PLOS | ONE

Unbiased Estimation of Mutation Rates under Fluctuating Final Counts

Bernard Ycart[1,2]*, **Nicolas Veziris**[3,4,5,6]

1 Laboratoire Jean Kuntzmann, Univ. Grenoble Alpes, Grenoble, France, **2** Laboratoire d'Excellence "TOUCAN" (Toulouse Cancer), Toulouse, France, **3** Sorbonne Universités, UPMC Univ. Paris 06, CR7, Centre d'Immunologie et des Maladies Infectieuses, CIMI, Team E13 (Bacteriology), Paris, France, **4** INSERM, U1135, Centre d'Immunologie et des Maladies Infectieuses, CIMI, Team E13 (Bacteriology), Paris, France, **5** AP-HP, Hôpital Pitié-Salpêtrière, Centre National de Référence des Mycobactéries et de la Résistance des Mycobactéries aux Antituberculeux, Laboratoire de Bactériologie-Hygiène, Paris, France, **6** Mycobacteria Research Laboratories, Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado, United States of America

Abstract

Estimation methods for mutation rates (or probabilities) in Luria-Delbrück fluctuation analysis usually assume that the final number of cells remains constant from one culture to another. We show that this leads to systematically underestimate the mutation rate. Two levels of information on final numbers are considered: either the coefficient of variation has been independently estimated, or the final number of cells in each culture is known. In both cases, unbiased estimation methods are proposed. Their statistical properties are assessed both theoretically and through Monte-Carlo simulation. As an application, the data from two well known fluctuation analysis studies on *Mycobacterium tuberculosis* are reexamined.

Citation: Ycart B, Veziris N (2014) Unbiased Estimation of Mutation Rates under Fluctuating Final Counts. PLoS ONE 9(7): e101434. doi:10.1371/journal.pone.0101434

Editor: Robert Belshaw, Plymouth University, United Kingdom

Received February 4, 2014; **Accepted** June 6, 2014; **Published** July 2, 2014

Funding: Research supported by Laboratoire d'Excellence TOUCAN (Toulouse Cancer), by Fondation du Souffle and Fonds de Dotation Recherche en Santé Respiratoire. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: Bernard.Ycart@imag.fr

Introduction

Since the pioneering work of Luria and Delbrück [1], fluctuation analysis has been the object of many studies: see [2–7] for reviews. In the past twenty years, the stress has been put on the estimation of the expected number of mutations, for which reliable methods are now available [8–15]. However, as Stewart puts it (p. 1140 of [4]):

> The parameter Λ [expected number of mutations] is not, in itself, of biological interest because the experimenter can vary it at will simply by changing the size of the culture vessel or the richness of the medium. What he really wants to know is not Λ, but the mutation rate.

Deriving a mutation rate (i.e. the probability for a mutation to occur upon any given cell division) from an expected number of mutations seems easy: the former is the quotient of the latter by the final number of cells at the end of the experiment. The problem is the definition given to "final number of cells". The simplest view is expressed by Kendal and Frost (p. 1062 of [2]).

> N is obtained by averaging the final number of cells from each parallel culture.

Other authors have developed a more cautious approach, like Foster (p. 198 of [5]).

> The validity of the mutation rate calculation requires that N_t be the same in each culture. Usually, but not always, this can be accomplished by growing cells to saturation. If achieving an uniform N_t is a problem, the cell number in each culture can be monitored before mutant selection by measuring the optical density or by counting cells microscopically (e.g. using a Petroff-Hausser chamber). Because there is currently no valid method to correct for different N_t's, deviant cultures must be eliminated from the analysis.

Even under the most careful monitoring, final numbers of cells vary [16]. Yet, final number data are rarely reported in fluctuation analysis experiments, although exceptions exist such as [17,18]. Theoretical models considering variations in the population size have previously been proposed by Angerer [19] and Komarova *et al.* [20]. Yet, to the best of our knowledge, Foster's assertion that "there is currently no valid method to correct for different N_t's" remains true to this date. This paper proposes several such methods.

As we shall see, dividing an estimated expected number of mutations by a mean final number of cells, induces a negative bias on mutation rates. Not only the mutation rate, but also the variance of the estimator are underestimated, thus potentially inducing wrong conclusions in statistical testing. Two levels of knowledge on the fluctuations of final numbers are considered. Either the mean and variance of final numbers have been estimated separately, or the final number is known for each culture. In the first case, if $\hat{\pi}$ denotes the estimate of the mutation rate assuming constant final numbers, the unbiased estimate $\hat{\pi}_{\mathrm{ub}}$ is obtained by:

$$\hat{\pi}_{\text{ub}} = \hat{\pi}\left(1 + \frac{\hat{\pi}\mu C^2}{2}\right), \qquad (1)$$

where μ and C denote the mean and coefficient of variation of the final number of cells. When final numbers are known for all cultures, better results are obtained by the Maximum Likelihood method. The qualities of the proposed estimators have been assessed on a simulation study. The impact on real experiments is discussed, using *Mycobacterium tuberculosis* data published by David [17], and Werngren & Hoffner [21]. Our R [22] implementation of the simulation function and the different estimators is provided in File S1.

Results

Simulation experiments

Six different estimates of π were computed on 1000 simulated samples of 50 couples mutant counts – final numbers. Our choice for the sample size was motivated by two opposite reasons. On the one hand, sample sizes in practice rarely exceed a few tens. On the other hand, confidence interval calculations are all based on asymptotic normality, which requires the sample size to be large enough. A sample size of 50 seemed a reasonable compromise. Boxplots for the estimates are represented on Figure 1. The first boxplot corresponds to the 1000 estimates by the p_0-method, assuming the mean final number is known; it is negatively biased as predicted by the theory. The next boxplot represents estimates from the Maximum Likelihood method with known mean final number; it is coherent with the previous one, and similarly biased as expected. On the next two boxplots, the estimates have been multiplied by the unbiasing factor (1). The unbiasing is correct for both methods. For the last two boxplots, each estimate has been computed using the 50 couples with no prior knowledge on the mean and coefficient of variation of final numbers. The best results are obtained by the maximum likelihood method (last boxplot). The p_0-method (label MLP0) performs nearly as well. Since the last two boxplots do not use any prior information, one could have expected their dispersions to be higher than those of the first four. This was not the case, which proves that prior knowledge on the distribution of N is not a real improvement over measuring final numbers for each culture.

Each estimation method returns a (theoretical) standard deviation, from which confidence intervals can be computed. It is is based on a large sample approximation. The sample size in current fluctuation analysis experiments usually ranges from 20 to 50. Since the estimated standard deviation is of high importance for statistical decision, it was necessary to check whether theoretical standard deviations matched observations. On the same samples, the empirical standard deviation of the 1000 estimates was computed, and compared to the mean value of theoretical standard deviations. For each of the estimators, the theoretical standard deviation was smaller than the observed one; yet, the relative error was smaller than 5%, which validates the theoretical value. For instance, the empirical standard deviation for the maximum likelihood estimate (rightmost boxplot of Figure 1) was 1.85×10^{-10}, whereas the theoretical value was 1.80×10^{-10}.

Published data sets

In the two references studied here [17,21], the authors used Luria & Delbrück's method of the mean. Luria & Delbrück [1] themselves had remarked that the method is very sensitive to the size of jackpots and induces important biases; see also Lea & Coulson [23], and Pope *et al.* [6] for a more recent reference.

Table 1 reports mutation rate estimates for the data in Table 1 of David [17]. Since detailed data were not avaible, only the p_0-method could be used. The second column contains the author's estimates. The next two columns contain the unbiased p_0-estimate and its 95% confidence interval. Observe that, even though confidence intervals are large due to the small sample sizes, the author's estimates are outside the confidence interval in 5 cases out

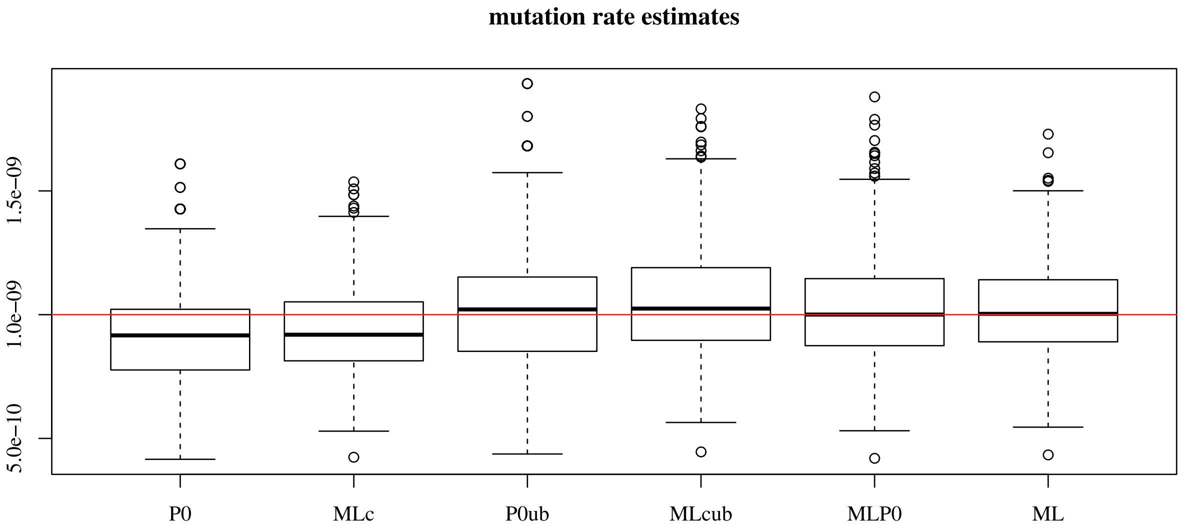

Figure 1. Estimates of a mutation rate on 1000 samples of size 50 of pairs mutant counts – final counts. The horizontal line marks the true value. The first two boxplots correspond the traditional p_0- and ML methods, which estimate the expected number of mutations from the sample of mutant counts, then divide by the final number of cells, supposed as known. On the next two boxplots, the estimates have been multiplied by the unbiasing factor (1). The last two boxplots use the full samples of pairs but no prior knowledge on final numbers. The best results are obtained by the maximum likelihood method (last boxplot). The p_0-method (label MLP0) performs nearly as well.
doi:10.1371/journal.pone.0101434.g001

Table 1. Mutation rate estimates from Table 1 of [17].

Determination	Author	p_0-method	Confidence interval
Isoniazid 1	1.84×10^{-8}	2.2×10^{-8}	$[5.8\times10^{-9};\ 3.8\times10^{-8}]$
Isoniazid 2	3.5×10^{-8}	1.1×10^{-8}	$[5.3\times10^{-9};\ 1.7\times10^{-8}]$
Isoniazid 3	1.7×10^{-8}	1.3×10^{-8}	$[5.1\times10^{-9};\ 2.1\times10^{-8}]$
Isoniazid 4	3.2×10^{-8}	8.6×10^{-9}	$[4.7\times10^{-9};\ 1.2\times10^{-8}]$
Streptomycin 1	0.9×10^{-8}	5.2×10^{-9}	$[1.9\times10^{-9};\ 8.5\times10^{-9}]$
Streptomycin 2	5.0×10^{-8}	6.6×10^{-9}	$[3.9\times10^{-9};\ 9.2\times10^{-8}]$
Rifampin 1	1.8×10^{-10}	3.2×10^{-10}	$[0.0\times10^{-10};\ 9.4\times10^{-10}]$
Rifampin 2	2.7×10^{-10}	2.9×10^{-10}	$[0.0\times10^{-10};\ 5.8\times10^{-10}]$
Ethambutol 1	0.7×10^{-7}	3.3×10^{-9}	$[9.1\times10^{-10};\ 5.6\times10^{-9}]$
Ethambutol 2	1.3×10^{-7}	3.8×10^{-9}	$[2.3\times10^{-9};\ 5.3\times10^{-9}]$

The author's estimates were calculated by Luria and Delbrück's method of the mean. Our estimates were calculated by the p_0-method. The bias correction (1) was applied, with a coefficient of variation $C=0.35$ on final numbers. The 95% confidence interval is given in the last column.

doi:10.1371/journal.pone.0101434.t001

of 10. The most important discrepancies are due the author's use of a strongly biased estimation method: when large jackpots appear in the mutant counts, as in the Ethambutol cases (last two lines of Table 1), the method of the mean may overestimate π by several orders of magnitude. The main conclusion of [17] was a significant difference in mutation rates, depending on the drug (Isoniazid, Streptomycin, Rifampin, or Ethambutol). Indeed that difference is confirmed by an ANOVA of the estimated mutation rates ($P=0.012$).

Table 2 of David [17] contains two paired samples of mutant counts and final numbers. All possible estimates were computed. Values ranged between 1.81×10^{-10} and 2.14×10^{-10}. The two values that we consider most reliable, obtained by the maximum likelihood method, were very similar: 1.98×10^{-10} and 1.97×10^{-10}. The estimate reported by the author is 7.53×10^{-10}. Again, the difference is due to the bias induced by the author's estimation method.

Table 2 reports mutation rate estimates by the ML method, from data in Table 1 of Werngren & Hoffner [21]. The second column contains the authors' estimates, calculated by Luria & Delbrück method of the mean. The next two columns contain the unbiased ML estimate and its 95% confidence interval. Except for two strains, the authors' estimate is outside the confidence interval. Here, the method of the mean used by the authors has underestimated the mutation rate, because of the very small number of jackpots in the data. The main conclusion of [21] was that no significant difference had been observed between non-Beijing strains (first seven lines) and Beijing strains (last six lines). Actually, the average mutation rate over the first seven lines is 4.37×10^{-8}, over the last six lines it is 2.69×10^{-8}. The difference is significant at threshold 5% (Welsh Two Sample t-test, $P=0.047$).

Discussion

In any estimation problem, three levels must be distinguished: the reality which is and will remain unknown, the mathematical model which involves more or less realistic hypotheses, and the estimation method. Minimal requirements for an estimator are consistence (outputs should be close to the unknown value of the parameter), and a computable asymptotic variance (to allow statistical inference). Since there is no way to validate all mathematical hypotheses that define the model, another quality is desirable: robustness. Indeed, designing an estimator for a given model and applying it to a different one usually induces a bias: the smaller the bias, the more robust the estimator. For mutation rate estimates, several sources of bias have been identified, such as cell deaths [19,24–26], unknown division time distribution [15], etc. Since there is no way to double check estimates on real data, the usual approach for evaluating an estimation method consists in repeating in silico experiments, i.e. simulate mathematical models for a given value of the parameter, estimate that value repeatedly, and study the distribution of the obtained estimates. A general simulation algorithm described in [15] permits extensive Monte-Carlo experiments.

Usually, only the expected number of mutations is considered as the parameter of interest. Among the many estimation procedures that have been proposed, we have focused on the p_0-method and the maximum likelihood (ML); they satisfy the basic requirements of statistical inference. As for most other parametric estimation problems, the ML method is the most precise. Provided cell deaths are neglected, the p_0-method stands out as the most robust.

All estimation methods are valid only if all observed mutant counts come from the same Luria-Delbrück distribution, i.e. if they have been obtained under a fixed expected number of mutations. However, the parameter of *real* interest which must be considered as fixed, is the mutation rate. For each culture the expected number of mutations is the product of the mutation rate by the final number of cells. Since final numbers vary from one culture to another, so do expected numbers of mutations. As shown here, applying the p_0- and ML procedures to the fluctuating final number case as if final numbers were constant, induces a bias. Two solutions have been proposed. In the case where the final numbers of each culture are unknown, but a coefficient of variation is available, an unbiasing factor has been defined, and validated on simulation experiments. The unbiasing factor (1) measures the error induced by neglecting final number fluctuations: the relative error is of order $\alpha C^2/2$ where $\alpha=\pi\mu$ is the expected number of mutations and C the coefficient of variation of final numbers.

The more favorable case is when final numbers are available. Of course measuring the final number of cells for each culture leads to reducing the volume of the culture in which the mutants are counted, and therefore underestimating mutations. This

Table 2. Mutation rate estimates from Table 1 of [21].

Strain	Authors	ML method	Confidence interval
H37Rv	8.6×10^{-9}	4.8×10^{-8}	$[3.0\times10^{-8};\ 6.6\times10^{-8}]$
E 865/94	2.4×10^{-8}	7.6×10^{-8}	$[4.3\times10^{-8};\ 1.1\times10^{-7}]$
E 729/94	9.6×10^{-9}	2.3×10^{-8}	$[1.3\times10^{-8};\ 3.3\times10^{-8}]$
E 740/94	1.1×10^{-8}	3.6×10^{-8}	$[2.2\times10^{-8};\ 5.0\times10^{-8}]$
E 1221/94	6.5×10^{-9}	1.3×10^{-8}	$[7.3\times10^{-9};\ 1.9\times10^{-8}]$
E 1449/94	1.5×10^{-8}	4.8×10^{-8}	$[2.9\times10^{-8};\ 6.8\times10^{-8}]$
Harlingen	1.4×10^{-8}	6.2×10^{-8}	$[3.8\times10^{-8};\ 8.6\times10^{-8}]$
E 26/95	1.3×10^{-8}	2.3×10^{-8}	$[1.3\times10^{-8};\ 3.4\times10^{-8}]$
E 80/95	7.9×10^{-9}	2.8×10^{-8}	$[1.6\times10^{-8};\ 4.0\times10^{-8}]$
E 55 94	1.0×10^{-8}	2.0×10^{-8}	$[9.7\times10^{-9};\ 3.1\times10^{-8}]$
E 26/94	9.4×10^{-9}	3.2×10^{-8}	$[2.2\times10^{-8};\ 4.3\times10^{-8}]$
E 3942/94	1.5×10^{-8}	3.9×10^{-8}	$[2.2\times10^{-8};\ 5.6\times10^{-8}]$
E 47/94	1.2×10^{-8}	1.8×10^{-8}	$[9.2\times10^{-9};\ 2.8\times10^{-8}]$

The authors' estimates were calculated by Luria and Delbrück's method of the mean. Our estimates were calculated by the maximum likelihood method under exponential division times. The bias correction (1) was applied, using a coefficient of variation $C=0.44$ on final numbers. The 95% confidence interval is given in the last column.

doi:10.1371/journal.pone.0101434.t002

should be accounted for, by proportionally adjusting the estimates of final numbers. When coupled mutant counts – final numbers have been collected, variants of the p_0- and ML methods are available. Both yield quite precise estimates. As in the constant final number case, the p_0-method is more robust, and almost as precise as the ML method. Only the ML method can output relative fitness estimates.

Does the correction for fluctuating final numbers have an impact on the interpretation of the data? We have reexamined the data in two examples chosen from the literature. In both cases, important discrepancies were oberved, that do not only come from neglecting final numbers: they are essentially due to the author's use of Luria-Delbrück's method of the mean, which is very sensitive to jackpots, and can bias the mutation rate estimate by several orders of magnitude. In David's paper, the ethambutol mutation rate had been estimated around 10^{-7} whereas our estimation is of order 10^{-9}. The demonstration is even more striking in Werngren and Hoffner's paper. They compared mutation rate between Beijing and non Beijing *M. tuberculosis* strains and concluded that it was not different and thus could not explain the strong association between Beijing strains and multidrug resistance phenotype. However we re-calcutated the mutation rate and showed that it was significantly higher for Beijing vs. non-Beijing strains. This result is consistent with a recent paper [27] showing that lineage 2 (Beijing) *M. tuberculosis* strains have a higher mutation rate than lineage 4 (non-Beijing) strains. Given the importance of mutation rates on the risk of selection of drug resistant mutants, an accurate evaluation is very important. We hope that our results will help improving precision in the evaluation of mutation rates.

Conclusion

Dealing with classical estimation methods, Foster [5] was right in recommending that cultures with deviant final numbers be eliminated from fluctuation analysis. Indeed, under varying final numbers those methods underestimate mutation rates, and the relative bias is proportional to the squared coefficient of variation of final numbers. Yet, instead of being discarded as a nuisance, variations in final numbers should be added to the available information to improve estimation: the best mutation rates estimates are obtained when couples mutation count – final number are used.

Two possibilities exist. If mutant counts contain enough zeros (say 10% or more), the p_0-method gives reliable results in virtually null computer time, and is robust both to relative fitness and division time distribution changes. If mutant counts do not contain enough zeros, or if an estimate of relative fitness is sought for, then the joint estimation of the mutation rate and relative fitness should be carried through by the maximum likelihood method.

We are currently working on an optimized implementation of these methods into a forthcoming R [22] package that will be made freely available.

Methods

Here, N denotes the final number of cells in a Luria-Delbrück fluctuation analysis experiment. Contrarily to the traditional point of view [5], fluctuations on N are considered, i.e. N is viewed as a random variable. In the following subsections, different levels of information are assumed on the distribution of N: either its Laplace transform is known, or only its expectation and variance are known, or nothing is known, but the final numbers of cells have been measured together with mutant counts for each experiment. Notations for the different parameters are summarized in Table 3.

As usual, adding a 'hat' to the notation of a parameter denotes an estimator of that parameter. We shall consider only strongly consistent, asymptotically Gaussian estimators. If θ is any parameter, and s denotes the sample size, then $\sqrt{s}(\hat{\theta}-\theta)$ converges to a centered Gaussian distribution as s tends to infinity. The variance of that distribution, called asymptotic variance of $\hat{\theta}$, will be denoted by $v_{\hat{\theta}}$.

In the next four subsections, the focus is on the so-called p_0-method, introduced by Luria and Delbrück [1] (see also [5,28]). The problem of jointly estimating the mutation rate π and the relative fitness ρ by he maximum likelihood method will be treated after.

Unbiasing p_0-estimates

The final number of cells N is viewed as a random variable with probability distribution function G on $[0,+\infty)$. The distribution of N is supposed to be known and its Laplace transform is denoted by $\mathcal{L}$.

$$\mathcal{L}(\pi) = \mathbb{E}\left[\mathrm{e}^{-\pi N}\right] = \int_0^{+\infty} \mathrm{e}^{-\pi t}\,\mathrm{d}G(t)\ .$$

The expectation and variance of N are denoted by μ and σ^2 respectively. Let U be a random variable, with uniform distribution on [0,1], independent from N. The indicator X for the mutant count being null is defined as:

$$X = \mathbb{I}_{U < \mathrm{e}^{-\pi N}}\ ,$$

where $\mathbb{I}_A$ denotes the indicator of event A (1 if A is true, 0 else). Therefore:

$$\mathbb{P}[X=1 \mid N=t] = \mathrm{e}^{-\pi t}\ ,$$

and

$$p_0 = \mathbb{P}[X=1] = \mathcal{L}(\pi).$$

Consider a sample of size s, i.e. s independent copies of X: $(X_1, \ldots, X_s)$. Denote by $\hat{p}_0$ the empirical mean of the X_i's, i.e. the relative frequency of zeros among mutant counts.

$$\hat{p}_0 = \frac{1}{s}\sum_{i=1}^{s} X_i\ .$$

By the central limit theorem, $\sqrt{s}(\hat{p}_0 - p_0)$ converges in distribution to the centered Gaussian distribution with variance $p_0(1-p_0)$, i.e. $\hat{p}_0$ has asymptotic variance $v_{\hat{p}_0} = p_0(1-p_0)$.

The p_0-method consists of estimating the mean number of mutations α by the negative logarithm of $\hat{p}_0$, then divide by μ to obtain an estimate of π.

$$\hat{\alpha}_0 = -\log(\hat{p}_0) \quad \text{and} \quad \hat{\pi}_0 = \frac{\hat{\alpha}_0}{\mu}\ .$$

Actually, $\hat{\alpha}_0$ is a consistent estimator of:

$$-\log(p_0) = -\log(\mathcal{L}(\pi))\ .$$

If N is constant, then $\mathcal{L}(\pi) = \mathrm{e}^{-\pi\mu} = \mathrm{e}^{-\alpha}$, and $-\log(p_0) = \alpha$: in that case $\hat{\alpha}_0$ is asymptotically unbiased. If N is not constant, because of the convexity of the exponential, and by Jensen's inequality, $-\log(\mathcal{L}(\pi))$ is smaller than α, i.e. $\hat{\alpha}_0$ underestimates α, and therefore $\hat{\pi}_0$ underestimates π.

Denote by $\mathcal{L}^{-1}$ the inverse of $\mathcal{L}$ (assumed to be injective). Define a new estimator of π by:

$$\hat{\pi}_{\mathrm{ub}} = \mathcal{L}^{-1}\left(\mathrm{e}^{-\mu\hat{\pi}_0}\right) = \mathcal{L}^{-1}(\hat{p}_0)\ . \qquad (2)$$

By construction, $\hat{\pi}_{\mathrm{ub}}$ is a strongly consistent estimator of π, and therefore it is asymptotically unbiased. Its asymptotic variance is obtained by the traditional delta-method (see e.g. [29]): $\sqrt{s}(\hat{\pi}_{\mathrm{ub}} - \pi)$ converges in distribution to the univariate centered Gaussian distribution with variance:

$$v_{\hat{\pi}_{\mathrm{ub}}} = (\mathcal{L}'(\pi))^{-2} p_0(1-p_0)\ .$$

As expected, if N is constant at μ, then $p_0 = \mathcal{L}(\pi) = \mathrm{e}^{-\pi\mu}$, $\hat{\pi}_{\mathrm{ub}} = \pi_0$, and

$$v_{\hat{\pi}_{\mathrm{ub}}} = v_{\hat{\pi}_0} = \frac{1-p_0}{\mu^2 p_0}$$

This formula is not new: the asymptotic variance of $\hat{\pi}_0$ appeared as formula 35, p. 276 of Lea & Coulson [23]; see also [5,28].

Table 3. Parameters and notations for the mathematical model.

known parameters	
N	random final number of cells
$\mathcal{L}(\pi) = \mathbb{E}[\mathrm{e}^{-\pi N}]$	Laplace transform of N
$\mu = \mathbb{E}[N]$	expectation of N
$\sigma = \sqrt{\mathrm{Var}[N]}$	standard-deviation of N
$C = \sigma/\mu$	coefficient of variation of N
unknown parameters	
π	mutation rate
$\alpha = \pi\mu$	expected number of mutations
$p_0 = \mathrm{e}^{-\alpha}$	probability of zero mutant
ρ	relative fitness of normal cells compared to mutants

Notations for known and unknown parameters: N denotes a generic random final number of cells.

doi:10.1371/journal.pone.0101434.t003

Families of distributions for which explicit expressions of $\hat{\pi}_{\text{ub}}$ and $v_{\hat{\pi}_{\text{ub}}}$ can be obtained are scarce. Two examples are given below.

Gamma distributions. They depend on two parameters, usually denoted by a and λ. The expectation and variance are:

$$\mu = \frac{a}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{a}{\lambda^2} \ .$$

The squared coefficient of variation is the inverse of the shape parameter: $C^2 = 1/a$. The Laplace transform at π is:

$$\mathcal{L}(\pi) = \left(\frac{\lambda}{\lambda+\pi}\right)^a \ .$$

One gets:

$$\hat{\pi}_{\text{ub}} = \lambda\left(\hat{p}_0^{-\frac{1}{a}} - 1\right) \quad \text{and} \quad v_{\hat{\pi}_{\text{ub}}} = p_0^{-\frac{2}{a}} v_{\hat{\pi}_0} \ .$$

Expressed in terms of $\hat{\alpha}_0$, μ and C:

$$\hat{\pi}_{\text{ub}} = \frac{1}{\mu C^2}\left(\exp(\hat{\alpha}_0 C^2) - 1\right) \quad \text{and} \quad v_{\hat{\pi}_{\text{ub}}} = \exp(\alpha C^2)\, v_{\hat{\pi}_0} \ .$$

Inverse Gaussian distributions. They depend on two parameters, λ and μ. The parameter μ is the expectation, and the variance is $\sigma^2 = \mu^3/\lambda$. The squared coefficient of variation is $C^2 = \mu/\lambda$. The Laplace transform at π is:

$$\mathcal{L}(\pi) = \exp\left(\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2\pi}{\lambda}}\right)\right) \ .$$

One gets:

$$\hat{\pi}_{\text{ub}} = -\frac{\log(\hat{p}_0)}{\mu} + \frac{\log^2(\hat{p}_0)}{2\lambda} \ ,$$

and

$$v_{\hat{\pi}_{\text{ub}}} = \left(1 - \frac{\mu}{\lambda}\log(p_0)\right)^2 \frac{1-p}{\mu^2 p_0} = \left(1 - \frac{\mu}{\lambda}\log(p_0)\right)^2 v_{\hat{\pi}_0} \ .$$

Expressed in terms of $\hat{\alpha}_0$ and C^2, these expressions become:

$$\hat{\pi}_{\text{ub}} = \hat{\pi}_0\left(1 + \frac{\hat{\alpha}_0 C^2}{2}\right) \ ,$$

and

$$v_{\hat{\pi}_{\text{ub}}} = \left(1 + \alpha C^2\right)^2 v_{\hat{\pi}_0} \ .$$

As we shall see in the next subsection, the last two expressions, which are exact for inverse Gaussian distributions, hold as a first order approximation for any distribution.

First order approximation

If the probability distribution of N is known, the bias can be exactly corrected by inverting the Laplace transform of N. However, this is only a theoretical viewpoint. The best that can be hoped for in practice is an estimate of the expectation of N together with its variance. It turns out that whatever the distribution of N, and provided the product of the coefficient of variation by the expected number of mutations remains relatively small, the bias can be corrected. Here, we only assume that the first two moments of N, μ and σ^2 are known, but the full distribution of N, and in particular its Laplace transform, remains unknown. As we have seen, the expectation of $\hat{\alpha}_0$ is $-\log(\mathcal{L}(\pi))$. Consider the terms of the series expansion of $\mathcal{L}(\pi)$ in π up to order 2 (see e.g. [30]):

$$\mathcal{L}(\pi) = 1 - \mu\pi + \frac{\mathbb{E}[N^2]}{2}\pi^2 + \cdots$$

Taking negative logarithm,

$$-\frac{\log(\mathcal{L}(\pi))}{\mu} = \pi - \frac{\sigma^2}{2\mu}\pi^2 + \cdots$$

Expressed in terms of α and C^2, the relative bias is:

$$1 - \frac{\sigma^2}{2\mu}\pi = 1 - \frac{\alpha C^2}{2} \ .$$

To unbias $\hat{\pi}_0$, one must divide by the relative bias or (as a first order approximation), multiply by $1 + \frac{\hat{\alpha}_0 C^2}{2}$. Hence (1):

$$\hat{\pi}_{\text{ub}} = \hat{\pi}_0\left(1 + \frac{\hat{\alpha}_0 C^2}{2}\right) \ .$$

The asymptotic variance, obtained through the delta-method is:

$$v_{\hat{\pi}_{\text{ub}}} = \left(1 + \alpha C^2\right)^2 v_{\hat{\pi}_0} \ . \tag{3}$$

These expressions are exact for inverse Gaussian distributions, only approximations for any other distribution.

To assess the validity range of the unbiasing factor, a simulation experiment was conducted. For the same value of $\pi = 10^{-9}$, samples of final numbers were simulated with a log-normal distribution with mean $\mu = \alpha/\pi$ and coefficient of variation C. The values of α ranged from 0.1 to 2, those of C from 0 to 1. The results are shown on Figure 2. Red curves show the actual relative bias of the p_0-method; for blue curves, the bias has been corrected by the unbiasing factor (1). The correction maintains the bias under acceptable values even for relatively large α and C.

The p_0-method by maximum likelihood

In this section, nothing is assumed about the distribution of N. A couple (X,N) of random variables is considered, where X represents the indicator of a null mutant count, and N the total number of cells at the end of the experiment. The conditional distribution of X knowing $N=n$, is defined as before:

$$\mathbb{P}[X=1\,|\,N=n]=\mathrm{e}^{-\pi n}\;.$$

Assume that s experiments have been repeated independently, yielding s couples (x_i,n_i), where x_i is 1 or 0 according to whether zero or a positive number of mutants have been counted, and n_i is the final number of cells. The likelihood is the probability of the observation:

$$L(\pi)=\prod_{i=1}^{s}(\mathrm{e}^{-\pi n_i})^{x_i}\,(1-\mathrm{e}^{-\pi n_i})^{1-x_i}\;.$$

The likelihood depends only on the products πn_i. If all $n_i^{'}s$ are divided by a given constant, then the maximum likelihood estimator will be multiplied by the same constant. Since the n_i's are very large and π very small, rescaling both can make the calculation numerically more stable.

The log-likelihood and its derivatives are:

$$\begin{aligned}
\ell(\pi) &= \sum_{i=1}^{s}-\pi n_i x_i+(1-x_i)\log(1-\mathrm{e}^{-\pi n_i})\;,\\
\frac{\mathrm{d}\ell}{\mathrm{d}\pi}(\pi) &= \sum_{i=1}^{s}-n_i x_i+\frac{(1-x_i)n_i}{\mathrm{e}^{\pi n_i}-1}\;,\\
\frac{\mathrm{d}^2\ell}{\mathrm{d}\pi^2}(\pi) &= -\sum_{i=1}^{s}\frac{(1-x_i)n_i^2\mathrm{e}^{\pi n_i}}{(\mathrm{e}^{\pi n_i}-1)^2}\;.
\end{aligned}$$

The maximum likelihood estimator $\hat{\pi}_{\mathrm{ml}}$ is the solution of $\frac{\mathrm{d}\ell}{\mathrm{d}\pi}(\hat{\pi}_{\mathrm{ml}})=0$, and its asymptotic variance is computed from $\left(-\frac{\mathrm{d}^2\ell}{\mathrm{d}\pi^2}(\pi)\right)^{-1}$ (see [29]). This is essentially the method used by de la Iglesia et al. [18] in a similar case.

Bivariate maximum likelihood estimation

In cases where no null mutant counts have been observed, or if an estimate of the relative fitness is desired together with the mutation rate, another procedure must be used. Estimating the two parameters of a classical Luria-Delbrück distribution by the method of maximum likelihood was proposed long ago [8,12,31,32]. Using well known explicit formulas, the method has been implemented [11,14,33]. In [15] it was shown that similar algorithms apply not only to the classical Luria-Delbrück distribution (in which division times are exponentially distributed), but also to the so-called Haldane model in which distribution times are supposed constant [34,35]. The situation here is only slightly different. Instead of being considered as a sample of a fixed Luria-Delbrück distribution, mutant counts can be viewed as independent realizations of different distributions. Denote by $LD(\alpha,\rho)$ a Luria-Delbrück distribution with expected number of mutations α and relative fitness ρ. If a pair mutant count – final number (m,n) has been observed, m is viewed as a realization of the $LD(\pi n,\rho)$, and the likelihood is computed accordingly. Thus the pair (π,ρ) is jointly estimated, as the pair (α,ρ) in the constant final number case.

Here is the mathematical model: for each experiment a pair of numbers giving the number of mutants and the final number of cells is obtained. An experiment is modelled by a couple (M,N) of random variables, where M represents the number of mutants and N the total number of cells at the end of the experiment. The conditional distribution of M knowing $N=n$ is assumed to be the generalized Luria-Delbrück distribution $GLD(\pi n,\rho,F)$. The notation is that of [15]: the expected number of mutations α is the product of π by the expected final number of cells, the relative fitness (ratio of the growth rate of the population of normal cells divided by that of mutants) is ρ, and the distribution of mutant division times is given by F. As in [15], we assume that a model has been chosen for the distribution of division times, so that only the mutation probability π and the relative fitness ρ are to be estimated.

The sample size being s, for $i=1,\ldots,s$ experiment number i has yielded a couple (m_i,n_i), where m_i is the mutant count and n_i is the final number of cells. As in [14,15], we denote by $q_m(\alpha,\rho)$ the probability of a mutant count equal to m, under the Luria-Delbrück distribution with parameters α (expected number of mutations) and ρ (relative fitness). The computation algorithms of the $q_m(\alpha,\rho)$ are well known and will not be reproduced here: see [12,14,15]. With that notation, the mutant count at the end of the i-th experiment is equal to m_i with probability $q_{m_i}(\pi n_i,\rho)$. No assumption being made on the final counts, we consider the s-tuple of mutant counts $(m_i)_{i=1,\ldots,s}$ as a realization of a sample of independent random variables.

The log-likelihood is:

$$\ell(\pi,\rho)=\sum_{i=1}^{s}\log\left(q_{m_i}(\pi n_i,\rho)\right)\;. \tag{4}$$

The computation of the gradient and Hessian of $\ell(\pi,\rho)$ are only slightly different from those needed for the calculation of the maximum likelihood estimates of α and ρ in the classical case [12,14]. In the formulas below, be shall omit the dependence in (π,ρ) for clarity. The first and second derivatives of ℓ are evaluated at (π,ρ), those of q_{m_i} are evaluated at $(\pi n_i,\rho)$. The gradient is computed by:

$$\frac{\partial\ell}{\partial\pi}=\sum_{i=1}^{s}\frac{n_i}{q_{m_i}}\frac{\partial q_{m_i}}{\partial\alpha}\;,\quad \frac{\partial\ell}{\partial\rho}=\sum_{i=1}^{s}\frac{1}{q_{m_i}}\frac{\partial q_{m_i}}{\partial\rho}\;. \tag{5}$$

The Hessian is computed by:

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\pi^2} &= \sum_{i=1}^{s}\frac{n_i^2}{q_{m_i}}\frac{\partial^2 q_{m_i}}{\partial\alpha^2}-\frac{n_i^2}{q_{m_i}^2}\left(\frac{\partial q_{m_i}}{\partial\alpha}\right)^2,\\
\frac{\partial^2\ell}{\partial\pi\partial\rho} &= \sum_{i=1}^{s}\frac{n_i}{q_{m_i}}\frac{\partial^2 q_{m_i}}{\partial\alpha\partial\rho}-\frac{n_i}{q_{m_i}^2}\left(\frac{\partial q_{m_i}}{\partial\alpha}\right)\left(\frac{\partial q_{m_i}}{\partial\rho}\right),\\
\frac{\partial^2\ell}{\partial\rho^2} &= \sum_{i=1}^{s}\frac{1}{q_{m_i}}\frac{\partial^2 q_{m_i}}{\partial\rho^2}-\frac{1}{q_{m_i}^2}\left(\frac{\partial q_{m_i}}{\partial\rho}\right)^2.
\end{aligned} \tag{6}$$

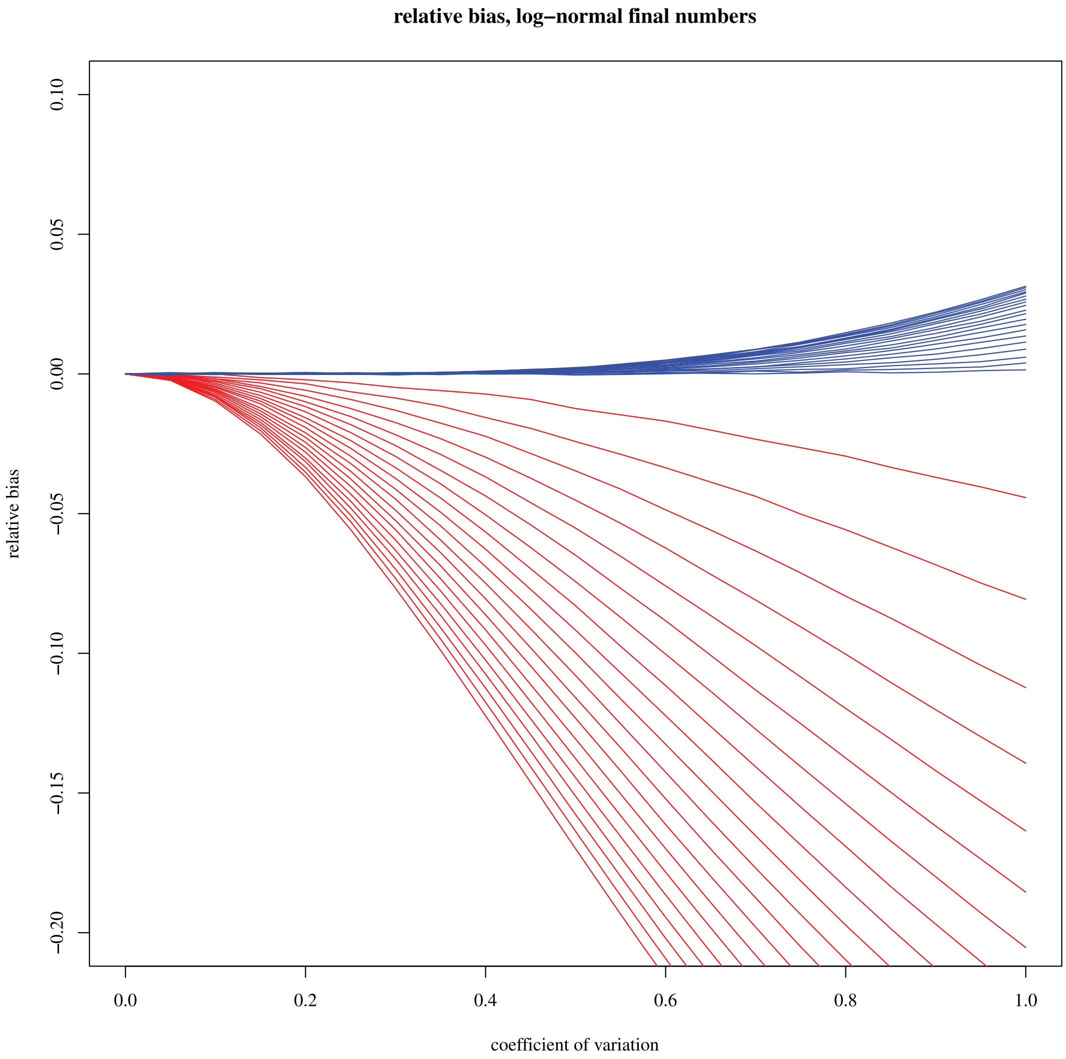

Figure 2. Relative biases on estimates of a mutation rate. Relative biases are plotted as a function of the coefficient of variation C. The different curves correspond to 20 values of $\alpha = \mu\pi$ from 0.1 to 2. Red curves show biases of the p_0-method. For blue curves, the bias has been corrected by the unbiasing factor (1). The correction maintains the bias under acceptable values even for relatively large α and C.
doi:10.1371/journal.pone.0101434.g002

The first and second derivatives of $q_m(\alpha,\rho)$ in α and ρ are obtained by recursive algorithms that will not be reproduced here [12,14].

It is a well known fact in statistics, that the most easy looking maximum likelihood problem usually conceals algorithmic difficulties: numeric instability, bad conditionning of the Hessian, etc. [36]. Here, the procedure looks straightforward from (5) and (6): solving the gradient by a quasi-Newton or conjugate gradient method should be done quite efficiently at low computing cost. However, depending on the values in the sample, some optimization techniques may be more efficient than others. For the results described in this article, we have used the statistical software R [22], and compared several optimization algorithms: quasi-Newton, BFGS, conjugate gradient, simulated annealing [37]. The calculation of the Hessian at the maximum likelihood solution, which is needed to output asymptotic variances poses a numerical problem, already signalled in [14]. For the results of the article a numeric evaluation of the Hessian was used instead of (6) [37]. In File S1, only the simplest method has been included: it consists in solving the gradient by the Raphson-Newton method, from (5) and (6). It is not the best method by far. We are presently working on an optimized implementation, to be included in a forthcoming R package.

Model for simulations

In the simulation study reported in the Results section, we have chosen to draw samples of final numbers according to a log-normal distribution with fixed expectation μ and coefficient of variation C. Other similarly shaped distributions could have been used: gamma, inverse Gaussian, Weibull, etc. Our choice of the log-normal was motivated by fitting real data, and by previously published results: see [16] and references therein.

If some value of the mutation rate π has been fixed, and the final number of cells N has been simulated, a mutant count can be drawn according to a Luria-Delbrück distribution with expected number of mutations $\alpha = \pi N$ and relative fitness ρ. As explained in [15], an additional choice must be made: that of a probability distribution for division times. Neither of the two extreme choices that leads to computable versions of the Luria-Delbrück distribution (exponential and constant division times) is realistic. We have chosen the same distribution as in [15]: the best adjustment on Kelly and Rahn's observation on *Bacterium aerogenes* [38].

Simulations have been conducted for different sets of parameters. Results are reported for the following values, considered as representative:

$$\pi = 10^{-9}\,,\ \mu = 10^{9}\,,\ C = 0.5\,,\ \rho = 1\ .$$

One thousand samples of size 50 of pairs (mutant counts – final numbers) were simulated. For each sample, six estimates of π were computed, together with their theoretical standard deviation.

- *Classical methods*: the estimate of the expected number of mutations α was computed by two different methods: the p_0-method [1,5,28], and the maximum likelihood (ML) method [12,31,32], both applied to the sample of mutant counts. Dividing by the expected final number μ, assumed to be known, leads to two estimates for π.
- *unbiased estimates*: to each of the two previous estimates, the unbiasing formulas (1) and (3) were applied, assuming that the true value of the coefficient of variation was known which lead to two more estimates of π. There again, the expected final number μ was supposed to be known, as well as the coefficient of variation C.
- *p_0-method on the pairs*: no prior information being assumed, the maximum likelihood determination of π by the p_0-method was applied to the sample of pairs mutant counts – final numbers.
- *maximum likelihood for π and ρ*: taking againg the sample of pairs with no prior information, a joint estimation for π was obtained.

Treatment for published datasets

We have reexamined data from David [17], and Werngren & Hoffner [21].

The data in Table 1 of [17] are not detailed, so only the p_0-method could be applied. The bias correction (1) was applied, using a coefficient of variation of 0.35 (estimated from Table 2 in the same reference).

Table 2 of [17] shows 10 pairs mutant counts – final numbers. All possible estimates were computed together with their confidence intervals. However, it must be remarked that standard deviation computations rely upon asymptotic results, and do not apply to such a small sample.

In Table 1 of [21] mutant counts are explicitly given. The maximum likelihood estimate with exponential division time was computed, then unbiased using a coefficient of variation of 0.44 (estimated from the given final counts).

Supporting Information

File S1 File S1 is a script of the R functions that have been used for the simulation experiments described here. It is a preliminary version of a forthcoming R package. The functions have not been protected nor optimized.
(R)

Acknowledgments

We are grateful to Santiago Elena for kindly sending us the data of [18].

Author Contributions

Conceived and designed the experiments: BY NV. Performed the experiments: BY. Analyzed the data: BY NV. Contributed reagents/materials/analysis tools: BY NV. Wrote the paper: BY NV. Designed the software: BY. Analyzed published data: NV.

References

1. Luria DE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28: 491–511.
2. Kendal WS, Frost P (1988) Pitfalls and practice of Luria-Delbrück fluctuation analysis: a review. Cancer Res 48: 1060–1065.
3. Stewart FM, Gordon DM, Levin BR (1990) Fluctuation analysis: the probability distribution of the number of mutants under different conditions. Genetics 124: 175–185.
4. Stewart FM (1994) Fluctuation tests: how reliable are the estimates of mutation rates? Genetics 137: 1139–1146.
5. Foster PL (2006) Methods for determining spontaneous mutation rates. Methods Enzymol 409: 195–213.
6. Pope CF, O'Sullivan DM, McHugh TD, Gillespie SH (2008) A practical guide to measuring mutation rates in antibiotic resistance. Antimicrob Agents Chemother 52: 1209–1214.
7. Jin JL, Wei G, Yang WQ, Zhang HQ, Gao PJ (2012) Discussion on research methods of bacterial resistant mutation mechanisms under selective culture-uncertainty analysis of data from the Luria-Delbrück fluctuation experiment. Science China, Life sciences 55: 1007–1021.
8. Sarkar S, Ma WT, v H Sandri G (1992) On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. Genetica 85: 173–179.
9. Jones ME (1994) Luria-Delbrück fluctuation experiments; accounting simultaneously for plating efficiency and differential growth rate. J Theo Biol 166: 355–363.
10. Jaeger G, Sarkar S (1995) On the distribution of bacterial mutants: the effects of differential fitness of mutants and non-mutants. Genetica 96: 217–223.
11. Zheng Q (2002) Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. Math Biosci 176: 237–252.
12. Zheng Q (2005) New algorithms for Luria-Delbrück fluctuation analysis. Math Biosci 196: 198–214.
13. Gerrish PJ (2008) A simple formula for obtaining markedly improved mutation rates estimates. Genetics 180: 1773–1778.
14. Hamon A, Ycart B (2012) Statistics for the Luria-Delbrück distribution. Elect J Statist 6: 1251–1272.
15. Ycart B (2013) Fluctuation analysis: can estimates be trusted? PLoS One 8: e80958.
16. Koutsoumanis KP, Lianou A (2013) Stochasticity in colonial growth dynamics of individual bacterial cells. Appl Environ Microbiol 79: 2294–2301.
17. David HL (1970) Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis*. Appl Microbiol 20: 810–814.
18. de la Iglesia F, Martínez F, Hillung J, Cuevas JM, Gerrish PJ, et al. (2012) Luria-Delbrück estimation of turnip mosaic virus mutation rate in vivo. J Virol 86: 3386–3388.
19. Angerer WP (2001) An explicit representation of the Luria-Delbrück distribution. J Math Biol 42: 145–174.
20. Komarova NL, Wu L, Baldi P (2007) The fixed-size Luria-Delbrück model with a nonzero death rate. Math Biosci 210: 253–290.
21. Werngren J, Hoffner SE (2003) Drug susceptible *Mycobacterium tuberculosis* Beijing genotype does not develop motation-conferred resistance to Rifampin at an elevated rate. J Clin Microbiol 41: 1520–1524.
22. R Development Core Team (2008) R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.
23. Lea DE, Coulson CA (1949) The distribution of the number of mutants in bacterial populations. J Genetics 49: 264–285.
24. Tan WY (1982) On distribution theories for the number of mutants in cell populations. SIAM J Appl Math 42: 719–730.
25. Dewanji A, Luebeck EG, Moolgavkar SH (2005) A generalized Luria-Delbrück model. Math Biosci 197: 140–152.
26. Ycart B (2014) Fluctuation analysis with cell deaths. J Appl Probab Statist 9: 12–28.
27. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, et al. (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature Genetics 45: 784–790.
28. Fu J, Li IC, Chu EHY (1982) The parameters for quantitative analysis of mutation rates with cultured mammalian somatic cells. Mut Research 105: 363–370.
29. Wasserman L (2004) All of statistics: a concise course in statistical inference. Springer, New York.
30. Dyke P (2001) An introduction to Laplace transforms and Fourier series. Springer, London.

31. Ma WT, v H Sandri G, Sarkar S (1992) Analysis of the Luria-Delbrück distribution using discrete convolution powers. J Appl Probab 29: 255–267.
32. Jones ME, Wheldrake J, Rogers A (1993) Luria-Delbrück fluctuation analysis: estimating the Poisson parameter in a compound Poisson distribution. Comput Biol Med 23: 525–534.
33. Hall BM, Ma C, Liang P, Singh KK (2009) Fluctuation Analysis CalculatOR (FALCOR): a web tool for the determination of mutation rate using Luria-Delbrück fluctuation analysis. Bioinformatics 25: 1564–1565.
34. Sarkar S (1991) Haldane's solution of the Luria-Delbrück distribution. Genetics 127: 257–261.
35. Zheng Q (2007) On Haldane's formulation of the Luria-Delbrück mutation model. Math Biosci 209: 237–252.
36. Gupta NK, Mehra RK (1974) Computational aspects of maximum likelihood: estimation and reduction in sensitivity function calculations. IEEE Trans Automatic Control 19: 774–783.
37. Nocedal J, Wright S (2006) Numerical optimization. Springer, New-York, 2nd edition.
38. Kelly CD, Rahn O (1932) The growth rate of individual bacterial cells. J Bacteriol 23: 147–153.