Published in final edited form as:

Stat Appl Genet Mol Biol. ; 11(4): . doi:10.1515/1544-6115.1788.

Estimating the Number of One-step Beneficial Mutations

Andrzej J. Wojtowitz,

University of Idaho

Craig R. Miller, and

University of Idaho

Paul Joyce

University of Idaho

Abstract

Mutations that confer a selective advantage to an organism are the raw material upon which natural selection acts. The number of such mutations that are available is a central quantity of interest for understanding the tempo and trajectory of adaptive evolution. While this quantity is typically unknown, it can be estimated with varying levels of accuracy based on data obtained experimentally. We propose a method for estimating the number of beneficial mutations that accounts for the evolutionary forces that generate the data. Our model-based parametric approach is compared to an adjusted nonparametric abundance-based coverage estimator. We show that, in general, our estimator performs better. When the number of mutations is small, however, the performances of the two estimators are similar.

Keywords

one-step beneficial mutations; experimental evolution; adaptive evolution; natural selection; strong selection weak mutation; flask transfer experiments; estimating the number of classes

1 Introduction

The total number of beneficial mutations that can arise on the background of the same microbial genotype is a central quantity of interest for understanding microevolutionary processes. One approach for estimating the total number of one-step beneficial mutations is to employ experimental evolution coupled with the appropriate statistical methodology. In order to determine this quantity, an adaptation experiment must be replicated, and each mutation that arises as the first step of adaptation documented. Since the number of experiments that can be carried out is limited, usually some one-step beneficial mutations are missed.

Viruses and bacteria are examples of microorganisms that are commonly used in evolutionary experiments. One-step beneficial mutations are typically observed as a result of flask transfer experiments. In such experiments, a population is grown for a specific amount of time and then it is bottlenecked by transferring a fixed fraction (e.g. 10^{-4} of the population) to fresh media where growth starts again. This cycle is repeated and at some point a one-step beneficial mutation arises and sweeps to fixation, and it is evident by an

increase in fitness of the population (e.g. Rokyta, Joyce, Caudle, and Wichman 2005) or a rapid change in frequency of genetic marker assumed to be hitchhiking with a beneficial mutation (e.g. Rozen, de Visser, and Gerrish 2002, Barrett, MacLean, and Bell 2006, Perfeito, Fernandes, Mota, and Gordo 2007). Subsequently, a fitness assay is conducted to determine fitness of the mutant. An alternate method is to identify gain-of-function mutations by subjecting large, independent populations to a non-permissive growth condition such as high temperature (Bull, Badgett, and Wichman, 2000), a novel host (Ferris, Joyce, and Burch, 2007) or the presence of an antibiotic (Kassen and Bataillon, 2006, MacLean and Buckling, 2009, McDonald, Cooper, Beaumont, and Rainey, 2011). Again, fitness of isolated mutations is determined by fitness assay. Regardless of how beneficial mutations are obtained, if the number of beneficial mutations is to be estimated, the identity of the mutations must be observed through complete or partial genome sequencing (e.g. Rokyta et al. 2005, Bull et al. 2000, Ferris et al. 2007, McDonald et al. 2011).

In the case of flask transfer experiments adaptive evolution favors the more fit mutations, making them more likely to fix and thus more likely to be sampled than less fit alleles. In addition, transition / transversion rate differences cause some adaptive mutations to occur more often than others. Thus, there is a need for developing an estimation procedure that explicitly accounts for these unequal sampling probabilities. Here we develop methods for estimating the number of beneficial mutations specifically for flask passaging experiments. These methods complement previous work focused on transition / transversion rates differences (Ferris et al., 2007) that applies to gain-of-function experiments.

Estimating the number of beneficial mutations is an example of well studied problem of estimating the number of classes. Bunge and Fitzpatrick (1993) and Chao (2005) review a whole family of methods applicable to this problem. They distinguish two main groups of methods. The first group are methods based on sampling models. Among the sampling models listed by Bunge and Fitzpatrick, the *Multinomial Model's* assumption of sampling with replacement and fixed sample size is consistent with the procedure of sampling beneficial mutations by replicating an evolutionary experiment where one mutation is observed each time. Another group of methods that could be used for estimating the number of classes are data-analytic methods. These methods are based on curve fitting and extrapolation of a curve in order to determine the number of classes that were missed. However, these methods require large sample sizes and large number of classes for fairly accurate inference. In the case of estimation of the number of beneficial mutations, sample sizes and number of classes are usually too small to use such methods. Therefore, application of data-analytic methods for estimation of the number of beneficial mutations would be very limited.

Methods of estimation of the number of classes have been widely applied in wildlife biology for estimation of population sizes (e.g. Miller, Joyce, and Waits 2005). They have also been used in other areas such as numismatics (e.g. Holst 1981) and linguistics (e.g. Efron and Thisted 1976). Based on the methods of estimation of population sizes in wildlife biology, Ferris et al. (2007) derived estimators of the number of one-step adaptive mutations for gain-of-function experiments. In those experiments sampling probabilities of mutations do not

depend on their selective advantage, but only on their mutation rates. In effect, estimators derived by Ferris et al. (2007) do not account for differences in sampling probabilities due to natural selection and they cannot be applied to the data resulting from flask transfer experiments.

Here we present two parametric methods of estimation of the number of one-step beneficial mutations which are based on the assumptions of the Gillespie's model of adaptation (Gillespie, 1983, 1984, 1991). Both proposed estimators account for the effect of natural selection as well as unequal mutation rates. We compare their performance to the nonparametric abundance-based coverage estimator (Chao and Lee, 1992) that has been adjusted to the assumed sampling model and is therefore expected to perform better than other applicable nonparametric methods.

2 The model of sampling of beneficial mutations

Gillespie developed a model of adaptation (Gillespie, 1983, 1984, 1991), which was later extended by Orr (2002, 2003, 2005) and Joyce (Joyce, Rokyta, Beisel, and Orr, 2008). This model assumes "strong selection and weak mutation" (SSWM) conditions, meaning that $Ns \gg 1$ and $N\mu < 1$ (N is population size, s is the selection coefficient, and μ is the per site per generation mutation rate). Under SSWM conditions, adaptation involves the fixation of single-step nucleotide changes.

For a given DNA sequence, let r be the number of available one-step beneficial mutations. In a single process of adaptation only one of those r mutations fixes. As a result of replicating an experiment in which one-step mutations arise and fix, we observe a subset of the whole set of r available beneficial mutations.

Let Y be a random variable that denotes the number of beneficial mutations observed in n experiments. The distribution of Y is dependent on the sampling probabilities of available one-step beneficial mutations.

2.1 Simplified model

Assuming SSWM conditions, Gillespie (1984) derived the following formula for the probability of sampling the j -th beneficial mutation out of r available beneficial mutations

$$P_{r+1,j} = \frac{s_j}{\sum_{k=1}^r s_k} \quad (1)$$

where $\mathbf{s} = (s_1, s_2, \dots, s_r)$ are selection coefficients of mutations. Following Ohta (1977), Kimura (1979), and Gillespie (1983) we assumed that selection coefficients of mutations are a random sample from a statistical distribution. Gillespie pointed out that beneficial mutations fall in the right tail of the fitness distribution of all available one-step mutations, and therefore the distribution of fitness of beneficial mutations is subject to extreme value theory (EVT). According to EVT there are three domains of attraction that correspond to three types of tail distributions. Of these, Gillespie assumed that the distribution of fitness always belongs to the Gumbel domain. This assumption implies that fitness effects of

beneficial mutations follow the exponential distribution. However, Beisel, Rokyta, Wichman, and Joyce (2007) and Joyce et al. (2008) argued that two other domains of attraction (Weibull and Fréchet) can also occur, and the distribution of selection coefficients of beneficial mutations is then described by the Generalized Pareto Distribution (GPD), where the exponential distribution is a special case. Nevertheless, Joyce et al. (2008) discovered that some general predictions about one-step adaptation derived by Orr (2002) under assumption of the Gumbel domain are robust when the actual distribution of fitness belongs to the Weibull domain. Taking into consideration those results, we derived estimators of r under the assumption of the Gumbel domain of attraction; then we checked if they are robust when this assumption is violated.

Since selection coefficients are assumed to be a sample from the exponential distribution, sampling probabilities of beneficial mutations given by equation (1) have the symmetric Dirichlet distribution with parameter 1

$$P_{r+1,1}, P_{r+1,2}, \dots, P_{r+1,r} \sim \text{Dir}(1).$$

Therefore, the sampling model based on Gillespie's formula given by equation (1) turns out to be equivalent to the *broken stick* model of species abundances (Pielou, 1977). The sampling model based on equation (1) accounts for the effect of natural selection, but it does not account for unequal mutation rates. Hence, this is a simplified model of sampling.

2.2 Full model

The full model of sampling accounts for both natural selection and unequal mutation rates. It is based on the adjusted formula obtained by Rokyta et al. (2005)

$$P_{r+1,j} = \frac{s_j \mu_j}{\sum_{k=1}^r s_k \mu_k} \quad (2)$$

where μ_j is the mutation rate of j -th mutation. We assume that mutation rates $\mu_1, \mu_2, \dots, \mu_r$ are independent, identically distributed random variables which can assume one of two possible values: μ_{TI} or μ_{TV} , corresponding to transitions and transversions, respectively. For convenience, we use relative mutation rates scaled by μ_{TV} . As a result, the relative mutation rate that corresponds to transversions equals 1, while the relative mutation rate for

transitions equals $\frac{\mu_{TI}}{\mu_{TV}}$ and is denoted by the parameter m . Therefore, we assume that mutation rates are assigned using the following distribution

$$\mu_j \sim X_j(m-1)+1 \quad (3)$$

where $X_j \sim \text{Bernoulli}\left(\frac{1}{3}\right)$. The parameter of the Bernoulli distribution is equal to $\frac{1}{3}$ because there are twice as many possible transversions than transitions, and therefore the

probability that a randomly selected mutation is a transition equals $\frac{1}{3}$. Typically, parameter m is estimated using external data (Rokyta et al., 2005, Ferris et al., 2007).

In this section we have covered most of the notation for parameters and random variables used in our models. Table 1 provides a summary of these along with several important terms introduced later in the paper.

3 Estimators for the simplified model

The first proposed estimator of r is a method of moments estimator (MME). Let $g(r)$ be a formula representing the expected value of Y , i.e. $E(Y) = g(r)$, then by the method of moments principle we assume that

$$g(\hat{r}) = Y. \quad (4)$$

To obtain the MME for r , the above equation has to be solved for $\hat{r}$, resulting in $\hat{r} = g^{-1}(Y)$. However, under the full model of sampling which assumes both natural selection and unequal mutation rates, the exact formula for $g(r)$ could not be obtained. Instead, we derived the MME using the formula for the expected value of Y under the simplified model. Since this estimator is expected to be biased when applied to data generated under the full model, in the next section we derive a bias correction term that improves its performance.

The expected value of Y under the simplified model is given by the following formula (Keener, Rothman, and Starr 1987; see also appendices A, B, and C for details of the derivation)

$$E(Y) = \frac{rn}{r+n-1}. \quad (5)$$

To obtain the MME, we substituted the left side of equation (4) into formula (5) and we solved the resulting equation for r . In effect, the MME for the simplified model of sampling has the following form

$$\hat{r} = \frac{(n-1)Y}{n-Y}. \quad (6)$$

Note that the MME is equivalent to the estimator derived by Huillet and Paroissin (2009).

Keener et al. (1987) derived an estimator that applies in a more general case, when species total abundances or selection coefficients are assumed to be a sample from the gamma distribution. In the special case when selection coefficients are drawn from the exponential distribution, their estimator has the following form (Huillet and Paroissin, 2009)

$$\hat{r}_{\text{Keener}} = \frac{nY}{n - Y + 1} \quad (7)$$

which is similar to the MME. Both the MME and the Keener estimator are based on Y , which is a sufficient statistic for r under the simplified model (Keener et al., 1987). Under this model, the Keener estimator is the UMVUE (uniformly minimum variance unbiased estimator) if $n > r$. Computer simulations confirmed that the Keener estimator outperforms the MME when the simplified model is true, but they also showed that difference is small. Moreover, if data are generated under the full model, the MME turns out to be more robust and it performs slightly better than the Keener estimator.

In addition, in appendix D we derived the second moment of Y for the simplified model, using the formula obtained by Harris (1959)

$$E(Y^2) = \frac{rn}{r+n-1} \left(1 + \frac{(r-1)(n-1)}{r+n-2} \right).$$

Then, the variance of Y is given by the following formula

$$\text{Var}(Y) = \frac{rn}{r+n-1} \left(1 + \frac{(r-1)(n-1)}{r+n-2} - \frac{rn}{r+n-1} \right) \quad (8)$$

which is equivalent to the variance formula derived by Keener et al. (1987). We applied this formula in the interval estimation procedure described in section 8.

4 Bias correction for the MME

As stated earlier, the MME given by equation (6) and the Keener estimator given by equation (7) were derived under the simplified model, and therefore they do not account for unequal mutation rates. When applied to the data generated under the full model these estimators are likely to be biased, which means that the expected value of $\hat{r}$ is not equal to the true value of r . To improve performance of the MME, we determined approximate level of bias of that estimator using simulations and then we used the results to correct the MME for bias. The same approach might be used in the case of the Keener estimator. However, these two estimators are similar and they perform similarly. Therefore, we focused on one of them - the MME - which is slightly more robust when data is generated under the full model.

Denote the initial estimator of r given by equation (6) as $\hat{r}_0$ and the estimator with reduced bias as $\hat{r}$, then

$$\hat{r} = \hat{r}_0 - b(\hat{r}_0, m) = \frac{(n-1)Y}{n-Y} - b(\hat{r}_0, m)$$

where $b(\hat{r}_0, m)$ is a bias correction term.

Simulations to determine the level of bias were run for combinations of r , m , and n that were assumed to include those values usually occurring in real experiments. In particular, we determined bias for combinations of $r = 1, 2, \dots, 50$, $m = 1, 2, \dots, 20$, and $n = 20, 30, 40, 50$. Based on the results we concluded that bias is almost invariant when n changes from 20 to 50. Therefore, we assumed that bias is a function of r and m , and does not depend on n . It was also observed that if $m < 3$, then bias is low and bias correction is not necessary.

The next step was to define a functional relationship between bias and r and m . Based on the plots of bias against r (Appendix E), it was established that for fixed m and $r < 20$ bias can be approximated by a quadratic function of r

$$b(r) = \alpha_1 r^2 + \alpha_2 r + \alpha_3$$

and for $r \geq 20$ bias can be approximated by a linear function of r

$$b(r) = \beta_1 r + \beta_2.$$

Since the coefficients of the above equations are dependent on m , we used plots to determine the functional relationship between these coefficients and m (see Appendix F). Based on the plots we concluded that the coefficients can be approximated using rational or quadratic functions of m . To obtain the final version of the bias correction formula we applied the least squares method to fit data to the selected functions. Below we present a complete procedure of estimation of r using the MME with a bias correction term that accounts for unequal mutation rates.

1. Let $\hat{r}_0 = \frac{(n-1)Y}{n-Y}$.
2. Determine m using external data.
3. If $m < 3$ then $\hat{r} = \hat{r}_0$.
4. If $m \geq 3$ and $\hat{r}_0 < 20$ then

$$\hat{r} = \hat{r}_0 - \frac{(m-2.1)(m-67.24)}{3876(m+3.25)} \hat{r}_0^2 - \frac{(m-69.47)(m-1.92)}{3211} \hat{r}_0 + \frac{(m-56.29)(m-1.74)}{1247}.$$

(9)

5. If $m \geq 3$ and $\hat{r}_0 \geq 20$ then

$$\hat{r} = \hat{r}_0 + \frac{0.71(m-2.29)}{m+1.81} \hat{r}_0 + \frac{0.133(m-2.52)(m-44.27)}{(m+0.37)}. \quad (10)$$

Note that the bias correction term is a function of r . Here r is an unknown parameter and we substitute it by the estimator $\hat{r}_0$. But $\hat{r}_0$ underestimates r . Since bias increases with r , the bias correction term based on $\hat{r}_0$ does not remove bias completely. Further reduction of bias is possible using either an iterative procedure in which $\hat{r}_0$ in the bias correction term is substituted by an estimate from the previous iteration, or by multiplying an estimate by a factor that can be determined using simulations. However, further reduction of bias leads to increase of mean squared error (MSE) due to increase of variance of the estimator. As MSE is our primary criterion for evaluation of performance of the estimators, an estimator with lower MSE, i.e. the MME with a single bias correction, is preferred.

5 Simulation based maximum likelihood estimator

The second proposed method is the maximum likelihood estimator (MLE). The MLE is a value $\hat{r}$ such that the likelihood function of r attains the maximum at $\hat{r}$. The likelihood function of r is given by

$$L(r) = P(Y|r)$$

where r is an integer, and $r \leq Y$. For the simple model assuming equal mutation rates, the likelihood function derived by Keener et al. (1987) can be applied. But for the full model we use simulations to obtain an approximate value of the likelihood. In order to find the maximum of the likelihood function, our strategy is to calculate $P(Y|r=k)$ for consecutive values of k , starting from $k=Y$, and terminate the procedure when the maximum is identified.

The $P(Y|r=k)$ is approximated using a large number of simulated data sets. For example, in the evaluation procedure described in section 7 we simulated 3000 Y 's in each iteration. Below we present the procedure of simulating a single Y :

1. There are $r=k$ available one-step beneficial mutations. Draw selection coefficients of these mutations from the exponential distribution or any other distribution that is assumed to be a distribution of selection coefficients.
2. Assign mutation rates by sampling from the distribution defined by formula (3).
3. Calculate sampling probabilities of mutations using formula (2).
4. Draw a sample from the multinomial distribution with n trials, k outcomes, and sampling probabilities calculated in the previous step.
5. Let Y_{sim} be the number of mutations sampled at least once.

For each k this procedure is replicated and then, the approximation of $P(Y|r=k)$ is obtained by dividing the number of simulations for which $Y_{sim} = Y$ by the total number of simulations.

6 Abundance-based coverage estimator

Chao and Lee (1992) proposed a nonparametric abundance-based coverage estimator (ACE), which applies in the case when class sizes are unequal. Chao and Lee used the concept of sample coverage (C) to derive this estimator. Sample coverage is defined as the sum of the probabilities of the observed classes

$$C = \sum_{i=1}^r J_i \pi_i$$

where J_i is an indicator function such that

$$J_i = \begin{cases} 1 & \text{if } i\text{-th class was observed at least once} \\ 0 & \text{otherwise} \end{cases}$$

and π_i denotes the probability of sampling the i -th class. Note that J_i is a random variable, and therefore C is also a random variable. Chao and Lee used the “estimator” of C derived by Good (1953)

$$\hat{C} = 1 - \frac{f_1}{n}$$

where f_1 is number of observed singletons. Since C is a random variable, $\hat{C}$ is not an estimator in a classical sense. It is an estimator in the sense that $E(\hat{C}) \approx E(C)$. The estimator of the number of classes proposed by Chao and Lee (1992) has the following form

$$\hat{r}_{ACE} = \frac{Y}{\hat{C}} + \frac{n(1 - \hat{C})}{\hat{C}} \gamma^2.$$

The first term ($Y/\hat{C}$) is the coverage based estimator of the number of classes for the equal class sizes case. It is a lower bound for the estimated number of classes in the case when class sizes are unequal (Chao, 1984, 1987). The second term is a bias correction that depends on the coefficient of variation of class sizes (γ). Chao and Lee (1992) provide two nonparametric estimators of γ . However, coefficient of variation of class sizes can be determined analytically based on the assumed model of sampling of beneficial mutations. For the simplified model assuming equal mutation rates the coefficient of variation is equal to 1 because selection coefficients are drawn from the exponential distribution. Under the full model where sampling probabilities are given by formula (2) and mutation rates are drawn from distribution (3), the coefficient of variation of class sizes is given by the following equation

$$\gamma_m = \frac{\sqrt{5m^2 - 4m + 8}}{m + 2}.$$

Our simulations showed, however, that if $\gamma = 1$ is used instead of γ_m in the case when data is generated under the full model, then mean squared error of the ACE is lower. Since the ACE with $\gamma = 1$ performs better, we used this version of the estimator for evaluation described in the next section.

7 Evaluation of the proposed estimators

In this section we evaluate the method of moments estimator with bias correction (MME) and simulation based maximum likelihood estimator (MLE). We also compare their performance to the abundance-based coverage estimator (ACE), adjusted as described in the previous section. Our primary criterion for evaluation of the estimators is mean squared error (MSE), which is defined as an average squared deviation of an estimate from the true value of a parameter. This is equivalent to the sum of variance and the squared bias of an estimator. Additionally, we examine bias of the estimators and verify robustness of the MME and MLE in the case when the estimate of m is inaccurate.

We used simulations to determine MSE of the estimators for combinations of $m = 8$, $n = \{20, 30, 40, 50\}$, and $r = \{10, 40\}$. For each combination we simulated at least 10000 data sets using the procedure described in section 5, and then we estimated r for each simulated Y . Next, we used estimates of r to calculate approximated values of both components of MSE, i.e. variance and squared bias. In the graphs demonstrating the results of those simulations (figures 1 and 2) we present coefficient of variation (CV) which is defined as the square root of the MSE divided by the true value of r .

First, we determined MSE of the estimators using data simulated under the exponential distribution of selection coefficients (figure 1). According to the results, if r is small then MSEs of the MME and ACE are similar, while the MLE performs slightly worse. If r is large, the MME performs better than ACE and MLE, except the case when the number of replicates is much smaller than r . In such case the MSEs of the MME and ACE are about the same.

Next, we checked if the estimators are robust when the assumption of the exponential distribution of selection coefficients is violated. A plausible alternative to this assumption is the uniform distribution (Rokyta, Beisel, Joyce, Ferris, Burch, and Wichman, 2008). Therefore, we simulated selection coefficients of beneficial mutations using the uniform distribution and then determined MSE of the estimators using the same approach as earlier. According to the results presented in figure 2, the MME is robust since its performance does not worsen when the assumption of the exponential distribution of selection coefficients is not met. Unlike the MME, the MLE is not robust when the distribution of selection coefficients is different than exponential. It can be observed that when $r = 40$, MSE of the MLE is much higher compared to the case when the selection coefficients were generated using the exponential distribution. Also, note that the change of the distribution of selection coefficients from exponential to uniform leads to decrease of MSE of the ACE if r is large.

We next examined bias of the estimators. In figures 3 and 4 we present approximated level of bias divided by the true value of r . Notice that if the total number of available beneficial

mutations is small ($r = 10$), then bias of all estimators is low. If r is large ($r = 40$), differences in bias level are more substantial. When the assumption of the exponential distribution selection coefficients is met (figure 3), the MLE is the least biased estimator, the MME is moderately biased, and the adjusted ACE is characterized by strong negative bias. If selection coefficients are sampled from the uniform distribution (figure 4, right panel), bias of the MME and ACE estimator decreases, while bias of the MLE increases substantially.

In addition, we verified robustness of the MME and MLE when the estimate of m was inaccurate. We simulated data under $m = 8$, but we estimated r assuming that the estimate of m was equal to 4, 6, 10, or 12. This analysis indicated that the MME was robust even if the estimate of m was off by 4, i.e. $m = 4$ and $m = 12$ was used instead of $m = 8$ (detailed results not shown). In such cases MSE of the MME was increased usually by no more than 15%. In the case where the estimate of m was off by 2 ($m = 6$ and $m = 10$ was used while the actual value of m was 8), MSE was increased by at most a few percent. Moreover, despite using inaccurate estimates of m , the MME still performed better than the ACE when r was large, and similar to the ACE when r was small. Unlike the MME, the MLE was robust only if m was underestimated. In fact, underestimating of m led to decreasing MSE of the MLE. But if m was overestimated, MSE of the MLE increased considerably. According to the results of our simulations, when $m = 10$ was used while the actual m was 8, MSE of the MLE was increased by about 30%. In the same case when $m = 12$ was applied, MSE was increased by about 60%. Therefore, when the MLE is applied, we recommend using the lower bound on m .

Since the overall performance of bias corrected the MME is better than the other considered estimators, we applied it to a bacteriophage dataset (Rokyta et al., 2005). Rokyta et al. observed 9 different beneficial mutations in 20 replicates, while the mutation rate of transitions was estimated to be 8.5 times higher than the mutation rate of transversions. In this case the estimate of r obtained with the MME with bias correction equals 19.22. For comparison, the adjusted ACE applied to the same data set gave an estimate equal to 21.43, and the ML estimate was equal to 25.

8 Interval estimation

We present two methods of interval estimation for r . In the first method the 95% confidence interval for r is approximated using an analytical procedure, while the second method is an iterative procedure based on simulations of the distribution of Y given r . Below we describe these two methods and compare their performances.

In the first method, the 95% confidence interval for r is approximated using 2.5 and 97.5 percentiles of the distribution of $\hat{r}$ given r . Since $\hat{r}$ is a function of Y , the distribution of $\hat{r}$ given r is dependent on the distribution of Y given r . Therefore, as a first step the 2.5 and 97.5 percentiles of the distribution of Y given r are determined under assumption that $r = \hat{r}$. Then, the MM estimates of r are calculated for the obtained values. The resulting values are the estimates of the confidence limits. Since the distribution of Y given r is nearly symmetric, unimodal, and bell-shaped, 2.5 and 97.5 percentiles of that distribution are

approximated by the values that are two standard deviations from observed Y . The standard deviation of Y comes from taking the square root of the value calculated with formula (8).

Preliminary simulations showed that the coverage of confidence intervals obtained by strictly applying the method described above is close to 95% only in the case when mutations rates are equal ($m = 1$) and otherwise the actual coverage is often much below the intended 95%. This problem is caused by two factors. First, the formula for variance of Y was derived under the assumption of equal mutation rate, while the actual variance of Y is higher if this assumption does not hold. Second, the MME (which is used to calculate the confidence limits and is based on appropriate percentiles of the distribution of Y given r) is a biased estimator despite partial bias reduction with bias correction term. Therefore, confidence limits obtained with this method might also be biased. To circumvent these problems we adjusted the MME and variance of Y by multiplying them by factors. The following values for the factors were determined and validated using simulations:

1. Multiply estimate of variance of Y (formula (8)) by 1.2 if $2 \leq m < 6$ or by 1.3 if $m \geq 6$.
2. Multiply the MME with bias correction (formula (9) or (10)) by 1.1 if $3 \leq m < 5$, by 1.2 if $5 \leq m < 9$, by 1.3 if $9 \leq m < 13$, or by 1.4 if $m \geq 13$.

After applying these adjustments, the coverage is approximately 95% (see table 2 as an example).

The following provides a summary of our first method for estimating 95% confidence intervals for r :

1. Calculate variance of Y using formula (8) and adjust the obtained value as described above. In the formula for variance of Y substitute r by $\hat{r}$ adjusted as recommended above. Calculate s_Y , the standard deviation of Y , as a square root of variance of Y .
2. Calculate 2.5 and 97.5 percentiles of the distribution of Y given $r = \hat{r}$: $Y_L = Y_{obs} - 2s_Y$, $Y_U = Y_{obs} + 2s_Y$. If $Y_U > n - 1$, then let $Y_U = n - 1$.
3. Determine lower and upper confidence limits for r using the MME adjusted as described above: $r_L = \hat{r}(Y_L)$ and $r_U = \hat{r}(Y_U)$. Round r_L and r_U to the nearest integers. If $r_L < Y_{obs}$, then let $r_L = Y_{obs}$.

The second method of interval estimation for r is based on the following theorem (Bain and Engelhardt, 1992): if $G(Y|r)$ is a cumulative density function of Y and $G(Y|r)$ is an increasing function of r , then $100(1-\alpha)\%$ confidence limit for r is given by r_U , such that $G(Y|r_U) = \alpha$. In order to determine a specific value of r_U , an iterative procedure has to be run. Additionally, in each iteration $G(Y|r)$ has to be determined by simulations. Below we present an algorithm that applies the above theorem for estimation of $100(1-\alpha)\%$ confidence interval for r :

1. Let $r = \hat{r}$. Round r to the nearest integer.
2. Increase r by 1 and simulate the distribution of Y under current value of r , using the procedure of simulating Y described in section 5.

3. Calculate $P(Y \leq Y_{obs.})$ using the distribution of Y simulated in previous step.
4. If $P(Y \leq Y_{obs.}) \leq \alpha$ then assume that r_U is equal to the current value of r . Otherwise, go back to 2.

This procedure provides a one-sided upper tail confidence interval for r , while Y is a lower bound for r . A similar procedure might be used to determine two-sided confidence interval. However, according to the results of the preliminary evaluation, one-sided upper tail confidence intervals obtained with this method are usually shorter than two-sided confidence intervals. Therefore, in the case of the second method we recommend employing one-sided intervals.

We compared performance of the two methods described above using simulated data sets. In particular, we checked length and coverage of the intervals. Since the mean length of the confidence intervals is influenced by extremely long intervals occurring in rare cases when Y is close to n , we used median confidence interval as a criterion for comparison. In tables 2 and 3 we present the results. Table 2 refers to the case where the assumption of the exponential distribution of selection coefficients is met, while table 3 presents the results for the case where the data were simulated under the uniform distribution of selection coefficients. Notice that the median length of confidence intervals obtained with the first method is, in most cases, lower than the median length of the intervals resulting from the second method. However, confidence intervals estimated with the second method are conservative and their coverage is not lower than 95%, while the coverage of the intervals obtained using the first method is sometimes slightly lower than 95%.

According to the table 3 both methods are robust in the case when the assumption of the exponential distribution of the selection coefficients is violated. Median length of the interval is usually higher, but the coverage also increases which means that confidence intervals are more conservative.

In a typical case where $n = 50$ and $r = 50$, we recommend using the first method, which is simple and convenient, and the coverage of resulting confidence intervals is approximately equal to 95%. The second method requires running simulations in order to estimate a confidence interval, and the resulting confidence intervals are, in some cases, excessively conservative. This method is recommended for the cases when n is much greater than 50, i.e. in the range in which the MME with bias correction should not be used for point estimation because the bias correction term has only been validated for $n = 50$ and $r = 50$. In those instances, the adjusted ACE or the MLE can be used for point estimation. In fact, the algorithm for the MLE involves generating the distribution of Y given r for consecutive values of r ; therefore the same procedure can be used for both point estimation and interval estimation.

Finally, we applied the first method to the phage *ID11* data (Rokyta et al., 2005). According to the results, the lower confidence limit of the 95% confidence interval for the number of available one-step beneficial mutations equals 9, while the upper confidence limit equals 53.

9 Final conclusions

We have developed two estimators of the number of beneficial mutations that account for unequal sampling probabilities due to natural selection and unequal mutation rates. The first of these two estimators, the MME with bias correction, was designed to be applied in most typical cases, when the number of replicates of an experiment and the total number of available mutations do not exceed 50. If r is large ($r = 40$) it performs better than the other estimators. When r is small ($r = 10$) its performance is similar to the adjusted ACE, but is less biased than this estimator. The MME is convenient to use since a direct formula for the estimator is provided and there is no need to use computer simulations. Moreover, the MME is robust when the assumption of the exponential distribution of selection coefficients is violated, and it is not required that the distribution of the selection coefficients be determined before applying the MME. Furthermore, the performance of the MME is also robust to inaccuracies in the estimate of the parameter m .

Because the bias correction term of the MME was determined for the number of replicate experiments $n = 50$ and the available number of first-step mutations $r = 50$, this estimator is not recommended to be used far outside of this range. In rare cases when n is much greater than 50, we recommend the simulation-based ML estimator that we developed in section 5, or the adjusted ACE. While the adjusted ACE is more convenient to use since direct formula for the estimator is provided, the MLE is expected to be less biased than the adjusted ACE when the assumption of the exponential distribution of selection coefficients is met. However, the MLE is sensitive to departures from the assumed distribution of selection coefficients. Therefore, it is suggested that the analysis is preceded by inference about this distribution, and then the appropriate distribution is used in the estimation procedure. An available procedure for inferring the distribution of selection coefficients from observed data based on a likelihood ratio test (LRT) comes from Beisel et al. (2007).

We have also presented two methods of interval estimation for r . The first method is a procedure based on approximation of 2.5 and 97.5 percentiles of the distribution of $\hat{r}$, where $\hat{r}$ is a function of Y . This method performs better than the second presented method, which is an iterative procedure based on simulations of the distribution of Y . However, the first method involves using the MME formula with the bias correction, which has only been validated when $n = 50$ and $r = 50$. Therefore, if the number of replicates of an experiment is much greater than 50, we recommend the second method.

Acknowledgments

We thank Holly A. Wichman for helpful comments on this manuscript. This work was supported by a grant from the National Institutes of Health R01 GM076040-01. Computer simulations were carried out on the IBEST Bioinformatics Core supported by NIH/NCRR P20RR16448 and P20RR016454 grants.

Appendix

A Expected value of Y , given vector of selection coefficients

Result

$$E(Y|\mathbf{s}) = r - \sum_{j=1}^r \left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n$$

Proof—If $I_{j,n}$ is an indicator function, such that

$$I_{j,n} = \begin{cases} 1 & \text{if } j\text{-th mutant was observed at least once in } n \text{ replicates} \\ 0 & \text{if } j\text{-th mutant was observed in } n \text{ replicates} \end{cases}$$

then

$$Y = \sum_{j=1}^r I_{j,n}.$$

The probability of not sampling j -th mutant in n trials is equal to $(1 - P_{r+1,j})^n$, where $P_{r+1,j}$ is the probability of sampling j -th mutation in one experiment given by equation (1). Therefore

$$P(I_{j,n}=0) = (1 - P_{r+1,j})^n = \left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n$$

and

$$P(I_{j,n}=1) = 1 - P(I_{j,n}=0) = 1 - \left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n.$$

The above formula is equivalent to the expected value of $I_{j,n}$ given $\mathbf{s}$:

$$E(I_{j,n}|\mathbf{s}) = 1 - \left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n.$$

Based on that, we obtained the expected value of Y given $\mathbf{s}$

$$E(Y|\mathbf{s}) = r - \sum_{j=1}^r E(I_{j,n}|\mathbf{s}) = r - \sum_{j=1}^r \left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n.$$

B First moment of Y

Result

$$E(Y) = \frac{rn}{r+n-1}$$

Proof—In order to derive the expected value of Y , we used the approach proposed by Harris (1959). In the appendix A we derived $E(Y|\mathbf{s})$ to be

$$E(Y|\mathbf{s}) = r - \sum_{j=1}^r \left(1 - \frac{s_j}{\sum_{k=1}^r s_k} \right)^n.$$

Since selection coefficients are random variables, the expectation with respect to $\mathbf{s}$ has to be taken to obtain $E(Y)$

$$E(Y) = r - \sum_{j=1}^r E \left(1 - \frac{s_j}{\sum_{k=1}^r s_k} \right)^n.$$

To simplify the above formula we used the symmetry argument analogous to Joyce et al. (2008)

$$E(Y) = r - r E \left(1 - \frac{s_j}{\sum_{k=1}^r s_k} \right)^n. \quad (11)$$

In order to derive the expectation in the right side of the above equation, the first step was to determine the distribution of the term $s_j / \sum_{k=1}^r s_k$. This term can be written as $\frac{U}{U+V}$, where $U = s_j$ and $V = \sum_{k \neq j}^r s_k$. Since $s_1, s_2, \dots, s_r \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, then $U \sim \text{Gamma}(1, \lambda)$ and $V \sim \text{Gamma}(r-1, \lambda)$. Next, we applied the following theorem: $\frac{U}{U+V} \sim \text{Beta}(a, b)$ if $U \sim \text{Gamma}(a, \lambda)$ and $V \sim \text{Gamma}(b, \lambda)$ (Rizzo 2008, theorem 5 on page 58). Therefore,

$$\frac{s_j}{\sum_{k=1}^r s_k} \sim \text{Beta}(1, r-1)$$

and

$$1 - \frac{s_j}{\sum_{k=1}^r s_k} \sim \text{Beta}(r-1, 1).$$

Next, we derived the distribution of $(1 - P_{r+1, j})^n$ (see appendix C)

$$\left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n \sim \text{Beta}\left(\frac{r-1}{n}, 1\right).$$

For any beta distributed random variable B

$$E(B) = \frac{a}{a+b}$$

where a and b denote shape parameters of the beta distribution. Therefore,

$$E\left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n = \frac{r-1}{r+n-1}.$$

Then, by equation (11)

$$E(Y) = r - \frac{r(r-1)}{r+n-1} = \frac{rn}{r+n-1}.$$

C Distribution of $(1 - P_{r+1,j})^n$

Result

If $s_1, s_2, \dots, s_r \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, then

$$\left(1 - \frac{s_j}{\sum_{k=1}^r s_k}\right)^n \sim \text{Beta}\left(\frac{r-1}{n}, 1\right).$$

Proof—As it was shown in appendix B

$$1 - \frac{s_j}{\sum_{k=1}^r s_k} \sim \text{Beta}(r-1, 1).$$

Let $X = 1 - \frac{s_j}{\sum_{k=1}^r s_k}$ and $Z = X^n$. Using the CDF technique for transformation of random variables, we obtained the probability density function of Z :

$$F_Z(z) = P(X^n \leq z) = P\left(X \leq z^{\frac{1}{n}}\right) = F_X\left(z^{\frac{1}{n}}\right)$$

$$f_Z(z) = \frac{d}{dz} F_X\left(z^{\frac{1}{n}}\right) = f_X\left(z^{\frac{1}{n}}\right) \frac{1}{n} z^{\frac{1}{n}-1} = (r-1) z^{\frac{r-1}{n}-\frac{1}{n}} \frac{1}{n} z^{\frac{1}{n}-1} = \frac{r-1}{n} z^{\frac{r-1}{n}-1}$$

which means that $Z = (1 - P_{r+1,j})^n \sim \text{Beta}\left(\frac{r-1}{n}, 1\right)$.

D Second moment of Y

Result

If $s_1, s_2, \dots, s_r \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, then

$$E(Y^2) = \frac{rn}{r+n-1} \left(1 + \frac{(r-1)(n-1)}{r+n-2} \right).$$

Proof—According to the formula derived by Harris (1959)

$$E(Y^2 | \mathbf{s}) = \sum_{j=1}^r [1 - (1 - P_{r+1,j})^n] + \sum_{k \neq j} [1 - (1 - P_{r+1,j})^n - (1 - P_{r+1,i})^n + (1 - P_{r+1,j} - P_{r+1,i})^n].$$

As it was shown in appendices B and C, if $s_1, s_2, \dots, s_r \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, then

$$\frac{s_j}{\sum_{k=1}^r s_k} \sim \text{Beta}(1, r-1), \quad 1 - \frac{s_j}{\sum_{k=1}^r s_k} \sim \text{Beta}(r-1, 1),$$

$$\left(1 - \frac{s_j}{\sum_{k=1}^r s_k} \right)^n \sim \text{Beta}\left(\frac{r-1}{n}, 1\right).$$

Using the symmetry argument analogous to Joyce et al. (2008), we obtained the formula for expected value of Y^2 :

$$\begin{aligned} E(Y^2) &= rE[1 - (1 - P_{r+1,j})^n] + r(r-1)E[1 - (1 - P_{r+1,j})^n - (1 - P_{r+1,i})^n + \\ &\quad + (1 - P_{r+1,j} - P_{r+1,i})^n] = \\ &= \frac{rn}{r+n-1} + r(r-1) \left(1 - \frac{2(r-1)}{r+n-1} + E(1 - P_{r+1,j} - P_{r+1,i})^n \right). \end{aligned}$$

The n -th moment of $1 - P_{r+1,j} - P_{r+1,i}$ given by

$$(1 - P_{r+1,j} - P_{r+1,i})^n = \left(1 - \frac{s_j}{\sum_{k=1}^r s_k} - \frac{s_i}{\sum_{k=1}^r s_k} \right)^n = \left(1 - \frac{s_j + s_i}{\sum_{k=1}^r s_k} \right)^n.$$

had to be derived separately. Since $s_j + s_i \sim \text{Gamma}(2, \lambda)$, then

$$\frac{s_j + s_i}{\sum_{k=1}^r s_k} \sim \text{Beta}(2, r-2),$$

$$1 - \frac{s_j + s_i}{\sum_{k=1}^r s_k} \sim \text{Beta}(r - 2, 2).$$

Using the well known formula for the n -th moment of a beta distributed random variable, we

derived a formula for $E \left(1 - \frac{s_j + s_i}{\sum_{k=1}^r s_k} \right)^n$

$$E \left(1 - \frac{s_j + s_i}{\sum_{k=1}^r s_k} \right)^n = \frac{(r-2)(r-1) \dots (r+n-3)}{r(r+1) \dots (r+n-1)} = \frac{(r-1)(r-2)}{(r+n-1)(r+n-2)}.$$

Finally, we derived the second moment of Y

$$\begin{aligned} E(Y^2) &= \frac{rn}{r+n-1} + r(r-1) \left(1 - \frac{2(r-1)}{r+n-1} + \frac{(r-1)(r-2)}{(r+n-1)(r+n-2)} \right) \\ &= \frac{rn}{r+n-1} + r(r-1) \left(\frac{n(n-1)}{(r+n-1)(r+n-2)} \right) \\ &= \frac{rn}{r+n-1} \left(1 - \frac{(r-1)(n-1)}{r+n-2} \right). \end{aligned}$$

E Functional relationship between bias of the MME and r

To determine the bias correction term, a functional relationship between bias and r had to be identified for different values of m . Bias was assessed using simulations. Then, based on plotting bias versus r for different values of m , it was concluded that for $r < 20$, the relationship between bias and r can be described by a quadratic function; for $r \geq 20$ it can be described by a linear function (figure 5).

F Functional relationship between coefficients of the bias correction term for the MME and m

After concluding that bias of the MME may be approximated by a quadratic or linear function of r (appendix E), the next step was to find the coefficients of these functions for different values of m . The coefficients of quadratic and linear functions were plotted against m and the relationship between m and the coefficients was identified as a rational function $(\alpha_1, \beta_1, \beta_2)$ or a quadratic function (α_2, α_3) (figure 6).

References

- Bain, L.J., Engelhardt, M. Introduction to probability and mathematical statistics. Duxbury: 1992.
- Barrett RDH, MacLean RC, Bell G. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biology Letters*. 2006; 2:236–238. [PubMed: 17148371]
- 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]
- Bull JJ, Badgett MR, Wichman HA. Big-benefit mutations in a bacteriophage inhibited with heat. *Molecular Biology and Evolution*. 2000; 17:942–950. [PubMed: 10833201]

- Bunge J, Fitzpatrick M. Estimating the number of species: review. *Journal of the American Statistical Association*. 1993; 88:364–373.
- Chao A. Nonparametric estimation of the number of the classes in a population. *Scandinavian Journal of Statistics*. 1984; 11:265–270.
- Chao A. Estimating population size for capture-recapture data with unequal catchability. *Biometrics*. 1987; 43:783–791. [PubMed: 3427163]
- Chao A. Species estimation and applications. In: Kotz, S. Balakrishnan, N. Read, C., Vidakovic, B., editors. *Encyclopedia of Statistical Sciences*. Wiley; 2005. p. 7907–7916.
- Chao A, Lee S-M. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*. 1992; 87:210–217.
- Efron B, Thisted R. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*. 1976; 63:435–447.
- Ferris MT, Joyce P, Burch CL. High frequency of mutations that expand the host range of an RNA virus. *Genetics*. 2007; 176:1013–1022. [PubMed: 17409090]
- Gillespie JH. A simple stochastic gene substitution model. *Theoretical Population Biology*. 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.
- Good IJ. The population frequencies of species and the estimation of population parameters. *Biometrika*. 1953; 40:237–264.
- Harris B. Determining bounds on integrals with application to cataloging problems. *The Annals of Mathematical Statistics*. 1959; 30:521–548.
- Holst L. Some asymptotic results for incomplete multinomial or Poisson samples. *Scandinavian Journal of Statistics*. 1981; 8:243–246.
- Huillet T, Paroissin C. Sampling from Dirichlet partitions: estimating the number of species. *Environmetrics*. 2009; 20:853–876.
- 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]
- Kassen R, Bataillon T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics*. 2006; 38:484–488. [PubMed: 16550173]
- Keener R, Rothman E, Starr N. Distributions on partitions. *The Annals of Statistics*. 1987; 15:1466–1481.
- Kimura M. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the USA*. 1979; 76:3440–3444. [PubMed: 16592684]
- MacLean RC, Buckling A. The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. *PLoS Genetics*. 2009; 5:e1000406. [PubMed: 19266075]
- McDonald MJ, Cooper TF, Beaumont HJE, Rainey PB. The distribution of fitness effects of new beneficial mutations in *Pseudomonas fluorescens*. *Biology Letters*. 2011; 7:98–100. [PubMed: 20659918]
- Miller CR, Joyce P, Waits LP. A new method for estimating the size of small populations from genetic mark-recapture data. *Molecular Ecology*. 2005; 14:1991–2005. [PubMed: 15910321]
- Ohta, T. *Molecular Evolution and Polymorphism*. National Institute of Genetics; Mishima: 1977.
- Orr HA. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution*. 2002; 56:1317–1330. [PubMed: 12206234]
- Orr HA. The distribution of fitness effects among beneficial mutations. *Genetics*. 2003; 163:1519–1526. [PubMed: 12702694]
- Orr HA. The probability of parallel evolution. *Evolution*. 2005; 59:216–220. [PubMed: 15792240]
- Perfeito L, Fernandes L, Mota C, Gordo I. Adaptive mutations in bacteria: high rate and small effects. *Science*. 2007; 317:813–815. [PubMed: 17690297]
- Pielou, EC. *Mathematical Ecology*. John Wiley & Sons; 1977.

- Rizzo, ML. Statistical Computing with R. Chapman & Hall/CRC; Boca Raton, FL: 2008.
- Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA. Beneficial fitness effects are not exponential for two viruses. *Journal of Molecular Evolution*. 2008; 67:368–276. [PubMed: 18779988]
- Rokyta DR, Joyce P, Caudle SB, Wichman HA. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nature Genetics*. 2005; 37:441–444. [PubMed: 15778707]
- Rozen DE, de Visser JAGM, Gerrish PJ. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology*. 2002; 12:1040–1045. [PubMed: 12123580]

Figure 1.

Coefficient of variation ($\sqrt{MSE}/r$) of the estimators of number of available one-step beneficial mutations (r) in the case where beneficial mutations are exponentially distributed. Mean squared error (MSE) was determined using simulations. Left panel corresponds to a small number of available beneficial mutations ($r = 10$) and right panel to a large number of mutations ($r = 40$); n denotes the number of replicates of an experiment. The method of moments estimator (MME) outperforms the other estimators when r is large, but if r is small, then MSE of the adjusted abundance-based coverage estimator (ACE) is similar to MSE of the MME.

Figure 2.

Coefficient of variation ($\sqrt{MSE}/r$) of the estimators of number of available one-step beneficial mutations (r) in the case where beneficial mutations are uniformly distributed. Mean squared error (MSE) was determined using simulations. Left panel corresponds to a small number of available beneficial mutations ($r=10$) and right panel to a large number of mutations ($r=40$); n denotes the number of replicates of an experiment. The assumption of the exponential distribution of selection coefficients is violated and performance of the maximum likelihood estimator (MLE) worsens. By contrast, the method of moments estimator (MME) is a robust estimator and its performance is not affected. Also, performance of the adjusted abundance-based coverage estimator (ACE) improves.

Figure 3.

Scaled bias of the estimators of number of available one-step beneficial mutations (r) in the case where beneficial mutations are exponentially distributed. Bias was determined using simulations. Left panel corresponds to a small number of mutations ($r = 10$) and right panel to a large number of mutations ($r = 40$); n denotes the number of replicates of an experiment. If r is small, bias of all estimators is low. When r is large, the maximum likelihood estimator (MLE) is the least biased estimator, while the adjusted abundance-based coverage (ACE) estimator has the largest bias.

Figure 4.

Scaled bias of the estimators of number of available one-step beneficial mutations (r) in the case where beneficial mutations are uniformly distributed. Bias was determined using simulations. Left panel corresponds to a small number of mutations ($r = 10$) and right panel to a large number of mutations ($r = 40$); n denotes the number of replicates of an experiment. When r is large, violation of the assumption of the exponential distribution of selection coefficients leads to decrease in bias of the method of moments estimator (MME) and the adjusted abundance-based coverage estimator (ACE), and substantial increase in bias of the maximum likelihood estimator (MLE).

Figure 5. Bias of the MME as a function of r (for $m = 4, 8, 12$). Dots denote the level of bias determined by simulations while curves represent functions fitted to the data.

Figure 6. Coefficients of the bias correction term for the MME as functions of m . Dots denote values of the coefficients calculated on the basis of bias determined by simulations while curves represent fitted functions.

Table 1

Notation used in the text.

r	number of available one-step beneficial mutations
Y	number of observed one-step beneficial mutations
n	number of replicates of adaptation experiment
m	relative mutation rate of transitions to transversions
s_j	selection coefficient of j -th mutation
μ_j	mutation rate of j -th mutation
$\alpha_1, \alpha_2, \alpha_3$	coefficients of bias correction term when $r < 20$
β_1, β_2	coefficients of bias correction term when $r \geq 20$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Comparison of performance of two methods of interval estimation for r when the assumption of the exponential distribution of selection coefficients is met and $m = 8$. Confidence levels and median lengths of the confidence intervals were determined using 1 million simulations.

Table 2

		Analytical, approximate method					Iterative, simulation based method				
r		n=20	n=30	n=40	n=50		n=20	n=30	n=40	n=50	
median length of the CI	10	16	14	11	10		21	17	13	10	
	40	75	58	49	46		81	66	55	51	
actual conf. level (%)	10	94.7	92.1	96.2	98.0		99.3	98.4	99.3	98.0	
	40	96.5	91.8	95.7	96.7		97.4	97.8	97.1	98.1	

Comparison of performance of two methods of interval estimation for r when the selection coefficients are uniformly distributed and $m = 8$. Confidence levels and median lengths of the confidence intervals were determined using 1 million simulations.

Table 3

		Analytical, approximate method					Iterative, simulation based method				
r		n=20	n=30	n=40	n=50		n=20	n=30	n=40	n=50	
median length of the CI	10	16	14	14	12		21	17	15	12	
	40	102	69	55	50		107	79	63	58	
actual conf. level (%)	10	98.0	97.4	99.2	99.7		99.9	99.7	99.9	99.7	
	40	97.2	93.8	95.9	95.7		99.7	99.8	99.8	99.9	