REPOP: bacterial population quantification from plate counts

Pedro Pessoa[1,2], Carol Lu[1,3,4], Stanimir Asenov Tashev[1,2], Rory Kruithoff[1,2], Douglas P. Shepherd[1,2], Steve Pressé[1,2,5]
[1]Center for Biological Physics, [2]Department of Physics,
[3]School of Biological and Health Systems Engineering,
[4]School of Mathematical and Statistical Sciences,
[5]School of Molecular Sciences,
Arizona State University, Tempe, AZ, USA

Abstract

Bacterial counts from native environments, such as soil or the animal gut, often show substantial variability across replicate samples. This heterogeneity is typically attributed to genetic or environmental factors. A common approach to estimating bacterial populations involves successive dilution and plating, followed by multiplying colony counts by dilution factors. This method, however, overestimates the heterogeneity in bacterial population because it conflates the inherent uncertainty in drawing a subsample from the total population with the uncertainty in the sample arising from biological origins. In other words, this approach may obscure features that may otherwise be present in the data hinting at the presence of genuine subpopulations. For example, in plate counting applied to *C. elegans* gut microbiota, observed multimodality is often interpreted as large host-to-host variance, while the randomness introduced by measurement is frequently ignored. To explicitly account for the uncertainty introduced by dilution and plating randomness, we introduce REPOP, a PyTorch-based library to REconstruct POpulations from Plates within a Bayesian framework. Beyond simple cases, REPOP addresses more complex scenarios, including multimodal populations and correcting the mathematically subtle, but experimentally relevant, bias introduced by excluding plates deemed too crowded to distinguish individual colonies. We demonstrate REPOP's ability to resolve distinct population peaks otherwise obscured by standard multiplication methods. Applications to both simulated and experimental datasets, including bacterial samples of different concentrations and ones from the gut microbiota of *C. elegans*, show that REPOP accurately recovers the underlying multimodality by properly accounting for error propagation, where naive multiplication fails. REPOP is available on GitHub: `https://github.com/PessoaP/REPOP`.

1 Introduction

Plate counting is fundamental to microbiology in estimating the number of living microorganisms in a sample [1,2]. For example, it is essential in food safety testing [3], water quality assessments [4,5], amongst many other applications [6–9] that require flexible quantification of microbial loads.

Briefly, plate counting involves serially diluting an original sample to ensure an appropriate colony density, allowing for the accurate enumeration of colonies for each sample (as summarized in Fig. 1a). Subsequently, the diluted samples are spread over nutrient-rich agar plates and incubated. After incubation, each viable organism (typically a bacterium) spread over the plate forms a colony, or in some cases a plaque [10], which are then enumerated as colony forming units (CFUs) or plaque forming units (PFUs), respectively. The degree of dilution (represented by the dilution factor) is typically set in such a way as to collect a statistically precise number of colonies but simultaneously avoiding colony overlap – usually aiming for 30 to 300 CFUs in a plate with a diameter of 10 cm. However, when studying a poorly characterized or highly variable bacterial population, collecting plates that are overcrowded is inevitable. Challenges related to counting crowded plates were only recently mitigated by AI-inspired tools [11–14].

Beyond simple quantification, plate counting is often used to assess population variation within a single microbial species, the degree of phenotypic variability [15–18] and environmental resource availability [19,20]. This stems from interpreting multimodality – several samples exhibiting high and several others low counts – or general heterogeneity in plate counts as reflecting population differences in colony formation, growth rates, or stochastic expression of traits [21,22]. This heterogeneity can then be used to understand ecological dynamics of bacterial populations [19,23–26] such as the competition of populations inside the gut of *C. elegans* [27].

On account of this, learning the variability from sample to sample is critical in accurately capturing a population's heterogeneity. Thus, in order to draw biological conclusions, one would have to disentangle the effects of the stochasticity involved in the measurement (e.g., the randomness involved in the process of dilution and plating) from the heterogeneity in the original sample's population arising from the underlying biological or ecological variation.

Naively, in dilution and plating, the total number of viable bacteria in the original sample is usually estimated by multiplying the number of colonies by the total dilution factor for each plate, resulting in a histogram of the population in the original samples. This process significantly overestimates the true variability in the bacterial count within the samples and can even obscure the multimodality in reconstructed population count histograms, as demonstrated in Fig. 1b and c.

While all methods – including those based on flow cytometry or DNA quantification [28–35] – have their own limitations, plate counting is considered the gold standard in several fields of microbiology [4,36]. It has, nevertheless, the critical limitation of propagating error introduced via dilution.

To overcome the limitations of traditional plate count estimation, we developed REPOP (REconstruct POpulations from Plates), a software package that employs a Bayesian approach that rigorously accounts for the stochasticity introduced by dilution and plating. By modeling these statistical processes, REPOP infers both the full probability distribution of the original samples, preserving its inherent heterogeneity, as demonstrated in Fig. 1b. Notably, REPOP can detect the multimodal structure of the data even when such multimodality is not readily apparent from dilution count histograms, as shown in Fig. 1c. REPOP is available on GitHub [37]

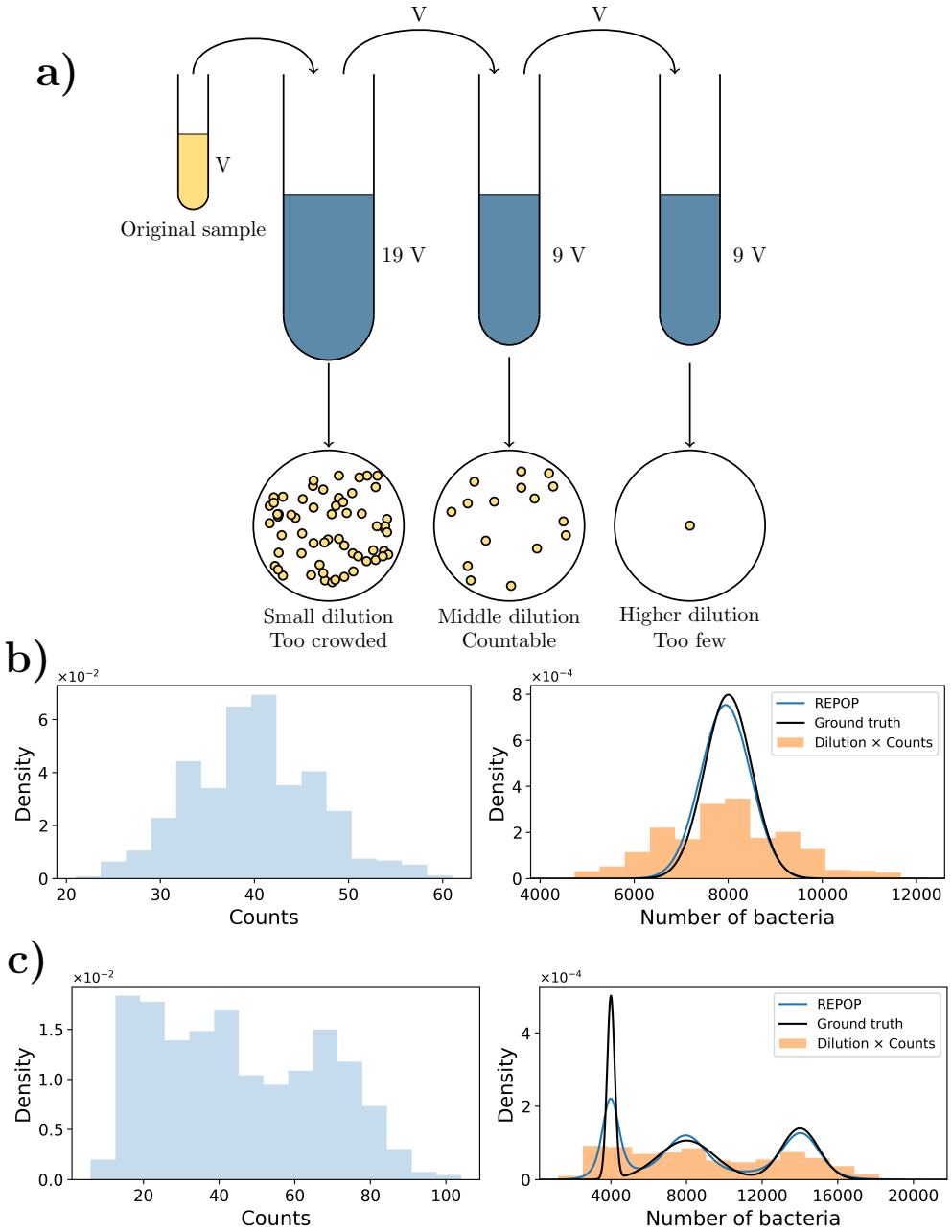

Figure 1: **Visual description of the dilution and plate counting process and the need for its analysis**. a) A sample of volume V is mixed with sterile medium in a larger vial (*e.g.*, $19V$, for a total dilution of 20). A portion of volume V of this diluted sample is then spread on a plate and incubated. Further tenfold dilutions are performed by transferring V into a vial with $9V$ of medium before plating. The process continues until a reasonable number of counts is obtained. b) The histogram on the left shows simulated counts across 1000 plates, obtained by diluting samples from the ground truth distribution by a factor of 200. The usual method of multiplying the counts by the total dilution factor produces the orange histogram (in the right) which appears significantly broader than the ground truth (black line, right) due to additional stochasticity introduced by dilution and plating. The ground truth distribution is a Gaussian with a mean of 8000 and a standard deviation of 500. REPOP's reconstruction (blue) corrects the broadening. c) A similar simulation to (b) where the ground truth population distribution is trimodal (parameters in Fig. 2). Here the stochasticity introduced dilution and plating obscure the three peaks, making them difficult to distinguish in the observed counts. Nevertheless, REPOP successfully reconstructs the trimodal distribution.

When presenting our method, we address a mathematically subtle point that may qualitatively affect the inferred counts: the cutoff for "uncountable" on a plate. That is to say, when multiple dilutions are performed in experiments to obtain plates with a large quantity of CFUs, plates that exceed a preset colony cutoff – a number beyond which colonies are considered too numerous to count – are excluded. While this practice is common, failing to properly account the exclusion of such plates can introduce bias in the inferred population distribution.

Within the present paper, we will explain how REPOP is able integrate the previously ignored over-populated plates exceeding the cutoff in a more sophisticated albeit requisite iteration of our simpler method. Why requisite? No matter how low we set our dilution factor, a plate may probabilistically always exceed the countable cutoff, though increasingly rarely. Thus, if we can deal with over-populated plates, then we may collect better statistics on colony counts by avoiding too large a dilution.

To demonstrate REPOP's applicability to real-world datasets, we analyzed two distinct experimental systems. First, we applied REPOP to plate counting data obtained from controlled mixtures of E. coli populations at different optical densities. Without prior knowledge of vial identity, REPOP successfully reconstructed the underlying distribution. Second, we employed REPOP to study bacterial colonization dynamics in individual C. elegans hosts. By analyzing plate counts of gut bacterial populations over several days post-feeding, REPOP revealed the progressive increase in bacterial load and captured the multimodal structure associated with host heterogeneity. These results highlight REPOP's ability to disentangle true biological variability from stochastic sampling noise in complex, real experimental contexts.

2 Methods

2.1 Statistical description of plate counting

We first describe the process of serial dilutions in mathematical terms. From Fig. 1a we observe that, on each plate, a fraction of the volume in the initial sample, equal to the inverse of their respective dilution factor (ϕ), is spread over the plate. As such, supposing the population in the original sample to be well-stirred, each individual bacterium has probability $1/\phi$ of being transferred onto that plate.

We denote the total number of bacteria in the original sample as n. Assuming each bacterium is independently and equally likely to be transferred, the number of bacteria transferred and forming colonies, denoted as k, follows a binomial distribution with parameters n (the number of trials) and $1/\phi$ (the probability of success in each trial) expressed as

$$p(k|n,\phi) = \binom{n}{k}\left(\frac{1}{\phi}\right)^k\left(1-\frac{1}{\phi}\right)^{n-k} . \tag{1}$$

With this description of the dilution and plating process, we now describe how we use Bayesian inference to reconstruct the population distribution.

2.2 Bayesian inference

For now we assume that each sample (drawn from the sample at the previous dilution) gives rise to one plate. The dataset obtained consists of the sequence of plate counts, k_i and their respective dilution factors ϕ_i for each replicate i. We denote the complete dataset as $D = (\bar{k}, \bar{\phi})$ with $\bar{k} = \{k_1, k_2, \ldots, k_I\}$ and $\bar{\phi} = \{\phi_1, \phi_2, \ldots, \phi_I\}$, with each pair (k_i, ϕ_i) generated independently as they arise from different samples.

Models	Plate counting method	Population distribution $p(n\|\theta)$	
Model 1	No cutoff	Unimodal	(3)
Model 2	No cutoff	Multimodal	(4)
Model 3	Cutoff	Multimodal	(4)

Table 1: For ease of reference, we summarize the differences among the three models taken into account in the present article.

Our overall goal is to learn the the probability distribution for the number of bacteria in the original samples, n, from the dataset D. To achieve this, we assume *a priori* that the distribution over n follows a model of population distribution with a set of parameters θ, written as $p(n|\theta)$.

The following subsection (2.3) is aimed at considering both unimodal and multimodal population distributions, $p(n|\theta)$. First, assuming that the population distribution is unimodal, we represent $p(n|\theta)$ as a simple Gaussian distribution, where θ reflects the mean and standard deviation. Second, in order to model a multimodal distribution, we invoke non-parametric models which provide a means to describe multimodal probability distributions for which the number of modes is unknown *a priori* [38–43]. More specifically, we assume a non-parametric Gaussian mixture, where θ represents a (potentially infinite) set of means, standard deviations, and mixture weights. Importantly, in large datasets, non-parametric models can flexibly adapt to approximate probabilities that do not strictly follow a predefined functional form [43–45].

The goal of reconstructing the distribution of bacteria in the original samples then becomes one of learning θ from D. In other words, we want to learn the probability of θ conditioned in D, $p(\theta|D)$. This probability is called a posterior. As we will see later, calculating the posterior will require the use of $p(n|\theta)$, which is uni- or multimodal, and the use of $p(\phi_i|n)$ when cutoffs on countability of plates are used in order to eliminate plates containing too many colonies.

REPOP is applied here to three cases of increasing complexity: the first (Model 1) is the simplest, assumes unimodal $p(n|\theta)$ without cutoffs. In Model 2 we keep no cutoff in counts, but generalize $p(n|\theta)$ to a multimodal (which simplifies to the unimodal Model when all but one component is considered). In Model 3 we keep the most general multimodal distribution for $p(n|\theta)$, but expand the counting method to account for the cutoff resulting in some plates not being counted and simply exceeding the cutoff. We summarize the differences among the three mechanisms in Table 1.

For all models highlighted above, in order to compute $p(\theta|D)$, we need the likelihood, $p(D|\theta)$, then related to $p(\theta|D)$ through Bayes' theorem

$$p(\theta|D) \propto p(\theta)\ p(D|\theta)\ . \tag{2}$$

Here $p(\theta)$ is the prior. Further details on the prior are given in the supplemental information A. Once we calculate probability $p(\theta|D)$, we may maximize this probability to obtain the maximum *a posteriori* (MAP) value, θ^{MAP}. More details are provided later in Sec. 2.5.

2.3 Models for the population distribution

2.3.1 Unimodal distribution (Model 1)

For Model 1, we can mathematically represent the probability of the true population count n given the distribution parameters θ as:

$$p(n|\theta) = \frac{\exp\left(-\frac{(n-\mu)^2}{2\sigma^2}\right)}{\sum\limits_{n'=0}^{\infty} \exp\left(-\frac{(n'-\mu)^2}{2\sigma^2}\right)} \ . \tag{3}$$

Here, θ includes the mean, μ, and standard deviation, σ, of the integer Gaussian $p(n|\theta)$, $\theta = \{\mu, \sigma\}$. The denominator above is necessary because we are dealing with (positive) integer Gaussians and will approach the usual $\sqrt{2\pi}\sigma$ when $\mu - 5\sigma \gg 1$. For computational reasons, we assume that the probability of having a number of bacteria, n, larger than twice the largest number found by multiplying CFU counts per dilution is approximately zero. Thus, the infinite upper bound over the sum in the denominator of (3) is substituted for $N^* = 2\max\limits_i(k_i\phi_i)$. Note that, by using the integer Gaussian in (3), we truncate the negative support of a Gaussian distribution, meaning we assume that even if we have a pair of μ and σ that assigns non-negligible probability to counts $n \leq 0$, those probabilities are set to zero, and the distribution is renormalized accordingly.

2.3.2 Multimodal distribution non-parametric inference (Models 2 and 3)

Naturally, the choice of modeling n through a single Gaussian component, Model 1, already commits us to assuming a unimodal distribution. However, if bacteria are present across replicates in a multimodal distribution — such as observed in some studies of *C.elegans* gut microbiota [8,27,46], where some replicates have few bacteria and others have more based on rare colonization events [27,46] — how many modes should we consider?

This consideration immediately forces upon us an important motivation for considering a much more general choice for $p(n|\theta)$, namely a non-parametric Gaussian mixture. Loosely speaking, non-parametric Gaussian mixture models do not assume a fixed number of parameters (or components) beforehand. Instead, they adapt and determine the appropriate number of components as warranted by the data. Knowledge of non-parametric mixture models, while subtle and detailed in Ref. [43] and references therein, are not required for running REPOP. In such a model, θ encompasses the means $\bar{\mu} = (\mu_1, \mu_2, \ldots)$, standard deviations $\bar{\sigma} = (\sigma_1^2, \sigma_2^2, \ldots)$, and mixture weights (the relative weights of each Gaussian component with the weights summing to unity) $\bar{\rho} = (\rho_1, \rho_2, \ldots)$, such that $\theta = \{\bar{\mu}, \bar{\sigma}, \bar{\rho}\}$. In this non-parametric model, the probability of having a sample of n bacteria is given as

$$p(n|\theta) = \sum_{\alpha=1}^{\infty} \rho_\alpha \left[\frac{\exp\left(-\frac{(n-\mu_\alpha)^2}{2\sigma_\alpha^2}\right)}{\left(\sum\limits_{n'=0}^{\infty} \exp\left(-\frac{(n'-\mu_\alpha)^2}{2\sigma_\alpha^2}\right)\right)} \right] \ . \tag{4}$$

When performing non-parametric inference, the choice of prior $p(\theta)$ is key. A well-chosen prior can add computational efficiency and avoid runaway pathologies associated with assigning every datapoint a different mixture component. A common choice of prior for the infinite mixture model satisfying both of these criteria is the Dirichlet process prior [43]; see supplemental information A for more details. While a finite truncation in the number of

components can be avoided [47, 48], for computational reasons, here we use a finite truncation (defaulted to the smallest value among 25 and the square root of the total number of samples, though modifiable, in REPOP) for the sum over the components, α, in (4).

2.4 Calculating the likelihood

Here we explain how to calculate the likelihood, $p(D|\theta)$, appearing in (2) and start with Models 1 and 2 and move to Model 3 where the cutoff is taken into consideration.

2.4.1 Likelihood for Models 1 and 2

In the Models where cutoffs are not taken into consideration (*i.e.*, where it is assumed that the colonies on a plate can always be counted and plates are never rejected for having too many colonies), the calculation of the total likelihood, for all counted plates from the original number of bacteria in one sample, n, follows the scheme described in Sec. 2.1. However, n_i is itself a realization of the population distribution $p(n|\theta)$ described in the previous section (2.3). Thus the probability that a plate with k_i counts were generated from a dilution ϕ_i is given by marginalizing over n_i as

$$p(k_i|\phi_i, \theta) = \sum_{n_i=0}^{\infty} p(k_i|\phi_i, n_i)p(n_i|\theta) \ . \tag{5}$$

With each factor $p(k_i|\phi_i, n_i)$ within the summation given by (1) and $p(n_i|\theta)$ given by the respective population model as presented in Sec. 2.3.

This allows us to write Bayes' theorem (2) as

$$\begin{aligned} p(\theta|D) &\propto p(\theta) \prod_{i=1}^{I} p(k_i|\phi_i, \theta) \\ &\propto p(\theta) \prod_{i=1}^{I} \sum_{n_i=0}^{\infty} p(k_i|\phi_i, n_i)p(n_i|\theta) \end{aligned} \tag{6}$$

where I is defined as the total number of replicates. Note that while this summation is truncated to $N^* = 2\max_i(k_i\phi_i)$, these sums can still remain computationally intensive. Thus, while REPOP can run on CPUs, we automatically leverage GPU acceleration to enable faster evaluation of these summations if GPU is available.

2.4.2 Likelihood for Model 3

For Model 3, we assume that the number of colonies may exceed a certain cutoff, k_{CO}, and that such plates are not counted. Typically, for each sample, only one of the plates generated in the process described in Fig. 1a is reported. In other words, for each collected sample only the counts (and dilution) of one of the plates at the first dilution resulting in a countable number of colonies (below a pre-defined cutoff) is recorded as a datapoint. This is illustrated in Fig. 1a.

This process is different from the previous models. If a dilution different from the first dilution in the dilution schedule is reported, this implies that all lower dilutions therefore resulted in plates whose colony count (a stochastic variable) already exceeded the cutoff k_{CO}.

To obtain the posterior over θ, it is necessary to know the dilution schedule from the plating experiment. We refer to the full set of these as $\Phi = \{\phi^1, \phi^2, \ldots, \phi^M\}$ ordered from

lowest to highest dilution factor. In the example given in Fig. 1a we have $\phi^1 = 20$, $\phi^2 = 200$, and $\phi^3 = 2000$. While, mathematically, we could extend this indefinitely by sequentially plating dilutions that are factors of 10 larger, in practice only a finite number of plates can be produced, thus justifying the upper limit M. It is expected that the dilution schedule leads to dilution values sufficiently large that the approximate number of colonies recovered approaches zero. Expanding upon the previous models, it is now important to report not only the counts but also the associated plate index. The datapoint is now reported as (k_i, m_i), which means that for the i-th sample, the plate diluted by a total factor of ϕ^{m_i} had k_i counts and all plates with smaller dilution had counts larger than the cutoff k_{CO}.

With the set of all possible dilutions Φ explicitly considered, we incorporate this dependence into the calculation of all relevant probabilities. Thus, the total posterior for the full dataset is given by

$$\begin{aligned} p(\theta|D,\Phi) &\propto p(\theta|\Phi) \prod_{i=1}^{I} p(k_i, m_i|\theta,\Phi) \\ &\propto p(\theta|\Phi) \prod_{i=1}^{I} \sum_{n_i=0}^{\infty} p(k_i, m_i|n_i,\Phi) p(n_i|\theta,\Phi) \\ &\propto p(\theta) \prod_{i=1}^{I} \sum_{n_i=0}^{\infty} p(k_i, m_i|n_i,\Phi) p(n_i|\theta) \ . \end{aligned} \tag{7}$$

Here, we used the fact that the population distribution and its parameters, θ, are independent of the dilutions selected to plate, Φ. Therefore, $p(\theta|\Phi) = p(\theta)$ is the original prior and $p(n_i|\theta,\Phi) = p(n_i|\theta)$ is given by the respective population model of Sec. 2.3. The remainder of this subsection focuses on how to calculate the likelihood of each datapoint, $p(k_i, m_i|n_i,\Phi)$, within the summation above. To obtain this likelihood, we revisit the dilution and plating process. As described in Fig. 1a, a plate is generated for each dilution within the set of dilutions Φ. For each sample, labeled i, the number of colonies on the plate with dilution ϕ^m is denoted as k_i^m.

Since all plates originate from the same original sample (with a finite number of bacteria, n_i), the values of k_i^m for the same i are, in principle, not independent. I nstead, the set of all k_i^m is jointly sampled from a single multinomial distribution. However, as each bacterium has an independent probability of $1/\phi^m$ of appearing on the m-th plate, the correlations between the multiple k_i^m arise solely due to the finite number of total bacteria in the sample. Consequently, as long as the fraction of plated bacteria remains significantly smaller than the total sample size (*i.e.*, $\sum_m 1/\phi^m \ll 1$), the number of plated bacteria can be treated as independent binomial random variables.

Under this assumption, each count k_i^m can be modeled as a binomial realization of the original number of bacteria n_i with dilution factor ϕ^m, as described in Eq. (1). Thus, the likelihood of observing k_i^m is written as:

$$p(k_i^m|n_i,\Phi) = \binom{n_i}{k_i^m} \left(\frac{1}{\phi^m}\right)^{k_i^m} \left(1 - \frac{1}{\phi^m}\right)^{n_i - k_i^m} . \tag{8}$$

As previously described, the datapoint (k_i, m_i) corresponds to the counts and the index of the first value within the sequence $(k_i^1, k_i^2, \ldots, k_i^M)$ that satisfies the condition $k_i^m \leq k_{\mathrm{CO}}$, where k_{CO} is the cutoff. To formalize this, we define m_i as the (superscript) index of the first plate where the counts fall within the cutoff:

$$m_i = \min\{m \mid k_i^m \leq k_{\mathrm{CO}}\}, \tag{9}$$

from which the observed datapoint (k_i, m_i) is obtained with:

$$k_i = k_i^{m_i}. \tag{10}$$

With this in mind, it becomes possible to rewrite the likelihood, $p(k_i, m_i|n_i, \Phi)$, as

$$p(k_i, m_i|n_i, \Phi) = p(k_i|m_i, n_i, \Phi)p(m_i|n_i, \Phi) \ . \tag{11}$$

Once more, we proceed by showing how to calculate each factor within the product above. Starting from the last factor, $p(m_i|n_i, \Phi)$, we note that if the plate with index m_i is reported, we must have that both $k_i^{m_i} \leq k_{\text{CO}}$ and $k_i^m > k_{\text{CO}}$ for all $m < m_i$,

$$p(m_i|n_i, \Phi) = P(\{k_i^{m_i} \leq k_{\text{CO}}\}|n_i, \Phi) \prod_{m=1}^{m_i-1} P(\{k_i^m > k_{\text{CO}}\}|n_i, \Phi) \ . \tag{12}$$

Here we use the capitalized P to represent that we are calculating the probability that a variable is within a set, rather than the probability of a single possible value. In this specific model, what we are asking is the probability that k_i^m is smaller than or equal to k_{CO}. This is given by summing the probability of every value of k_i^m up to k_{CO}, meaning

$$P(\{k_i^m \leq k_{\text{CO}}\}|n_i, \Phi) = \sum_{k_i^m=0}^{k_{\text{CO}}} p(k_i^m|n_i, \Phi) \ , \tag{13}$$

so (12) becomes

$$p(m_i|n_i, \Phi) = \left(\sum_{k_i^{m_i}=0}^{k_{\text{CO}}} p(k_i^{m_i}|n_i, \Phi) \right) \left[\prod_{m=1}^{m_i-1} \left(1 - \sum_{k_i^m=1}^{k_{\text{CO}}} p(k_i^m|n_i, \Phi) \right) \right] \tag{14}$$

with each $p(k_i^m|n_i, \Phi)$ given by (8).

Then, for the remaining factor in (11), $p(k_i|m_i, n_i, \Phi)$, since counts above the cutoff are not reported as data, the likelihood of finding k_i colonies in a plate diluted by a total factor ϕ^{m_i} is re-scaled to account for the fact it was truncated at the cutoff

$$p(k_i|m_i, n_i, \Phi) = \frac{\binom{n_i}{k_i} \left(\frac{1}{\phi^{m_i}}\right)^{k_i} \left(1 - \frac{1}{\phi^{m_i}}\right)^{n_i - k_i}}{\sum_{k'=0}^{k_{\text{CO}}} \binom{n_i}{k'} \left(\frac{1}{\phi^{m_i}}\right)^{k'} \left(1 - \frac{1}{\phi^{m_i}}\right)^{n_i - k'}} \ , \tag{15}$$

for $k_i \leq k_{\text{CO}}$ and zero otherwise.

Thus, in order to obtain the posterior (7), we need to substitute the likelihood per datapoint, $p(k_i, m_i|n_i, \Phi)$ in (11), which is computed using (15) and (14).

2.5 Bayesian MAP

In Bayesian inference, we compute the posterior distribution $p(\theta|D)$ as it captures the updated probability over the model parameters after observing the data. However, posteriors are often complex and high-dimensional (technically infinite in non-parametric models), making direct computation and visualization challenging. To address this, we often resort to sampling methods to approximate the posterior. However, these samplers can also require long computation.

Since our objective here is to create a tool that is quick and easy to run, we resort to maximum *a posteriori* (MAP) estimation. This reduces the problem of knowing the distribution over θ to identifying the optimal θ. In this model, we maximize $p(\theta|D)$ and denote the maximum value as θ^{MAP},

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|D) \ , \tag{16}$$

with $p(\theta|D)$ given by (6) for Models 1 and 2, or (7) for Model 3. For our purposes here, we say that $p(n|\theta^{\mathrm{MAP}})$ is the distribution of n we learned from the data. In REPOP, the maximum value θ^{MAP} is found using optimization libraries within PyTorch [49].

3 Results

In the previous section we have constructed the posterior $p(\theta|D)$ across unimodal and multimodal population distributions, and by using a likelihood that either assumes all plates are countable (Model 1-2) or considers plates exceeding the cutoff (Model 3).

3.1 When all plates are considered countable

For the first model, we generate simulated data drawn from a single Gaussian with equal dilution factors across all samples. Meaning, this assumes that the dilution factor used (200) is able to generate a countable number of colonies in all samples. We also assume a unimodal Gaussian, as in Model 1. The reconstruction obtained by REPOP was shown earlier in Fig. 1b where we see that REPOP corrects for the overestimation of variance introduced by the stochasticity in dilution and plating.

In the more complex multimodal Gaussian case (Model 2), we further observe that REPOP uncovers the true underlying distribution's peaks with greater resolution and sensitivity to multimodality than by naive multiplication. As an example, here we generate simulated data for a mixture of Gaussians with three modes. In Fig. 2a, we show a relatively simple dataset, where the three peaks are already discernible in the histogram obtained via naive multiplication. Nonetheless, REPOP improves the resolution of these peaks, particularly as the number of samples increases. In Fig. 2b, we present a more challenging case where the modes are less distinct under the naive multiplication approach; the additional stochasticity introduced by dilution causes the resulting histogram to appear as a single broad peak. Nonetheless, REPOP is able to accurately recover the underlying multimodality. In both cases, the performance of REPOP improves with larger datasets, as evidenced by the progressively sharper resolution of each peak.

3.2 When the cutoff is taken into account

For our final demonstration with simulated data, we show results when the data follows a cutoff in reporting, as described in Sec. 2.4.2. We have generated simulated data according to the following scheme: starting from the lowest dilution (20, as shown in Fig.1a). If the counts are larger than the cutoff, we used k_{CO}, we sample again at a dilution factor 10 times higher until we find a plate that has fewer colonies than k_{CO}. We then record that number of counts and that dilution factor as the i-th datapoint (k_i, ϕ_i), as would be done by rigorously following the cutoff. If we know the dilution schedule, this is equivalent to having the data as the counts and the index of the reported dilution (k_i, m_i) as discussed in Sec. 2.4.2, with each reported dilution matched to the corresponding index m_i, meaning the dilution reported

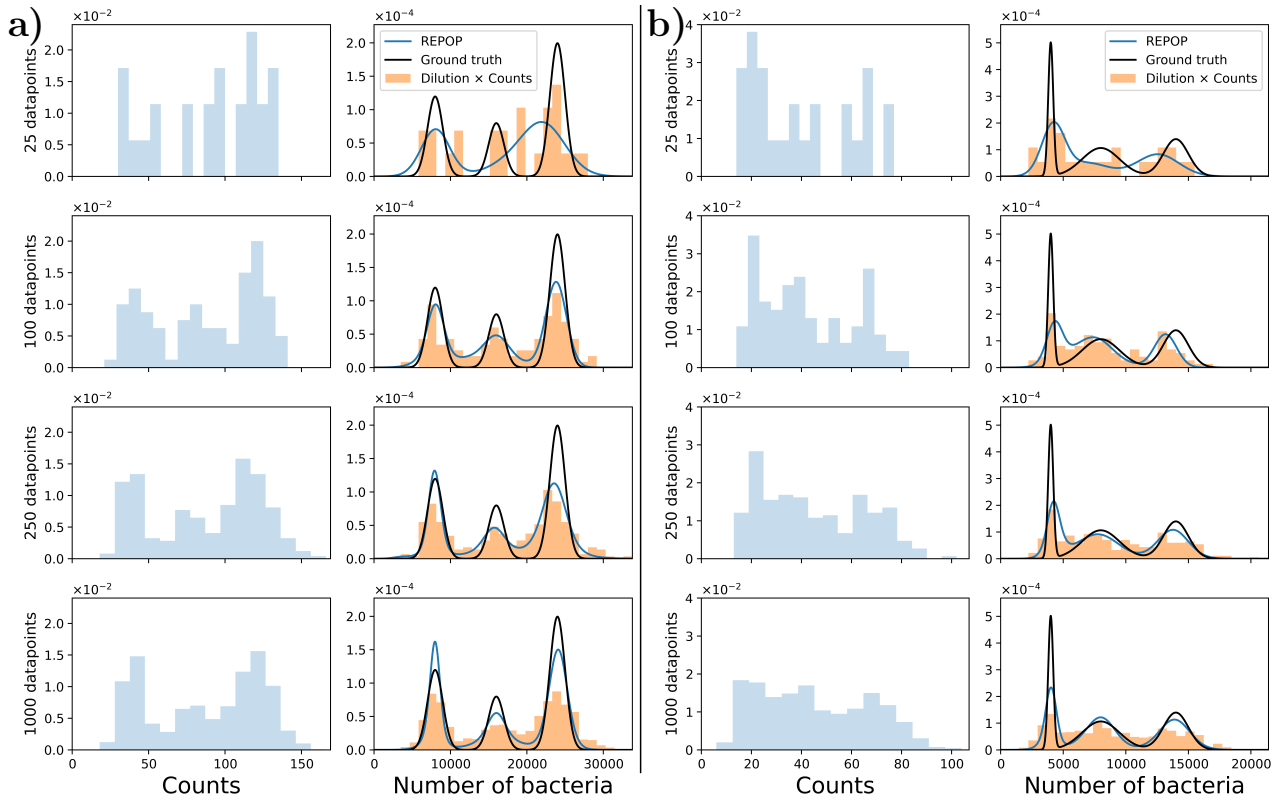

Figure 2: **REPOP can reconstruct multimodal populations even when they are not directly observable by naively multiplying counts by dilution.** Similar to Fig. 1b and c, this figure shows the observed counts, the naïve estimate obtained by multiplying counts by dilution, and the reconstruction produced by REPOP, compared to the ground truth. In both cases, the population n is drawn from a ground truth mixture of three Gaussians, shown as the black curves. In a), the mixture has means (4000, 8000, 14,000), standard deviations (200, 1500, 1000), and weights (0.25, 0.4, 0.35), while in b), it has means (8000, 16,000, 24,000), standard deviations (1000, 1000, 1000), and weights (0.3, 0.2, 0.5). Observed counts are sampled from (1), diluted by a factor of 200, and shown as histograms on the left. In a), a trimodal structure is visible in the observed counts histogram. However, in (b), the trimodality is heavily obscured by the stochasticity introduced by dilution and plating, making it indiscernible even in large datasets. Despite this, when applying REPOP to datasets of increasing size, we see that while a small dataset of 25 plates may result in missed modes, larger datasets allow REPOP to accurately recover the underlying multimodal structure in both cases. This demonstrates REPOP's ability to infer the true population despite the stochastic noise introduced by dilution and plating.

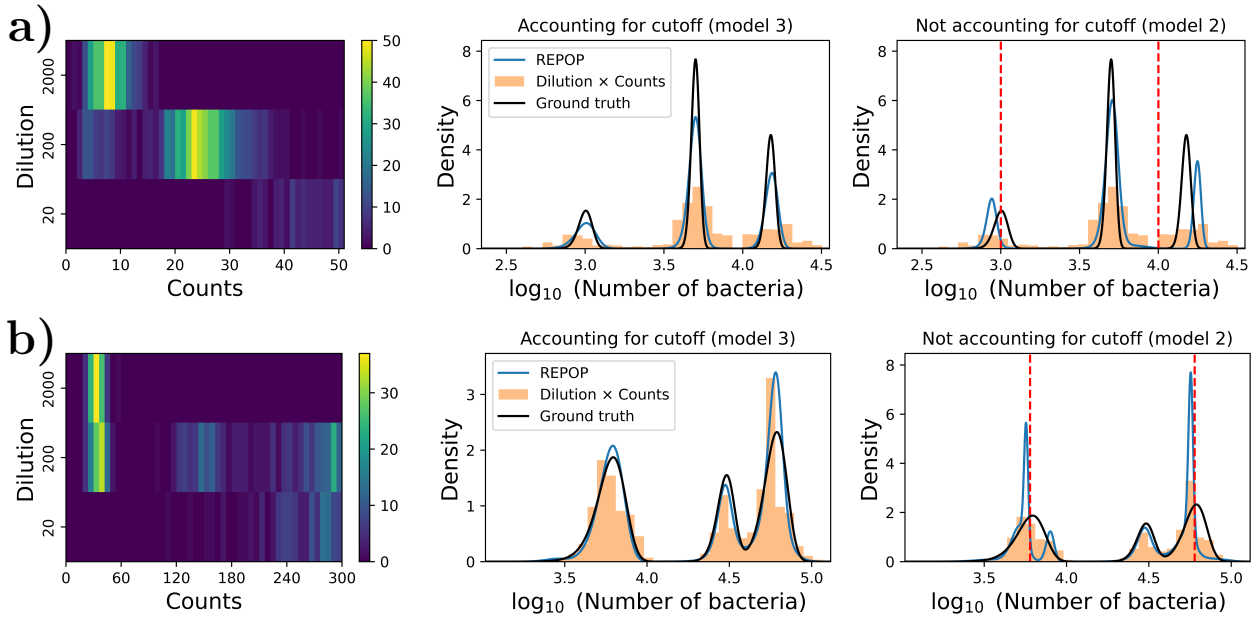

Figure 3: **Not accounting for the cutoff can lead to incorrectly attributed multimodality.** This figure compares population reconstructions with and without accounting for the cutoff effect, (using Model 3 and Model 2 respectively). a) simulated data generated with a cutoff $k_{\mathrm{CO}} = 50$. The ground truth population consists of three Gaussian components with means $\bar{\mu} = (1000, 5000, 15000)$, standard deviations $\bar{\sigma} = (100, 300, 1000)$, and weights $\rho = (\frac{1}{6}, \frac{1}{2}, \frac{1}{3})$. b) A more realistic cutoff $k_{\mathrm{CO}} = 300$, with ground truth parameters $\bar{\mu} = (6000, 30000, 60000)$, $\bar{\sigma} = (1200, 3600, 12000)$, and $\rho = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$. The middle column shows reconstructions using Model 3 (which accounts for the cutoff), while the rightmost column shows results from Model 2 (which ignores it). The rightmost column also highlights key cutoff crossings, using red dashed lines. These lines indicate where bacterial counts approach the cutoff at different dilutions. Notably, failing to account for the cutoff leads to incorrect mode placement including the emergence of an extra peak observed in b). As discussed in Sec. 3.2, when the true distribution has substantial probability mass near these points, meaning there is a non-negligible way to find a sample with these numbers, failing to account for the cutoff effects results in inaccurate reconstructions.

is the m_i-th in the schedule $\phi_i = \phi^{m_i}$. In Sec. 2.4.2 we also discuss how taking these cutoffs into account changes the likelihood calculation method.

We visualize the data generated this way and show the reconstruction obtained for Model 3 in Fig. 3. In the rightmost column, we demonstrate how failing to account for the cutoff leads to an artificial peak (*e.g.*, an extra mode near the leftmost peak in the ground truth, Fig. 3b). This occurs because counts near the cutoff k_{CO} are ambiguously assigned to incorrect dilutions.

These errors become particularly pronounced when colony counts approach k_{CO}, as the cutoff procedure forces certain samples to be reported at different dilutions. For example, consider a cutoff $k_{\mathrm{CO}} = 300$ and two possible dilution factors, $\phi^1 = 20$ and $\phi^2 = 200$. A sample with an original bacterial count n_i near $k_{\mathrm{CO}}\phi^1 = 6000$ may yield counts at dilution ϕ^1 that either exceed or fall below the cutoff due to variability in the plating process. When the cutoff effect is ignored, the algorithm mistakenly interprets these counts as coming from separate populations, leading to incorrect peak formation, as seen in the rightmost column of Fig. 3b. Beyond this qualitative observation, we also assess the magnitude of these errors

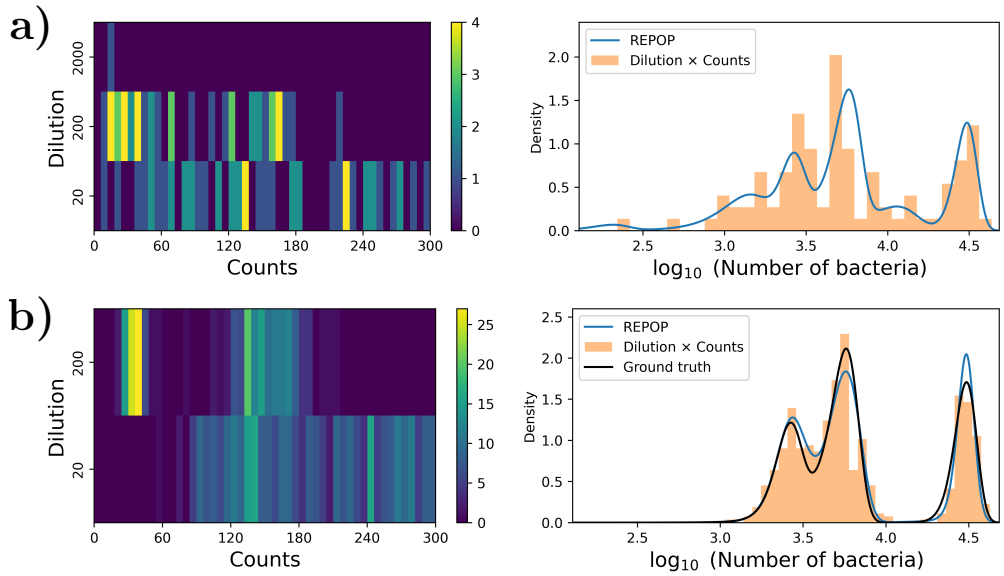

Figure 4: **Reconstructing, from real experimental plate counting, the population from vials with different optical densities.** a) Distribution of observed colony counts obtained as described in Sec. 3.3. The data is passed to the system without vial identification, aiming to reconstruct the underlying tetramodal structure. b) Results from a simulated dataset with 500 datapoints (plates). The increased sample size improves the estimation of the four main components, yielding a more accurate reconstruction.

by computing the relative deviation between the estimated and ground truth means for each component (see Supplemental Information B).

3.3 Results for experimental data: Discerning different true concentrations of bacteria

While we have demonstrated in the previous subsections how to resolve overlapping distributions in simulated datasets, here we move on to obtaining a similar case built with real-world plate counting. In particular, we construct a case that emulates multiple peaks at different dilutions, as shown in Fig. 3, by generating three different sample sources of bacteria.

We prepared vials with four different concentrations of *E.coli* measured by optical density, and consider samples of a volume equivalent to $0.2n$L from these vials. These samples are further diluted according to the dilution schedule $\Phi = 20, 200, 2000$. However, in practice, we do not directly dilute $0.2n$L. Instead, we take 1μL and dilute it by factors of 10^5, 10^6, and 10^7 – further details in Supplemental Information C.1. This approach ensures that while the original vials can be assumed to be have neglegible variability, the small $0.2n$L subsamples may no longer be, resulting in a unimodal but with a larger relative variance population for each vial.

The colony counts obtained are processed without identifying their original vial, aiming to reconstruct the underlying tetramodal structure. However, the dataset remains relatively small, consisting of only 97 total samples (see details in Supplemental Information C.1). This dataset and the reconstructed distribution are presented in Fig. 4a.

Although the limited sample size results in an imperfect match, the method successfully captures the four main components – corresponding to the estimated values of μ_α and σ_α associated with the four highest values of ρ_α. To further validate the approach, we supplement

the dataset by generating 800 synthetic datapoints from the mixture distribution made of the four main components, leading to a dataset considerably larger than what is typical of plate counting. The results, shown in Fig. 4b, demonstrate a significant improvement in capturing the underlying structure of the data.

3.4 Application: Capturing host differences through examining gut microbiota population

Here, we use REPOP to studying heterogeneity in more complex biological systems, such as microbial ecology within a host organism. Thus, we apply our framework to analyze gut bacterial counts from individual *C. elegans*.

Heterogeneity in bacterial counts among individual *C. elegans* has been reported and is typically attributed to rare colonization events and size variability [27, 46, 50]. However, to derive meaningful insights into gut ecology, it is essential to distinguish true population heterogeneity from the stochasticity introduced by dilution and plating.

To minimize confounding factors, we restrict our analysis to the OP50 strain of *E. coli*, which is well-characterized in laboratory settings and, under idealized conditions, is expected to produce a unimodal population distribution. Any deviations from this expected pattern may reflect individual *C. elegans* variability, either due to differences in the gut microenvironment or the fact that gut colonization is a rare event [27].

Due to this, we conducted an experiment in which *C. elegans* were synchronized, cleared of gut bacteria at three days post-hatching and subsequently fed live *E. coli* for controlled durations (1, 3, 5, 7, or 9 days). At each studied time point, 25 individual worms were picked, cleaned of all non-gut adhered bacteria, and crushed. The worm homogenate was then serially diluted and plated with a dilution schedule, $\Phi = \{22, 222, 2222\}$. The plates were counted after a 24 hour incubation. More detailed experimental methods may be found in the Supplemental Information C.2.

The results presented in Fig. 5 demonstrate a clear trend in bacterial population dynamics within the *C. elegans* gut. The population distribution shifts consistently to higher values with each successive day, aligning with our expectation that bacterial numbers increase over time due to colonization and replication inside the gut.

The multimodality exhibited in Fig. 5 can be explained by the heterogineity of the host. Recently, this has been collaborated by *in vivo* experiments [51]. This phenomenon could be related to the well-established stratification in heat-shock protein (HSP) expression, which is also correlated with aging and lifespan in nematodes [52–54]. Thus the variation in the population could be due to the fact that ingestion and egestion exhibit an age-related change [55, 56] or due to state switching [51] in feeding and digestion, akin to genetic state switching [57, 58]. The results in Fig. 5 support the conclusion that biologically meaningful differences in population counts are being captured, rather than the data being dominated by instrumentation error. REPOP provides a robust method for distinguishing multiple peaks in the population distribution, indicating that intrinsic variability among individual *C. elegans* can be deconvolved from measurement noise.

4 Conclusion

In accurately estimating bacterial counts, simply multiplying counts by the dilution factor provides a starting point. However, this method alone often fails to capture the original individual counts across samples, as straightforward multiplication of colony counts by the

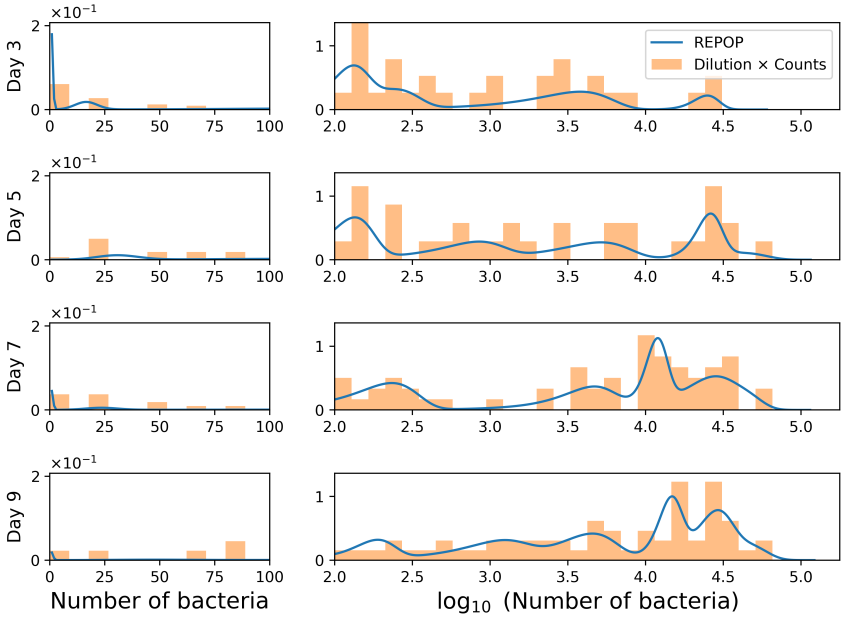

Figure 5: **Results for the population of *E.coli*-fed *C. elegans*' gut at days 3, 5, 7, and 9 of adulthood**. Here we are able to observe the shift in the population distribution obtained with plate counting shifts to higher population for longer-living *C. elegans*. To better visualize the distribution and model fits, we use a split-log x-axis: bacterial counts from 0 to 100 are shown on a linear scale, and counts above 100 on a logarithmic scale. The population distribution shifts toward higher values each day, as expected from colonization and replication within the gut. If biological significance, such as different *C. elegans* types, as proposed in *e.g.* [51] is to be inferred from this multimodality, a rigorous separation of population heterogeneity and plating stochasticity, as implemented by REPOP, is essential.

total dilution factor typically leads to an overestimation of the spread (Fig. 1b) that may obscure multimodal behavior (Fig. 1c). Conversaly, inferring population growth dynamics [8, 15,19,46,51] and learning features driving population heterogeneity across samples [19,25,26], demands a deeper analysis of the statistics involved in dilution and counting. Indeed, such a treatment avoids attributing ecological (or other) significance to the randomness introduced by plating.

The Bayesian reconstruction presented here and implemented through REPOP provides rigorous method to obtain the distribution of populations from plate counts, including observing the correctly multimodality of the population, as seen in Figs. 1 – 3.

The study of the multimodal case accounting for cutoff (Model 3) also demonstrates that retaining all data, regardless of their individual importance in reconstructing the final population size distribution, remains crucial. As shown in Fig. 3, ignoring the cutoff can lead to misidentifying multimodality, especially for plates with colony counts near the cutoff. Consequently, REPOP's results emphasize that explicitly reporting both the dilution schedule and the cutoff values used should be a fundamental part of data reporting, as they impact the form in which likelihoods are calculated.

As a clear extension of the current work, it is possible to combine the two modular steps involved in counting bacteria in plates with the current work of establishing the original population size based on plated samples. For instance, a notion of possible error rate, such as false positive or negative rate in counting that may even appear at high dilution (where there is always a chance for bacterial colony overlap), could help in further reducing the breadth in the original counts over bacteria. For instance, intuitively, we may imagine the probability of miscounting two or more colonies as a single one to grow as we reach colony counts approaching the cutoff. Simple simulations, for instance, considering typical colony and plate size, may be used to gain initial insight on potential false negative rates, and thus its effect on the original sample size.

Finally, REPOP provides a means by which to test hypothesis [8, 51] relating observed multimodality in gut microbial content of *C. elegans*, as shown in Fig. 5, to distinct categories of hosts. What is more, REPOP can also be used to study other organisms, exhibiting similar multimodatlity in their microbiota [59–61]. More than just a large dataset, identifying these multiple modes requires a theoretical framework maximizing the information extracted from each plate while remaining flexible enough to accommodate different numbers of modes. REPOP fulfills both of these requirements.

Acknowledgments

SP acknowledges support from the NIH (R35GM148237), ARO (W911NF-23-1-0304), and NSF (Grant No. 2310610). We also thank Ayush Saurabh, Max Schweiger, Lance (W. Q.) Xu, and Ioannis Sgouralis for insightful discussions in the development of this work.

References

[1] J. E. L. Corry, G. D. W. Curtis, R. M. Baird, Handbook of Culture Media for Food and Water Microbiology, The Royal Society of Chemistry, 2011. doi:10.1039/9781847551450.

[2] H. Smith, A. Brown, Benson's microbiological applications laboratory manual, 15th Edition, McGraw-Hill Education, 2021.

[3] S. L. Vial, D. R. Doerscher, C. W. Hedberg, W. A. Stone, S. J. Whisenant, C. M. Schroeder, Microbiological testing results of boneless and ground beef purchased for the U.S. national school lunch program, school years 2015 to 2018, Journal of Food Protection 82 (2019) 1761–1768. doi:10.4315/0362-028x.jfp-19-241.

[4] J. Bartram, J. Cotruvo, M. Exner, C. Fricker, A. Glasmacher, Heterotropic plate counts and drinking-water safety, Emerging issues in water and infectious disease series, World Health Organization, Genève, Switzerland, 2003.

[5] S. Some, R. Mondal, D. Mitra, D. Jain, D. Verma, S. Das, Microbial pollution of water with special reference to coliform bacteria and their nexus with environment, Energy Nexus 1 (2021) 100008. doi:10.1016/j.nexus.2021.100008.

[6] X. Wang, Z. Cao, M. Zhang, L. Meng, Z. Ming, J. Liu, Bioinspired oral delivery of gut microbiota by self-coating with biofilms, Science Advances 6 (2020) 1472–1476. doi:10.1126/sciadv.abb1952.

[7] A. Tamargo, N. Molinero, J. J. Reinosa, V. Alcolea-Rodriguez, R. Portela, M. A. Bañares, J. F. Fernández, M. V. Moreno-Arribas, PET microplastics affect human gut microbiota communities during simulated gastrointestinal digestion, first evidence of plausible polymer biodegradation during human digestion, Scientific Reports 12 (2022) 528. doi:10.1038/s41598-021-04489-w.

[8] K. M. Martini, S. S. Boddu, I. Nemenman, N. M. Vega, Maximum likelihood estimators for colony forming units, Microbiology Spectrum 12 (2024) e03946–23. doi:10.1128/spectrum.03946-23.

[9] M. N. Taylor, S. Spandana Boddu, N. M. Vega, Using single-worm data to quantify heterogeneity in Caenorhabditis elegans-bacterial interactions, Journal of Visualized Experiments 185 (2022) e64027. doi:10.3791/64027.

[10] E. Marine, D. S. Milner, C. Lambert, R. E. Sockett, K. M. Pos, A novel method to determine antibiotic sensitivity in Bdellovibrio bacteriovorus reveals a DHFR-dependent natural trimethoprim resistance, Scientific Reports 10 (2020) 5315. doi:10.1038/s41598-020-62014-x.

[11] Q. Geissmann, OpenCFU, a new free and open-source software to count cell colonies and other circular objects, PLOS ONE 8 (2013) 1–10. doi:10.1371/journal.pone.0054072.

[12] A. u. M. Khan, A. Torelli, I. Wolf, N. Gretz, AutoCellSeg: robust automatic colony forming unit (CFU)/cell analysis using adaptive image segmentation and easy-to-use post-editing techniques, Scientific Reports 8 (2018) 7302. doi:10.1038/s41598-018-24916-9.

[13] S. A. Albaradei, F. Napolitano, M. Uludag, M. Thafar, S. Napolitano, M. Essack, V. B. Bajic, X. Gao, Automated counting of colony forming units using deep transfer learning from a model for congested scenes analysis, IEEE Access 8 (2020) 164340–164346. doi:10.1109/ACCESS.2020.3021656.

[14] J. Zhang, C. Li, M. M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, M. Grzegorzek, A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches, Artificial Intelligence Review 55 (2021) 2875–2944. doi:10.1007/s10462-021-10082-4.

[15] D. D. Martino, F. Capuani, A. D. Martino, Growth against entropy in bacterial metabolism: the phenotypic trade-off behind empirical growth rate distributions ine.coli, Physical Biology 13 (2016) 036005. doi:10.1088/1478-3975/13/3/036005.

[16] D. De Martino, Maximum entropy modeling of metabolic networks by constraining growth-rate moments predicts coexistence of phenotypes, Physical Review E 96 (2017) 060401. doi:10.1103/physreve.96.060401.

[17] T. Pecht, A. C. Aschenbrenner, T. Ulas, A. Succurro, Modeling population heterogeneity from microbial communities to immune response in cells, Cellular and Molecular Life Sciences 77 (2019) 415–432. doi:10.1007/s00018-019-03378-w.

[18] A. P. Muntoni, A. Braunstein, A. Pagnani, D. De Martino, A. De Martino, Relationship between fitness and heterogeneity in exponentially growing microbial populations, Biophysical Journal 121 (2022) 1919–1930. doi:10.1016/j.bpj.2022.04.012.

[19] D. W. Armitage, S. E. Jones, How sample heterogeneity can obscure the signal of microbial interactions, The ISME Journal 13 (2019) 2639–2646. doi:10.1038/s41396-019-0463-3.

[20] F. G. Blanchet, K. Cazelles, D. Gravel, Co-occurrence is not evidence of ecological interactions, Ecology Letters 23 (2020) 1050–1063. doi:10.1111/ele.13525.

[21] J.-W. Veening, W. K. Smits, O. P. Kuipers, Bistability, epigenetics, and bet-hedging in bacteria, Annual Review of Microbiology 62 (2008) 193–210. doi:10.1146/annurev.micro.62.081307.163002.

[22] L. S. Moore, E. Stolovicki, E. Braun, Population dynamics of metastable growth-rate phenotypes, PLoS ONE 8 (2013) e81671. doi:10.1371/journal.pone.0081671.

[23] J. Mao, T. Lu, Population-dynamic modeling of bacterial horizontal gene transfer by natural transformation, Biophysical Journal 110 (2016) 258–268. doi:10.1016/j.bpj.2015.11.033.

[24] J. Werner, T. Pietsch, F. M. Hilker, H. Arndt, Intrinsic nonlinear dynamics drive single-species systems, Proceedings of the National Academy of Sciences 119 (2022) e2209601119. doi:10.1073/pnas.2209601119.

[25] M. Goberna, M. Verdú, Cautionary notes on the use of co-occurrence networks in soil ecology, Soil Biology and Biochemistry 166 (2022) 108534. doi:10.1016/j.soilbio.2021.108534.

[26] S. Pinto, E. Benincà, E. H. van Nes, M. Scheffer, J. A. Bogaards, Species abundance correlations carry limited information about microbial network interactions, PLoS Computational Biology 18 (2022) e1010491. doi:10.1371/journal.pcbi.1010491.

[27] N. M. Vega, J. Gore, Stochastic assembly produces heterogeneous communities in the caenorhabditis elegans intestine, PLOS Biology 15 (2017) e2000633. doi:10.1371/journal.pbio.2000633.

[28] S. Van Nevel, S. Koetzsch, C. Proctor, M. Besmer, E. Prest, J. Vrouwenvelder, A. Knezev, N. Boon, F. Hammes, Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring, Water Research 113 (2017) 191–206. doi:10.1016/j.watres.2017.01.065.

[29] M. G. Wilkinson, Flow cytometry as a potential method of measuring bacterial viability in probiotic products: A review, Trends in Food Science and Technology 78 (2018) 1–10. doi:10.1016/j.tifs.2018.05.006.

[30] M. M. Taimur, R. Acevedo, L. A. Sklar, K. Thomsen, G. P. Tegos, A. K. Camper, Rapid identification of bacterial viability, culturable and non-culturable stages by using flow cytometry, Protocol Exchange (2020). doi:10.21203/rs.3.pex-1153/v1.

[31] H. Davey, S. Guyot, Estimation of microbial viability using flow cytometry, Current Protocols in Cytometry 93 (2020) e72. doi:https://doi.org/10.1002/cpcy.72.

[32] S. J. Z. Hansen, P. Tang, A. Kiefer, K. Galles, C. Wong, W. Morovic, Droplet digital pcr is an improved alternative method for high-quality enumeration of viable probiotic strains, Frontiers in Microbiology 10 (2020) 3025. doi:10.3389/fmicb.2019.03025.

[33] M. L. J. Weitzel, C. S. Vegge, M. Pane, V. S. Goldman, B. Koshy, C. H. Porsby, P. Burguière, J. L. Schoeni, Improving and comparing probiotic plate count methods by analytical procedure lifecycle management, Frontiers in Microbiology 12 (2021) 693066. doi:10.3389/fmicb.2021.693066.

[34] H. Tracey, N. Coates, E. Hulme, D. John, D. R. Michael, S. F. Plummer, Insights into the enumeration of mixtures of probiotic bacteria by flow cytometry, BMC Microbiology 23 (2023) 48. doi:10.1186/s12866-023-02792-2.

[35] Q. Sun, K. Huynh, D. Muratovic, N. J. Gunn, A. R. Zelmer, L. B. Solomon, G. J. Atkins, D. Yang, Beyond the colony-forming-unit: Rapid bacterial evaluation in osteomyelitis, eLife 13 (2024) RP93698. doi:10.7554/elife.93698.2.

[36] International Organization for Standardization, Microbiology of the food chain -— Horizontal method for the enumeration of microorganisms —- Part 2: Colony count at 30 degrees C by the surface plating technique, Standard, Geneva, CH (2013).

[37] P. Pessoa, REPOP - Reconstructing population from plates, `https://github.com/PessoaP/platereconstruction` (2024).

[38] S. Pressé, K. Ghosh, R. Phillips, K. A. Dill, Dynamical fluctuations in biochemical reactions and cycles, Physical Review E 82 (2010) 031905. doi:10.1103/PhysRevE.82.031905.

[39] S. Jazani, I. Sgouralis, S. Pressé, A method for single molecule tracking using a conventional single-focus confocal setup, The Journal of Chemical Physics 150 (2019) 114108. doi:10.1063/1.5083869.

[40] M. Tavakoli, S. Jazani, I. Sgouralis, W. Heo, K. Ishii, T. Tahara, S. Pressé, Direct photon-by-photon analysis of time-resolved pulsed excitation data using bayesian nonparametrics, Cell Reports Physical Science 1 (2020) 100234. doi:https://doi.org/10.1016/j.xcrp.2020.100234.

[41] Z. Kilic, I. Sgouralis, S. Pressé, Generalizing HMMs to Continuous Time for Fast Kinetics: Hidden Markov Jump Processes, Biophysical Journal 120 (2021) 409–423. doi:10.1016/j.bpj.2020.12.022.

[42] I. Sgouralis, L. W. Q. Xu, A. P. Jalihal, Z. Kilic, N. G. Walter, S. Pressé, BNP-Track: a framework for superresolved tracking, Nature Methods 21 (2024) 1716–1724. doi:10.1038/s41592-024-02349-9.

[43] S. Pressé, I. Sgouralis, Data Modeling for the Sciences: Applications, Basics, Computations, Cambridge University Press, 2023.

[44] I. Bryan, J. Shepard, I. Sgouralis, S. Pressé, Inferring effective forces for Langevin dynamics using gaussian processes, The Journal of Chemical Physics 152 (2020) 124106. doi:10.1063/1.5144523.

[45] P. Pessoa, M. Schweiger, L. W. Q. Xu, T. Manha, A. Saurabh, J. A. Camarena, S. Pressé, Avoiding subtraction and division of stochastic signals using normalizing flows: NFdeconvolve (2025). arXiv:2501.08288, doi:10.48550/arXiv.2501.08288.

[46] M. N. Taylor, N. M. Vega, Sample pooling inflates error rates in between-sample comparisons: an empirical investigation of the statistical properties of count-based data, bioRxiv (2024). doi:10.1101/2022.07.25.501406.

[47] S. G. Walker, Sampling the dirichlet mixture model with slices, Communications in Statistics - Simulation and Computation 36 (2007) 45–54. doi:10.1080/03610910601096262.

[48] M. Kalli, J. E. Griffin, S. G. Walker, Slice sampling mixture models, Statistics and Computing 21 (2009) 93–105. doi:10.1007/s11222-009-9150-y.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[50] D. M. Eisenmann, Wormbook (2005).
URL `http://www.wormbook.org/`

[51] S. S. Boddu, K. M. Martini, I. Nemenman, N. M. Vega, Variance in c. elegans gut bacterial load suggests complex host-microbe dynamics, bioRxiv (2024). doi:10.1101/2024.12.09.627557.

[52] A. I. Yashin, J. W. Cypser, T. E. Johnson, A. I. Michalski, S. I. Boyko, V. N. Novoseltsev, Heat shock changes the heterogeneity distribution in populations of caenorhabditis elegans: does it tell us anything about the biological mechanism of stress response?, The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 57 (2002) B83–B92. doi:10.1093/gerona/57.3.b83.

[53] D. Wu, S. L. Rea, A. I. Yashin, T. E. Johnson, Visualizing hidden heterogeneity in isogenic populations of c. elegans, Experimental gerontology 41 (2006) 261–270. doi:10.1016/j.exger.2006.01.003.

[54] N. Vertti-Quintero, S. Berger, X. Casadevall i Solvas, C. Statzer, J. Annis, P. Ruppen, S. Stavrakis, C. Y. Ewald, R. Gunawan, A. J. DeMello, Stochastic and age-dependent proteostasis decline underlies heterogeneity in heat-shock response dynamics, Small 17 (2021) 2102145. doi:10.1002/smll.202102145.

[55] M. A. Bolanowski, R. L. Russell, L. A. Jacobson, Quantitative measures of aging in the nematode caenorhabditis elegans. i. population and longitudinal studies of two behavioral parameters, Mechanisms of ageing and development 15 (1981) 279–295. doi:10.1016/0047-6374(81)90136-6.

[56] C. Huang, C. Xiong, K. Kornfeld, Measurements of age-related changes of physiological processes that predict lifespan of caenorhabditis elegans, Proceedings of the National Academy of Sciences 101 (2004) 8084–8089. doi:10.1073/pnas.0400848101.

[57] P. Pessoa, M. Schweiger, S. Pressé, Avoiding matrix exponentials for large transition rate matrices, The Journal of Chemical Physics 160 (2024) 094109. doi:10.1063/5.0190527.

[58] Z. Kilic, M. Schweiger, C. Moyer, D. Shepherd, S. Pressé, Gene expression model inference from snapshot rna data using bayesian non-parametrics, Nature Computational Science 3 (2023) 174–183. doi:10.1038/s43588-022-00392-0.

[59] B. Acuña Hidalgo, S. A. Armitage, Host resistance to bacterial infection varies over time, but is not affected by a previous exposure to the same pathogen, Frontiers in Physiology 13 (2022) 860875. doi:10.3389/fphys.2022.860875.

[60] A. W. Clemmons, S. A. Lindsay, S. A. Wasserman, An effector peptide family required for drosophila toll-mediated immunity, PLoS Pathogens 11 (2015) e1004876. doi:10.1371/journal.ppat.1004876.

[61] J. Z. Chen, Z. Kwong, N. M. Gerardo, N. M. Vega, Ecological drift during colonization drives within-host and between-host heterogeneity in an animal-associated symbiont, Plos Biology 22 (2024) e3002304. doi:10.1371/journal.pbio.3002304.

Supplemental Information

A Prior selection across the uni-modal and non-parametric model

In Sec. 2.3.2 we allude to a selection of a prior, $p(\theta)$, capable of handling multi-modal data. This will be explained in detail here.

As mentioned in the main text, this prior encompasses all of the parameters needed to construct the probability distribution for n, given by Eq.(3) in the main text. Therefore, θ must include the means $\bar{\mu} = (\mu_1, \mu_2, \ldots)$, standard deviations $\bar{\sigma} = (\sigma_1^2, \sigma_2^2, \ldots)$, and mixture weights $\bar{\rho} = (\rho_1, \rho_2, \ldots)$ of the Gaussian mixture, $\theta = \{\bar{\mu}, \bar{\sigma}, \bar{\rho}\}$. As mentioned in the main text, for practical purposes we assume a weak limit, which we refer to as A and equivalent to the truncation on the number of components in Eq. (4) in the main text (defaulted as the smallest value among 25 and the square root of the total number of samples) although it is changeable in our software.

Let us separate each piece of the prior $p(\theta) = p(\bar{\mu}, \bar{\sigma}, \bar{\rho})$. First, we consider the prior for the weights. In a uni-modal distribution, case 1, the result is trivial: we have $\rho_1 = 1$ with probability 1. However, in a multi-modal cases (2-3) where the number of components is unknown, we use a Dirichlet distribution, a common choice for mixture models in Bayesian non-parametrics [43], given by:

$$\begin{aligned} \bar{\rho} &\sim \text{Dirichlet}(\bar{\xi}) \ , \\ p(\bar{\rho}) &= \frac{1}{B(\bar{\xi})} \prod_{\alpha=1}^{A} \rho_\alpha^{\xi_\alpha - 1} \ , \end{aligned} \tag{17}$$

where $\bar{\xi}$ is a sequence of the same shape as $\bar{\rho}$, $\bar{\xi} = \{\xi_1, \xi_2, \ldots, \xi_A\}$ and $B(\bar{\xi}) = \frac{\prod_{\alpha=1}^{A} \Gamma(\xi_\alpha)}{\Gamma\left(\sum_{\alpha=1}^{A} \xi_\alpha\right)}$ with Γ representing the gamma function. In order to prevent overfitting is an usual choice to have an exponentially decreasing. Let us write $\bar{\xi} = \lambda \bar{\eta}$ with $\bar{\eta}$ being a sequence $\bar{\eta} = \{\eta_1, \eta_2, \ldots, \eta_A\}$ where the parameters decay exponentially by a factor q, such that $\eta_\alpha = q^\alpha$, and λ being a scalar (we use $\lambda = \frac{1.1}{q^A}$). The expected value of ρ_α sampled from (17) will be $\langle \rho_\alpha \rangle = \frac{\eta_\alpha}{\sum_{\alpha'} \eta_{\alpha'}}$.

With the prior for $\bar{\rho}$ established, we proceed to define the prior for the means $\bar{\mu}$ and standard deviations $\bar{\sigma}$. Here, each component α is independent from each other. Initially, for the means, we select a distribution that avoids biasing the choice by being nearly uniform, but preventing too small numbers and adhering to a sensible cutoff in n. The mean of each Gaussian must be smaller than twice the largest number found by naively multiplying counts per dilution; we denote this largest cutoff as $N^* = 2 \max_i \phi_i k_i$. Thus, the prior for each mean, μ_α, is given by

$$\begin{aligned} \frac{\mu_\alpha}{N^*} &\sim \text{Beta}\,(1.1, 1.1) \ , \\ p(\mu_\alpha) &= \frac{1}{N^*} \left(\frac{\mu_\alpha}{N^*}\right)^{0.1} \left(1 - \frac{\mu_\alpha}{N^*}\right)^{0.1} \ . \end{aligned} \tag{18}$$

For the prior in $\bar{\sigma}$ we write a prior scales with its respective means but is still broad, we

choose a lognormal prior here

$$\sigma_\alpha|\mu_\alpha \sim \text{LogNormal}(\log \mu_\alpha, 1) \ ,$$
$$p(\sigma_\alpha|\mu_\alpha) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_\alpha} \exp\left[-\frac{(\log \sigma_\alpha/\mu_\alpha)^2}{2}\right] \ . \tag{19}$$

Thus, the prior used to generate the figures in the main text is defined as:

$$p(\theta) = p(\bar{\rho}) \prod_{\alpha=1}^{A} p(\mu_\alpha)\ p(\sigma_\alpha|\mu_\alpha) \ , \tag{20}$$

with $p(\bar{\rho}), p(\mu_\alpha)$, and $p(\sigma_\alpha|\mu_\alpha)$ given by (17), (18), and (19) respectively.

B Relative error between the means

When analyzing the error when ignoring the cutoff in Sec. 3.2, we used a metric that is the relative error between the means. When comparing two distributions, $p(n|\theta)$ and $p(n|\theta')$, in terms of how the means match, we use the metric

$$M(\theta, \theta') = \sum_{\alpha=1}^{A} \rho_\alpha \ \min_{\alpha'} \left| \frac{\mu_\alpha - \mu_{\alpha'}}{\mu_{\alpha'}} \right| \ . \tag{21}$$

given as percentages where α are indexes to the components of θ, while α' are the indexes of the components of θ'. The term within the absolute value is equivalent to the relative error between two means μ_α and $\mu_{\alpha'}$. We take the minimum to account for cases where the labels for different components do not match across the distributions. Throughout the text, we have presented the metric M in percentages. In the main text, Sec. 3.2, θ represents the reconstruction obtained by the method while θ' represents the ground truth.

For the dataset shown in Fig.3a, the relative error is 0.82% when the cutoff is accounted for (model 3). In contrast, when the cutoff is ignored (model 2), the error increases significantly to 7.50%. Similarly, for the dataset shown in Fig.3b, the relative error is 2.10% when the cutoff is considered, compared to 11.1% when it is not. Full details of these calculations are available on our GitHub repository [37].

C Experimental Methodology

C.1 E. coli culture for simulating different expected population peaks

An overnight culture of the OP50 strain of *Escherichia coli* was grown in lysogeny broth (LB) media, and its optical density at 600 nm (OD_{600}) was measured against an LB blank. The culture was concentrated to an expected 5X of OD_{600}=1.0 by centrifugation and resuspension in phosphate-buffered saline (PBS). The 5X suspension was then diluted to an expected OD_{600} of 0.5, 0.1,, and two at 0.01, with empirical OD_{600} values recorded against a PBS blank, which were 0.804, 0.188., 0.025, and 0.02. From the first three vials, we obtained 25 samples (each with 3 dilutions), while the last vial yielded 22 samples.

For each OD_{600} condition, serial dilutions were prepared to 1:100,000, 1:1,000,000, and 1:10,000,000 in PBS. A total of 90 μL from each dilution was plated onto LB agar, resulting in 75 plates per OD condition. Plates were incubated at 37°C for 18–24 hours before counting colony-forming units (CFUs) to estimate bacterial concentration.

C.2 C. elegans culture

The temperature-sensitive, reproductively sterile strain AU37 of Caenorhabditis elegans (C. elegans) was cultivated and synchronized according to standard protocols [50]. Following the protocol described by Gore and Vega [27], synchronized eggs were cultivated at the non-permissive temperature of 25°C to produce sterile adult worms. These worms were then washed and transferred to a heat-killed E. coli OP50 suspension in S media supplemented with gentamicin for 24 hours to clear gut bacteria.

Once cleared of gut bacteria, worms were washed once with 10 mL M9 buffer containing 0.1% Triton X and twice with 10 mL M9 buffer. They were then transferred to a live bacterial suspension (E. coli OP50 at $OD_{600} = 1.0$ in 1 mL S media) for a duration of 3, 5, 7, or 9 days. Fresh bacterial suspension was provided at 24, 72, 120, and 168 hours (approximately every 48 hours after the initial 24-hour feeding).

On the designated sampling days for gut bacterial enumeration, worms were washed to remove non-adhered surface bacteria and manually disrupted following the protocol outlined in the section titled *Disruption of Worms and Plating of Intestinal Populations* of Ref. [27]. The resulting homogenate was serially diluted using a 10-fold dilution series, with final expected dilution factors of 1:22, 1:222, and 1:2222 before plating. Colony-forming units were enumerated after incubating the plates for 24 hours.