Computational and Structural Biotechnology Journal 18 (2020) 791–804

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Selection or drift: The population biology underlying transposon insertion sequencing experiments

Anel Mahmutovic [a], Pia Abel zur Wiesch [a,b,c,d,1], Sören Abel [a,b,d,e,*,1]

[a] Department of Pharmacy, Faculty of Health Sciences, UiT - The Arctic University of Norway, 9037 Tromsø, Norway
[b] Centre for Molecular Medicine Norway, Nordic EMBL Partnership, 0318 Oslo, Norway
[c] Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA
[d] Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA
[e] Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, PA 16802, USA

ARTICLE INFO

Article history:
Received 15 October 2019
Received in revised form 6 March 2020
Accepted 22 March 2020
Available online 25 March 2020

Keywords:
Tn-seq
Transposon insertion sequencing
Population biology
Random birth-death process
Multinomial random sampling
Bottleneck
Distribution of fitness effects
DFE
Drift
Selection

ABSTRACT

Transposon insertion sequencing methods such as Tn-seq revolutionized microbiology by allowing the identification of genomic loci that are critical for viability in a specific environment on a genome-wide scale. While powerful, transposon insertion sequencing suffers from limited reproducibility when different analysis methods are compared. From the perspective of population biology, this may be explained by changes in mutant frequency due to chance (drift) rather than differential fitness (selection).

Here, we develop a mathematical model of the population biology of transposon insertion sequencing experiments, i.e. the changes in size and composition of the transposon-mutagenized population during the experiment. We use this model to investigate mutagenesis, the growth of the mutant library, and its passage through bottlenecks. Specifically, we study how these processes can lead to extinction of individual mutants depending on their fitness and the distribution of fitness effects (DFE) of the entire mutant population.

We find that in typical in vitro experiments few mutants with high fitness go extinct. However, bottlenecks of a size that is common in animal infection models lead to so much random extinction that a large number of viable mutants would be misclassified. While mutants with low fitness are more likely to be lost during the experiment, mutants with intermediate fitness are expected to be much more abundant and can constitute a large proportion of detected hits, i.e. false positives. Thus, incorporating the DFEs of randomly generated mutations in the analysis may improve the reproducibility of transposon insertion experiments, especially when strong bottlenecks are encountered.

1. Introduction

Transposon insertion sequencing is a powerful method to detect genomic loci (e.g., genes) that contribute to growth and survival in a given environment. It relies on genome-wide random disruptions of loci by transposon insertion in a bacterial population and the detection of mutants with transposons at specific insertion sites by sequencing. Mutants with transposon insertions in loci that are important for survival in the tested environment are assumed to be underrepresented in the population of mutated cells. This is the central paradigm of transposon insertion sequencing: The number of sequence reads per locus is correlated with mutant fitness in the tested environment.

In the last decade this technique has been very successfully used in a wide variety of organisms and strains to determine the gene products essential for virulence [1–9], tumorigenesis [10], persistence [11,12], biofilm formation [13] and antibiotic resistance [14,15] both in vivo and in vitro (see [16–18] for a more comprehensive list of applications). While several transposon insertion methods have been developed e.g., Tn-seq [19], Tn-seq Circle [20], INSeq [21], TraDIS [22] and HITS [23], the primary distinction between them is the protocol employed to amplify the transposon-genome junction to identify the transposon insertion site. For simplicity, we will refer to transposon insertion sequencing as Tn-seq in this work.

The fundamental steps shared by all Tn-seq methods are transposon mutagenesis (Fig. 1 – Step 1), growth in a selective environ-

* Corresponding author at: Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA 18602, USA.
E-mail address: soeren.abel@psu.edu (S. Abel).
[1] Authors contributed equally.

https://doi.org/10.1016/j.csbj.2020.03.021

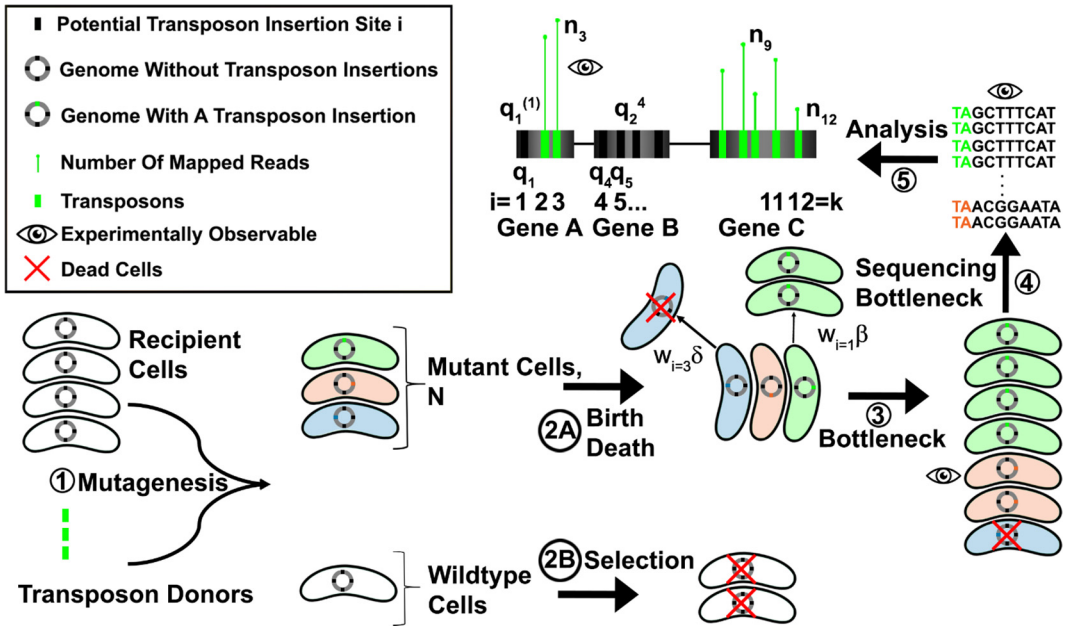

Fig. 1. Schematic of transposon insertion sequencing workflow. Description of individual steps to create a transposon insertion library and/or define essential genes in a specific condition. Not all steps can be easily observed experimentally; we highlight routinely measured quantities (eye symbol). In the first step (1), transposons (colored rectangles) are delivered to recipient bacteria and integrated into the genome (rings) at different positions (i) out of all possible insertion sites (k; black rectangles), resulting in N mutant cells. Wild-type cells grow with a division rate β and a death rate δ. The transposons disrupt genes which can result in altered division rates ($w_i\beta$) and altered death rates ($w_i\delta$) that are specific for the cells bearing a transposon at site i. Typically, experimental constraints lead to inadvertent (or sometimes intended) bacterial growth and death before the library can be analyzed (2AB). This typically serves to select against the wild-type (2B) (dead cells are marked by red x) and leads to a distortion of the mutant frequencies present in the library created by mutagenesis. Sampling of cells (3) can lead to additional distortions. Sampling includes various experimental processes for example harvest of the cells, genomic DNA preparation, and the small amount of genomic DNA subjected to PCR amplification. During the last experimental steps (4), the transposon-genome junctions are prepared for sequencing (exact protocol varies by technique) and then sequenced. Since sequencing capacity is typically limiting, the sequencing bottleneck is another sampling event. Finally, the sequencing data are analyzed (5) by mapping them to the genome and by quantifying the number of sequences per transposon insertion site (n_i) (green bars). The probability of no reads for a transposon insertion site i is given as q_i. The probability of no reads in m sites is $q_j^{(m)}$ (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ment (Fig. 1 – Step 2), selection for mutants (Fig. 1 – Step 3), preparation of the transposon-genome junction for sequencing (depending on the technique) and sequencing (Fig. 1 – Step 4), mapping the sequence reads to the wild-type genome and tallying the number of reads for each transposon insertion site. Afterwards, the reads are analyzed (Fig. 1 – Step 5). During the analysis, the sequence reads are mapped to the genome to identify the respective transposon insertion site, tallied and the loci are categorized based on the tally. Essential insertion sites are those for which the mutants display a strong growth defect relative to the wild-type population, for example when no sequence reads are mapped to all transposon insertion sites within a gene. Accordingly, mutants displaying no growth defect are neutral and mutants that show a strong growth advantage over the wild-type population are categorized as advantageous. Hence, the objective of a Tn-seq experiment is to assess fitness costs where the basic premise is that the number of reads is proportional to fitness, i.e. that changes in mutant frequencies are due to selection.

However, the results are frequently not clear-cut. Repetitions of very similar (if not identical) Tn-seq experiments by different laboratories often have poor overlap [24,25]. Part of the problem is that a number of factors contribute to noisy results and obscure the correlation between the number of reads and fitness. Among them are random birth-death processes and sampling events/bottlenecks. Both can lead to random distortions of mutant frequencies, i.e. drift. Depending on the experimental setup, bottlenecks can randomly remove substantial proportions of the transposon library, especially for in vivo experiments. For example, ~99.99% of an infectious dose of *Vibrio cholerae* is lost during infections of a rabbit model host [26] and ~99.9999% of *Listeria monocytogenes* does not survive orogastric inoculation of mice [27]. In addition, sequencing itself can become a bottleneck if a (too) low number of sequence reads is acquired to analyze an experiment. Statistical methods have been developed to analyze noisy reads and assign essentiality to transposon insertion sites [28–38]. In general, transposon insertion experiments have been greatly successful to generate lists of genomic loci enriched for relevant hits from which researchers pick individual genes or groups of genes with related function. These are then confirmed independently and analyzed further e.g. in [39,40].

Since Tn-seq is a global genetic screening technique, it allows in principle to test all genes in a single experiment simultaneously. However, a global screen poses much higher demands on understanding experimental noise: from the perspective of a single gene or pathway, a much higher false positive rate is acceptable than in a global screen where the false positive rate has to be multiplied with the number of all genes that are tested. Experimental noise, i.e. distortion of mutant frequencies, can be understood with population biological models. Birth-death processes and bottlenecks are well understood in population biology, and both bottlenecks [25] and bacterial growth have been modeled implicitly and explicitly [32,41]. However, understanding random distortions of mutant frequencies requires a population biological model encompassing all processes that add noise.

In addition, when screening globally for a comprehensive list of genes that are crucial for viability in the given environment, it is important to carefully formulate the goal of the screen. After random mutagenesis, most mutants will have fitness defects, even though they may be on average mild. The distribution of fitness effects of random mutagenesis is a matter of intensive research [42]. Since the vast majority of mutations is detrimental, the goal of a Tn-seq experiment is to enrich for those mutants that are sig-

nificantly more impaired than an average random mutant. To assess this, the distribution of fitness effects of random mutations has to be taken into account.

Here, we develop a theoretical modeling framework that can describe the effects of mutagenesis, random birth-death processes and bottlenecks on the composition of a library created by transposon mutagenesis. In contrast to previous approaches, we do take the distribution of fitness effects (DFE) of random mutations into account, i.e. assume that the fitness of the generated mutants is on average slightly lower than the one of the wild-type, with some mutants having very low fitness and only few mutations being beneficial. Our model allows us to address the question whether mutant frequencies changed because of selection or drift. For simplicity and to get conservative estimates, we focus on the extreme case of mutant extinction, i.e. zero reads in a specific locus. While we find that in the absence of strong bottlenecks, mutants that go extinct typically have strongly reduced fitness, we also find that bottlenecks of the size to be expected in animal models add so much random extinction that the fitness of extinct mutants is not substantially larger than what can be expected of any random mutant.

2. Results

2.1. Mutagenesis

A standard transposon insertion experiment proceeds by first creating a pool of insertion mutants whereby transposons are randomly inserted into the genome of cells by means of transposon mutagenesis (Fig. 1). We model mutagenesis so we can characterize the composition of the starting population of N mutant cells by making some simplifying assumptions. The primary assumption is that mutagenesis is completely random where the distribution of mutants over the potential insertion sites is uniform. A uniform distribution of mutants over potential insertion sites requires neglecting the influence of potential genomic cold-/hotspots at which transposons are more or less likely to insert [43]. In addition, this requires that the procedure to create mutants itself does not distort the mutant proportions, for example by more growth of mutants that are created early in contrast to mutants that are created late in the mutagenesis process. Finally, we assume that at most one transposon integrates into the genome. This can for example be experimentally achieved by having an excess of wild-type recipients over transposon donors such that the chance of more than one transposon donor transferring a transposon to a particular wild-type cell is negligible.

Given these assumptions, the mutagenesis process is equivalent to a multinomial random sampling process where N mutants are picked from an infinite pool of uniformly distributed mutants (Fig. 2A). Equivalence is here taken to mean that both the experimental process of mutagenesis and the model result in a uniform distribution of N mutants over the potential insertion sites. A consequence of the uniform distribution is that the chance of picking a mutant cell corresponding to any transposon insertion site i is $1/k$ where k is the number of potential transposon insertion sites. The number of mutant cells where the transposon has inserted into insertion site i independently n time is described by n_i. The probability distribution of obtaining n_i mutants is binomial in the multinomial random sampling model where the average and the variance over repetitions of the mutagenesis experiment is $\mu = N/k$ and $\mu(1 - 1/k)$, respectively.

At this stage, we approximate the binomial distribution as a Poisson distribution for which the variance is equal to the mean, μ. This approximation works well for large N and small probabilities for picking a mutant cell ($1/k$). Typical transposon insertion experiments are particularly amenable to this approximation. For example, for the *Himar1 mariner* transposon k is of the order 10^5 and while it depends on the organism, an N of the order 10^5-10^6 can be achieved in many bacteria. In the context of the Poisson model, the probability C of sampling at least one mutant corresponding to transposon insertion site i is

$$C = 1 - e^{-\mu} \tag{1}$$

In supplementary figure S1 we compare and validate this equation against random sampling simulations. See table 1 for an explanation of all variables in this paper.

Because mutants are uniformly distributed over k, repetitions over experiments are equivalent to repetitions over transposon insertion sites. Therefore, the probability of a given number of independent mutants with transposon insertions at site i is equal to the proportion of mutants with that given number in a single experiment. Therefore, C is the proportion of transposon insertion sites with at least one mutant, often referred to as the library complexity [16]. The library complexity C is experimentally easily observable and serves a good measure for how comprehensive the Tn-seq screen is. Finally, since C is the probability of sampling at least one mutant for transposon insertion site i, 1-C is the probability of sampling zero mutants. The relationship between the number of mutants (N), the library complexity (C) and the potential number of transposon insertion sites (k) (eq. (1)) is illustrated in (Fig. 2B) and validated against Monte-Carlo simulations in (Fig. S1).

2.2. Growth of mutant library, a random birth-death process

The next step in the workflow of a transposon insertion sequencing experiment involves growing the mutant library in a bacterial growth medium (Fig. 1) and is typically required to select against wild-type cells that did not receive a transposon. While this step can be used to simply prepare a mutant library for later investigation, it can also be used to identify the genetic elements for which disruption has a strong fitness effect on the mutant strain under the respective growth conditions, often called "essential gene analysis" [44–50]. Over time, mutants with a low fitness will decrease in frequency, while mutants with a high fitness will increase. The growth process can be understood as random birth-death events and changes the abundance of mutant cells. This can even lead to extinction of mutants, especially when the initial mutant population size is small, either because the mutant was not created often during the mutagenesis process in the first place or because its abundance decreased over time. For simplicity and clarity, we will first describe the changes in mutant frequencies due to fitness differences (Section 2.2.1). In a second step, we will investigate how mutants may disappear (i.e. lead to zero reads) (Sections 2.2.2 and 2.3).

2.2.1. Frequency changes due to fitness costs

The effect of a transposon insertion at site i can in general be complex and can affect the baseline (wild-type) division rate, β, and the baseline death rate, δ, differentially. Here, the meaning of the division rate β is the inverse of the average division time of a cell and conversely, the death rate δ is the inverse average of the time it takes for a wild-type cell to die. Usually, bacterial division and death rates are unknown for cells and would need to be determined for all mutants by complex setups, e.g. by single cell microscopy [51] or the plasmid segregation method [52]. Hence, we do not distinguish between the fitness effects on division and death, but investigate the net division rate, i.e. the net change of the bacterial population size over time of the transposon mutant. In contrast to the division and the death rate, the net division rate is easily observable and is described as $r = \beta$-δ for wild-type cells.

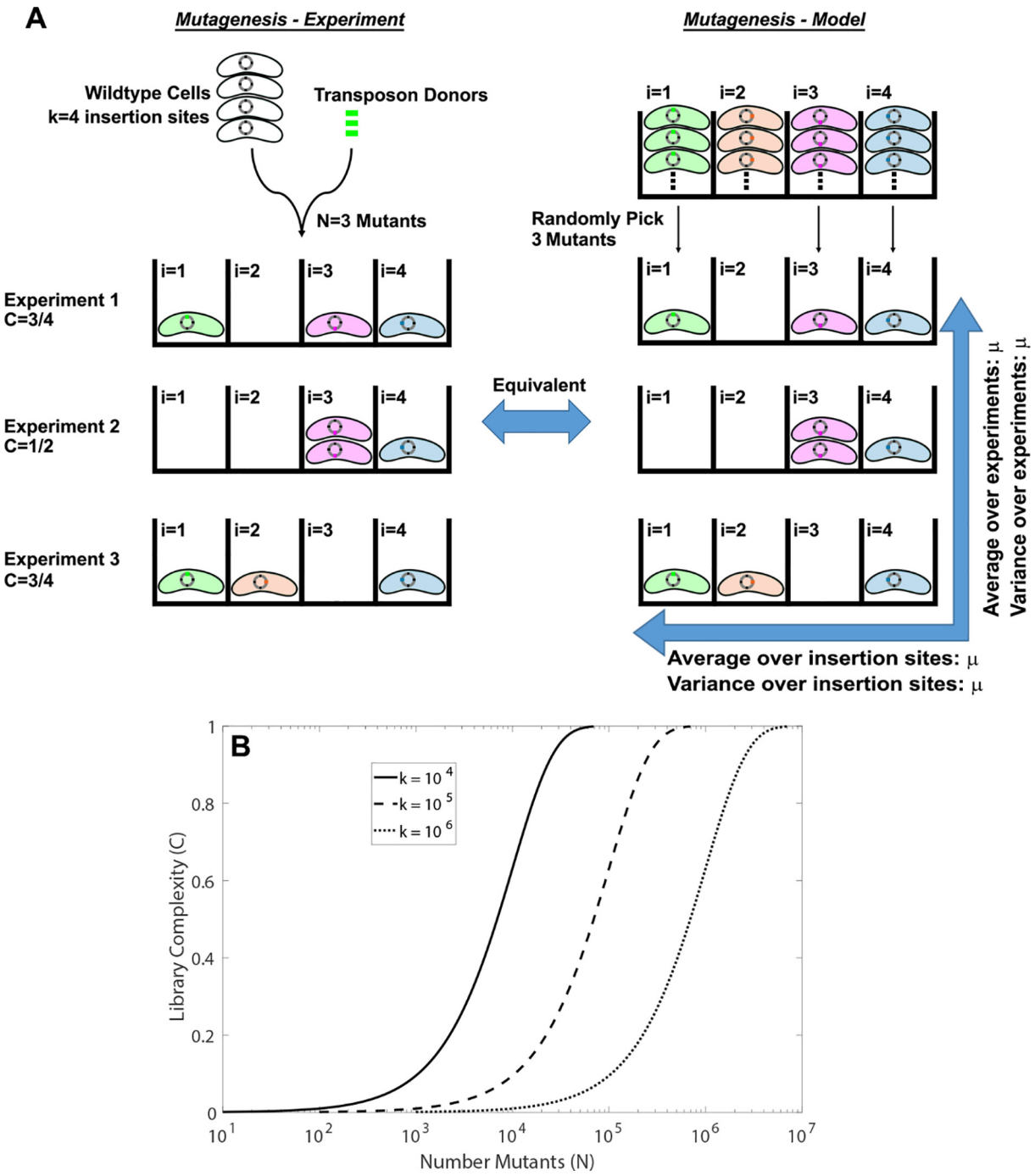

Fig. 2. Illustration of mutagenesis model and the resulting correlation between library complexity and number of mutants. (A) Transposon mutagenesis leads to a pool of uniformly distributed mutant cells over transposon insertion sites *i*. Illustration of the experimental mutagenesis and the multinomial random sampling model that leads to the same distribution. We show three examples of experiments in which three independent transposon mutants *N* were created. In this example the transposon (colored rectangle) can insert in any of four potential transposon insertion sites *k* (black rectangles) on the genome (rings) and depending on the insertion site, can lead to different library complexities *C*. This is equivalent to randomly picking *N* mutants from an infinite pool (triple dots) of uniformly distributed mutants. Both the experiment and the model lead to the same distribution of cells with mean μ and variance μ. Because the probability of a transposon to insert to any transposons insertion site is equal, the mean and the variance can be determined either over multiple repetitions of the same experiment or over the different transposon insertion sites within a single experiment (arrows). (B) The relationship between the number of mutants after mutagenesis, *N*, and the library complexity, *C*, for *k* potential insertion sites where *k* was set to 10^4 (solid), 10^5 (dashed) and 10^6 (dotted). Eq. (1) was used and solved for the number of mutants $N = k\ln(1\text{-}C)$ to generate the plot where the range of *C* was set to 0.001 to 0.999. The larger the number of potential insertion sites, the more mutants are needed to reach a given library complexity shown as a shift of the graph to the right for larger *k*.

For the mutant population at insertion site i this becomes $w_i r$. Accordingly, the average number of mutants for insertion site i at time t is,

$$< n_i(t) >= n_i(0)e^{w_i rt} \quad (2)$$

where $n_i(0)$ is the mutant population size corresponding to transposon insertion site i at the start of the birth-death process. The fitness coefficients in Eq. (2) are real numbers and can be either positive or negative. For example, the meaning of a negative fitness coefficient w_i is that the mutant subpopulation for transposon

Table 1
A summary of the variables used in this work.

Variable	Meaning	Comments
i	A potential transposon insertion site in the genome of the cell.	The range is i = 1,2,. . .,k. When focusing on extinction probabilities within a gene: i = 1,2,. . .,$m = k_G$. In the main paper, we consider complete extinctions of all sites within a gene. In the supplementary material, we show the equations for quantifying extinctions for a subset of insertion sites, m, with k_G being the potential number of transposon insertion sites within a gene (Supplementary data S1).
k_G	Number of insertion sites per gene/locus	This variable is used in the supplementary material to quantify extinctions for a subset of insertion sites, m, with k_G being the potential number of transposon insertion sites within a gene (Supplementary data S1).
k	Total number of potential transposon insertion sites in the genome.	The value of k depends on the specific transposon used in the experiment and the wild-type organism and strain.
N [a]	Number of mutants in the mutagenesis step.	
β	Division rate of wild-type cells.	Defined as the inverse average time it takes for a wild-type cell to divide.
δ	Death rate of wild-type cells.	Defined as the inverse average time it takes for a wild-type cell to die.
w_i	Fitness coefficient.	Throughout the paper we assume that the effect of inserting a transposon into the genome of wild-type cells is to modify the net growth rate by $w_i r$ with $w_i > 0$ and $r > 0$.
$<n_i(t)>$ [b]	Average number of mutants with a transpons insertion at site i at time t.	
m	Subset of insertion sites in one locus that are simultaneously extinct	
$q_{i,growth/bottleneck}^{m}$	Extinction probability of all mutants with transposon insertions in gene i in all m sites.	We use $q_{i,growth}$ if the extinction is due to a random birth–death process or $q_{i,bottleneck}$ if the extinction is due to a random sampling event.
μ	Number of mutants per potential number of transposon insertion sites.	
C	Library complexity after mutagenesis.	Defined as the number of transposon insertion sites with at least one transposon insertion divided by the potential number of transposon insertion sites.
r	Net growth rate of wild-type cells.	The net growth rate is the difference between the division rate β and the death rate δ.
t	Time of growth	
s	Selection coefficient $s = w$-1	
$<N(t)>$	Total average number of mutants at time t.	
$f_i(t)$	The proportion of mutants with a transposon insertion at site i.	$<f_i>$ denotes the average proportion of mutants where the average is taken over realizations.
Z_m	Average number of zero reads over m transposon insertion sites.	Z_k, i.e. the average number of zero reads over all transposon insertion sites in the genome, is used to calculate the reduction in library complexity due to random birth–death events and bottlenecks.
n_s	Sampling size.	
b	Bottleneck size	Sample size relative to the total mutant population size N.
C'	The library complexity after growth/death or sampling.	
a_i	The base of Eq. (5): $a_i = \frac{\delta - \delta e^{w_i r t}}{\delta - \beta e^{w_i r t}}$	Introduced for notational simplicity to express Eq. (9) in an easily accessible form. The variable a_i carries the interpretation of the extinction probability of a mutant population consisting of a single cell for transposon insertion site i.

[a] All cell numbers are implicitly expressed as per unit volume.
[b] Averages over repetitions are denoted with angular brackets <>.

insertion site i is dying if the wild-type population of cells is on average growing (r > 0). Likewise, if $w_i > 0$ and $r > 0$ then a small value for w_i means that the mutant cells grow at a slower rate than wild-type cells.

We only consider mutants that would be able to grow under the measured conditions, because only those will contribute to the population dynamics in the long term. Depending on the definition of essentiality, mutants that are able to grow but go extinct during the experiment can be regarded as false-positives. It has been argued that random mutations with positive fitness coefficients follow a gamma distribution [53], and consequently we sample the fitness of our mutant library from a theoretical gamma distribution. The modelling approach is illustrated in (Fig. 3). However, the quantitative method presented here works just as well for any real numbers w_i and r.

Eq. (2) is based on well-established theory on stochastic birth-death processes [54] where $<n_i(t)>$ is the average over multiple stochastic trajectories. This equation formalizes the assumption underlying all Tn-seq analysis: The number of sequence reads are proportional to the fitness cost of a transposon mutant [19]. When the population size goes to infinity (and the dynamics can be described by deterministic models), stochastic fluctuations do not exist and we do not need to take the average of several realizations. Hence $<n_i(t)>$ can be replaced with $n_i(t)$. This is valid for large population sizes for example during exponential growth. The net growth rates for the mutant subpopulations in Eq. (2) are $w_i r$ where $r = \beta$-δ is the net growth rate for wild-type cells. Mutant cells therefore divide with a uniformly scaled division rate $w_i\beta$ and die with a death rate $w_i\delta$. Developing Eq. (2) a bit further, the total average population size $<N(t)>$ is simply the sum of the average mutant subpopulation sizes (eq. (2)) over all potential transposon insertion sites,

$$< N(t) >= \sum_{i=1}^{k} n_i(0) e^{w_i r t}. \tag{3}$$

Hence the average proportion of mutants at insertion site i ($<f_i(t)>$) is approximately $< f_i(t)>= <n_i(t)/N(t)>\approx<n_i(t)>/<N(t)>$ for which the equation reads after substituting in Eqs. (2)–(3),

$$< f_i(t) >\approx \frac{n_i(0) e^{w_i r t}}{\sum_{i=1}^{k} n_i(0) e^{w_i r t}}. \tag{4}$$

The approximation is a consequence of the fact that the population sizes in a stochastic birth-death process are random for which the average of a ratio of random variables is not exactly equal to the ratio of the averages. Eq. (4) becomes an equality for infinite

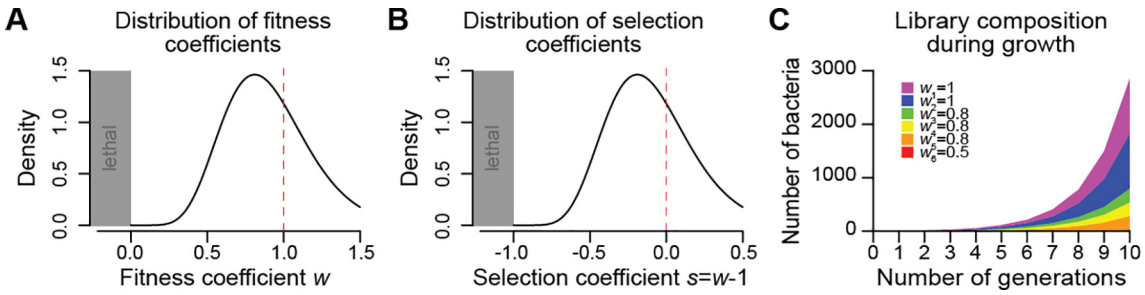

Fig. 3. Modelling growth of mutant library. (A) Illustration of the distribution of fitness coefficients (distribution 1). We assume a gamma-distribution for a mutant library created by transposon insertion [53] with a shape parameter of 10 and a scale parameter of 0.09. The fitness of the wild-type, $w = 1$, is highlighted by the red dashed line. We do not model lethal mutants with a fitness below 0 that would have a negative net growth rate. (B) Same as (A), illustrating the distribution of selection coefficients $s = w\text{-}1$. (C) Illustration of mutant composition of an arbitrary mutant library during exponential growth. The x-axis shows the time in bacterial doubling times, the y-axis shows the number of bacteria. There are six mutants with fitnesses $w_1 = w_2 = 1$ (violet and blue), $w_3 = w_4 = w_5 = 0.8$ (green, yellow and orange), $w_6 = 0.5$ (red). At $t = 0$, the simulation starts with one mutant of each of the six genotypes and follows them for 10 generations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

population sizes, i.e. for deterministic models. Note that while the mutant subpopulations grow independently of each other, the average proportion of mutants at insertion site i depend on the population sizes of the mutants at all insertion sites. To rephrase this, the average proportion of mutants at insertion site i depends on the fitness distribution in the entire population. As such, calculated quantities that depend on the average proportion of mutants will be sensitive to the fitness distribution of the whole population. This can for example complicate the analysis of Tn-seq experiments that compare two mutant libraries generated in different parental backgrounds.

2.2.2. Random extinction due to birth-death events

One of the major readouts of a Tn-seq experiment is the absence of any insertions in a gene. This is often taken as a sign that the gene in question is essential under the conditions in which the experiment was conducted. To further improve our quantitative understanding of the underlying birth-death processes we investigate mutant extinction due to stochastic fluctuations and quantify the average number of extinct mutant populations over m sites, (Z_m). For instance, Z_{50}, is the average number of extinct mutant populations in a gene with 50 potential transposon insertion sites. Based on random birth-death processes, the extinction probability of all mutants for one transposon insertion site i ($q_{i,growth}$) reads,

$$q_{i,growth} = \left\{\frac{\delta - \delta e^{w_i rt}}{\delta - \beta e^{w_i rt}}\right\}^{n_i(0)} \tag{5}$$

where $n_i(0)$ is again the mutant population size corresponding to transposon insertion site i at the start of the birth-death process. The extinction probability corresponding to a particular transposon insertion site i within a gene is dependent on only the fitness cost of the transposon insertion in that particular site. In other words, the extinction probability of each individual mutant due to stochastic fluctuations is independent of the fitness distribution of fitness values of all mutants (DFE). This stems from the assumption that all mutant subpopulations grow and die independently of each other. Additionally, the extinction probability depends on not only the net growth rate for the wild-type population but also the baseline turnover of cells under the investigated conditions, i.e. the division rate and death rate of wild-type cells. Therefore, to quantitatively answer whether zero transposon insertion reads are due to a significant fitness cost (i.e., the gene is "essential") would require knowing the wild-type division rate β, the wild-type death rate δ, and the mutant subpopulation sizes $n_i(0)$ at the start of the birth-death process.

We get Z_m, the average number of extinct mutant populations over m sites in a given gene, by recognizing that the extinction probability of mutants for insertion site i (q_i) is equal to the average number of times that the mutants go extinct for site i. Therefore, the average number of extinct mutant populations over m sites is

$$Z_m = \sum_{i=1}^{m} q_{i,growth} \tag{6}$$

where we have labeled the insertion sites within a gene as $i = 1,2,\ldots, m$. Moreover the extinction probability of all mutants in a gene with m potential insertion sites is $q^m_{i,growth}$. For the sake of simplicity we will focus on complete extinctions of all sites within a gene and use $q^m_{i,growth}$ to calculate their extinction probability. In the supplementary material (Supplementary data S1, Figure Supplementary figure 1, Supplementary figure 2 and Supplementary figure 3), we show how to get the extinction probability of mutants corresponding to m sites in a gene with k_G potential insertion sites where the combinatorics of counting the number of ways that m extinction events can be realized is taken into account.

In summary, this section sheds light on the factors that influence extinction probabilities due to a random birth-death process while cells grow in liquid culture. The following example illustrates the importance of the baseline division and death rates. Two identical mutant populations, i.e. the same mutant population sizes $n_i(0)$ with the same relative fitness w_i, are grown for the same time span with different baseline division and death rates, for example by growing them in media with different nutrient content. Even though the mutants have the same relative fitness compared to the wild-type at the respective growth conditions, the extinction probabilities of each mutant differ in the two environments because of their dependence on baseline division and death rates.

Another example illustrates that the extinction probability during a birth death process is independent of the fitness of other cells, i.e. the fitness distribution in the entire population (DFE). We simulate the growth of two different mutant libraries in the same media for two hours and for four hours (Fig. 4). These libraries need not have the same fitness distribution, however, we assume that the wild-type populations grow with the same division rate (β) and the same death rate (δ). Over time the mutant populations grow, i.e. the total number of cells increases, seen as an increase in the area of the distribution (Fig. 4ABC). Simultaneously, the relative abundance of a specific mutant changes over time because fitter mutants, larger w_i, grow faster than mutants with higher fitness costs, (w_i-1). How the relative abundance changes depends on the fitness distribution in the entire population. In our example, this is evident by the change in the binned mean fitness values of the mutant library over time (Fig. 4ABC). This also means that

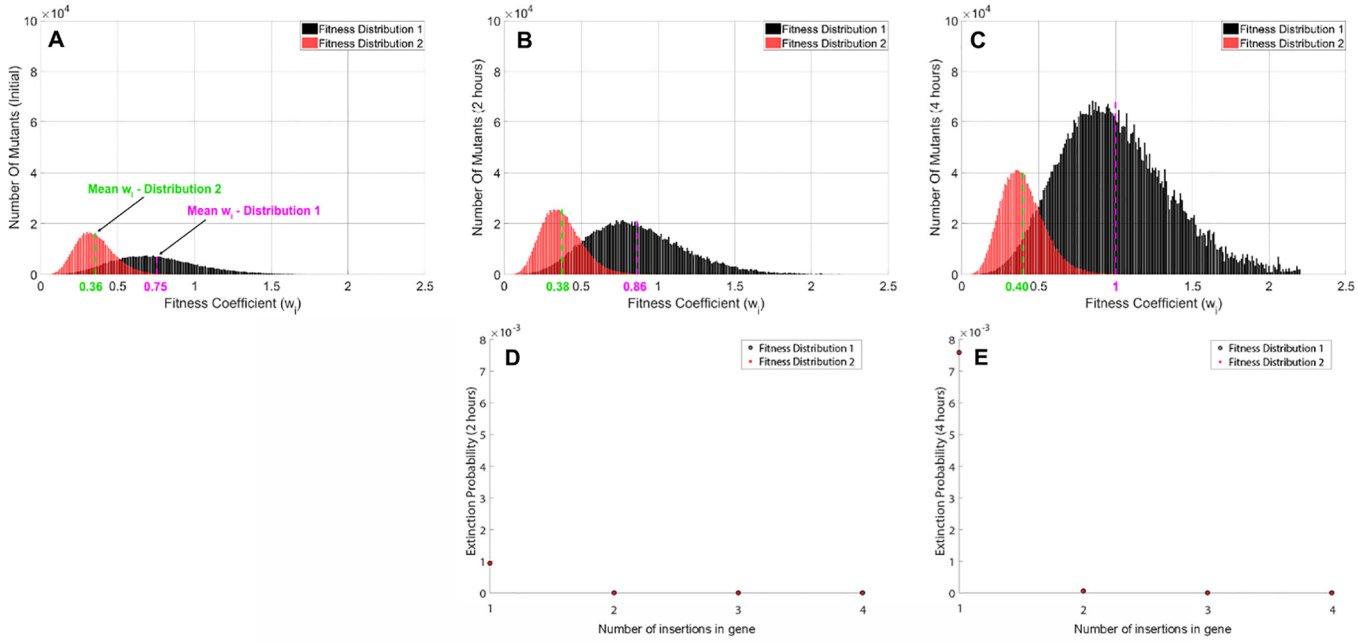

Fig. 4. Random birth/death process and extinctions for two different DFEs. In this graph we compare the dynamics of two mutant libraries with different DFEs. The DFEs are gamma-distributed with shape parameter 10 and scale parameter 0.09 (distribution 1, black) and 0.04 (distribution 2, red). The number of potential transposon insertion sites (k) was set to 10^5 with the number of mutant cells set to 5 at $t = 0$ for $i = 1,2,\ldots, 10^5$. In the top panel, the fitness coefficient w_i is shown on the x-axis and the number of mutants i with the corresponding fitness is shown on the y axis. The fitness coefficients were binned using a bin width of 0.01 for w_i values between 0.01 and 2.5 for both distributions. The binned mean fitness values (magenta and green dashed vertical lines) were calculated by summing $w_i f_i$ over i where i is the number of bins (250) and f_i is the proportion of mutant cells in bin i. (A) The number of mutants present at the start of a birth-death process. (B) The distribution of the number of mutants over the fitness coefficients after 2 h of growth (Eq. (2)) with a baseline division rate set to $\beta = 0.03\ \text{min}^{-1}$ and a baseline death rate set to $\delta = 0.02\ \text{min}^{-1}$. (C) The distribution of the number of mutants over the fitness coefficients after 4 h of growth with the same rates as in (B). The bottom panel shows the extinction probability (y-axis) for all mutants corresponding to 1–4 insertion sites within a gene (x-axis). Eq. (5) was used to calculate the extinction probabilities where w_i was either sampled from distribution 1 (black) or distribution 2 (red). All insertion sites within an essential gene have the same fitness cost with w_j arbitrarily chosen and set to 0.15 to represent a gene with high fitness costs. (D) Extinction probabilities after 2 h of growth. (E) Same as (D) except the mutant cells have been growing for 4 h. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a specific mutant with the same fitness in both mutant libraries will be present in different proportions at the same time-point. Nevertheless, the extinction probabilities for two mutants with the same relative fitness would be the same for both mutant libraries as a consequence of the extinction probability being independent of the fitness distribution (Fig. 4DE). Even for mutants with a substantially reduced fitness ($w_i = 0.15$), extinction is very unlikely. After two hours (Fig. 4D), extinction probabilities for the mutants corresponding to 1, 2, 3 and 4 insertion sites within a gene are 9.3×10^{-4}, 8.7×10^{-7}, 8.2×10^{-10} and 7.6×10^{-13} respectively. Hence, on average a single insertion site in a gene goes extinct every $1/9.3 \times 10^{-4} \approx 1000$ repetitions of the experiment with more than one extinction event being orders of magnitude less likely. The extinction probability of all mutants within a gene increases with time, though it remains independent of the fitness distribution for a constant baseline division rate and death rate. After four hours, the extinction probabilities corresponding to 1, 2, 3 and 4 insertion sites within a gene are 7.6×10^{-3}, 5.8×10^{-5}, 4.4×10^{-7} and 3.3×10^{-9}, respectively. Therefore, growing the cells for two additional hours increases the extinction probability by approximately one order of magnitude where there is, on average, one extinction event of a single insertion site in $1/3.4 \times 10^{-3} \approx 130$ repetitions of the experiment.

2.3. Random sampling events (bottlenecks)

In addition to stochastic fluctuations during bacterial growth and death, sampling events (also known as bottlenecks) contribute to changes in mutant frequencies and may lead to extinction of mutants. Bottlenecks are frequently encountered during several steps of a typical Tn-seq experiment; some of which are unavoidable. For example, pipetting and sequencing can constitute bottlenecks. During pipetting, often a small volume is taken from a larger volume and during sequencing, the number of sequences acquired is limited. In addition, when Tn-seq studies are performed in vivo in animal models, the host defenses of the animal will impose additional bottlenecks [55]. For example, only one in 10^4 *V. cholerae* will contribute to colonization in a rabbit model after inoculation [56].

Generally speaking, all bottlenecks result in random distortions of mutant frequencies and extinction of mutants, both of which are independent of mutant fitness. Since the underlying assumption of all Tn-seq experiments is that mutant frequencies change depend on mutant fitness, bottlenecks add noise to the experimental readouts. This can be illustrated with an extreme example: When only one cell makes it through a bottleneck, the absence of all other mutants at the end of the experiment does then not state much about their fitness.

The aim of this section is to formalize this intuitive reasoning and to predict mutant extinction due to random sampling. To model bottlenecks we use the multinomial random sampling model and the Poisson approximation for the probability distribution of sampling n_i mutants for insertion site i. The multinomial random sampling model and the accompanying Poisson approximation are elaborated upon in the mutagenesis Section 2.1. In contrast to the mutagenesis model however, the total population size is finite. As a consequence, the multinomial random sampling model will only be accurate if small samples, n_s, are taken from a

very large population size N such that perturbations in the proportions of mutants for insertion sites i are negligible. Based on the Poisson model the probability $q_{i,bottleneck}$ of all mutants for transposon insertion site i going extinct reads

$$q_{i,bottleneck} = e^{-n_s f_i} \quad (7)$$

where f_i is the proportion of mutants for insertion site i prior to sampling and $n_s f_i$ is the average number of mutant cells for transposon insertion site i after sampling. The extinction probability due to a bottleneck is a random variable if the frequency of the mutant i is itself a random variable due to preceding random processes. In our framework (Fig. 1) this could be a birth-death process and/or mutagenesis (or potentially another sampling event). To account for this, we move from the extinction probability of mutant i in a single experiment, $q_{i,bottleneck}$, to the average extinction probability when repeating experiments, $\langle q_{i,bottleneck} \rangle$. The average extinction probability can be estimated by using the error propagation method for which the simplest estimate is a substitution of the proportion of mutants f_i with the average proportion of mutants $\langle f_i \rangle$ in Eq. (7). In the context of library complexity reduction (Section 2.4), we include higher order terms that depends on the variances and covariances in the proportion of mutants to estimate the average extinction probability (Supplementary data S2).

The interpretation of Eq. (7) and the connection to experiments depend on the experimental protocol prescribing how a sample is to be taken. Here, we distinguish between absolute and fractional bottlenecks [55]. In an absolute bottleneck the number of mutants that are present after a bottleneck is constant and independent of the pre-bottleneck population size. An example of such a bottleneck is the sequencing step, when a limited and constant number of sequences (often around 10^{6-7}) are read. When a population undergoes a fractional bottleneck, the number of mutant after a sampling event is always the same fraction of the original population, i.e. directly proportional. Such a bottleneck is encountered for example during pipetting where a fixed fraction of the total volume, independently from the concentration of the cells, genomic DNA, etc., is processed.

The effect of absolute bottlenecks on mutant frequencies is illustrated in (Fig. 5), where we take the two mutant libraries that underwent a birth-death process from Fig. 4 and sampled 10^6 cells from them. The initial mutant population (Figs. 4A and 5A) for 10^5 potential insertion sites (k) contains 5 mutants for each individual transposon insertion site i with a fitness coefficient drawn from two different gamma distributions (red and black). This distribution and the total number of cells changes during the two hours or four hours long birth-death process (Fig. 4BC). While bottlenecks result in random distortions of mutant frequencies for single realizations of Tn-seq experiments they do not change the average frequencies. Since the only difference between Figs. 4 and 5 is an added bottleneck, the mean fitness of the resulting distributions are identical. However, the extinction probability of individual mutants might strongly differ. In Fig. 5DE, we show the extinction probabilities of an arbitrary mutant with a fitness coefficient of $w_i = 0.15$ after a random birth-death process and a subsequent bottleneck as shown in Fig. 5BC.

After two hours of growth, the extinction probabilities for the mutants corresponding to 1,2,3 and 4 insertion sites within a gene are 0.02186, 4.8×10^{-4}, 1.0×10^{-5} and 2.3×10^{-7} for the first distribution (black) and 6.7×10^{-4}, 4.4×10^{-7}, 2.9×10^{-10} and 1.9×10^{-13} for the second distribution (red) (Fig. 5D). Hence, there is on average one extinction event per transposon insertion site in $1/0.02186 \approx 46$ repetitions of the experiment for the first distribution with more than one extinction event being orders of magnitude less likely. The extinction probabilities increase with additional time to grow (Fig. 5E). The extinction probabilities corresponding to 1, 2, 3 and 4 insertion sites within a gene are 0.2828, 0.08, 0.02263 and 0.0064 for the first distribution and 0.005,

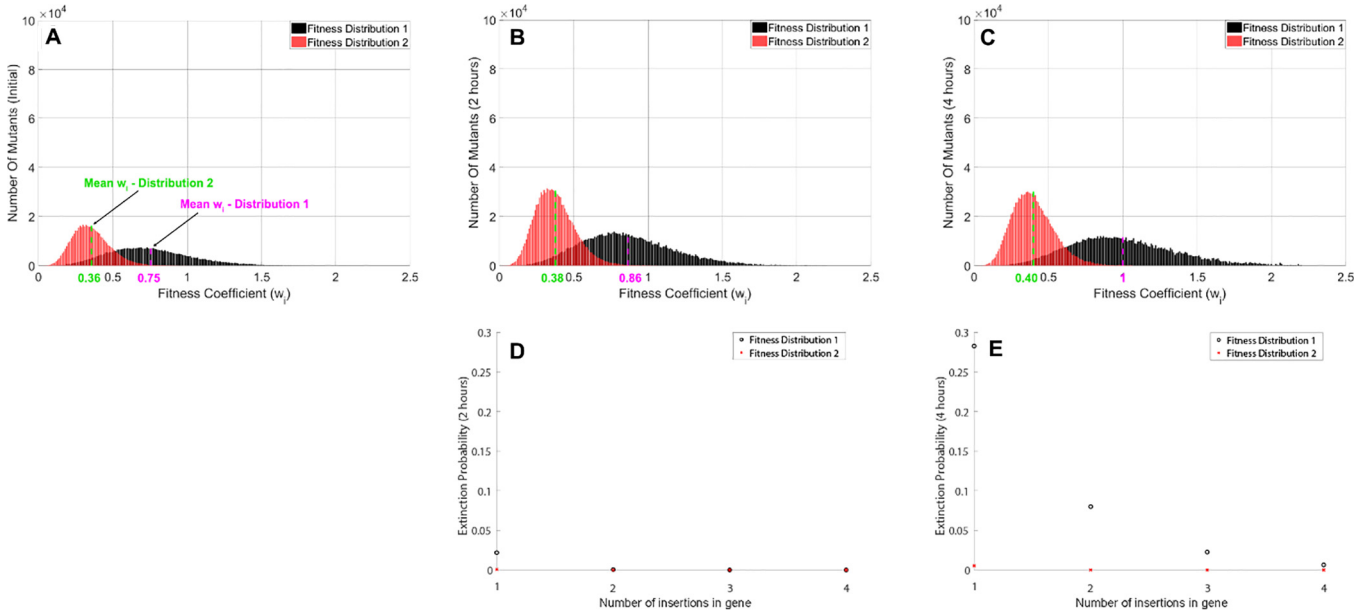

Fig. 5. The effects of bottlenecks on mutant extinction. (A) The fitness distributions (DFE) of two mutant populations with high average (black, distribution 1) and low average (red, distribution 2) fitness at the start of the birth-death process and before passage through a bottleneck. This figure is equivalent to Fig. 4A. The binned mean fitness is indicated in dashed magenta for distribution 1 and dashed green for distribution 2. (B) The distribution of the number of mutants over the fitness coefficients after a sample of 10^6 mutants following 2 h of growth. The distribution of the pre-bottleneck population is shown in Fig. 4B. (C) The distribution of the number of mutants over the fitness coefficients after a sample of 10^6 mutants following 4 h of growth. The distribution of the pre-bottleneck population is shown in Fig. 4C. (D) The extinction probability (y-axis) for all mutants corresponding to 1–4 insertion sites within a gene (x-axis) after sampling 10^6 mutants following 2 h of growth. Eq. (7) was used to calculate the extinction probabilities where w_i was either sampled from distribution 1 (black) or distribution 2 (red). All insertion sites within a gene have the same fitness cost where we show an example with w_i arbitrarily chosen and set to 0.15. (E) Same as (D) except the mutant cells have been growing for 4 h after which 10^6 mutant cells are sampled. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.9×10^{-5}, 1.6×10^{-7} and 8.6×10^{-10} for the second distribution. Therefore, growing the cells for two additional hours prior to sampling increases the extinction probability for one event by ~2 orders of magnitude where there is, on average, one extinction event in 1/0.2828 ≈ 3–4 repetitions of the experiment for the first distribution.

These numbers illustrate that even when two mutants from the two distributions have the same fitness, the extinction probability differs. The intuitive explanation is that mutants "compete" with others in the same library for a "spot" among the 10^6 cells that are let through the bottleneck. In our example, a mutant with the low fitness value of w_i = 0.15 will have a higher extinction probability and more extinctions over m sites for distribution 1 (higher average fitness) than for distribution 2 (lower average fitness). A large part of the differences in extinction probabilities can be explained by the bottleneck after four hours being relatively more restrictive than after two hours. Most cells for distribution 1 grow much faster than cells from distribution 2 because of their higher fitness. Therefore, a smaller fraction of distribution 1 is sampled when keeping the sampled cells constant at 10^6.

However, if the number of sampled cells is adjusted to the total bacterial population size, i.e. a fractional bottleneck, the extinction probabilities between a two hours and four hours sampling time point would be equal. This is because the mutant frequencies do not change that much, compared to the total population size which changes dramatically. Mathematically, this can be explained by looking at the exponent in Eq. (7): $-n_s\langle f_i\rangle$. For fractional bottlenecks, a constant fraction b of the total mutant population size N is sampled at both time points. Hence, the sample size is $n_s = b\langle N\rangle$. Moreover $\langle f_i\rangle \approx \langle n_i\rangle/\langle N\rangle$, such that $n_s\langle f_i\rangle = b\langle n_i\rangle$ and is independent of the total bacterial population size.

2.4. Comparison of effects of birth-death events and bottlenecks on library complexity

So far, we have quantified the effect of random birth-death processes (Section 2.2) and sampling bottlenecks (Section 2.3) via extinction probabilities q_i and the average number of mutant subpopulation extinctions over m sites, q_i^m. In order to better quantify the relative impact of extinctions due to random sampling events and random birth-death processes we aggregate the extinction probabilities into a single measure, the library complexity C. To do this, we use Eq. (6) where we set $m = Ck$ to get Z_{Ck} which is the average number of mutant extinctions over the transposon insertion sites with at least one mutant cell after mutagenesis due to stochastic fluctuations if Eq. (5) is used ($q_{i,growth}$) or random sampling events if Eq. (7) is used ($q_{i,bottleneck}$). Since Ck is the number of transposon insertion sites with at least one mutant, Ck-Z_{Ck} is the number of transposon insertion sites with at least one mutant after a random birth-death process or sampling. By dividing Ck-Z_{Ck} with k and applying the average operator, <>, we get the average proportion of insertion sites with at least one mutant after a sampling bottleneck or random birth-death process, $\langle C'\rangle$,

$$\langle C' \rangle = \langle C \rangle - \left\langle \frac{Z_{Ck}}{k} \right\rangle \tag{8}$$

where the average is taken with respect to repetitions of the experiment. In different words, mutants for Ck transposon insertion sites emerge from mutagenesis and subsequently undergo extinction events due to either a birth-death process or a random sampling event in a single repeat of the experiment. The implication is that only successful integrations at the very beginning of the experiment should be counted when taking the average of all steps to get the average reduced library complexity. In supplementary text S2 we take this into account and use the error propagation method to derive the average extinction probabilities $\langle q_{i,growth}\rangle$,

$$\langle q_{i,growth} \rangle = a_i^{\frac{\mu}{C}} + \frac{\mu \ln(a_i)^2}{2C}\left(1 + \mu - \frac{\mu}{C}\right)a_i^{\frac{\mu}{C}} \tag{9}$$

due to a birth-death process where a_i has been introduced for notational simplicity and is equal to the base of Eq. (5). For a fractional random sampling event, the average extinction probability reads,

$$\langle q_{i,bottleneck} \rangle = e^{-\frac{\mu b}{C}} + \mu\left(\frac{b}{Ck}\right)^2\left(1 + \mu - \frac{\mu}{C}\right)\left(C - \frac{1}{k}\right)e^{-\frac{\mu b}{C}} \tag{10}$$

The average extinction probabilities in Eqs. (9) and (10) are subsequently used to calculate $\langle Z_{Ck}\rangle$ (eq. (6)) and the average reduced library complexity, $\langle C'\rangle$ (eq. (8)).

Fig. 6 illustrates the library reduction due to stochastic fluctuations in a random birth-death process (Fig. 6A) and due to a random sampling event (Fig. 6B). The library complexity reduction due to stochastic fluctuations becomes negligible as μ becomes bigger than 10. However, the library complexity can become as low as 35% if on average, there is one mutant per insertion site (μ = 1), for which the initial library complexity is 63% according to Eq. (1). It is therefore advisable to have at least one order of magnitude more mutants after mutagenesis than the potential number of transposon insertion sites to minimize the chance of extinctions due to chance fluctuations caused by random birth-death events. In addition, the time the mutant population of cells is allowed to grow should be carefully evaluated. Depending on the underlying DFE, the baseline division rate, and death rate, the proportion of mutants with the lowest fitness cost will become overrepresented in the population. Sequencing will then act as a bottleneck and will select for the mutant cells present in the highest proportions, which could potentially lead to a significant reduction in library complexity. For example, if the number of mutants following mutagenesis is 10^6, which corresponds to a library complexity of 1 according to Eq. (1). By iteratively applying Eq. (8) to a birth-death step and then a sequencing bottleneck of size $n_s = 2 \times 10^7$ we get an average library complexity of 2.92%. The duration of the birth-death step was chosen to be 24 h, otherwise we used the same parameter values as in Fig. 6 and a gamma distributed DFE with shape parameter 10 and scale parameter 0.09. If the scale parameter is reduced to 0.04 to decrease the likelihood of introducing an advantageous mutant, the library complexity jumps to 90.58%. Hence, the library complexity can be very sensitive to the shape of the underlying DFE, where the sensitivity is proportional to the baseline net growth rate and the time that the mutant cells spend in exponential growth phase. If the growth time is reduced from 24 h to 12 h, the library complexity becomes 78.82% for scale parameter 0.09 and 97.44% for scale parameter 0.04. Therefore, it is advisable to minimize the time the cells spend in exponential growth and maximize the number of mutants from mutagenesis when creating a transposon insertion library.

Fig. 6B illustrates the reduction in library complexity due to random sampling events for two bottleneck sizes, 1% and 10%. Bottleneck effects could potentially have a significant impact on the reduction in the library complexity depending on the bottleneck size and the number of mutant cells. In particular, bottleneck effects become more severe when preceded by a birth-death process as discussed above. Even at very large population sizes, the library complexity is severely reduced when a sufficiently stringent bottleneck is imposed. Importantly, the effects of bottlenecks are much more severe than those of random extinctions during growth for a population of mutant cells with $w_i > 0$ and $r > 0$.

2.5. Bottlenecks are a major source of false positives during Tn-seq experiments

The main concern with fitness-independent disappearance of mutants due to bottlenecks is that genes that are extinct at the

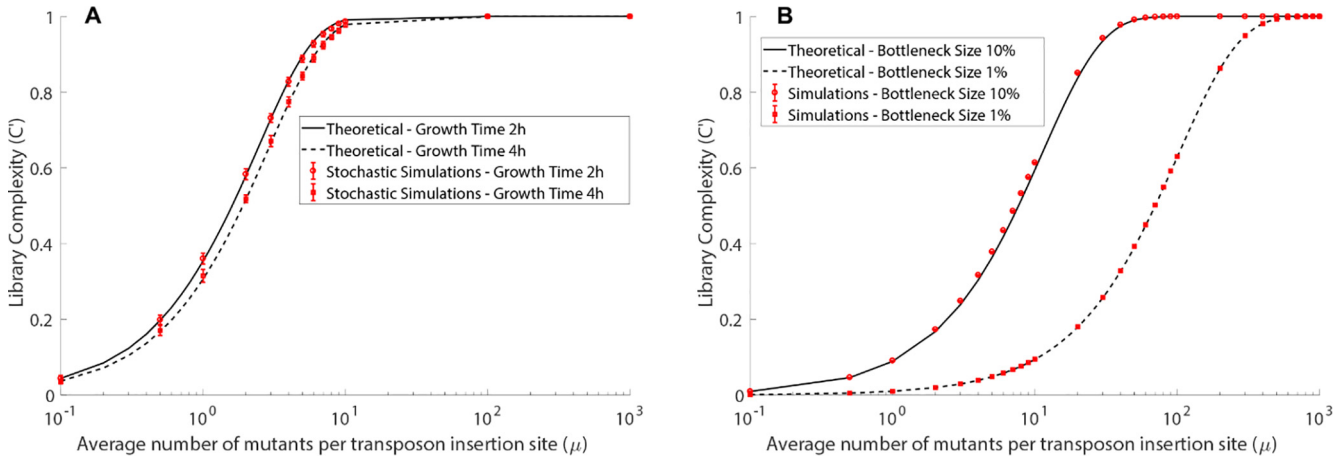

Fig. 6. Library complexity reduction due to random birth-death events and random sampling events. The initial mutant population emerges from the process of mutagenesis with an average number of mutant cells per potential transposon insertion site μ (x-axis) related to the initial library complexity *C* through Eq. (1). (A) The mutant population grows on average with a baseline division rate β set to 0.03 min^{-1} and a baseline death rate δ set to 0.02 min^{-1} with fitness coefficients sampled from distribution 1 (shape = 10, scale = 0.09). The mutant subpopulations have a chance to go extinct as a consequence of stochastic fluctuations shown as a reduction in library complexity, *C'* (y-axis). The theoretical results were calculated using Eq. (9). The black solid line shows the library complexity reduction due to stochastic fluctuations after 2 h of growth and the black dashed line after 4 h of growth. Stochastic tau-leaping simulations were ran for 20 iterations for each value of μ where for each iteration the number of cells for site $i = 1,2,\ldots,10^3$ was drawn from a Poisson distribution with mean μ. The mean and the standard error in the library complexity was subsequently calculated and plotted as red circles for cells growing for 2 h and red squares for cells growing for 4 h. (B) Library complexity reduction after sampling 10% (black solid line) and 1% (black dashed line) of the initial mutant population that emerges from the process of mutagenesis. The theoretical results were calculated using Eq. (10). Multinomial random sampling simulations were ran for 20 iterations where for each iteration the number of cells for site $i = 1,2,\ldots,10^5$ was drawn from a Poisson distribution with mean μ. The mean and the standard error in the library complexity was subsequently calculated and plotted as red circles and red squares for the 10% bottleneck case and the 1% bottleneck case, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

end of the experiment are classified as essential even though they may not be. Most mutants with a random mutation will have reduced fitness [53]. The aim of a Tn-seq experiment is to identify those that have a very large effect on fitness, ideally significantly larger than an average random mutant. In this section, we investigate how many mutants with fitness larger than zero, i.e. that would be able to grow under the selected conditions, go extinct by chance and how the fitness of these mutations is distributed.

Since we have seen in Fig. 6 that birth-death processes generally are not the main source of fitness-independent extinctions, we now simplify our approach and model the growth of individual mutants *i* as deterministic exponential growth (deterministic version of Eq. (2)). When we substitute n_i in Eq. (7) with Eq. (2), we obtain an extinction probability q_i

$$q_i = e^{-bn_i(0)e^{w_i rt}} \tag{11}$$

with *b* as the bottleneck size, $n_i(0)$ as the initial post-mutagenesis number of mutants *i*, w_i as the fitness coefficient, *r* as the net growth rate, and *t* as the time. If we now do not only look at the extinction of a specific insertion site *i*, but require all insertion sites *m* being extinct, the extinction probability of an entire gene, q_i^m becomes

$$q_i^m = e^{-mbn_i(0)e^{w_i rt}} \tag{12}$$

where we assume that the fitness cost of all transposon insertion within a gene are equal. Note that requiring all insertion sites per gene being hit is mathematically the same as focusing on one individual insertion site, repeating the experiment *m* times and always observing extinction.

Eq. (12) describes an approach that is often used to assign essentiality: Quantifying the probability of random extinction. Many approaches exist to do so, and many use more sophisticated frameworks [16,41]. However, the prior fitness distribution of random mutations (DFE) is neglected. We illustrate the impact of the prior distribution in Fig. 7, where we derive the fitness distribution and size of the population of extinct mutants.

Because completely non-viable mutants would not contribute to the population, we focus only on mutants with a fitness larger than zero, i.e. we focus on "false positives" that would be classified as "essential" because of their absence in the sequencing data despite the fact that they can still grow under the experimental conditions. We furthermore assume that all insertion sites were hit ($C = 1$) and that $n_i(0)$ is uniformly distributed, i.e. the same for all mutants *i*. We also assume that the number of insertion sites that are grouped and go extinct simultaneously (e.g. per gene), *m*, is the same for all genes. This leads to conservative estimates that underreport extinction, which would be more likely if some mutants were only present infrequently or some genes had only few insertion sites (e.g. short genes). Under these assumptions, we can multiply the gamma-distributed DFE (Fig. 7A) with Eq. (12) (Fig. 7B and D) to obtain the distribution of fitness values for genes that randomly disappeared (Fig. 7C and E). This shows that even though the extinction probability per site is highest for mutants that have very low fitness, the majority of mutants that randomly disappear actually have intermediate fitness. This is because mutants with intermediate fitness were more frequent in the original population after mutagenesis. The percentage of entire genes that disappear randomly depends strongly on the bottleneck size and on the number of insertion sites per gene (compare Fig. 7BC to DE). Other factors that influence the percentage of "false positive" genes are the prior fitness distribution (Fig. S2) and the number of mutants per site *i* at the beginning of the experiment, $n_i(0)$ (Fig. S2). In our example, depending on the number of insertion sites per locus, 0.8%-10% of all genes would be misclassified as essential after a 0.01% bottleneck which was reported for *V. cholerae* infection models [26]. On an *E. coli* or *V. cholerae* genome scale (with approx. 4000 genes), this means 32–400 false positive hits. If a comprehensive answer to which genes are essential was the goal, the repletion would have to increase until the expected number of false positives falls below one. In this case, one would have to repeat a gene-wide analysis (approx. 30 insertion sites per gene, [16]) for eight experimental replicates, thereby bringing the number of repetitions to $30 \times 8 = 240$ (Fig. S2). We additionally

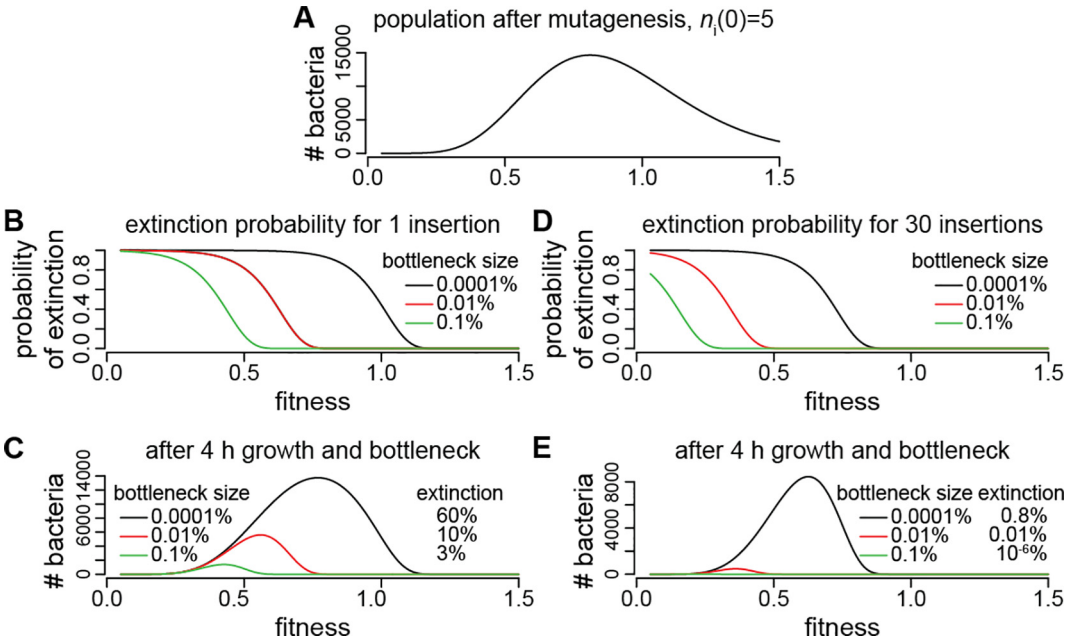

Fig. 7. Bottlenecks are a major source of false positives during Tn-seq experiments. (A) Distribution of fitness coefficients (DFE) for a library of random mutants. The DFE is the same as distribution 1 in Figs. 4 & 5, i.e. gamma-distributed with shape parameter 10 and scale parameter 0.09. Here, we scaled this distribution to a population size of 10^6 cells after mutagenesis and assume five mutants per transposon insertion site at the beginning of the experiment. (B) Extinction probabilities (y-axis) of a single insertion site depending on the fitness coefficient (x-axis) and different bottleneck sizes (0.0001% black, 0.01% red, 0.1% green) as predicted by Eq. (12). We assume that bacteria grew for 4 h with a doubling time of 20 min ($r = \ln(2)/20\ \text{min}^{-1}$). (C) Fitness distribution of the population of mutants that went extinct, i.e. had zero sequence reads at the end of the experiment. This graph was obtained by multiplying the gamma distribution in (A) with the extinction probability in (B). (B) and (C) focus on a single transposon insertion site. The percentage of loci going extinct for a given bottleneck is specified in the graph. (D) Same as (B) for the simultaneous extinction of 30 insertion sites with the same fitness, i.e. located in the same gene. (E) Same as (C), also for 30 insertion sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analyzed the number of false positives in our full stochastic model to ascertain that our main conclusions hold true under more realistic scenarios (Fig. S3). Again, we find that the majority of false positive mutants that go extinct have intermediate fitness costs.

3. Discussion

Tn-seq is a powerful method to identify genes that are important for bacterial viability in a given environment. It has been successfully used to identify genes that are critical for specific processes and determine important phenotypes including virulence [1–9], persistence [12,13], biofilm formation [14] and antibiotic resistance [15,16]. Since Tn-seq is a global genetic screening technique, it allows in principle to test all genes in a single experiment simultaneously. In reality, several aspects of the experimental setup are limiting and repetitions of very similar (if not identical) Tn-seq experiments by different laboratories have poor overlap [25].

From a population biological perspective, this may be explained by the fact that drift, i.e. random extinction, plays a larger role than anticipated. All steps in the experimental process have their limitations and may obscure the correlation between sequence reads and mutant fitness that underlies all Tn-seq experiments. For example, it is difficult to ascertain that a transposon library is complex enough that all loci that are to be investigated are sufficiently covered. Random events during growth and passage through bottlenecks distort mutant frequencies and thereby add noise. In addition, the analysis is complicated by the fact that most mutants in any bacterial genome will result in a reduced fitness of the organism [53]. All this makes it difficult to achieve the goal of a Tn-seq experiment: to identify loci in which mutation leads to severe fitness defects, which are significantly larger than the mild effects of an average random mutation. Several analysis methods have approached these limitations from different angles. Some studies employ successive sequencing to track mutant populations over time [32,41], implicitly or explicitly modelling bacterial growth. Also the effect of bottlenecks have been modeled by multinomial sampling [25]. However, a concise population biological framework encompassing all steps is still missing. Most importantly, while the effects of random mutagenesis on the distribution of fitness in a population (DFE) have been a matter of intensive research [42,53,57,58], this underlying or prior distribution at the very start of a Tn-seq experiment has received little attention.

In this paper, we investigate the population biology of a Tn-seq experiment, i.e. the changes in size and composition of a population of mutants created by transposon mutagenesis. Specifically, we investigate the population biological mechanisms that underlie experimental noise. We illustrate the effects of experimental noise and focus on a very conservative case, when mutants have zero reads at the end of the experiment despite being able to grow in the given environment. If "essential gene" is taken to mean that a cell cannot grow at all, i.e. the mutation is lethal, we focus on the dynamics of false positives. We take a two-pronged approach to investigate the population biology of a Tn-seq library: we develop both an analytical approach using basic population dynamical and probability theory concepts and a framework based on stochastic tau-leaping simulations. Both frameworks encompass the following steps of a Tn-seq experimental workflow: mutagenesis, a random birth-death process and random sampling events or bottlenecks (including sequencing). This yields the size and distribution of fitness values in those bacteria that go extinct (false-positives). To use the model developed in this work to fully describe a Tn-seq experiment, one requires careful measurements of parameters that influence library complexity. Thus, growth times (in exponential phase), OD values, the sampled volume sizes and the volume sizes of the media from which the cells are sam-

pled, and DNA concentrations before and after each step during which DNA can be lost (shearing, adaptor ligation, ...) must be carefully determined. In addition, one would need an estimated shape of the DFE and an estimate of the division rate and death rate of wild-type cells. In addition, care should be taken in how to sample fitness coefficients from the DFE. It is reasonable to assume that, in general, the fitness cost of insertions within the same gene are similar. Given this assumption, the DFE could be sampled on a gene level. Then the sampled fitness coefficients would need to be repeated a number of times corresponding to the number of potential Tn insertion sites in the genes. Taking this approach of integrating the model developed in this work with a specific Tn-seq protocol can aid in minimizing the discrepancies of current statistical methods, which do not fully incorporate a quantitative understanding of the process that generates the sequence reads in the first place.

For the mutagenesis step, we find a simple relationship between the library complexity (the fraction of transposon insertion sites with at least one mutant), the total number of mutant cells after mutagenesis, and the potential number of transposon insertion sites. Our approach can be used to plan experiments such that a sufficient library complexity can be ensured. In addition, it serves as a null hypothesis regarding transposon distribution that can be used to identify preferential transposon insertion regions, i.e. genomic cold/hot spots.

When modelling bacterial growth after mutagenesis with a birth-death process (a well-established method in population biology [54]), we find that fitter mutants are enriched and that the mean fitness of the population shifts towards higher values. However, we also find that this is not due to pervasive extinction of mutants (which is rare during the birth-death process in our model), but due to substantial growth where the fitter mutants grow faster. While we found little extinction and therefore little contribution to false-positives while bacteria grow, bottlenecks by definition lead to "mass extinctions". The random disappearance of many mutants makes it very difficult to find out which mutants went extinct by chance and which loci are truly essential. With very stringent bottlenecks, the fitness distribution of the extinct mutants will be exactly the same as the one right after mutagenesis (minus the few cells that survived). Random extinction due to bottlenecks is especially problematic if genes have few insertion sites (e.g., short genes) or if subgenomic locations that only encompass few insertion sites are investigated.

Random mutations such as those created by transposon mutagenesis [53] are usually deleterious, with only few being advantageous. The distribution of fitness effects (DFE) peaks at mildly deleterious fitness values with a long tail to strongly deleterious mutations and a sharp decline after mildly beneficial mutations. As stated above, the aim of a Tn-seq experiment is to identify those loci that are critical, i.e. a mutation in this loci is strongly detrimental. To assess in how far this can be achieved in the presence of bottlenecks, we investigate the distribution of fitness values in the population that goes extinct during the experiment. Importantly, we assess an idealized situation where the library complexity is perfect and no extinctions happen during the birth-death process. In this case, the extinction probability depending on mutant fitness can be described with Eq. (11). This probability also depends on the time bacteria grow, their doubling time, the bottleneck stringency, how many insertion sites are grouped together (e.g. located in one gene), and the number of individual mutants per insertion site after mutagenesis. When multiplying the number of loci expected with a certain fitness (from the DFE) with the extinction probability, we obtain both the size and the distribution of fitness values of the extinct mutants. Thus, this equation can serve as a tool to assess false positive rates and the mean fitness of false positives after a bottleneck.

We find that permissive bottlenecks (0.1%, e.g. sequencing 10^6 reads from 1 ml bacterial culture with 10^9 bacteria, i.e. OD ~1) are only problematic when insertion sites are investigated individually and not grouped to larger loci. If we only count genes as "essential" when not a single insertion site in that gene was hit, the false positive rate for an average gene with 30 insertion sites starts to rise when only approx. 0.01% of the original population survives a bottleneck (such as in rabbit models of cholera infections [26]). For such a stringent bottleneck, one would have to repeat the experiment 8 times to achieve less than 1 false positive, thereby achieving a comprehensive, genome-wide list of essential loci. In extreme cases such as e.g. the colonization by listeria [27], there are so few cells that survive the host defenses, that the absence or presence of all other mutants is not conclusive of essentiality or even low fitness. For an average gene with 30 insertion sites, we would expect 30% of all genes to go extinct, i.e. recorded as false positives. These would have a mean fitness close to the mean fitness of a random mutant.

While this work incorporates explicit population biological models of many steps of Tn-seq experiments, we had to make several simplifying assumptions, frequently because of a lack of experimental data. For example, it is conceivable that a mutation affects bacterial division, but not bacterial death. However, since only the net growth rate is routinely measured in batch culture and the rates of bacterial division and death are largely unknown, we here assume that the fitness coefficient scales the division and death rates equally. This has no effect on population turnover when the baseline death rate is 0 (as often assumed in population biological models, [52]). If we however assume there is bacterial death, lower fitness would result both in a lower growth rate and in lower bacterial death, thereby underestimating population turnover and extinction. We also assume for simplicity that bacteria replicate exponentially. Our assumption is only true as long as the culture density is far away from carrying capacity, such that our model is best suited to describe experiments that start with an abundance of nutrients. However, the assumption is helpful because resource limitations lead to competition between mutants of different fitnesses, which in turn leads to complex dynamics that would again depend on the unknown bacterial division and death rates [59]. Finally, the DFEs of bacteria are increasingly investigated, but not well-characterized for different strains or species. This work therefore calls for more determinations of DFEs, better measurements of bacterial division and death rates or supplementing Tn-seq with e.g. the plasmid segregation method [52] or single cell microscopy [51]. Furthermore, bottlenecks during the experimental setup need to be assessed and ideally measured [56].

As stated above, we are taking a relatively conservative approach to assess how many viable mutants would go extinct by chance. While we find that typical sampling during experiments should not be very problematic, our results do show that most mutants that are lost in the presence of strong bottlenecks have average fitness values. This is despite the probability of extinction being dependent on mutant fitness. This probability is not the sole determinant of the DFE of extinct mutants, but the DFE of the population after mutagenesis has to be taken into account by multiplying probabilities. If there are few mutants with low fitness to being with, few will be lost. Among the many mutants with intermediate fitness, some will be lost even though the chance of extinction per mutant is low. Thus, the previous disregard of the prior DFE and the conditional probabilities of extinction may shed light on the problem of reproducibility of Tn-seq experiments, especially in animal models.

This reasoning is very similar to Bayesian reasoning in interpreting diagnostic tests. A famous example would be a diagnostic test with 95% accuracy taken by a person who is from a population with a low prevalence (0.1%) of the disease in question [60]. The

probability that a randomly selected person who tests positive then actually has the disease is not 95%, but 2%, necessitating further tests [60]. Therefore, the fitness of all mutants extinct in Tn-seq experiments with animal models or other systems with strong bottlenecks should be tested individually against the wild-type, as for example in [5,9]. The same conclusions can be drawn for any kind of screen that relies on random mutagenesis, such as CRISPR interference (CRISPRi) in eukaryotes [61].

4. Methods

Codes for (Figs. 3, 7) and (Fig. S2) were implemented in statistical software package R (version 3.4.4, The R Foundation for Statistical Computing, Vienna, Austria). Codes for all other figures were implemented in MATLAB (R2017b, The MathWorks, Natick, MA, USA). The stochastic tau-leaping simulation results (Figs. 6 and S3) were generated using StochKit2 [62] with a time step of 0.01 with additional code in Matlab to create an input file to StochKit2. All code for reproducing the figures and the Matlab code for generating the input file to StochKit2 for running the stochastic simulations are available as supplementary data.

All scripts can be accessed via Mendeley Data under the terms of the Creative Commons (CC) license CC by NC 3.0.

CRediT authorship contribution statement

Anel Mahmutovic: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Pia Abel zur Wiesch:** Conceptualization, Formal analysis, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Sören Abel:** Conceptualization, Formal analysis, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by Research Council of Norway (NFR) Grant 262686 (to P.AzW.) and 249979 (to S.A.), and Helse-Nord Grant 14796 (to S.A.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.03.021.

References

[1] Karlinsey JE, Stepien TA, Mayho M, Singletary LA, Bingham-Ramos LK, Brehm MA, et al. Genome-wide analysis of Salmonella enterica serovar typhi in humanized mice reveals key virulence features. Cell Host Microbe 2019;26(3):426–434.e6. https://doi.org/10.1016/j.chom.2019.08.001.

[2] Cowley LA, Low AS, Pickard D, Boinett CJ, Dallman TJ, Day M, et al. Transposon insertion sequencing elucidates novel gene involvement in susceptibility and resistance to phages T4 and T7 in Escherichia coli O157. MBio. 2018;9(4). Epub 2018/07/26. doi: 10.1128/mBio.00705-18. PubMed PMID: 30042196; PubMed Central PMCID: PMCPMC6058288.

[3] Gao B, Vorwerk H, Huber C, Lara-Tejero M, Mohr J, Goodman AL, et al. Metabolic and fitness determinants for in vitro growth and intestinal colonization of the bacterial pathogen Campylobacter jejuni. PLoS Biol. 2017;15(5):e2001390. Epub 2017/05/26. doi: 10.1371/journal.pbio.2001390. PubMed PMID: 28542173; PubMed Central PMCID: PMCPMC5438104.

[4] Fulton BO, Sachs D, Schwarz MC, Palese P, Evans MJ. Transposon Mutagenesis of the Zika Virus Genome Highlights Regions Essential for RNA Replication and Restricted for Immune Evasion. J Virol. 2017;91(15). Epub 2017/05/19. doi: 10.1128/JVI.00698-17. PubMed PMID: 28515302; PubMed Central PMCID: PMCPMC5512254.

[5] Hubbard TP, Chao MC, Abel S, Blondel CJ, Abel Zur Wiesch P, Zhou X, et al. Genetic analysis of Vibrio parahaemolyticus intestinal colonization. Proc Natl Acad Sci USA. 2016;113(22):6283–8. Epub 2016/05/18. doi: 10.1073/pnas.1601718113. PubMed PMID: 27185914; PubMed Central PMCID: PMCPMC4896720.

[6] Bachman MA, Breen P, Deornellas V, Mu Q, Zhao L, Wu W, et al. Genome-Wide Identification of Klebsiella pneumoniae Fitness Genes during Lung Infection. MBio. 2015;6(3):e00775. Epub 2015/06/11. doi: 10.1128/mBio.00775-15. PubMed PMID: 26060277; PubMed Central PMCID: PMCPMC4462621.

[7] Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, et al. Comparative analysis of Klebsiella pneumoniae genomes identifies a phospholipase D family protein as a novel virulence factor. BMC Biol. 2014;12:41. Epub 2014/06/03. doi: 10.1186/1741-7007-12-41. PubMed PMID: 24885329; PubMed Central PMCID: PMCPMC4068068.

[8] Fu Y, Waldor MK, Mekalanos JJ. Tn-Seq analysis of Vibrio cholerae intestinal colonization reveals a role for T6SS-mediated antibacterial activity in the host. Cell Host Microbe. 2013;14(6):652–63. Epub 2013/12/18. doi: 10.1016/j.chom.2013.11.001. PubMed PMID: 24331463; PubMed Central PMCID: PMCPMC3951154.

[9] Warr AR, Hubbard TP, Munera D, Blondel CJ, Abel Zur Wiesch P, Abel S, et al. Transposon-insertion sequencing screens unveil requirements for EHEC growth and intestinal colonization. PLoS Pathog. 2019;15(8):e1007652. Epub 2019/08/14. doi: 10.1371/journal.ppat.1007652. PubMed PMID: 31404118; PubMed Central PMCID: PMCPMC6705877.

[10] Takeda H, Wei Z, Koso H, Rust AG, Yew CC, Mann MB, et al. Transposon mutagenesis identifies genes and evolutionary forces driving gastrointestinal tract tumor progression. Nat Genet 2015;47(2):142–50. https://doi.org/10.1038/ng.3175. PubMed PMID: 25559195.

[11] Cameron DR, Shan Y, Zalis EA, Isabella V, Lewis K. A Genetic Determinant of Persister Cell Formation in Bacterial Pathogens. J Bacteriol. 2018;200(17). Epub 2018/06/27. doi: 10.1128/JB.00303-18. PubMed PMID: 29941425; PubMed Central PMCID: PMCPMC6088157.

[12] Shan Y, Lazinski D, Rowe S, Camilli A, Lewis K, Bush K. Genetic basis of persister tolerance to aminoglycosides in Escherichia coli. mBio 2015;6(2). https://doi.org/10.1128/mBio.00078-15. PubMed PMID: 25852159; PubMed Central PMCID: PMCPMC4453570.

[13] Yan F, Yu Y, Gozzi K, Chen Y, Guo JH, Chai Y, Schottel JL. Genome-wide investigation of biofilm formation in Bacillus cereus. Appl. Environ. Microbiol. 2017;83(13). https://doi.org/10.1128/AEM.00561-17. PubMed PMID: 28432092; PubMed Central PMCID: PMCPMC5478996.

[14] Dorr T, Delgado F, Umans BD, Gerding MA, Davis BM, Waldor MK. A Transposon Screen Identifies Genetic Determinants of Vibrio cholerae Resistance to High-Molecular-Weight Antibiotics. Antimicrob Agents Chemother. 2016;60(8):4757–63. Epub 2016/05/25. doi: 10.1128/AAC.00576-16. PubMed PMID: 27216069; PubMed Central PMCID: PMCPMC4958186.

[15] Roux D, Danilchanka O, Guillard T, Cattoir V, Aschard H, Fu Y, et al. Fitness cost of antibiotic susceptibility during bacterial infection. Sci Transl Med. 2015;7(297):297ra114. Epub 2015/07/24. doi: 10.1126/scitranslmed.aab1621. PubMed PMID: 26203082.

[16] Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nat Rev Microbiol 2016;14(2):119–28. https://doi.org/10.1038/nrmicro.2015.7. PubMed PMID: 26775926; PubMed Central PMCID: PMCPMC5099075.

[17] Kwon YM, Ricke SC, Mandal RK. Transposon sequencing: methods and expanding applications. Appl Microbiol Biotechnol 2016;100(1):31–43. https://doi.org/10.1007/s00253-015-7037-8. PubMed PMID: 26476650.

[18] Shields RC, Jensen PA. The bare necessities: uncovering essential and condition-critical genes with transposon sequencing. Mol Oral Microbiol 2019;34(2):39–50. https://doi.org/10.1111/omi.2019.34.issue-2. PubMed PMID: 30739386.

[19] van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 2009;6(10):767–72. https://doi.org/10.1038/nmeth.1377. PubMed PMID: 19767758; PubMed Central PMCID: PMCPMC2957483.

[20] Gallagher LA, Shendure J, Manoil C, Mekalanos J. Genome-scale identification of resistance functions in pseudomonas aeruginosa using Tn-seq. mBio 2011;2(1). https://doi.org/10.1128/mBio.00315-10. PubMed PMID: 21253457; PubMed Central PMCID: PMCPMC3023915.

[21] Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, Knight R, Gordon JI. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 2009;6(3):279–89. https://doi.org/10.1016/j.chom.2009.08.003. PubMed PMID: 19748469; PubMed Central PMCID: PMCPMC2895552.

[22] Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res 2009;19(12):2308–16. https://doi.org/10.1101/gr.097097.109. PubMed PMID: 19826075; PubMed Central PMCID: PMCPMC2792183.

[23] Gawronski JD, Wong SMS, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proc Natl Acad Sci 2009;106(38):16422–7. https://doi.org/10.1073/pnas.0906627106. PubMed PMID: 19805314; PubMed Central PMCID: PMCPMC2752563.

[24] Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, Papin JA. Reconstruction of the metabolic network of Pseudomonas aeruginosa to interrogate virulence factor synthesis. Nat Commun 2017;8(1). https://doi.org/10.1038/ncomms14631. PubMed PMID: 28266498; PubMed Central PMCID: PMCPMC5344303.

[25] Pritchard JR, Chao MC, Abel S, Davis BM, Baranowski C, Zhang YJ, et al. ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. Plos Genet. 2014;10(11):e1004782. Epub 2014/11/07. doi: 10.1371/journal.pgen.1004782. PubMed PMID: 25375795; PubMed Central PMCID: PMCPMC4222735.

[26] Abel S, Abel zur Wiesch P, Chang HH, Davis BM, Lipsitch M, Waldor MK. Sequence tag-based analysis of microbial population dynamics. Nat Methods. 2015;12(3):223-6, 3 p following 6. Epub 2015/01/20. doi: 10.1038/nmeth.3253. PubMed PMID: 25599549; PubMed Central PMCID: PMCPMC4344388.

[27] Zhang T, Abel S, Abel Zur Wiesch P, Sasabe J, Davis BM, Higgins DE, et al. Deciphering the landscape of host barriers to Listeria monocytogenes infection. Proc Natl Acad Sci U S A. 2017;114(24):6334–9. Epub 2017/06/01. doi: 10.1073/pnas.1702077114. PubMed PMID: 28559314; PubMed Central PMCID: PMCPMC5474794.

[28] Li K, Chen R, Lindsey W, Best A, DeJongh M, Henry C, et al. Implementing and evaluating a Gaussian mixture framework for identifying gene function from TnSeq data. Pac Symp Biocomput 2019;24:172–83. Epub 2019/03/14 PubMed PMID: 30864320.

[29] Hubbard TP, D'Gama JD, Billings G, Davis BM, Waldor MK. Unsupervised Learning Approach for Comparing Multiple Transposon Insertion Sequencing Studies. mSphere. 2019;4(1). Epub 2019/02/23. doi: 10.1128/mSphere.00031-19. PubMed PMID: 30787116; PubMed Central PMCID: PMCPMC6382967.

[30] Segal ES, Gritsenko V, Levitan A, Yadav B, Dror N, Steenwyk JL, et al. Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a Stable Haploid Isolate of Candida albicans. MBio. 2018;9(5). Epub 2018/11/01. doi: 10.1128/mBio.02048-18. PubMed PMID: 30377286; PubMed Central PMCID: PMCPMC6212825.

[31] Zhao L, Anderson MT, Wu W, Mobley HLT, Bachman MA. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinf 2017;18(1). https://doi.org/10.1186/s12859-017-1745-2. PubMed PMID: 28683752; PubMed Central PMCID: PMCPMC5500955.

[32] Yang G, Billings G, Hubbard TP, Park JS, Yin Leung K, Liu Q, et al. Time-resolved transposon insertion sequencing reveals genome-wide fitness dynamics during infection. mBio 2017;8(5). https://doi.org/10.1128/mBio.01581-17. PubMed PMID: 28974620; PubMed Central PMCID: PMCPMC5626973.

[33] DeJesus MA, Nambi S, Smith CM, Baker RE, Sassetti CM, Ioerger TR. Statistical analysis of genetic interactions in Tn-Seq data. Nucleic Acids Res. 2017;45(11):e93. Epub 2017/03/24. doi: 10.1093/nar/gkx128. PubMed PMID: 28334803; PubMed Central PMCID: PMCPMC5499643.

[34] Liu F, Wang C, Wu Z, Zhang Q, Liu P. A zero-inflated Poisson model for insertion tolerance analysis of genes based on Tn-seq data. Bioinformatics 2016;32(11):1701–8. https://doi.org/10.1093/bioinformatics/btw061. PubMed PMID: 26833344.

[35] DeJesus MA, Ioerger TR. Capturing uncertainty by modeling local transposon insertion frequencies improves discrimination of essential genes. IEEE/ACM Trans. Comput. Biol. and Bioinf. 2015;12(1):92–102. https://doi.org/10.1109/TCBB.2014.2326857. PubMed PMID: 26357081.

[36] Deng J, Su S, Lin X, Hassett DJ, Lu LJ. A statistical framework for improving genomic annotations of prokaryotic essential genes. PLoS One. 2013;8(3):e58178. Epub 2013/03/23. doi: 10.1371/journal.pone.0058178. PubMed PMID: 23520492; PubMed Central PMCID: PMCPMC3592911.

[37] DeJesus MA, Zhang YJ, Sassetti CM, Rubin EJ, Sacchettini JC, Ioerger TR. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. Bioinformatics. 2013;29(6):695–703. Epub 2013/01/31. doi: 10.1093/bioinformatics/btt043. PubMed PMID: 23361328; PubMed Central PMCID: PMCPMC3597147.

[38] DeJesus MA, Ioerger TR. A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinf 2013;14(1). https://doi.org/10.1186/1471-2105-14-303. PubMed PMID: 24103077; PubMed Central PMCID: PMCPMC3854130.

[39] Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell 2016;165(6):1493–506. https://doi.org/10.1016/j.cell.2016.05.003. PubMed PMID: 27238023; PubMed Central PMCID: PMCPMC4894308.

[40] Liu X, Gallay C, Kjos M, Domenech A, Slager J, van Kessel SP, et al. High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. Mol Syst Biol. 2017;13(5):931. Epub 2017/05/12. doi: 10.15252/msb.20167449. PubMed PMID: 28490437; PubMed Central PMCID: PMCPMC5448163.

[41] van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 2013;11(7):435–42. https://doi.org/10.1038/nrmicro3033. PubMed PMID: 23712350; PubMed Central PMCID: PMCPMC3842022.

[42] Robert L, Ollion J, Robert J, Song X, Matic I, Elez M. Mutation dynamics and fitness effects followed in single cells. Science 2018;359(6381):1283–6. https://doi.org/10.1126/science.aan0797. PubMed PMID: 29590079.

[43] Kimura S, Hubbard TP, Davis BM, Waldor MK. The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. mBio 2016;7(4). https://doi.org/10.1128/mBio.01351-16. PubMed PMID: 27578758; PubMed Central PMCID: PMCPMC4999555.

[44] Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, et al. Defining the core essential genome of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A. 2019;116(20):10072-80. Epub 2019/05/01. doi: 10.1073/pnas.1900570116. PubMed PMID: 31036669; PubMed Central PMCID: PMCPMC6525520.

[45] Zhu J, Gong R, Zhu Q, He Q, Xu N, Xu Y, et al. Genome-wide determination of gene essentiality by transposon insertion sequencing in yeast Pichia pastoris. Sci Rep 2018;8(1). https://doi.org/10.1038/s41598-018-28217-z.

[46] Willcocks SJ, Stabler RA, Atkins HS, Oyston PF, Wren BW. High-throughput analysis of Yersinia pseudotuberculosis gene essentiality in optimised in vitro conditions, and implications for the speciation of Yersinia pestis. BMC Microbiol 2018;18(1). https://doi.org/10.1186/s12866-018-1189-5. PubMed PMID: 29855259; PubMed Central PMCID: PMCPMC5984423.

[47] Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, et al. The essential genome of Escherichia coli K-12. mBio 2018;9(1). https://doi.org/10.1128/mBio.02096-17. PubMed PMID: 29463657; PubMed Central PMCID: PMCPMC5821084.

[48] DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. MBio. 2017;8(1). Epub 2017/01/18. doi: 10.1128/mBio.02133-16. PubMed PMID: 28096490; PubMed Central PMCID: PMCPMC5241402.

[49] De Y, Dong C, Cao Y, Wang X, Yang X, Wang N, et al. Genome-wide sequence transposon insertion sites and analyze the essential genes of Brucella melitensis. Microb Pathog. 2017;112:97–102. Epub 2017/09/11. 10.1016/j.micpath.2017.09.005 PubMed PMID: 28888882.

[50] Burger BT, Imam S, Scarborough MJ, Noguera DR, Donohue TJ. Combining Genome-Scale Experimental and Computational Methods To Identify Essential Genes in Rhodobacter sphaeroides. mSystems. 2017;2(3). Epub 2017/07/27. doi: 10.1128/mSystems.00015-17. PubMed PMID: 28744485; PubMed Central PMCID: PMCPMC5513736.

[51] Ocampo PS, Lazar V, Papp B, Arnoldini M, zur Wiesch PA, Busa-Fekete R, et al. Antagonism between Bacteriostatic and Bactericidal Antibiotics Is Prevalent. Antimicrob Agents Ch. 2014;58(8):4573-82. doi: 10.1128/Aac.02463-14. PubMed PMID: WOS:000339259200038.

[52] Frenoy A, Bonhoeffer S. Death and population dynamics affect mutation rate estimates and evolvability under stress in bacteria. PLoS Biol. 2018;16(5):e2005056. Epub 2018/05/12. doi: 10.1371/journal.pbio.2005056. PubMed PMID: 29750784; PubMed Central PMCID: PMCPMC5966242.

[53] Elena SF, Ekunwe L, Hajela N, Oden SA, Lenski RE. Distribution of fitness effects caused by random insertion mutations in Escherichia coli. Genetica 1998;102–103(1–6):349–58. Epub 1998/08/28 PubMed PMID: 9720287.

[54] Crawford FW, Ho LST, Suchard MA. Computational methods for birth-death processes. WIREs Comput Stat 2018;10(2):e1423. https://doi.org/10.1002/wics.2018.10.issue-2.10.1002/wics.1423. PubMed PMID: 29942419; PubMed Central PMCID: PMCPMC6014701.

[55] Abel S, Abel zur Wiesch P, Davis BM, Waldor MK. Analysis of Bottlenecks in Experimental Models of Infection. PLoS Pathog. 2015;11(6):e1004823. Epub 2015/06/13. doi: 10.1371/journal.ppat.1004823. PubMed PMID: 26066486; PubMed Central PMCID: PMCPMC4465827.

[56] Abel S, Wiesch PAZ, Chang HH, Davis BM, Lipsitch M, Waldor MK. Sequence tag–based analysis of microbial population dynamics. Nat Methods 2015;12(3):223–6. https://doi.org/10.1038/nmeth.3253. PubMed PMID: WOS:000350670300021.

[57] Kassen R, Bataillon T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 2006;38(4):484–8. https://doi.org/10.1038/ng1751. PubMed PMID: WOS:000236340500024.

[58] Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet 2007;8(8):610–8. https://doi.org/10.1038/nrg2146. PubMed PMID: WOS:000248170700015.

[59] Ayala FJ, Gilpin ME, Ehrenfeld JG. Competition between species – theoretical models and experimental tests. Theor Popul Biol 1973;4(3):331–56. Doi 10.1016/0040-5809(73)90014-2. PubMed PMID: WOS:A1973Q774300006.

[60] Westbury CF. Bayes' rule for clinicians: an introduction. Front Psychol. 2010;1:192. Epub 2010/01/01. doi: 10.3389/fpsyg.2010.00192. PubMed PMID: 21833252; PubMed Central PMCID: PMCPMC3153801.

[61] Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet 2015;16(5):299–311. https://doi.org/10.1038/nrg3899. PubMed PMID: WOS:000353164800011.

[62] Sanft KR, Wu S, Roh M, Fu J, Lim RK, Petzold LR. StochKit2: software for discrete stochastic simulation of biochemical systems with events. Bioinformatics 2011;27(17):2457–8. https://doi.org/10.1093/bioinformatics/btr401. PubMed Central PMCID: PMCPMC3157925.