Inferring Within-Host Bottleneck Size: A Bayesian Approach

R. Dybowski[1,*], O. Restif[1], D.J. Price[1], P. Mastroeni[1]

[1] Department of Veterinary Medicine, University of Cambridge,
Madingley Road, Cambridge, CB3 0ES, UK
*Corresponding author: rd460@cam.ac.uk

Abstract

Recent technical developments in microbiology have led to new discoveries on the within-host dynamics of bacterial infections in laboratory animals. In particular, they have highlighted the importance of stochastic bottlenecks at the onset of invasive disease.

A number of approaches exist for bottleneck-size estimation with respect to within-host bacterial infections; however, some are more appropriate than others under certain circumstances. A Bayesian comparison of several approaches is made in terms of the availability of isogenic multitype bacteria (e.g., WITS), knowledge of post-bottleneck dynamics, and the suitability of dilution with monotype bacteria. A sampling approach to bottleneck-size estimation is also introduced.

The results are summarised by a guiding flowchart, which we hope will promote the use of quantitative models in microbiology to refine the analysis of animal experiment data.

Keywords

Bottlenecks
Bayesian inference
Wildtype isogenic tagged strains (WITS)
Salmonella

1 Introduction

The outcome of an infection process is usually underlain by a fine balance between the virulence mechanisms of the pathogen and the resistance of the host. The presence of physical, immunological or therapeutic barriers poses constraints to the ability of bacteria to divide and disseminate within the host organism. Suppose that we have a population $\mathcal{P}_0$ of bacteria. When a subset of $\mathcal{P}_0$ is inactivated by an antibiotic or an immune response, or subject to an anatomical barrier to transmission, this can result in a substantially smaller population $\mathcal{P}_1$ (commonly known as a *bottleneck*) and, after the occurrence of the bottleneck, the bacteria of $\mathcal{P}_1$ can grow to form a new population $\mathcal{P}_2$.

Understanding the site, nature and size of bottlenecks in infectious disease processes is important to rationally design prevention strategies and treatments to control the spread of the infection within a given host. In fact, classes of vaccines and therapeutic compounds differ significantly in how they restrain an infection process with respect to the control, for example, of microbial killing or division rates and the spread within and between organs.

The inability to monitor bacterial dynamics in real time within a living host hindered the detection of bottlenecks (and more generally, a proper quantification of infection dynamics) throughout the 20th century (Smith, 2000). This has changed dramatically with the development of genetic engineering technology allowing the identification of multiple sub-populations of bacteria in the last 15 years. In particular, bottlenecks have been the focus of a number of articles, and Abel et al. (2015) provide a biologically motivated introduction to bottlenecks. Specific experimental studies that have shown bottlenecks using isogenic tagged strains include Grant et al. (2008) in the early stage of salmonellosis in mice, Schwartz et al. (2011) in urinary tract *Escherichia coli* infection in mice, Lowe et al. (2013) during *Bacillus anthracis* colonisation in mice, Kaiser et al. (2014) with *Salmonella* Typhimurium crossing the intestinal barrier in mice, Lim et al. (2014) also with *Salmonella*, Gerlini et al. (2014) and Kono et al. (2016) with invasive *Streptococcus pneumoniae*, and Abel et al. (2015) with *Vibrio cholerae* in the intestinal tract. However, in spite of the number of studies that have been conducted involving within-host bacterial bottlenecks, there has not been a unified study of the various analytical methods used for tagged/multitype experimental studies, which is the motivation of this study.

Table 1 lists the symbols used in this article, and Figure 2 provides an overview of the various methods described.

[Table 1 about here.]

2 Monotype populations

2.1 Posterior bottleneck distributions

Of interest is estimating the size of a bottleneck given observations made after the bottleneck, and possibly also before it, in a Bayesian framework. This can be expressed as the posterior probabilities

$$p(\text{bottleneck size}|\text{observation after bottleneck}) \tag{1}$$

and

$$\begin{aligned} p(\text{bottleneck size}|&\text{observation before bottleneck \&} \\ &\text{observation after bottleneck}) \ . \end{aligned} \tag{2}$$

The benefit of taking the Bayesian approach of using posterior probability distributions is that such distributions not only give estimates for the most probable bottleneck sizes (in terms of the modes of the distribution) but they also express the uncertainty through the variance of the distributions.

[Figure 1 about here.]

Abel et al. (2015) estimated bottleneck size with respect to multitype populations by equating it to the effective population size as estimated by Krimbas and Tsakas (1971), which uses the standardised covariance of allele frequency due to a bottleneck. Although Abel et al. found this approach successful in the context of the within-host dynamics of *Vibro cholera*, both Pamilo and Varvio-Aho (1980) and Sourdis and Krimbas (1980) caution that bottleneck-size estimation by the Krimbas-Tsakas method can be unreliable unless sample size is sufficiently large.

Let n_1 be the size of a bottleneck $\mathcal{P}_1$ and n_2 the size of a post-bottleneck population $\mathcal{P}_2$ after it (Figure 1). For the posterior distribution $p(n_1|n_2)$, Bayes' theorem gives

$$p(n_1|n_2) \propto p(n_1)p(n_2|n_1) \ ,$$

and if we assume that the priors $p(n_1)$ are equally likely we then have the expression

$$p(n_1|n_2) = \frac{p(n_2|n_1)}{\sum_{n_1} p(n_2|n_1)} \ . \tag{3}$$

Consider $p(n_2|n_1)$ for (3). Suppose that we can simulate the occurrence of n_2 resulting stochastically from n_1 after time interval Δt (Figure 1) according to a set of parameters $\boldsymbol{\theta}$ for the dynamics; for example, in the context of the birth-death-migration process shown in Figure 5 (Grant et al., 2008; Kaiser et al., 2014; Coward et al., 2014; Dybowski et al., 2015). Such a simulation can be achieved by a Gillespie stochastic simulation algorithm (Gillespie, 1997). If we obtain a finite number of such simulations,

$$(n_2)_1, (n_2)_2, \ldots, (n_2)_m \overset{i.i.d}{\sim} \mathit{Gillespie}(n_1, \Delta t, \boldsymbol{\theta}) \ ,$$

our task is then to estimate the probability mass function $p(n_2|n_1, \boldsymbol{\theta})$ from $\{(n_2)_1, (n_2)_2, \ldots, (n_2)_m\}$, which can then be used for (3).

Estimating $p(n_2|\cdot,\cdot)$ from the sparse sample $\{(n_2)_1, (n_2)_2, \ldots, (n_2)_m\}$ can be attempted using local polynomial smoothing (Simonoff, 1996). A recent development in this area is the local polynomial smoothing proposed by Jacob and Oliveira (2011), which is effective for small sample sizes.

Let $\omega = \{(n_2)_1, (n_2)_2, \ldots, (n_2)_m\}$, and let the values of ω be placed in k successive cells $C_1, \ldots, C_k$, with all occurrences of $\min(\omega)$ being placed in cell C_1, all occurrences of $\min(\omega) + 1$ in cell C_2, ..., and all occurrences of $\max(\omega)$ in C_k. The aim is to estimate the true cell probability π_l for each cell C_l based on the finite observations ω. A straightforward estimator of π_l is, of course, the relative frequency N_l/m, where N_l is the number of values occupying cell C_l; however, using local polynomials of degree d, a more accurate estimation due to smoothing is provided by Jacob and Oliveira (2011):

$$\widehat{\pi}_l(d) = \frac{1}{kh} \sum_{j=1}^{k} L_{l,d}\left(\frac{x_j - x_l}{h}\right) \frac{N_j}{m} \ , \tag{4}$$

where $L_{l,d}(\cdot)$ is the local d-degree polynomial estimator for the probability of cell C_l, and $x_j = (j - 1/2)/k$ for $j = 1, \ldots, k$. Based on the work by Ruppert and Wand (1994) on locally weighted regression, Aerts et al. (1997a, 1997b) express $L_{l,d}(\cdot)$ by

$$L_{l,d}(u) = \frac{\left|M_{l,d}(u)\right|}{\left|N_{l,d}\right|} K(u) \ ,$$

where $|\cdot|$ is the determinant, with $N_{l,d}$ the $(d+1)\times(d+1)$-matrix having the (r,s) entry given by

$$(N_{l,d})_{r,s} = \frac{1}{kh}\sum_{i=1}^{k}\left(\frac{x_i - x_l}{h}\right)^{r+s-2} K\left(\frac{x_i - x_l}{h}\right) .$$

The matrix $M_{l,d}(u)$ is the same as $N_{l,d}$, but with the first column replaced by $(1, u, \ldots, u^d)^T$. The width of density function $K(\cdot)$ is controlled by h.

Both this technique for local polynomial smoothing and Gillespie simulation were used in Algorithm 1 for the estimation of $p(n_1|n_2)$, and the relationship between this approach and the other strategies we describe is shown by **Box A** of Figure 2. Note that the implementation of Algorithm 1 assumes that the values within $\boldsymbol{\theta}$ are the same across all the bacteria of a bottleneck.

[Figure 2 about here.]

Algorithm 1 Estimation of $p(n_1|n_2)$.

Input: Post-bottleneck population size n_2, and Gillespie simulation parameters Δt and $\boldsymbol{\theta}$.

Output: An estimate of $p(n_1|n_2)$ for $n_1 = 1, \ldots, n_{max}$ ($n_{max} \le n_2$).

```
posts ← [ ]
for $n_1 \in \{1, \ldots, n_{max}\}$ do
    $\{(n_2)_j\}_{j=1,\ldots,100} \overset{i.i.d}{\sim} Gillespie(n_1, \Delta t, \boldsymbol{\theta})$
    get estimate $\widehat{p}(n_2|n_1)$ from $\{(n_2)_j\}_{j=1,\ldots,100}$        ▷ using (4)
    append $\widehat{p}(n_2|n_1)$ to array posts
end for
$[\widehat{p}(n_1 = 1|n_2), \ldots, \widehat{p}(n_1 = n_{max}|n_2)] \leftarrow posts/sum(posts)$
return array $[\widehat{p}(n_1 = 1|n_2), \ldots, \widehat{p}(n_1 = n_{max}|n_2)]$
```

[Figure 3 about here.]

[Figure 4 about here.]

2.2 Example

Salmonella enterica is the main cause of salmonellosis, and its pathogenesis is under extensive research thanks to well established murine experimental models (Dougan & Baker, 2014). When *S. enterica* enters the bloodstream of a host, bacteria spread to a number of organs including the liver and the spleen (Figure 5). In researching this process using a mouse model, Grant et al. (2008) estimated the division, death, clearance and immigration rates of the bacterium at different times in the absence of medical treatment.

[Figure 5 about here.]

Here we build on previous models by introducing a hypothetical bottleneck caused by a dose of antibiotics. To demonstrate the efficacy of Algorithm 1 to detect bottlenecks of size $n_1^\star = 1$, 2, 4, 8, 80, 800 and 1600 in an organ such as the liver of a mouse, a value $n_2^\star$ for n_2 was first derived from $n_1^\star$ and then the posterior distribution $p(n_1|n_2^\star)$ was estimated from $n_2^\star$ using Algorithm 1 as follows:

```
1: Choose a target bottleneck size $n_1^\star \in \{1, 2, 4, 8, 80, 800, 1600\}$
2: $\{(n_2)_j\}_{j=1,\dots,101} \overset{i.i.d}{\sim} \mathit{Gillespie}(n_1^\star, \Delta t, \boldsymbol{\theta})$   ▷ 101 $n_2$ values derived from $n_1^\star$
3: Set $n_2^\star$ to the median of $\{(n_2)_j\}_{j=1,\dots,101}$
4: Obtain $\widehat{p}(n_1|n_2^\star)$ using Algorithm 1
5: Compare $\widehat{p}(n_1|n_2^\star)$ with target $n_1^\star$
```

The 101 Gillespie simulations were conducted assuming an inoculum of 1000 bacteria injected intravenously at $t = 0$ (equivalent to using 101 mice). An infection was allowed to progress to $t = 24$ hours according to $\boldsymbol{\theta}$ using the parameter values published by Grant et al. (2008) (i.e., time-varying per capita division, death, emigration and immigration rates estimated by iteratively fitting rate equations to observed data), but at $t = 24$ hours, the number of bacteria in the liver was changed to n_1 within the Gillespie simulation. The growth period Δt was set to 12 hours to allow sufficient time for significant growth of the bacteria before infection can be detected and treatment applied. Estimates of $p(n_2|n_1, \boldsymbol{\theta})$ for (3) were provided by (4) using local polynomial smoothing with an Epanechnikov kernel density function and local polynomials of degree $d = 1$.

Figure 3 presents the resulting estimated posterior probabilities, and Table 2 shows that the medians of the posteriors agreed within 1% of the true bottleneck sizes.

[Table 2 about here.]

2.3 Inclusion of a pre-bottleneck population

What if the size n_0 of the pre-bottleneck population $\mathcal{P}_0$ is also available to us, or is assumed? The posterior distribution for n_1 becomes

$$\begin{aligned} p(n_1|n_0, n_2) &\propto p(n_1|n_0)p(n_2|n_0, n_1) \\ &= p(n_1|n_0)p(n_2|n_1) \ , \end{aligned}$$

where $p(n_2|n_1)$ is estimated as before.

Bacteria are either killed by an antibiotic or not; therefore, with regard to $p(n_1|n_0)$, a simple assumption is that this probability is given by the binomial probability distribution

$$p(n_1|n_0; \pi) = \binom{n_0}{n_1} \pi^{n_1} (1 - \pi)^{n_0 - n_1} \tag{5}$$

where π is the probability that a bacterium will be included in the bottleneck; however, π is not known *a priori*. Furthermore, (5) implies that the expected size n_1 of the bottleneck is a linear function of n_0 for all n_0,

$$\mathbb{E}[n_1|n_0; \pi] = n_0\pi \ ,$$

but Abel et al. (2015) warn that alternative scenarios could exist. One alternative scenario suggested by Abel et al. (an “absolute bottleneck”) is when the bottleneck reaches a plateau with respect to inoculum size; another (a “cooperative bottleneck”) is when bacteria cannot pass into a bottleneck unless a sufficient number of organisms are present; however, data are currently lacking to validate those scenarios.

3 Multitype populations

A multitype population of bacteria is possible by using phenotypically identical bacterial strains where each strain carries a different DNA signature tag in the same noncoding region of the chromosome (Crimmins & Isberg, 2012). An example of this is the use of wild-type isogenic tagged strains (WITS) (Grant et al., 2008) which we will consider herein.

Suppose now that bacterial pre-bottleneck population $\mathcal{P}_0$ is composed of eight WITS: $w_0^{[1]}, \ldots, w_0^{[8]}$, where $w_0^{[i]} \geq 0$ is the number of bacteria tagged with the i-th WITS tag at $t = 0$. The population is reduced to bottleneck $\mathcal{P}_1$ with WITS distribution $w_1^{[1]}, \ldots, w_1^{[8]}$, and population $\mathcal{P}_2$ resulting from the growth of $\mathcal{P}_1$ has WITS distribution $w_2^{[1]}, \ldots, w_2^{[8]}$. As shown in Figure 6, the distribution of WITS in $\mathcal{P}_2$ can be very different to that present in $\mathcal{P}_0$ because of the stochastic variation of $\mathcal{P}_1$ (Abel et al., 2015), and possibly also from $\mathcal{P}_1$ to $\mathcal{P}_2$.

Given the WITS distribution $w_2^{[1]}, \ldots, w_2^{[8]}$ of a post-bottleneck population, estimated posterior distributions $\widehat{p}(w_1^{[1]}|w_2^{[1]}), \ldots, \widehat{p}(w_1^{[8]}|w_2^{[8]})$ can be obtained for each of the WITS independently of each other using Algorithm 1, with $w_1^{[j]}$ being used in the algorithm place of n_1, and $w_2^{[j]}$ in place of n_2.

The total size n_1 of a bottleneck composed of WITS is given by the sum $w_1^{[1]} + \cdots + w_1^{[8]}$. There are two approaches to estimating n_1: (a) use the sum $w_2^{[1]} + \cdots + w_2^{[8]}$ for n_2, ignore the WITS tags and estimate $p(n_1|n_2)$ using Algorithm 1; (b) determine the posterior mass function for the sum $n_1 = w_1^{[1]} + \cdots + w_1^{[8]}$ by applying convolution successively to the individual WITS posterior distributions $\widehat{p}(w_1^{[1]}|w_2^{[1]}), \ldots, \widehat{p}(w_1^{[8]}|w_2^{[8]})$.

If X_1 and X_2 are two independent integer-valued random variables with distribution functions $p_1 = p(X_1)$ and $p_2 = p(X_2)$, and $Z = X_1 + X_2$, the distribution function $p(Z)$ is given by

$$p(Z = z) = (p_1 \otimes p_2)(z) = \sum_x p(X_1 = x) p(X_2 = z - x) \ ,$$

where $\otimes$ is the convolution operator.

Because the convolution operator is commutative, we can extend its use to sums of more than two random variables, $Z = X_1 + X_2 + \cdots + X_m$, by repeatedly applying the operator:

$$\begin{aligned} p(Z = z) &= (p_1 \otimes p_2 \otimes \cdots \otimes p_m)(z) \\ &= ((\cdots (p_1 \otimes p_2)(z) \otimes \cdots) \otimes p_m)(z) \end{aligned}$$

To examine the effect of using convolution, we used target value $n_1^\star = 800$ with $w_1^{[1]\star} = 100, \ldots, w_1^{[8]\star} = 100$. Convolution was applied to the estimated posteriors $\widehat{p}(w_1^{[1]}|w_2^{[1]\star}), \ldots, \widehat{p}(w_1^{[8]}|w_2^{[8]\star})$ via Fourier transformation, but the resulting estimated posterior $\widehat{p}(n_1|w_2^{[1]\star},\ldots,w_2^{[8]\star}) = \widehat{p}(w_1^{[1]} + \cdots + w_1^{[8]}|w_2^{[1]\star},\ldots,w_2^{[8]\star})$ was no better than the posterior $\widehat{p}(n_1|n_2)$ obtained by ignoring the WITS other than being smoother (Figure 4). However, the convolutional approach used eight estimated posterior probabilities $\widehat{p}(w_1^{[i]}|w_2^{[i]\star})$ $(i = 1,\ldots,8)$; in contrast, the non-convolutional approach would use only one estimated posterior, $\widehat{p}(n_1|n_2)$. This suggests that the convolutional approach is potentially more prone to error.

3.1 Inclusion of a multitype pre-bottleneck population

Suppose that we know, or are able to assume, the composition $\mathbf{w}_0 = \{w_0^{[1]},\ldots,w_0^{[8]}\}$ of a pre-bottleneck multitype population as well as that $(\mathbf{w}_2)$ of a post-bottleneck population. In this case, the posterior of the bottleneck size is $p(n_1|\mathbf{w}_0,\mathbf{w}_2)$.

[Figure 6 about here.]

Now,

$$\begin{aligned} p(n_1|\mathbf{w}_0,\mathbf{w}_2) &= p(\bigvee_{\substack{\mathbf{w}_1 \\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} \mathbf{w}_1 \quad |\mathbf{w}_0,\mathbf{w}_2) \\ &= \sum_{\substack{\mathbf{w}_1 \\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0,\mathbf{w}_2), \end{aligned}$$

where $\vee$ denotes logical disjunction and $sum(\mathbf{w}_1) = \sum_i w_1^{[i]}$. From Bayes' theorem,

$$\begin{aligned} p(\mathbf{w}_1|\mathbf{w}_0,\mathbf{w}_2) &= \frac{p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1,\mathbf{w}_0)}{\sum\limits_{\mathbf{w}_1} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1,\mathbf{w}_0)} \\ &= \frac{p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)}{\sum\limits_{\mathbf{w}_1} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)}; \qquad (6) \end{aligned}$$

thus

$$p(n_1|\mathbf{w}_0,\mathbf{w}_2) = \frac{\sum\limits_{\substack{\mathbf{w}_1\\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)}{\sum\limits_{\mathbf{w}_1} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)}$$

$$= \frac{\sum\limits_{\substack{\mathbf{w}_1\\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)}{\sum\limits_{n_1} \sum\limits_{\substack{\mathbf{w}_1\\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)} . \tag{7}$$

If the WITS are phenotypically identical, each bacterium has the same probability of surviving an antibiotic. Under this assumption, a distribution $\mathbf{w}_1$ of n_1 WITS in a bottleneck $\mathcal{P}_1$ can be regarded as a sample resulting from a random selection (without replacement) of n_1 WITS from the pre-bottleneck population $\mathcal{P}_0$ with distribution $\mathbf{w}_0$. The probability of selecting $\mathbf{w}_1$ from $\mathbf{w}_0$ without replacement such that $sum(\mathbf{w}_1) = n_1$ is given by the multivariate hypergeometric probability distribution:

$$p(\mathbf{w}_1|\mathbf{w}_0, sum(\mathbf{w}_1) = n_1) = \frac{\binom{w_0^{[1]}}{w_1^{[1]}}\binom{w_0^{[2]}}{w_1^{[2]}}\cdots\binom{w_0^{[8]}}{w_1^{[8]}}}{\binom{w_0^{[1]}+w_0^{[2]}+\cdots+w_0^{[8]}}{w_1^{[1]}+w_1^{[2]}+\cdots+w_1^{[8]}}} . \tag{8}$$

As for $p(\mathbf{w}_2|\mathbf{w}_1)$, the independence between the WITS enables us to factorise $p(\mathbf{w}_2|\mathbf{w}_1,\boldsymbol{\theta})$ as follows:

$$p(\mathbf{w}_2|\mathbf{w}_1,\boldsymbol{\theta}) = \prod_i p(w_2^{[i]}|\mathbf{w}_1,\boldsymbol{\theta}) = \prod_i p(w_2^{[i]}|w_1^{[i]},\boldsymbol{\theta}) , \tag{9}$$

and estimation of $p(w_2^{[i]}|w_1^{[i]},\boldsymbol{\theta})$ for (9) can be performed in the same manner as described for $p(n_2|n_1,\boldsymbol{\theta})$ using Algorithm 1.

See Box B in Figure 2 for the relationship between this approach and the others we describe.

3.2 On assuming proportionality

What if we make the simplifying assumption that the distribution of the frequencies of $\mathbf{w}_2$ are proportional to those of $\mathbf{w}_1$ by the same amount ψ? That is, $\mathbf{w}_2 = \psi\mathbf{w}_1$. where ψ is a positive integer.

Through Bayes' theorem, we have

$$\begin{aligned} p(n_1|\boldsymbol{\pi}_0, \mathbf{w}_2) &\propto p(n_1|\boldsymbol{\pi}_0)p(\mathbf{w}_2|n_1, \boldsymbol{\pi}_0) \\ &= p(n_1)p(\mathbf{w}_2|n_1, \boldsymbol{\pi}_0) \ , \end{aligned}$$

where the elements of $\boldsymbol{\pi}_0$ are those of $\mathbf{w}_0$ expressed as relative frequencies.

Moreover, if we assume $p(n_1)$ to be equiprobable for all n_1 then

$$p(n_1|\boldsymbol{\pi}_0, \mathbf{w}_2) \propto p(\mathbf{w}_2|n_1, \boldsymbol{\pi}_0) \ . \tag{10}$$

Suppose we make the further assumption that the elements of $\mathbf{w}_2$ developed proportionality from those in $\mathbf{w}_1$: $\mathbf{w}_2 = \psi \mathbf{w}_1$ for some positive integer ψ. This assumption implies that $\psi n_1 = sum(\mathbf{w}_2)$ and, as $sum(\mathbf{w}_2)$ is constant for a given $\mathbf{w}_2$, it follows that $\mathbf{w}_2$ can be derived from $\mathbf{w}_1$ using a range of ψ values such that $\psi = sum(\mathbf{w}_2)/n_1$. But is one ψ more likely than another?

If $p(\mathbf{w}_2|n_1, \boldsymbol{\pi}_0)$ is defined by a multinomial distribution then

$$\begin{aligned} p(\mathbf{w}_2|n_1, \boldsymbol{\pi}_0) &= p(\langle w_2^{[1]}, w_2^{[2]}, \ldots, w_2^{[8]} \rangle | n_1, \langle \pi_0^{[1]}, \pi_0^{[2]}, \ldots, \pi_0^{[8]} \rangle) \\ &= \binom{w_2^{[1]} + w_2^{[2]} + \cdots + w_2^{[8]}}{w_2^{[1]}, w_2^{[2]}, \ldots, w_2^{[8]}} \prod_{i=1}^{8} \left(\pi_0^{[i]}\right)^{w_2^{[i]}} \ , \end{aligned}$$

where $\binom{\alpha}{\beta_1, \ldots, \beta_m}$ is the multinomial coefficient $\frac{\alpha!}{\beta_1! \cdots \beta_m!}$; however,

$$\binom{\psi w_1^{[1]} + \psi w_1^{[2]} + \cdots + \psi w_1^{[8]}}{\psi w_1^{[1]}, \psi w_1^{[2]}, \ldots, \psi w_1^{[8]}} \prod_{i=1}^{8} \left(\pi_0^{[i]}\right)^{\psi w_1^{[i]}} < \binom{w_1^{[1]} + w_1^{[2]} + \cdots + w_1^{[8]}}{w_1^{[1]}, w_1^{[2]}, \ldots, w_1^{[8]}} \prod_{i=1}^{8} \left(\pi_0^{[i]}\right)^{w_1^{[i]}}$$

for any positive integer ψ (note that $\pi_0^{[i]}$ is the same on both sides of the inequality), thus

$$p(\psi \mathbf{w}_1 | n_1 = \psi sum(\mathbf{w}_1), \boldsymbol{\pi}_0) < p(\mathbf{w}_1 | n_1 = sum(\mathbf{w}_1), \boldsymbol{\pi}_0) \ .$$

Consequently, if $\mathbf{w}_2 = \psi \mathbf{w}_1$ then $sum(\mathbf{w}_1)$ is the most probable value for n_1. Put another way, if proportionality is assumed then the most probable value for n_1 is $sum(\mathbf{w}_2)$ divided by the highest common factor for the elements of $\mathbf{w}_2$.

3.3 A sampling approach

In the previous section, we have shown how to obtain values for $p(\mathbf{w}_1|\mathbf{w}_0)$ and $p(\mathbf{w}_2|\mathbf{w}_1)$ that are required for (7), but (7) also requires us to determine the summands for every possible $\mathbf{w}_1$ such that $\sum_i w_1^{[i]} = n_1$. The problem with this is that the number of possible $\mathbf{w}_1$ for a given value of n_1 grows super-exponentially with n_1 (Charalambides, 2002, p.138); for example, the number of possible $\mathbf{w}_1$ when selecting 100 microbes from $\mathbf{w}_0 = \langle 1000^{[1]}, 1000^{[2]}, \ldots, 1000^{[8]} \rangle$ is more than 26 thousand million. As this is combinatorially (and thus computationally) challenging, an alternative approach is required.

To circumvent the combinatorial issue, one could consider restricting the summations of (7) to the more probable configurations of $\mathbf{w}_1$, such as the modes of $\mathbf{w}_1$. An algorithm for the generation of all the modes of a multivariate hypergeometric distribution has been proposed by Requena and Cludad (2003), but a simpler approach is to randomly sample points $\mathbf{w}_1$, say 1000 times, from $p(\mathbf{w}_1|\mathbf{w}_0)$ as a multivariate hypergeometric distribution, given that most of these points would be expected to be in the vicinity of the modes. With multiple modes, sampling takes place proportionally across the modes.

Our implementation of the sampling approach is shown in Algorithm 2, and its efficacy was tested using the following steps of a toy experiment:

```
Set $\mathbf{w}_0$ to $\langle 600^{[1]}, 600^{[2]}, \ldots, 600^{[8]} \rangle$
Choose a target bottleneck size $n_1^\star \in \{80, 800, 1600\}$
In order to choose a $\mathbf{w}_1$ associated with target $n_1^\star$, select a mode $\mathbf{w}_1^\star$ from the multivariate hypergeometric distribution associated with random samples of size $n_1^\star$ taken from $\mathbf{w}_0$ (Requena & Cludad, 2003)
In order to choose a $\mathbf{w}_2$ resulting from $\mathbf{w}_1^\star$, first do
    $\{(\mathbf{w}_2)_j\}_{j=1,\ldots,101} \overset{i.i.d}{\sim} \mathit{Gillespie}(\mathbf{w}_1^\star, \Delta t, \boldsymbol{\theta})$,
then set $\mathbf{w}_2^\star$ to the median of $\{(\mathbf{w}_2)_j\}_{j=1,\ldots,101}$
Obtain $\widehat{p}(n_1|\mathbf{w}_0, \mathbf{w}_2^\star)$ using Algorithm 2
Compare $\widehat{p}(n_1|\mathbf{w}_0, \mathbf{w}_2^\star)$ with target $n_1^\star$
```

Figure 7 displays the resulting posterior distributions, which have median accuracies similar to those shown for the estimation of $p(n_1|n_2)$ derived by Algorithm 1.

[Figure 7 about here.]

4 Patterns of missing WITS

An assumption made when estimating $p(n_2|n_1, \boldsymbol{\theta})$ via Gillespie simulation is that parameters $\boldsymbol{\theta}$ are known and are not influenced by the presence of an antibiotic, but this is not necessarily always the case (Kaiser et al., 2014). Consequently, how can we estimate bottleneck size when $\boldsymbol{\theta}$ is not known to us?

Consider (7) written as

$$p(n_1|\mathbf{w}_0, \mathbf{w}_2) \propto \sum_{\substack{\mathbf{w}_1 \\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0)p(\mathbf{w}_2|\mathbf{w}_1)\,' \tag{11}$$

and suppose we replace $\mathbf{w}_2$ with a vector ξ_2 denoting which WITS in $\mathbf{w}_2$ are missing, then $p(\mathbf{w}_2|\mathbf{w}_1)$, in turn, becomes replaced by $p(\xi_2|\mathbf{w}_1)$. Furthermore, if we assume that missingness pattern ξ_2 is equal to the missingness pattern ξ_1 of $\mathbf{w}_1$ then ξ_2 is implied by $\mathbf{w}_1$ and there is no need to consider post-bottleneck dynamics. Using this approach, (11) simplifies to

$$p(n_1|\mathbf{w}_0, \xi_2) \propto \sum_{\substack{\mathbf{w}_1 \\ \text{s.t. } sum(\mathbf{w}_1)=n_1}} p(\mathbf{w}_1|\mathbf{w}_0)\ \mathbb{1}(\mathbf{w}_1 \Rightarrow \xi_2)\ , \tag{12}$$

where $\mathbb{1}(\cdot)$ is the indicator function. However, the assumption that $\xi_2 = \xi_1$, and thus that $\mathbf{w}_1 \Rightarrow \xi_2$, may not hold if a few WITS are randomly lost soon after a bottleneck due to (a) a small bottleneck, (b) a large number of WITS, or (c) high post-bottleneck replication and death rates, or any combination of these three.

In order to compare the estimate provided by (12) with the correct value given by (11), an experiment was used based on the following scenario. A bottleneck WITS population $\mathbf{w}_1$ of size n_1 is assumed to have been sampled from $\mathbf{w}_1 = \langle 4^{[1]}, 4^{[2]}, \ldots, 4^{[8]} \rangle$ without replacement. Each element $w_1^{[i]}$ of $\mathbf{w}_1$ then gives rise to an element $w_2^{[i]}$ of $\mathbf{w}_2$ by sampling from a Poisson distribution with Poisson parameter $\lambda = 10w_1^{[i]}$. This scenario is the basis for the following toy experiment:

```
1: Set $\mathbf{w}_0$ to $\langle 4^{[1]}, 4^{[2]}, \ldots, 4^{[8]} \rangle$
2: Choose a target bottleneck size $n_1^\star \in \{3, 7, 20\}$
3: In order to choose a $\mathbf{w}_1$ associated with target $n_1^\star$, select a mode $\mathbf{w}_1^\star$
   from the multivariate hypergeometric distribution associated with random samples of size $n_1^\star$ taken from $\mathbf{w}_0$
```

```
For $\mathbf{w}_2$, use $\mathbf{w}_2^\star = 10\mathbf{w}_1^\star$ (i.e., vector of expected values from the Poisson distributions)
Get missingness pattern $\xi_2$ corresponding to $\mathbf{w}_2^\star$
Obtain $p(n_1|\mathbf{w}_0, \mathbf{w}_2^\star)$ using (11)
Compare $p(n_1|\mathbf{w}_0, \mathbf{w}_2^\star)$ with target $n_1^\star$
Obtain $p(n_1|\mathbf{w}_0, \xi_2)$ using (12)
Compare $p(n_1|\mathbf{w}_0, \xi_2)$ with target $n_1^\star$
```

The results of the experiment are shown in Figures 9 and 10. In the case of the probability mass functions for $p(n_1|\mathbf{w}_0, \mathbf{w}_2^\star)$, the modes coincided exactly with the target values, but this was not the case for $p(n_1|\mathbf{w}_0, \xi_2)$. When at least one WITS was missing, the mode for $p(n_1|\mathbf{w}_0, \xi_2)$ was greater than the target value. This is associated with the observation that the mean value of $p(\mathbf{w}_2|\mathbf{w}_1)$ as encountered in (11) tended to be less than the mean value for $\mathbb{1}(\mathbf{w}_1 \Rightarrow \xi_2)$ in (12), which is equal to $p(\mathbf{w}_1 \Rightarrow \xi_2)$. When no WITS were missing, the resulting probability mass function for $p(n_1|\mathbf{w}_0, \xi_2)$ exhibited a plateau as n_1 increased. This can be explained as follows: it is increasingly unlikely that no WITS are missing as n_1 decreases; on the other hand, the absence of missing WITS can be explained by the occurrence of large n_1 values up to and including the complete absence of a bottleneck.

When $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$ has a plateau, a lower bound for n_1 can be set equal to the lower bound of the 95% highest density interval with respect to $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$, which is a type of one-sided credible interval.

Note that the above toy experiment uses Equation (12) exactly so as to display the resulting distributions. In reality, the inoculum size would be far greater than 4×8 and, in such circumstances, the sampling approach of Section 3.3 would be used instead. Figure 8 shows the result of using sampling when ξ_2 is used in place of $\mathbf{w}_2$, with $n_1 = 600 \times 8$. Note also that, in order to use patterns of missing WITS, it is not necessary for the isotypes to be in equal amounts in the pre-bottleneck population.

As seen by comparing Figures 9 with 10, although the use of missingness patterns decreases model complexity and computation time (no Gillespie simulations required), accuracy is also decreased.

The relationship between using missingness patterns (without dilution) and the other methods we describe is shown by **Box C** in Figure 2.

[Figure 8 about here.]

4.1 Dilution of WITS with monotypes

In Figure 10 (c), the bottleneck of size 20 could not be estimated because of the presence of a plateau instead of a mode. This issue can be overcome by diluting the WITS with untagged (i.e., monotype) isogenic bacteria. The justification for this is that, for a fixed n_1, the probability of at least one WITS being missing increases as the pre-bottleneck population $\mathbf{w}_0$ becomes more dilute.

Let $\mathbf{w}_0^+$ represent $\mathbf{w}_0$ augmented with u untagged bacteria. For example, if we add 10 untagged bacteria to $\mathbf{w}_0 = \langle 4^{[1]}, \ldots, 4^{[8]} \rangle$ then $\mathbf{w}_0^+ = \langle 4^{[1]}, \ldots, 4^{[8]}, 10 \rangle$. Upon using $\mathbf{w}_0^+$ in place of $\mathbf{w}_0$, expression (12) becomes

$$p(n_1 | \mathbf{w}_0^+, \xi_2) \propto \sum_{\substack{\mathbf{w}_1^+ \\ \text{s.t. } \mathit{sum}(\mathbf{w}_1^+) = n_1}} p(\mathbf{w}_1^+ | \mathbf{w}_0^+) \; \mathbb{1}(\mathbf{w}_1^+ \Rightarrow \xi_2) \; , \tag{13}$$

where $\mathbf{w}_1^+$ allows for the possibility that untagged bacteria can be present in the bottleneck. The implication that $\mathbf{w}_1^+ \Rightarrow \xi_2$ in (13) is based only on the WITS component of $\mathbf{w}_1^+$; the untagged bacteria in $\mathbf{w}_1^+$ are ignored. See **Box D** in Figure 2.

By way of example, suppose that we add $u = 13$ untagged bacteria to $\mathbf{w}_0 = \langle 4^{[1]}, \ldots, 4^{[8]} \rangle$ in order to perform the following toy experiment:

1: Set $\mathbf{w}_0^+$ to $\langle 4^{[1]}, 4^{[2]}, \ldots, 4^{[8]}, u \rangle$
2: Set number of untagged bacteria $u = 13$
3: Set target bottleneck size $n_1^\star = 20$
4: In order to choose a $\mathbf{w}_1^+$ associated with target $n_1^\star$, first select a mode $\mathbf{w}_1^\star$ from the multivariate hypergeometric distribution associated with random samples of size $n_1^\star - u$ taken from $\langle 4^{[1]}, 4^{[2]}, \ldots, 4^{[8]} \rangle$, and then set $\mathbf{w}_1^{+\star} = \langle \mathbf{w}_1^\star, u \rangle$
5: For $\mathbf{w}_2$, use $\mathbf{w}_2^\star = 10\mathbf{w}_1^{+\star}$
6: Get missingness pattern ξ_2 corresponding to $\mathbf{w}_2^\star$
7: Obtain $p(n_1 | \mathbf{w}_0^+, \xi_2)$ using (13)

Figure 11 shows that dilution with untagged bacteria has allowed an estimate of the bottleneck size to be performed. But a note of caution is due. Using $u = 13$ allowed the bottleneck size to be estimated as 18 (whereas $u = 12$ gave a plateau), but the position of the mode for $p(n_1 | \mathbf{w}_0^+, \xi_2)$ is influenced by the choice of u, with the mode decreasing as u increases. For example, the mode was 13 when $u = 14$ and 10 when $u = 15$. A similar behaviour has been observed when using other target values for n_1. This suggests that the best approach to estimating bottleneck size via this method is to use the smallest possible value for u that permits a mode to appear instead of a plateau.

The above WITS dilution technique was used by Maier et al. (2014) to estimate the size of gut luminal bottlenecks during *Salmonella* Typhimurium colitis. The inoculum consisted of seven WITS in equal proportions, which was increasingly diluted with an untagged isogenic wild-type strain until a loss of at least one WITS was first detected in a post-bottleneck population. This point occurred at a dilution of 1:7000. The size of a bottleneck was then estimated using a likelihood function based on binomial selection.

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

Lim et al. (2014) also developed a method to estimate the size of a bottleneck from an observation of missing WITS, but their approach was more granular in that it was restricted to just those cases where at least one WITS was missing but without considering the precise the number of missing WITS. Let $\zeta \in \{\mathit{true}, \mathit{false}\}$ denote the state that at least one WITS is missing from $\mathbf{w}_2$ (and thus assumably from $\mathbf{w}_1$). In this context, the posterior for bottleneck size is $p(n_1|\mathbf{w}_0, \zeta)$. The computational approach used by Lim et al. differed from (12) in that they derived a plot of $\widehat{p}(\zeta|n_1, \mathbf{w}_0)$ as a function of n_1 using computer simulations and then assumed that

$$p(n_1|\mathbf{w}_0, \zeta) \propto p(\zeta|n_1, \mathbf{w}_0) \ .$$

4.2 On increasing the number of WITS

All the examples shown so far have been based on the use of eight WITS, but what if a larger number of WITS are used? Intuitively, if we keep the bottleneck size n_1 constant but increase the number $|\mathcal{W}|$ of WITS available then the number of missing WITS is expected to increase. Furthermore, the rate of change in the number of missing WITS as n_1 decreases is expected to

Algorithm 2 Estimation of $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$.

Input: Pre-bottleneck population $\mathbf{w}_0$, post-bottleneck population $\mathbf{w}_2$, and Gillespie simulation parameters Δt and $\boldsymbol{\theta}$.

Output: An estimate of $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$ for $n_1 = 1, \ldots, nmax$.

```
$posts \leftarrow [\,]$
for $n_1 \in \{1, \ldots, nmax\}$ do
    $sum2 \leftarrow 0$
    loop 1000 times
        $\mathbf{w}_1 \sim MultivariateHypergeometric(n_1, \mathbf{w}_0)$
        $a \leftarrow p(\mathbf{w}_1|\mathbf{w}_0)$                                ▷ according to (8)
        $b \leftarrow 1$
        for $w_1^{[i]} \in \mathbf{w}_1$ do
            if $w_1^{[i]} = 0$ then
                if $w_2^{[i]} = 0$ then
                    $p \leftarrow 1$
                else
                    $p \leftarrow 0$
            else
                $\{(w_2^{[i]})_j\}_{j=1,\ldots,100} \overset{i.i.d}{\sim} Gillespie(w_1^{[i]}, \Delta t, \boldsymbol{\theta})$
                get estimate $\widehat{p}(w_2^{[i]}|w_1^{[i]})$ from $\{(w_2^{[i]})_j\}_{j=1,\ldots,100}$
                $p \leftarrow \widehat{p}(w_2^{[i]}|w_1^{[i]})$
            $b \leftarrow b \times p$                                ▷ final $b$ is $\widehat{p}(\mathbf{w}_2|\mathbf{w}_1)$
        end for
        $sum2 \leftarrow sum2 + a \times b$          ▷ final $sum2$ is approx numerator of (7)
        append $sum2$ to array $posts$
    end loop
end for
$\left[\widehat{p}(n_1 = 1|\mathbf{w}_0, \mathbf{w}_2), \ldots, \widehat{p}(n_1 = nmax|\mathbf{w}_0, \mathbf{w}_2)\right] \leftarrow posts/sum(sum1)$
return array $\left[\widehat{p}(n_1 = 1|\mathbf{w}_0, \mathbf{w}_2), \ldots, \widehat{p}(n_1 = nmax|\mathbf{w}_0, \mathbf{w}_2)\right]$
```

increase as $|\mathcal{W}|$ increases. This suggests that accuracy in the estimation of n_1 from missingness patterns ξ_2 should improve with larger $|\mathcal{W}|$. This argument is supported by the results of the computer simulations conducted by Lim et al. (2014) using different values for $|\mathcal{W}|$ in which a significant improvement occurs on going from $|\mathcal{W}| = 10$ to $|\mathcal{W}| = 40$.

As a further demonstration, the posterior distribution $p(n_1|\mathbf{w}_0, \xi_2)$ shown in Figure 10 when $n_1 = 7$, which is based on 8 WITS, was recalculated using 12 WITS resulting in a decrease in variance (Figure 12).

[Figure 12 about here.]

5 Discussion

We describe Bayesian approaches to estimating the size of bacterial bottlenecks given observation of a post-bottleneck population (either monotype or multitype), but size estimation with respect to monotypes is possible only when post-bottleneck dynamics are known (or assumed). Unlike previous studies which were tailored to single experiments, we demonstrated how this framework can be applied to a variety of situations. In particular, we presented analyses inspired by several published studies on *Salmonella enterica* in mice, covering a range of bottleneck sizes, bacterial dynamic regimes and technical constraints.

The use of multitype isogenic bacteria in the form of WITS allows the composition of pre-bottleneck populations to be included in the analysis. Furthermore, the use of WITS enables bottleneck sizes to be estimated when post-bottleneck dynamics is not known. This is done through observation of patterns of missing WITS; however, this can require dilution of an inoculum with isogenic untagged bacteria. Figure 2 provides a flowchart that summarises these observations.

In our analysis, we have assumed that $\boldsymbol{\theta}$ is homogeneous across all the bacteria in a bottleneck population $\mathcal{P}_1$, but this may not be the case for some species of bacteria. Indeed, recent experimental studies have shown that, following inoculation into mice, genetically identical *S. enterica* bacteria segregate into heterogeneous subpopulations with different division rates and sensitivities to antibiotics (Claudi et al., 2014; Helaine & Kugelberg, 2014). If we can differentiate between the subpopulations having different $\boldsymbol{\theta}$, and if the dynamics of the individual subpopulations are known, then the posterior probability distribution of n_1 can be determined separately – using either interpolation or extrapolation as stated in Figure 2 – and combined. On the other hand, if only a single $\boldsymbol{\theta}$ is available, and its values are incorrect for $\mathcal{P}_1$ then the estimated bottleneck size will be incorrect. In this case, the use of missingness patterns with dilution should be considered but using the largest number of WITS possible in order to maximise accuracy.

Our work has focused on intrahost bacterial bottlenecks, but a further line of enquiry is to extend the framework to the assessment of other pathogens (e.g., viruses) or types of bottlenecks (e.g., transmission bottlenecks). Models for viruses already exist, but they rely on naturally occurring genetic diversity through mutations in RNA viruses; for example, the case with influenza (Sobel Leonard et al., 2017).

Author contributions

RD, PM and OR conceived the project. RD performed the mathematical analyses and simulations, which were checked by DJP. All the authors contributed to the preparation of the manuscript.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

RD was supported by BBSRC grant BB/I002189/1 awarded to PM. DJP was supported by BBSRC grant BB/M020193/1 awarded to OR.

References

Abel, S., Abel zur Wiesch, P., Davis, B., & Waldor, M. (2015). Analysis of bottlenecks in experimental models of infection. *PLoS Pathogens*, *11*(6), e1004823.

Aerts, M., Augustyns, I., & Janssen, P. (1997a). Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*, *8*(2), 127-147.

Aerts, M., Augustyns, I., & Janssen, P. (1997b). Sparse consistency and smoothing for multinomial data. *Statistics and Probability Letters*, *33*, 41-48.

Charalambides, C. (2002). *Enumerative combinatorics.* Boca Raton, Fl: Chapman & Hall/CRC.

Claudi, B., Spröte, P., Chirkova, A., Personnic, N., Zankl, J., Schürmann, N., et al. (2014). Phenotypic variation of Salmonella in host tissues delays eradication by antimicrobial chemotherapy. *Cell*, *158*(4), 722-733.

Coward, C., Restif, O., Dybowski, R., Grant, A., Maskell, D., & Mastroeni, P. (2014). The effects of vaccination and immunity on bacterial infection dynamics *in vivo*. *PLoS Pathogens*, *10*(9), e1004359.

Crimmins, G., & Isberg, R. (2012). Analyzing microbial disease at high resolution: following the fate of the bacterium during infection. *Current Opinion in Microbiology*, *15*(1), 23-27.

Dougan, G., & Baker, S. (2014). *Salmonella enterica* serovar Typhi and the pathogenesis of typhoid fever. *Annual Review of Microbiology*, *68*, 317-336.

Dybowski, R., Restif, O., Goupy, A., Maskell, D., Mastroeni, P., & Grant, A. (2015). Single passage in mouse organs enhances the survival and spread of *Salmonella enterica*. *Journal of the Royal Society Interface*, *12*, 20150702.

Gerlini, A., Colomba, L., Furi, L., Braccini, T., Manso, A. S., Pammolli, A., et al. (2014). The role of host and microbial factors in the pathogenesis of pneumococcal bacteraemia arising from a single bacterial cell bottleneck. *PLoS Pathogens*, *10*(3), e1004026.

Gillespie, D. (1997). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, *81*, 2340-236.

Grant, A., Restif, O., McKinley, T., Sheppard, M., Maskell, D., & Mastroeni, P. (2008). Modelling within-host spatiotemporal dynamics of invasive bacterial disease. *PLoS Biology*, *6*(4), e74.

Helaine, S., & Kugelberg, E. (2014). Bacterial persisters: formation, eradication, and experimental systems. *Trends in Microbiology*, *22*, 417-424.

Jacob, P., & Oliveira, P. (2011). Relative smoothing of discrete distributions with sparse observations. *Journal of Statistical Computation and Simulation*, *81*(1), 109-121.

Kaiser, P., Regoes, R., Dolowschiak, T., Wotzka, S., Lengefeld, J., Slack, E., et al. (2014). Cecum lymph node dendritic cells harbor slow-growing bacteria phenotypically tolerant to antibiotic treatment. *PLoS Biology*, *12*(2), e1001793.

Kono, M., Zafar, M., Zuniga, M., Roche, A., Hamaguchi, S., & Weiser, J. (2016). Single cell bottlenecks in the pathogenesis of *Streptococcus pneumoniae*. *PLoS Pathogens*, *12*(10), e1005887.

Krimbas, C., & Tsakas, S. (1971). The genetics of *Dacus okae*. V. Changes of esterase polymorphism in a natural populatian following insecticide control-selection or drift? *Evolution*, *25*, 454-460.

Lim, C., Voedisch, S., Wahl, B., Rouf, S., Geffers, R., Rhen, M., et al. (2014).

Independent bottlenecks characterize colonization of systemic compartments and gut lymphoid tissue by *Salmonella*. *PLoS Pathogens*, *10*(7), e1004270.

Lowe, D., Ernst, S., Zito, C., Ya, J., & Glomski, I. (2013). Bacillus anthracis has two independent bottlenecks that are dependent on the portal of entry in an intranasal model of inhalational infection. *Infection & Immunity*, *81*(12), 4408-4420.

Maier, L., Diard, M., Sellin, M., Chouffane, E.-S., Trautwein-Weidner, K., Periaswamy, B., et al. (2014). Granulocytes impose a tight bottleneck upon the gut luminal pathogen population during *Salmonella* Typhimurium colitis. *PLoS Pathogens*, *10*(12), e1004557.

Pamilo, P., & Varvio-Aho, S. (1980). On the estimation of population size from allele frequency changes [Letter to the Editor]. *Genetics*, *95*(4), 1055-1057.

Requena, F., & Cludad, M. (2003). The maximum probability 2 x c contingency tables and the maximum probability points of the multivariate hypergeometric distribution. *Communications in Statistics - Theory and Methods*, *32*(9), 1737-1752.

Ruppert, D., & Wand, M. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, *22*, 1346-1370.

Schwartz, D., Chen, S., Hultgren, S., & Seed, P. (2011). Population dynamics and niche distribution of uropathogenic *Escherichia coli* during acute and chronic urinary tract infection. *Infection & Immunity*, *79*(10), 4250-4259.

Simonoff, J. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

Smith, H. (2000). Questions about the behaviour of bacterial pathogens *in vivo*. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *355*, 551-564.

Sobel Leonard, A., Weissman, D., Greenbaum, B., Ghedin, E., & Koelle, K. (2017). Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *Journal of Virology*. Available from `http://jvi.asm.org/content/early/2017/04/27/JVI.00171-17.abstract`

Sourdis, J., & Krimbas, C. (1980). On Pamilo and Varvio-Aho's note about the estimation of effective population size [Letter to the Editor]. *Genetics*, *96*(2), 561-563.

List of Figures

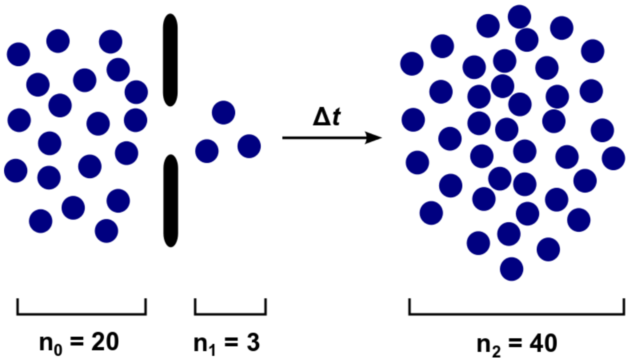


Figure 1: Diagrammatic representation of a bottleneck. The bottleneck is derived from a pre-bottleneck population by random sampling without replacement. A post-bottleneck population arises after time Δt according to a defined growth mechanism. n_1 is the size of the bottleneck, n_0 the size of the pre-bottleneck population, and n_2 the size of the post-bottleneck population.

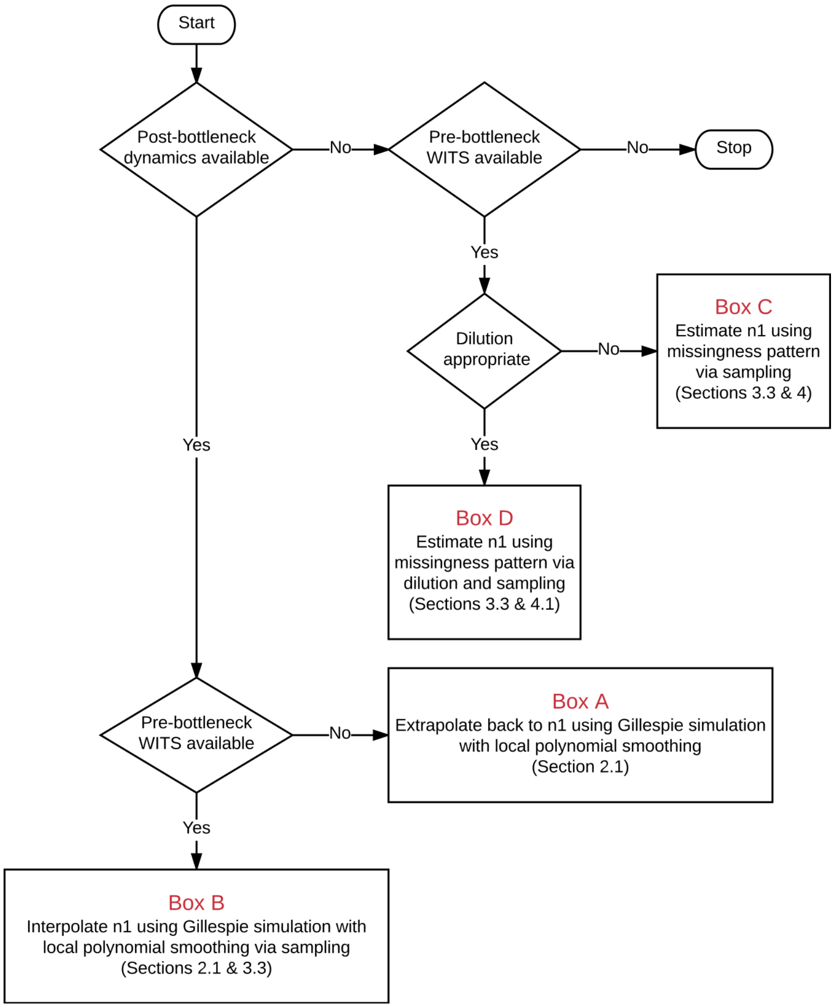
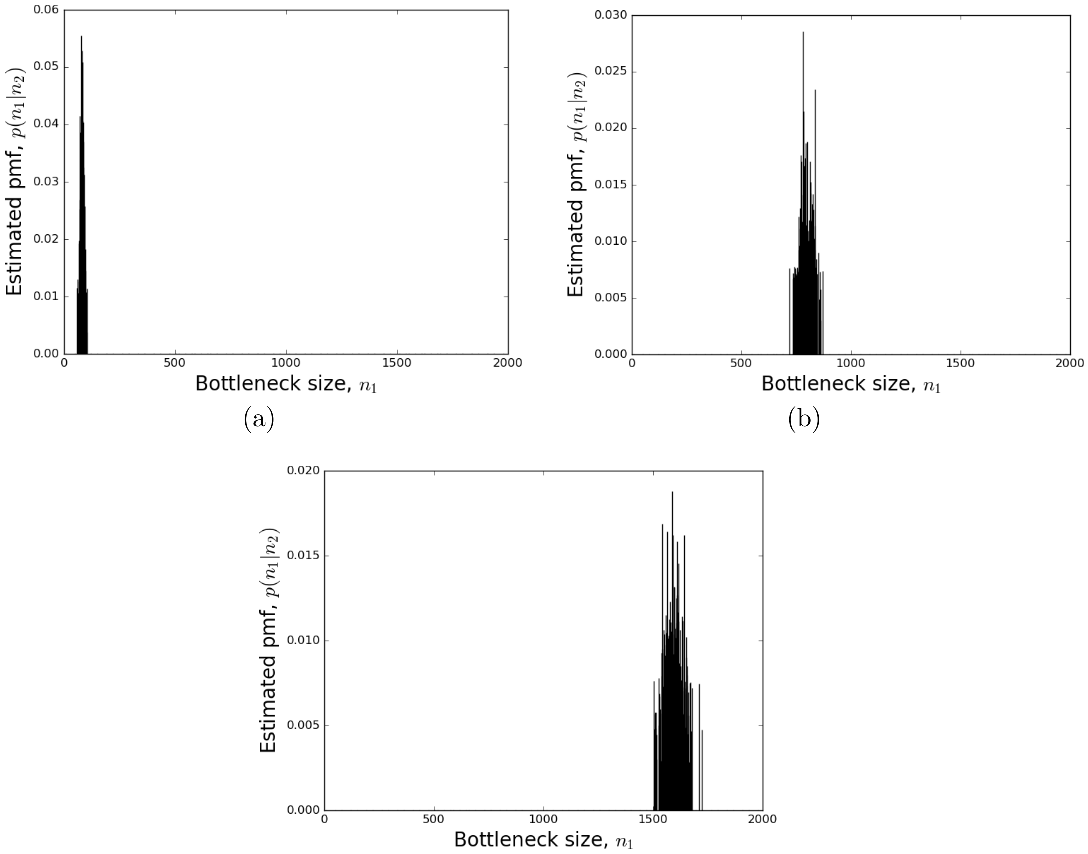

Start
Post-bottleneck dynamics available
No
Pre-bottleneck WITS available
No
Stop
Yes
Dilution appropriate
No
Box C
Estimate n1 using missingness pattern via sampling (Sections 3.3 & 4)
Yes
Box D
Estimate n1 using missingness pattern via dilution and sampling (Sections 3.3 & 4.1)
Yes
Pre-bottleneck WITS available
No
Box A
Extrapolate back to n1 using Gillespie simulation with local polynomial smoothing (Section 2.1)
Yes
Box B
Interpolate n1 using Gillespie simulation with local polynomial smoothing via sampling (Sections 2.1 & 3.3)

Figure 2: Flowchart summarizing alternative strategies to bottleneck-size estimation depending on circumstances. The box labels enable relevant parts of the main text to be linked to this flowchart.



(c)

Figure 3: Posterior bottleneck-size distributions $p(n_1|n_2)$ estimated using Algorithm 1. Bottlenecks of size (a) $n_1 = 80$, (b) $n_1 = 800$, and (c) $n_1 = 1600$, were artificially induced.

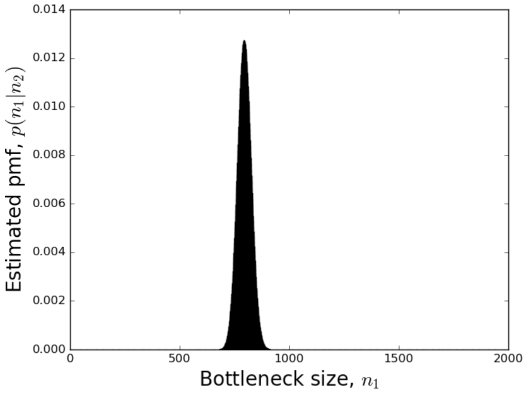


Figure 4: Posterior bottleneck-size distribution estimated using convolution of eight posterior distributions associated with the eight WITS. Target $n_1 = 800$. Summary statistics are mode 793, mean 793.8, median 795 and variance 980.1. See Section 3 for details.

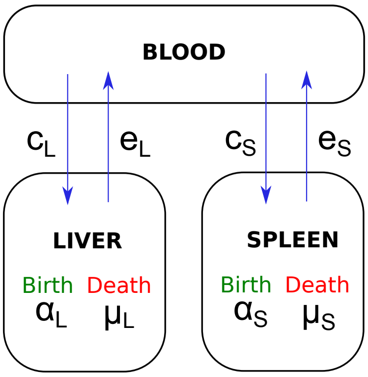


Figure 5: Schematic representation of the spread of *S. enterica* from the blood and between the liver and the spleen. The dynamics is governed by a set of parameters $\boldsymbol{\theta}$; namely, the per capita division rates (α_L, α_S), death rates (μ_L, μ_S), immigration rates (c_L, c_S) and clearance rates (e_L, e_S).

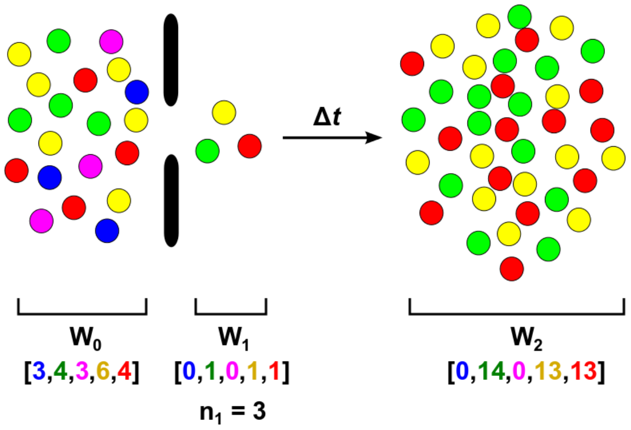


Figure 6: Diagrammatic representation of a bottleneck and its propensity to stochastic variability in terms of both its size (n_1) and the composition of its population ($\mathbf{w}_1$). The array of numbers below each population states the number of tagged bacteria present in that population.

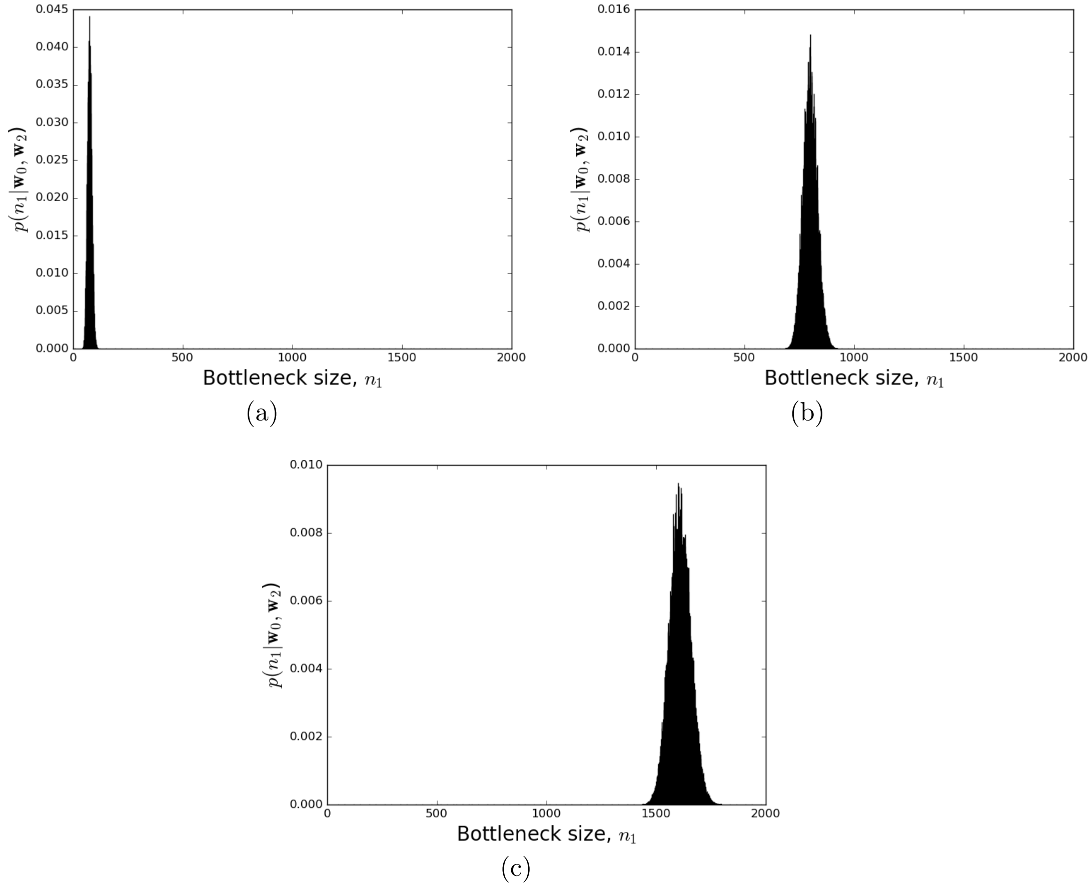
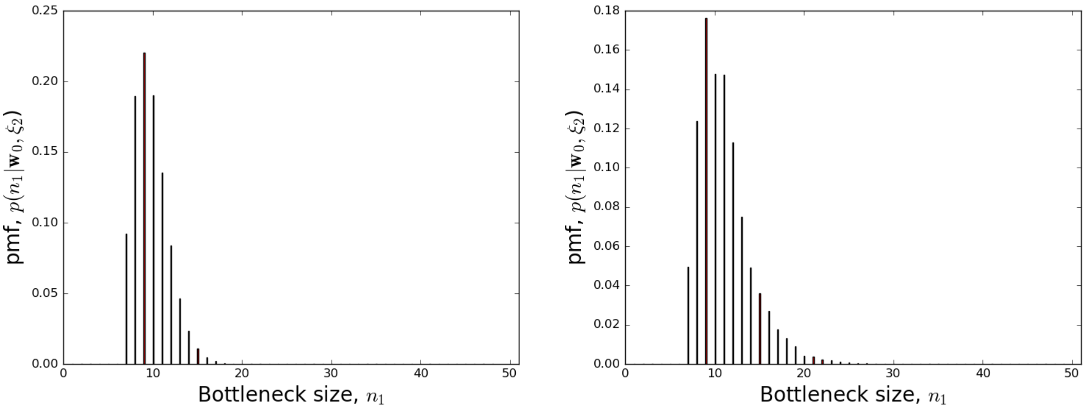


Figure 7: Posterior bottleneck-size distributions $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$ estimated using Algorithm 2. Target bottlenecks of size (a) $n_1 = 80$, (b) $n_1 = 800$, and (c) $n_1 = 1600$, were artificially induced. See Section 3.3 for details.



(a) (b)

Figure 8: Posterior bottleneck-size distributions obtained (a) exactly using equation (12) with $\mathbf{w}_0 = \langle 4^{[1]}, 4^{[2]}, \ldots, 4^{[8]} \rangle$, and (b) estimated using 1000 samples with $\mathbf{w}_0 = \langle 600^{[1]}, 600^{[2]}, \ldots, 600^{[8]} \rangle$. Target is $n_1 = 7$ in both cases.

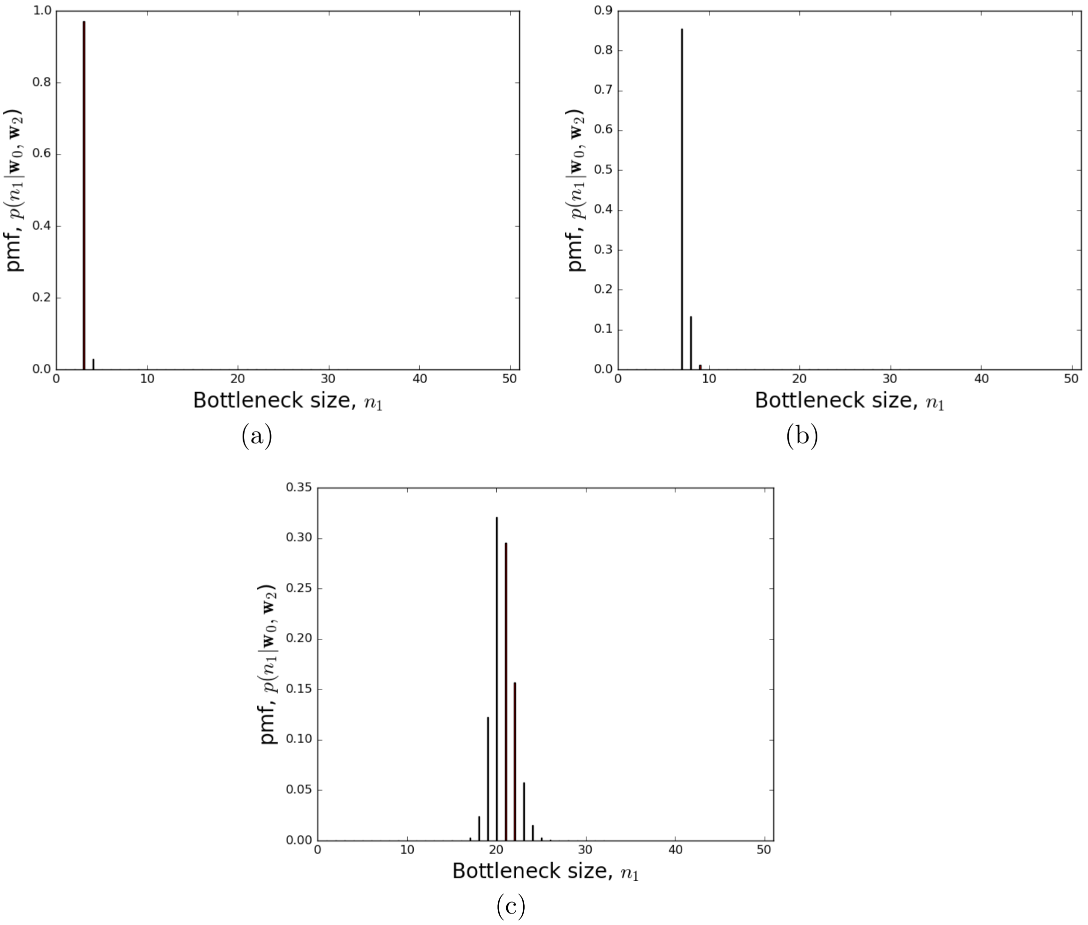
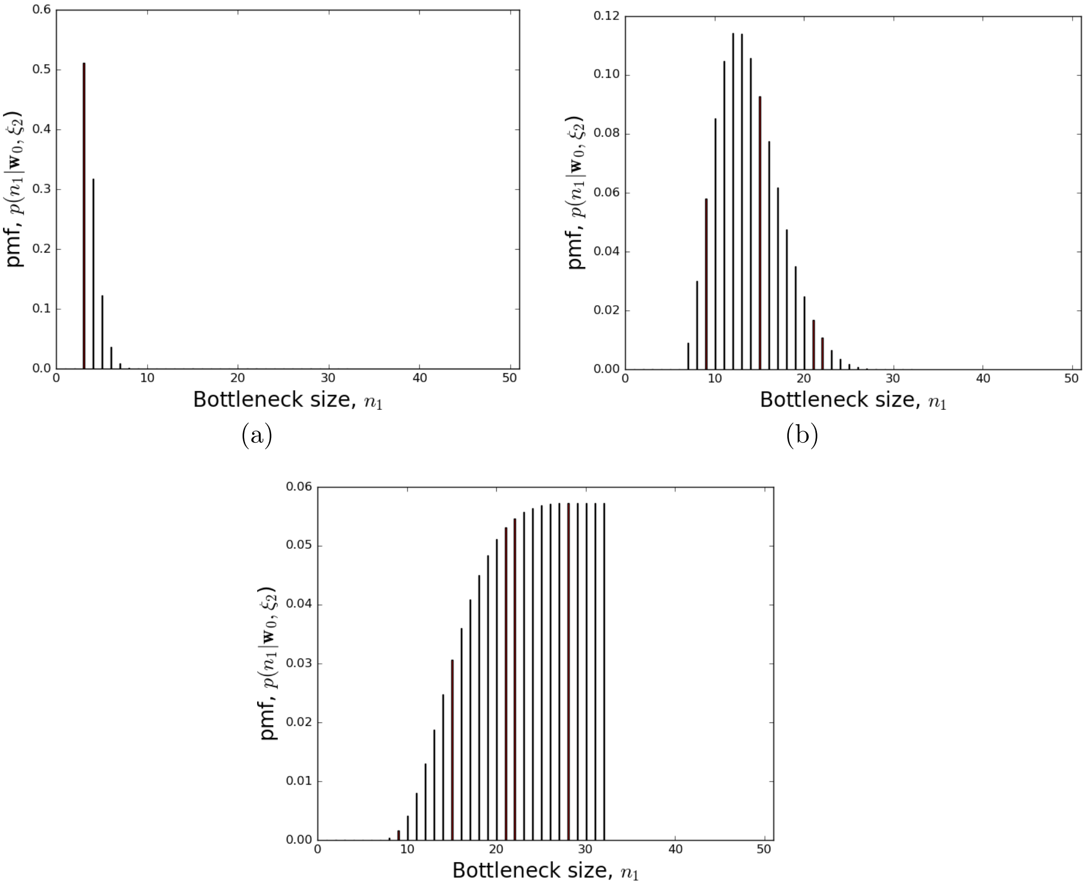


Figure 9: Posterior bottleneck-size distributions $p(n_1|\mathbf{w}_0, \mathbf{w}_2)$ determined accurately using (11). Target bottlenecks of size (a) $n_1 = 3$, (b) $n_1 = 7$, and (c) $n_1 = 20$, were used. See Section 4 for details.

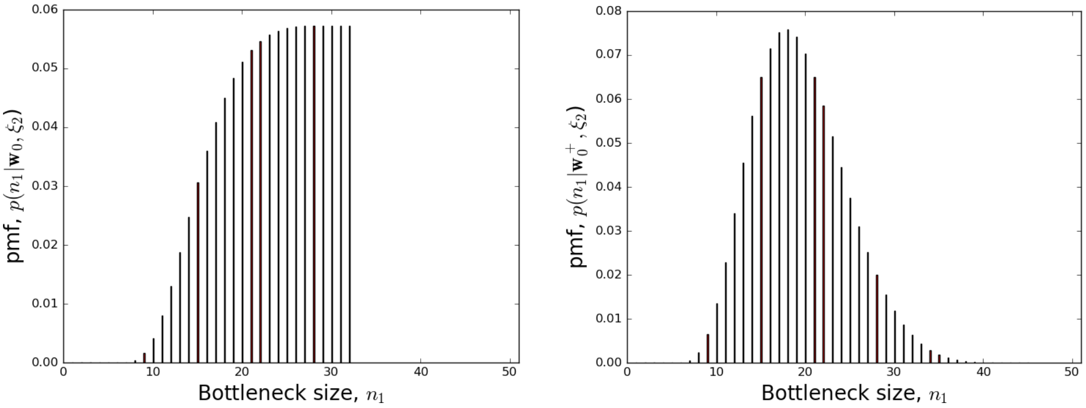


(a)

(b)

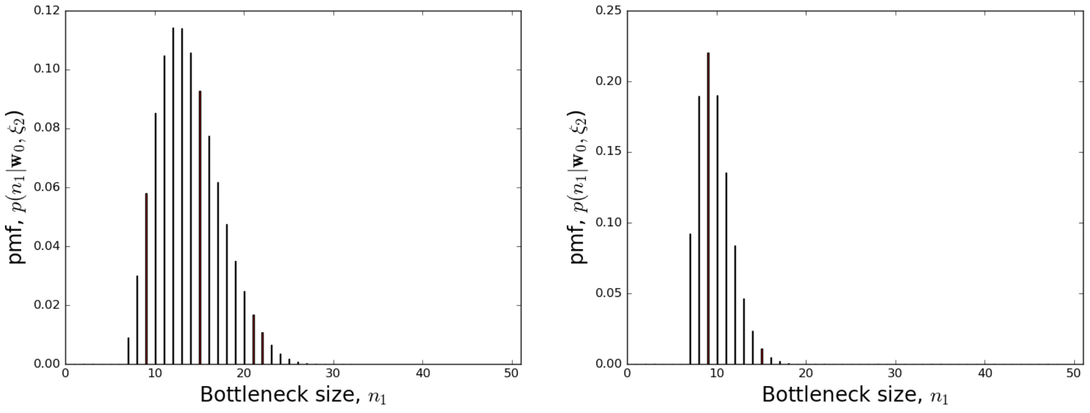
(c)

Figure 10: Posterior bottleneck-size distributions $p(n_1|\mathbf{w}_0, \xi_2)$ determined using (12). Target bottlenecks of size (a) $n_1 = 3$, (b) $n_1 = 7$, and (c) $n_1 = 20$, were used. See Section 4 for details.



(a) (b)

Figure 11: Posterior bottleneck-size distributions (a) without dilution and (b) with dilution. Distributions $p(n_1|\mathbf{w}_0, \xi_2)$ and $p(n_1|\mathbf{w}_0^+, \xi_2)$ were determined using (12) and (13), respectively. The target bottleneck size was $n_1 = 20$ in both cases. For the dilution, $u = 13$ untagged bacteria were present in $\mathbf{w}_0^+$. See Section 4.1 for details.



(a)

(b)

Figure 12: Posterior bottleneck-size distributions obtained (a) with 8 WITS and (b) with 12 WITS. Note the decrease in variance.

List of Tables

Table 1: Symbols used in this article.

Symbol(s)	Meaning
$\mathcal{P}_1$	Bottleneck population.
n_1	Size of $\mathcal{P}_1$.
$\mathcal{P}_0$ and $\mathcal{P}_1$	Pre- and post-bottleneck populations.
n_0 and n_2	Sizes of $\mathcal{P}_0$ and $\mathcal{P}_2$.
Δt	Time period from $\mathcal{P}_1$ and $\mathcal{P}_2$.
α_L and μ_L	Per capita division and death rates in liver.
c_L and e_L	Per capita immigration and clearance rates in liver.
α_S and μ_S	Per capita division and death rates in spleen.
c_S and e_S	Per capita immigration and clearance rates in spleen.
$\boldsymbol{\theta}$	Parameter set $\{\alpha_L, \mu_L, c_L, e_L, \alpha_S, \mu_S, c_S, e_S\}$ for $\mathcal{P}_1$.
$w_j^{[i]}$	Number of bacteria in $\mathcal{P}_j$ with the i-th WITS tag.
$\mathbf{w}_j$	The set $\{w_j^{[1]}, \ldots, w_j^{[8]}\}$.
$\pi_0^{[i]}$	Probability of an i-th tagged bacterium in $\mathcal{P}_0$ continuing into $\mathcal{P}_1$.
$\boldsymbol{\pi}_0$	The set $\{\pi_0^{[1]}, \ldots, \pi_0^{[8]}\}$.
ξ_2	Vector indication of which WITS in $\mathbf{w}_2$ are missing.

Table 2: The accuracy of Algorithm 1 in terms of summary statistics for the estimated posterior distribution $\widehat{p}(n_1|n_2^\star)$. See Section 2.2 for details.

Target $n_1^\star$	$n_2^\star$	Mode	Mean	Median	SD
1	3	1	1.9	1	1.1
2	8	2	2.7	2	1.4
4	17	4	4.5	4	1.9
8	35	7	8.4	8	3.2
80	376	76	79.2	81	9.5
800	3731	782	794.9	794	30.8
1600	7507	1587	1595.1	1597	42.3