Published in final edited form as:

Stat Appl Genet Mol Biol. 2015 February ; 14(1): 65–81. doi:10.1515/sagmb-2014-0030.

Inference for one-step beneficial mutations using next generation sequencing

Andrzej J. Wojtowicz,

Department of Animal Sciences, Washington State University, Pullman, WA 99164, USA, awojtowicz@alumni.uidaho.edu; and Department of Mathematics, University of Idaho, Moscow, ID 83844, USA; Department of Statistics, University of Idaho, Moscow, ID 83844, USA; and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA

Craig R. Miller, and

Department of Mathematics, University of Idaho, Moscow, ID 83844, USA; Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA; and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA

Paul Joyce

Department of Mathematics, University of Idaho, Moscow, ID 83844, USA; Department of Statistics, University of Idaho, Moscow, ID 83844, USA; and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA

Abstract

Experimental evolution is an important research method that allows for the study of evolutionary processes occurring in microorganisms. Here we present a novel approach to experimental evolution that is based on application of next generation sequencing. Under this approach population level sequencing is applied to an evolving population in which multiple first-step beneficial mutations occur concurrently. As a result, frequencies of multiple beneficial mutations are observed in each replicate of an experiment. For this new type of data we develop methods of statistical inference. In particular, we propose a method for imputing selection coefficients of first-step beneficial mutations. The imputed selection coefficient are then used for testing the distribution of first-step beneficial mutations and for estimation of mean selection coefficient. In the case when selection coefficients are uniformly distributed, collected data may also be used to estimate the total number of available first-step beneficial mutations.

Keywords

adaptive evolution; experimental evolution; next generation sequencing; one-step beneficial mutations; selection coefficients

Introduction

Next generation sequencing (NGS) technology is a new powerful tool for collecting genomic data. One of the areas where NGS has been applied are studies of experimental microbial evolution. Application of NGS in these types of studies has been reviewed by Brockhurst et al. (2011). The authors conclude that the main benefit of applying NGS in experimental evolution is ability to sequence larger, more complex genomes than in the case of traditional sequencing techniques. Here we describe a novel application of NGS in experimental evolution studies that allows for inference about first-step beneficial mutations in adaptation experiments.

Typically, adaptation experiments designed for detecting first-step beneficial have been conducted under conditions known as *strong selection weak mutation* (SSWM) (Gillespie, 1983, 1984, 1991). Under SSWM population size is small and most beneficial mutations that arise are lost due to drift. As a consequence, each beneficial mutation that survives drift grows to fixation, without interference with other first-step beneficial mutations. Therefore, in each replicate of an experiment conducted under SSWM only one first-step beneficial mutation is isolated (Rokyta et al., 2005; Barrett et al., 2006; Barrick et al., 2010), which then needs to be sequenced to determine its identity. Subsequently, a fitness assay has to be conducted to assess its selection coefficient. Here we consider a case where not one, but multiple first-step beneficial mutations are identified in a single replicate of an experiment. In order for this to occur, the experimental population must be large enough for multiple first-step beneficial mutations to occur concurrently. Such population dynamics is known as *clonal interference* (Gerrish and Lenski, 1998) or *concurrent-mutations regime* (Desai and Fisher, 2007). The application of NGS allows for identification of multiple first-step beneficial mutations; their selective advantage can be inferred from observed frequencies. This novel approach is expected to be more cost-effective and less time consuming than the traditional approach where only one mutation is isolated in one replicate.

The key assumption of the evolutionary model we employ is that the effect of clonal interference on evolutionary trajectories of each beneficial mutation is negligible since the sum of proportions of lineages carrying beneficial mutations is still small when sampling occurs. This assumption simplifies the model of population dynamics and allows us to consider each beneficial mutation independent of the other beneficial mutations.

In addition, our methods assume that the per site per generation total mutation rate (i.e., frequency of nucleotide substitutions regardless of their effect on fitness) is a known quantity. This may be estimated using external data or predicted from the studies of mutation rates in similar organisms (Drake, 1991; Drake et al., 1998). We also assume that selection coefficients of beneficial mutations are a sample from a probability distribution (Ohta, 1977; Kimura, 1979; Gillespie, 1983). To infer this distribution we propose application of the likelihood ratio test (Beisel et al., 2007) to imputed selection coefficients.

Since the experimental population is large, not only do beneficial mutations increase their frequencies, but neutral mutations also constantly accumulate throughout the experiment. Therefore, the proposed approach applies when selection is strong and, as a result, top

beneficial mutations increase their frequencies much faster than neutral mutations. This contrasts with *weak selection* where a large number of neutral mutations would be observed in addition to beneficial mutations. We do not recommend applying the proposed method when selection is mild.

Outline of the experimental protocol

Here we describe an evolutionary experiment that may be used to obtain a microbial population containing a proportion of individuals carrying first-step beneficial mutations. To begin, an initial inoculum that consists of the wild type individuals is placed in an altered environment where the current wild type has lower fitness than some number of potential one-step mutations. Then, a serial transfer protocol that involves periodic bottlenecks is applied in order to maintain population size at a certain level (e.g., Rokyta et al., 2005; Miller et al. 2011). In order to allow multiple one-step beneficial mutations to occur concurrently, the bottleneck size has to be large. Based on the results of simulations presented in a further section we determined that $N\mu$ has to be >1 (where N is the bottleneck size, and μ is per site per generation total mutation rate) in order to maintain population dynamics that meets one of the important assumptions of the model. As a result, bottleneck size in the experiment must be $>\mu^{-1}$.

We assume that population level sequencing occurs after sustaining a population for a fixed number of generations. Moreover, we assume that the sequencing detects mutations, but not their linkage. Thus, if two substitutions should arise in the same genome, the sequencing data cannot distinguish this from two substitutions in separate genomes. For this reason we focus on inference of first-step beneficial mutations rather than multiple-step mutations. To reduce the chance of observing multiple-step mutations we propose sampling the population when the proportion of first-step mutations is still small. In that case it can be assumed that the great majority of observed mutations are first-step mutations because the probability of multiple-step beneficial mutations arising and being observed is low.

Another important issue is composition of the initial inoculum. Ideally, the initial inoculum should contain only wild type individuals. However, since it is obtained by reproducing individuals, mutations may occur here as well. To minimize the proportion of beneficial mutations in the inoculum we recommend preparing it under conditions where the wild type is as near the optimum as possible, such that first-step mutations will either be neutral or deleterious.

Structure of the data

The data resulting from replicating n times the protocol described in the previous section can be summarized by the following matrix

$$\mathbf{X} = \begin{bmatrix} X_{1,1}=x_{1,1} & X_{1,2}=x_{1,2} & \dots & X_{1,n}=x_{1,n} \\ X_{2,1}=x_{2,1} & X_{2,2}=x_{2,2} & \dots & X_{2,n}=x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{Y,1}=x_{Y,1} & X_{Y,2}=x_{Y,2} & \dots & X_{Y,n}=x_{Y,n} \\ \dots & \dots & \dots & \dots \\ X_{Y+1,1}=0 & X_{Y+1,2}=0 & \dots & X_{Y+1,n}=0 \\ \vdots & \vdots & \ddots & \vdots \\ X_{r,1}=0 & X_{r,2}=0 & \dots & X_{r,n}=0 \end{bmatrix}.$$

Each entry in this matrix is a number representing observed frequency of a particular mutation in one replicate of an experiment. These frequencies are random variables, each of them denoted by $X_{i,j}$. Columns of the matrix $\mathbf{X}$ correspond to replicates of an experiment, while rows correspond to first-step beneficial mutations. Therefore, $\mathbf{X}$ has size n by r , where n is the number of replicates and r is the number of available first-step beneficial mutations. Among all r mutations, Y mutations were observed at least once in one or more replicates. Thus, in each row from 1 to Y there is at least one non-zero entry. Rows $Y+1$ to r represent beneficial mutations that were not observed in any replicate and those rows contain only zeros.

The model of population dynamics and sampling

Currently available NGS technology provides sequence of multiple fragments that cover the whole genome of an organism. For each fragment multiple reads are obtained. The number of times a region is sampled by a read is approximately independent of the number of times any other region is sampled. The number of reads at a particular site varies between replicates. However, for simplicity we assume that this number is constant and equal to the mean number of reads at this site over the n replicates (we denote it by k_j), while the frequency of a mutation is based on the actual number of reads with this mutation. In addition, for the unobserved first-step beneficial mutations we assume that the number of reads is equal to the mean number of reads across the genome (this number is denoted by k).

In a typical case, beneficial mutations are expected to be located on different fragments, which implies that their observed frequencies are independent, binomially distributed random variables. However, occasionally two or more first-step beneficial mutations can occur on the same fragment. In such a case the frequencies of mutations on the same fragment have multinomial distribution. Nevertheless, we make a simplifying assumption that distributions of all beneficial mutations are independent. Under these assumptions $X_{i,j}$ in the matrix $\mathbf{X}$ has a binomial distribution with parameters k_j and $\pi_{i,j}(t)$

$$X_{i,j} \sim \text{Bin}(k_j, \pi_{i,j}(t)),$$

where $\pi_{i,j}(t)$ is the probability that a read which covers the site of the i -th mutation contains that mutation. Note that $\pi_{i,j}(t)$ is equal to the proportion of the i -th mutation in the population in j -th replicate of an experiment at the time when sampling occurs. As a

consequence of the assumption that mutants grow independently, we consider each $\pi_{i,j}(t)$ as a random variable independent of the proportions of other beneficial mutations.

The distribution of $\pi_{i,j}(t)$ remains unknown, but we approximate it by its expected value. In order to obtain $E(\pi_{i,j}(t))$, we modeled the dynamics of the i -th mutant genotype in the absence of other beneficial genotypes. From the results of computer simulations of population dynamics (see Appendix A for the detailed procedure) we learned that the expected value of $\pi_{i,j}(t)$ converges to a fixed curve if $N\mu > 10$ and is close to that curve if $10 > N\mu > 1$ (Figure 1). We also observed that if N increases, then the influence of drift declines and variance of $\pi_{i,j}(t)$ decreases (Figure 2). Therefore, the results of simulations indicate that as $N\mu \rightarrow \infty$, $\text{Var}(\pi_{i,j}(t)) \rightarrow 0$ and $\pi_{i,j}(t) \approx E(\pi_{i,j}(t))$. This result is consistent with transition from stochastic mutation-selection dynamics of a population to deterministic dynamics described by Jain and Krug (2007). Since the dynamics of a population is deterministic when $N\mu \rightarrow \infty$, $E(\pi_{i,j}(t))$ can be derived using a deterministic model with no drift.

Our deterministic model for a single beneficial mutation i with selection coefficient s_i assumes discrete, non-overlapping generations. Competition between the mutation i and the wild type occurs in the absence of other beneficial mutations. In addition, the model does not account for second-step, deleterious, and reverse mutations. Letting $\pi_{D,i}(t)$ denote the proportion of the mutant i in a population at generation t , then the proportion of the wild type is given by $1 - \pi_{D,i}(t)$. The total number of individuals with the mutation i in generation t is equal to $N\pi_{D,i}(t)$ and the number of wild type individuals equals $N(1 - \pi_{D,i}(t))$. Let q denote the number of progeny produced by a wild type individual at the end of each generation and let $q = 2^{w_{wt}}$ (where w_{wt} is fitness of the wild type). Wild type individuals produce $N(1 - \pi_{D,i}(t))q$ progeny and individuals with mutation i produce $N\pi_{D,i}(t)q^{1+s_i}$ progeny at the end of generation t . There are $N(1 - \pi_{D,i}(t))q\mu$ individuals with mutation i among the progeny of the wild type. Therefore, the total number of progeny carrying the mutation i is equal to $N\pi_{D,i}(t)q^{1+s_i} + N(1 - \pi_{D,i}(t))q\mu$ and the total number of progeny of both individuals with mutation i and wild type individuals equals $N\pi_{D,i}(t)q^{1+s_i} + N(1 - \pi_{D,i}(t))q$. Since there is no drift, the proportion of individuals with mutation i in generation $t + 1$ is equal to the proportion of individuals with mutation i in progeny produced at the end of generation t . Hence, the proportion of genotype with mutation i at generation $t + 1$ is given by the following difference equation

$$\pi_{D,i}(t+1) = \frac{N\pi_{D,i}(t)q^{1+s_i} + N(1 - \pi_{D,i}(t))q\mu}{N\pi_{D,i}(t)q^{1+s_i} + N(1 - \pi_{D,i}(t))q} = \frac{\pi_{D,i}(t)q^{s_i} - \mu + \mu}{\pi_{D,i}(t)(q^{s_i} - 1) + 1},$$

which has a solution given by the following formula (see Appendix B)

$$\pi_{D,i}(t) = \frac{1}{q^{s_i} - 1} \left(\frac{q^{s_i t} (q^{s_i} + \mu - 1) (\mu + \pi_{D,i}(0) (q^{s_i} - 1))}{(q^{s_i} - 1) (1 - \pi_{D,i}(0)) (1 - \mu)^t + q^{s_i t} (\mu + \pi_{D,i}(0) (q^{s_i} - 1))} - \mu \right), \quad (1)$$

where $\pi_{D,i}(0)$ is the proportion of individuals with mutation i in the initial inoculum. As stated earlier, the above formula is equivalent to the expected proportion of individuals with mutation i , denoted later by $g(s_i)$, and we use it to approximate the observed proportion $\pi_{i,j}(t)$

$$\pi_{i,j}(t) \approx E(\pi_{i,j}(t)) = \pi_{D,i}(t) = g(s_i).$$

Note that this approximation becomes more accurate when $N\mu$ is large, because the variance of $\pi_{i,j}(t)$ decreases as $N\mu$ increases.

Imputation of selection coefficients and inference based on imputed values

In traditional experiments selection coefficients are determined using separate fitness assays preceded by isolation and sequencing of each mutation individually. Here we present a procedure that is much less complex. Under the proposed approach, selection coefficients are inferred from frequency data obtained through application of the NGS technology, and no additional assays are required. The procedure of imputing selection coefficients is independent of their underlying distribution and may be applied regardless of what that distribution is. Our model assumes that evolutionary trajectories of first-step mutations are independent, and therefore each selection coefficient can be imputed independently. This assumption is valid because the sum of proportions of all beneficial mutations at the time when sampling occurs is small. If the total proportion of beneficial mutations was large, individual trajectories would be significantly altered by competition with other beneficial mutations and those trajectories could no longer be considered independent.

The function $g(s_i)$ given by equation (1) is equal to the expected proportion of the i -th mutation in generation t , while the observed frequency of the i -th mutation in the j -th replicate ($x_{i,j}$) divided by the number of reads k_j is an estimate of the actual frequency in the j -th replicate. By the method of moments principle we assume that the expected proportion of a mutation with selection coefficient s_i is equal to the average observed proportion of that mutation over n replicates

$$g(s_i) = \frac{\sum_{j=1}^n x_{i,j}}{nk_i}.$$

To impute s_i the above equation has to be solved for s_i using a numerical method.

The next step in the analysis may include testing the distribution of fitness effects using the likelihood ratio test (LRT) (Beisel et al., 2007). This test assumes that the distribution of fitness effects is a tail distribution and therefore is subject to extreme value theory (EVT). According to EVT, any tail distribution is described by the Generalized Pareto Distribution (GPD) (Pickands, 1975) with shape parameter κ and scale parameter τ . The LRT tests whether fitness effects of beneficial mutations are a sample from the GPD with $\kappa=0$, which is equivalent to the exponential distribution, or from either truncated ($\kappa<0$) or heavy tailed ($\kappa>0$) distribution from the GPD family. A special case is the uniform distribution, which is

equivalent to GPD with $\kappa=-1$. Since low-effect beneficial mutations have very small probabilities of being sampled, the shifting procedure proposed by Beisel et al. (2007) needs to be applied for valid inference. In the original version, shifting was defined as moving the threshold of the tail distribution to the selection coefficient with the smallest observed effect. As a result, that mutation was removed from the sample while selection coefficients of the other beneficial mutations were decreased by the value of the shift. Here we shift the threshold of the tail distribution to the selection coefficient corresponding to a mutation with lowest observed frequency, i.e., a singleton. As a result, all singletons are removed from the sample and selection coefficients of remaining beneficial mutations are decreased by the value of selection coefficient of a singleton. However, since mutation rates of transitions and transversions differ, thresholds for these two types of mutations are not equal. Therefore, selection coefficients of transitions and transversions cannot be simply merged into one sample. One approach is to apply the transversion threshold to transitions as well so the shift is the same. The two sets of selection coefficients can then be combined and tested with the LRT. However, this approach involves removing non-singleton transitions that are below the transversion threshold. Alternatively, the LRT might be applied only to transitions, but in this case all transversions are excluded. Since both approaches involve eliminating additional mutations, we recommend using the one that removes fewer mutations in a particular case.

Imputed selection coefficients may also be used for estimation of the mean selection coefficient δ , which is equivalent to the first moment of the GPD given by (Joyce et al., 2008)

$$\delta = \frac{\tau}{1 - \kappa}.$$

Here τ may be estimated using shifted data. The scale parameter of the shifted distribution is given by (Castillo and Hadi, 1997)

$$\tau_{\text{shift}} = \tau + \kappa u,$$

where u is the size of shift. Since τ_{shift} can be estimated using the method of moments approach

$$\hat{\tau}_{\text{shift}} = \bar{s}(1 - \kappa),$$

the estimator of τ is given by

$$\hat{\tau} = \bar{s}(1 - \kappa) - \kappa u,$$

and therefore

$$\hat{\delta} = \bar{s} - \frac{\kappa u}{1 - \kappa}. \quad (2)$$

Since u is different for transitions and transversions, we calculate the final estimate of δ as a weighted mean of estimates of τ for these two types of mutations.

Maximum likelihood estimator of the number of beneficial mutations

The collected data can also be used for estimation of the total number of available first-step beneficial mutations (denoted by r). We derived the log likelihood function of r and δ for data represented by the matrix $\mathbf{X}$ (Appendix C)

$$\begin{aligned} \ln L(r, \delta) = & \ln \binom{r}{Y} + \sum_{a=1}^Y \sum_{b=1}^n \ln \left(\frac{k_a}{x_{a,b}} \right) + (r - Y) \ln \int_0^\infty (1 - g(s))^{\text{nk}} f(s) ds \\ & + \sum_{i=1}^Y \ln \int_0^\infty g(s_i)^{\sum_{j=1}^n x_{i,j}} (1 - g(s_i))^{\text{nk}_i - \sum_{j=1}^n x_{i,j}} f(s_i) ds_i. \end{aligned} \quad (3)$$

To obtain estimates of r and δ , the log likelihood function has to be maximized using a numerical method. This method assumes that κ is known. Therefore, the estimation procedure needs to be preceded by the LRT on imputed selection coefficients.

The log likelihood function given by equation (3) was derived under the assumption that mutation rate is equal for all first-step beneficial mutations. In order to account for differences in the mutation rate, the number of transition and the number of transversions can be estimated jointly as two separate parameters r_{TI} and r_{TV} . The likelihood function of these parameters is defined as a product of the likelihood function of r_{TI} and δ and the likelihood function of r_{TV} and δ . Therefore, the joint log likelihood function is given by

$$\ln L_{\text{TI,TV}}(r_{\text{TI}}, r_{\text{TV}}, \delta) = \ln L(r_{\text{TI}}, \delta) + \ln L(r_{\text{TV}}, \delta),$$

where L is given by equation (3).

The likelihood function contains integrals that have to be evaluated numerically in each iteration of the optimization procedure. In order to simplify the procedure of estimation, we can assume that δ is equal to $\hat{\delta}$ obtained with equation (2). Under this assumption, the term that contains the sum of the logarithm of integrals becomes a constant. As a result, a maximum likelihood estimate of r can be obtained by maximization of the sum of two terms

$$\ln L(r|\delta) \propto \ln \binom{r}{Y} + (r - Y) \ln \int_0^\infty (1 - g(s))^{\text{nk}} f(s) ds$$

or by solving the following equation for r

$$\sum_{i=0}^{Y-1} \frac{1}{r-i} + \ln \int_0^{\infty} (1-g(s))^{\text{nk}} f(s) ds = 0.$$

While the proposed MLE is designed for inference about first-step beneficial mutations and assumes that only mutations of this type are observed, neutral or second-step beneficial mutations may sometimes be sampled and mistakenly categorized as first-step beneficial mutations. Similar problem may also arise from sequencing errors. Since including observations that are not first-step beneficial mutations will bias estimates of r and δ , we developed an approach to address the problem. We assume that all neutral, second-step beneficial mutations, and sequencing errors are rare and they are observed only as singletons, i.e., they are sampled only once in the whole experiment. Since it is not possible to distinguish which singletons are first-step beneficial mutations, we remove all observed singletons from the analysis. Next, we apply the EM algorithm (Dempster et al., 1977), where the expected number of singletons is calculated under the current values of the parameters using the following formula (see Appendix D for derivation)

$$E(Z) = \text{nk} \int_0^{\infty} g(s) (1-g(s))^{\text{nk}-1} f(s) ds,$$

and then this number of singletons is appended to the data. In the next step, new estimates of the parameters are obtained by maximizing the likelihood function, and then the next iteration begins. The algorithm is terminated when $E(Z)$ does not change after updating the parameters.

Selecting parameters of the experimental protocol

Before conducting an experiment, parametric values of the laboratory protocol must be selected. These include: sampling time (t), the number of replicates (n), the target mean number of sequence reads per site (k), and the bottleneck size (N). The choice of these parameters is very important for data quality and performance of the estimation procedure. In particular, these parameters determine whether the assumptions of the estimation method are met. In this section we discuss how to select values of these parameters in order to obtain relevant data.

Number of replicates and number of reads

The number of replicates (n) and the number of reads (k) determine frequencies of first-step beneficial mutations in the observed data. Increasing both n and k results in collecting more data, but it simultaneously raises the probability of sampling neutral and second-step beneficial mutations as non-singletons.

In order to determine whether particular values of n and k may lead to sampling a large number of neutral mutations, we derived an approximate formula for the expected number of neutral mutations observed as non-singletons. We assume that each neutral mutation accumulates deterministically at the rate μ per generation. Therefore, the expected proportion of a neutral mutation in generation t is equal to $t\mu + \pi(0)$, where $\pi(0)$ is a

proportion of that mutation in the initial inoculum. Then, the expected number of neutral mutations observed more than once is given by the following formula (see Appendix E for derivation)

$$E(V) = r_{\text{neu}} [1 - (1 - t\mu - \pi(0))^{\text{nk}} - \text{nk}(t\mu + \pi(0))(1 - t\mu - \pi(0))^{\text{nk}-1}],$$

where r_{neu} is the total number of available neutral mutations. This formula is dependent on the total number of available neutral mutations (r_{neu}), which is typically an unknown quantity. Nevertheless, some assumptions about r_{neu} can be made based on external studies. For example, according to one study of random mutations in an RNA virus, 27% of examined mutations were neutral (Sanjuán et al., 2004).

The dynamics of second-step beneficial mutations is much more complex and therefore we do not provide a direct formula for assessing the expected number of observed second-step mutations. It is possible to simulate population dynamics that includes both first-step and second-step beneficial mutations and then calculate the average number of non-singleton second-step mutations over a large number of replicates. However, this approach is not expected to be effective since it would require making assumptions about multiple unknown parameters such as the number of first-step and second-step beneficial mutations, their mean selection coefficients and the shape parameter of the distribution.

Sampling time

Sampling time – measured in generations – directly affects the total proportion of beneficial mutations in a collected sample. If a population is sampled at the stage when the total proportion of beneficial mutations is very small, then no mutations or only a small number of mutations with low frequencies is expected to be observed. Furthermore, at that early stage, frequencies of beneficial mutations may still be similar to frequencies of neutral mutations, making them difficult to distinguish. On the other hand, if the total proportion of beneficial mutations is large, then clonal interference occurs, and the assumption of independent growth of mutations is violated. In such cases imputed values of selection coefficients will be biased. In addition, when the proportion of first-step mutations is high, the second-step mutations are more likely to occur. Excluding singletons from the analysis can help if almost all second-step beneficial mutations are observed as singletons. However, as the proportion of first-step beneficial mutations becomes larger, the probability of observing second-step mutations as non-singletons increases as well. As a result, second-step beneficial mutations may be observed as non-singletons if the proportion of first-step beneficial mutations is very large. In order to minimize bias of imputed selection coefficients and to decrease the probability of sampling second-step beneficial and neutral mutations, we therefore recommend sampling when the total proportion of mutations is in the range of 10%–25%. We used simulations to explore scenarios in which sampling occurred when the total proportion of mutations was much >25%, and we concluded that our approach was still robust in those cases (detailed results not shown). However, we recommend sampling when the total proportion of beneficial mutations is <25% since it is safely below the level where clonal interference and presence of second-step mutations may cause problems.

To select sampling time for a particular experiment we propose that researchers conduct a pilot study where the total proportion of mutations is determined across a series of time points. Based on the results, a time point at which the observed proportion is close to the desired proportion or exceeds some preselected threshold is chosen as a sampling time in the main experiment. For example, in the simulation study described in the next section, sampling time was selected as the first time point at which the total proportion of mutations exceeded 15%.

Population size

As stated earlier, population size should be $>1/\mu$ so that the mean proportion of a mutation converges to $g(s_i)$ given by formula (1). It was also noted that if N is much higher than $1/\mu$, the variance of proportions of beneficial mutations is low and approximation of these proportions by formula (1) becomes more accurate. However, if N is closer to $1/\mu$, the effect of drift reducing the number of second-step beneficial mutations in a population is stronger. For this reason we recommend choosing a bottleneck size that is not much $>1/\mu$, where μ refers to the lower of the two relevant mutation rates (i.e., mutation rate of transversions).

Evaluation of the proposed methods of inference

We evaluated performance of the proposed methods using data simulated with the algorithm analogous to the procedure described in Appendix A, but modified to emulate real experimental evolution. Specifically, instead of simulating single mutations with binomial sampling during bottlenecks, we generated multiple mutations and used multinomial sampling. Selection coefficients of beneficial mutations were generated from a probability distribution. Since experimental studies have suggested that fitness effects of beneficial mutation may commonly be exponentially (Sanjuán et al., 2004) or uniformly (Rokyta et al., 2008) distributed, we used these two distributions in our evaluation procedure. In order to make the evaluation procedure more realistic, we also generated neutral and second-step beneficial mutations. To account for differences in mutation rates between transitions and transversions, we classified mutations as transitions with the probability 0.294, reflecting the proportion of transitions among all possible non-synonymous single nucleotide mutations. We also varied the number of reads at a particular site in different replicates as well as the mean number of reads at different sites. To do this, the mean number of reads at the i -th site (denoted by k_i) was drawn from a normal distribution with mean k and standard deviation $0.1 k$, where k denotes the mean number of reads for all sites. The number of reads at the i -th site in j -th replicate was generated from a Poisson distribution with parameter k_i . Sampling time t was selected as a first time point in a preliminary run of the procedure at which the total proportion of mutations exceeded 15%. Additional information about specific values used in the simulation procedure is provided in Table 1.

We first evaluated the method of imputation of selection coefficients. In Figure 3 we present plots of imputed selection coefficients against their true values. These plots indicate that accuracy of imputation is good and it improves as the effect of mutations increases. In both uniform and exponential cases increasing the number of reads allows for detection of additional medium and low effect beneficial mutations. In the next set of simulations we

determined the mean squared error (MSE) of imputed selection coefficients. To obtain the MSE for a specific selection coefficient s , in each replicate of the simulation procedure we fixed the selection coefficient of one mutation at s , while all the other selection coefficients were generated from a probability distribution, either uniform or exponential. The MSE was then calculated based on imputed values obtained in 10,000 replicates of the procedure. In Figure 4 we present the results for mutations observed as non-singletons in at least 10% replicates of the procedure. In order to simplify interpretation of the results we present coefficient of variation (CV), which is defined as the square root of MSE divided by the actual value of selection coefficient. Here CV represents the mean error of imputation expressed as a percentage of the true value of selection coefficient. The results presented in Figure 4 confirm that performance of the proposed method is good, particularly when the effects of mutations are large. However, since the proposed method assumes that the observed frequency is equal to the expected frequency, imputation for low effect mutations with expected frequency < 2 is biased. These mutations are observed only in the cases when the actual frequency is greater than the expected and therefore their imputed selection coefficients are always inflated. This leads to the increase of MSE which is observed on the left side of the curves in Figure 4 and is also apparent in the departure of the observed points from the lines in Figure 3 at low values of s .

We also checked performance of the LRT applied to imputed selection coefficients. As noted earlier, the null hypothesis of the LRT is that the distribution of selection coefficients is exponential. Therefore, in order to verify type I error, we applied the LRT to data generated under the exponential distribution of selection coefficients. The actual type I error rate was determined by calculating the proportion of data sets for which the null hypothesis was rejected. The results of the simulation procedure (Figure 5, left panel) indicate that type I error of the LRT is moderately inflated, but it decreases as the number of reads increases. To evaluate the power of the LRT we used a similar procedure, but in this case data were generated under the alternative hypothesis. The alternative hypothesis of the LRT assumes that selection coefficients are distributed according to the GPD with $\kappa < 0$ or $\kappa > 0$. We evaluated the power of the LRT only for the uniform distribution (equivalent to GPD with $\kappa = -1$), which is a relevant and plausible alternative distribution of selection coefficients (Rokyta et al., 2008; Miller et al., 2011). According to the results presented in the right panel of Figure 5, the power of the LRT is high if selection coefficients are uniformly distributed.

Next, we checked performance of the estimators of r . The results of simulation suggest that accuracy is low in the exponential case. With CV often $> 50\%$, we do not recommend applying estimators of r when selection coefficients are exponentially distributed. However, the proposed estimators perform well when selection coefficients are sampled from a uniform distribution (Figure 6). Among the two estimators we considered here, the MLE based on the joint likelihood function of r and δ was more accurate. However, the computation procedure for this estimator is much more complex than for the simplified estimator. We observed that when frequencies of sampled mutations were large, computational problems often prohibited estimation. This did not occur when the other estimator was applied. Therefore, we recommend use of the MLE based on joint likelihood function, but if the computation procedure cannot be completed, the simplified MLE can be used instead.

Finally, we evaluated estimators of δ . The results of simulations indicate that in both exponential and uniform cases performance of the estimators does not change significantly as the number of reads increases from 100 to 500. When selection coefficients were generated from an exponential distribution, CV of the simplified estimator based on imputed selection coefficients [given by equation (2)] varied between 23% and 28%. Application of the estimator based on the joint likelihood function of r and δ was limited due to the computational problems described above. In the uniform case performance of the estimators of δ was better than in the exponential case. CV of the simplified estimator was between 6% and 6.7% for $r=50$ and between 4.5% and 5.4% for $r=100$. When δ was estimated jointly with r , CV was in the range of 3% to 4% for both values of r .

Conclusion

We have presented a novel approach to experimental evolution of microorganisms and inference about first-step beneficial mutations. We have briefly described the experimental protocol that involves application of the next generation sequencing technology. This protocol allows detection of multiple beneficial mutations and their frequencies in a single replicate of an experiment. We developed a theoretical model of population dynamics for this type of experiment. Based on this model we derived methods of statistical inference about first-step beneficial mutations. We proposed a method of imputing selection coefficients based on frequency data. Most importantly, our new method does not require conducting additional fitness assays in order to determine selection coefficients. The imputed selection coefficients may be used for testing the distribution of selection coefficients and for estimation of the mean selection coefficient. If the underlying distribution of selection coefficient is uniform, the number of available first-step beneficial mutations may be estimated as well.

We evaluated the proposed methods of inference using simulated data. The results indicate that, in most cases, these methods perform well. Therefore, despite several simplifying assumptions, the presented approach is robust. The logical next step will be to conduct an evolutionary experiment and analyze the data with the proposed method to assess how well it works and what modifications are needed.

An experiment in which population level sequencing was applied to an evolving population of yeast has already been conducted by Lang et al. (2013). However, their protocol was different from the protocol described here because $N\mu$ was much <1 . As a result, population dynamics observed in their experiment was highly affected by drift and more random compared to dynamics expected to be observed in an experiment proposed here. Therefore, our statistical methods cannot be applied to data obtained in that experiment. In addition, since selection was fairly weak, application of our methods most likely would not be possible even if $N\mu$ was >1 .

Acknowledgments

The authors would like to thank Holly A. Wichman for her comments on this research. This work was supported by a grant from the National Institutes of Health R01 GM076040-01.

Appendix

A Simulation of population dynamics for a single first-step beneficial mutation

Below we present a procedure of simulation of a population dynamics in which one beneficial mutation competes against the wild type. The relative frequency of a mutation is recorded in consecutive generations.

1. Let $N_{wt}(t)$ denote the count of the wild type and let $N_{mut}(t)$ denote the count of the mutant in generation t . Let $N_{wt}=N$ and $N_{mut}=0$ at $t=0$.
2. The wild type individuals produce $G_{wt}(t)=N_{wt}(t)2^{wt}$ progeny and mutants produce $G_{mut}(t)=N_{mut}(t)2^{wt(1+s)}$ progeny.
3. The probability that an offspring of the wild type is a mutant equals μ . Let $M(t)$ denote the number of mutants among the progeny of the wild type, then $M(t) \sim \text{Bin}(G_{wt}(t), \mu)$.
4. The population size is constant and equal to N . Therefore, at the end of each generation the number of progeny needs to be reduced to N by binomial sampling. Let

$$N_{mut}(t+1) \sim \text{Bin} \left(N, \frac{G_{mut}(t) + M(t)}{G_{wt}(t) + G_{mut}(t)} \right)$$

and

$$N_{wt}(t+1) = N - N_{mut}(t+1).$$

5. Record current relative frequency of a mutant. Go back to step 2.

B Solution to the difference equation for the mean proportion of a mutation

Here we provide a solution to the following difference equation

$$\pi(t+1) = \frac{\pi(t)(q^s - \mu) + \mu}{\pi(t)(q^s - 1) + 1}.$$

Let $a=q^s-\mu$, $b=\mu$, $c=q^s-1$, and $d=1$. Since

$$(d-a)^2 + 4bc = ((1-\mu) - q^s)^2$$

is always positive, we can apply the following method. Let

$$\rho = \sqrt{(d-a)^2 + 4bc} = \sqrt{(1-q^s+\mu)^2 + 4\mu(q^s-1)} = \sqrt{((1-\mu) - q^s)^2} = q^s + \mu - 1.$$

Note that in our case ρ is always positive since $q^s > 1 - \mu$, because $s > 0$, $w_{wt} > 0$, and $0 < \mu < 1$. Let

$$\eta = \frac{d - a + \rho}{2c} = \frac{1 - q^s + \mu + q^S - 1 + \mu}{2(q^S - 1)} = \frac{\mu}{q^S - 1}.$$

Then

$$\pi(t) = \frac{1}{x(t)} - \eta. \quad (4)$$

To calculate $x(t)$ the following first-order difference equation has to be solved

$$x(t+1) = \frac{d - \eta c}{\eta c + a} x(t) + \frac{c}{\eta c + a}.$$

Let

$$\alpha = \frac{d - \eta c}{\eta c + a}$$

$$\beta = \frac{c}{\eta c + a}$$

then

$$x(t) = \alpha^t x(0) + \frac{\beta(1 - \alpha^t)}{1 - \alpha},$$

where

$$x(0) = \frac{1}{\eta + \pi(0)}.$$

Therefore,

$$x(t) = \frac{\alpha^t}{\eta} + \frac{\beta(1 - \alpha^t)}{1 - \alpha} = \frac{(\eta c + a - d - c\pi(0))(d - \eta c)^t + c(\eta c + a)^t + c(\eta c + a)^t(\eta + \pi(0))}{(\eta c + a)^t(\eta + \pi(0))(2\eta c + a - d)}.$$

Next, using equation (4) we obtained the final formula for $\pi(t)$

$$\pi(t) = \frac{(\eta c + a)^t (\eta + \pi(0)) (2\eta c + a - d)}{(\eta c + a - d - c\pi(0))(d - \eta c)^t + c(\eta c + a)^t + c(\eta c + a)^t (\eta + \pi(0))} - \eta$$

$$= \frac{1}{q^s - 1} \left(\frac{q^{\text{st}} (q^s + \mu - 1) (\mu + \pi(0) (q^s - 1))}{(q^s - 1)(1 - \pi(0))(1 - \mu)^t + q^{\text{st}} (\mu + \pi(0) (q^s - 1))} - \mu \right).$$

C Likelihood function of r and δ

Result

The log likelihood function of r and δ is given by:

$$\ln L(r, \delta) = \ln \binom{r}{Y}$$

$$+ \sum_{a=1}^Y \sum_{b=1}^n \ln \binom{k_a}{x_{a,b}} + (r - Y) \ln \int_0^\infty (1 - g(s))^{\text{nk}} f(s) ds$$

$$+ \sum_{i=1}^Y \ln \int_0^\infty g(s_i)^{\sum_{j=1}^n x_{i,j}} (1 - g(s_i))^{\text{nk}_i - \sum_{j=1}^n x_{i,j}} f(s_i) ds_i.$$

Proof

Since all entries in the matrix $\mathbf{X}$ are assumed to be independent random variables distributed according to a binomial distribution, the probability of observing a particular matrix $\mathbf{X}$ is equal to the product of binomial probabilities of cells in the matrix multiplied by the combinatorial term that accounts for different possible order of observations. This probability is given by the following formula

$$P(X_{1,1}=x_{1,1}, X_{1,2}=x_{1,2}, \dots, X_{Y,n}=x_{Y,n}, X_{Y+1,1}=0, \dots, X_{r,n}=0 | s_1, \dots, s_r)$$

$$= \binom{r}{Y} P(X_{1,1}=x_{1,1}) P(X_{1,2}=x_{1,2}) \dots P(X_{Y,n}=x_{Y,n}) P(X_{Y+1,1}=0) \dots P(X_{r,n}=0)$$

$$= \binom{r}{Y} \prod_{a=1}^Y \prod_{b=1}^n \binom{k_a}{x_{a,b}} \prod_{i=1}^Y g(s_i)^{\sum_{j=1}^n x_{i,j}} (1 - g(s_i))^{\text{nk}_i - \sum_{j=1}^n x_{i,j}} \prod_{l=Y+1}^r (1 - g(s_l))^{\text{nk}},$$

where $g(s_j)$ is given by equation (1). Here we assume that selection coefficients $s_1, s_2, \dots, s_r$ are unobserved random variables with probability density function $f(s_j)$. Therefore, we define the joint probability of a matrix $\mathbf{X}$ and a vector of selection coefficients as

$$P(X_{1,1}=x_{1,1}, X_{1,2}=x_{1,2}, \dots, X_{Y,n}=x_{Y,n}, X_{Y+1,1}=0, \dots, X_{r,n}=0, s_1, \dots, s_r)$$

$$= \binom{r}{Y} \prod_{a=1}^Y \prod_{b=1}^n \binom{k_a}{x_{a,b}} \prod_{i=1}^Y g(s_i)^{\sum_{j=1}^n x_{i,j}} (1 - g(s_i))^{\text{nk}_i - \sum_{j=1}^n x_{i,j}} f(s_i) \prod_{l=Y+1}^r (1 - g(s_l))^{\text{nk}} f(s_l).$$

To obtain the likelihood function of r and δ we integrated out selection coefficients. Here we present the log likelihood function

$$\begin{aligned}\ln L(r, \delta) = & \ln \binom{r}{Y} \\ & + \sum_{a=1}^Y \sum_{b=1}^n \ln \binom{k_a}{x_{a,b}} + (r - Y) \ln \int_0^\infty (1 - g(s))^{\text{nk}} f(s) \text{d}s \\ & + \sum_{i=1}^Y \ln \int_0^\infty g(s_i) \sum_{j=1}^n x_{i,j} (1 - g(s_i))^{\text{nk}_i - \sum_{j=1}^n x_{i,j}} f(s_i) \text{d}s_i.\end{aligned}$$

D Expected number of singletons

Result

The expected number of singletons is given by the following formula

$$E(Z) = \text{nk} \int_0^\infty g(s) (1 - g(s))^{\text{nk}-1} f(s) \text{d}s.$$

Proof

Denote the number singletons by Z , then

$$Z = \sum_{i=1}^r I_i,$$

where I_i is an indicator function such that

$$I_i = \begin{cases} 1 & \text{if the } i\text{-th mutant was observed only once in one replicate} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(I_i | s_i) = P(I_i = 1) = P\left(\sum_{j=1}^n X_{i,j} = 1\right).$$

Since

$$\sum_{j=1}^n X_{i,j} \sim \text{Bin}(\text{nk}, g(s_i)),$$

where $g(s_i)$ is given by formula (1),

$$E(I_i | s_i) = \text{nk} g(s_i) (1 - g(s_i))^{\text{nk}-1}.$$

To obtain the expected value of I_i , the selection coefficient s_i has to be integrated out from the above formula. Therefore,

$$E(I_i) = nk \int_0^\infty g(s_i) (1 - g(s_i))^{nk-1} f(s_i) ds_i,$$

where $f(s_i)$ is the probability density function of s_i . Based on this result we obtained the expected value of the number of singletons

$$E(Z) = \sum_{i=1}^r E(I_i) = nk \int_0^\infty g(s) (1 - g(s))^{nk-1} f(s) ds.$$

E Expected number of neutral mutations observed more than once

Result

The expected number of neutral mutations observed more than once is given by the following formula

$$E(V) = r_{\text{neu}} (1 - (1 - t\mu - \pi(0))^{nk} - nk(t\mu + \pi(0))(1 - t\mu - \pi(0))^{nk-1}).$$

Proof

We assume that a neutral mutation accumulates in a population at the rate μ . Therefore, under the deterministic model, after t generations proportion of that mutation in a population is equal to $t\mu + \pi(0)$, where $\pi(0)$ is the proportion of that mutation in the initial inoculum. Since a site where a particular neutral mutation can occur is sampled nk times and the probability of sampling this mutation in one trial equals $t\mu + \pi(0)$, the number of times it is observed in n replicates is a binomially distributed random variable, which we denote by W

$$W \sim \text{Bin}(nk, t\mu + \pi(0)).$$

Then, the probability of sampling that mutation more than once (denoted by γ) is given by

$$\gamma = P(W > 1) = 1 - [P(W=0) + P(W=1)] = 1 - (1 - t\mu - \pi(0))^{nk} - nk(t\mu + \pi(0))(1 - t\mu - \pi(0))^{nk-1}.$$

Assuming that there are r_{neu} available neutral mutations, the number of neutral mutations observed more than once (denoted by V) has a binomial distribution

$$V \sim \text{Bin}(r_{\text{neu}}, \gamma).$$

Therefore, the expected number of neutral mutations observed more than once is given by

$$E(V) = r_{\text{neu}} \gamma = r_{\text{neu}} (1 - (1 - t\mu - \pi(0))^{nk} - nk(t\mu + \pi(0))(1 - t\mu - \pi(0))^{nk-1}).$$

References

- Barrett RDH, MacLean RC, Bell G. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biol. Lett.* 2006; 2:236–238. [PubMed: 17148371]
- Barrick JE, Kauth MR, Streltsov CC, Lenski RE. *Escherichia coli* rpoB mutants have increased evolvability in proportion to their fitness defects. *Mol. Biol. Evol.* 2010; 27:1338–1347. [PubMed: 20106907]
- Beisel CJ, Rokyta DR, Wichman HA, Joyce P. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. *Genetics.* 2007; 176:2441–2449. [PubMed: 17565958]
- Brockhurst MA, Colegrave N, Rozen DE. Next-generation sequencing as a tool to study microbial evolution. *Mol. Ecol.* 2011; 20:972–980.
- Castillo E, Hadi AS. Fitting the generalized Pareto distribution to data. *J. Am. Stat. Assoc.* 1997; 92:1609–1620.
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Met.).* 1977; 39:1–38.
- Desai MM, Fisher DS. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics.* 2007; 176:1759–1798. [PubMed: 17483432]
- Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA.* 1991; 88:7160–7164. [PubMed: 1831267]
- Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. *Genetics.* 1998; 148:1667–1686. [PubMed: 9560386]
- Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. *Genetica.* 1998; 102–103:127–144.
- Gillespie JH. A simple stochastic gene substitution model. *Theor. Popul. Biol.* 1983; 23:202–215. [PubMed: 6612632]
- Gillespie JH. Molecular evolution over the mutational landscape. *Evolution.* 1984; 38:1116–1129. [PubMed: 28555784]
- Gillespie, JH. The causes of molecular evolution. Oxford University Press; New York: 1991.
- Jain K, Krug J. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics.* 2007; 175:1275–1288. [PubMed: 17179085]
- Joyce P, Rokyta DR, Beisel CJ, Orr HA. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics.* 2008; 180:1627–1643. [PubMed: 18791255]
- Kimura M. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. USA.* 1979; 76:3440–3444. [PubMed: 16592684]
- Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature.* 2013; 500:571–574. [PubMed: 23873039]
- Miller CR, Joyce P, Wichman HA. Mutational effects and population dynamics during viral adaptation challenge current model. *Genetics.* 2011; 187:185–202. [PubMed: 21041559]
- Ohta, T. Molecular evolution and polymorphism. National Institute of Genetics; Mishima: 1977.
- Pickands J III. Statistical inference using extreme order statistics. *Ann. Stat.* 1975; 3:119–131.
- Rokyta DR, Joyce P, Caudle SB, Wichman HA. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genetics.* 2005; 37:441–444. [PubMed: 15778707]
- Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA. Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* 2008; 67:368–376. [PubMed: 18779988]
- Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA.* 2004; 101:8396–8401. [PubMed: 15159545]

Figure 1.

Dynamics of the mean proportion of a beneficial mutation under different values of $N\mu$. Data obtained by simulation in which a single genotype with selection coefficient $s=0.1$ competed against the wild type. According to the results, the mean proportion of a mutation converges to the function $g(s_i)$ when $N\mu \geq 10$.

Figure 2.

Dynamics of the standard deviation of the proportion of a beneficial mutation under different values of $N\mu$. Data obtained by simulation in which a single genotype with selection coefficient $s_f=0.1$ competed against the wild type. According to the results, as $N\mu$ increases, variance of the proportion decreases. It can be concluded that when $N\mu \rightarrow \infty$, the standard deviation goes to 0 and the observed proportion converges to its expected value which is given by the function $g(s_f)$.

Figure 3.

Plots of imputed selection coefficient vs. their true values. Selection coefficients were generated from either a uniform or an exponential distribution. Each plot contains combined results from 25 (uniform case) or 50 (exponential case) simulations. Selection coefficients were imputed for mutations observed 2 times. Performance of the proposed method is good since, in most cases, imputed values are close to the actual values of selection coefficients. Accuracy of imputation increases as the effects of mutations become larger.

Figure 4.

Coefficient of variation ($\sqrt{\text{MSE}}/s$) of the imputed selection coefficients. The results are based on 10,000 replicates of the simulation procedure. The presented curves are left-truncated at the point where mutations are observed in <10% (or 1000) replicates. Accuracy of imputation improves as the effects of mutations become larger. Increasing the number of reads from $k=100$ to $k=500$ allows for observing and imputing additional low or medium effect mutations, but it does not improve accuracy of imputation of big effect mutations since these are already accurately estimated when $k=100$.

Figure 5.

Type I error and power of the LRT when the test is applied to imputed selection coefficients. The power of the test was examined under the uniform distribution of selection coefficients. The test was conducted if the sample size was 5. According to the left panel, type I error of the test is moderately inflated. The right panel indicates that the power of the LRT is high when selection coefficients are uniformly distributed.

Figure 6.

Coefficient of variation ($\sqrt{\text{MSE}}/r$) of estimators of r when selection coefficients are a sample from a uniform distribution. The approach based on the joint likelihood function for r and δ provides more accurate estimates than the simplified method, but as the number of reads (k) increases the difference becomes smaller.

Table 1

Parameters used in the procedure of simulating data for evaluation of the estimators.

Number of available first-step beneficial mutations	$t=50,100$
Number of available second-step beneficial mutations	$r_2=\frac{2}{3}r$
Number of available neutral mutations	$r_{neu}=1000$
Fitness of the wild type (in doublings per generation)	$w_w=5$
Mean selection coefficient when $s_f\sim\text{UNIF}$	$\delta=0.15$
Mean selection coefficient when $s_f\sim\text{EXP}$	$\delta=0.1$
Mean selection coefficient of second-step mutations	$\delta_2=\frac{\delta}{2}$
Mean mutation rate	$\mu=10^{-6}$
Relation of mutation rate of transitions to mutation rate of transversions	$\mu_{TF}=8\ \mu_{TV}$
Mean number of reads per site	$k=100, 300, 500$
Number of replicates of an experiment	$n=10$
Bottleneck size	$N=10^7$
Initial inoculum size	$N_{ini}=10^7$
Number of generations between bottlenecks	1