A Bayesian model of acquisition and clearance of bacterial colonization incorporating within-host variation

Marko Järvenpää[1,2], Mohamad R. Abdul Sater[2], Georgia K. Lagoudas[3,4], Paul C. Blainey[3,4], Loren G. Miller[5], James A. McKinnell[5,6], Susan S. Huang[7], Yonatan H. Grad[2‡*], Pekka Marttinen[1‡*]

1 Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland
2 Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Boston, MA, USA
3 Department of Biological Engineering, MIT, Cambridge, MA, USA
4 Broad Institute, MIT and Harvard, Cambridge, MA, USA
5 Infectious Disease Clinical Outcomes Research Unit (ID-CORE), Division of Infectious Diseases, Los Angeles Biomedical Research Institute, Harbor-UCLA Medical Center, Torrance, CA
6 Los Angeles County Department of Public Health, Acute Communicable Disease Control Unit
7 Division of Infectious Diseases and Health Policy Research Institute, University of California, Irvine School of Medicine, Irvine, CA, USA

‡These authors contributed equally to this work.
* ygrad@hsph.harvard.edu, pekka.marttinen@aalto.fi

Abstract

Bacterial populations that colonize a host can play important roles in host health, including serving as a reservoir that transmits to other hosts and from which invasive strains emerge, thus emphasizing the importance of understanding rates of acquisition and clearance of colonizing populations. Studies of colonization dynamics have been based on assessment of whether serial samples represent a single population or distinct colonization events. With the use of whole genome sequencing to determine genetic distance between isolates, a common solution to estimate acquisition and clearance rates has been to assume a fixed genetic distance threshold below which isolates are considered to represent the same strain. However, this approach is often inadequate to account for the diversity of the underlying within-host evolving population, the time intervals between consecutive measurements, and the uncertainty in the estimated acquisition and clearance rates. Here, we present a fully Bayesian model that provides probabilities of whether two strains should be considered the same, allowing us to determine bacterial clearance and acquisition from genomes sampled over time. Our method explicitly models the within-host variation using population genetic simulation, and the inference is done using a combination of Approximate Bayesian Computation (ABC) and Markov Chain Monte Carlo (MCMC). We validate the method with multiple carefully conducted simulations and demonstrate its use in practice by analyzing a collection of methicillin resistant Staphylococcus aureus (MRSA) isolates from a large recently completed longitudinal clinical study. An R-code implementation of the method is freely available at: `https://github.com/mjarvenpaa/bacterial-colonization-model.git`.

Author summary

As colonizing bacterial populations are the source for much transmission and a reservoir for infection, they are a major focus of interest clinically and epidemiologically. Understanding the dynamics of colonization depends on being able to confidently identify acquisition and clearance events given intermittent sampling of hosts. To do so, we need a model of within-host bacterial population evolution from acquisition through the time of sampling that enables estimation of whether two samples are derived from the same population. Past efforts have frequently relied on empirical genetic distance thresholds that forgo an underlying model or employ a simple molecular clock model. Here, we present an inferential method that accounts for the timing of sample collection and population diversification, to provide a probabilistic estimate for whether two isolates represent the same colonizing strain. This method has implications for understanding the dynamics of acquisition and clearance of colonizing bacteria, and the impact on these rates by factors such as sensitivity of the sampling method, pathogen genotype, competition with other carriage bacteria, host immune response, and antibiotic exposure.

Introduction

Colonizing bacterial populations are often the source of infecting strains and transmission to new hosts [1–5], making it important to understand the dynamics of these populations and the factors that contribute to persistent colonization and to the success or failure of clinical decolonization protocols. The study of colonization dynamics is based on inferring whether bacteria from samples collected over time represent the same population or distinct colonization events, thereby permitting calculation of rates of acquisition and clearance [6, 7]. Whole genome sequencing has provided a detailed measure of genetic distance between isolates, which can then be used to infer the relationship between them [8–11]. While to date most studies have used genetic distance thresholds as the basis for determining the relationship between isolates [8, 10], here we improve on these heuristic strategies and present a robust and accurate fully Bayesian model that provides probabilities of whether two strains should be considered the same, allowing us to determine bacterial clearance and acquisition from genomes sampled over time.

An example of a typical individual-level longitudinally sampled data set from a study population is shown in Fig 1: each ’row’ represents a patient, x-axis is time, and dots are the genomes sampled at multiple time points. Dot color refers to different, easily distinguishable, sequence types (ST). The coloured number between two consecutive samples reflects the distance between the genomes, and we see that even within the same ST the distances may vary considerably, and, therefore, determining whether the changes can be explained by within-host evolution only, is challenging. Intuitively, if two genomes are very similar, we interpret this as a single strain colonizing the host. On the other hand, two very different genomes, even if the same ST, are interpreted as two different strains, obtained either jointly or separately as two acquisitions. With these data, we would like to address questions including: to what extent are people persistently colonized, cleared, and recolonized? If recolonized, what is the likelihood that it is the same or a distinct strain? To address these questions, previous works have relied on using a threshold number of single nucleotide polymorphisms (SNPs) to define a strain. Optimally, however, the SNP distance between the genomes observed and the interval between the sampling time defines a probability that the two genomes represent the same strain. Such data are critical for understanding within-host dynamics, response to interventions, and transmission.

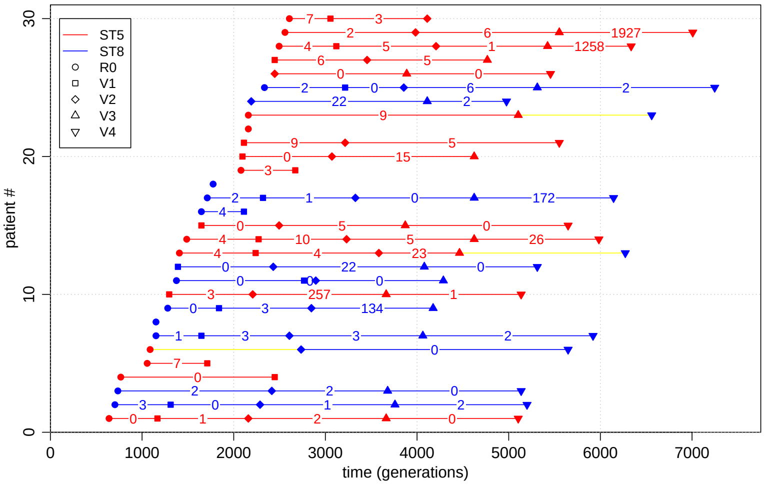

Fig 1. Illustration of a subset of the data used in the study. Each row corresponds to one patient and only the first 30 patients are shown. R0 is the initial hospital visit and V1, V2 etc. are the further visits. Red colour refers to ST5 and blue to ST8 and the coloured numbers are the amount of mutations d_i. Yellow colour highlights the cases where the ST changes from ST 5 to ST 8.

Previously, transitions between different colonizing bacteria have been modeled using hidden Markov models [12] with states corresponding to different colonizing STs. However, this approach is not suitable for modeling within a single ST, where acquisition and clearance must be determined based on a small number of mutations. Crucial for interpreting such small differences is a model for within-host variation [8, 13], specifying the number of mutations expected by evolution within the host. Population genetic models can be used for understanding the variation in an evolving population [14]. A major difficulty in fitting such models to data like those shown in Fig 1 is that the information contained by the data is extremely limited regarding the variation within the host: a single time point is summarized with just a single (or a few) genomes, and must serve to represent the whole within-host population. While some studies use genome sequence from multiple isolates to achieve a more complete characterization of within-host diversity [3, 10], these tend to be limited in terms of the number of time points and/or patients.

The Bayesian statistical framework can be used to combine information from multiple data sources. In the Bayesian approach, a prior distribution is updated using the laws of probability into a posterior distribution in the light of the observations, and this can be repeated multiple times with different data sets [15, 16]. Approximate Bayesian computation (ABC) is particularly useful with population genetic models, where the likelihood function may be difficult to specify explicitly, but simulating the model is straightforward [17, 18]. ABC has recently been introduced in bacterial population genetics [19–22]. Here, we present a Bayesian model for determining whether two genomes should be considered the same strain, enabling a strategy grounded in population genetics to make inferences about acquisition and clearance from data of closely related genomes. Benefits of the fully Bayesian analysis include: rigorous quantification of uncertainty, explicit statement of modeling assumptions (open for criticism and further development when needed), and straightforward utilization of multiple data sources. We demonstrate these benefits by analyzing a large collection longitudinally collected methicillin resistant *Staphylococcus aureus* (MRSA) genomes, obtained through a clinical trial (Project CLEAR) to evaluate the effectiveness of an MRSA decolonization protocol [23]. This method for identifying strains with explicit assessment of uncertainty will enable studies of the characteristics–both host and pathogen–that impact colonization in the presence and absence of interventions.

Methods

Overview of the model

One input data item for our model consists of a pair of genomes that are of the same ST, sampled from the same individual at two consecutive time points (or possibly with an intervening time point with no samples or a sample of a different ST). Each of these data items (i.e. pairs of consecutive genomes) is summarized in terms of two quantities: the distance between the genomes and the difference between their sampling times (see Fig 1). Hence, the observed data D can be written as consisting of pairs (d_i, t_i), $i = 1, \ldots, N$, where $t_i > 0$ is the time between the sampling of the genomes, $d_i \in \{0, 1, 2, \ldots\}$ is the observed distance, and N the total number of genome pairs that satisfy the criteria (i.e. same patient, same ST, consecutive time points or possibly with an intervening time point with no samples or a sample of a different ST). The restriction to genome pairs of the same ST stems from the fact that different STs will always be considered different strains anyway.

There are two possible explanations for the observed distances. If the genomes are from the same strain, we expect their distance to be relatively small. If the genomes are from different strains, we expect a greater distance. Below we define two probabilistic models that represent these two alternative explanations. These models are then combined into one overall mixture model, which assumes that the distance between a certain pair of genomes is generated either from the 'same strain' model or the 'different strain' model, and enables calculation of the probabilities of these two alternatives for each genome pair, rather than relying on a fixed threshold to distinguish between them.

An essential part of our approach is a population genetic simulation which allows us to model the within-host variation, and hence make probabilistic statements of the plausibilities of the 'same strain' vs. 'different strain' models. For this purpose, we adopt the common Wright-Fisher (W-F) simulation model, see e.g. [24], with a constant mutation rate and population size, which are estimated from the data. The simulation is started with all genomes being the same, which corresponds to a biological scenario according to which a colonization begins with a single isolate multiplying rapidly until reaching the maximum 'capacity', followed by slow diversification of the population. This assumption is supported by the fact that in the distance distribution, in cases where the acquisition time was known and had happened recently, very little variation was observed in the population. See the Discussion section for more details on the modeling assumptions. Overview of the approach, including data sets, models, and methods for inference, is outlined in Fig 2 and discussed below in detail.

Model p_S: Same strain

Let (s_{i1}, s_{i2}) denote a pair of genomes with distance d_i, sampled from a patient at two consecutive time points (see the previous section) with time t_i between taking the samples. Here we present a model, i.e., a probability distribution $p_S(d_i \,|\, t_i, n_{\text{eff}}, \mu)$, which tells what kind of distances we should expect if the genomes are from the same strain. The parameter n_{eff} is the effective population size and μ is the mutation rate. We model d_i as

$$d_i = d_{i1} + d_{i2} \tag{1}$$

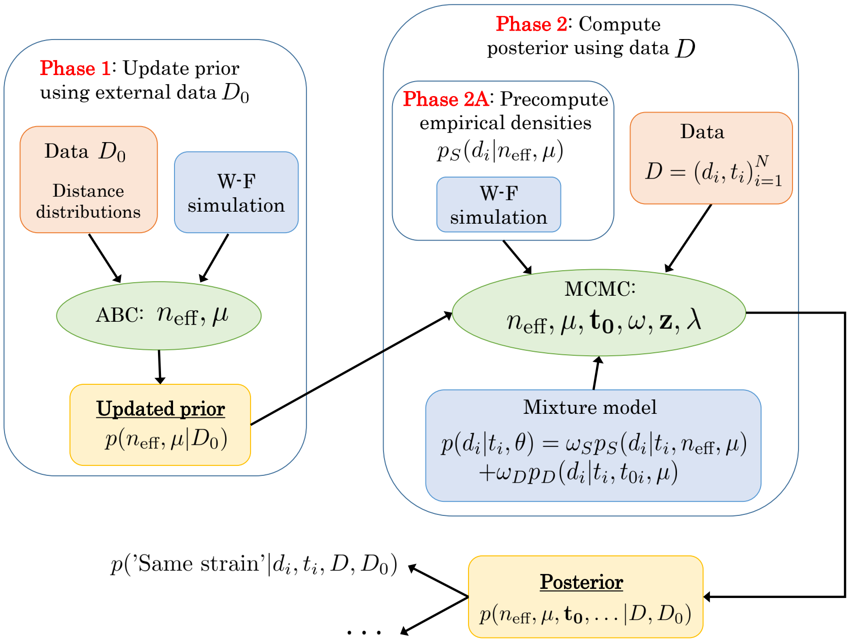
Phase 1: Update prior using external data D_0
Data D_0
Distance distributions
W-F simulation
ABC: n_{eff}, μ
Updated prior
$p(n_{\text{eff}}, \mu | D_0)$
Phase 2: Compute posterior using data D
Phase 2A: Precompute empirical densities $p_S(d_i | n_{\text{eff}}, \mu)$
W-F simulation
Data
$D = (d_i, t_i)_{i=1}^N$
MCMC:
$n_{\text{eff}}, \mu, \mathbf{t_0}, \omega, \mathbf{z}, \lambda$
Mixture model
$p(d_i|t_i,\theta) = \omega_S p_S(d_i|t_i, n_{\text{eff}}, \mu)$
$+\omega_D p_D(d_i|t_i, t_{0i}, \mu)$
$p(\text{'Same strain'}|d_i, t_i, D, D_0)$
...
Posterior
$p(n_{\text{eff}}, \mu, \mathbf{t_0}, \ldots |D, D_0)$

Fig 2. Overview of the modeling and data fitting steps. In Phase 1 we update our prior information on parameters (n_{eff}, μ) based on external data D_0. In phase 2 we estimate all the parameters of the (mixture) model using MCMC, precomputed distance distributions p_S and the information obtained in Phase 1. The fitted model can be used to e.g. obtain the same strain probability for a new (future) measurement.

where we have defined

$$d_{i1} = \text{dist}(s_{i1}, s_{i*}) \quad \text{and} \quad d_{i2} = \text{dist}(s_{i*}, s_{i2}), \tag{2}$$

where dist$(\cdot,\cdot)$ is a distance function that tells the number of mutations between its arguments, and s_{i*} is the unique ancestor of s_{i2} that was present in the host when s_{i1} was sampled, and which has descended within the host from the same genome as s_{i1} (see Fig 3A). The Equation 1 is valid when mutations between s_{i1} and s_{i*}, and s_{i*} and s_{i2} have occurred in different sites, which is true with a high probability when the genomes are long (millions of bps) compared to the number of mutations (dozens or a few hundred at most). The probability distribution of d_{i1} which we will denote by $p_{\text{sim}}(d_{i1} \,|\, n_{\text{eff}}, \mu)$, and which is not available analytically and does not depend on t_i, represents the within-host variation at a single time point, and we approximate it as

$$p_{\text{sim}}(d_{i1} \,|\, \mu, n_{\text{eff}}) = \text{WF-simulator}(d_{i1} \,|\, \mu, n_{\text{eff}}). \tag{3}$$

The distribution of d_{i2} is assumed to be

$$d_{i2} \,|\, \mu, t_i \sim \text{Poisson}(d_{i2} \,|\, \mu t_i), \tag{4}$$

that is, mutations are assumed to occur according to a Poisson process with the rate parameter μ.

Model p_D: Different strains

Model p_D represents the case that the genomes s_{i1} and s_{i2} are from different strains, which we define to mean that their most recent common ancestor (MRCA), denoted by

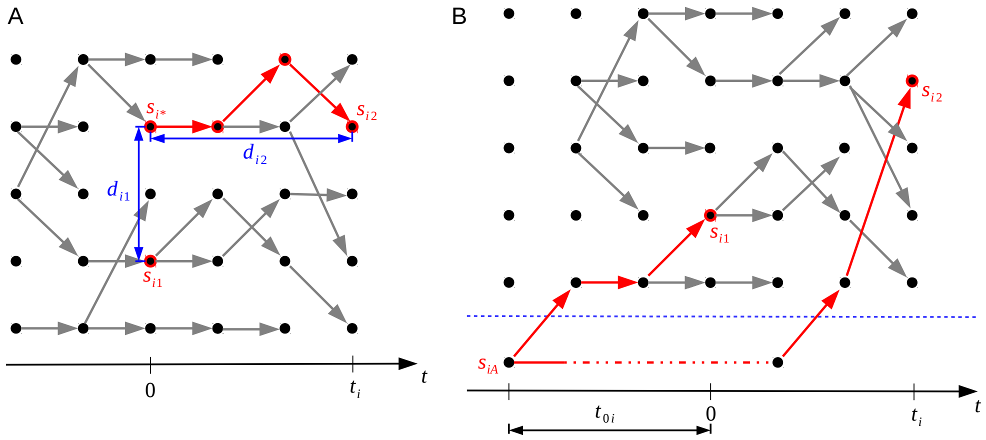

Fig 3. Outline of the 'same strain' and 'different strain' models. Model p_D on the left (panel A) represents the situation where the genomes denoted by s_{i1} and s_{i2} are of the same strain. Model p_S on the right (panel B) shows the case where these genomes are of different strains. Time flows from left to right in the figures, the dots represent individual genomes, and the edges parent-offspring relationships.

s_{iA}, resided outside the host. The time between s_{iA} and s_{i1} is denoted by t_{0i} (see Fig 3B). Under model p_D, we assume that the distribution of the distance d_i is

$$p_D(d_i \,|\, \mu, t_i, t_{0i}) = \text{Poisson}(d_i \,|\, \mu(2t_{0i} + t_i)), \tag{5}$$

where the values of t_{0i} are unknown and will be estimated, but let us assume for now that they are known. One difference between the same strain model p_S (defined by Equations 1, 3, 4) and the different strain model p_D (Equation 5) is that the former uses Wright-Fisher simulation, whereas the latter does not. The reason is that the within-host variation is bounded, occasionally increasing and decreasing, which is reflected by the constant population size of the Wright-Fisher simulation in the same strain model. On the other hand, in the different strain model the distance between s_{i1} and s_{i2} can in principle increase without bound, given enough time since their common ancestor, because they diverged and evolved outside the host.

Mixture model

With the two alternative models for the distance, we can write the full model, which assumes that each distance observation is distributed according to

$$p\,(d_i \,|\, t_i, \boldsymbol{\theta}) = \omega_S p_S(d_i \,|\, t_i, n_{\text{eff}}, \mu) + \omega_D p_D(d_i \,|\, t_i, t_{0i}, \mu), \quad i = 1, \ldots, N, \tag{6}$$

where $\boldsymbol{\theta}$ denotes jointly all the parameters of the models, i.e., $\boldsymbol{\theta} = (n_{\text{eff}}, \mu, \omega_S, \omega_D, t_{01}, \ldots, t_{0N})$. The parameter ω_S represents the proportion of pairs from the same strain and ω_D is the proportion of pairs from different strains, such that $\omega_S + \omega_D = 1$. To learn the unknown parameters $\boldsymbol{\theta}$, we need to fit the model to data, but before going into details, we discuss how to use an external data set to update the prior distribution about the mutation rate μ and the effective sample size n_{eff}. This updated distribution will itself be used as the prior in the mixture model.

ABC inference to update the prior using external data

Simulations with the W-F model are used in our approach for two purposes: 1) to incorporate information from an external data set to update the prior on the mutation rate μ and the effective sample size n_{eff}, and 2) to define empirically the distribution $p_S(d_i|t_i, n_{\text{eff}}, \mu)$ required in the mixture model. Here we discuss the first task.

As external data we use measurements from eight patients colonised with MSSA [3], comprising nasal swabs from two time points for each patient, such that the acquisition is known to have happened approximately just before the first swab. Multiple genomes were sequenced from each sample, and the distributions of pairwise distances between the genomes provide snapshots to the within-host variability at the two time points for each individual, and these distance distributions are used as data. We exclude one patient (number 1219) because according to [3] this patient was likely infected already long before the first sample. The data set also contains observations from an additional 13 patients from [13], denoted by letters from A to M in [3]. For these patients, distance distributions from only one time point are available, and the acquisition times are unknown. The data comprising the distance distributions from the 7 patients (two time points) and the additional 13 patients (a single time point) are jointly denoted by D_0.

To learn about the unknown parameters n_{eff} and μ, we first note that their values affect the distance distribution of a population resulting from a W-F simulation with the specified values (Fig. 4). To estimate these parameters, we try to find such values for them which make the output of the W-F similar to the observed distance distributions D_0. Since the corresponding likelihood function is unavailable, standard statistical techniques for model fitting do not apply. Therefore, we use Approximate Bayesian Computation (ABC), a class of methods for Bayesian inference when the likelihood is either unavailable or too expensive to evaluate but simulating the model is feasible, see [17, 18, 25, 26] for an overview on ABC. The basic ABC rejection sampler algorithm for the model fitting consists of the following steps:

1. Simulate a parameter vector (n_{eff}, μ) from the prior distribution $p(n_{\text{eff}}, \mu)$.

2. Generate a pseudo-data similar to the observed data D_0 by running the W-F model separately for each patient using the parameter (n_{eff}, μ).

3. Accept the parameter (n_{eff}, μ) as a sample from the (approximate) posterior distribution if the discrepancy between the observed and simulated data is smaller than a specified threshold ε.

The quality of the resulting ABC approximation depends on the selection of the discrepancy function, the threshold ε and the number of accepted samples. Broadly speaking, if the discrepancy summarizes the information in the data completely (e.g. it is a function of the sufficient statistics) and ε is arbitrarily small, the approximation error becomes negligible and the samples are generated from the exact posterior. In practice, choosing ε very small makes the algorithm inefficient since many simulations are needed to obtain an accepted sample even with the optimal value of the parameter. Also, finding a good discrepancy function may be difficult because sufficient statistics are typically unavailable. Many sophisticated ABC variants exist, see e.g. [18, 26] and the references therein, but as we need to estimate only two parameters (one of which is discrete) and because running the simulations in parallel is straightforward with the basic algorithm, we use a the ABC rejection sampler outlined above, with details discussed below.

In [13], MRSA evolution was simulated using parameters derived from the following estimates: 8 mutations per genome per year and generation length of 90 minutes (the whole year is thus 5840 generations). This gives mutation rate of 0.0019 per genome per generation, approximately 6.3×10^{-10} mutations per site per generation assuming the genome length of 3 Mbp. We also use the generation time of 90 minutes, originally derived by [13] from the estimated doubling time of *Staphylococcus aureus* [27]. We use independent uniform priors for the parameters of the W-F model, so that

$$n_{\text{eff}} \sim \mathcal{U}(\{20, 21, \ldots, 10000\}), \quad \mu \sim \mathcal{U}([a_\mu, b_\mu]) \tag{7}$$

with $a_\mu = 0.00005$ and $b_\mu = 0.005$ mutations per genome per generation.

We argue that reasonable parameters should produce populations with similar histograms of the pairwise distances compared to the observations at the corresponding times. Consequently, we use the discrepancy Δ defined as

$$\Delta = \sum_{i=1}^{7}\sum_{j=1}^{2} l_1(\widehat{p}_{ij}(n_{\text{eff}}, \mu), \widehat{p}_{ij}^{\text{obs}}) + \sum_{i \in \{A,B,\ldots,M\}} \min_j \{l_1(\widehat{p}_{ij}(n_{\text{eff}}, \mu), \widehat{p}_{i1}^{\text{obs}})\}, \tag{8}$$

where $\widehat{p}_{ij}(n_{\text{eff}}, \mu)$ and $\widehat{p}_{ij}^{\text{obs}}$ are the simulated and observed empirical distributions of pairwise distances for patient i with time point j, respectively, and $l_1(\cdot, \cdot)$ denotes the L^1 distance between the distributions. In principle, the unknown acquisition times for the 13 patients (A-M) could be estimated by making each of them an additional parameter. However, ABC in the resulting 15 dimensional parameter space would be difficult due to the curse of dimensionality. Instead, as shown in the Eq 8, we use these data such that we supplement the unknown times with values that produce the minimum discrepancy. This way, parameters that never produce enough variability to match the observations will increase the discrepancy, allowing us to gain evidence against such unreasonable values, even if the exact times are unknown and too computationally costly to infer.

Instead of simulating (n_{eff}, μ) samples from the prior we perform equivalent grid-based computations. That is, we consider an equidistant 50×50 grid of (n_{eff}, μ) values and simulate the model $1,000$ times at each grid point. However, in preliminary experiments we noticed that if n_{eff} and μ are simultaneously large, the amount of mutations produced by the model increases rapidly and it is clear that the simulated pairwise distances are always greater than in the observed data, and also the computation time and memory usage become prohibitive. Thus, we do not run the full set of 1000 simulations in this parameter region because it is clear that the posterior density would be negligible. Finally, the threshold ε is chosen such that $5,000$ out of the total of almost 1 million simulations are below the threshold, corresponding to the acceptance probability of 0.0057.

Details of the mixture model

We now discuss the mixture model in detail and then derive an efficient algorithm to estimate its parameters. Because the values of t_{0i} in Eq 5, denoting the times to the MRCAs in case the sequences are different strains, are unknown, we model them as random variables and give each of them a prior distribution

$$t_{0i} \,|\, k, \lambda \sim \text{Gamma}(k, \lambda), \quad i = 1, \ldots, N. \tag{9}$$

We further specify a weakly informative prior for λ such that

$$\lambda \sim \text{Gamma}(\alpha, \beta). \tag{10}$$

The parameter λ is thus shared between different t_{0i} which allows us to learn about its distribution.

If $k = 1$, then the Gamma distribution in Eq 9 reduces to the Exponential, which, however, does not reflect our prior understanding of reasonable value of t_{0i} because the

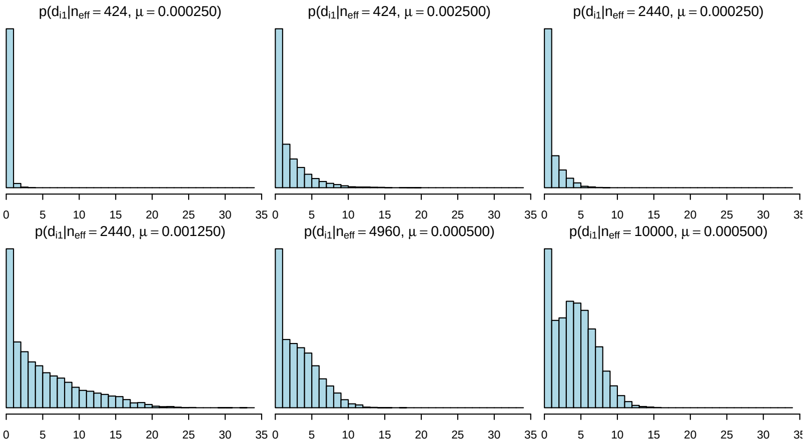

Fig 4. Distributions of pairwise distances for populations simulated with different parameters. The histograms show the estimated probability mass functions $\widehat{p}_{\mathrm{sim}}(d_{i1} \mid n_{\mathrm{eff}}, \mu)$ with selected parameter vectors (n_{eff}, μ). Increasing μ and/or n_{eff} tends to increase the distances. The distance distribution can also be bimodal as the subfigure in lower right corner shows. Each histogram represent variability in a simulated population at a single time point 6,000 generations after the beginning of the simulation.

mode of the resulting distribution is at zero, corresponding to a very recent common ancestor for genomes considered to be from different strains. Instead, we set $k = 5$, $\alpha = 2.5$, and $\beta = 1600$, which approximately correspond to the mean and standard deviation of 5800 and 8400 generations, respectively. This weakly informative prior reflects the notion that different strains diverged on average approximately a year ago, but with a large variance. Furthermore, if the time between samples, t_i, is three months, the prior translates to an expectation that, if the sampled genomes are from different strains, they are on average 30 mutations apart, with a large standard deviation of 50 mutations. Moreover, the density has a heavy tail to account for some possibly much greater distances. The formulas used to compute these values and other useful facts about the prior are provided in the supplementary material.

An equivalent way of writing the mixture model in Eq 6, which also simplifies the computations, is to introduce hidden labels which specify the component which generated each observation d_i, see [28]. We thus define latent variables

$$\mathbf{z}_i = (z_{i1}, z_{i2})^T = \begin{cases} (1,0)^T, & \text{if } d_i \text{ has distribution } p_S \\ (0,1)^T, & \text{if } d_i \text{ has distribution } p_D. \end{cases} \tag{11}$$

The prior density for the latent variables $\mathbf{z}$ is

$$p(\mathbf{z} \mid \boldsymbol{\omega}) = \prod_{i=1}^{N} p(\mathbf{z}_i \mid \boldsymbol{\omega}) = \prod_{i=1}^{N} \omega_S^{z_{i1}} \omega_D^{z_{i2}}, \tag{12}$$

where we have used vector notation $\mathbf{t} = (t_1, \ldots, t_N)^T$, $\mathbf{d} = (d_1, \ldots, d_N)^T$, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^T$, $\mathbf{t}_0 = (t_{01}, \ldots, t_{0N})^T$ and $\boldsymbol{\omega} = (\omega_S, \omega_D)^T$. We augment the parameter $\boldsymbol{\theta}$ to represent jointly all model parameters in Eq 6 and the prior densities specified in Eq 9 and 10, i.e., $\boldsymbol{\theta} = (n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{z}, \mathbf{t}_0, \lambda)^T$. To complete the model specification, we must specify the prior for $\boldsymbol{\omega}$, n_{eff} and μ. We use

$$\boldsymbol{\omega} \sim \mathrm{Dir}(\boldsymbol{\gamma}), \tag{13}$$

that is, a Dirichlet distribution with parameter $\boldsymbol{\gamma} = (1,1)^T$. We use the posterior $p(n_{\mathrm{eff}}, \mu \,|\, D_0)$, obtained by ABC using the external data D_0 as discussed in the previous section, as the (joint) prior for (n_{eff}, μ).

Bayesian inference for the mixture model

We now show how the mixture model can be fit efficiently to data. The joint probability distribution for the data $\mathbf{d}$ and the parameters $\boldsymbol{\theta}$ can be now written as

$$\begin{aligned} p(\mathbf{d}, \boldsymbol{\theta} \,|\, \mathbf{t}, D_0) &= p(\mathbf{d}, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{z}, \mathbf{t}_0, \lambda \,|\, \mathbf{t}, D_0) \\ &= p(\mathbf{d}, \mathbf{z} \,|\, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{t}_0, \lambda, \mathbf{t})\, p(n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{t}_0, \lambda \,|\, D_0) \qquad (14) \\ &= \prod_{i=1}^{N} p(d_i \,|\, \mathbf{z}_i, n_{\mathrm{eff}}, \mu, t_{0i}, \lambda, t_i)\, p(\mathbf{z}_i \,|\, \boldsymbol{\omega})\, p(n_{\mathrm{eff}}, \mu \,|\, D_0)\, p(\boldsymbol{\omega}) \prod_{i=1}^{N} p(\mathbf{t}_{0i} \,|\, \lambda)\, p(\lambda) \qquad (15) \end{aligned}$$

We use Gibbs sampling, which is an MCMC algorithm, to sample from the posterior density. The algorithm exploits the hierarchical structure of the model and it proceeds by iteratively sampling from the conditional density of each variable (or a block of variables) at a time [29]. In the following we derive the conditional densities for the Gibbs sampling algorithm. We observed that some of the parameters $\boldsymbol{\theta}$ are highly correlated which causes slow mixing of the resulting Markov chain and thus inefficient exploration of the parameter space. To make the algorithm more efficient, we reparametrise the model by defining new parameters $\boldsymbol{\theta}' = (n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, \lambda)$ via the transformation $\boldsymbol{\eta} = \mu \mathbf{t}_0$ and we use the Gibbs sampler for the transformed parameters $\boldsymbol{\theta}'$. This common strategy [29] resolves the problem arising from correlations between t_{0i} and μ, because the magnitudes of all η_i can now be changed simultaneously by a single μ update. The original variables t_{0i} can be obtained from the generated samples as $t_{0i} = \eta_i/\mu$.

The joint probability in Eq 15 for the transformed parameters then becomes

$$\begin{aligned} &p(\mathbf{d}, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, \lambda \,|\, \mathbf{t}, D_0) \\ &\quad = \prod_{i=1}^{N} p(d_i \,|\, \mathbf{z}_i, n_{\mathrm{eff}}, \mu, \mu^{-1}\boldsymbol{\eta}, \lambda, t_i)\, p(\mathbf{z}_i \,|\, \boldsymbol{\omega})\, p(n_{\mathrm{eff}}, \mu \,|\, D_0)\, p(\boldsymbol{\omega}) \prod_{i=1}^{N} p(\mu^{-1}\boldsymbol{\eta} \,|\, \lambda)\, p(\lambda) \mu^{-N} \\ &\quad = \prod_{i=1}^{N} [\omega_S^{z_{i1}} p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu)^{z_{i1}} \omega_D^{z_{i2}} p_D(d_i \,|\, t_i, \eta_i/\mu, \mu)^{z_{i2}}\, \mathrm{Gamma}(\eta_i/\mu \,|\, k, \lambda)] \\ &\qquad \cdot \mathrm{Gamma}(\lambda \,|\, \alpha, \beta)\, \mathrm{Dir}(\boldsymbol{\omega} \,|\, \boldsymbol{\gamma}) p(n_{\mathrm{eff}}, \mu \,|\, D_0) \mu^{-N}, \qquad (16) \end{aligned}$$

where μ^{-N} is the determinant of the Jacobian of the inverse transformation. Computing the conditional density of parameter $\boldsymbol{\omega}$ is straightforward. We neglect those terms in Eq 16 that do not depend on $\boldsymbol{\omega}$ and recognise the resulting formula as an unnormalised Dirichlet distribution. We then obtain

$$p(\boldsymbol{\omega} \,|\, \mathbf{z}, D) = \mathrm{Dir}(\boldsymbol{\omega} \,|\, \mathbf{n} + \boldsymbol{\gamma}), \qquad (17)$$

with $\mathbf{n} = (n_1, n_2)^T$, where $n_1 = \sum_{i=1}^{N} z_{i1}$ and $n_2 = \sum_{i=1}^{N} z_{i2}$. Next we consider the latent variables $\mathbf{z}_i$. We see that the conditional distribution of $\mathbf{z}_i$ for any $i = 1, \ldots, N$ does not depend on other latent variables $\mathbf{z}_j, j \neq i$. Specifically, we obtain

$$\mathbb{P}(z_{i1} = 1 \,|\, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \boldsymbol{\eta}, D) \propto \omega_S p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu), \qquad (18)$$

$$\mathbb{P}(z_{i2} = 1 \,|\, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \boldsymbol{\eta}, D) \propto \omega_D \frac{(2\eta_i + \mu t_i)^{d_i} e^{-(2\eta_i + \mu t_i)}}{d_i!}. \qquad (19)$$

We expect the effective sample size n_{eff} and the mutation parameter μ to be correlated a posteriori so we include them to the same block and update them together. We also include λ to this block as it also tends to be correlated with n_{eff} and μ. It is convenient to replace the sampling step from $p(n_{\mathrm{eff}}, \mu, \lambda \,|\, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, D, D_0)$ with the following two consecutive sampling steps: first sample from $p(n_{\mathrm{eff}}, \mu \,|\, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, D, D_0) = \int p(n_{\mathrm{eff}}, \mu, \lambda \,|\, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, D, D_0)\, \mathrm{d}\lambda$ and then sample from $p(\lambda \,|\, n_{\mathrm{eff}}, \mu, \boldsymbol{\omega}, \mathbf{z}, \boldsymbol{\eta}, D, D_0)$. From Eq 16 we observe that

$$\begin{aligned} p(n_{\mathrm{eff}}, \mu, \lambda \,|\, \mathbf{z}, \boldsymbol{\eta}, D, D_0) \propto \prod_{i=1}^{N} \left[p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu)^{z_{i1}} (2\eta_i + \mu t_i)^{z_{i2} d_i} \right] \frac{e^{-\mu \sum_{i=1}^{N} z_{i2} t_i}}{\mu^{Nk}} \\ \cdot\, p(n_{\mathrm{eff}}, \mu \,|\, D_0) \lambda^{Nk+\alpha-1} e^{-\lambda(\mu^{-1} \sum_{i=1}^{N} \eta_i + \beta)}. \end{aligned} \tag{20}$$

The above formula is recognised to be proportional to a Gamma density as a function of λ. We can thus marginalise λ easily to obtain the following density for the first step

$$\begin{aligned} & p(n_{\mathrm{eff}}, \mu \,|\, \mathbf{z}, \boldsymbol{\eta}, D, D_0) \\ & \propto \prod_{i=1}^{N} \left[p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu)^{z_{i1}} (2\eta_i + \mu t_i)^{z_{i2} d_i} \right] \frac{e^{-\mu \sum_{i=1}^{N} z_{i2} t_i} p(n_{\mathrm{eff}}, \mu \,|\, D_0)}{\mu^{Nk} (\mu^{-1} \sum_{i=1}^{N} \eta_i + \beta)^{Nk+\alpha}}. \end{aligned} \tag{21}$$

In the second step, we sample λ from the probability density

$$p(\lambda \,|\, \mu, \boldsymbol{\eta}, D) = \mathrm{Gamma}\left(\lambda \,\middle|\, Nk + \alpha, \beta + \frac{1}{\mu} \sum_{i=1}^{N} \eta_i \right). \tag{22}$$

This formula follows directly from Eq 20.

Sampling from Eq 21 and sampling $\mathbf{z}$ using Eq 18 are challenging because p_S is defined implicitly via the W-F simulation model. Consequently, we will consider an approximation that allows to compute $p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu)$ for any proposed point (n_{eff}, μ) and all values of d_i and t_i in the data. Since $d_i = d_{i1} + d_{i2}$, we can use the convolution formula for a sum of discrete random variables to see that

$$p_S(d_i \,|\, t_i, n_{\mathrm{eff}}, \mu) = \sum_{j=\max\{0, d_i - d_m\}}^{d_i} \mathrm{Poisson}(j \,|\, \mu t_i) p_{\mathrm{sim}}(d_i - j \,|\, n_{\mathrm{eff}}, \mu), \tag{23}$$

where p_{sim} specifies the distribution for a distance between two genomes as in Eq 3 and d_m is the maximum distance that can be obtained from p_{sim}.

Since $p_{\mathrm{sim}}(d_{i1} \,|\, n_{\mathrm{eff}}, \mu)$ is not available analytically, we estimate this probability mass function by simulation. A special case is if we know that there is no variation in the population at the time of taking the first sample s_{i1}, which can happen if we know that the acquisition happened just before the first sample. In this case, $d_{i1} = 0$, and we do not need the simulation. Since this is usually not the case, we use a general solution as follows: for each (n_{eff}, μ) value, we sample independently $d_{i1}^{(j)} \sim p_{\mathrm{sim}}(\cdot \,|\, n_{\mathrm{eff}}, \mu)$ by simulating the W-F model, sample a pair of genomes at a fixed time t from the simulated population, and compute their distance $d_{i1}^{(j)}$. This is repeated for $j = 1, \ldots, s$. Since d_{i1} is discrete, we approximate

$$p_{\mathrm{sim}}(d_{i1} \,|\, n_{\mathrm{eff}}, \mu) \approx \widehat{p}_{\mathrm{sim}}(d_{i1} \,|\, n_{\mathrm{eff}}, \mu) = \frac{1}{s} \sum_{j=1}^{s} \mathbb{1}_{d_{i1}^{(j)} = d_{i1}}, \tag{24}$$

for all i. Since in data D we do not know the acquisition times, we set $t = 6000$ generations and use this same value for all i. This large value represents a steady state of the simulation, where the variation in the population occasionally increases and decreases as new lineages emerge and old ones die out, which can be seen as corresponding to a reasonable default expectation about population variability when the true acquisition time is unknown. While this assumption was introduced for computational necessity, it can be justified by considering its impact on the inferences: the simplification may cause slightly overestimated distances d_{i1} if many acquisitions in reality happened very recently. The consequence is that the criterion for reporting new acquisitions becomes more *conservative*, because now the 'same strain' model will place some probability mass on occasional greater distances, and hence better accommodate also distant genomes which might otherwise have been considered as different strains.

Some of the resulting probability mass functions $\widehat{p}_{\mathrm{sim}}(d_{i1} \mid n_{\mathrm{eff}}, \mu)$ were already shown in Fig 4. In practice, the computations above are done using logarithms and the fact $\log \sum_i e^{a_i} = \max_i\{a_i\} + \log \sum_i e^{a_i - \max_i\{a_i\}}$, to avoid numerical underflow, which can occur whenever $a_i \ll 0$. The finite sample size s causes some numerical error, but, because the distances are usually small enough that the number of values we need to consider is limited, s can be made large enough without too extensive computation, making this error small in general. The above procedure allows computation of the conditional density in Eq 21 for any (n_{eff}, μ), and we can use a Metropolis update for (n_{eff}, μ). We marginalised λ in Eq 21 to improve the mixing of the chain and to be able to use the analytical formula in Eq 22, and in the supplementary material we justify that this algorithm is valid under the assumption that a new λ parameter is sampled only if the corresponding proposed value (n_{eff}, μ) has been accepted.

Whenever a new (n_{eff}, μ)-parameter is proposed, we need to compute p_{sim} at this point to check the acceptance condition. This value is also needed when sampling $\mathbf{z}$. However, computing p_{sim} on each MCMC iteration as described earlier makes the algorithm slow. Consequently, we instead precompute the values of p_{sim} in a dense grid of (n_{eff}, μ)-points which can be done in a parallel manner on a computer cluster. Given the grid values, we use bilinear interpolation to approximate p_{sim} at each proposed point $(n^*_{\mathrm{eff}}, \mu^*)$. We proceed similarly also with the prior density $p(n_{\mathrm{eff}}, \mu \mid D_0)$. This approach also allows one to fit the mixture model using different modelling assumptions or different data sets without need to repeat the costly W-F simulations.

Finally, we see that the probability density of η_i conditioned on the other variables does not depend on $\eta_j, j \neq i$. Specifically, we obtain

$$p(\eta_i \mid \mu, \mathbf{z}_i, \lambda, D) = \begin{cases} \mathrm{Gamma}(\eta_i \mid k, \lambda/\mu), & \text{if } z_{i2} = 0 \\ \sum_{j=0}^{d_i} w_j \, \mathrm{Gamma}(\eta_i \mid k+j, 2+\lambda/\mu), & \text{if } z_{i2} = 1 \end{cases} \tag{25}$$

for $i = 1, \ldots, N$. Derivation of this result, the formula for the mixture weights w_j and a special algorithm (Algorithm 2) to generate random values from this density are shown in the supplementary material.

The resulting Gibbs sampler is presented as Algorithm 1. It could be alternatively called a Metropolis-within-Gibbs sampler since some of the parameters (n_{eff} and μ) are sampled using a Metropolis-Hastings step using a proposal density that is denoted as q. Because n_{eff} is a discrete random variable, (n_{eff}, μ) is a mixed random vector and we cannot use the standard Gaussian proposal. Instead, we consider the distribution

$$q((n^*_{\mathrm{eff}}, \mu^*) \mid (n_{\mathrm{eff}}, \mu)) \propto \sum_{n \in \mathbb{Z}} \exp\left(-\frac{(\mu^* - \mu)^2}{2\sigma^2_{q,\mu}} - \frac{(n^*_{\mathrm{eff}} - n_{\mathrm{eff}})^2}{2\sigma^2_{q,n_{\mathrm{eff}}}} \right) \delta(n - n^*_{\mathrm{eff}}), \tag{26}$$

where $\sigma^2_{q,\mu}$ and $\sigma^2_{q,n_{\mathrm{eff}}}$ are chosen to produce acceptance probability of the Metropolis step close to 0.25 and $\delta(\cdot)$ is the Dirac delta function. The first element of a random sample from q in Eq 26 is an integer, and this proposal is also symmetric. We truncate the tails of q with respect to n_{eff} to be able to sample the discrete element from q

efficiently. In practice we then use a proposal q that is a mixture density where the components are as in Eq 26 but with different variance parameters $\sigma^2_{q,\mu}$ and $\sigma^2_{q,n_{\text{eff}}}$ to occasionally propose large steps to increase the exploration of the parameter space.

Algorithm 1 MH-within-Gibbs sampling algorithm for the mixture model

```
select an initial parameter $\boldsymbol{\theta}'^{(0)}$ (e.g. by sampling from the prior $p(\boldsymbol{\theta}')$), proposal $q$ and the number of samples $s$
for $i = 1, \ldots, s$ do
    sample $(n^*_{\text{eff}}, \mu^*) \sim q(\cdot \mid (n^{(i-1)}_{\text{eff}}, \mu^{(i-1)}))$ and $u \sim \mathcal{U}([0,1])$
    compute $\rho = \min\left\{1, \frac{p(n^*_{\text{eff}},\mu^* \mid \mathbf{z}^{(i-1)},\boldsymbol{\eta}^{(i-1)},D,D_0)q((n^{(i-1)}_{\text{eff}},\mu^{(i-1)}) \mid (n^*_{\text{eff}},\mu^*))}{p(n^{(i-1)}_{\text{eff}},\mu^{(i-1)} \mid \mathbf{z}^{(i-1)},\boldsymbol{\eta}^{(i-1)},D,D_0)q((n^*_{\text{eff}},\mu^*) \mid (n^{(i-1)}_{\text{eff}},\mu^{(i-1)}))}\right\}$ using Eq 21
    if $\rho < u$ then
        set $(n^{(i)}_{\text{eff}}, \mu^{(i)}) \leftarrow (n^*_{\text{eff}}, \mu^*)$
        sample $\lambda^{(i)} \sim p(\cdot \mid \mu^{(i)}, \boldsymbol{\eta}^{(i-1)}, D)$ using Eq 22
    else
        set $(n^{(i)}_{\text{eff}}, \mu^{(i)}, \lambda^{(i)}) \leftarrow (n^{(i-1)}_{\text{eff}}, \mu^{(i-1)}, \lambda^{(i-1)})$
    end if
    for $j = 1, \ldots, N$ do
        sample $\eta^{(i)}_j$ using the Algorithm 2 with $\mu = \mu^{(i)}, \mathbf{z} = \mathbf{z}^{(i-1)}, \lambda = \lambda^{(i)}$
    end for
    for $j = 1, \ldots, N$ do
        sample $\mathbf{z}^{(i)}_j \sim p(\cdot \mid n^{(i)}_{\text{eff}}, \mu^{(i)}, \boldsymbol{\omega}^{(i-1)}, \boldsymbol{\eta}^{(i)}, D)$ using Eq 18
    end for
    sample $\boldsymbol{\omega}^{(i)} \sim p(\cdot \mid \mathbf{z}^{(i)}, D)$ using Eq 17
end for
return samples $\{(n^{(i)}_{\text{eff}}, \mu^{(i)}, \boldsymbol{\omega}^{(i)}, \mathbf{z}^{(i)}, \boldsymbol{\eta}^{(i)}, \lambda^{(i)})\}^s_{i=1}$
```

Posterior predictive distribution

Given a new (future) data point (d^*, t^*) from a new patient, we would like to compute the probability of whether this case is of the same strain. This can be computed from the posterior of the model fitted to data D, D_0 as follows. We denote the original parameter vector with $\boldsymbol{\theta}$ as before and additional parameters related to the new data point $D^* = \{(d^*, t^*)\}$ as $\mathbf{z}^* \in \{(1,0),(0,1)\}$ and $t^*_0 > 0$. The updated posterior after considering the new data point D^* is then

$$\begin{aligned} p(\mathbf{z}^*, t^*_0, \boldsymbol{\theta} \mid D^*, D, D_0) &\propto p(\mathbf{z}^*, t^*_0, \boldsymbol{\theta}) p(D^*, D, D_0 \mid \boldsymbol{\theta}, \mathbf{z}^*, t^*_0) && (27) \\ &= p(\boldsymbol{\theta}) p(\mathbf{z}^*, t^*_0, \mid \boldsymbol{\theta}) p(D, D_0 \mid \boldsymbol{\theta}) p(d^* \mid t^*, \mathbf{z}^*, t^*_0, \boldsymbol{\theta}) && (28) \\ &\propto p(d^* \mid t^*, \mathbf{z}^*, t^*_0, \boldsymbol{\theta}) p(\mathbf{z}^*, t^*_0 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D, D_0), && (29) \end{aligned}$$

where $p(\boldsymbol{\theta} \mid D, D_0)$ is the posterior based on our original data D, D_0. We marginalise the set of parameters least contributory to the aim to obtain

$$\begin{aligned} p(\mathbf{z}^* \mid D^*, D, D_0) &\propto \int_{\boldsymbol{\theta}} \int_{t^*_0} p(d^* \mid t^*, \mathbf{z}^*, t^*_0, \boldsymbol{\theta}) p(t^*_0 \mid \lambda) p(\mathbf{z}^* \mid \boldsymbol{\omega}) p(\boldsymbol{\theta} \mid D, D_0) \, \mathrm{d}t^*_0 \, \mathrm{d}\boldsymbol{\theta} && (30) \\ &\approx \frac{1}{s} \sum_{i=1}^{s} \left(\omega^{(i)}_S p_S(d^* \mid n^{(i)}_{\text{eff}}, \mu^{(i)}, t^*)\right)^{z^*_1} \left(\omega^{(i)}_D p_D(d^* \mid \mu^{(i)}, {t^*_0}^{(i)}, t^*)\right)^{z^*_2}, && (31) \end{aligned}$$

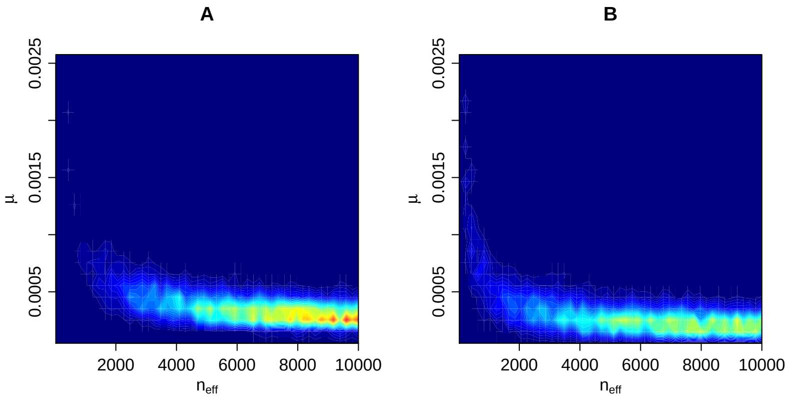

Fig 5. ABC posterior distribution for $(n_{\mathbf{eff}}, \mu)$. The ABC posterior distribution i.e. the updated prior for parameters (n_{eff}, μ), the effective population size and mutation rate, given data D_0. Panel A shows the result with the full data and panel B the corresponding result with only a subset of the data (see text for details).

where $(t_0^{*(i)}, \boldsymbol{\theta}^{(i)}) \sim p(t_0^* \mid \lambda) p(\boldsymbol{\theta} \mid D, D_0)$ for $i = 1, \ldots, s$. The probability of the new measurement point (d^*, t^*) being of the same strain, based on the previously observed data D, D_0 is obtained from Eq. 31.

Results

In this section we fit the W-F model to the external data D_0 as discussed in Section ABC inference to update the prior using external data. We then verify that the proposed Gibbs sampling algorithm for fitting the mixture model from Section Bayesian inference for the mixture model is consistent based on experiments with simulated data. Subsequently, we fit the mixture model to the MRSA data and discuss the results. Finally, we assess the quality of the model fit.

Updating the prior using ABC inference

The ABC posterior based on the external data D_0 and the discrepancy in Eq 8, is shown in Fig 5A. We also repeated the computations so that we omitted a subset, patients A-M, from the analysis i.e. the second summation term in Eq 8 was set to zero. This was done to assess the effect of patients A-M, which have measurements from one time point only, and an unknown time since acquisition. This extra analysis resulted in an ABC posterior approximation shown in Fig 5B. We see that in both cases large parts of the parameter space have been ruled out as having negligible posterior probability. As expected, the posterior distribution based on the subset (Fig 5B) is slightly more dispersed than with the full data D_0 (Fig 5A). Using the full data causes the estimated mutation rate to be slightly greater than with the subset, likely because the model needs to accommodate the higher variability in the patients A-M. In addition, small effective sample sizes ($n_{\text{eff}} < 2000$) are less probable based on the full data D_0.

Overall, we see that the effective sample size n_{eff} cannot be well identified based on the external data D_0 alone. We also see that if the upper bound of the prior density of n_{eff} was increased from $10,000$, higher values would likely have non-negligible posterior probability also; however, this constraint will have a negligible impact on the resulting posterior from the mixture model as is seen later. The mutation rate μ, on the other hand, is smaller than 0.001 mutations per genome per generation with high probability and cannot be arbitrarily small.

Validation of the mixture model using simulated data

To empirically investigate the identifiability of the mixture model parameters and the correctness and consistency of our MCMC algorithm under the assumption that the model is specified correctly, we first fit the mixture model to simulated data. We generate artificial data from the mixture model with parameter values similar to the estimates for the observed data D from the next section. Specifically, we choose $n_{\text{eff}} = 2,137, \mu = 0.0011, \omega_S = 0.8, \lambda = 0.0001$ and we repeat the analysis with various data sizes N. We use otherwise similar priors as for the real data in the next section except that, for simplicity, instead of using the prior obtained from the ABC inference, we use a uniform prior in Eq 7. We then fit the mixture model to the simulated data sets to investigate if the true parameters can be recovered (identifiability) and whether the posterior becomes concentrated around their true values when the amount of data increases (consistency).

Results are illustrated in Fig 6. We see that the (marginal) posterior of (n_{eff}, μ) is concentrated around the true parameter value that was used to generate the data (green diamond in the figure). Also, despite the fact that the number of parameters increases as a function of data size N (because each data point (d_i, t_i) has its own class indicator $\mathbf{z}_i$ and time to the most recent common ancestor t_{0i} parameter), the marginal posterior distribution of (n_{eff}, μ) can be identified and appears to converge to the true value as N increases. On the other hand, we cannot learn each t_{0i} accurately since essentially only the data point to which the parameter corresponds provides information about its value. However, precise estimates of these nuisance parameters are not needed for using the model or obtaining useful estimates of the other unknown parameters as demonstrated in Fig 6.

The panel in the lower right corner of Fig 6 shows results from an additional simulation experiment where the mixture model is fitted to data generated with different values for the ω_S parameter, which represents the proportion of pairs that are from the same strain. Other than that and the fact that we fixed $N = 150$, the experimental design is the same as above. The results show that the estimated ω_S values generally agree well with the true values. Interestingly, ω_S is slightly overestimated when its true value is close to 0, and slightly underestimated when the true value is close to 1, which may reflect the regularizing effect of the prior, drawing the estimates away from the extreme values. Furthermore, when the true value of ω_S is around 0.5, the variance of the estimate tends to be higher than with ω_S values close to 0 or 1. This observation may be explained by the fact that there are more data points that overlap both mixture model components when ω_S is around 0.5 which makes the inference task more challenging and causes higher posterior variance.

Analysis of the Project CLEAR MRSA data

The following settings are used to analyse longitudinally-sampled *S. aureus* nares isolates from the control arm of Project CLEAR [23]. We generate 4 MCMC chains, each of length 25, 000, initialized randomly from the prior density, whose first halves are discarded as “burn-in”. We use the Gelman and Rubin’s convergence diagnostic in R-package `coda` and visual checks to assess the convergence of the MCMC algorithm. We use 100×100 equidistant grid for numerical computation with the (n_{eff}, μ) values and $s = 10,000$ in Eq 24. The ABC posterior obtained in Section Updating the prior using ABC inference and visualised in Fig 5A is used as the prior for (n_{eff}, μ).

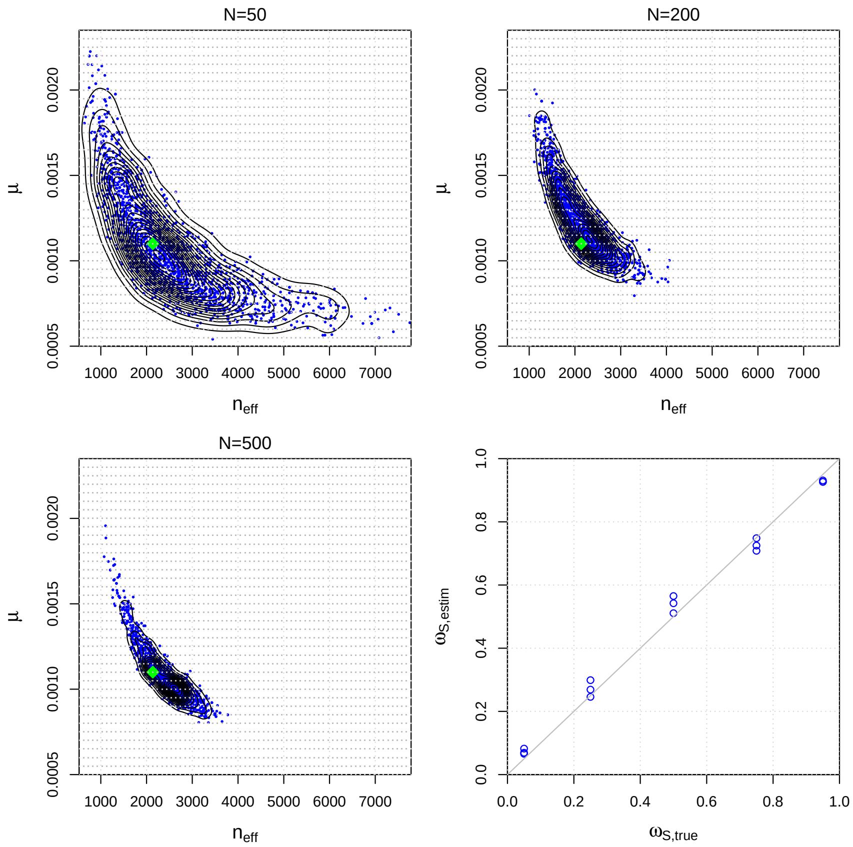

Fig 6. Accuracy and consistency with synthetic data. The first three panels show the estimated posterior distributions for parameters (n_{eff}, μ) of the mixture model using simulated data of different sizes N. The green diamond shows the true value used to generate the simulated data and the light grey dots denote the grid point locations needed for numerical computations. The bottom right panel shows the estimated vs. the true ω_S parameter in a set of additional simulation experiments.

The parameter vector $\boldsymbol{\theta}$ consists of the 'global' parameters $n_{\text{eff}}, \mu, \boldsymbol{\omega}, \lambda$, as well as a large number of nuisance parameters ($\mathbf{z}$ and $\mathbf{t}_0$) related to each data point. The estimated global parameters are presented in Table 1. We also repeated the analysis using a uniform prior on (n_{eff}, μ). While the uniform prior is non-informative about the parameters (n_{eff}, μ), the results are nevertheless surprisingly similar (Table 1). In other words, the additional data D_0 used to update the prior has only a small effect on the estimated parameters of the mixture model. This was unexpected because the data set D used to train the mixture model has only one genome per sampled time point, and yet, impressively, the model is able to learn about the parameters (n_{eff}, μ) which effectively define the variability in the whole population. This further demonstrates the robustness of the mixture model to the prior used. We observe, however, that incorporating the prior from the ABC slightly shifts the probability distribution for n_{eff} towards larger values, although there is no clear conflict between the two results. For example, as seen in Table 1, the 95% credible interval (CI) for n_{eff}, $[1200, 2200]$, gets updated to $[1300, 2200]$ when the extra prior information is included.

Fig 7 shows the posterior predictive distribution for the probability of the same strain case for a (hypothetical future) observation with distance d^* and time difference

Table 1. Posterior mean and 95% credible interval (CI) for the 'global' parameters of the mixture model.

	Informative prior (ABC, data D_0)		Uniform prior	
parameter	mean	95% CI	mean	95% CI
n_{eff}	1700	$[1300, 2200]$	1700	$[1200, 2200]$
μ	0.00076	$[0.00060, 0.00092]$	0.00080	$[0.00064, 0.00095]$
ω_S	0.87	$[0.83, 0.91]$	0.88	$[0.83, 0.92]$
ω_D	0.13	$[0.09, 0.17]$	0.12	$[0.08, 0.17]$
λ $(\times 10^5)$	7.3	$[5.8, 9.0]$	7.5	$[5.9, 9.3]$

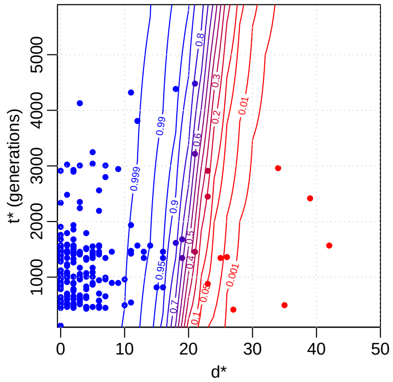

Fig 7. Results for the Project CLEAR MRSA data. Contour plot for same strain probability of a distance d^* and time interval t^* based on the fitted model. The coloured points denote the observations that were used to fit the model. Blue colour indicates large same strain probability. Distances greater than 50 are not shown and are classified as different strains with probability one. $6,000$ generations on the y-axis correspond to approximately one year.

t^*. Blue colour in the figure denotes high probability of the same strain. The corresponding 50% classification curve is (almost) a straight line with a steep positive slope. This is as expected since the same strain model can explain a greater number of mutations when more time has passed. Approximately 20 mutations draws the line between the same strain and different strains cases within the time difference up to 6000 generations. The uncertainty in the classification occurs because there is overlap in the two explanations (around $d^* \approx 20$) and because of the posterior uncertainty in the model parameters $\boldsymbol{\theta}$.

We also analysed explicitly all observed patterns where: 1) two genomes of the same ST from the same patient are interleaved with a missing observation, i.e. the colonization appears to disappear and then re-emerge, and 2) two genomes of the same ST from the same patient are interleaved with an observation of a different ST. The numbers for the two genomes being from the same or different strain in these patterns are shown in Table 2. The credible intervals for the 'same strain' proportion combine uncertainty from the limited number of samples with the posterior uncertainty of whether a sample is from the same strain or not (see the Supplementary material for further details). From Table 2 we see that approximately 58% of genome pairs in pattern 1) are from the same strain. This is only a little smaller than the same strain proportion when there are no missing observations in between (84%). Therefore, a plausible explanation for most of the missing in-between observations is that in reality the same strain has been colonizing the patient throughout, and the missing observation reflects the limited sensitivity of the sampling, rather than a clearance followed by a novel acquisition. Similarly, even if interleaved with a different ST (pattern 2), the surrounding genomes often, in 63% of cases, appear to be from the same strain. This suggests that in these cases the patient has been colonized by the surrounding strain throughout, and co-colonized by two different STs at the time of observing the divergent ST in the middle.

Table 2. The estimated numbers (mean, 95% CI in parenthesis) of cases with genomes in the beginning and in the end of the pattern being from the same or different strain, for three different patterns in the Project CLEAR MRSA data, and the estimated proportion of the same strain cases.

	same strain/n	diff. strain/n	same strain prop.
ST A → ST A	$190(187, 192)/224$	$34(32, 37)/224$	$0.84(0.78, 0.89)$
ST A → $\emptyset$ → ... → ST A	$17(16, 19)/29$	$12(10, 13)/29$	$0.58(0.36, 0.76)$
ST A → ST B → ... → ST A	$12(10, 12)/18$	$6(6, 8)/18$	$0.63(0.34, 0.81)$

"ST A → ST A" denotes the case where the ST does not change between two genomes at consecutive samples, "ST A → $\emptyset$ → ... → ST A" is the pattern 1) where one or more negative samples are seen between the same ST and "ST A → ST B → ... → ST A" is the pattern 2) where a sample with different ST is observed between two samples of the same ST. n denotes the number of data points in each alternative.

Finally, we compute acquisition and clearance rates using our model, and compare those to the ones obtained with the common strategy of using a fixed distance threshold. For the purposes of this exposition, we define the acquisition r_{acq} and clearance rates r_{clear} informally as

$$r_{\text{acq}} = \frac{B + C + E}{G}, \quad r_{\text{clear}} = \frac{D}{A + B + C + D}, \tag{32}$$

where the quantities A, B, C, D and E denote the numbers of possible events in consecutive samples (e.g. acquisition, replacement, clearance, or no change) defined in detail in Table 3. Also, G is the total number of possible events over the whole data. The quantities A, B, D and E are random variables that depend on the same/different strain posterior probabilities and, consequently, we also compute the uncertainty estimates for these quantities in Eq 32. Number C is a constant because an observed change of ST always indicates an actual change of ST as well. For cases with one or more negative samples (denoted by $\emptyset$) between two positive samples, we do not know when the clearance and acquisition events took place and whether the negative samples are "false negatives". To handle these cases, we parsimoniously assume that a missing observation between two positive samples that are inferred to come from the same strain is a false negative (i.e. that the same strain was present also in the middle, even if it was not detected), and record these events in the groups A-E accordingly. Details on how we unambiguously determine the group for all special cases is provided in the Supplementary material.

The estimated acquisition and clearance rates with 95% credible intervals are shown on the last two lines of Table 3. For comparison, we also computed these rates otherwise similarly but using a fixed distance threshold of 40 mutations, a value used in [10], to determine if two genomes are from the same strain or not. We see that the threshold-based estimates are relatively similar to, and only slightly smaller than the estimates from our model. The explanation for the similarity of summaries such as the acquisition and deletion rates is that, when estimating these quantities across the whole data set, the uncertainty gets averaged out, even if individual data points exhibit a lot of uncertainty regarding whether they are the same strain or not (see Fig 7). Importantly, while being consistent with the previous results, our model bypasses the task of heuristically choosing a single threshold and adds uncertainty estimates around the point estimates, crucial for drawing rigorous conclusions.

Table 3. Estimated numbers (posterior means) of different patterns A-E of consecutive samples and the estimated acquisition and clearance rates (mean, 95% CI in parenthesis).

event	expected number	
A: ST A, str X → ST A, str X	231	
B: ST A, str X → ST A, str Y	34	
C: ST A → ST B	45	
D: ST A, str X → $\emptyset$	104	
E: $\emptyset$ → ST A, str X	21	
rate parameter	post. estimate	threshold-based estimate
acquisition rate r_{acq}	$0.18(0.17, 0.19)$	0.16
clearance rate r_{clear}	$0.25(0.24, 0.25)$	0.24

Above, ST denotes sequence type as before, str denotes the strain and symbol $\emptyset$ denotes a negative sample i.e. no bacteria detected.

Assessing the goodness-of-fit of the model for the Project CLEAR MRSA data

As the last part of our analysis, we use posterior predictive checks to assess the quality of the model, see e.g. [15] for further details. Briefly, this consists of simulating replicated data sets $D^{\text{rep},(j)}$ from the fitted mixture model and comparing these to the observed data D for any systematic deviations. Any discrepancies between the observed and simulated data can be used to criticise the model and understand how the model could be improved. In practice, simulating replicate data is done by simulating a parameter vector $\boldsymbol{\theta}^{(j)}$ from the posterior (by using the existing MCMC chain) and simulating a new set of distance-time difference pairs $(\tilde{d}_i^{(j)}, \tilde{t}_i^{(j)}), i = 1, \ldots, N$ in $D^{\text{rep},(j)}$ from the model using $\boldsymbol{\theta}^{(j)}$. To obtain M replicates this procedure is repeated for $j = 1, \ldots, M$.

Example replicate data sets are shown in Fig 8. Overall, the simulated distances are similar to the corresponding observations. There is a clear peak at $d_i = 0$, and as the distance is increased the frequency starts to decrease. Occasional large distances ($d_i > 20$) occur only rarely, in keeping with the observed data. A minor discrepancy is that the fitted model tends to underestimate the frequency of distance zero while small positive distances tend to occur more frequently than observed. This could happen because we estimated the empirical densities $p_{\text{sim}}(d_{i1} \,|\, n_{\text{eff}}, \mu)$ using a constant time of $6,000$ (i.e. 1 year) since the acquisition (as discussed in Section Bayesian inference for the mixture model), which may lead to a slight overestimation of the distances. To explore the impact of this assumption further, we repeated the analysis so that we computed the densities $p_{\text{sim}}(d_{i1} \,|\, n_{\text{eff}}, \mu)$ at a constant time of $1,000$ generations. However, the mismatch did not disappear completely and the estimated mutation rate increased as a result to compensate for the occurrence of greater distances, in disagreement with the prior density from the ABC analysis and data D_0. We thus believe that the current model is adequate.

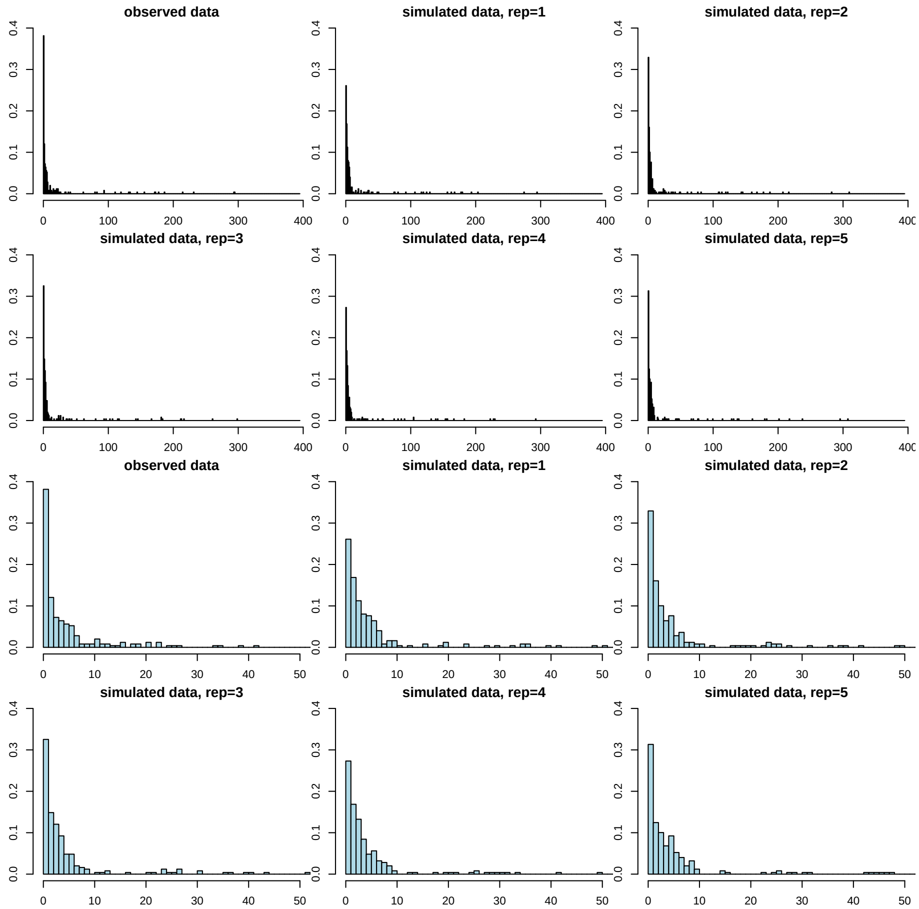

Fig 8. Model validation using posterior predictive checking. The histogram in the upper left corner shows the observed distance distribution in the Project CLEAR MRSA data, the other figures in the top two rows show the corresponding distances in replicate data sets simulated from the fitted model. The bottom two rows show the same histograms zoomed to range $[0, 50]$. The replicate data sets look overall similar to the observed data, demonstrating the adequacy of the model. However, the amount of zero distances is underestimated and the frequencies of small positive distances tend to be slightly overestimated.

Discussion

We presented a new model for the analysis of clearance and acquisition of bacterial colonization, which, unlike previous approaches, does not rely on a heuristic fixed distance threshold to determine whether genomes observed at different times points are from the same or different acquisition. Fully probabilistic, the model automatically provides uncertainty estimates for all relevant quantities. Furthermore, it takes into account the variation in the time intervals between pairs of consecutive samples. Another benefit is that the model can easily incorporate additional external data to inform about the values of the parameters. To fit the model, we developed an innovative combination of ABC and MCMC, based on an underlying mixture model where one of the component distributions was formulated empirically by simulation.

We demonstrated the model using data on *S. aureus* genomes sampled longitudinally from multiple patients. Our analysis provided evidence for occasional co-colonization and identified likely false negative samples. The output of the model consists of the same vs. different strain probability for any pair of genomes, and, by using this information to decide (probabilistically) when and where the colonizing strain had changed, the acquisition and clearance rates were easy to calculate. Estimates of these parameters were found to be in agreement with previous estimates derived using a fixed threshold, but now we were able to provide confidence intervals, essential for drawing rigorously supported conclusions. We believe such analyses are common enough that our method should be useful for many, and, consequently, we provide it as an easy-to-use R-code. The code includes tools for both the ABC-inference to incorporate external data of distance distributions between multiple samples at a given time point (or two time points), and the MCMC-algorithm. We note that our method does not assume recombination, which was not relevant with the present data. If this is an issue, we recommend removing recombinations by preprocessing the genomes with one of the standard methods [30–32]. While our analysis demonstrated that the external data may reduce uncertainty in the resulting posterior, we also saw that the method may work without such data. In the latter case the input is simply a list of distance-time difference pairs for genomes sampled from the same patient at consecutive time points, and it is sufficient to run the MCMC, which is efficient and fast in typical cases.

A central component of our approach is a model for within-host variation, required to determine how much variation can be expected if the genomes at different time points have evolved from the same strain obtained in a single acquisition. We selected for this purpose the basic Wright-Fisher model assuming constant population size and mutation rate with the understanding that these assumptions are expected to be violated to some extent in any realistic data set, but the benefits of simplicity include robustness of the conclusions to prior distributions and identifiability of the parameters from the available data. More complex models have been fitted to the distance distributions (our external data D_0), assuming the population size first increases and then decreases [13]. However, our model can fit the same data with fewer parameters, which justifies the simpler alternative. Furthermore, the constant population size may also be seen as a sensible model for persistent colonization. An interesting future research question is what additional data should be collected in order to be able to fit one of the possible extensions of the basic model. Another direction that we are currently pursuing is to extend the model to cover genomes sampled from multiple body sites.

Supporting information

S1 File. Derivations and further details of the model. We provide some further derivations and details related to our MCMC algorithm. To guide the selection of prior hyperparameters, we also derive the explicit prior distribution and some of its summaries for the parameter $\mathbf{t}_0$ and the mean and variance for the prior predictive distribution for the distance. We also describe further details on computing the acquisition and clearance rates.

Acknowledgments

We acknowledge the computational resources provided by Aalto Science-IT project.

References

1. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proceedings of the National Academy of Sciences. 2012;109(12):4550–4555.

2. Young BC, Wu CH, Gordon NC, Cole K, Price JR, Liu E, et al. Severe infections emerge from commensal bacteria by adaptive evolution. elife. 2017;6:e30637.

3. Gordon N, Pichon B, Golubchik T, Wilson D, Paul J, Blanc D, et al. Whole-genome sequencing reveals the contribution of long-term carriers in Staphylococcus aureus outbreak investigation. Journal of Clinical Microbiology. 2017;55(7):2188–2197.

4. Alam MT, Read TD, Petit RA, Boyle-Vavra S, Miller LG, Eells SJ, et al. Transmission and microevolution of USA300 MRSA in US households: evidence from whole-genome sequencing. MBio. 2015;6(2):e00054–15.

5. Coll F, Harrison EM, Toleman MS, Reuter S, Raven KE, Blane B, et al. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. Science Translational Medicine. 2017;9(413):eaak9745.

6. Calderwood MS. Editorial commentary: Duration of colonization with methicillin-resistant Staphylococcus aureus: A question with many answers. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2015;60(10):1497.

7. Miller RR, Walker AS, Godwin H, Fung R, Votintseva A, Bowden R, et al. Dynamics of acquisition and loss of carriage of Staphylococcus aureus strains in the community: the effect of clonal complex. Journal of Infection. 2014;68(5):426–439.

8. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nature Reviews Microbiology. 2016;14(3):150.

9. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Computational Biology. 2014;10(3):e1003549.

10. Price JR, Cole K, Bexley A, Kostiou V, Eyre DW, Golubchik T, et al. Transmission of Staphylococcus aureus between health-care workers, the environment, and patients in an intensive care unit: a longitudinal cohort study based on whole-genome sequencing. The Lancet Infectious Diseases. 2017;17(2):207–214.

11. Uhlemann AC, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MT, et al. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. Proceedings of the National Academy of Sciences. 2014;111(18):6738–6743.

12. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife. 2017;6.

13. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, et al. Within-host evolution of Staphylococcus aureus during asymptomatic carriage. PLoS One. 2013;8(5):e61319.

14. Robinson DA, Feil EJ, Falush D. Bacterial population genetics in infectious disease. John Wiley & Sons; 2010.

15. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Chapman & Hall/CRC Texts in Statistical Science; 2013.

16. O'Hagan A, Forster J. Advanced Theory of Statistics, Bayesian inference. 2nd ed. London, UK: Arnold; 2004.

17. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–2035.

18. Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J. Fundamentals and Recent Developments in Approximate Bayesian Computation. Systematic biology. 2017;66(1):e66–e82.

19. Ansari MA, Didelot X. Inference of the properties of the recombination process from whole bacterial genomes. Genetics. 2014;196(1):253–265.

20. Numminen E, Cheng L, Gyllenberg M, Corander J. Estimating the transmission dynamics of Streptococcus pneumoniae from strain prevalence data. Biometrics. 2013;69(3):748–757.

21. Järvenpää M, Gutmann M, Vehtari A, Marttinen P. Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria. Annals of Applied Statistics. 2018;.

22. De Maio N, Wilson DJ. The bacterial sequential Markov coalescent. Genetics. 2017;206(1):333–343.

23. Agency for Healthcare Research and Quality (AHRQ). Project CLEAR (Changing Lives by Eradicating Antibiotic Resistance) Trial; 2018. `https://clinicaltrials.gov/ct2/show/NCT01209234`. Last accessed 08-23-2018.

24. Schierup MH, Wiuf C. The coalescent of bacterial populations. Bacterial Population Genetics in Infectious Disease. 2010; p. 1–18.

25. Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. Statistical inference for stochastic simulation models–theory and application. Ecology Letters. 2011;14(8):816–27.

26. Marin JM, Pudlo P, Robert CP, Ryder RJ. Approximate Bayesian computational methods. Statistics and Computing. 2012;22(6):1167–1180.

27. Wertheim HF, Walsh E, Choudhurry R, Melles DC, Boelens HA, Miajlovic H, et al. Key role for clumping factor B in Staphylococcus aureus nasal colonization of humans. PLoS medicine. 2008;5(1):e17.

28. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc.; 2006.

29. Robert CP, Casella G. Monte Carlo Statistical Methods. 2nd ed. New York: Springer; 2004.

30. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research. 2014;43(3):e15–e15.

31. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS computational biology. 2015;11(2):e1004041.

32. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient inference of recent and ancestral recombination within bacterial populations. Molecular biology and evolution. 2017;34(5):1167–1182.