Reviewed Preprint
v1 • July 9, 2025
Not revised

Reviewed Preprint
v2 • September 8, 2026
Revised by authors

REPOP: bacterial population quantification from plate counts

Pedro Pessoa[1,2,3], **Carol Yunxiao Lu**[1,3,4,5,6], **Stanimir Asenov Tashev**[1,2,3], **Rory Kruithoff**[1,2], **Douglas P Shepherd**[1,2], **Steve Pressé**[1,2,3,4] ✉

[1]Center for Biological Physics, Arizona State University, Tempe, United States • [2]Department of Physics, Arizona State University, Tempe, United States • [3]Center for Single Molecule Biophysics – Biodesign Institute, Arizona State University, Tempe, United States • [4]School of Molecular Sciences, Arizona State University, Tempe, United States • [5]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, United States • [6]School of Biological and Health Systems Engineering, Arizona State University, Tempe, United States

For correspondence: spresse@asu.edu

Competing interests: No competing interests declared

Funding: See page 25

Reviewing editor: Anne-Florence Bitbol, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

© 2025, Pessoa et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife Assessment

This **important** study introduces a Bayesian method to determine bacterial counts that accounts for the experimental noise inherent to dilution and plating methods and distinguishes it from biological uncertainty. The evidence supporting the conclusions is **compelling**, combining simulated data and experimental data. The method will be of interest to microbial ecologists, and potentially to the broader community interested in inference from biological data, even more so if the domain of application and the limitations are further clarified.

https://doi.org/10.7554/eLife.107122.2.sa3

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/LabPresse/REPOP*.

Introduction

Plate counting is fundamental to microbiology in estimating the number of living microorganisms in a sample (Corry et al., 2011; Smith and Brown, 2021). For example, it is essential in food safety testing (Vial et al., 2019), water quality assessments (Bartram et al., 2003; Some et al., 2021), amongst many other applications (Wang et al., 2020; Tamargo et al., 2022; Martini et al., 2024; Taylor et al., 2022) 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 (Marine et al., 2020). All of the colonies or plaques on the plate 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 (Geissmann, 2013; Khan et al., 2018; Albaradei et al., 2020; Zhang et al., 2021).

Beyond simple quantification, plate counting is often used to assess population variation within a single microbial species, the degree of phenotypic variability (Martino et al., 2016; De Martino, 2017; Pecht et al., 2019; Muntoni et al., 2022) and environmental resource availability (Armitage and Jones, 2019; Blanchet et al., 2020). 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 (Veening et al., 2008; Moore et al., 2013). This heterogeneity can then be used to understand ecological dynamics of bacterial populations (Mao and Lu, 2016; Armitage and Jones, 2019; Werner et al., 2022; Goberna and Verdú, 2022; Pinto et al., 2022) such as the competition of populations inside the gut of *Caenorhabditis elegans* (*C.elegans*) (Lu et al., 2026; Vega and Gore, 2017).

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 must 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 sample. 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.

Indeed, this higher variance introduced in quantification is a major disadvantage of plate counting when compared to other methods, such as those based on flow cytometry or DNA quantification (Van Nevel et al., 2017; Wilkinson, 2018; Hansen et al., 2020; Weitzel et al., 2021; Tracey et al., 2023; Sun et al., 2024), which have the ability to count cells without dilution. Yet, plate counting remains the gold standard in several areas of microbiological research due to its ability to quantify only viable cells (Davey and Guyot, 2020; International Organization for Standardization, 2013; Bartram et al., 2003). In contrast, flow cytometry can only distinguish viable cells under specific conditions. It typically relies on fluorophore-based methods, assuming that a high intracellular concentration of non-membrane-permeable fluorescent probes indicates

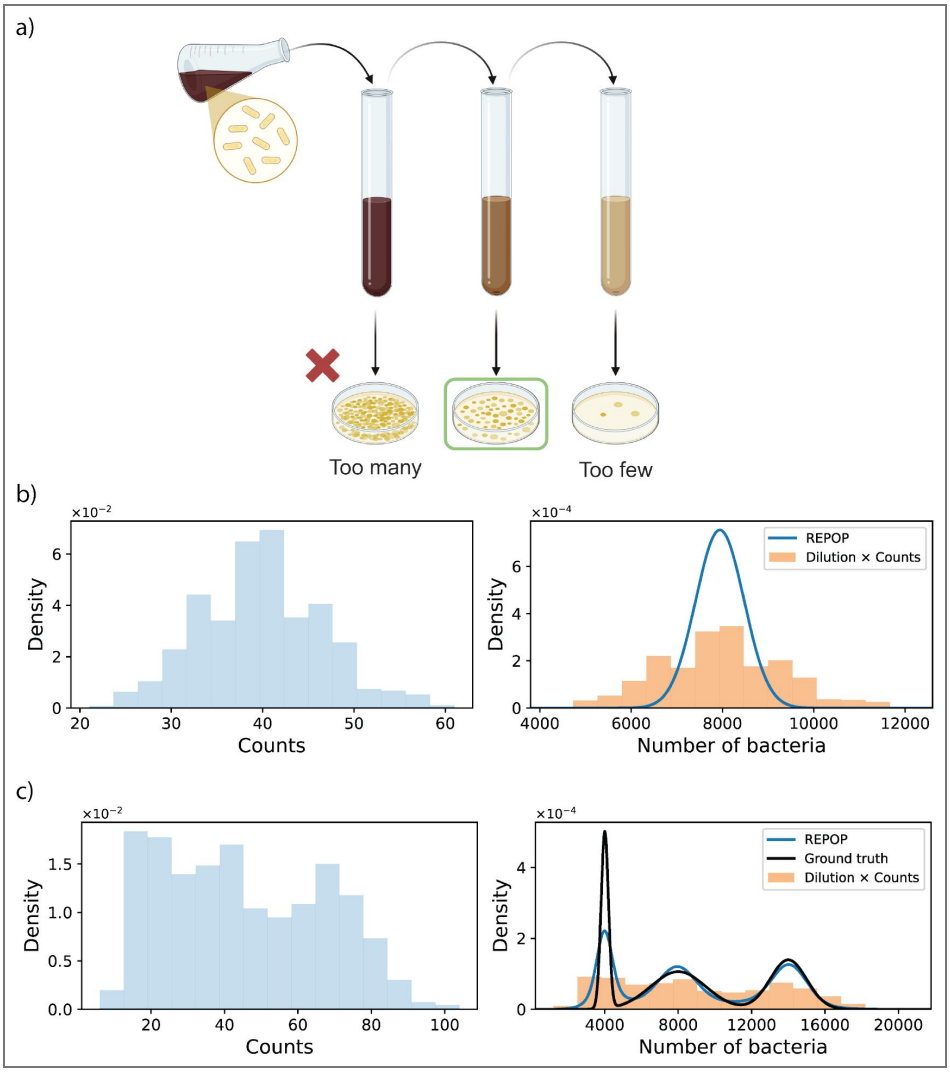


Figure 1. Visual description of the dilution and plate counting process and the need for its analysis.

a) A sample of volume *V* is mixed with sterile medium in a larger vial (*e.g.*, 19*V*, 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 9*V* of medium before plating. The process continues until a reasonable number of counts is obtained. Panel created using BioRender (*https://BioRender.com/espzw5y*). 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 is 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. 3). Here, the stochasticity introduced by dilution and plating obscures the three peaks, making them difficult to distinguish in the observed counts. Nevertheless, REPOP successfully reconstructs the trimodal distribution.

a compromised membrane and, therefore, a lack of viability (Taimur et al., 2020; Davey and Guyot, 2020). These challenges underscore why plate counting remains relevant, and improvements in the technique have wide downstream implications.

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 their 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 developed in Python and built on PyTorch (Paszke et al., 2019). Although PyTorch was originally developed for neural networks, it is also well suited for the type of optimization and GPU-accelerated calculations required here. REPOP is pip-installable and can take advantage of a GPU when available.

REPOP is available on GitHub (Pessoa, 2025), where users can find recipe-like tutorials, including examples of how to reconstruct the underlying population distribution and obtain quantities such as quantiles. A schematic overview of the workflow is shown in Fig. 2. Briefly, the input data consist of a sequence of observed colony counts together with their respective dilution factors, where each pair corresponds to a single biological sample. These data are passed to the inference pipeline, which evaluates the likelihood and reconstructs the distribution of the original bacterial population across samples. The main output is the inferred population distribution, along with summary statistics such as moments or quantiles.

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 for the exclusion of such plates can introduce bias in the inferred population distribution.

Here, we will explain how REPOP is able to integrate previously ignored over-populated plates exceeding the cutoff in a more sophisticated albeit requisite iteration of our simpler method.

Why requisite? Under the very general assumption that there is a nonzero probability of the population sizes being larger than the cutoff, no matter how low we set our dilution factor, a plate may still probabilistically 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 *Escherichia coli* (*E. coli*) populations at different optical densities. Without being provided the optical density condition from which each individual sample was drawn from, REPOP successfully reconstructed the mixture distribution of bacterial abundances across the different conditions. 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.

Methods

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 of the initial sample, equal to the inverse of the dilution factor (Φ), is spread over the plate.

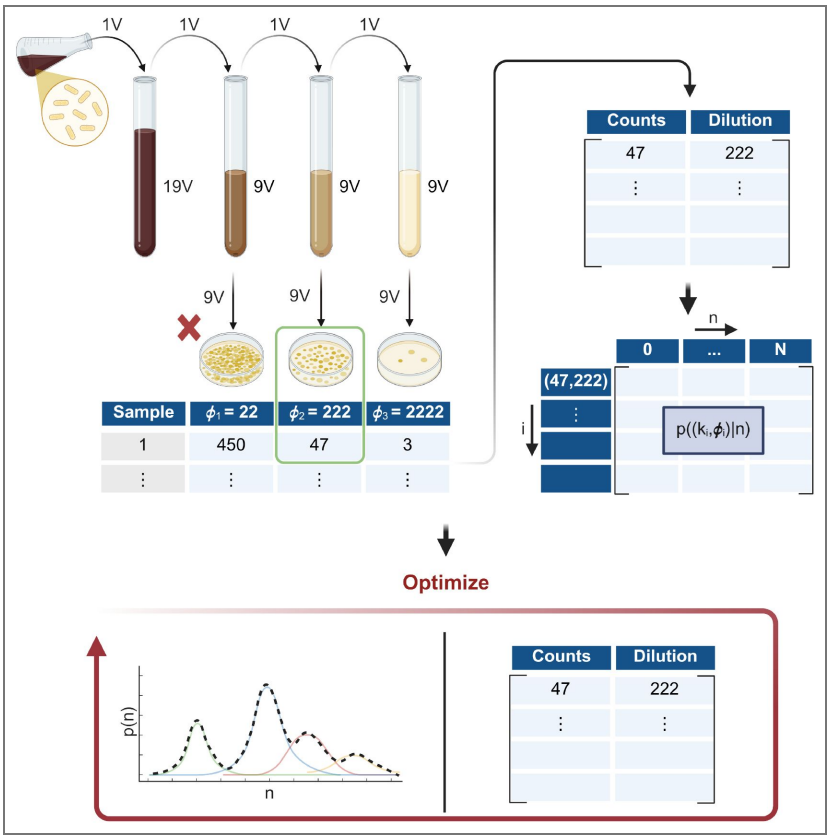


Figure 2. REPOP workflow from plate counts to reconstructed population statistics.

The input consists of observed colony counts and their corresponding dilution factors, organized as pairs (k_i, Φ_i), each arising from a single biological sample. From these data, REPOP evaluates the probability of each observed dilution count pair conditioned on the underlying bacterial number, $p(k_i, \Phi_i \mid n_i)$, thereby capturing the stochasticity introduced by dilution and plating. These likelihood contributions are then combined to reconstruct the distribution of the original bacterial population across samples. Icons in the top-left are from BioRender (*https://BioRender.com/espzw5y*).

In practice, a sample is first collected from the biological system and then mixed as part of the dilution procedure prior to plating, resulting in a well-mixed solution from which a subvolume corresponding to a fraction $^1/_{\Phi}$ of the total volume is transferred to the plate. Consequently, each individual bacterium in the diluted sample has probability $^1/_{\Phi}$ of being transferred onto the 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 to the final plate and form a colony, the resulting number of colonies, denoted as k, follows a binomial distribution with parameters n (the number of trials) and $\frac{1}{\phi}$ (the probability of success in each trial), expressed as

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

Here, for the numbers usually used in plate counting, the stochasticity introduced by the dilution is far from negligible. From (1), the variance of k is given by $\frac{n}{\phi}\left(1 - \frac{1}{\phi}\right)$, while the mean is $\frac{n}{\phi}$. Thus, for large dilution factors, the variance is approximately equal to the mean.

In practice, we expect colony counts to remain below 300. For example, if the expected count is 150, then the binomial variance is approximately 150, corresponding to a standard deviation of $\sqrt{150} \approx 12$. Therefore, the observed count k will typically lie within one standard deviation of the mean, approximately in the interval [138, 162]. This corresponds to a relative error of approximately 8%. Similarly, for an expected count of 50 (smaller but still reasonable) the standard deviation of k is approximately $\sqrt{50} \approx 7$. Therefore, for a case where the expected number is 50, one standard deviation corresponds roughly to the interval [43, 57], which gives a relative error of approximately 14%.

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 $\phi = \{\phi_1, \phi_2, \ldots, \phi_I\}$, with each pair ($k_i$, Φ_i) generated independently as they arise from different samples.

Our overall goal is to learn 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 $p(n \mid \theta)$.

Later, we will consider both unimodal and multimodal population distributions, $p(n \mid \theta)$. First, assuming that the population distribution is unimodal, we represent $p(n \mid \theta)$ 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* (Pressé et al., 2010; Jazani et al., 2019; Tavakoli et al., 2020; Kilic et al., 2021; Sgouralis et al., 2024; Pressé and Sgouralis, 2023). 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 (Bryan IV et al., 2020; Pressé and Sgouralis, 2023; Pessoa et al., 2025).

The goal of reconstructing the distribution of bacteria in the original samples then becomes one of learning the underlying parameters (θ) from plate count data (D). In other words, we want to learn the probability of θ conditioned on D, $p(\theta \mid D)$. This probability is called a posterior. As we will see later, calculating the posterior will require the use of $p(n \mid \theta)$, which is uni- or multimodal, and the use of $p(\Phi_i \mid n)$ 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 $p(n \mid \theta)$ without cutoffs. In Model 2, we keep no cutoff in counts, but generalize $p(n \mid \theta)$ 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 $p(n \mid \theta)$, but expand the counting method to account for the cutoff, resulting in some plates not being counted due to exceeding the cutoff. We summarize the differences among the three mechanisms in Table 1.

Models	Plate counting method	Population distribution $p(n\|\theta)$
Model 1	No cutoff	Unimodal (5)
Model 2	No cutoff	Multimodal (6)
Model 3	Cutoff	Multimodal (6)

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

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

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

Here $p(\theta)$ is the prior. Further details on the prior are given in the Appendix 1. Once we calculate probability $p(\theta \mid D)$, we may maximize this probability to obtain the maximum *a posteriori* (MAP) value, θ^{MAP} (as we describe later).

Population versus measurement stochasticity

In the following sections, we describe how the likelihood, $p(D \mid \theta)$, is computed. We denote the dataset as $D = \{(k_i, \phi_i)\}_{i=1}^{I}$, where each pair corresponds to the colony count and associated dilution factor for sample *i*. Assuming samples are uncorrelated, except in sharing the same underlying ecology, they may be treated as independent and identically distributed, so that the likelihood factorizes as

$$p(D \mid \theta) = \prod_{i=1}^{T} p\left(k_i, \phi_i \mid \theta\right). \tag{3}$$

However, each observed plate count arises from an underlying (and unobserved) number of bacteria in the original sample, which we denote by n_i. This latent variable captures the true population size for sample i and is the quantity directly informed by the ecological variability. Thus, we can rewrite each term in the likelihood by marginalizing over n_i,

$$p\left(k_i, \phi_i \mid \theta\right) = \sum_{n_i=0}^{\infty} p\left(k_i, \phi_i \mid n_i\right) p\left(n_i \mid \theta\right). \tag{4}$$

This decomposition makes explicit the two sources of stochasticity. The term $p(n_i \mid \theta)$ describes the population distribution, encoding the intrinsic variability across samples, while $p(k_i, \Phi_i \mid n_i)$ describes the measurement process, capturing the stochastic effects of dilution and plating. The following two subsections explain these two ingredients in detail. In particular, the population model $p(n_i \mid \theta)$ is flexible and must be chosen to reflect the underlying biological or ecological structure of interest, whereas the measurement model

$p(k_i, \Phi_i \mid n_i)$, explained in the following subsection, is dictated by the experimental protocol itself.

In the following sections, we introduce a hierarchy of models for $p(n_i \mid \theta)$, ranging from simple unimodal distributions to flexible non-parametric multimodal representations. While REPOP provides general purpose choices for these models, users interested in specific ecological hypotheses may substitute alternative forms for $p(n_i \mid \theta)$ without modifying the measurement model. That is, REPOP can also be used to extract only the measurement component, $p(k_i, \Phi_i \mid n_i)$, whose implementation is described in detail on GitHub (Pessoa, 2025).

Models for the population distribution

Unimodal distribution (Model 1)

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

$$p\left(n_i \mid \theta\right) = \frac{\exp\left(-\frac{\left(n_i-\mu\right)^2}{2\sigma^2}\right)}{\sum_{n'=0}^{\infty}\exp\left(-\frac{\left(n'-\mu\right)^2}{2\sigma^2}\right)}. \tag{5}$$

Here, θ includes the mean, μ, and standard deviation, σ, of the integer Gaussian $p(n_i \mid \theta)$, $\theta = \{\mu, \sigma\}$.

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 (5) is substituted for $N^* = 2\max(k_i\Phi_i)$. Note that, by using the integer Gaussian in (5), 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 $n_i \leq 0$, those probabilities are set to zero, and the distribution is renormalized accordingly.

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 (Vega and Gore, 2017; Taylor and Vega, 2024; Martini et al., 2024), where some replicates have few bacteria and others have more based on rare colonization events Vega and Gore (2017); Taylor and Vega (2024)) how many modes should we consider?

This consideration immediately forces upon us an important motivation for considering a much more general choice for $p(n \mid \theta)$, 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 Pressé and Sgouralis (2023) and references therein, is not required for running REPOP. In such a model, θ encompasses the means $\bar{\mu} = (\mu_1, \mu_2, \ldots)$, standard deviations $\bar{\sigma} = (\sigma_1, \sigma_2, \ldots)$, 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

$$p\left(n_i \mid \theta\right) = \sum_{\alpha=1}^{\infty}\rho_\alpha\left[\frac{\exp\left(-\frac{\left(n_i-\mu_\alpha\right)^2}{2\sigma_\alpha^2}\right)}{\sum_{n'=0}^{\infty}\exp\left(-\frac{\left(n'-\mu_\alpha\right)^2}{2\sigma_\alpha^2}\right)}\right]. \tag{6}$$

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 (Pressé and Sgouralis, 2023); see Appendix 1 for more details. While a finite truncation in the number of components can be avoided (Walker, 2007; Kalli et al., 2009), 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 (6).

Calculating the likelihood

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

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, is given by (1). However, n_i is itself a realization of the population distribution $p(n|\theta)$. Thus, the probability that a plate with k_i counts was generated from a (fixed) dilution Φ_i is given by

$$p\left(k_i \mid \phi_i, \theta\right) = \sum_{n_i=0}^{\infty} p\left(k_i \mid \phi_i, n_i\right) p\left(n_i \mid \theta\right). \tag{7}$$

Here, the dilution factor is treated as fixed rather than as a random variable because it is assumed that the colonies can always be counted. Within the summation in (7), each factor $p(k_i|\Phi_i, n_i)$ within the summation given by (1) and $p(n_i \mid \theta)$ given by the respective population model as presented in Models 1–3 in the previous subsection.

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

$$\begin{aligned} p(\theta \mid D) &\propto p(\theta) \prod_{i=1}^{I} p\left(k_i \mid \phi_i, \theta\right) \\ &\propto p(\theta) \prod_{i=1}^{I} \left[\sum_{n_i=0}^{\infty} p\left(k_i \mid \phi_i, n_i\right) p\left(n_i \mid \theta\right)\right] \end{aligned} \tag{8}$$

where I is defined as the total number of replicates. Note that while this summation is truncated to twice the naively reconstructed maximum from observed CFU counts $N^* = 2\max_i\left(k_i\phi_i\right)$, 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 a GPU is available.

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 $\Phi = \{\Phi^1, \Phi^2, ..., \Phi^M\}$ ordered from lowest to highest dilution factor. In the example given in Fig. 1a, we have $\Phi^1 = 22$, $\Phi^2 = 222$, and $\Phi^3 = 2222$. 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

$$\begin{aligned} p(\theta \mid D, \Phi) &\propto p(\theta \mid \Phi) \prod_{i=1}^{I} p\left(k_i, m_i \mid \theta, \Phi\right) \\ &\propto p(\theta \mid \Phi) \prod_{i=1}^{I} \sum_{n_i=0}^{\infty} p\left(k_i, m_i \mid n_i, \Phi\right) p\left(n_i \mid \theta, \Phi\right) \\ &\propto p(\theta) \prod_{i=1}^{I} \sum_{n_i=0}^{\infty} p\left(k_i, m_i \mid n_i, \Phi\right) p\left(n_i \mid \theta\right) \end{aligned} \tag{9}$$

Here, we used the fact that the population distribution and its parameters, θ, are independent of the dilutions selected to plate, Φ. Therefore, $p(\theta \mid \Phi) = p(\theta)$ is the original prior and $p(n_i \mid \theta, \Phi) = p(n_i \mid \theta)$ is given by the respective population model. The remainder of this subsection focuses on how to calculate the likelihood of each datapoint, $p(k_i, m_i \mid n_i, \Phi)$, 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. Instead, 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:

$$p\left(k_i^m \mid n_i, \Phi\right) = \binom{n_i}{k_i^m} \left(\frac{1}{\phi^m}\right)^{k_i^m} \left(1 - \frac{1}{\phi^m}\right)^{n_i - k_i^m}. \tag{10}$$

As previously described, the datapoint (k_i, m_i) corresponds to the counts and the index of the first value within the sequence $\left(k_i^1, k_i^2, \ldots, k_i^M\right)$ 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:

$$m_i = \min\left\{m \mid k_i^m \leq k_{\mathrm{CO}}\right\}, \tag{11}$$

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

$$k_i = k_i^{m_i} \quad \text{and} \quad \phi_i = \phi^{m_i}. \tag{12}$$

To avoid confusion, hereinafter we express the likelihood in terms of the plate index m_i rather than the corresponding dilution factor Φ_i, since the ordered dilution schedule Φ establishes a one-to-one correspondence between them.

Within this convention. the likelihood, $p(k_i, m_i \mid n_i, \Phi)$, is written as

$$p\left(k_i, m_i \mid n_i, \Phi\right) = p\left(k_i \mid m_i, n_i, \Phi\right) p\left(m_i \mid n_i, \Phi\right). \tag{13}$$

Once more, we proceed by showing how to calculate each factor within the product above. Starting from the last factor, $p(m_i \mid n_i, \Phi)$, we note that if the plate with index m_i is reported, we must have that both $m < m_i$,

$$p\left(m_i \mid n_i, \Phi\right) = P\left(\{k_i^{m_i} \leq k_{\mathrm{CO}}\} \mid n_i, \Phi\right) \prod_{m=1}^{m_i-1} P\left(\{k_i^m > k_{\mathrm{CO}}\} \mid n_i, \Phi\right). \tag{14}$$

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

$$P\left(\{k_i^m \leq k_{\mathrm{CO}}\} \mid n_i, \Phi\right) = \sum_{k_i^m=0}^{k_{\mathrm{CO}}} p\left(k_i^m \mid n_i, \Phi\right), \tag{15}$$

so (14) becomes

$$p\left(m_i \mid n_i, \Phi\right) = \left(\sum_{k_i^{m_i}=0}^{k_{\mathrm{CO}}} p\left(k_i^{m_i} \mid n_i, \Phi\right)\right) \left[\prod_{m=1}^{m_i-1} \left(1 - \sum_{k_i^m=0}^{k_{\mathrm{CO}}} p\left(k_i^m \mid n_i, \Phi\right)\right)\right] \tag{16}$$

with each $p\left(k_i^m \mid n_i, \Phi\right)$ given by (10).

Then, for the remaining factor in (13), $p(k_i \mid m_i, n_i, \Phi)$, 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

$$p\left(k_i \mid m_i, n_i, \Phi\right) = \frac{\binom{n_i}{k_i}\left(\frac{1}{\phi^{m_i}}\right)^{k_i}\left(1-\frac{1}{\phi^{m_i}}\right)^{n_i-k_i}}{\sum_{k'=0}^{k_{\mathrm{CO}}}\binom{n_i}{k'}\left(\frac{1}{\phi^{m_i}}\right)^{k'}\left(1-\frac{1}{\phi^{m_i}}\right)^{n_i-k'}}. \tag{17}$$

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

Thus, in order to obtain the posterior (9), we need to substitute the likelihood per datapoint, $p(k_i, m_i \mid n_i, \Phi)$ in (13), which is computed using (17) and (16).

Bayesian MAP

In Bayesian inference, we compute the posterior distribution $p(\theta \mid D)$ 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},

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} p(\theta \mid D), \tag{18}$$

with $p(\theta|D)$ given by (8) for Models 1 and 2, or (9) 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 (Paszke et al., 2019).

Results

In the previous section, we have constructed the posterior $p(\theta \mid D)$ 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).

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. 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 using naive multiplication. As an example, in Fig. 3 we generate simulated data for a mixture of Gaussians with three modes.

In Fig. 3a, 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. 3b, 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. We quantify this improvement in Appendix 2, where we analyze the reconstruction error in the mixture means and the Kullback-Leibler (KL) divergence between the reconstructed and ground-truth distributions as functions of the number of plates.

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 the likelihood for Model 3. 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 in (17), with each reported dilution matched to the corresponding index m_i, meaning the dilution reported is the m_i-th in the schedule $\phi_i = \phi^{m_i}$.

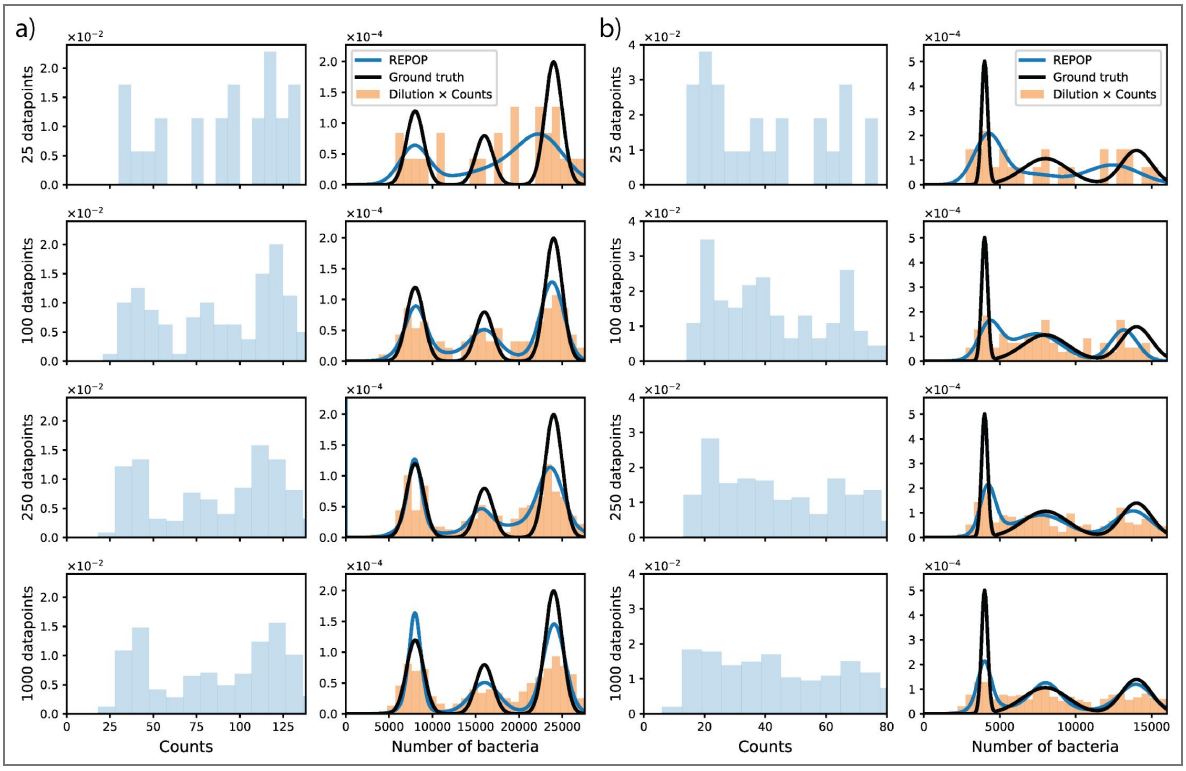


Figure 3. REPOP can reconstruct multimodal populations even when they are not directly observable by naively multiplying colony counts by dilution.

Similar to Fig. 1b and c, we show above the observed colony counts, the naïve estimate obtained by multiplying counts by the dilution factor, and the reconstruction produced by REPOP, compared to the ground truth. In both cases, the population size n is drawn from a mixture of three Gaussians, shown as the black curves. In a), the mixture has means (4000, 8000, 14000), standard deviations (200, 1500, 1000), and weights (0.25, 0.4, 0.35). In b), the mixture has means (8000, 16000, 24000), standard deviations (1000, 1000, 1000), and weights (0.3, 0.2, 0.5). Observed counts are sampled from (1), with dilution factor 200, and are shown as histograms on the left. In (a), a trimodal structure is visible in the observed count histogram. However, in b), the trimodal structure is strongly obscured by stochasticity from dilution and plating, making it difficult to discern even in large datasets. Despite this, applying REPOP to datasets of increasing size shows that, although small datasets of 25 plates may miss some modes, larger datasets allow REPOP to accurately recover the underlying multimodal structure in both cases. We quantify this behavior further in Appendix 2, where we analyze the error between the reconstructed and ground-truth means, as well as the relative entropy between the reconstructed and ground-truth distributions, as a function of the number of plates. Together, these results demonstrate REPOP's ability to infer the true population despite stochastic noise introduced by dilution and plating.

We visualize the data generated this way and show the reconstruction obtained for Model 3 in Fig. 4. 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. 4b). 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 $k_{CO} = 300$ and two possible dilution factors, $\Phi^1 = 20$ and $\Phi^2 = 200$. A sample with an original bacterial count n_i near $k_{CO}\Phi^1 = 6000$ 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. 4b. Beyond this qualitative observation, we also assess the magnitude of these errors by computing the relative deviation between the estimated and ground truth means for each component (see Appendix 2).

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. 4, by generating four 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 0.2 nL 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 0.2 nL. Instead, we take 1 μL and dilute it by factors of 10^5, 10^6, and 10^7 – further details in Appendix 3. This approach ensures that while the original vials can be assumed to have negligible variability, the small 0.2 nL 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 four components structure. However, the dataset remains relatively small, consisting of only 97 total samples (see details in Appendix 3). This dataset and the reconstructed distribution are presented in Fig. 5a.

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 750 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. 5b, demonstrate a significant improvement in capturing the underlying structure of the data.

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 (Vega and Gore, 2017; Taylor and Vega, 2024; Eisenmann, 2005). However, to derive meaningful insights into gut ecology, it is essential to distinguish true population heterogeneity from the stochasticity introduced by dilution and plating.

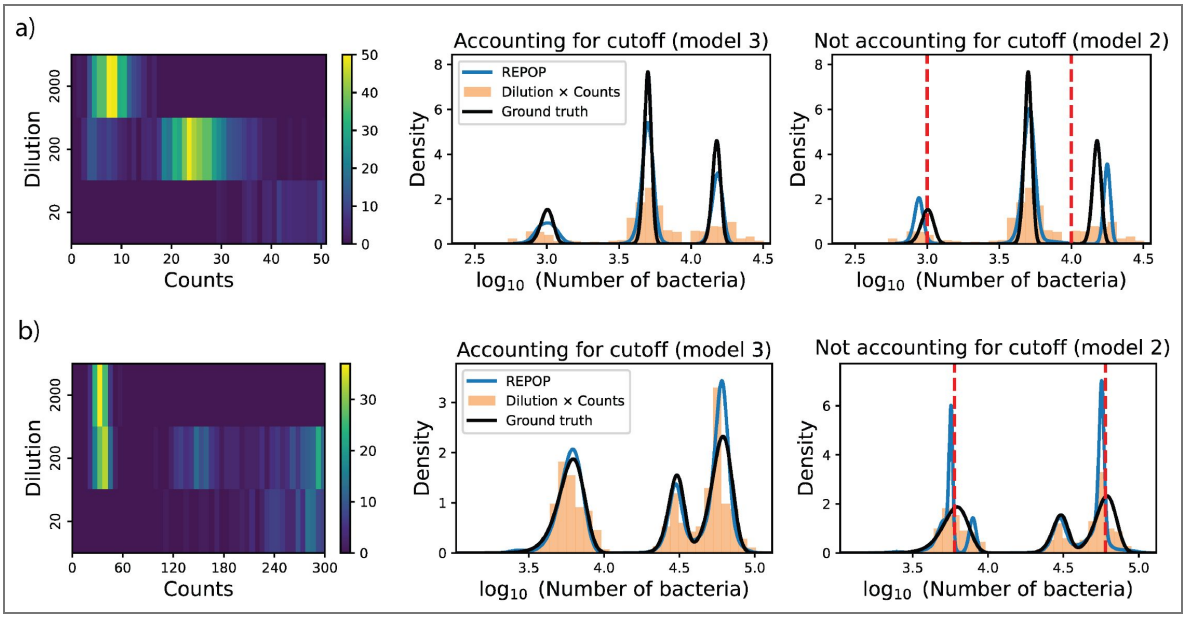


Figure 4. Not accounting for the cutoff can lead to incorrectly attributed multimodality.

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_{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 $\rho = \left(\frac{1}{6}, \frac{1}{2}, \frac{1}{3}\right)$ A more realistic cutoff k_{CO} = 300, with ground truth parameters $\bar{\mu} = (6000, 30000, 60000), \bar{\sigma} = (1200, 3600, 12000)$, and $\rho = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right)$. 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 the main text, when the true distribution has substantial probability near these points (meaning there is a non-negligible probability to find a sample with these number of bacteria close to those cutoff values) failing to account for the cutoff effects results in inaccurate reconstructions.

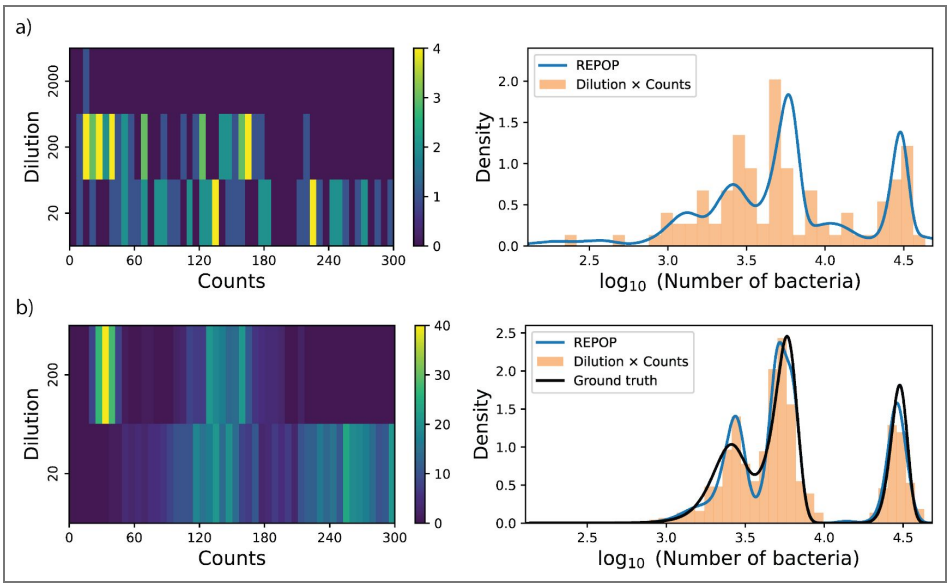


Figure 5. Reconstructing, from real experimental plate counting, the population from vials with different optical densities.

a) Distribution of observed colony counts obtained as described in text. The data is passed to the system without vial identification, aiming to reconstruct the underlying structure. b) Results from a simulated dataset with 750 datapoints (plates). The increased sample size improves the estimation of the four main components, (which appear as approximately three peaks because two components substantially overlap), yielding a more accurate reconstruction.

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 (Vega and Gore, 2017).

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 Appendix 3.

The results presented in Fig. 6 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. 6 can be explained by the heterogeneity of the host. Recently, this has been corroborated by *in vivo* experiments (Boddu et al., 2024). 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 (Yashin et al., 2002; Wu et al., 2006; Vertti-Quintero et al., 2021). Thus the variation in the population could be due to the fact that ingestion and egestion exhibit an age-related change (Bolanowski et al., 1981; Huang et al., 2004) or due to state switching (Boddu et al., 2024) in feeding and digestion, akin to genetic state switching (Pessoa et al., 2024; Kilic et al., 2023). The results in Fig. 6 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.

Resolving different species

One possible extension of plate counting is the quantification of multiple bacterial species or phenotypes within the same biological sample. In practice, this could be done when different bacteria produce distinguishable colony colors or morphologies (Maeda et al., 2017; Qamer et al., 2003; Haldeman and Amy, 1993). In such cases, the counts of different bacterial species can be obtained separately from the same sample.

To illustrate how the framework can be extended to this setting, we consider a synthetic dataset consisting of two correlated species. For each sample i, we denote by n_i^A and n_i^B the unobserved numbers of bacteria of species A and B, respectively, drawn from a joint distribution that captures their correlation. Correspondingly, the plate count measurements yield observations $\left(k_i^A, k_i^B, \phi_i\right)$. In this section, we use these synthetic data to examine how well the framework can recover the marginal population distributions of the two species from their respective plate count measurements.

In the example shown in Fig. 7, we generate a synthetic two-species population distribution with three modes. The mixture components have equal weights and means (4000, 24000), (14000, 8000), (8000, 16000), where the first and second coordinates denote the population sizes of species A and B, respectively. The covariance matrices are chosen so that the marginal distribution of species A matches the trimodal distribution used in Fig. 1c, while the marginal distribution of species B consists of three approximately equally spaced modes at 8000, 16000, and 24000, with equal variances. In the ground truth distribution, two of the mixture components include correlations between the populations of species A and B, as seen in the population samples shown by the black points in Fig. 7.

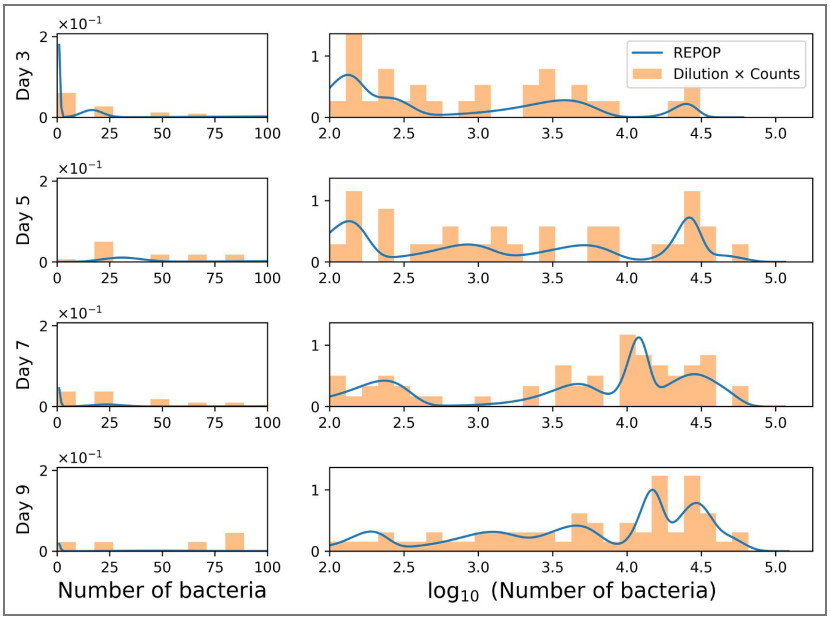


Figure 6. Results for the population of *E.coli*-fed *C. elegans*' gut at days 3, 5, 7, and 9 of adulthood.

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.* (Boddu et al., 2024) is to be inferred from this multimodality, a rigorous separation of population heterogeneity and plating stochasticity, as implemented by REPOP, is essential.

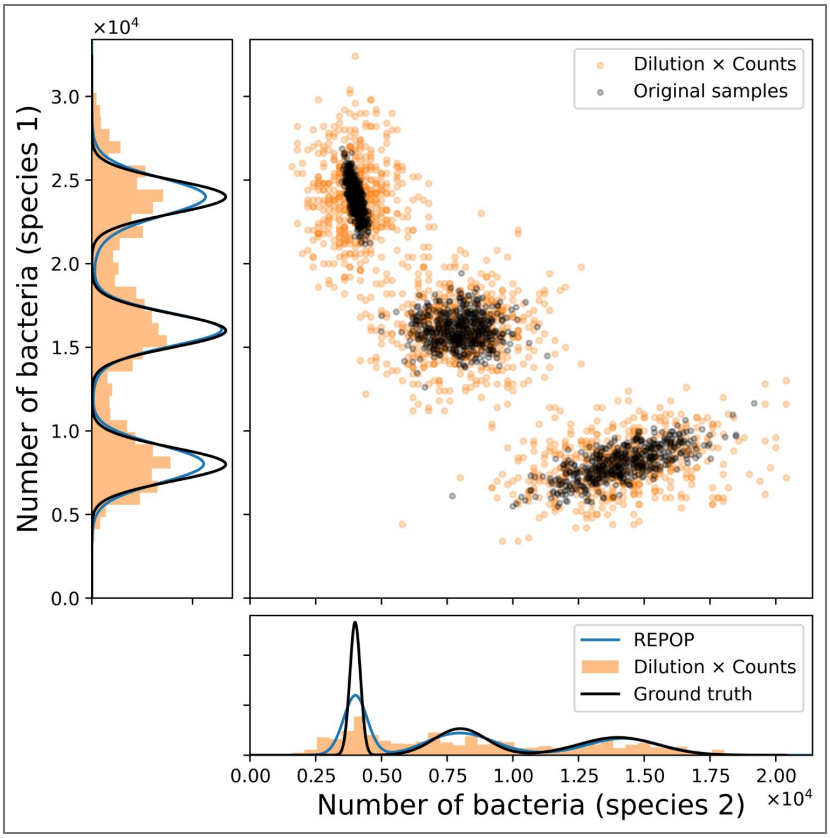


Figure 7. REPOP reconstructs marginal population distributions when individual plates distinguish two species.

A synthetic dataset of 1500 samples contains two correlated bacterial populations. Black points denote the true underlying populations (n^A, n^B), which are drawn from a joint distribution with three well-separated modes, as described in the text. Orange points denote the corresponding plate counts after a dilution factor of 200, naively reconstructed as counts multiplied by the dilution factor. Although the true population (black) modes are well separated, stochasticity introduced by dilution and plating causes the observed measurements(orange) to overlap. We assume that the two species can be distinguished on each plate, so that separate counts are obtained for species *A* and *B*. Applying REPOP separately to the counts from each species recovers the corresponding marginal population distributions, as shown in the side panels.

We then simulate plate count measurements from this population distribution using a dilution factor $\Phi_i = 200$, generating a synthetic dataset of 1500 samples. As shown in Fig. 7, the true population modes are well separated. However, after dilution and plating, the naively reconstructed observations, obtained by multiplying the plate counts by the dilution factor, are broadened and partially overlap. Nevertheless, applying REPOP separately to the species-specific counts recovers the marginal population distributions for both species.

Conclusion

In accurately estimating bacterial counts, simply multiplying counts by the dilution factor provides only a rough first approximation. As we showed here, this method alone often fails to capture the original individual counts across samples, as straightforward multiplication of colony counts by the total dilution factor typically leads to an overestimation of the spread (Fig. 1b) that may obscure multimodal behavior (Fig. 1c).

This is not merely a technical issue in quantification: it directly affects the biological interpretation of the data. For example, inferring population growth dynamics (Armitage and Jones, 2019; Martino et al., 2016; Taylor and Vega, 2024; Martini et al., 2024; Boddu et al., 2024) and learning features driving population heterogeneity across samples (Armitage and Jones, 2019; Goberna and Verdú, 2022; Pinto et al., 2022), 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 a rigorous method to obtain the distribution of populations from plate counts, including observing the correct multimodality of the population, as seen in Figs. 1 – 4. This allows the method to extract more information from each individual plate and to reconstruct population features that may be obscured in the raw observations, including multimodality and distributional differences across conditions. This setting is naturally suited to a Bayesian treatment because the quantity of interest is not the observed colony count itself, but the latent population distribution that gave rise to the observed colony counts.

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. 4, 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.

Moreover, due to REPOP's joint statistical description of dilution and plating, REPOP can be extended beyond the specific cases considered here. In particular, the same treatment of measurement stochasticity can be combined with alternative population models designed for more specific ecological hypotheses, as explored in related work by some of us (Lu et al., 2026). In that work, the joint measurement model was combined with a specific parameterized ecological model, allowing posterior inference over $p(\theta \mid D)$ and therefore uncertainty quantification over the model parameters. Specifically, the REPOP measurement model was used to provide the measurement probabilities $p(k_i, \Phi_i \mid n_i)$, as described in the section "Population versus measurement stochasticity."

Therefore, while the main implementation of REPOP focuses on finding the population distribution that maximizes the posterior, as described in (18), and is designed to remain model-agnostic by using Gaussian or mixtures of Gaussians as flexible models for the population distribution, its measurement model can also be incorporated into hypothesis-specific population models. This makes REPOP not only a practical reconstruction tool but, more importantly, a foundation for future inference problems involving plate counting data.

Furthermore, to help with computational scalability, REPOP can optionally as specified by user rely on GPU acceleration. This can become helpful as likelihood expressions in (8) and (9) require summations over all possible values of the underlying population size, which can become memory and time intensive. REPOP mitigates these computational demands through GPU parallelization, as performance can be improved substantially from CPU-only systems. Moreover, extending the framework to jointly infer multiple interacting populations in the future would increase computational cost considerably, since the relevant sums must then be performed over the product space of possible population values. In the species case, this leads to an effective quadratic scaling with population range. This quickly becomes prohibitive for larger populations, which is why, in Fig. 7, we focus on marginal rather than full joint distributions. Developing more efficient strategies for multiple species inference problems remains an important direction for future work.

Beyond that, a natural extension of the present framework is to refine the measurement model itself by incorporating additional sources of counting error beyond dilution and plating stochasticity. For instance, integrating the notion of additional error sources, such as false positive or negative rates in counting that due to *e.g*. 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 grows as we reach colony counts near 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 hypotheses (Boddu et al., 2024; Martini et al., 2024) relating observed multimodality in gut microbial content of *C. elegans*, as shown in Fig. 6, to distinct categories of hosts. What is more, REPOP can also be used to study other organisms, exhibiting similar multimodality in their microbiota (Acuña Hidalgo and Armitage, 2022; Clemmons et al., 2015; Chen et al., 2024). The central point is that identifying biologically meaningful heterogeneity requires more than collecting large datasets: it requires a framework that distinguishes true ecological structure from variability introduced by the measurement itself. Without such a framework, one risks assigning ecological significance to artifacts of dilution and plating. REPOP is designed precisely to avoid those biases while remaining flexible enough to accommodate complex and multimodal population structure.

Data availability

The plate counting real data

Acknowledgements

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.

Appendix 1

Prior selection across the unimodal and non-parametric models

When estimating the non-parametric model by MAP, we use a prior $p(\theta)$ designed to accommodate multimodal population distributions. Here, we describe this prior choice in detail.

As mentioned in the main text, this prior encompasses all of the parameters needed to construct the probability distribution for n, given by Eq.(5) in the main text. Therefore, θ must include the means $\bar{\mu} = (\mu_1, \mu_2, ...)$, standard deviations $\bar{\sigma} = (\sigma_1^2, \sigma_2^2, ...)$, and mixture weights $\bar{\rho} = (\rho_1, \rho_2, ...)$ 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. (6) 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 Pressé and Sgouralis (2023), given by:

$$\begin{aligned} \bar{\rho} &\sim \text{Dirichlet}(\bar{\xi}), \\ p(\bar{\rho}) &= \frac{1}{B(\bar{\xi})} \prod_{\alpha=1}^{A} \rho_\alpha^{\xi_\alpha - 1}, \end{aligned} \tag{19}$$

Where ξ 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 $\xi = \lambda\bar{\eta}$ with η 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 (19) will be $\langle \rho_\alpha \rangle = \frac{\eta_\alpha}{\sum_{\alpha'} \eta_{\alpha'}}$.

With the prior for ρ established, we proceed to define the prior for the means μ 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

$$\begin{aligned} \frac{\mu_\alpha}{N^*} &\sim \text{Beta}(1.1, 1.1), \\ p(\mu_\alpha) &\propto \frac{1}{N^*} \left(\frac{\mu_\alpha}{N^*}\right)^{0.1} \left(1 - \frac{\mu_\alpha}{N^*}\right)^{0.1}. \end{aligned} \tag{20}$$

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

$$\begin{aligned} \sigma_\alpha \mid \mu_\alpha &\sim \text{LogNormal}\ (\log \mu_\alpha, 1)\,, \\ p(\sigma_\alpha \mid \mu_\alpha) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_\alpha} \exp\left[-\frac{(\log \sigma_\alpha / \mu_\alpha)^2}{2}\right]. \end{aligned} \tag{21}$$

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

$$p(\theta) = p(\bar{\rho}) \prod_{\alpha=1}^{A} p(\mu_\alpha)\, p(\sigma_\alpha \mid \mu_\alpha)\,, \tag{22}$$

with $p(\bar{\rho}), p(\mu_\alpha)$, and $p(\sigma_\alpha|\mu_\alpha)$ given by (19), (20), and (21) respectively.

Appendix 2

Error metrics between reconstructed and ground-truth distributions

In Figs. 3 and 4, we showed that REPOP can reconstruct underlying population distributions obscured by dilution and plating stochasticity. Here, we make this comparison quantitative using two complementary error metrics. The first measures how well the inferred mixture components recover the locations of the ground-truth modes. The second compares the full reconstructed and ground-truth distributions using the KL divergence.

To compare two distributions, $p(n_i|\theta)$ and $p(n_i|\theta_{GT})$, in terms of how well their component means match, we use the relative error between the means defined as

$$M(\theta,\theta_{\mathrm{GT}}) = \sum_{\alpha=1}^{A} \rho_\alpha \min_{\alpha_{\mathrm{GT}}} \left| \frac{\mu_\alpha - \mu_{\alpha_{\mathrm{GT}}}}{\mu_{\alpha_{\mathrm{GT}}}} \right|. \tag{23}$$

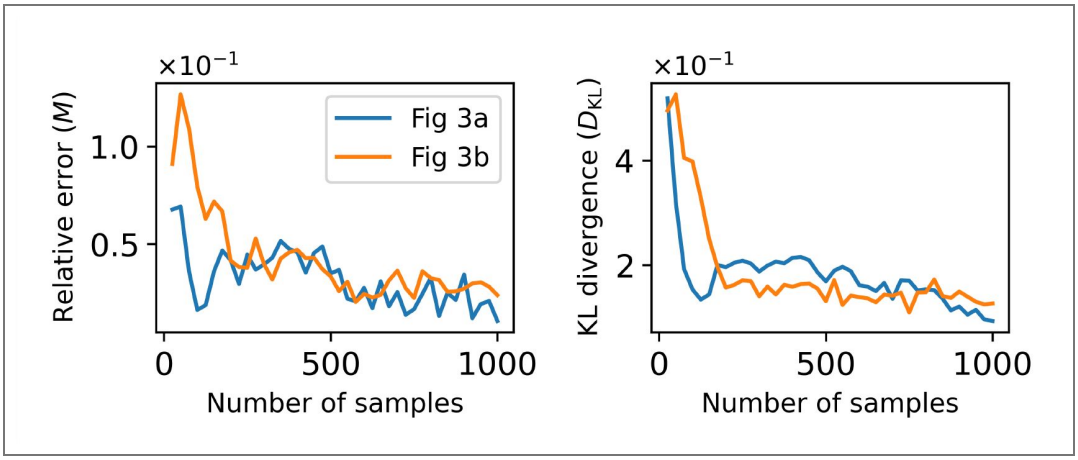
Here, $p(n|\theta)$ is the reconstructed distribution (through REPOP) and $p(n|\theta_{\mathrm{GT}})$ is the groundtruth distribution. The index α labels the components of the reconstructed distribution, while α_{GT} labels the components of the ground-truth distribution. The quantity inside the absolute value is the relative error between the reconstructed component mean μ_α and a ground-truth component mean $\mu_{\alpha_{\mathrm{GT}}}$. We take the minimum over α_{GT} to account for cases in which component labels do not match across the two distributions. The weights ρ_α are the reconstructed mixture weights, so components with larger inferred probability mass contribute more strongly to the metric.

To compare the full reconstructed and ground-truth distributions, rather than only the locations of their modes, we use a standard metric for comparing probability distributions: the relative entropy, also known as the Kullback-Leibler (KL) divergence. For each reconstructed distribution $p(n|\theta)$ and corresponding ground truth distribution $p(n|\theta_{\mathrm{GT}})$, we define

$$D_{\mathrm{KL}}\left[p(n\mid\theta_{\mathrm{GT}})\,\|p(n\mid\theta)\right] = \sum_n p(n\mid\theta_{\mathrm{GT}}) \log \frac{p(n\mid\theta_{\mathrm{GT}})}{p(n\mid\theta)}. \tag{24}$$

The sum is taken over the population size n. Here, $p(n \mid \theta_{\mathrm{GT}})$ is the known ground truth mixture distribution used to generate the simulated data, while $p(n|\theta)$ is the mixture distribution reconstructed by REPOP. Roughly, KL divergence measures how much information is lost when the reconstructed distribution is used as an approximation to the ground truth distribution. In the figure below, we show how both metrics decay as the dataset size increases for the same examples shown in Fig. 3.

Analogously, for the dataset shown in Fig. 4a, the relative error M in (23) is 1.50% when the cutoff in the total number of colonies is accounted for (model 3). When the cutoff is ignored (model 2), the error increases substantially to 7.47%. The same trend is observed for the KL divergence in (24), which increases from 0.13 when the cutoff is included to 1.73 when the cutoff is ignored. Similarly, for the dataset shown in Fig. 4b, the relative error is 2.07% when the cutoff is considered, compared with 11.5% when it is not. The KL divergence likewise increases from 0.03 to 0.64 when the cutoff is ignored. Together, these results show that accounting for the cutoff is necessary to accurately reconstruct both the mode locations and the full population distribution. Full details of these calculations are available in our GitHub repository Pessoa (2025).



Appendix 2—figure 1. Decay of error metrics as the dataset size increases for the distributions shown in Fig. 3. We show the relative error defined in (23) (left) and the KL divergence defined in (24) (right) for the distributions in Fig. 3a and Fig. 3b. Both metrics decrease as the number of plates increases, indicating improved agreement between the REPOP reconstruction and the ground-truth.

Appendix 3

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. elegans culture

The temperature-sensitive, reproductively sterile strain AU37 of *Caenorhabditis elegans* (*C. elegans*) was cultivated and synchronized according to standard protocols (Eisenmann, 2005). Following the protocol described by Gore and Vega (Vega and Gore, 2017), 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 Vega and Gore (2017). 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.

Additional information

Funding

Funder	Grant reference number	Author
HHS | National Institutes of Health (NIH)	35GM148237)	Steve Presse
DOW | USA | AFC | CCDC | Army Research Office (ARO)	W911NF-23-1-0304)	Steve Presse
National Science Foundation (NSF)	2310610	Steve Presse

Author ORCID iDs

Pedro Pessoa: https://orcid.org/0000-0002-6258-5679
Stanimir Asenov Tashev: https://orcid.org/0000-0002-0119-4412
Douglas P Shepherd: https://orcid.org/0000-0001-9087-0832

References

Acuña Hidalgo B, Armitage SA (2022) Host resistance to bacterial infection varies over time, but is not affected by a previous exposure to the same pathogen. *Frontiers in Physiology* **13**:860875 https://doi.org/10.3389/fphys.2022.860875 | PubMed

Albaradei SA, Napolitano F, Uludag M, Thafar M, Napolitano S, Essack M, Bajic VB, Gao X (2020) Automated Counting of Colony Forming Units Using Deep Transfer Learning From a Model for Congested Scenes Analysis. *IEEE Access* **8**:164340 https://doi.org/10.1109/ACCESS.2020.3021656

Armitage DW, Jones SE (2019) How sample heterogeneity can obscure the signal of microbial interactions. *The ISME Journal* **13**:2639 https://doi.org/10.1038/s41396-019-0463-3 | PubMed

Bartram J, Cotruvo J, Exner M, Fricker C, Glasmacher A (2003) Heterotrophic plate counts and drinking-water safety. Emerging issues in water and infectious disease series. Genève, Switzerland: World Health Organization.

Blanchet FG, Cazelles K, Gravel D (2020) Co-occurrence is not evidence of ecological interactions. *Ecology Letters* **23**:1050 https://doi.org/10.1111/ele.13525 | PubMed

Boddu SS, Martini KM, Nemenman I, Vega NM (2024) Variance in C. elegans gut bacterial load suggests complex host-microbe dynamics. *bioRxiv* https://doi.org/10.1101/2024.12.09.627557

Bolanowski MA, Russell RL, Jacobson LA (1981) Quantitative measures of aging in the nematode Caenorhabditis elegans. I. Population and longitudinal studies of two behavioral parameters. *Mechanisms of ageing and development* **15**:279 https://doi.org/10.1016/0047-6374(81)90136-6 | PubMed

Bryan IV JS, Sgouralis I, Pressé S (2020) Inferring effective forces for Langevin dynamics using Gaussian processes. *The Journal of Chemical Physics* **152**:124106 https://doi.org/10.1063/1.5144523 | PubMed

Chen JZ, Kwong Z, Gerardo NM, Vega NM (2024) Ecological drift during colonization drives within-host and between-host heterogeneity in an animal-associated symbiont. *PLoS Biology* **22**:e3002304 https://doi.org/10.1371/journal.pbio.3002304 | PubMed

Clemmons AW, Lindsay SA, Wasserman SA (2015) An effector peptide family required for Drosophila Toll-mediated immunity. *PLoS Pathogens* **11**:e1004876 https://doi.org/10.1371/journal.ppat.1004876 | PubMed

Corry JEL, Curtis GDW, Baird RM (2011) *Handbook of Culture Media for Food and Water Microbiology* The Royal Society of Chemistry.

Davey H, Guyot S (2020) Estimation of Microbial Viability Using Flow Cytometry. *Current Protocols in Cytometry* **93**:e72 https://doi.org/10.1002/cpcy.72 | PubMed

De Martino D (2017) Maximum entropy modeling of metabolic networks by constraining growth-rate moments predicts coexistence of phenotypes. *Physical Review E* **96**:060401 https://doi.org/10.1103/physreve.96.060401 | PubMed

Eisenmann DM (2005) *WormBook* http://www.wormbook.org/

Geissmann Q (2013) OpenCFU, a New Free and Open-Source Software to Count Cell Colonies and Other Circular Objects. *PLOS One* **8**:1 https://doi.org/10.1371/journal.pone.0054072 | PubMed

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

Haldeman DL, Amy PS (1993) Diversity within a Colony Morphotype: Implications for Ecological Research. *Applied and Environmental Microbiology* **59**:933 https://doi.org/10.1128/aem.59.3.933-935.1993 | PubMed

Hansen SJZ, Tang P, Kiefer A, Galles K, Wong C, Morovic W (2020) Droplet Digital PCR Is an Improved Alternative Method for High-Quality Enumeration of Viable Probiotic Strains. *Frontiers in Microbiology* **10**:3025 https://doi.org/10.3389/fmicb.2019.03025 | PubMed

Huang C, Xiong C, Kornfeld K (2004) Measurements of age-related changes of physiological processes that predict lifespan of Caenorhabditis elegans. *Proceedings of the National Academy of Sciences* **101**:8084 https://doi.org/10.1073/pnas.0400848101 | PubMed

International Organization for Standardization (2013) 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.

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

Kalli M, Griffin JE, Walker SG (2009) Slice sampling mixture models. *Statistics and Computing* **21**:93 https://doi.org/10.1007/s11222-009-9150-y

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

Kilic Z, Schweiger M, Moyer C, Shepherd D, Pressé S (2023) Gene expression model inference from snapshot RNA data using Bayesian non-parametrics. *Nature Computational Science* **3**:174 https://doi.org/10.1038/s43588-022-00392-0 | PubMed

Kilic Z, Sgouralis I, Pressé S (2021) Generalizing HMMs to Continuous Time for Fast Kinetics: Hidden Markov Jump Processes. *Biophysical Journal* **120**:409 https://doi.org/10.1016/j.bpj.2020.12.022 | PubMed

Lu CY, Tashev SA, Pessoa P, Kruithoff R, Shepherd DP, Presse S (2026) Stochastic colonization and host-to-host transmission shape gut bacterial variability. *bioRxiv* https://doi.org/10.64898/2026.05.11.724410 | PubMed

Maeda Y, Dobashi H, Sugiyama Y, Saeki T, Tk Lim, Harada M, Matsunaga T, Yoshino T, Tanaka T (2017) Colony fingerprint for discrimination of microbial species based on lensless imaging of microcolonies. *PLOS One* **12**:1 https://doi.org/10.1371/journal.pone.0174723 | PubMed

Mao J, Lu T (2016) Population-Dynamic Modeling of Bacterial Horizontal Gene Transfer by Natural Transformation. *Biophysical Journal* **110**:258 https://doi.org/10.1016/j.bpj.2015.11.033 | PubMed

Marine E, Milner DS, Lambert C, Sockett RE, Pos KM (2020) A novel method to determine antibiotic sensitivity in Bdellovibrio bacteriovorus reveals a DHFR-dependent natural trimethoprim resistance. *Scientific Reports* **10**:5315 https://doi.org/10.1038/s41598-020-62014-x | PubMed

Martini K, Boddu SS, Nemenman I, Vega NM (2024) Maximum Likelihood Estimators For Colony Forming Units. *Microbiology Spectrum* **12**:e03946 https://doi.org/10.1128/spectrum.03946-23 | PubMed

Martino DD, Capuani F, Martino AD (2016) Growth against entropy in bacterial metabolism: the phenotypic trade-off behind empirical growth rate distributions inE.coli. *Physical Biology* **13**:036005 https://doi.org/10.1088/1478-3975/13/3/036005 | PubMed

Moore LS, Stolovicki E, Braun E (2013) Population Dynamics of Metastable Growth-Rate Phenotypes. *PLoS ONE* **8**:e81671 https://doi.org/10.1371/journal.pone.0081671 | PubMed

Muntoni AP, Braunstein A, Pagnani A, De Martino D, De Martino A (2022) Relationship between fitness and heterogeneity in exponentially growing microbial populations. *Biophysical Journal* **121**:1919 https://doi.org/10.1016/j.bpj.2022.04.012 | PubMed

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, *et al.* (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems. **32** pp. 8024

Pecht T, Aschenbrenner AC, Ulas T, Succurro A (2019) Modeling population heterogeneity from microbial communities to immune response in cells. *Cellular and Molecular Life Sciences* **77**:415 https://doi.org/10.1007/s00018-019-03378-w | PubMed

Pessoa P (2025) REPOP -Reconstructing population from plates. GitHub. https://github.com/LabPresse/REPOP

Pessoa P, Schweiger M, Pressé S (2024) Avoiding matrix exponentials for large transition rate matrices. *The Journal of Chemical Physics* **160**:094109 https://doi.org/10.1063/5.0190527 | PubMed

Pessoa P, Schweiger M, Xu LWQ, Manha T, Saurabh A, Camarena JA, Pressé S (2025) Avoiding subtraction and division of stochastic signals using normalizing flows: NFdeconvolve. *iScience* **28**:113823 https://doi.org/10.1016/j.isci.2025.113823 | PubMed

Pinto S, Benincà E, van Nes EH, Scheffer M, Bogaards JA (2022) Species abundance correlations carry limited information about microbial network interactions. *PLoS Computational Biology* **18**:e1010491 https://doi.org/10.1371/journal.pcbi.1010491 | PubMed

Pressé S, Ghosh K, Phillips R, Dill KA (2010) Dynamical fluctuations in biochemical reactions and cycles. *Physical Review E* **82**:031905 https://doi.org/10.1103/PhysRevE.82.031905 | PubMed

Pressé S, Sgouralis I (2023) *Data Modeling for the Sciences: Applications, Basics, Computations* Cambridge University Press.

Qamer S, Sandoe JAT, Kerr KG (2003) Use of Colony Morphology To Distinguish Different Enterococcal Strains and Species in Mixed Culture from Clinical Specimens. *Journal of Clinical Microbiology* **41**:2644 https://doi.org/10.1128/jcm.41.6.2644-2646.2003 | PubMed

Sgouralis I, Xu LWQ, Jalihal AP, Kilic Z, Walter NG, Pressé S (2024) BNP-Track: a framework for superresolved tracking. *Nature Methods* **21**:1716 https://doi.org/10.1038/s41592-024-02349-9 | PubMed

Smith H, Brown A (2021) *Benson's microbiological applications laboratory manual* (15) McGraw-Hill Education.

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

Sun Q, Huynh K, Muratovic D, Gunn NJ, Zelmer AR, Solomon LB, Atkins GJ, Yang D (2024) Beyond the Colony-Forming-Unit: Rapid Bacterial Evaluation in Osteomyelitis. *eLife* **13**:RP93698 https://doi.org/10.7554/elife.93698.2

Taimur MM, Acevedo R, Sklar LA, Thomsen K, Tegos GP, Camper AK (2020) Rapid identification of bacterial viability, culturable and non-culturable stages by using flow cytometry. *Protocol Exchange* https://doi.org/10.21203/rs.3.pex-1153/v1

Tamargo A, Molinero N, Reinosa JJ, Alcolea-Rodriguez V, Portela R, Bañares MA, Fernández JF, Moreno-Arribas MV (2022) PET microplastics affect human gut microbiota communities during simulated gastrointestinal digestion, first evidence of plausible polymer biodegradation during human digestion. *Scientific Reports* **12**:528 https://doi.org/10.1038/s41598-021-04489-w | PubMed

Tavakoli M, Jazani S, Sgouralis I, Heo W, Ishii K, Tahara T, Pressé S (2020) Direct Photon-by-Photon Analysis of Time-Resolved Pulsed Excitation Data using Bayesian Nonparametrics. *Cell Reports Physical Science* **1**:100234 https://doi.org/10.1016/j.xcrp.2020.100234 | PubMed

Taylor MN, Spandana Boddu S, Vega NM (2022) Using Single-Worm Data to Quantify Heterogeneity in Caenorhabditis elegans-Bacterial Interactions. *Journal of Visualized Experiments* **185**:e64027 https://doi.org/10.3791/64027 | PubMed

Taylor MN, Vega NM (2024) Sample pooling inflates error rates in between-sample comparisons: an empirical investigation of the statistical properties of count-based data. *bioRxiv* https://doi.org/10.1101/2022.07.25.501406

Tracey H, Coates N, Hulme E, John D, Michael DR, Plummer SF (2023) Insights into the enumeration of mixtures of probiotic bacteria by flow cytometry. *BMC Microbiology* **23**:48 https://doi.org/10.1186/s12866-023-02792-2 | PubMed

Van Nevel S, Koetzsch S, Proctor CR, Besmer MD, Prest EI, Vrouwenvelder JS, Knezev A, Boon N, Hammes F (2017) Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring. *Water Research* **113**:191 https://doi.org/10.1016/j.watres.2017.01.065 | PubMed

Veening JW, Smits WK, Kuipers OP (2008) Bistability, Epigenetics, and Bet-Hedging in Bacteria. *Annual Review of Microbiology* **62**:193 https://doi.org/10.1146/annurev.micro.62.081307.163002 | PubMed

Vega NM, Gore J (2017) Stochastic assembly produces heterogeneous communities in the Caenorhabditis elegans intestine. *PLOS Biology* **15**:e2000633 https://doi.org/10.1371/journal.pbio.2000633 | PubMed

Vertti-Quintero N, Berger S, Casadevalli Solvas X, Statzer C, Annis J, Ruppen P, Stavrakis S, Ewald CY, Gunawan R, DeMello AJ (2021) Stochastic and Age-Dependent Proteostasis Decline Underlies Heterogeneity in Heat-Shock Response Dynamics. *Small* **17**:2102145 https://doi.org/10.1002/smll.202102145 | PubMed

Vial SL, Doerscher DR, Hedberg CW, Stone WA, Whisenant SJ, Schroeder CM (2019) 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**:1761 https://doi.org/10.4315/0362-028x.jfp-19-241 | PubMed

Walker SG (2007) Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics-Simulation and Computation* **36**:45 https://doi.org/10.1080/03610910601096262

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

Weitzel MLJ, Vegge CS, Pane M, Goldman VS, Koshy B, Porsby CH, Burguière P, Schoeni JL (2021) Improving and Comparing Probiotic Plate Count Methods by Analytical Procedure Lifecycle Management. *Frontiers in Microbiology* **12**:693066 https://doi.org/10.3389/fmicb.2021.693066 | PubMed

Werner J, Pietsch T, Hilker FM, Arndt H (2022) Proceedings of the National Academy of Sciences. *Intrinsic nonlinear dynamics drive single-species systems* **119**:e2209601119 https://doi.org/10.1073/pnas.2209601119 | PubMed

Wilkinson MG (2018) Flow cytometry as a potential method of measuring bacterial viability in probiotic products: A review. *Trends in Food Science and Technology* **78**:1 https://doi.org/10.1016/j.tifs.2018.05.006

Wu D, Rea SL, Yashin AI, Johnson TE (2006) Visualizing hidden heterogeneity in isogenic populations of C. elegans. *Experimental gerontology* **41**:261 https://doi.org/10.1016/j.exger.2006.01.003 | PubMed

Yashin AI, Cypser JW, Johnson TE, Michalski AI, Boyko SI, Novoseltsev VN (2002) 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**:B83 https://doi.org/10.1093/gerona/57.3.b83 | PubMed

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

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors developed a novel theoretical/computational procedure to count bacterial populations without introducing artificial randomness effects due to dilution. Surprisingly, this very important aspect of studies of bacterial systems has been overlooked. The proposed method provides a simple and transparent approach to eliminate the randomness of bacterial accounting procedures, allowing now to fully concentrate on the intrinsic effects of the studied systems.

Strengths:

A very simple and clear procedure is introduced and explained in full detail. This elegant approach finds an excellent compromise between mathematical rigor and computational efficiency, which is important for practical applications. The provided examples are convincing beyond a doubt, clearly indicating the potential strong impact of the proposed framework. Various complications and possible issues are also discussed and analyzed. This seems to be a very powerful novel method that should significantly advance the analysis of complex biological systems.

Weaknesses:

The only minor weakness that I found is the assumption of independence of bacterial species, which is expressed as the well-stirred approximation. One could imagine that bacterial species might cooperate, leading to non-uniform distributions that are real. How to distinguish such situations?

I believe that this method can be extended to determine if this is the case or not before the application. For example, if the bacteria species are independent of each other and one can use the binomial distributions - then the Fano factor would be proportional to the overall relative fraction of bacterial species. Maybe a simple test can be added to test it before the application of REPOP. However, I believe that this is a minor issue.

Comments on revised version.

I am satisfied with the correction proposed by the authors. The method is already quite impressive, and there is no need to complicate it at this stage.

https://doi.org/10.7554/eLife.107122.2.sa2

Reviewer #2 (Public review):

I appreciate the thorough responses from the authors, which address my concerns. The expansion of Appendix B as well as the addition of text discussing how the REPOP method interacts with data collection efforts are very useful. These new sections show that relative error decreases with increasing samples, as expected, yet these error metrics, including KL divergence, describing the fit of the full distributions not just the modes, drop off fairly

quickly with increasing number of samples showing that REPOP likely minimizes discrepancies between estimated and true distributions even at lower sampling efforts.

Additionally, the extension of the REPOP method to the quantification of multiple bacterial species or phenotypes shows the potential utility of the method in contexts beyond basic plate counts. Between this example and the additional information on how to implement REPOP, I believe this workflow will be attractive and accessible to the audience.

https://doi.org/10.7554/eLife.107122.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *(R1C1) The only minor weakness that I found is the assumption of independence of bacterial species, which is expressed as the well-stirred approximation. One could imagine that bacterial species might cooperate, leading to non-uniform distributions that are real. How to distinguish such situations?*
>
> *I believe that this method can be extended to determine if this is the case or not before the application. For example, if the bacteria species are independent of each other and one can use the binomial distributions, then the Fano factor would be proportional to the overall relative fraction of bacterial species. Maybe a simple test can be added to test it before the application of REPOP. However, I believe that this is a minor issue.*

This is an interesting point raised by the reviewer.

First, we need to clarify an important point: we do not make a well-stirred assumption. Samples can be drawn and plated from any region of space however small and that region's population can be quantified using our method. The stirring only occurs after we collect a sample in order to dilute the contents and pour the solution homogeneously over the plate.

As such, learning multiple independent species is possible and not impacted by the dilution ("well-stirred" assumption). In the new first paragraph of the methods section, we made it clear that this assumption concerns the dilution process. REPOP is designed to recover the true underlying heterogeneity in species abundance (even from limited data) by leveraging a Bayesian framework that remains valid regardless of whether species are independent or correlated.

If the method is applied to multiple species as currently implemented, REPOP can recover the marginal distribution of each species, provided that the species are either selectively cultured or produce sufficiently distinguishable colonies on the same plate. To demonstrate this, we have added a new Results subsection with a synthetic two-species example in which the species abundances are correlated across samples.

However, in order to learn the joint distribution and capture correlations between species within samples, the method would need to be extended. At present, in Eq. 5 we sum the likelihood over all values of n, using a data-driven cutoff (twice the largest naïvely estimated count times the dilution factor). Extending this to multiple species adding up to (n_1,n_2), while retain the generality of the method, would require quadratically scaling memory with this cutoff in the population number. For this reason while we comment on this in the new paragraph in the conclusion, it is not implemented as part of REPOP.

> ***Reviewer #2 (Public review):***
>
> *(R2C1) A more thorough discussion of when and by how much estimated microbial population abundance distributions differ from the ground truth would be helpful in determining the best practices for applying this method. Not only would this allow researchers to understand the sampling effort necessary to achieve the results presented here, but it would also contextualize the experimental results presented in the paper. Particularly, there is a disconnect between the discussion of the large sample sizes necessary to achieve accurate multimodal distribution estimates and the small sample sizes used in both experiments.*

That is a great suggestion from the reviewer. To address it, we expanded Appendix B. We know report (1) the relative error in the estimated means (as already done for Fig. 4 formally 3), and (2) the Kullback-Leibler (KL) divergence between the reconstructed and ground-truth distributions. These metrics will are show as a function of the size of the dataset, for the examples in Fig 3. enabling a direct assessment of how the sampling effort affects the precision of the inference.

That said, we now highlight in the Conclusion that, by explicitly modeling the dilution process within a Bayesian framework, REPOP extracts the maximum information available from each individual sample at a given sample size. This strategy therefore enables more accurate inference with fewer measurements, which is particularly important in applications such as plate counting, where data acquisition is labour-intensive.

> ***Reviewer #3 (Public review):***
>
> *(R3C1) While the study is promising, there are a few areas where the paper could be strengthened to increase its impact and usability. First, the extent to which dilution and plating introduce noise is not fully explored. Could this noise significantly affect experimental conclusions? And under what conditions does it matter most? Does it depend on experimental design or specific parameter values? Clarifying this would help readers appreciate when and why REPOP should be used.*

We agree with the reviewer that this is an important point, and we expanded Appendix B to include a quantitative analysis using simulated data (Fig. 3, formely 2), reporting both relative error and KL divergence as a function of dataset size. This complements our response to R2C1 clarifying when REPOP offers the greatest benefit.

In addition, we will expand the discussion on how modeling dilution noise becomes essential when learning population dynamics. In particular, we emphasize? the role of Model 3, especially relevant when working with multiple plates and approaching the asymptotic regime; an aspect that was alluded to in Fig. 3 but not fully explored.

> *(R3C2) Second, more practical details about the tool itself would be very helpful. Simply stating that it is available on GitHub may not be enough. Readers will want to know what programming language it uses, what the input data should look like, and ideally, see a step-by-step diagram of the workflow. Packaging the tool as an easy-to-use resource, perhaps even submitting it to CRAN or including example scripts, would go a long way, especially since microbiologists tend to favor user-friendly, recipe-like solutions.*

In the new paragraphs of the introduction, we made clear that REPOP is written in Python (PyTorch), installable via pip, and designed for ease of use. We are also expanding the tutorials to include clearer guidance on data formatting and common workflows. The new workflow figure (Fig 2) better illustrates the full process.

> *(R3C3) Third, it would be great to see the method tested on existing datasets, such as those from Nic Vega and Jeff Gore (2017), which explore how colonization frequency impacts abundance fluctuation distributions. Even if the general conclusions remain unchanged, showing that REPOP can better match observed patterns would strengthen the paper's real-world relevance.*

We thank the reviewer for this interesting suggestion. We agree that applying REPOP to additional existing datasets would make REPOP's relevance clearer. However, the Vega and Gore datasets lack the information required. REPOP requires the plate count measurement process to be specified, including the dilution factors used for each measurement. Furthermore, we can leverage on additional information about the experimental procedure when the colony cutoffs and dilution schedules used are reported. Without the dilution factors, the likelihood connecting the observed colony counts to the underlying population size is not possible. We hope this clarification will help make future datasets made available publicly more useful for purposes of uncertainty propagation.

> *(R3C4) Lastly, it would be helpful for the authors to briefly discuss the limitations of their method, as no approach is without its constraints. Acknowledging these would provide a more balanced and transparent perspective.*

We agree with the reviewer. We have added two new paragraphs to the conclusion highlighting important current constraints and future development directions of the framework. In particular, we now discuss that, in its present implementation, REPOP focuses on the population distribution that maximizes the posterior, rather than returning posterior uncertainty over the reconstructed distributions themselves. We also note the computational demands of the method, making GPU acceleration highly beneficial and more complex multi-population inference computationally challenging. This discussion synthesizes points raised throughout our response to R1C1 and the reviewers and provides a more balanced perspective on the current scope of the method.

https://doi.org/10.7554/eLife.107122.2.sa0