The Distribution of Pairwise Genetic Distances: A Tool for Investigating Disease Transmission

Colin J. Worby,[1] Hsiao-Han Chang, William P. Hanage,[2] and Marc Lipsitch[2]
Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115

ABSTRACT Whole-genome sequencing of pathogens has recently been used to investigate disease outbreaks and is likely to play a growing role in real-time epidemiological studies. Methods to analyze high-resolution genomic data in this context are still lacking, and inferring transmission dynamics from such data typically requires many assumptions. While recent studies have proposed methods to infer who infected whom based on genetic distance between isolates from different individuals, the link between epidemiological relationship and genetic distance is still not well understood. In this study, we investigated the distribution of pairwise genetic distances between samples taken from infected hosts during an outbreak. We proposed an analytically tractable approximation to this distribution, which provides a framework to evaluate the likelihood of particular transmission routes. Our method accounts for the transmission of a genetically diverse inoculum, a possibility overlooked in most analyses. We demonstrated that our approximation can provide a robust estimation of the posterior probability of transmission routes in an outbreak and may be used to rule out transmission events at a particular probability threshold. We applied our method to data collected during an outbreak of methicillin-resistant *Staphylococcus aureus*, ruling out several potential transmission links. Our study sheds light on the accumulation of mutations in a pathogen during an epidemic and provides tools to investigate transmission dynamics, avoiding the intensive computation necessary in many existing methods.

PATHOGEN genomic data are rapidly becoming abundant, and there is a demand for statistical methods to extract meaningful conclusions from the wealth of information these data provide. One of the most basic and frequently used—yet imperfectly understood—comparative tools is the genetic distance between two samples [commonly defined as the number of single-nucleotide polymorphisms (SNPs) between the isolates]. In the context of epidemiological investigations, genetic distance can be used as a discriminatory value to determine whether infected individuals belong to the same outbreak or cluster or to rule out potential transmission events.

Genetic distance is central to the inference of transmission routes—intuitively, the greater the similarity is between samples taken from two different hosts, the more likely they are to have been involved in a transmission event. While in some cases it may suffice to identify the carrier of the genetically closest pathogen isolate as the source of infection (Jombart *et al.* 2011), this approach lacks any measure of uncertainty and may result in a high false positive rate; it has been demonstrated that estimation of a transmission network using genetic distance data alone is associated with much uncertainty, making the estimation of individual transmission routes impossible (Worby *et al.* 2014). However, with a probabilistic interpretation of genetic distances, given the relationship between the hosts of pathogen samples, one can quantify the uncertainty surrounding each potential transmission source and establish general trends of transmission in the epidemic. Furthermore, probabilistically weighted transmission routes may also lead to improved estimates of heterogeneous transmission rates from different subpopulations.

Many studies to date have developed methods to infer routes of transmission based on genomic and epidemiological data (Cottam *et al.* 2008; Jombart *et al.* 2011; Morelli *et al.* 2012; Ypma *et al.* 2012, 2013; Didelot *et al.* 2014; Jombart *et al.* 2014). Each method utilizes a likelihood component that describes the probability that a set of mutations occurs between two pathogen samples from different hosts,

doi: 10.1534/genetics.114.171538
Manuscript received July 25, 2014; accepted for publication October 7, 2014; published Early Online October 13, 2014.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171538/-/DC1.
[1]Corresponding author: Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115. E-mail: cworby@hsph.harvard.edu
[2]These authors contributed equally to this work.

given their epidemiological relationship. These are often based on strong assumptions (*e.g.*, transmission bottleneck size of 1 or mutation occurring only at the time of transmission), and many are highly computationally intensive.

The distribution of pairwise genetic distances between samples taken from epidemiologically linked carriers depends on numerous factors, such as the mutation rate, the within-host pathogen population dynamics, and the transmission bottleneck size. It is of interest to understand how each of these factors affects observed genetic distance.

In this study, we aimed to investigate the distribution of pairwise genetic distances to better understand how diversity accumulates during a disease outbreak. In particular, we developed an approximation to this distribution and investigated its use as a tool to assess the likelihood of transmission routes. We used simulated data and real outbreak data, collected during a hospital outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA), to demonstrate the ability of our method to rule out several patient-to-patient transmission routes.

Methods

The distribution of genetic distance between two samples taken during an outbreak

Consider a disease outbreak, consisting of n cases, where case 1 is the origin, and cases $2,\ldots,n$ each have a source of infection from within the population. Let t_j^{I} be the infection time of case j, and $t_1^{\mathrm{I}} = 0$. Each case is observed, and we initially assume that one pathogen specimen is taken for sequencing at time t_j^{s} with genotype g_j. Table 1 describes notations used in this article.

We consider the unobserved transmission network, which consists of infection routes and times. Let c_j be the vector of transmission ancestry for person j, such that the first element is the transmission source of j, and each successive element is the source of the preceding element. Since the network is fully connected, the final element of this vector for any given host will be the outbreak origin, and the vector will have length equal to the number of hosts in the transmission chain from the origin to j. Let $s_{ij} = c_i \cap c_j$ be the vector of ancestry common to both i and j, such that the first element $s_{ij}^{(1)}$ is the most recent common transmission source of both i and j, and the last element is 1.

Now consider the genealogy of the sampled isolates. This tree is not necessarily identical to, but must be consistent with, the transmission tree (Ypma *et al.* 2013). The time of coalescence for samples g_i and g_j, denoted $m(g_i,g_j)$, must occur prior to the divergence of the transmission tree branches to which persons i and j belong and will belong within one of the hosts in s_{ij}. The ancestries of the samples coexist in the same host or chain of hosts for a period of time, before one lineage is transmitted to another person and exists independently of the other. Let $d(i,j)$ be the time of lineage divergence, the time at which the lineages cease to exist within the same host (see Figure 1).

Table 1 Notation used in this article

Notation	Definition
$i \to j$	Transmission route from person i to person j
t_j^{I}	Time of infection of person j
t_j^{S}	Time of genome sampling from person j
s_{ij}	Vector of transmission ancestry common to persons i and j
$d(i,j)$	Time of lineage divergence
μ	Mutation rate per genome per generation
$\psi(a,b)$	Genetic distance (no. SNPs) between genomes a and b
$m(a,b)$	Coalescence time of isolates a and b
m_t	Time between coalescence and observation time t
$N(t)$	Effective pathogen population size at time t
N_{B}	Effective transmission bottleneck size

Let $\psi(g_i,g_j)$ denote the genetic distance between samples g_i and g_j, measured by the number of SNPs. The mutations could have arisen in two distinct periods—first, during the time between observations t_i^{S}, t_j^{S} and lineage divergence $d(i,j)$, and second, during the (earlier) time between lineage divergence and coalescence $m(g_i,g_j)$. The number of SNPs $\psi(g_i,g_j)$ is then equal to the sum of two random variables, $\psi(g_i,g_j) = X + Y$, where X represents mutations occurring between lineage divergence and observation, and Y represents mutations occurring prior to lineage divergence. For the former, we can assume that the number of SNPs arising from the time of lineage divergence $d(i,j)$ until observation follows a Poisson distribution with mean $\mu(t_i^{\mathrm{s}} + t_j^{\mathrm{s}} - 2d(i,j))$. For the latter, with a known time of coalescence, $m(g_i,g_j)$, the number of SNPs accumulating between coalescence and divergence is again a Poisson-distributed random variable,

$$Y|m(g_i,g_j) \sim \mathrm{Pois}(2\mu(d(i,j) - m(g_i,g_j))). \qquad (1)$$

However, the time of coalescence for two samples is generally unknown, although it must lie in the interval $0 \le m(g_i,g_j) < d(i,j)$. If the size of the transmitted inoculum is equal to one, then $t^{\mathrm{I}}_{s_{ij}^{(1)}} \le m(g_i,g_j) < d(i,j)$; in the scenario depicted in Figure 1, coalescence would have to occur within the host (rectangle) highlighted in a thick black line.

Most epidemic models describe nonlinear dynamics, and estimating the rate of coalescence between two pathogen samples during an outbreak is highly dependent on the demographic model used (Koelle and Rasmussen 2012; Volz 2012). However, in this study, interest lies in the individual-level rather than the population-wide dynamics. Under an assumed or hypothesized set of transmission routes, the time of lineage divergence $d(i,j)$ is known, and the rate of lineage coalescence can be derived from the specification of a model of within-host population dynamics and transmission.

Assuming a constant population size of N, the time to coalescence for two randomly sampled lineages at time t, m_t, is exponentially distributed with rate $1/N$. Under this assumption, it can be shown that the number of SNPs separating two randomly sampled lineages at time t follows a $\mathrm{Geom}((1/N)/(1/N + 2\mu))$ distribution, equivalent to $\mathrm{Geom}(1/(1+\theta))$, where $\theta = 2N\mu$ (Watterson 1975).

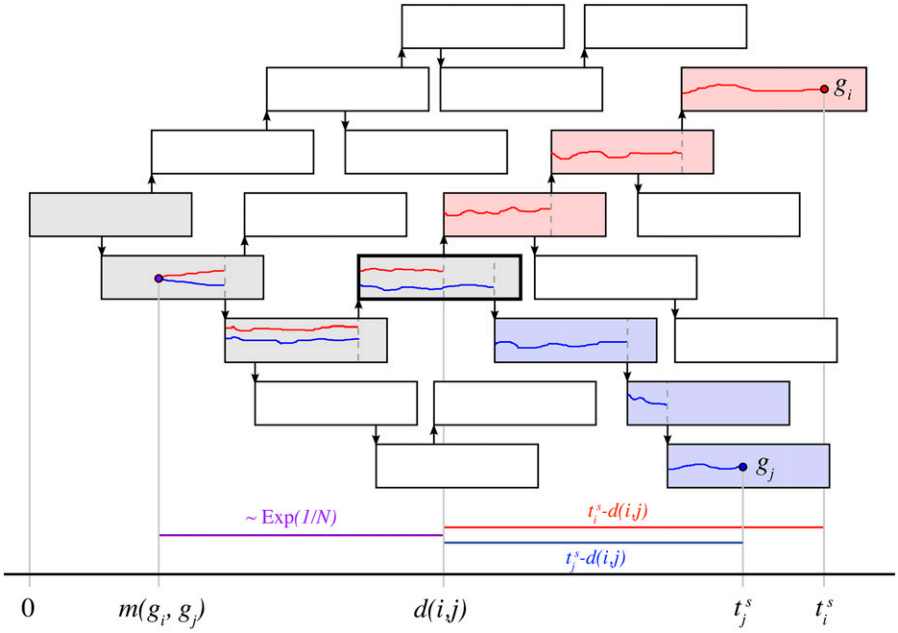

Figure 1 Two isolates sampled from infected cases during an outbreak. Each infected case is depicted by a rectangle, corresponding to its infectious period. Arrows denote transmission events. Samples g_i (red circle) and g_j (blue circle) are taken from persons i and j, respectively. The colored lines indicate the ancestry of each isolate back to its most recent common ancestor at time $m(g_i, g_j)$. Hosts shaded in gray denote the shared ancestry s_{ij}, while blue and red denote the lineages of the genotypes g_i and g_j, respectively. The colored bars at the bottom of the diagram show the distinct time periods in which mutations may occur—between divergence and observation (blue and red) and from divergence to coalescence (purple), which is exponentially distributed, assuming a constant population N.

As such, by assuming a constant mutation rate and effective population size prior to lineage divergence, we have

$$X \sim \text{Pois}(\mu(t_i^s + t_j^s - 2d(i,j))), \tag{2}$$

and

$$Y \sim \text{Geom}\left(\frac{1}{1+2N\mu}\right). \tag{3}$$

However, as the lineage is transmitted from one host to another, the population experiences repeated bottlenecks, violating the assumption of constant population size. We hence considered an approximation to the true population dynamics, using a discrete-time population model. The effective population size remains constant at size N, except during transmission, at which time it spends one generation in a bottleneck of size N_B, before recovering to its previous level. The expected time to coalescence under such a model is

$$E(m_t) = \sum_{k=0}^{t} k\left(1-\frac{1}{N}\right)^{k-\phi(k)-1}\left(1-\frac{1}{N_B}\right)^{\phi(k)}\left(\frac{1}{N(k)}\right), \tag{4}$$

where $\phi(k)$ is the number of bottlenecks a lineage must pass through between times 0 and k, and $N(k)$ is the effective population size at time k and is equal to either N or N_B. We note that $N(k)$ represents the short-term effective population size that takes into account nonrandom sampling during the bottleneck and stochastic variation, while $N_e^* = 1/E[m_{d(i,j)}]$ is the long-term effective population size that also considers the changes in short-term effective population sizes over time. We can then either assume that the time of coalescence is fixed at $\overline{m(g_i,g_j)} = d(i,j) - E(m_{d(i,j)})$ and that

$$\begin{aligned}\psi(g_i,g_j) &\sim \text{Pois}(\mu(t_i^s + t_j^s - 2\overline{m(g_i,g_j)})) \\ &= \text{Pois}(\mu(t_i^s + t_j^s - 2(d(i,j) - E(m_{d(i,j)})))\end{aligned} \tag{5}$$

[the sum of random variables (1) and (2)] or that the effective population size N_e^* prior to divergence is fixed at $1/E[m_{d(i,j)}]$ and that

$$\begin{aligned}\psi(g_i,g_j) \sim\ & \text{Geom}\left(\frac{1}{1+2E[m_{d(i,j)}]\mu}\right) \\ &+ \text{Pois}(\mu(t_i^s + t_j^s - 2d(i,j)))\end{aligned} \tag{6}$$

[the sum of random variables (2) and (3)], which we refer to as the geometric-Poisson approximation. Finally, we can derive the posterior probability of any transmission route $(i \to j)$, given the genetic distance between sampled isolates g_i and g_j and associated parameters $\omega = \{\mu, E[m_{d(i,j)}]\}$,

$$\begin{aligned}\pi(i\to j|\psi(g_i,g_j),\omega) &= \frac{\pi(\psi(g_i,g_j)|i\to j,\omega)\pi(i\to j|\omega)}{\pi(\psi(g_i,g_j)|\omega)} \\ &= \frac{\pi(\psi(g_i,g_j)|i\to j,\omega)}{\sum_{k\in S(j)}\pi(\psi(g_k,g_j)|k\to j,\omega)},\end{aligned} \tag{7}$$

assuming equal prior probabilities of potential transmission routes, where $S(j)$ is the set of all potential infection sources for individual j.

Simulation studies

We generated the empirical distribution of genetic distances by simulating within-host dynamics on top of a transmission process. We compared the resulting empirical distributions with the geometric-Poisson approximation given in Equation

6, as well as the Poisson approximation in Equation 5. The index case of the disease outbreak is infected with a clonal population of bacteria, and this is allowed to grow under a discrete-time neutral evolutionary process. At each generation, $x \sim \text{Binom}(N(t), N(t)/2N)$ cells die, and the remaining $N(t) - x$ cells are replicated, where $N(t)$ denotes the census population size at time t. We impose the restriction $x < N(t)$ to prevent the population from going extinct. Each replicated cell has a probability μ of being a mutation. All mutations are assumed to be neutral, and back mutations are allowed. A transmission event involves a bottleneck: N_B cells are randomly sampled from the host and passed to the susceptible individual. In reality, this inoculum is unlikely to be a truly random sample from the pathogen population, since a host is not a well-mixed vessel. However, N_B can be thought of as an effective bottleneck size.

Initially, we considered the simple example of a transmission chain, in which each infected individual infects exactly one new person. Transmission events occur at equidistant intervals, and the time from infection to sampling is constant. For each scenario under given parameters, we repeated the transmission chain 100 times and considered the average distribution of pairwise distance across these simulations.

We also simulated more general susceptible–infectious–removed (SIR) outbreaks in an initially susceptible population, using the R package "seedy" version 0.1 (Worby 2014). Genotypes were sampled randomly from the host at regular intervals, and person-to-person mixing in the population was assumed to be homogeneous. Outbreaks were simulated with $R_0 = 2$. We investigated the effect of varying the bottleneck size N_B, the equilibrium effective population size N_{eq}, and the mutation rate μ.

Data

We applied our approximations to a data set collected during an outbreak of MRSA. Colonization of MRSA strain type ST2371 was detected in a total of 15 newborn infants during an outbreak in a special care baby unit (SCBU) in Cambridge, United Kingdom. A single genome sampled from each of these individuals was sequenced, along with 20 isolates collected from a healthcare worker (HCW), who was found to be MRSA positive several weeks after the 15 cases were observed. The genetic similarity of the pathogen samples indicated potential transmission, (i) from patient to patient, via a transiently colonized HCW (transferring the bacteria from one patient to another, with carriage cleared upon hand washing); (ii) between persistently colonized HCW and patient; or (iii) from external sources. This study was described by Harris *et al.* (2013), and sequence data are available at the European Nucleotide Archive (www.ebi.ac.uk/ena).

Results

Within-host diversity

We first considered the distribution of pairwise genetic distances between isolates sampled from a single host. The distance between two isolates sampled at the same time point will be geometrically distributed according to the geometric-Poisson approximation (6), since the Poisson component is equal to zero. However, assuming infection with a single genotype, the empirical distribution generated from simulations can vary from this approximation (Figure 2A). This is a consequence of assuming a constant coalescent rate—under this simplification, it is assumed that the time to coalescence is exponentially distributed, while in reality, coalescence is much more likely to occur in the very early stages of infection, while the total within-host pathogen population is still expanding. With less uncertainty surrounding the coalescent time, pairwise genetic distance is approximately Poisson distributed, as in Equation 5. As the time since infection increases, the probability that coalescence occurred in the initial growth phase decreases, and the constant coalescent rate assumption of the geometric-Poisson approximation becomes more realistic.

For individuals infected with an inoculum containing multiple genotypes, the coalescence time of sampled lineages may occur within a previous host. As such, the initial diversity within a newly infected host is higher, and equilibrium levels of diversity are approached sooner than for a clonally infected host. This leads to better agreement between the empirical and geometric-Poisson distributions (Figure 2C).

The expected and empirical mean diversities are consistently similar, even when the empirical and expected distributions differ (Figure 3). However, for observations made soon after the time of infection, the approximate distribution may overestimate the frequency of genetically identical isolates. In situations where the timing of coalescence is more certain, for example, shortly after a bottleneck of size 1 (a "strict" bottleneck), a pure Poisson approximation (Equation 5) may be more appropriate (Figure 2B). We used Akaike's information criterion (AIC) to determine the better approximation at various time points after a strict bottleneck, finding the cutoff for the Poisson approximation to increase with population size N_{eq} (Supporting Information, Table S1).

Pairwise diversity along transmission chains

We next looked at the distribution of genetic distances arising from each pair of individuals in the transmission chain, simulated as described in *Methods*. Under most scenarios, the geometric-Poisson approximation correctly described the increasing mean and variance of the distribution as samples were taken farther down the transmission chain (Figure 3), with little apparent bias to the empirical mean (Figure S1). As the chain length increases, the genetic distributions reach an equilibrium, as the expected diversity of each transmission inoculum becomes constant.

Notably, there is considerable overlap between SNP distributions, meaning that the likelihood of observing a genetic distance between samples from two individuals will be similar for a range of transmission network configurations. This has

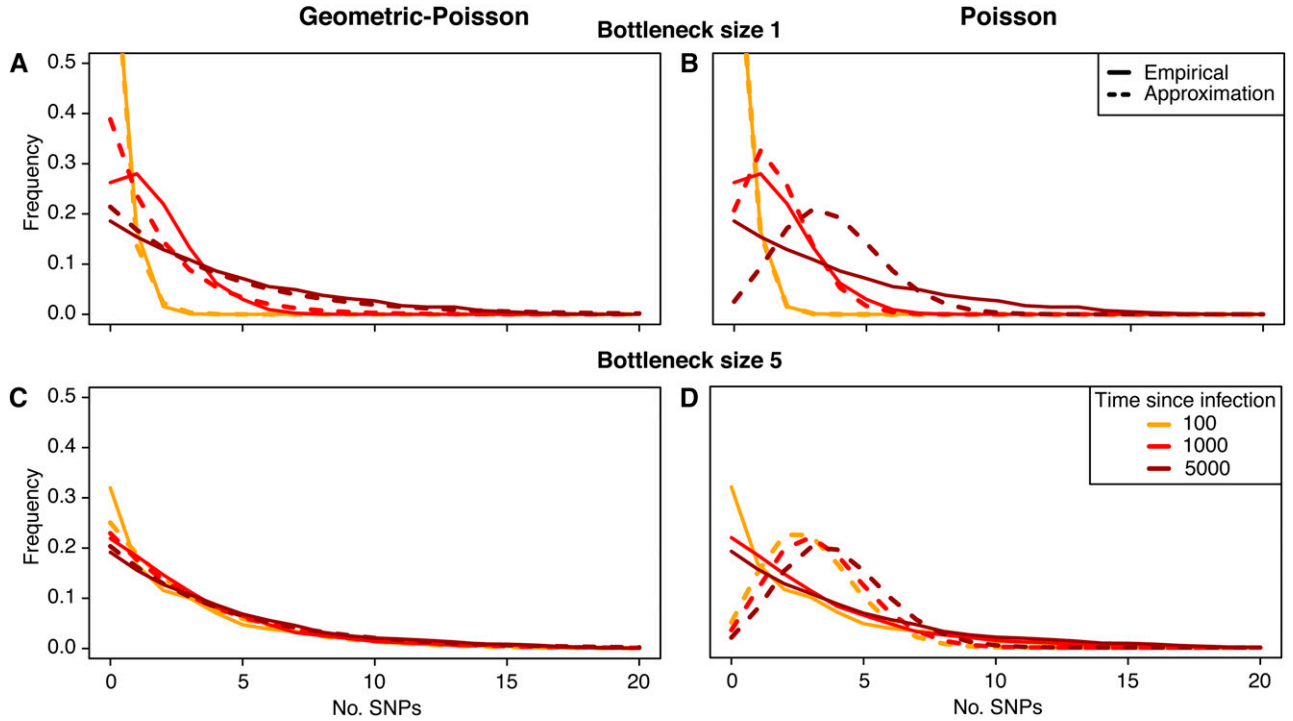

Figure 2 The empirical (solid lines) and estimated (dashed lines) distribution of genetic distances for sampling within host at specified times after infection. Both the geometric-Poisson approximation (A and C) and the simpler Poisson approximation (B and D) are shown. The infected host was infected by an inoculum of size 1 (A and B) and size 5 (C and D). The inoculum was a random sample from a bacterial population having evolved over a period of 5000 generations from an initial clonal population. Mutation rate is 0.002, and effective population size is 2000.

ramifications for identifying the source of infection, since the posterior probability of any particular transmission route will typically be low, and much uncertainty will be associated with the estimated network.

Identifying direct transmission

The geometric-Poisson distribution can be used to calculate the probability that an observed genetic distance arose from a direct transmission event. In the case where the transmission bottleneck is equal to one, the distribution of distances arising from samples taken from a transmission pair does not depend on the previous structure of the transmission network, so a probability for direct transmission can be derived independently of the outbreak structure.

We simulated SIR outbreaks and calculated the posterior probability of transmission for every pair of individuals given observed genetic distances, as derived in Equation 7. We found that the posterior probability of transmission routes corresponded well with the empirical probability calculated under repeated simulation (Figure 4). In File S1 and Figure S2, we describe a simulated disease outbreak and demonstrate the identification of potential transmission routes using the maximum likelihood, as well as the ability to rule out transmission routes at the 5% level.

To test the approximation as a tool for investigating transmission networks, we repeatedly simulated SIR outbreaks and assessed the likelihood of direct transmission between each pair of individuals, using a single sampled genotype from each host. Identification of the source of infection via maximum likelihood was consistently more successful than selection of the host with the genetically closest genotype. Furthermore, source identification was more successful for higher mutation rates. A heuristic approach, in which the infection route was selected if a potential source was both the maximum-likelihood estimate and the genetically closest host, was successful around one-third of the time (Table 2).

With a bottleneck size >1, the time of coalescence of the two sampled lineages may occur in previous hosts, and the expected time of coalescence depends on timing of bottlenecks in the bacterial population. Past population dynamics, and therefore previous transmission history, would be required to assess individual transmission links. To avoid conditioning on the remainder of the tree structure, we calculated the likelihood under the assumption that previous bottlenecks occurred at intervals equal to the expected serial interval. While we found that higher posterior probabilities were often underestimated using this approach (Figure S3), maximum-likelihood identification still consistently outperformed selection of the genetically closest host (Table S2).

We additionally compared our approach to the software "outbreaker" (Jombart *et al.* 2014) and "seqTrack" (Jombart *et al.* 2011) and found that it could identify more transmission routes correctly in many scenarios. However, differences in modeling assumptions mean the methods are not

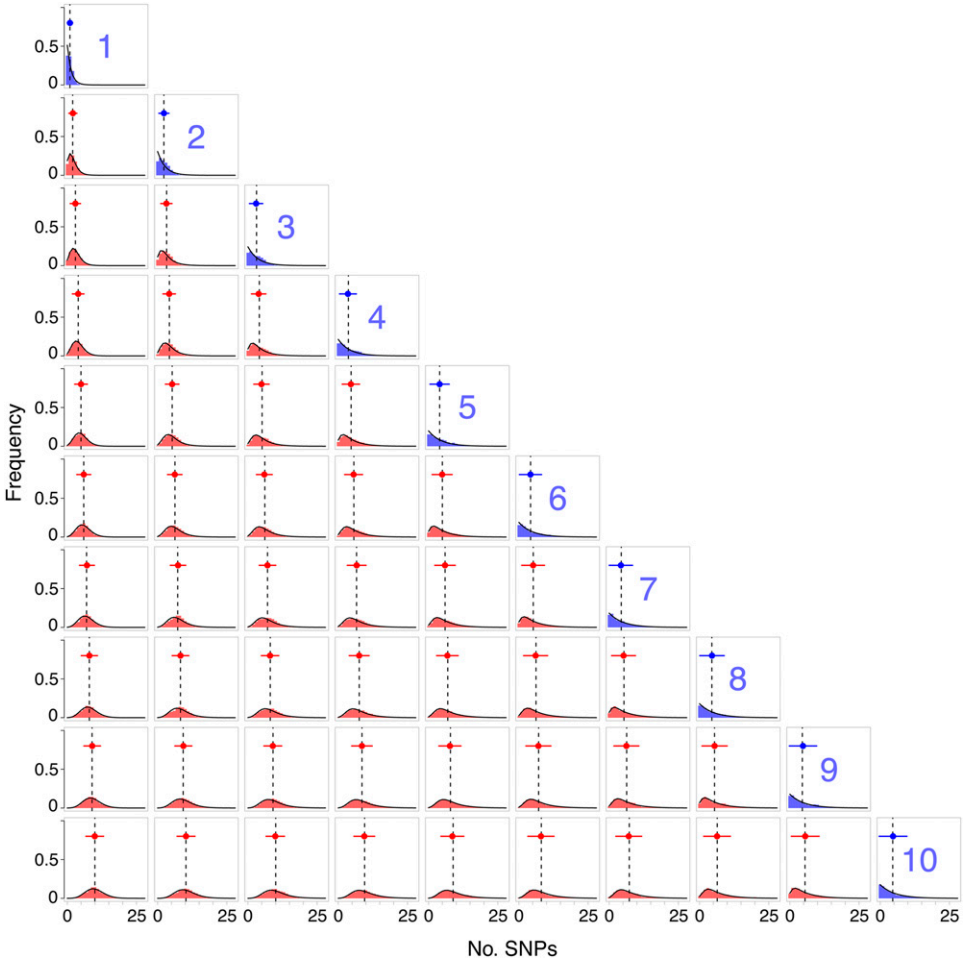

Figure 3 Genetic distance between each pair of cases in a transmission chain. The (i,j)th plot represents the empirical distribution of the genetic distance between samples taken from individuals i and j (red bars). The diagonal represents the within-host diversity for each of the 10 cases in the transmission chain (blue bars). Overlaid on each plot is the expected distribution (black line), based the geometric-Poisson approximation. The expected mean is marked with a dashed line, while the empirical mean and standard error bar are marked in red (blue for within host). The within-host equilibrium pathogen population was 10,000, with a bottleneck size of 5.

directly comparable. More details can be found in File S1 and Table S3.

Investigating transmission routes during a hospital MRSA outbreak

We used the MRSA data set described in *Methods* to investigate transmission routes in a real outbreak. We compared observed genetic distances to the geometric-Poisson approximation, to determine likely transmission routes. MRSA-positive patient episodes and swab times are shown in Figure 5A.

We initially investigated potential patient-to-patient transmission, ignoring the possibility that the HCW may have infected patients. We assumed a bacterial generation time of 30 min (Chang-Li *et al.* 1988; Dengremont and Membré 1995; Ender *et al.* 2004) and used the mutation rate of one SNP per 15 weeks (equivalent to 0.0002 per genome per generation) quoted in the study by Harris *et al.* (2013). We assumed a strict bottleneck. We found that, since the time from infection to sampling was typically short, the within-host effective population size made little difference to the approximated distributions. Five temporally consistent transmission routes could be ruled out at the 5% level, leaving five plausible transmission events (Figure 5C). Two of these form a cycle (between 11 and 12)—only one of these events could have occurred, but each route is equally plausible. The lack of any other observed and temporally consistent infection source within the ward suggests transmission from an external source or environmental contamination—however, since the infants in this study were non-ambulatory, this possibility was considered less likely.

We next supposed that the HCW could have been the source of infection for any of the patients in the SCBU. The observed mean pairwise distance between the samples collected from the HCW was 3.89 SNPs (Figure 5B), suggestive of a lengthy carriage time or a nonstrict bottleneck size. The time of HCW infection was estimated to be 23 days before the first patient case (Harris *et al.* 2013). We set the observed genetic distance from patient to HCW as the nearest integer to the mean of the genetic distances to each of the HCW's 20 samples. We found that all patients could plausibly have been infected by the HCW; however, in three cases this was not the most likely source of infection (Figure 5D). Assuming that infection must have a source from within the SCBU (including the HCW), we found that in addition to the six individuals with no other temporally consistent source, three patients had a posterior probability of >99% of acquiring infection from the HCW, while two others had a >50% probability. We additionally repeated the analysis, using each of the HCW's isolates individually (Figure S4). Furthermore, we ran the analysis using the

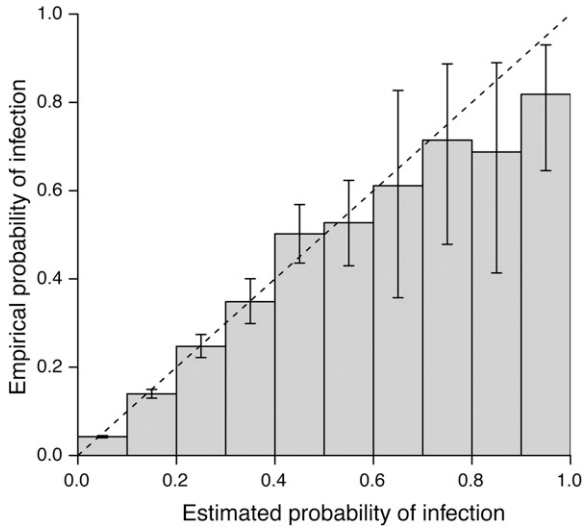

Figure 4 The empirical probability that a proposed transmission route is correct for a range of posterior probabilities calculated under the geometric-Poisson assumption. A total of 100 outbreaks were simulated and the posterior probability of direct transmission was calculated for every pair of infected individuals. Counts were collated into 10% probability bins and for each bin, the proportion of true transmission routes was calculated. Error bars depict the 95% exact binomial confidence interval.

Poisson approximation, finding little difference in transmission route probabilities (Figure S5).

We finally investigated the possibility that the HCW was infected by one of the patients on the ward. Assuming that the HCW was infected 2 days after the infection time of the potential source, we could rule out five patients as a source of infection for the HCW at the 5% level. If the HCW was infected by any one of the patients, the observed diversity within the HCW is greater than would be expected to accumulate in the period from infection to observation. At least 16% of the observed HCW within-host pairwise distances would be rejected at the 5% level under any patient–HCW transmission scenario (Table S4).

We found that, while most of our analyses were fairly robust to the specification of the effective population size, there was sensitivity to the choice of mutation rate and the time of HCW infection. We investigate these sensitivities in detail in File S1 and Table S5.

The methods we have described and implemented are for pairwise distances and, as such, cannot account for dependencies between several isolates. This is necessary when considering the transmission network as a whole, rather than just a set of pairwise connections. In addition, it is necessary to consider the conditional distribution of genetic distance to account for multiple samples per host. The degree of dependence varies considerably depending on the transmission bottleneck size (Figure S6, Figure S7). In File S1, we describe the conditional distribution for genetic distances.

Discussion

In this study, we have explored the distribution of the genetic distances arising from samples taken from infected hosts during an outbreak and investigated the impact of factors such as mutation rate, transmission dynamics, and within-host pathogen population dynamics on the expected value of such distances. Under most circumstances, a geometric-Poisson approximation is sufficient to describe genetic distances between samples taken during an outbreak. This allows the distribution to be approximated without knowing the coalescence time of two lineages. With known parameters of pathogen population dynamics, the likelihood of genetic distances arising between a host and various potential transmission sources may be compared, and certain links may be excluded. The transmission bottleneck size can have a large impact on the genetic distance distribution, and our methods can account for this.

The ability to assign a genetic distance threshold to rule out transmission events in a nonarbitrary fashion can be important in establishing distinct subgroups of the transmission tree, as well as identifying pathogen importation from outside of the studied population. This is of much importance when estimating transmission rates within a community, as incorrectly identified importations can introduce bias. Previous studies have used an arbitrary cutoff to determine potential transmission (*e.g.*, Jombart *et al.* 2014; Long *et al.* 2014).

We found that the geometric-Poisson approximation deviated from the empirical distribution to the greatest extent when sampling occurred shortly after infection with a clonal inoculum. While the expected genetic distance exhibited no apparent bias, and this deviation was minor for bottleneck sizes >1, it should be noted that this scenario may potentially be important in outbreaks of highly symptomatic pathogens, as samples are more likely to be taken in the earlier stages of infection, compared to asymptomatic, chronic infections. If a strict bottleneck is considered likely shortly before sampling, using the Poisson distribution (Equation 5) with fixed coalescent time is recommended.

Identification of transmission sources using this method is most successful with a high mutation rate. While higher mutation rates (and longer intervals between infection and onward transmission) can lead to more distinct distributions, potentially allowing one to rule out certain relationships, such as direct transmission, it is clear that even under extreme scenarios, uncertainty remains. We found that the success rate of identifying the source of infection was up to 33% better than selection of the genetically closest host, but still too low to identify transmission routes with confidence. We demonstrated that our approach could identify transmission routes more successfully than existing software packages, provided key values, such as mutation rate and infection times, are known. It has been shown previously that identification of transmission routes during an outbreak based on genomic data is likely to be challenging due to high levels of uncertainty (Worby *et al.* 2014), a finding also reflected in recent investigations (Didelot *et al.* 2014). The methods provided in this article are likely to be most valuable in the identification of a *group* of potential sources with a high likelihood, as well as the elimination of potential sources at a given probability

Table 2 Performance of geometric-Poisson distribution

Performance measure	Mutation rate ($\times 10^{-4}$)		
	1	3	5
Proportion of true infection sources identified by maximum likelihood	0.27	0.32	0.33
Proportion of true infection sources identified by closest genotype[a]	0.19	0.27	0.29
Proportion of potential links ruled out at 5% level	0.10	0.21	0.24
Proportion of true infection sources ruled out at 5% level	0.04	0.07	0.07
Proportion of cases identified as source by both maximum likelihood and genetic similarity found to be correct	0.27	0.33	0.35

SIR outbreaks with 30 initial susceptibles were simulated and a single genome sample was generated for each infective. Simulations with a final size <20 were discarded. For each infective, the maximum-likelihood source was calculated, and the genetically closest hosts were selected. All previously infected individuals were considered potential sources, regardless of removal times. Simulations for each scenario were repeated 100 times. Baseline parameters were infection rate 0.002, removal rate 0.001, and effective population size 5000.

[a] If the true source and other hosts are genetically equidistant, the true host is assumed to be identified with probability 1/(no. equidistant closest hosts).

level (discriminating, for example, between imported cases and within-population transmission events). Additional data sources, such as spatial location, contact patterns, and infectious periods, will increase the precision of estimates of infection paths (Ypma *et al.* 2012, 2013; Jombart *et al.* 2014).

We demonstrated the application of our methods to a data set collected during an MRSA hospital outbreak. We could rule out 5 of the 11 temporally consistent patient-to-patient transmission routes at the 5% level and found evidence supporting the important role played by the colonized HCW in the outbreak. However, our analysis was limited by a number of important parameter values that are uncertain or unknown. This work highlights the importance of deriving estimates for the transmission bottleneck size and gaining an improved understanding of within-host pathogen population dynamics. With less parameter uncertainty, it would be possible to draw more robust conclusions. Our analysis considered only sequence data, but other data sources could contribute valuable information to infection routes. For instance, we assume an uninformative prior distribution for infection sources, but contact patterns could potentially be factored into this, if such information were available.

While using our approximation to the genetic distance distribution can be useful to assess pairwise individuals for evidence of direct transmission, reconstruction of the full transmission network requires us to consider the conditional distribution of genetic distances and a framework to sample over the entire structure. Accounting for dependencies between genetic distances would require inference of the set of coalescent times. This approach has been described in a recent study (Ypma *et al.* 2013), which used sequence data directly, rather than genetic distance data. It may be possible to implement the distribution approximation described here, accounting for dependencies by conditioning on shared tree branches.

The transmission bottleneck size is important in the analysis of transmission dynamics, using genomic data. Most studies to date assume a strict bottleneck for convenience, as under this condition, the expected distance between two samples does not depend on pathogen population dynamics prior to the divergence of the lineages to different hosts. Previous studies have suggested a diverse transmission inoculum for influenza (Hughes *et al.* 2012; Murcia *et al.* 2012), while it is thought that the bottleneck size for bacterial transmission could vary dramatically (Balloux 2010). Conducting inference under the incorrect bottleneck size can generate misleading results. Our methods illustrate the degree to which the bottleneck size can affect the expected genetic distance between individuals and may potentially be used to assess whether a strict bottleneck is a realistic assumption.

There are several assumptions made in this work. First, we have assumed neutral evolution, such that no fitter mutant can arise and dominate the pathogen population. This may be a reasonable assumption in the short term, such as during individual carriage and in small outbreaks, but would have to be taken into account when considering epidemics over a long period of time. However, transmission route inference is most applicable to localized outbreaks within a community or a hospital, and the emergence of fitter variants may be of lesser importance. We have also assumed that the within-host pathogen population remains at equilibrium level and that this is identical for all infected individuals, while in reality this may be unrealistic, especially during antimicrobial use. Within-host pathogen dynamics are still poorly understood, and the effective size may fluctuate and vary considerably between hosts. We have primarily considered long-term bacterial infections, with a relatively stable within-host population, but alternative models could also be considered, provided the expected time of coalescence can be estimated at any given time. With appropriate sampling, methods exist to estimate the within-host effective population size, as well as the mutation rate (*e.g.*, Wang 2001; Minin *et al.* 2008). With known transmission routes, our approximation can also be used to estimate parameters of interest; however, these estimates are associated with some uncertainty (see File S1, Figure S8 and Figure S9). We assumed that the source of infection must come from the pool of observed infectives at the time of infection and furthermore that the time of infection is known. In some

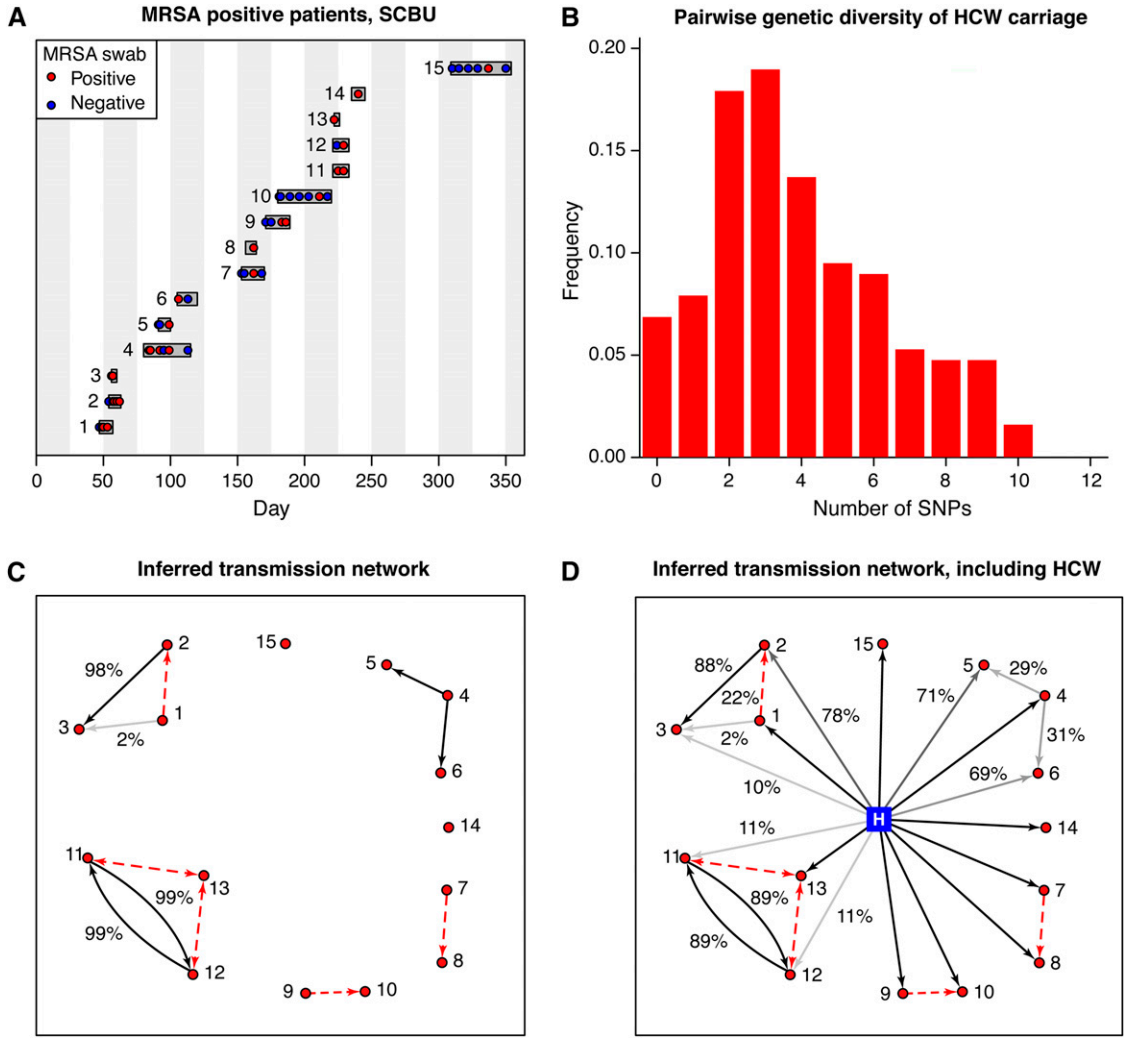

Figure 5 Data and transmission route inference for the MRSA outbreak in the SCBU. (A) Patient episodes are shown as horizontal bars, with colored circles representing positive and negative swab results. (B) The observed pairwise genetic distances between the 20 sequenced isolates collected from the HCW. (C) Inferred transmission routes are shown, excluding the possibility of HCW–patient transmission. Red dashed lines indicate routes excluded at the 5% level. All temporally consistent transmission routes are shown. Posterior probability is 100% unless stated. (D) Inferred transmission routes, including the HCW as a potential source. The HCW is marked as a blue square.

cases, particularly for outbreaks in large, well-mixing communities, it is unlikely that all infected cases will be identified and sampled. Nonetheless, evidence for an external source of infection can be seen when all potential observed sources are ruled out (for instance, cases 2 and 8 in the MRSA outbreak when not considering the HCW). In many cases, transmission times are unknown, although for many infections this can be estimated from the time of symptom onset or at least narrowed down by swabs for pathogen presence. Although one can test the hypothesis that an individual was infected at a certain time, this is a source of uncertainty, particularly for scenarios with a lengthy, asymptomatic infection period and/or a low pathogen mutation rate.

Genetic distances are an important and frequently used feature of genome sequence data, and our work contributes to a better understanding of how such distances arise during an outbreak. While sequence data provide a wealth of information regarding evolutionary history and relatedness of genotypes, the phylogeny derived from such data by itself may not be informative of transmission dynamics, and methods to combine this structure with the transmission tree are complex and computationally intensive (Ypma *et al.* 2013). Genetic distances offer a simple summary statistic of complex multidimensional data and may be more appropriate in comparative analyses of genomic samples. Genetic distances can crudely be used to determine direct transmission, via selection of the genetically closest host, but our simulations demonstrate that this approach may frequently be misleading. The geometric-Poisson approximation offers a less arbitrary method of quickly assessing the likelihood of direct transmission without requiring computationally intensive Monte Carlo sampling strategies. It may additionally provide an important component in the development of a full

transmission network reconstruction methodology based on genetic distance data.

Acknowledgments

We are grateful to R. J. F. Ypma for constructive comments and suggestions during the writing of this article and E. J. P. Cartwright for providing additional details on the MRSA outbreak data set. Research reported in this article was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award no. U54GM088558. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Literature Cited

Balloux, F., 2010 Demographic influences on bacterial population structure, pp. 103–120 in *Bacterial Population Genetics in Infectious Diseases*, edited by D. A. Robinson, D. Falush, and E. J. Feil. John Wiley & Sons, New York.

Chang-Li, X., T. Hou-Kuhan, S. Zhau-Hua, Q. Song-Sheng, L. Yao-Ting *et al.*, 1988 Microcalorimetric study of bacterial growth. Thermochim. Acta 123: 33–41.

Cottam, E. M., G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley *et al.*, 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc. R. Soc. Ser. B 275: 887–895.

Dengremont, E., and J. M. Membré, 1995 Statistical approach for comparison of the growth rates of five strains of Staphylococcus aureus. Appl. Environ. Microbiol. 61: 4389–4395.

Didelot, X., J. Gardy, and C. Colijn, 2014 Bayesian analysis of infectious disease transmission from whole genome sequence data. Mol. Biol. Evol. 31: 1869–1879.

Ender, M., N. McCallum, R. Adhikari, and B. Berger-Bächi, 2004 Fitness cost of SCCmec and methicillin resistance levels in Staphylococcus aureus. Antimicrob. Agents Chemother. 48: 2295–2297.

Harris, S. R., E. J. P. Cartwright, M. E. Török, M. T. G. Holden, N. M. Brown *et al.*, 2013 Whole-genome sequencing for analysis of an outbreak of methicillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect. Dis. 13: 130–136.

Hughes, J., R. C. Allen, M. Baguelin, K. Hampson, G. J. Baillie *et al.*, 2012 Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS Pathog. 8: e1003081.

Jombart, T., R. M. Eggo, P. J. Dodd, and F. Balloux, 2011 Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106: 383–390.

Jombart, T., A. Cori, X. Didelot, S. Cauchemez, C. Fraser *et al.*, 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput. Biol. 10: e1003457.

Koelle, K., and D. A. Rasmussen, 2012 Rates of coalescence for common epidemiological models at equilibrium. J. R. Soc. Interface 9: 997–1007.

Long, S. W., S. B. Beres, R. J. Olsen, and J. M. Musser, 2014 Absence of patient-to-patient intrahospital transmission of Staphylococcus aureus as determined by whole-genome sequencing. mBio 5: e01692-14.

Minin, V. N., E. W. Bloomquist, and M. A. Suchard, 2008 Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25: 1459–1471.

Morelli, M. J., G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon *et al.*, 2012 A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput. Biol. 8: e1002768.

Murcia, P. R., J. Hughes, P. Battista, L. Lloyd, G. J. Baillie *et al.*, 2012 Evolution of an Eurasian avian-like influenza virus in naïve and vaccinated pigs. PLoS Pathog. 8: e1002730.

Volz, E. M., 2012 Complex population dynamics and the coalescent under neutrality. Genetics 190: 187–201.

Wang, J., 2001 A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet. Res. 78: 243–257.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7: 256–276.

Worby, C. J., 2014 *Seedy: Simulation of Evolutionary and Epidemiological Dynamics*. Available at: CRAN: The Comprehensive R Archive Network (http://cran.r-project.org/web/packages/seedy/). Accessed July 14, 2014.

Worby, C. J., M. Lipsitch, and W. P. Hanage, 2014 Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput. Biol. 10: e1003549.

Ypma, R. J. F., A. M. A. Bataille, A. Stegeman, G. Koch, J. Wallinga *et al.*, 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc. R. Soc. Ser. B 279: 444–450.

Ypma, R. J. F., W. M. van Ballegooijen, and J. Wallinga, 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195: 1055–1062.

Communicating editor: J. D. Wall

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171538/-/DC1

The Distribution of Pairwise Genetic Distances: A Tool for Investigating Disease Transmission

Colin J. Worby, Hsiao-Han Chang, William P. Hanage, and Marc Lipsitch

DOI: 10.1534/genetics.114.171538

File S1

1. Estimation of parameters

While the Geometric-Poisson distribution appears to approximate the distance distribution under simulation well, this is under the assumption that several key parameters of interest are known – namely, the mutation rate, the equilibrium effective population size within-host, and the bottleneck size. With a known transmission structure (for instance, within a household (COWLING *et al.* 2010)), it is possible to estimate some of these quantities. We simulated an outbreak and assumed that a set of 25 transmission pairs was observed. Figure S8 shows the likelihood of these data under a range of values for mutation rate and effective population size. The estimate of the effective population size is uncertain, since the data are less informative of this parameter; in the most extreme case, where coalescence occurs immediately prior to the time of lineage divergence, the likelihood function depends only on the mutation rate.

The bottleneck size can additionally be estimated. Observation of multiple genotypes shortly after a bottleneck event suggests that the bottleneck must be large enough to allow diversity through; Figure S9 shows the likelihood of observing different numbers of SNPs within host shortly after transmission, for a range of potential bottleneck sizes. Again, such estimates are associated with very high levels of uncertainty, particularly for large bottleneck sizes. However, it may be possible to test the hypothesis that the bottleneck size is strict, an assumption frequently made in transmission network reconstruction methods.

2. Simulated outbreak

Figure S2 shows a simulated SIR outbreak with 25 infected individuals, 18 of which have a sampled genotype. We considered the relative likelihood of observing a genetic distance between two hosts, given direct transmission has occurred (Figure S2, bottom left). The maximum likelihood estimate of transmission source was correct in eight out of 17 transmission events. In comparison, selecting the genetically closest isolate as the source was correct in seven cases, although for some of these, multiple hosts were equally close.

For any given infected host, a genetic distance threshold may be specified, which may be used to rule out direct transmission to a given probability level. Consider the individual labelled 'N' in figure S2, with a sample at time 1000. Under the geometric-Poisson approximation with strict bottleneck, the probability of drawing a sample differing by 4 SNPs or greater at time 1000 from the true host is less than 5%. As such, six of the eleven previously infected individuals can be ruled out as transmission sources at this level. As the time between samples and/or the bottleneck size increase, this threshold also increases.

3. Comparison with transmission network estimation software packages.

'*Outbreaker*' is an R package for the investigation of individual-level transmission dynamics using genomic data (JOMBART *et al.* 2014), while '*seqTrack*' is an earlier and simpler method, implemented in the '*adegenet*' package (JOMBART *et al.* 2011). These software packages are arguably the most accessible tools for estimating a transmission network available at present, and as such, we wanted to compare their performance against our method. Given a user-specified infectivity distribution and one genomic sample per infected host, *outbreaker* implements an MCMC

algorithm which estimates the posterior edge probabilities of the network, along with several parameters of interest, including the mutation rate. Unlike our model, this approach therefore does not require infection times and mutation rate to be known (and can also be used to detect importations into a population), however, it operates on a less sophisticated model of within-host dynamics – mutations are assumed to be a feature of transmission, and an infected host is adequately represented by a single sequenced pathogen isolate. *seqTrack* identifies the genetically closest pathogen sample as the source, using the specified mutation rate to break ties. This approach also assumes that each host is represented by one genomic sample.

We simulated outbreaks under various assumptions, and attempted to identify the transmission network using our likelihood approach, as well as the *outbreaker* and *seqTrack* functions. While the *outbreaker* package can also be used to simulate outbreaks, this is performed under the assumptions mentioned previously, so we instead simulated the within-host pathogen dynamics explicitly, as described in *Methods*. We used the number of transmission routes to compare the two methods. We ran *outbreaker* with no spatial model, and detection of importations suppressed. Furthermore, we assumed a flat infectivity distribution. We emphasize that these approaches are not directly comparable, since *outbreaker* and *seqTrack* accommodate unknown infection times, and *outbreaker* furthermore estimates the mutation rate, giving our approach an advantage in this comparison. Results are presented in Table S2.

4. MRSA outbreak analysis

While the analysis provided in the main text provides estimates of transmission routes under plausible parameter values found in the literature, there is a great deal of uncertainty surrounding true within-host pathogen population dynamics, and as such, we repeated the analysis under a range of assumptions. The mutation rate used in the main analysis was given in the paper describing this dataset; the mutation rate of MRSA has previously been estimated to be higher (3×10^{-6} per nucleotide per year, equivalent to 5×10^{-4} per genome per generation (HARRIS *et al.* 2010; YOUNG *et al.* 2012)), so we repeated the analysis with this value. With this higher mutation rate, a larger range of genetic distances are plausible, and as such, fewer routes were excluded at the 5% level. The HCW was a plausible source for most patients on the ward, however, the genetic distance from patients 1 and 5 to the HCW were more similar than would be expected, given this infection route. No patient to HCW transmission route could be excluded at the 5% level.

Changing the effective population size had a limited effect on the estimated transmission route estimates. Values of 2000 and higher produced near identical posterior probabilities. Previous studies have estimated nasal carriage of *S. aureus* to have an effective population size in the range of 50-4000 (YOUNG *et al.* 2012; GOLUBCHIK *et al.* 2013). We experimented with an effective population size of 100, finding that five patient-HCW routes, and seven HCW-patient routes could be excluded at the 5% level.

Varying the time at which the HCW became infected had an impact on posterior transmission probabilities. Moving this value forward in time decreases the number of SNPs expected to accumulate by the time of observation. If the HCW infection time was 164 days after the first case, the upper bound of the range provided by (HARRIS *et al.* 2013), five patients remain temporally consistent with having become infected by the HCW. Two of these transmission routes can be excluded at the 5% level.

We repeated our analysis using the pure Poisson model. In general, this distribution has a shorter right tail than the geometric-Poisson distribution, and as such, can lead to more transmission routes being rejected at a given probability level. With the same assumptions as in the main text, the HCW-patient routes were typically given a higher posterior probability under the Poisson distribution, however, the most likely source of infection remained the same for all individuals (Figure S5).

5. Conditional distributions

We define a phylogenetic subtree to be the unique set of branch segments linking two isolates, originating at the time of their coalescence. Then the genetic distance $\psi(g_1,g_2)$ is dependent on another distance $\psi(g_3,g_4)$ by the intersection of the two phylogenetic subtrees. The conditional distribution of one genetic distance given another is

$$\psi(g_1,g_2)\,|\,\psi(g_3,g_4) \sim \mathrm{Bin}\left(\psi(g_3,g_4), \frac{\text{length of intersection}}{\text{length of subtree}(g_3,g_4)}\right) + \mathrm{Pois}\{\mu((\text{length of subtree}(g_1,g_2)) - (\text{length of intersection}))\} \quad (8)$$

Figure S7 shows two possible configurations of the phylogenetic and transmission tree with three infected cases. In both settings, $\psi(g_2,g_3)$ depends on $\psi(g_1,g_2)$ via the mutations occurring along branch b_3. If the sequences at the internal nodes are known, or can be inferred, this estimation is unnecessary, as the true number of mutations along any given branch segment can be calculated. However, since the genealogy is not typically observed, and does not necessarily correspond to the transmission network, even under a strict bottleneck (PYBUS and RAMBAUT 2009; YPMA *et al.* 2013), such an approximation may be useful for inference of the full network, and to account for multiple samples per host.

Transmission chains of length 3 were simulated to investigate conditional distributions of genetic distances. Times from infection to sampling and onward transmission were identical for all cases. With a strict bottleneck, $\psi(g_2,g_3)$ varies only minimally with $\psi(g_1,g_2)$, but $\psi(g_1,g_3)$ shows a clear dependency. Both distances increase with greater values of $\psi(g_1,g_2)$ under larger bottlenecks (Figure S6). With a strict bottleneck, the scenario in Figure S7B is impossible, and as such, the intersection of subtrees (g_1,g_2) and (g_2,g_3) is relatively small. With an increasing bottleneck size, the probability of scenario B, and therefore the potential length of subtree overlap, increases.

References

COWLING, B. J., K. H. CHAN, V. J. FANG, L. L. H. LAU, H. C. SO *et al.*, 2010 Comparative Epidemiology of Pandemic and Seasonal Influenza A in Households. New England Journal of Medicine 362: 2175-2184.

GOLUBCHIK, T., E. M. BATTY, R. R. MILLER, H. FARR, B. C. YOUNG *et al.*, 2013 Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. PLoS One 8: e61319.

HARRIS, S. R., E. J. P. CARTWRIGHT, M. E. TÖRÖK, M. T. G. HOLDEN, N. M. BROWN *et al.*, 2013 Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infectious Diseases 13: 130-136.

HARRIS, S. R., E. J. FEIL, M. T. G. HOLDEN, M. A. QUAIL, E. K. NICKERSON *et al.*, 2010 Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469-474.

JOMBART, T., A. CORI, X. DIDELOT, S. CAUCHEMEZ, C. FRASER *et al.*, 2014 Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Computational Biology 10: e1003457.

JOMBART, T., R. M. EGGO, P. J. DODD and F. BALLOUX, 2011 Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106: 383-390.

PYBUS, O. G., and A. RAMBAUT, 2009 Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10: 540-550.

YOUNG, B. C., T. GOLUBCHIK, E. M. BATTY, R. FUNG, H. LARNER-SVENSSON *et al.*, 2012 Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. PNAS 109: 4550-4555.

YPMA, R. J. F., W. M. VAN BALLEGOOIJEN and J. WALLINGA, 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195: 1055-1062.

SI Figures

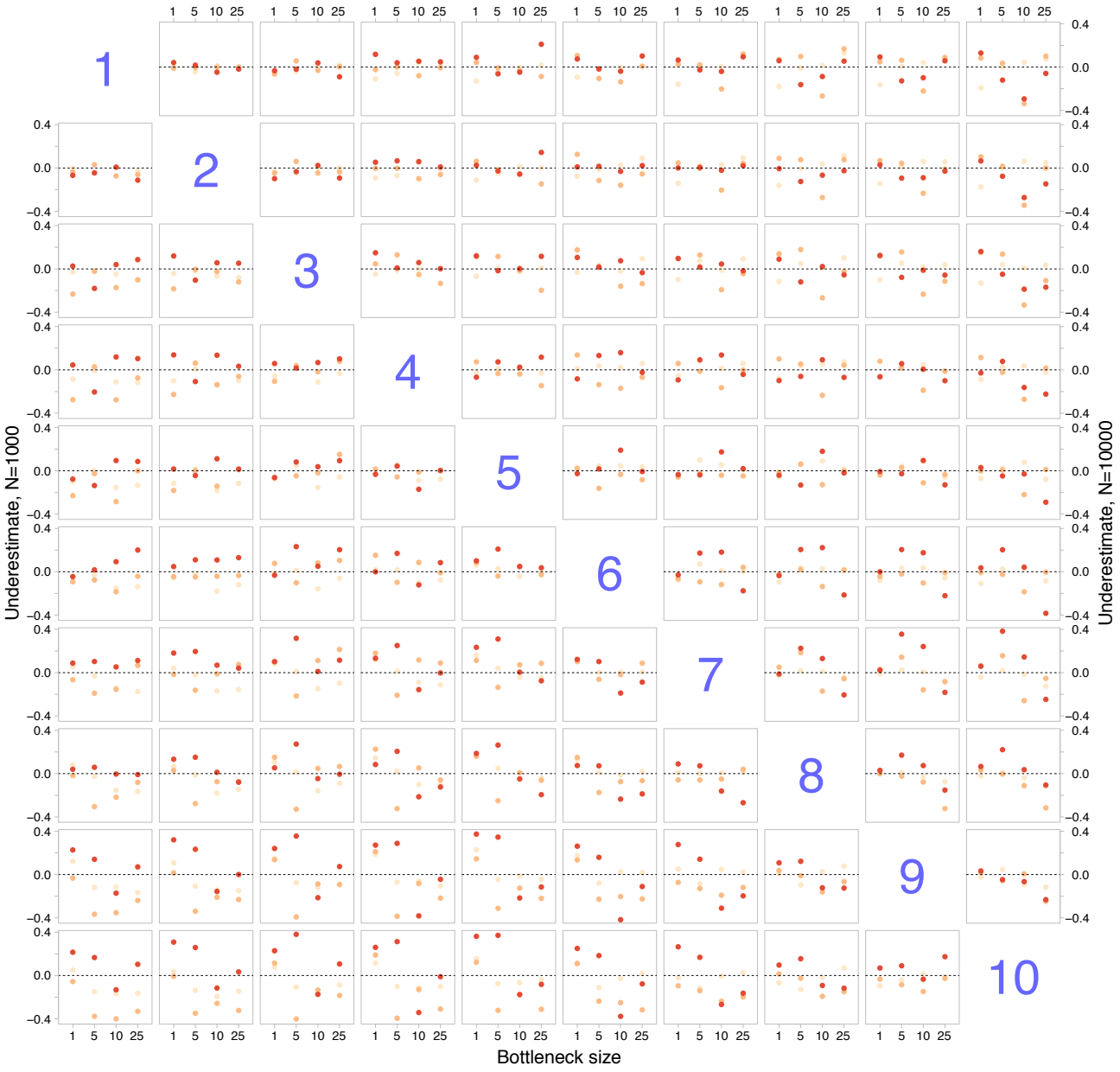

Figure S1. Differences between empirical and estimated pairwise genetic distances using the Geometric-Poisson approximation. The (i, j) th plot shows the difference between the empirical and simulated mean distance between samples taken from individuals i and j. Each plot shows the underestimate for various levels of bottleneck size and mutation rate (light, medium and dark points denote 1×10^{-4}, 3×10^{-4}, and 5×10^{-4} respectively). Plots above the diagonal show underestimates for equilibrium population size 10000, while below the diagonal, N_{eq}=1000.

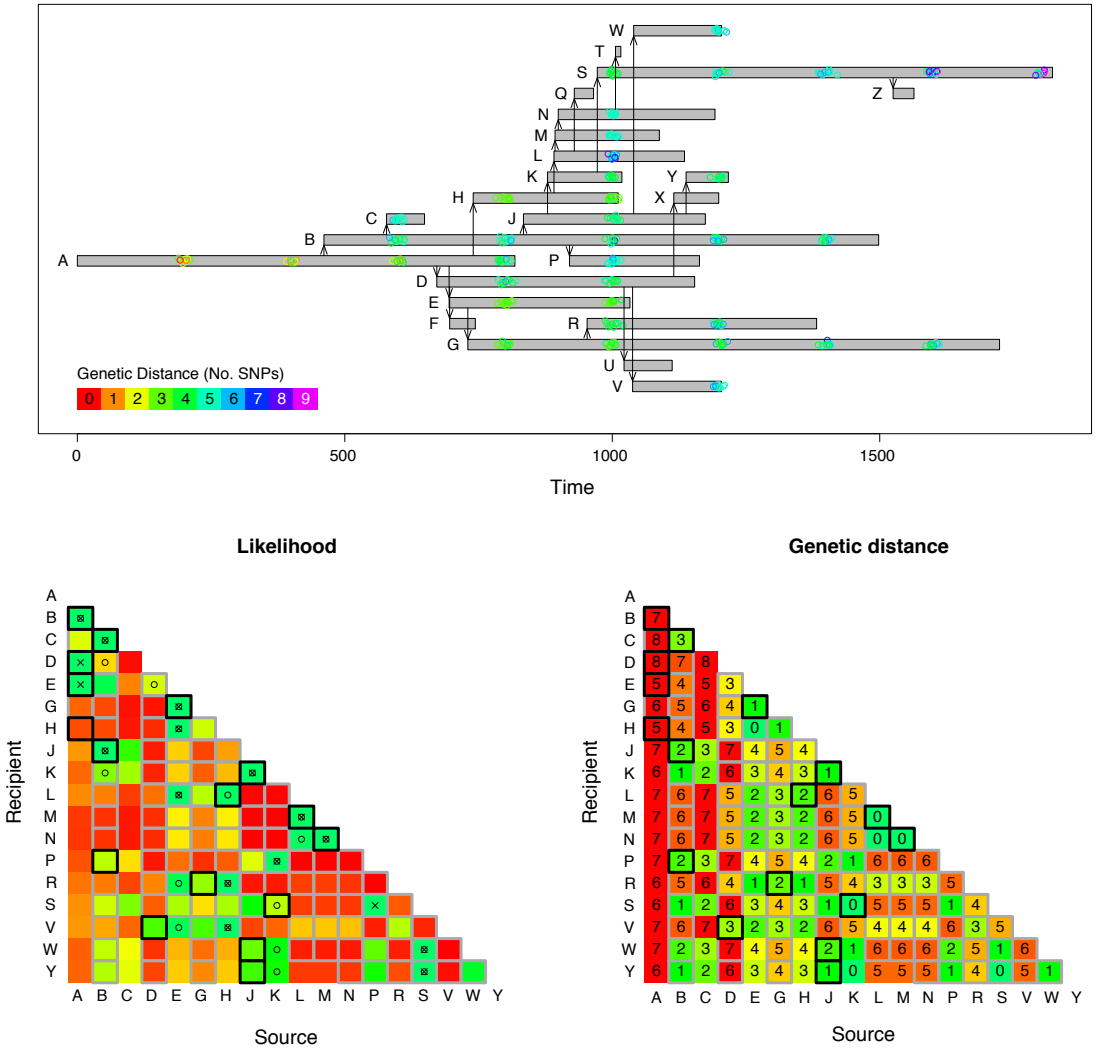

Figure S2. A simulated outbreak. 24 individuals are infected in a simulated SIR outbreak, of which 18 have sampled genotypes. Each individual has an infectious period shown as a gray bar, with genotypes shown as colored circles, the color denoting the genetic distance from the first sample (top). One randomly sampled genome for each individual is used to assess the likelihood of direct transmission from each other sampled individual. The pairwise genetic distances are shown (bottom right), with black boxes denoting the true source of infection, and gray boxes denoting presence at the time of infection. The relative likelihood of direct transmission using the geometric-Poisson approximation is shown for each pair (bottom left, green and red indicating high and low relative likelihood respectively). Crosses indicate the maximum likelihood estimate, while circles indicate the genetically closest isolate to each sample.

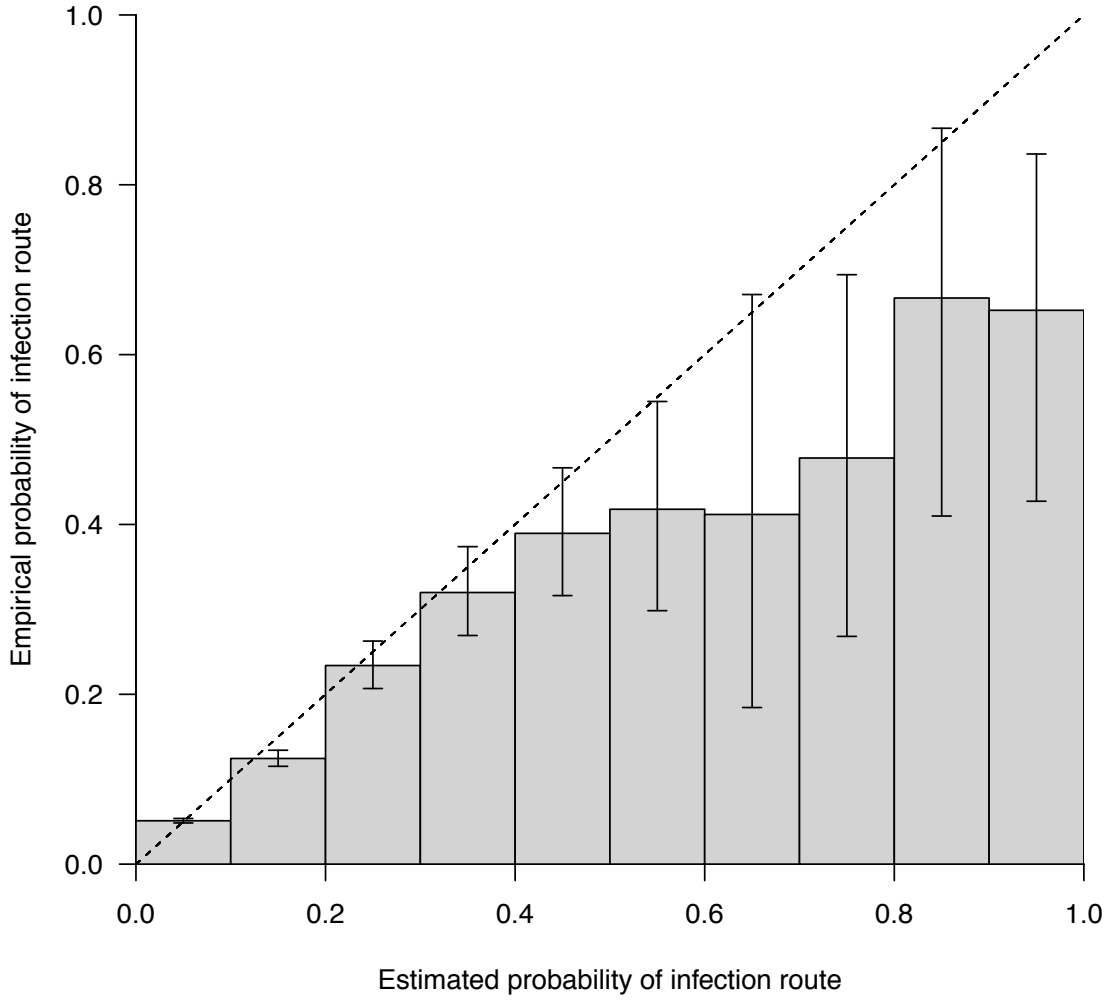

Figure S3. The empirical probability that a proposed transmission route correct for a range of posterior probabilities calculated under the geometric-Poisson assumption. A total of 100 outbreaks were simulated with a bottleneck size of 5; transmission events prior to the host were assumed to occur at intervals equal to the mean generation interval. The posterior probability of direct transmission was calculated for every pair of infected individuals. Counts were collated into 10% probability bins and for each, the proportion of true transmission routes calculated. Error bars depict the 95% exact binomial confidence interval.

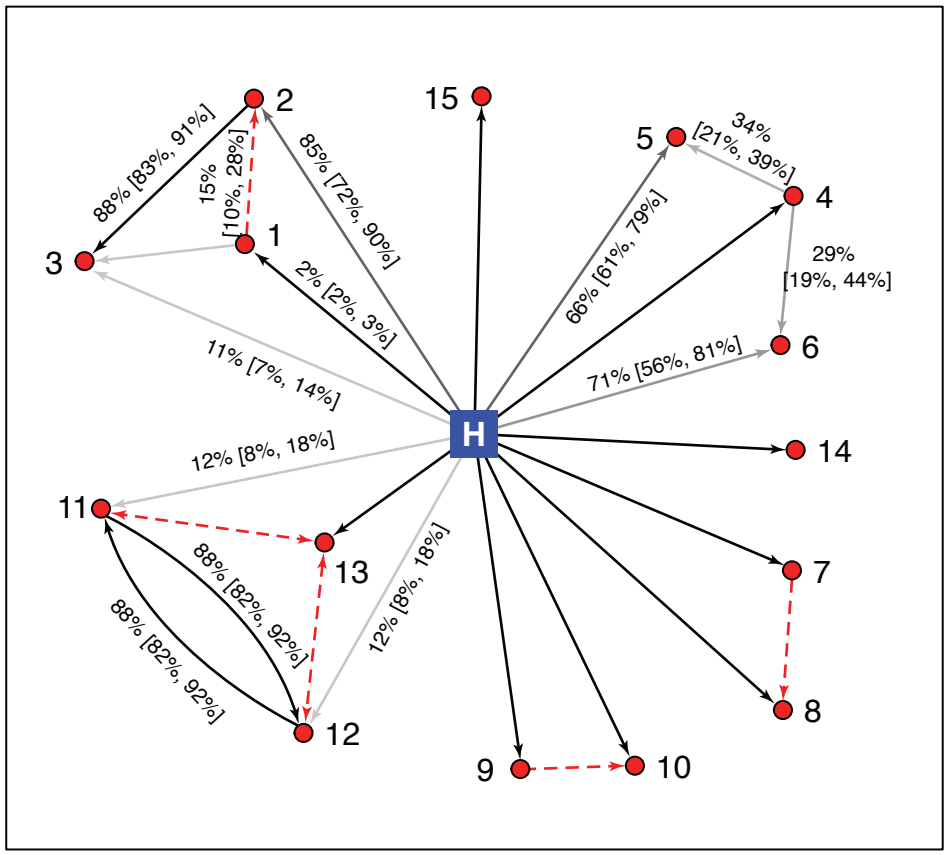

Figure S4. Transmission network in the SCBU, using each HCW isolate individually. HCW is shown as a blue square, potential transmission routes are shown as arrows. Red dashed arrows denote transmission routes rejected at the 5% level using the geometric-Poisson approximation. For each of the 20 HCW isolates, posterior transmission probabilities were calculated individually, and the mean and range of values are indicated on the plot.

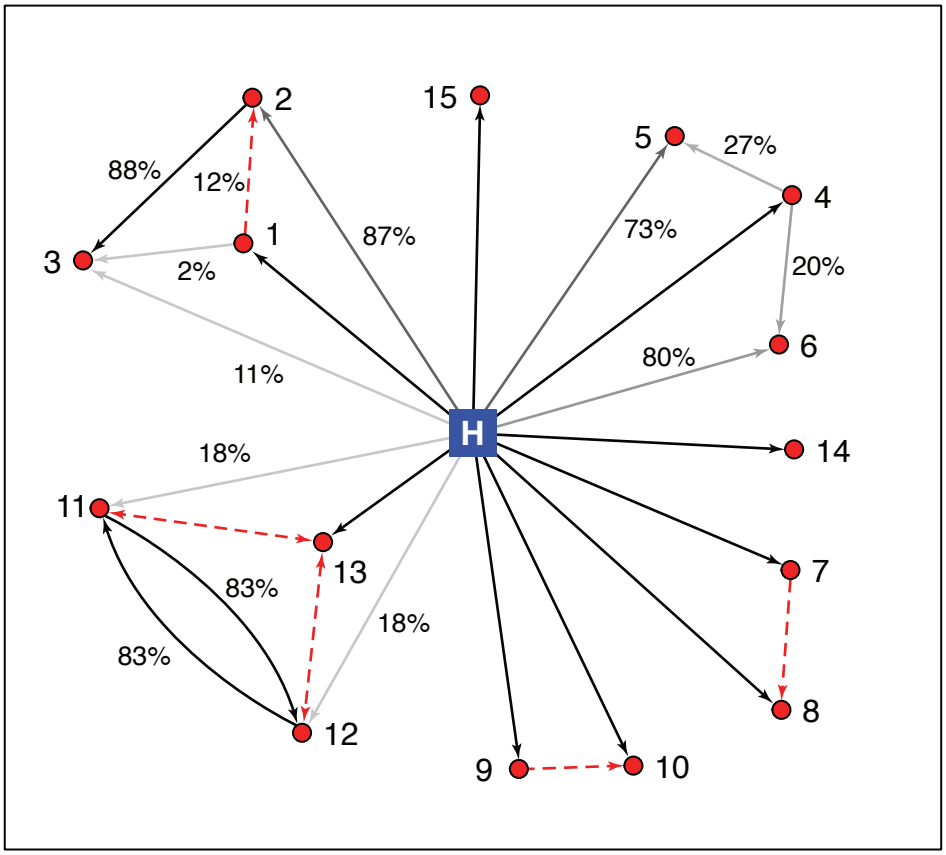

Figure S5. Transmission network in the SCBU, using the pure Poisson approximation. HCW is shown as a blue square, potential transmission routes are shown as arrows. Red dashed arrows denote transmission routes rejected at the 5% level using the Poisson approximation.

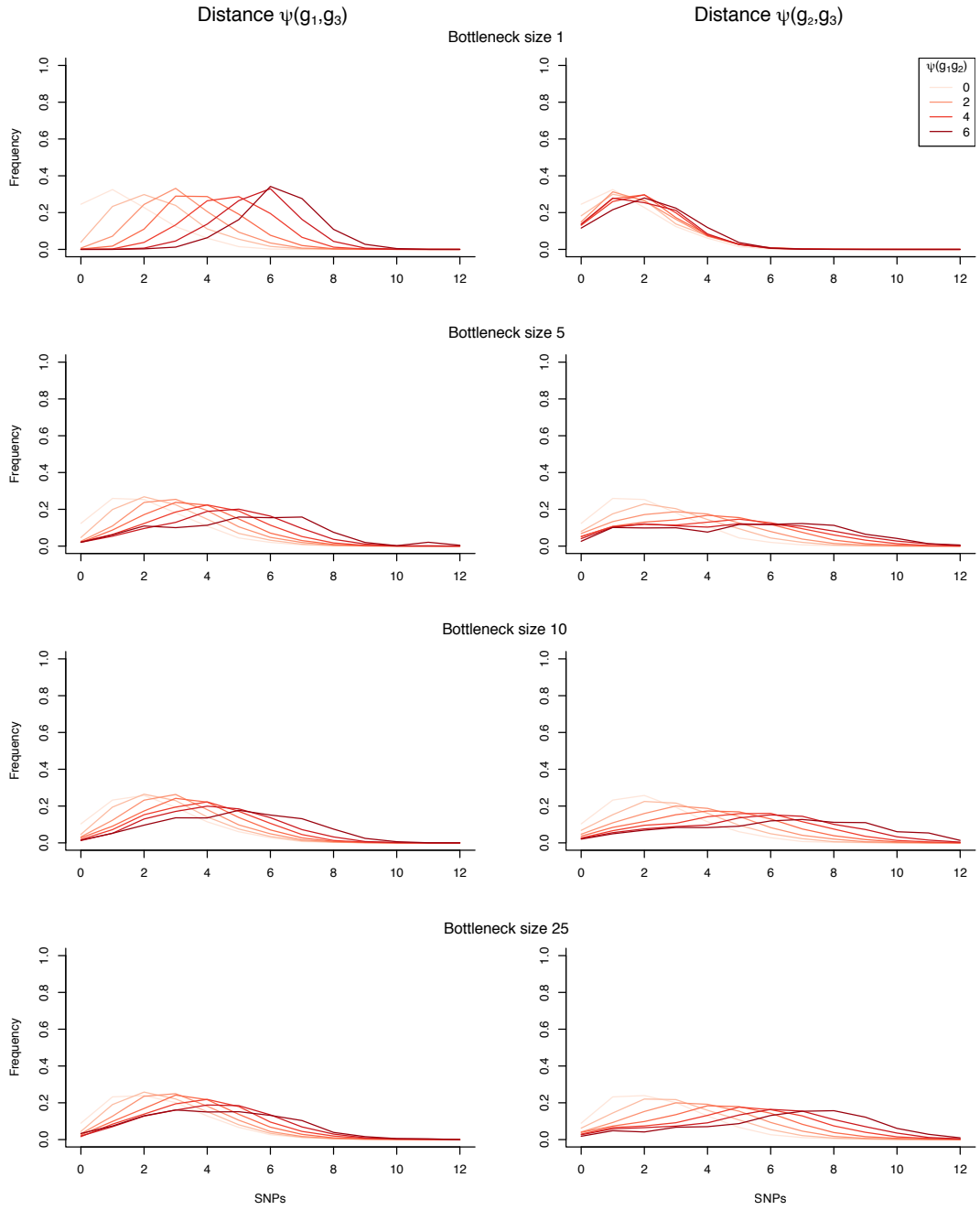

Figure S6. Simulated conditional distributions of genetic distances arising from a transmission chain of length 3. Each row shows plots for $\psi(g_1,g_3)$ and $\psi(g_2,g_3)$ given various levels of $\psi(g_1,g_2)$ (denoted by different colors). Bottleneck size varies by row. Equilibrium size was set to 10000, and mutation rate $\mu = 3\times10^{-4}$.

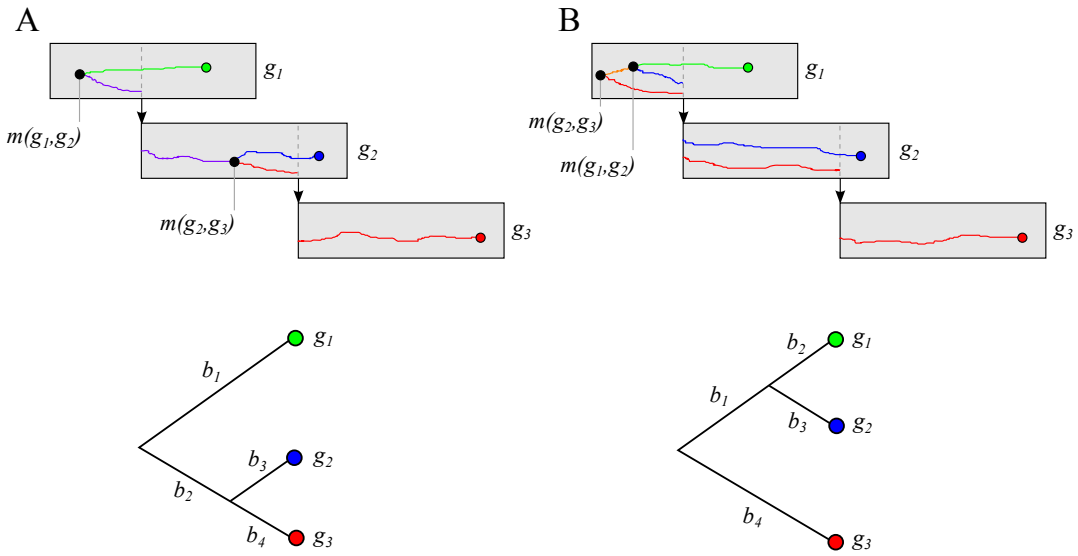

Figure S7. Two possible phylogenetic configurations in a transmission chain of length 3. (A) Lineages g_2 and g_3 coalesce within host 2. (B) Lineages g_2 and g_3 coalesce within host 1, prior to the coalescence of g_1 and g_2. This configuration is possible only with a bottleneck of size > 1.

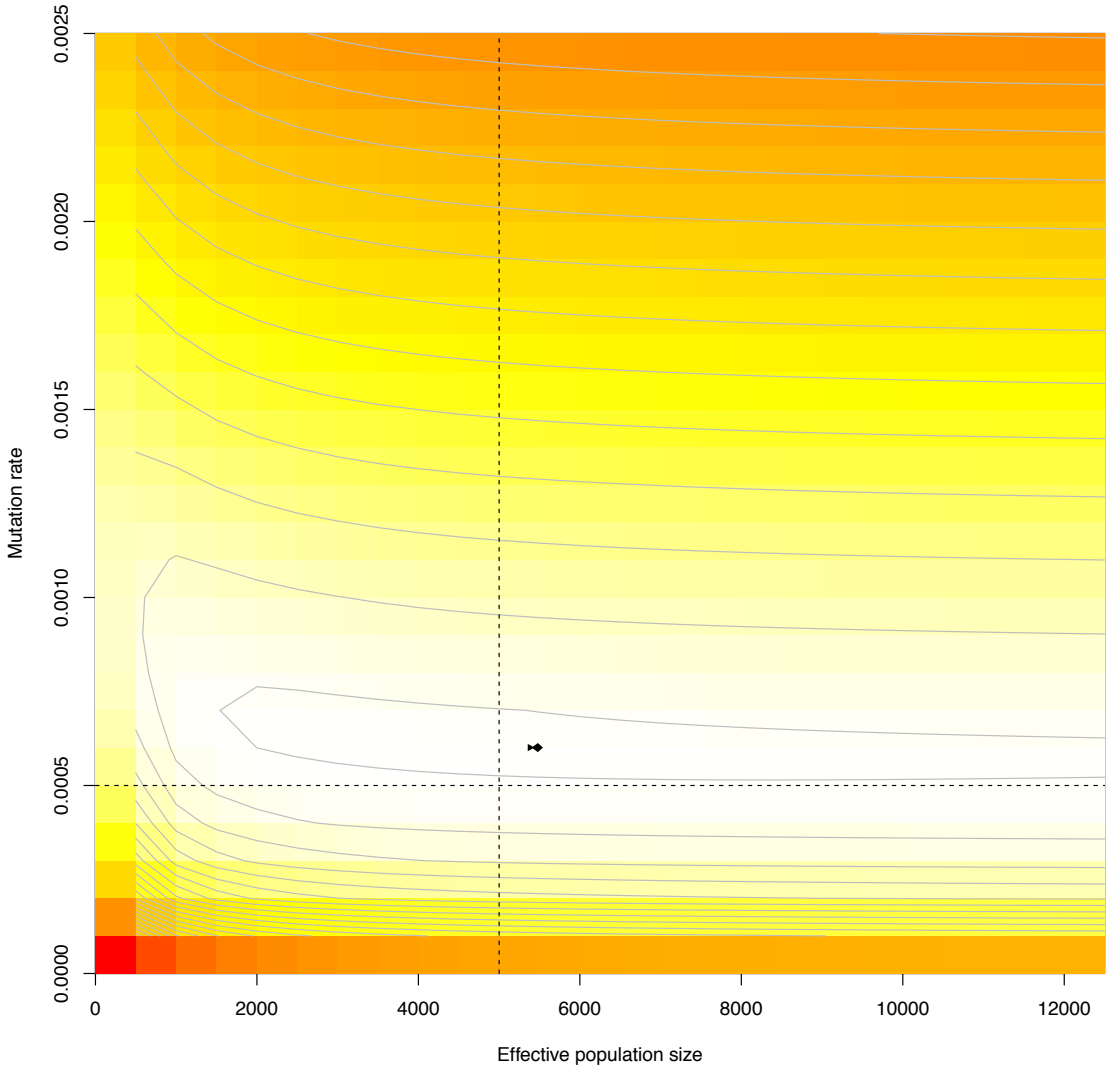

Figure S8. Likelihood of observing 28 pairwise genetic distances between known transmission pairs, given a range of values for the mutation rate and the effective population size. The dashed lines indicate parameter values under which the data were simulated, and the geometric-Poisson maximum likelihood value is marked. Maximum likelihood value calculated using the Nelder-Mead method in the 'optim' function in R.

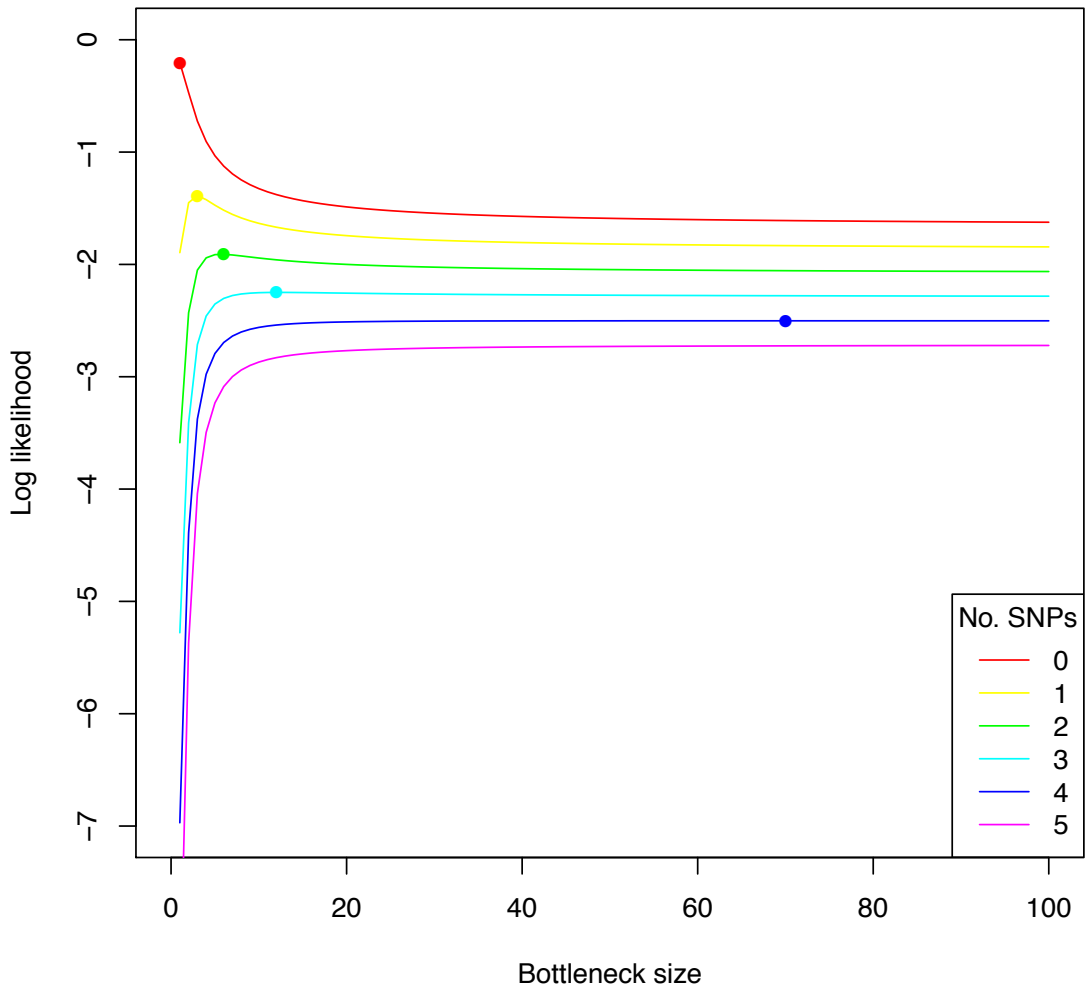

Figure S9. Likelihood curves for various within-host genetic distance observations, given a range of transmission bottleneck sizes. The effective population size and mutation rate are assumed to be known. The likelihood is calculated assuming samples are taken 50 generations after a transmission event; the maximum likelihood estimate of bottleneck size for each genetic distance is marked as a filled circle.

SI Tables

Table S1. The differences between approximated and empirical distributions for within-host genetic distances. For a range of μN_{eq} and times since clonal infection, Akaike's Information Criterion (AIC) is given for both the geometric-Poisson (GP) and the Poisson (P) approximation. 250 simulated pathogen populations were generated, and for each, 1000 pairwise distances were recorded at each of the sample times. Cells are shaded according to the lower AIC value – red for Poisson, green for geometric-Poisson. The mutation rate was 0.001 per genome per generation.

		Effective population size, N_{eq}					
		500	1000	2500	5000	7500	10000
Time since clonal infection	50	GP: 75341 P: 75216	GP: 80764 P: 80334	GP: 80729 P: 80462	GP: 80734 P: 80431	GP: 78318 P: 78008	GP: 84445 P: 84162
	100	GP: 115043 P: 114067	GP: 128371 P: 126955	GP: 131869 P: 130751	GP: 133561 P: 132260	GP: 129586 P: 128358	GP: 133905 P: 132656
	500	GP: 258951 P: 257189	GP: 297052 P: 291320	GP: 323116 P: 310162	GP: 343677 P: 324330	GP: 336449 P: 319142	GP: 340266 P: 322343
	1000	GP: 324557 P: 336776	GP: 384288 P: 386824	GP: 442356 P: 421016	GP: 455690 P: 421886	GP: 459908 P: 422279	GP: 464791 P: 424528
	2500	GP: 340205 P: 360382	GP: 455889 P: 499865	GP: 559643 P: 591170	GP: 616431 P: 602032	GP: 640539 P: 601454	GP: 648459 P: 583515
	5000	GP: 355353 P: 384607	GP: 470566 P: 555942	GP: 629747 P: 772276	GP: 730920 P: 844597	GP: 758704 P: 821443	GP: 781885 P: 804054
	7500	GP: 351289 P: 384044	GP: 489139 P: 599342	GP: 656489 P: 870263	GP: 755024 P: 994202	GP: 785749 P: 986565	GP: 801616 P: 947202
	10000	GP: 349955 P: 380901	GP: 477976 P: 567623	GP: 655821 P: 898879	GP: 708001 P: 1001501	GP: 708912 P: 984256	GP: 692577 P: 942683

Table S2. Proportion of true transmission routes identified by both maximum likelihood (ML) and genetic similarity. SIR outbreaks with 30 initial susceptibles were simulated and a single genome sample was generated for each infective. For scenarios with bottleneck size >1, it was assumed that transmission events prior to the infection of the source occurred at intervals equal to the mean generation interval. Simulations with a final size <20 were discarded. For each infective, the maximum likelihood source was calculated under the geometric-Poisson approximation, and the genetically closest hosts selected. Simulations for each scenario were repeated 100 times. Baseline parameters: infection rate 0.002, removal rate 0.001, effective population size 5000.

Mutation rate ($\times 10^{-4}$)	1			3			5		
Bottleneck size	1	5	25	1	5	25	1	5	25
Prop. routes identified by ML	0.27	0.21	0.21	0.32	0.23	0.22	0.33	0.24	0.21
Prop. routes identified by genetic similarity	0.19	0.17	0.15	0.27	0.20	0.18	0.29	0.22	0.19

Table S3. Proportion of correct transmission routes identified using the geometric Poisson likelihood, as well as with the 'outbreaker' and 'seqTrack' functions. A total of 25 outbreaks with 30 susceptible individuals were simulated for each scenario, with outbreaks terminating with fewer than 20 infections excluded. R_0 was set to be 2, with a within-population size 5000. In outbreaker, no spatial model was defined, importation identification was suppressed, and the infectivity distribution was specified to be uniform. In seqTrack, the mutation rate was provided.

[a] If the true source and other hosts are genetically equidistant, the true host is assumed to be identified with probability 1/(# equidistant closest hosts).

Parameters		Network identification method			
Mutation rate	Inoculum size	ML estimate	outbreaker	seqTrack	Closest genotype[a]
0.002	1	0.28	0.20	0.14	0.21
0.002	5	0.26	0.19	0.13	0.17
0.002	10	0.24	0.19	0.14	0.16
0.005	1	0.28	0.20	0.13	0.22
0.005	5	0.22	0.18	0.12	0.18
0.005	10	0.21	0.21	0.13	0.17

Table S4. Proportion of observed within-host pairwise distances rejected at the 5% level, under the assumption that HCW infection occurred 2 days after the infection time of the patient. Proportions were calculated under both the geometric-Poisson and the pure Poisson approximations.

Source of HCW infection	Proportion of within-host pairwise distances rejected at 5% level	
	Geometric-Poisson	Poisson
Patients 1-6	0.16	0.48
Patients 7-14	0.25	0.48
Patients 15	0.35	0.48

Table S5. Transmission routes excluded at the 5% level under a range of scenarios.

Mutation rate	Eff. Pop. Size	HCW infection time (relative to first case)	HCW ruled out as patient source	Patients ruled out as HCW source
0.0002	3000	-23	NA	8,9,10,13,14
0.0005	3000	-23	NA	NA
0.0002	10000	-23	NA	8,9,10,13,14
0.0002	100	-23	NA	8,9,10,13,14
0.0002	3000	164	1-10,13,14	–
0.0002	3000	-251	NA	–