Fast stochastic simulations and mutation rate inference for the fluctuation tests with death, fitness effect and sampling

Antoine Frénoy

1 Introduction

The study of the process of bacterial mutagenesis is one of the oldest problem in computational biology [21]. Infering the rate of a mutation in a growing bacterial population from the observed number of mutants after growth requires a non-trivial model of growth and mutation in a bacterial population. The original model derived by Salvador Luria and Max Delbrück [21] as well as most subsequents mathematical analyses, developments and refinements [19, 22] (reviewed in [36]) share a set of restrictive biological assumptions regarding population demographics and effect of the focal mutation: (1) bacteria grow (stochastically or deterministically) without death, (2) the mutation does not affect growth rate (and thus has no effect on fitness), and (3) the full population is assayed when counting the number of mutants (no sampling). This historical model is still the one predominantly used today to estimate mutation rate from fluctuation tests [7], and was for example implemented in the popular webtool Falcor [14].

However, departure from this set of hypotheses does not allow for estimation of mutation rate based on this standard model. This is becoming an important – but often ignored – limitation, as several current research questions lead to considering experimental scenarios where these hypotheses are not met. It has for example been found that some mutations classically scored in fluctuation tests are not neutral [9, 13, 25]; the effect of bactericidal stresses on mutation rate has been subject to much scrutinity [18, 4, 12, 8]; and dilution (sampling) before selective plating is required to obtain a countable number of mutants when selecting for loss-of-function mutations [3, 6] or when working with hypermutable strains [26]. Several authors thus proposed refined models and software able to circumvent one or several of these hypotheses. Mandelbrot [23] and Koch [17] considered the case where the mutant grows slower or faster than the wild-type, and their method is available in rSalvador [37]. Hamon and Ycart [15] proposed another method for the same problem, which has been implemented in bz-rates [11]. Stewart and collaborators proposed a method to take partial plating into account [28], which has been suggested by the authors of Falcor and implemented in bz-rates, but Zheng [35] found it to be inaccurate, and proposed another method implemented in rSalvador. Finally, Flan [24] aims at estimating mutation rate in presence of partial plating, differential growth of the mutant, and death of the mutant, but does not consider death of the wild-type.

To our knowledge, no existing software is fully able to work with the general case where none of these hypotheses stand, namely: (1) more complex demographics with non-zero (and possibly non-constant) death rate (of the mutant as well as of the wild-type), (2) the mutation affects fitness (growth and death) of the mutant bacteria, and (3) only a sample of the population is assayed (by selective plating) to detect mutants. We refer to the modeling or simulation of this problem as the general 'Birth-Death-Mutation-Selection-Sampling' (BDMSS) problem.

An analytical model predicting the probability density of the observed number of mutants for a given mutation rate under this general model may be difficult to obtain, hindering estimation of mutation rate by maximal likelihood in the BDMSS problem. Analytical mathematical models can sometimes be replaced by simulations for parameter inference [31, 16]. However the nature of the BDMSS problem raises an important challenge for simulations: mutations are discrete stochastic events, and thus deterministic and continuous models such as ordinary differential equations are not well-suited; but bacterial populations studied *in vitro* are usually large (from around 10^6 individuals for *Mycobacterium tuberculosis* to often more than 10^9 for *Escherichia coli*), rendering individual-based models or stochastic simulations inefficient.

In this work, we aim at (1) developping a stochastic simulation method for the BDMSS problem, several orders of magnitude faster than the standard Gillespie algorithm [10] for *M. Tuberculosis* and *E. coli*-inspired parameters, and (2) showing that approximate-bayesian-computing (ABC) techniques can use these simulations to estimate mutation rate in presence of death, fitness effect and sampling.

2 Very fast stochastic simulations for the BDMSS problem

2.1 Formalization of the problem (forward model)

The Birth-Death-Mutation-Selection-Sampling problem is represented on figure 1. It is an extension of the classical fluctuation test model able to consider the three scenario we discussed: death (of the wild-type and of the mutant), fitness effect of the mutation, and partial plating.

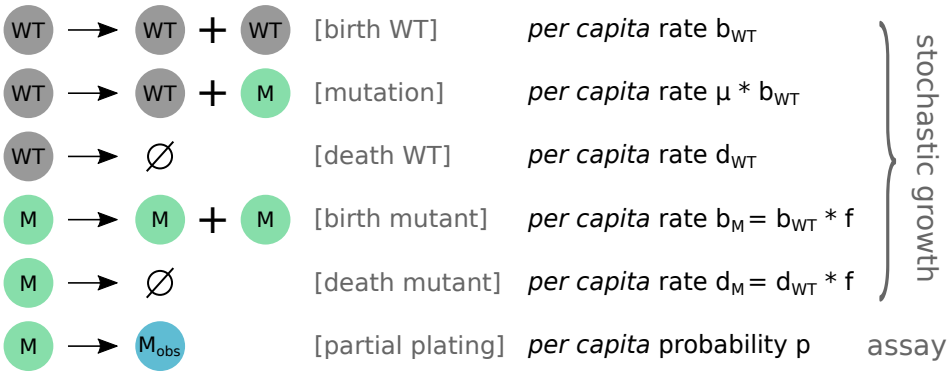

Figure 1: **Forward model.** WT and M represent wild-type and mutant individuals (as classical in fluctuation test, a single class of mutants is considered). M_{obs} represent the fraction of mutants which will be observed when assaying the population by selective plating, and is different from M in case of partial plating (only a sample p of the population is plated). μ is the rate of mutation from wt to m per genome per division. Birth and death rates of the mutant and the wild-type are linked by $b_m = b_{wt} \times f$ and $d_m = d_{wt} \times f$ where f is the fitness effect of the mutation.

Several standard simplifications are made in the above model, not because they are necessary for efficiency but because they are biologically relevant: the per capita rate of the WT birth reaction, formally equal to $(1-\mu)*b_{WT}$, is approximated by b_{WT}, as $\mu \ll 1$; and the reversal mutation is always neglected, as fluctuation tests only consider the case where the mutant population is small compared to the wild-type population.

The standard fluctuation test model can be recovered from our model by taking $p = 1$, $f = 1$, and $d_{WT} = 0$, as empirically verified further below by comparing our model with the Ma-Sandri-Sarkar probability distribution.

In presence of both cell death and fitness effect of the mutation, we chosed to have the same fitness effect term muliplicatively affecting rates of birth and death of the mutant. This could easily be changed: the fitness effect could be decoupled into the effect on birth rate and the effect on death rate. The rational behind coupling both stems from experimental observations that growth and death rates are often strongly linked in bacteria, between different strains [33] or between different growth conditions for the same strain [2], as well as when considering antibiotic-induced death in different physiological conditions [32, 20].

2.2 Fast simulations of the forward model (with constant parameters)

The simulations are initialized with a small value for the number of wild-type individuals N_{wt} and no mutant ($N_m = 0$), and continue until the total number of bacteria $N_{wt} + N_m$ reaches a user-provided value N_{final}. This corresponds to the experimental conditions of the fluctuation test, where populations are inoculated with a small number of wild-type bacteria and grown until reaching carrying capacity. The output of the simulation is the number of mutants observed in the final population (or a sample of the final population if $p \neq 1$). As these simulations are stochastic, when running several replicate simulations, a distribution of observed number of mutants is obtained.

As a reference method, we implemented the standard Gillespie algorithm [10] for this problem. With realistic parameters for fluctuation tests in *Escherichia coli* using antibiotic resistance markers ($N_{final} = 5 \times 10^9$, $\mu = 5 \times 10^{-10}$), the simulation of 100 populations takes 12,088 seconds (more than 3 hours), which is unreasonably slow for parameter inference based on simulation.

We were able to considerably speed up this process (several orders of magnitude, the simulation of 100 populations now taking 0.0287 seconds) by modifying the classical Gillespie simulation scheme, without noticeable changes in the obtained distribution. The key ingredient of this simulation scheme is implementing

a deterministic demographics (growth and death) for the wild-type population but a stochastic mutagenesis process and a stochastic demographics for the mutant population. This fast forward simulator, further refered as *Atreyu formard simulator*, is the basis of our simulation-based, likelihood-free inference method described further below.

The algorithm is described in details in section 4.1, and the source code from the simulation software is publicly available.

In the following part of this work, we apply this forward simulator to the case where death and mutation rates are constant throughout the experiment. The simulator is nonetheless able to simulate non-constant but piecewise constant death and mutation rates.

2.3 Comparaison with other methods

We compare the outcome of these fast stochastic simulations with standard Gillespie simulations; as well as with three pre-existing methods: sampling from the probability distribution given by analytical formula from Ma, Sandri and Sarkar [22] which was for example used by FALCOR [14]; the forward simulator from Flan [24]; and the forward simulator from rSalvador [37] based on the Mandelbrot-Koch model [23, 17].

Without death, fitness effect, or partial plating, the obtained distributions of number of mutants are perfectly similar for all methods (Atreyu forward simulator, Gillespie simulations, sampling from the Ma-Sandri-Sarkar formula, the Mandelbrot-Koch model simulated in rSalvador, and Flan forward simulator; as shown on Figure 2.

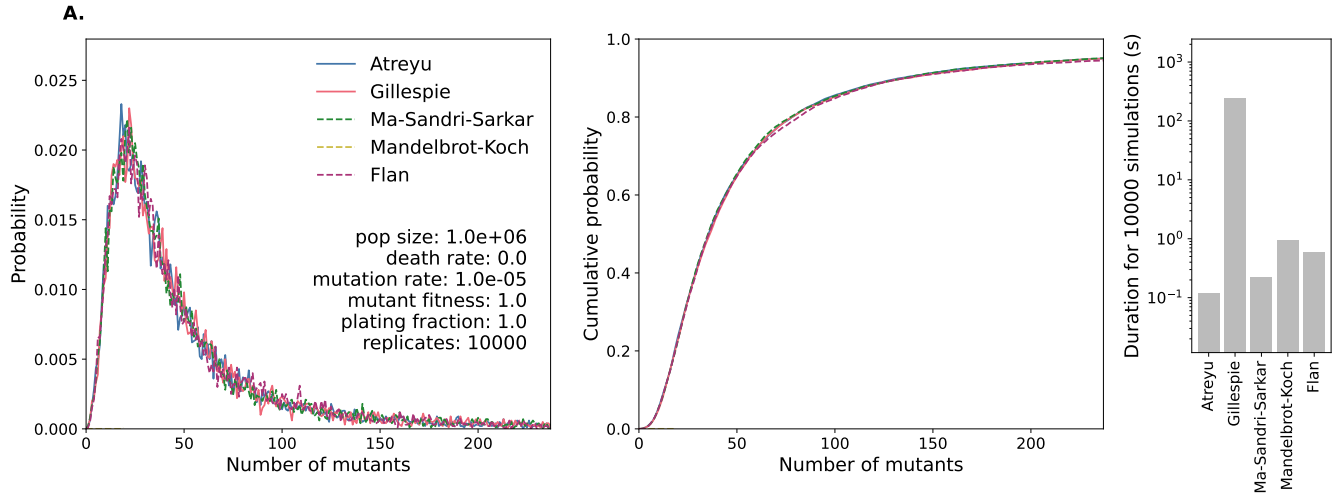

Figure 2: **Comparison of all forward simulators without death, fitness effect or partial plating: all simulators give the same distribution.** 10,000 replicate simulations were performed for each method. Both the probability distribution and the empirical cumulative probability distribution are shown. Continuous lines represent simulations from Atreyu (in blue) or Gillespie algorithm (in red). Dashed lines represent previously existing simulation methods: sampling from the Ma-Sandri-Sarkar probability distribution (in green), simulations of the Mandelbrot-Koch model in rSalvador (in yellow), and simulations from Flan (in purple).

With death, fitness effect or partial plating, some of the simulation methods are not usable anymore, or do not provide a valid output. Atreyu and Gillespie simulations can consider all of these three effects (possibly combined), unlike the other simulation methods.

Only Atreyu and Gillespie simulations can simulate non-zero death rates (Flan also has a death parameter, but it only considers death of the mutant, which is a very different biological model, as further discussed in supplementary materials). With death, the distributions obtained from Atreyu perfectly match those obtained from Gillespie simulations (Figure not shown), but as expected Flan produces a significantly different outcome (Kolmogorov-Smirnov test with distribution from Atreyu: $p < 10^{-9}$), since it does not consider death of the wild-type.

The Mandelbrot-Koch model and Flan can simulate mutations with a fitness effect of the mutation. With fitness effect, the same distributions are obtained with Atreyu, Gillespie simulations, the Mandelbrot-Koch model simulated in rSalvardor, and Flan forward simulator (Figure not shown).

Partial plating (sampling a fraction of the population for assay on selective medium) is not a real challenge for a forward simulator, as a Bernoulli sampling step – resulting in a binomial distribution often approximated by a Poisson distribution – can easily be incorporated in any simulator. It is however a surprinsingly heavy complication for the inference of mutation rate from experimental data. The high number of publications mentionning that overnight cultures were plated on selective medium after *appropriate dilution* without further details (eg [26]) shows how much the importance of this parameter have been classically underestimated.

When combining several or all of these effects for different parameter combinations, we always find that Atreyu and Gillespie simulations give undistinguisable distributions (Figures not shown. TODO: check more aggressively?). As seen on panel C of Figurce 2, Atreyu forward simulator is several orders of magnitude faster than Gillespie simulations for these realistic parameter sets.

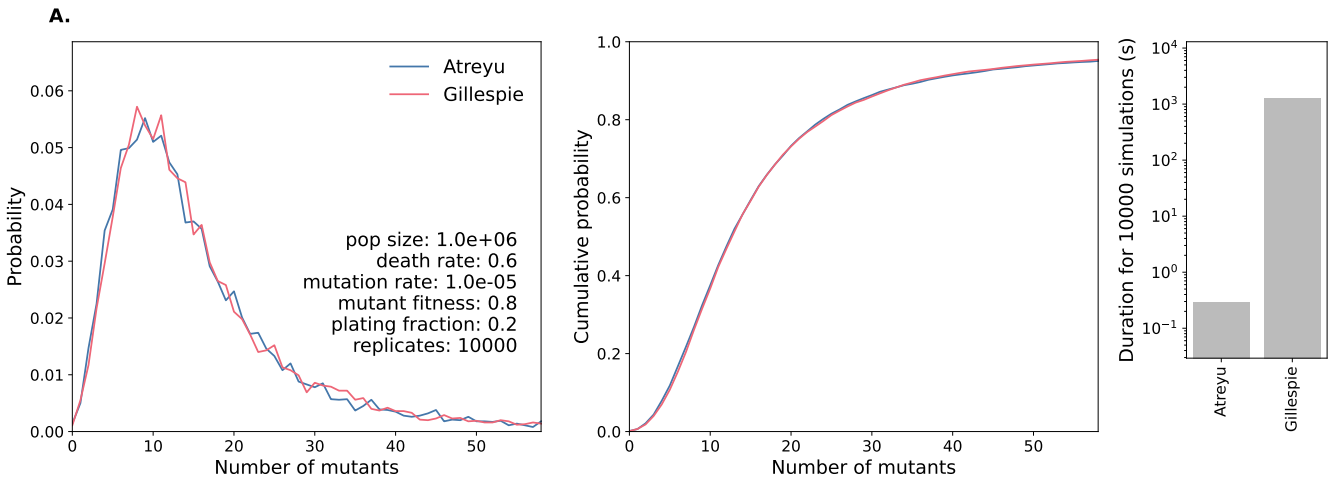

Figure 3: **Comparison of the forward simulators with death, fitness effect and partial plating: Atreyu and Gillespie simulations give the same distribution.** 10,000 replicate simulations were performed for each method. The color code is the same than on Figure 2.

3 Estimating mutation rate with Atreyu

Clasically, methods used to estimate mutation rate from experimental data use an analytical formulation of the distribution of the number of observed mutants as a function of mutation rate (the one given by Ma, Sandri and Sarkar [22] being the most popular). This permits estimation of mutation rate from observed mutant counts (for example using a maximum-likelihood approach, which often performs better than other classical estimators [27]). However, as discussed in the introduction, analytical formulations of the distribution of the number of observed mutants are not always available when the standard hypothesis are not met. Ignoring these discrepancies between the assumptions of the model and the studied biological process may result in important systematic biases (as for example shown by [35] for partial plating and fitness effect, and by [8] for death).

With our simulation method described above several orders of magnitude faster than the standard Gillespie method for the same outcome, replacing the analytical expression by simulations becomes feasable. Departure from the standard model are then no longer problematic, as the inference is based on simulations of the actual biological model describing the concrete experimental conditions instead of the standard model with unmet assumptions. It is therefore possible to infer mutation rate in any model that can be effectively simulated.

ABC is a family of methods to estimate parameter values of a model from experimental data, using simulations of the model rather than analytical expression of the likelihood [31, 29]. It has been used with success in several areas of ecology and evolutionary biology, for example for evaluating parameters of complex population genetics models based on genomics data (reviewed by [5]). In our case, the observed variable is the distribution of the number of mutants in several replicate populations. The known parameters are final population size, death rate, fitness effect, and sampling fraction. The unknown parameter to estimate

is mutation rate. The principle of ABC estimation is to find the parameter value which minimizes the discrepancy between the outcome of the simulation and the empirical value of the observed variable.

TODO: c'est dans un premier temps sur cette partie que vous êtes invités à proposer vos propres solutions

4 Methods

4.1 Fast forward simulator for Atreyu

The main loop simulates each mutation event:

1. The number of divisions $Nbirth_{wt}$ of the wild-type until the next mutation event is drawn from an exponential distribution with parameter μ (which is a fast approximation for a geometric distribution, accurate since $\mu \ll 1$) and is rounded up to the nearest integer.

2. The number of death events for the wild-type which occured during this waiting time to the next mutation is deterministically computed as $Ndeath_{wt} = Nbirth_{wt} \times d_{wt}$, rounded up to the nearest integer.

3. If this number of death and birth events for the wild-type population would yield a total population size higher than the final population size, then the simulation has to be stopped before the mutation: the number of births and deaths of the wild-type is deterministically adjusted to reach exactly final population size, and the numbers of births and deaths of the mutant population is then determined as below. [1]

4. The demographics (number of births and deaths) of the mutant population during this waiting time is computed using a standard Gillespie algorithm. The total time to simulate (expressed in number of synchronous generations) is the time needed for the wild-type population to achieve the already determined number of death and birth events, and is computed as $\frac{log(N_{wt}+Nbirth_{wt}-Ndeath_{wt})-log(N_{wt})}{1-d}$. Following the standard Gillespie algorithm, the loop of this inner simulation consists in the following steps:

    - Determining the propensity of occurence of each reaction (birth of the mutant and death of the mutant) – based on its rate per capita and on the size of the reactant population – and summing them to obtain the total reaction propensity
    - Drawing time to the next event from an exponential distribution whose parameter is the total reaction propensity, and updating the ellapsed time variable. If this ellapsed time becomes higher than the total time to simulate, the inner simulation is stopped.
    - Determining which of the two possible reactions occurs (random draw with probablitity proportional to the reaction propensity)
    - Updating the variable N_m. In case of extinction of the mutant population, the inner simulation is stopped even if the total time to simulate has not been reached.

5. The mutation is performed (except if we determined at step 3 that the simulation has to end before the mutation) and the variables are updated.

4.2 Details of the forward model and its parameters compared to other methods

By default, Flan forward simulator uses a non-standard model for lifetime distribution of mutant cells. In contrast, we use the more classical model in which the probability of division per unit of time is constant, and thus lifetime of cells (wild-type as well as mutant cells) are exponentially distributed. This is also the model underlying the Ma-Sandri-Sarkar estimator and the forward simulator of rSalvador. We thus enforce

[1] This implies that the simulation will not exactly stop at the target final population size, but this discrepancy is minor since the mutant population is considerably smaller than the wild-type population.

the use of this classical model in Flan by passing the parameter `dist=list(exp)` to the forward simulation function `rflan`.

Moreover, the fitness parameter in Flan indicates fitness of the wild-type relative to the mutant, instead of the opposite (fitness of the mutant relative to the wild-type) which is the standard in evolutionary biology. We thus perform the appropriate transformation of the fitness parameter passed to Flan.

Finally, the death parameter in Flan has a different meaning than our relative death rate, for two reasons. First, according to the manual and the associated publication [24], δ is defined such as after its lifetime, a mutant cell has δ chances of dying and $1-\delta$ chances of dividing. Thus $\delta = \frac{d}{1+d}$ where d is the relative death rate in our model. Second, in Flan death rate only applies to the mutant: this implies a radically different biological model, which can not be reconcilied with ours. Our model with death of the mutant and the wild-type stems from a large body of experimental work studying mutation rate under abiotic stress such as sub-inhibitory antibiotic treatments [18, 4, 12, 8]. Importantly, in these studies the applied stress is independent from the neutral mutation used to estimate mutation rate, meaning that the wild-type and the mutant are equally impacted by this stress, matching our model where death applies to both the mutant and the wild-type. The biological foundations of the model implemented in Flan are less clear. Interestingly, the papers [30, 1] cited by the theoretical work behind the death model in Flan [34] considered death of both the mutant and the wild-type (although not necessarily at the same rate).

5 References

[1] W. P. Angerer. An explicit representation of the Luria–Delbrück distribution. *Journal of Mathematical Biology* 42 (2001), 145–174. 10.1007/s002850000053 [sci-hub].

[2] E. Biselli, S. J. Schink, and U. Gerland. Slower growth of Escherichia coli leads to longer survival in carbon starvation due to a decrease in the maintenance rate. *Molecular Systems Biology* 16 (2020), e9478. 10.15252/msb.20209478 [sci-hub].

[3] I. Bjedov et al. Involvement of Escherichia coli DNA polymerase IV in tolerance of cytotoxic alkylating DNA lesions in vivo. *Genetics* 176 (2007), 1431–1440. 10.1534/genetics.107.072405 [sci-hub].

[4] J. Blázquez et al. Antimicrobials as promoters of genetic variation. *Current Opinion in Microbiology* 15 (2012), 561–569. 10.1016/j.mib.2012.07.007 [sci-hub].

[5] K. Csilléry et al. Approximate bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* 25 (2010), 410–418. 10.1016/j.tree.2010.04.001 [sci-hub].

[6] A. L. Decrulle et al. Engineering gene overlaps to sustain genetic constructs in vivo. *PLOS Computational Biology* 17 (2021), e1009475. 10.1371/journal.pcbi.1009475 [sci-hub].

[7] P. L. Foster. Methods for determining spontaneous mutation rates. *Methods in enzymology* 409 (2006), 195–213. 10.1016/S0076-6879(05)09012-9 [sci-hub].

[8] A. Frenoy and S. Bonhoeffer. Death and population dynamics affect mutation rate estimates and evolvability under stress in bacteria. *PLOS Biology* 16 (2018), e2005056. 10.1371/journal.pbio.2005056 [sci-hub].

[9] S. Gagneux et al. The Competitive Cost of Antibiotic Resistance in Mycobacterium tuberculosis. *Science* 312 (2006), 1944–1946. 10.1126/science.1124410 [sci-hub].

[10] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81 (1977), 2340–2361. 10.1021/j100540a008 [sci-hub].

[11] A. Gillet-Markowska, G. Louvel, and G. Fischer. Bz-rates : A web tool to estimate mutation rates from fluctuation analysis. *G3: Genes—Genomes—Genetics* 5 (2015), 2323–2327. 10.1534/g3.115.019836 [sci-hub].

[12] A. Gutierrez et al. B-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. *Nature communications* 4 (2013), 1610. 10.1038/ncomms2607 [sci-hub].

[13] A. R. Hall, J. C. Iles, and R. C. MacLean. The Fitness Cost of Rifampicin Resistance in Pseudomonas aeruginosa Depends on Demand for RNA Polymerase. *Genetics* 187 (2011), 817–822. 10.1534/genetics.110.124628 [sci-hub].

[14] B. M. Hall et al. Fluctuation AnaLysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* 25 (2009), 1564–1565. 10.1093/bioinformatics/btp253 [sci-hub].

[15] A. Hamon and B. Ycart. Statistics for the Luria-Delbrück distribution. *Electronic Journal of Statistics* 6 (2012), 1251–1272. 10.1214/12-EJS711 [sci-hub].

[16] F. Hartig et al. Statistical inference for stochastic simulation models – theory and application. *Ecology Letters* 14 (2011), 816–827. 10.1111/j.1461-0248.2011.01640.x [sci-hub].

[17] A. L. Koch. Mutation and growth rates from Luria-Delbrück fluctuation tests. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 95 (1982), 129–143. 10.1016/0027-5107(82)90252-4 [sci-hub].

[18] M. A. Kohanski, M. A. DePristo, and J. J. Collins. Sublethal antibiotic treatment leads to multidrug resistance via radical-induced mutagenesis. *Molecular cell* 37 (2010), 311–20. 10.1016/j.molcel.2010.01.003 [sci-hub].

[19] D. E. Lea and C. A. Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of genetics* 49 (1949), 264–285.

[20] A. J. Lee et al. Robust, linear correlations between growth rates and beta-lactam–mediated lysis rates. *Proceedings of the National Academy of Sciences* 115 (2018), 4069–4074. 10.1073/pnas.1719504115 [sci-hub].

[21] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28 (1943), 491.

[22] W. T. Ma, G. vH. Sandri, and S. Sarkar. Analysis of the Luria–Delbrück distribution using discrete convolution powers. *Journal of Applied Probability* 29 (1992), 255–267. 10.2307/3214564 [sci-hub].

[23] B. Mandelbrot. *Journal of Applied Probability* 11 (1974), 437–444. 10.2307/3212688 [sci-hub].

[24] A. Mazoyer et al. Flan: an R package for inference on mutation models. *The R Journal* 9 (2017), 334–351.

[25] Q. Qi, G. M. Preston, and R. C. MacLean. Linking System-Wide Impacts of RNA Polymerase Mutations to the Fitness Cost of Rifampin Resistance in Pseudomonas aeruginosa. *mBio* 5 (2014), e01562–14. 10.1128/mBio.01562-14 [sci-hub].

[26] R. M. Schaaper. Mechanisms of mutagenesis in the Escherichia coli mutator mutD5: role of DNA mismatch repair. *Proceedings of the National Academy of Sciences* 85 (1988), 8126–8130. 10.1073/pnas.85.21.8126 [sci-hub].

[27] F. M. Stewart. Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics* 137 (1994), 1139–1146. 10.1093/genetics/137.4.1139 [sci-hub].

[28] F. M. Stewart, D. M. Gordon, and B. R. Levin. *Genetics* 124 (1990), 175–85.

[29] M. Sunnåker et al. Approximate bayesian computation. *PLoS Computational Biology* 9 (2013), e1002803. 10.1371/journal.pcbi.1002803 [sci-hub].

[30] W. Y. Tan. On Distribution Theories for the Number of Mutants in Cell Populations. *SIAM Journal on Applied Mathematics* 42 (1982), 719–730. URL: http://www.jstor.org/stable/2101281 (visited on 03/24/2022).

[31] T. Toni et al. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* 6 (2009), 187–202. 10.1098/rsif.2008.0172 [sci-hub].

[32] E. Tuomanen et al. The Rate of Killing of Escherichia coli by Beta-Lactam Antibiotics Is Strictly Proportional to the Rate of Bacterial Growth. *Microbiology* 132 (1986), 1297–1304. 10.1099/00221287-132-5-1297 [sci-hub].

[33] Y. Yang et al. Temporal scaling of aging as an adaptive strategy of Escherichia coli. *Science Advances* 5 (2019), eaaw2069. 10.1126/sciadv.aaw2069 [sci-hub].

[34] B. Ycart. Fluctuation analysis with cell deaths. *Journal of Applied Probability and Statistics* 9 (2012).

[35] Q. Zheng. A new practical guide to the Luria–Delbrück protocol. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 781 (2015), 7–13. 10.1016/j.mrfmmm.2015.08.005 [sci-hub].

[36] Q. Zheng. Progress of a half century in the study of the Luria–Delbrück distribution. *Mathematical Biosciences* 162 (1999), 1–32. 10.1016/S0025-5564(99)00045-0 [sci-hub].

[37] Q. Zheng. rSalvador: an R package for the fluctuation experiment. *G3:Genes—Genomes—Genetics* 7 (2017), 3849–3856. 10.1534/g3.117.300120 [sci-hub].