s$, separately or as the product $N_e s$ (Bollback et al. 2008; Malaspinas et al. 2012; Foll et al. 2015; Ferrer-Admetlla et al. 2016). So far, the utility of these methods has been primarily confined to situations involving very small $N_e$ ($< 10^4$) and large $s$ (>0.05), conditions that do not commonly occur in natural populations of cellular organisms. A potential advantage of likelihood methods that arises with large $N_e s$ or $t/N_e$ is that these conditions raise the possibility of allele-frequency estimates of 0.0 or 1.0, especially if sample sizes are small. Such extremes lead to undefined method-of-moment estimators but can in principle be factored into likelihood functions that attempt to account for the full sampling distribution of alleles. Nevertheless, the latter types of methods do not supersede those outlined above or the statistical limitations that we have highlighted. Indeed, even when $s$ is estimated in a Bayesian framework, the background $N_e$ is still frequently estimated with a genome-wide application of the method of moments, equations (20–22), often without prior testing for the neutral behavior of the underlying polymorphisms.

## Acknowledgments

This work was supported by NIH (Grant Nos. R01-GM101672 and R35-GM122566) and NSF (Grant No. DEB-1257806). We thank J. Jensen for helpful comments.

## Appendix

### Sampling Variance of the Mean Selection Coefficient

As noted in the text, letting

$$\zeta_t = \ln\left(\frac{1-p_t}{p_t}\right), \tag{A1a}$$

the expected evolutionary trajectory of allele frequencies driven only by selection can be expressed as

$$\zeta_t = \zeta_0 - \sum_{i=0}^{t-1} s_i, \tag{A1b}$$

with $p_0$ being the frequency of the allele at time 0 (the first sampling point) and a positive value of $s_i$ implying that this allele is favored in interval $i$. In the following, it is assumed that there are $T$ consecutive samples of the allele frequency, indexed as $i = 0, 1, \cdots, T-1$, each based on a sample of $n$ diploid individuals. Throughout, estimates of parameters are designated by a $\widehat{}$ symbol, with the estimate $(\widehat{s})$ of the mean selection coefficient, $\mu_s$, being obtained by least-squares regression of the observed transformed variables $\widehat{\zeta}_t$ on $t$.

The concern here is with the variance in the estimate $\widehat{s}$ resulting from evolutionary variation in nature (which is a function of $\sigma_s^2$, the variance in $s$ around its expected value, assumed here to be randomly distributed over time with independent values across sampling time points) and sampling variance (which is determined by the number of sample points, $T$, and their sample size, $n$). From Lynch and Walsh (1998; eq. A1.20a, with a correction for sampling bias that reduces the effective sample size from $T$ to $T - 1$), the formula for the sampling variance of a regression coefficient leads to

$$\sigma^2(\widehat{s}) \simeq \frac{\sigma^2(\widehat{\zeta}_t)[1-\rho^2(\widehat{\zeta}_t, t)]}{(T-1)\sigma_t^2}, \tag{A2}$$

where $\sigma^2(\widehat{\zeta}_t)$ is the variance in the observed values of $\widehat{\zeta}_t$, $\rho(\widehat{\zeta}_t, t)$ is their correlation with $t$, and assuming $T$ consecutive samples,

$$\sigma_t^2 = \frac{(T-1)(T+1)}{12}. \tag{A3}$$

As the sample estimate $\widehat{\zeta}_t$ is equal to the sum of the population parameter $\zeta_t$ and a sampling deviation, its sampling variance has two components: 1) the evolutionary variance due to actual changes by selection and 2) the additional variance caused by finite numbers of individuals in the samples. The evolutionary variance is defined as

$$\sigma_e^2(\zeta) = E(\zeta^2) - [E(\zeta)]^2, \tag{A4}$$

where $E$ denotes an expectation. Using equation (A1b),

$$E(\zeta) = \frac{1}{T}\sum_{t=0}^{T-1} E(\zeta_t) = \zeta_0 - \frac{(T-1)\mu_s}{2}. \tag{A5}$$

The expected squared value is

$$E(\zeta^2) = \frac{1}{T}\sum_{t=0}^{T-1} E(\zeta_t^2) = \frac{1}{T}\sum_{t=0}^{T-1}\left(\zeta_0 - \sum_{i=0}^{t-1} s_i\right)^2, \tag{A6a}$$

and allowing for temporal variation in $s$, this reduces to

$$E(\zeta^2) = \zeta_0^2 - \zeta_0\mu_s(T-1) + \frac{\mu_s^2(T-1)(2T-1)}{6} + \frac{\sigma_s^2(T-1)}{2}. \tag{A6b}$$

Substituting equations (A5 and A6b) into equation (A4),

$$\sigma_e^2(\zeta) = \left(\frac{T-1}{2}\right)\left[\frac{(T+1)\mu_s^2}{6} + \sigma_s^2\right]. \tag{A7}$$

Note that this derivation assumes that drift has a negligible effect on allele-frequency change during the time interval $(0, T-1)$.

To obtain the additional variance in $\widehat{\zeta}_t$ associated with finite sample size, we make use of the fact that for each estimate $\widehat{\zeta}_t$,

$$\sigma_s^2(\widehat{\zeta}_t) \simeq \frac{1}{2np_t(1-p_t)} \tag{A8a}$$

(Lynch 1987), where $2n$ is the number of alleles assayed in a sample of $n$ diploid individuals, so that the average sampling variance over the time series is

$$\sigma_S^2(\widehat{\zeta}_t) \simeq \frac{1}{2Tn}\left[\sum_{t=0}^{T-1}\frac{1}{p_t(1-p_t)}\right]. \tag{A8b}$$

This expression can be further expanded by noting from equation (A1b) that

$$p_t = \frac{1}{1+f_t}, \tag{A9}$$

where $f_t = [(1-p_0)/p_0]e^{-\mu_s t}$. Using the fact that

$$\frac{1}{p_0(1-p_0)} = \frac{1}{p_0} + \frac{1}{1-p_0} \tag{A10}$$

then leads to

$$\sigma^2_S(\widehat{\zeta}_t) \simeq \frac{p_0}{2Tn(1-p_0)} \sum_{t=0}^{T-1} e^{st}(1+f_t)^2. \tag{A11a}$$

For weak selection such that the allele frequency does not change dramatically over time interval $T$,

$$\sigma^2_S(\widehat{\zeta}_t) \simeq \frac{T-1}{2Tnp_0(1-p_0)}. \tag{A11b}$$

The final component of equation (A2) requiring an expression involves the covariance between $\widehat{\zeta}_t$ and $t$,

$$\sigma(\widehat{\zeta}_t, t) = E(\widehat{\zeta}_t \cdot t) - E(\widehat{\zeta}_t) \cdot E(t). \tag{A12}$$

Noting that

$$E(\widehat{\zeta}_t \cdot t) = \frac{\zeta_0(T-1)}{2} - \frac{\mu_s(T-1)(2T-1)}{6}, \tag{A13}$$

and using equation (A5) and $E(t) = (T-1)/2$ leads to

$$\sigma(\widehat{\zeta}_t, t) = -\frac{\mu_s(T-1)(T+1)}{12}. \tag{A14}$$

One can see from equations (A3 and A14) that the slope of the regression $\sigma(\widehat{\zeta}_t, t)/\sigma^2_t$, has expected value $-\mu_s$.

Noting that equation (A2) expands to

$$\sigma^2(\widehat{s}) \simeq \left[\frac{1}{(T-1)\sigma^2_t}\right]\left[\sigma^2(\widehat{\zeta}_t) - \frac{\sigma^2(\widehat{\zeta}_t, t)}{\sigma^2_t}\right], \tag{A15}$$

and substituting with the above expressions leads to equation (10) in the main text.

## Literature Cited

Bollback JP, York TL, Nielsen R. 2008. Estimation of 2*Nes* from temporal allele frequency data. Genetics 179(1):497–502.

Buffalo V, Coop G. 2019. The linked selection signature of rapid adaptation in temporal genomic data. Genetics 213(3):1007–1045.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195–205.

Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper and Row.

Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D. 2016. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. Genetics 203(2):831–846.

Foll M, Shim H, Jensen JD. 2015. WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol Ecol Resour. 15(1):87–98.

Hill WG. 1972. Estimation of realised heritabilities from selection experiments. I. Divergent selection. Biometrics 28(3):747–765.

Hui TY, Burt A. 2015. Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. Genetics 200(1):285–293.

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

Kimura M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics 39(3):280–295.

King EG, Long AD. 2017. The Beavis effect in next-generation mapping panels in *Drosophila melanogaster*. G3 (Bethesda) 7:1643–1652.

Krimbas CB, Tsakas S. 1971. The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control-selection or drift? Evolution 25(3):454–460.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475(7357):493–496.

Liu X, Fu Y-X. 2015. Exploring population size changes using SNP frequency spectra. Nat Genet. 47(5):555–559.

Long H, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2(2):237–240.

Lynch M. 1987. Design and analysis of experiments on random drift and inbreeding. Genetics 120(3):791–807.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Assocs., Inc.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S. 2014. Population-genetic inference from pooled-sequencing data. Genome Biol Evol. 6(5):1210–1218.

Lynch M, et al. 2016. Genetic drift, selection, and evolution of the mutation rate. Nat Rev Genet. 17(11):704–714.

Lynch M, Walsh JB. 1998. Genetics and analysis of quantitative traits. Sunderland (MA): Sinauer Assocs., Inc.

Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. 2012. Estimating allele age and selection coefficient from time-serial data. Genetics 192(2):599–607.

Maruki T, Lynch M. 2015. Genotype-frequency estimation from high-throughput sequencing data. Genetics 201(2):473–486.

Nei M, Tajima F. 1981. Genetic drift and estimation of effective population size. Genetics 98(3):625–640.

Pearson K, Jeffery GB, Elderton EM. 1929. On the distribution of the first product moment-coefficient, in samples drawn from an indefinitely large normal population. Biometrika 21(1–4):164–201.

Pollak E. 1983. A new method for estimating the effective population size from allele frequency changes. Genetics 104(3):531–548.

Walsh JB, Lynch M. 2018. Evolution and selection of quantitative traits. Oxford: Oxford University Press.

Wang JL. 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet Res. 78(3):243–257.

Waples RS. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121(2):379–391.

Wright S. 1948. On the roles of directed and random changes in gene frequency in the genetics of populations. Evolution 2(4):279–294.

**Associate editor:** Brandon Gaut