LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-414337

A Statistical Framework for Microbial Source Attribution

Stephan P. Velsko, Jonathan E. Allen, Christoph T. Cunningham

June 30, 2009

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

A Statistical Framework for Microbial Source Attribution

Part 1: Forensic inferences on disease transmission networks

Stephan P. Velsko, Jonathan Allen, and Christopher Cunningham
Lawrence Livermore National Laboratory
April 30, 2009

Executive Summary

This report presents a general approach to inferring transmission and source relationships among microbial isolates from their genetic sequences. The outbreak transmission graph (also called the transmission tree or transmission network) is the fundamental structure which determines the statistical distributions relevant to source attribution. The nodes of this graph are infected individuals or aggregated sub-populations of individuals in which transmitted bacteria or viruses undergo clonal expansion, leading to a genetically heterogeneous population. Each edge of the graph represents a transmission event in which one or a small number of bacteria or virions infects another node thus increasing the size of the transmission network. Recombination and re-assortment events originate in nodes which are common to two distinct networks.

In order to calculate the probability that one node was infected by another, given the observed genetic sequences of microbial isolates sampled from them, we require two fundamental probability distributions. The first is the probability of obtaining the observed mutational differences between two isolates given that they are separated by M steps in a transmission network. The second is the probability that two nodes sampled randomly from an outbreak transmission network are separated by M transmission events. We show how these distributions can be obtained from the genetic sequences of isolates obtained by sampling from past outbreaks combined with data from contact tracing studies. Realistic examples are drawn from the SARS outbreak of 2003, the FMDV outbreak in Great Britain in 2001, and HIV transmission cases.

The likelihood estimators derived in this report, and the underlying probability distribution functions required to calculate them possess certain compelling general properties in the context of microbial forensics. These include the ability to quantify the significance of a sequence "*match*" or "*mismatch*" between two isolates; the ability to capture non-intuitive effects of network structure on inferential power, including the "*small world*" effect; the insensitivity of inferences to uncertainties in the underlying distributions; and the concept of *rescaling*, i.e. ability to collapse sub-networks into single nodes and examine transmission inferences on the rescaled network.

1. Introduction

Microbial forensics is often concerned with the problem of identifying the most probable source of an infecting pathogen based on genetic sequence comparisons with isolates of that pathogen obtained from a set of potential sources. Consider, for example, a case where a single person (the victim V) is infected with a virus, and we have a suspected source S. (The suspect source might be another individual who was infected with the virus during an outbreak, or an isolate that was collected from an infected person during some outbreak and held by a laboratory for legitimate research purposes.) As part of the investigation, isolates of the infecting virus are obtained from V and S, and the consensus genetic sequences are determined for the two isolates. A central question is how to *quantify* and express the degree of support that the sequence data provides for the hypothesis that the virus was directly transmitted to the victim from the suspect source.

Similarly, suppose there is an unusual outbreak of an infectious disease in some human or animal population, and the genetic sequences of isolates collected from the outbreak and a suspected reservoir are determined. Can we state, on the basis of the sequences, the *probability* that the suspected reservoir was the origin of the outbreak? More generally, is there a way based on genetic data to express our confidence *quantitatively* that the outbreak was probably natural, and not the consequence of a deliberate introduction of the disease (or *vice versa*)?

A typical approach to these questions would be to construct a phylogenetic tree to compare the genetic sequences of the victim's or outbreak's isolates with those of the suspected source and a number of "background" isolates. Unfortunately, phylogenetic constructions of this sort are not an adequate basis for determining the confidence levels associated with inferences made about source relationships. This is because a phylogenetic tree can provide information about the relative evolutionary relationships only among the set of compared isolates, and cannot provide a probabilistic measure that includes other possible, but un-sampled (and possibly unknown) sub-populations of that microbe. In most cases it is difficult to identify all potential sources, reservoirs, or relevant background isolates, or the relevant isolates and their genetic sequences may not be available. This and other limitations of phylogenetic construction as an inferential tool are discussed in more detail in Appendix 1.

In human DNA forensics, "source attribution" is based on explicit probabilities for finding the questioned genetic pattern in the human population, or (in paternity/kinship cases) under different mating scenarios[1,2]. In microbial forensics the consensus genetic sequence associated with a certain subpopulation of microbes (for example, the population of pathogens infecting an individual) are compared to sequences drawn from other subpopulations contained within other distinguishable sources (other individuals, some reservoir, or a flask in a laboratory.) In this case, a useful framework for attribution can be built by considering tests of a source hypothesis based on the similarity of the genetic sequences contained in the infectee and potential source (infector) sub-

populations. The key to this approach lies in defining the subpopulations correctly, and choosing suitable metrics to evaluate genetic similarity.

In this report we present a probabilistic approach to genetic inference that is based on the explicit consideration of certain statistical properties of microbial populations of infectious diseases. We begin with the observation that outbreaks are the fundamental objects of interest for understanding those aspects of microbial population genetics relevant to microbial forensics, and for framing the hypotheses that can be tested by genetic comparisons. Understanding the underlying natural structure of outbreak populations is critical for quantifying the confidence with which hypotheses about disease transmission and disease sources can be affirmed or refuted. This underlying structure is determined by the *transmission network* which connects all sub-populations of the microbe, and the relevant probability sampling distributions are those associated with the probability of observing genetic variation after a transmission step, and the chances of drawing at random two nodes from the network related by a certain number of transmission steps. Thus, the theory presented here weds two very active recent fields of research: microbial genetic evolution[3,4] and the modern network theory of infectious disease outbreaks[5-12].

While the theory is applicable to both viral and bacterial pathogens, epidemiologically referenced sequence data extensive enough to permit estimation of the required genetic sampling distributions is currently only available for viruses. (In fact, even this data is just barely adequate to illustrate how the framework can be applied.) In the near term (perhaps three to five years) this situation is likely to persist for bacterial pathogens, but given recent advances in sequencing technology it is likely that the available viral data will greatly expand, leading to significant improvements in the range of application and predictive accuracy of the theory. Therefore this report will primarily focus on viruses. As high throughput sequencing becomes even faster and cheaper, it will become feasible to apply the framework to bacterial pathogens as well.

The remainder of this report is organized into 6 sections. In Section 2 we derive the general framework and discuss how the probability distributions that are needed to perform calculations may be determined. In section 3 we illustrate the theory's application to several source attribution problems using data from the 2003 SARS outbreak and the 2001 FMDV outbreak in Great Britain. In section 4 we show how the theory can be extended to answer questions about natural sources of disease outbreaks, including how to differentiate a natural from a deliberate event. Section 5 describes how the framework can be utilized as the basis for a CODIS-like decision support system in microbial forensics. Section 6 discusses strategies for experimental and computational validation of the framework. Finally, Section 7 summarizes the current status of the framework, identifies outstanding problems that remain to be solved, and discusses possible future directions.

2. Genetic inference on transmission trees

General framework

Consider the case described in the introduction in which sequences from pathogen isolates that are obtained from a victim V and a putative source S are compared. The key to understanding how to generate a quantitative probabilistic measure of confidence that S is the source of V's infection is to recognize that all infectious disease outbreaks are characterized by a transmission tree, i.e. a graph (in the sense of graph theory[13]) where the nodes are infected entities and the edges represent transmission events. This is illustrated in Figure 1. Although our example involves individuals, nodes can be any unit within which the viral population is defined (individual hosts, flasks in laboratories, herds, cities, etc.) The edges may or may not be assigned a direction associated with the transmission event. In Fig. 1, our putative source is shown as a node in the outbreak transmission tree to which its viral sub-population is related. (Note that the actual source could be some laboratory isolate, but this had to have been obtained ultimately from some infected host in an outbreak. Strictly speaking the laboratory isolate forms a new node in the network, especially if laboratory culture passage is involved.)

Each pair of nodes in a transmission tree such as Fig. 1 is connected by an M-step transmission relationship. (For example, the two nodes marked with ✻ in Fig. 1 are separated by M = 7 steps; for reasons explained below we do not distinguish *direction* of transmission when calculating node-to-node distances.) In addition, each node can be categorized in terms of the number of steps between it and the "index case", i.e. the first node to be infected in the outbreak. This number of steps is denoted G, the number of *generations* between the node and the index case. In Fig. 1, for example, the putative source node is a member of the third generation (G = 3) of infection initiated by the index case.

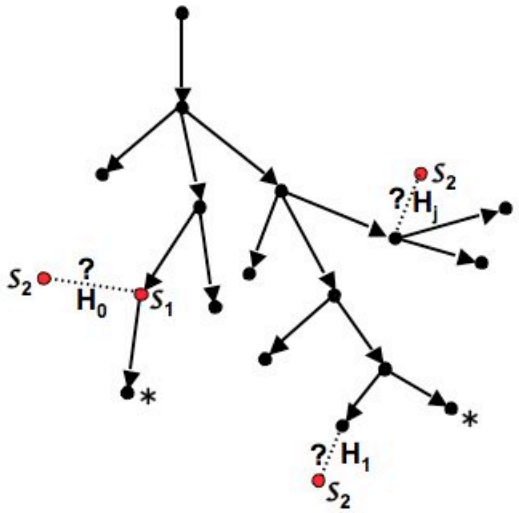


Figure 1. A notional disease transmission tree in an outbreak. Each node is an infected individual, and the "index case" is the uppermost node. S_1 and S_2 are sequences obtained from the victim and putative source nodes, marked in red. H_i represent different hypotheses about the source of the victim's infection. Asterisks mark two nodes separated by 7 transmission steps on this tree.

In general, it is highly unlikely that all of the nodes in the tree are known in any real outbreak, and even less likely that the interconnections between them are known. We may only have pathogen isolates from a few other nodes, whose relationship to S is unknown. Nonetheless, we can define H_0 to be the hypothesis that S is the source of the virus that infected V, and an alternative hypothesis that the source of the virus that infected V is a different node on the tree, not S. Inspection of the relationships illustrated in Figure 1 reveals that postulating that S is the source is equivalent to saying that there is only one transmission step (M = 1) between S and V. With all other possible source nodes, S_1 and S_2 are separated by more than one step. Thus, our hypothesis H_0 is equivalent to the hypothesis that M = 1, and the alternative hypothesis is M > 1.

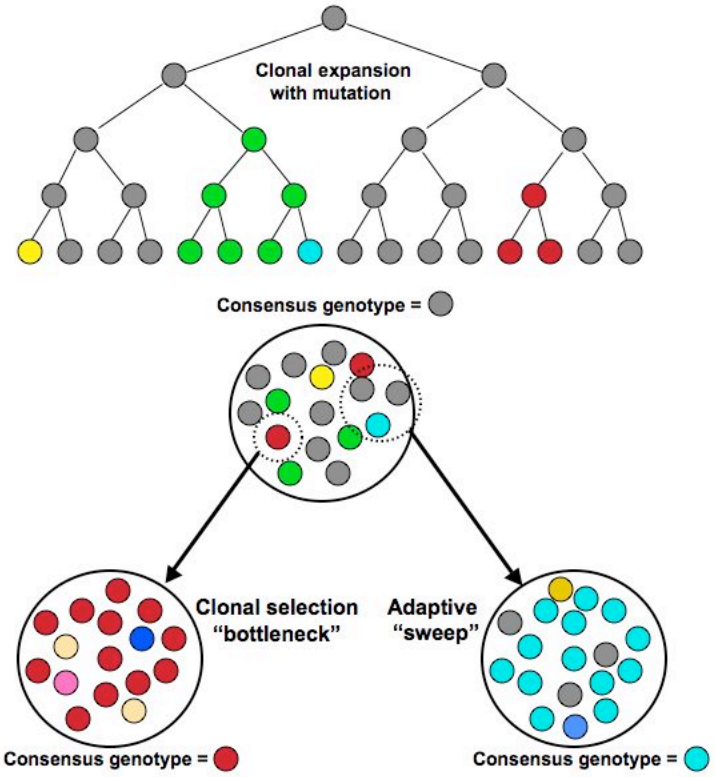


Figure 2. Basic genetic structure of a bacterial population during acute infection[3]. Each small circle represents a separate genomic sequence, with colors representing different genotypes[4]. Large circles represent isolated environments (e.g. a single infected host, colony, or fermentation vessel.) Clonal expansion with mutation leads to a mixed sub-population within that environment. Transfer of a portion of the subpopulation to a new environment (dotted circles, representing infective transmission to a new host or inoculation of a new culture vessel) can lead to a change in consensus genotype through statistical sampling ("bottlenecking") or selection (a "sweep".) The total population consists of all sub-populations including those in all infected hosts (human and non-human, active laboratory cultures, and laboratory isolates (in stasis) sampled from those hosts.

Each node-to-node transmission event in the transmission network leads to a potential difference between the sub-populations of genetic sequences contained within the parent and new nodes. We will assume that an isolate sampled from a node is sequenced, and a single representative sequence is used to characterize the isolate. In the examples discussed in this report, S_1 and S_2 are taken to be consensus sequences derived from each isolate, and *changes* to the consensus sequence from host to host constitute the primary

data of interest. However, with the advent of deep sequencing methods, it becomes possible to determine a most recent common ancestor (mrca) sequence from a set of clonal sequences from the intra-node pathogen population. The mrca sequence is generally a better representative of the population than the consensus sequence because it represents the population sequence close to the time of infection.

Figure 2 illustrates mechanisms that can change the representative sequence when a node is added to the network by disease transmission. Transmission of a pathogen involves transferring a small fraction of the pathogen population from one infected host to another, sometimes only a few organisms. Thus, node-to-node transfer can be thought of as a statistical sampling process. Acute infection primarily involves clonal expansion of the pathogen population within the new host. In this phase there are two mechanisms that can cause the consensus sequence in the new host to differ from that in the infecting host. Clonal expansion is accompanied by mutation, so sequence diversity increases as the population expands. Thus, sequences "sampled" from the infecting host in order to infect the new host can differ by chance from the consensus in that population. This phenomenon is sometimes referred to as a genetic "bottleneck"[2]. The probability of obtaining a different consensus sequence in the new host will depend on the pathogen population size and resulting diversity within the infecting host at the time of transmission. Adaptive sequence changes, for example in response to selective pressure from the new host's immune system, or tissue specific effects that modulate infectivity may also occur. Selection can shift the consensus even if the original "sample" of infecting pathogens is large enough to have essentially the same consensus sequence as the host population.

Long-term infection where the pathogen is maintained in the host with a relatively constant population size can bring into play additional mechanisms for changing the intra-host population of genetic sequences, and modifying the consensus sequence. First, Fischer-Wright (also known as "neutral") genetic drift may occur within the host as sequence lineages are removed from the population by chance[2,4]. Second, the population may change because of additional selective pressures (e.g. administration of anti-microbial therapeutics) that come into play in that particular host. A further complication to the theory can result when there are a high percentage of nodes that have been infected multiple times. In this circumstance, an infected node may support pathogen subpopulations from different origins that recombine genetically. For simplicity we will leave the treatment of such cases to a future report.

As the pathogen population propagates along the branches of the outbreak transmission tree, the process of genetic change is random and can be characterized by a distribution function that describes the probability of observing changes in the consensus (or mrca) sequence after M steps along a chain of infected nodes. The most general form for this distribution for a single step, M = 1, between any two nodes (here denoted 1 and 2 respectively), is:

$$P(S_1,S_2|M=1,\tau, t_1,t_2)$$

where S_1 and S_2 represent the consensus sequences of the microbial populations in each of the two nodes, τ represents the time between infection of node 1 and the transmission event between 1 and 2. The parameters t_1 and t_2 represent the time intervals between infection of each node and the time when isolates are obtained from each of them. (We assume that an isolate represents a sample of a node's population that is "frozen in time" with respect to the course of the infection. When isolates are derived from additional cell culture or animal passages, we consider the passaged samples new isolates associated with nodes that represent the populations of the pathogen in the culture vessel or animal host.)

We will postulate that inferences about the relationship between two nodes implied by sequences S_1 and S_2 are based on some quantitative comparison between S_1 and S_2. This quantitative comparison metric is some numerical function of the two sequences. In our derivation, we will assume that the comparison metric is a single scalar quantity, although there is no fundamental reason why it could not be multidimensional. We will denote the comparison metric by $\delta = \delta(S_1,S_2)$ and refer to δ as the "genetic distance" although it need not be a traditional genetic distance measure[4].

For simplicity, our inferential framework assumes, as do most other models of molecular evolution, that the random process is Markovian, and that the Markov process describing evolutionary change is time-reversible, which is also almost universally assumed in phylogenetic theory[14]. The assumption of reversibility has the effect of making the probability function depend only on the absolute number of steps that separate two nodes in the transmission tree. Thus, the transmission tree is regarded as an undirected graph with M computed as the number of edges connecting two nodes regardless of whether they are connected by a chain through intermediate nodes, or are descended from a common ancestor node.

In addition, two other random processes play a role in determining the probability of observing a particular δ value. These arise from the uncertainty in the times t_1 and t_2 that isolates are obtained from a node relative to the time the node is infected, and uncertainty in τ, the time that pathogen transmission occurs relative to the time the transmitting node was infected. These factors can affect the probability of observing a certain genetic difference between S_1 and S_2 because, the genetic diversity of the subpopulation of pathogens changes as the population size expands, and because of selective pressures and genetic drift during later stages of infection. To take these factors into account we can define probability distributions $P(t_1)$, $P(t_2)$, and $P(\tau)$, and average over them to obtain:

$$P(\delta|M) = \iiint P(\delta|M;t_1,t_2,\tau)\, P(t_1)\, P(t_2)\, P(\tau)\, dt_1 dt_2 d\tau \qquad (0)$$

This averaging accounts for the fact that the values of t_1, t_2 and τ are never known precisely. An explicit derivation of this equation is given in Appendix 2.

The transmission tree associated with an outbreak is also generated by a random process. Disease transmission depends on particular mechanisms (e.g. airborne transfer by droplets, or the oral-fecal route) that are mediated by various kinds of social contacts.

Each transmission tree generated in an actual outbreak can be thought of as a random sample from an ensemble of all possible outbreak trees that are consistent with the underlying mechanisms of transmission for that pathogen, and the underlying contact network for disease transmission. The probability $\mathcal{P}(M)$ that a pair of nodes drawn randomly from the tree will be related by M steps is defined on this ensemble of possible trees.

Consider an arbitrary sub-tree T drawn from the ensemble of outbreak trees {T} associated with outbreaks of the pathogen in question. Imagine that two nodes are chosen at random from this tree, the pathogen isolates from each node are sequenced and consensus or mrca sequences S_1 and S_2 are obtained, from which we calculate the value of $\delta(S_1, S_2)$. The joint probability of observing a particular δ value for a pair of nodes that are separated by M steps is given by:

$$P(\delta,M) = P(\delta \,|M)\bullet\mathcal{P}(M), \tag{1}$$

It must be noted that equation (1) implicitly assumes that the relationship between δ and M is independent of the particular tree, but is only a function of host-pathogen interactions and the host-host transmission mechanisms for the disease in question, and that every node and every transmission event in the tree is governed by the same probability distribution. Normalization clearly requires that $\sum_M \mathcal{P}(M) = 1$. The probability that two nodes are separated by more than M_0 steps is

$$\mathcal{P}(M > M_0) = \sum_J \mathcal{P}(J), \text{ where J runs from } M_0+1 \text{ to } \infty, \tag{2}$$

and the joint probability that two nodes exhibit a genetic difference δ and are separated by $M>M_0$ steps is

$$P(\delta,M > M_0) = \sum_J P(\delta|J)\mathcal{P}(J) \tag{3}$$

where J runs from $M = M_0+1$ to ∞.

Note that

$$\mathcal{P}(M \le M_0) = 1 - \mathcal{P}(M > M_0) \tag{4}$$

and

$$P(\delta,M \le M_0) = \sum_J P(\delta|J)\mathcal{P}(J) \tag{5}$$

where J runs from $M = 1$ to M_0

From equations (2) - (5) we can calculate the conditional probabilities

$$P(\delta|M > M_0) = P(\delta, M > M_0)/\mathcal{P}(M > M_0) \quad (6)$$

and

$$P(\delta\ |M \leq M_0) = P(\delta, M \leq M_0)/\mathcal{P}(M \leq M_0) \quad (7)$$

We can now use (6) and (7) and Bayes's theorem to calculate the probabilities that $M > M_0$ or $M \leq M_0$ *given* an observed δ value for isolates derived from the two nodes:

$$P(M>M_0|\delta) = P(\delta|M>M_0)\mathcal{P}(M>M_0)/[P(\delta|M>M_0)\mathcal{P}(M>M_0) + P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0)] \quad (8)$$

and

$$P(M\leq M_0|\delta) = P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0)/[P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0) + P(\delta|M>M_0)\mathcal{P}(M>M_0)]. \quad (9)$$

Equations (8) and (9) are the fundamental equations of the inferential framework offered in this report.

Referring to the previous discussion of Fig. 1, it is clear that equation (9) provides a weight-of-evidence expression relating the measured δ value for a pair of isolates to the probability that they were drawn from nodes related by a direct transmission event, i.e. H_0 is equivalent to setting $M_0 = 1$ in equation (9). (Strictly, the probability functions $P(\delta|M)$ and $\mathcal{P}(M)$ are not defined for the case $M = 0$ since they refer to two distinct nodes from the network, so the condition $M \leq 1$ is equivalent to $M = 1$.)

Equation (9) with the condition $M = 1$ can be re-written in the form:

$$P(M=1|s_1,s_2) = \left[\, 1 + \frac{P(s_1,s_2|M>1)}{P(s_1,s_2|M=1)} \times \frac{\mathcal{P}(M>1)}{\mathcal{P}(M=1)} \,\right]^{-1} \quad (10)$$

where we have made explicit the dependence on S_1 and S_2, the sequences determined for the "victim" and "suspect" nodes. This form is analogous to the equation used to determine the probability of paternity or other familial relations in human DNA forensics.

The fundamental expressions (8) and (9) are the formal basis for calculating probabilities within the framework for microbial genetic inference presented in this report. It is easy to see that other types of hypothesis tests can also be defined *mutis mutandis* within this framework. For example, the distribution $P(M=M_0|S_1,S_2)$, and its complement $P(M\neq M_0|S_1,S_2)$ where M_0 is an arbitrary number have utility for certain kinds of forensic

cases where entire transmission chains must be reconstructed. Regardless of the precise form of the hypothesis test, calculations of the posterior probability depend, through equations (3) – (7) on the sampling distributions P(δ|M) and $\mathcal{P}$(M).

Applying the formalism

To understand how equations 8 and 9 can be applied by an investigator in a case of biological terrorism or criminal activity, consider the following scenario:

There is an outbreak of SARS in a U.S. city, and phylogenetic analysis indicates that the sequence of the agent is "closely related to" published SARS CoV sequences from the Beijing outbreak of 2003. Published reports have concluded that the Beijing outbreak strains in question were derived from the SARS outbreak in Guangzhou. How does the investigator calculate the probability that the agent was derived from an isolate collected from the Beijing outbreak versus the Guangzhou outbreak?

In this scenario, the investigator has one or more isolates of SARS CoV from the U.S. outbreak, and a set of published (in this case, consensus) sequences associated with the historical SARS outbreak, including "representative" sequences from Beijing and Guangzhou. In addition, he can access published information on SARS transmission network topology and size (number of infected people) from the extensive literature on SARS epidemiology. From the existing historical sequence data, he can derive an estimate of P(δ|M), using methods explained below. The Beijing and Guangzhou outbreaks can be treated as statistically independent sub-trees of the worldwide SARS epidemic, whose size can be estimated from the number of recorded infections in each geographic region (corrected for the fraction of infections that go un-recorded, also known from epidemiology data.) Under this assumption, $\mathcal{P}$(M) can be estimated for each outbreak, using methods to be discussed shortly. From P(M) for the Beijing and Guangzhou outbreaks, the investigator can also determine a key parameter, the *diameter* of each sub-network, D_B and D_G. The network diameter is basically the value of M beyond which P(M) is negligible.

The investigator then computes δ between the U.S. outbreak sequences and those from the Beijing and Guangzhou outbreaks, which we will call δ_B and δ_G respectively. From P(δ|M) and $\mathcal{P}$(M), the investigator can then compute $P(M \le D_B|\delta_B)$ and $P(M \le D_G|\delta_G)$ from equation (9). $P(M \le D_B|\delta_B)$, for example, can be interpreted as the *probability that the U.S. sequence and the Beijing sequence both originated from within the Beijing outbreak*.

Note that the investigator can do this kind of comparison for each reference sequence he possesses from the Beijing and Guangzhou outbreaks, and the results will vary, depending on the relative location within the transmission network of the person from whom the isolate was obtained. Some cases may lead to ambiguous results, if an isolate came from a node very close to the node responsible for the transmission of SARS CoV between Guangzhou and Beijing.

Knowing that it is more probable that the U.S. strain originated from a node in the Beijing transmission tree would help focus investigative resources on obtaining additional isolates (or sequence data) from that outbreak (which might be held in laboratories in any part of the world.) The provenance can be further narrowed by computing $P(M \leq M_0|\delta_B)$ with $M_0 < D_B$. Ultimately, the investigator can state that the U.S. strain is less than M transmission steps away from a particular isolate obtained from the Beijing outbreak, with an explicitly calculated probabilty. For the SARS epidemic of 2003, if M is small (2 or 3) it will very often imply that the source isolate was obtained from a patient treated at a certain hospital, which might provide an additional clue for attribution.

Estimating P(δ|M)

In this section we examine some properties of the sampling distribution P(δ|M), and methods for determining it from sequence data determined from pairs of nodes with known epidemiological relationships. For any pair of nodes separated by M transmission steps in a completely connected outbreak tree, the observed δ value is a sample from its parent distribution. Unfortunately, a direct approach to determining P(δ|M) by random (or exhaustive) sampling of many nodes in the tree cannot generally be applied unless we have large, complete and accurate transmission trees so that the relationship between all the nodes is known. Instead, the data available to us in most cases is fragmentary, with some degree of uncertainty about the true transmission relationships among the samples. However, while P(δ|M) for arbitrary M is currently difficult to deduce accurately by purely empirical means, we can provide reasonable representations of P(δ|M=1) and its complement P(δ|M>1) using empirical data from three sources:

(1) The Singapore SARS outbreak of 2003, where whole genome sequences of SARS coronavirus isolates were obtained from an epidemiologically linked set of 12 patients[15,16].

(2) The 2001 FMDV outbreak in Great Britain, where whole genome sequences of FMDV virus were obtained from an epidemiologically linked set of 20 farms[17].

(3) A study by Trask, et. al. which contains partial genome sequences from a set of 63 pairs of HIV infected sexual partners who were married couples[18].

For purposes of this illustration, we will define δ to be the number of single nucleotide substitution differences observed between a pair of sequences, denoted k. For SARS and FMDV, each set of whole genome sequences was aligned and all substitution differences between each pairs were counted. For HIV, the *env* gene sequences were codon-aligned to the HXB 2 reference sequence[19], regions disrupted by multiple indels were excised, and only synonymous substitutions were scored. In the following discussion it is important to keep in mind that different definitions of δ can change the shape of P(δ|M).

Figures 3a,c and e show empirically derived histograms of k observed for pairs of sequences derived from isolates obtained from pairs of nodes related by M = 1 and M>1. These histograms are estimators for P(k|M=1) and P(k|M>1) respectively. Next to each

distribution figure we have plotted (Figs. 3b,d, and f) the Receiver Operating Characteristic (ROC) curve derived from the corresponding distributions. There are several general observations that can be made about these empirical distributions. First, the smaller the sample size, the "noisier" the empirical histograms, and the more uncertain our estimates of the distributions. Second, the larger the network sampled, the more separation there is between P(k|M=1) and P(k|M>1), and the sharper the ROC curve generated by the data. The empirical ROC curve generated this way provides a valid estimator for the likelihood ratio P(k|M=1)/P(k|M>1) in equation (10). The value of A quoted on each figure refers to the value of the likelihood ratio at zero false positive rate, which is a measure of how concave the ROC curve is. It is important to note, however, that these ROC curves only apply to transmission networks of the same size as the one used to generate the data.

Various types of error affect the distribution estimates in Figure 3. First, the putative transmission trees determined by contact tracing for the SARS and FMDV outbreaks probably contain errors. Thus, some of the identified M=1 pairs are actually M>1, and vice versa. Such errors always increase the overlap between the estimated P(δ|M=1) and P(δ|M>1), as illustrated in Figure 4. In outbreaks like SARS and FMDV, transmission involves the shedding of large numbers of viruses into the environment, and contact tracing can not always identify the true source of an infection. There can be unknown asymptomatic sub-clinical spreaders, and patient recollection may be faulty. In the case of HIV it is well known that sexual partners outside of established relationships are not always truthfully disclosed. Thus, the accuracy of empirical distributions of P(δ|M) determined from field data may be compromised.

Second, the empirically determined distributions are affected by the presence of errors in the sequence data or the sequence alignments. In most cases the sequence error rates are not determined or reported, so at least a few of the observed differences between genomes or genomic regions may be in error. Moreover, regions containing indels often lead to ambiguous local alignments that are inaccurately scored as substitutions. A third potential source of error occurs when infection density is high enough that a significant fraction of nodes are infected by two or more independent transmission events. Recombination can then introduce errors in the δ value, which is predicated on the clonal transmission model.

Third, random errors can be effectively introduced by variations in the time between infection and transmission, and between infection and the time isolates are drawn (i.e. the parameters t_1, t_2, and τ in equation (0).) These errors are expected to be of lowest concern for acute diseases like SARS or FMDV, where the distributions $P(t_1)$, $P(t_2)$ and $P(\tau)$ are very narrow relative to the timescale of genetic change. However, for long-term diseases such as HIV, these effects are very significant. When approximate times for infection and sampling are known, it is possible that methods for correcting the δ data can be developed. However, an accurate treatment of HIV transmission inferences when long time delays between sampling and infection are present awaits further research.

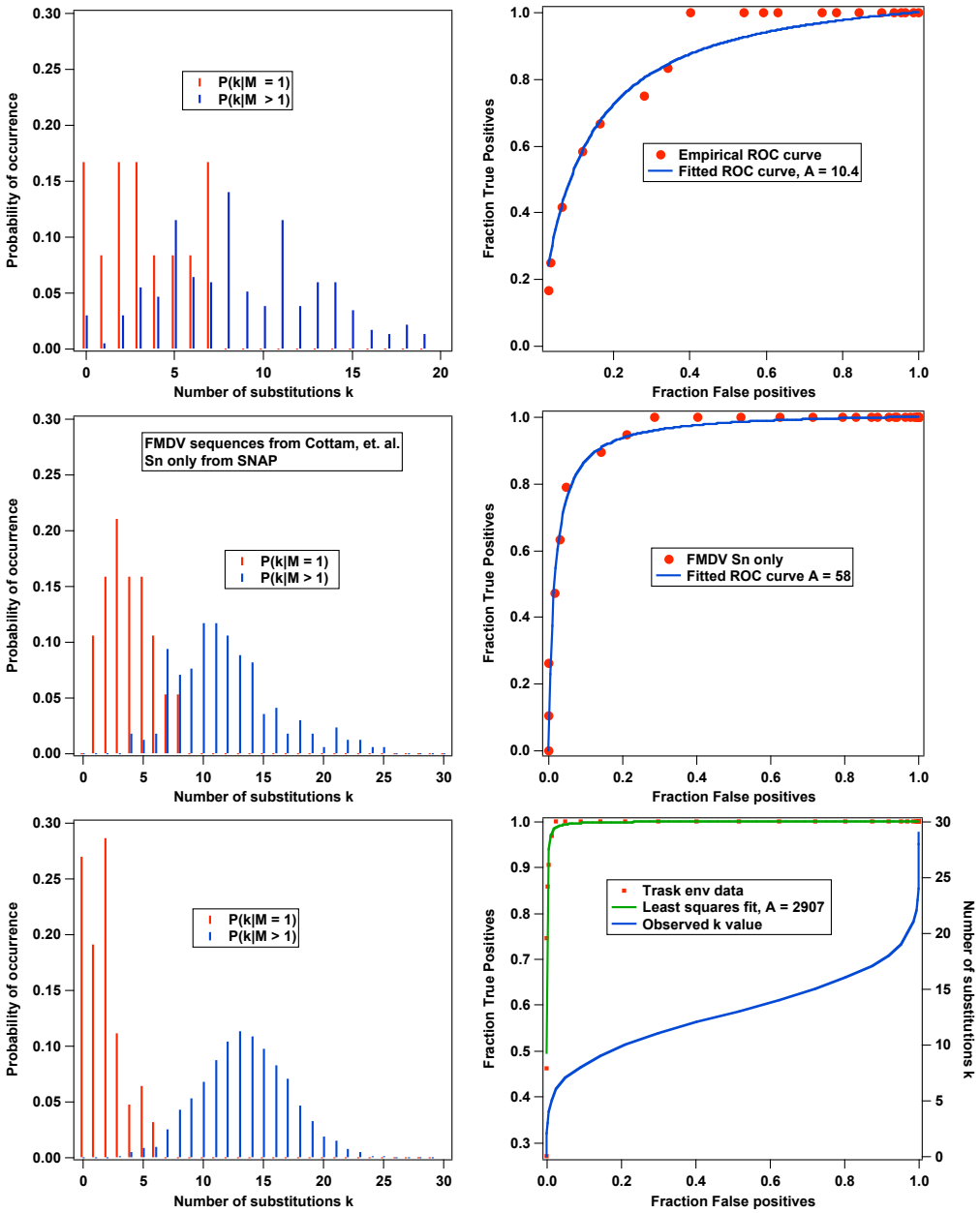


Figure 3. a, c, and e: Empirical distributions of the number of substitutions (k) observed between pairs of isolates related by a single transmission step (M=1). All substitutions over the entire genome were scored for SARS and FMDV; the HIV data was obtained using partial sequence data from the indicated gene regions. Figs b, d, and f: ROC curves based on the empirical distributions.

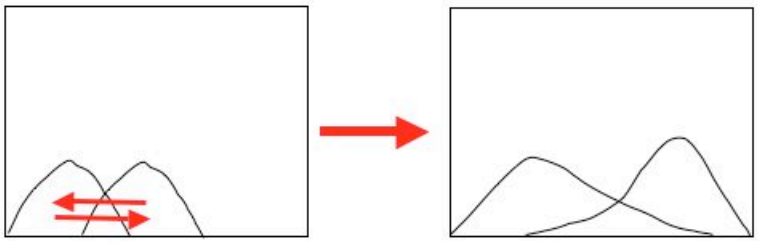


Figure 4. Epidemiological errors exchange δ values belonging to the P(δ|M=1) and P(δ|M>1) distributions, causing increased overlap.

As explained above, lack of complete accurate transmission network data generally precludes the direct empirical approach to determining P(δ|M) for arbitrary M. In such cases a more practical approach is to use statistical methods to estimate the parameters of some theory-based functional form that is then used to perform calculations. For example, one could fit simple functional forms to the M=1 data that automatically imply the behavior of the distribution for larger M values. In exploring this possibility we have performed maximum likelihood fits to Poisson and Negative Binomial distributions for the P(k|M=1) data shown in Figs 3a,c, and e. The Poisson form of the general distribution is then given by:

$$P(k|M) = (\gamma M)^k/k!\ e^{-\gamma M} \qquad (11)$$

The values of the γ parameters and their uncertainties derived from the fitting procedure[20] are given in Table 1. The form and fitting results for the Negative Binomial are given in Appendix 3. We will simply note here that the value of the parameter that characterizes the negative binomial dispersion implies that the Poisson distribution (with one fewer parameter) is as good a fit to these data. Although these well-known distributions have great appeal for describing substitution data in other contexts[21], statistical tests reveal that they do not actually fit the data in Figure 3 very well, as can be seen from the simple χ^2 test results shown in Table 1.

Table 1. Parameters and χ^2 results for fitting empirical P(k|M=1) data to a Poisson distribution.

Data	γ (per transmission step)	X^2	# points	Support for null hypothesis*
SARS	4.14 ± 0.77	15.3	7	Moderate against
FMDV	4.29 ± 0.50	3.91	17	No evidence against
HIV env	1.84 ± 0.16	15.5	70	Very strong against
HIV gag	2.04 ± 0.28	877	36	Very strong against

*Null hypothesis is that deviations from Poisson are due to sampling error.

If we possessed reasonable theoretical forms for P(δ|M) then we could generalize the fitting procedure to include all the genetic data collected from a set of linked nodes. A Bayesian or Maximum likelihood procedure that takes into account both missing links and uncertainty in the accuracy of the links could be used to estimate the parameters in these distributions[22]. A more general approach that is parameter free is based on the observation that if δ is a random variable distributed as P(δ|M=1) for a single transmission step, and each transmission event represents an independent sampling of the genomic distribution in the transmitting host, then δ after M transmission steps is distributed as the sum of M independent random variables each independently distributed as P(δ|M=1). Thus for larger M we may write:

$$P(\delta|M=M_0>1) = P(\delta|M=1) \otimes P(\delta|M=1) \otimes P(\delta|M=1) \otimes \ldots\ldots P(\delta|M=1) \qquad (11)$$

Where the right hand side of equation (11) is the M_0-fold auto-convolution of P(δ|M=1). Hence, it is only necessary to obtain P(δ|M=1) in order to estimate P(δ|M) for larger values of M. This convolution property is directly reflected in the function forms of the Poisson and Negative binomial distributions by the fact that the rate parameters scale as γ•M for arbitrary M, where γ is the observed substitution rate for M=1.

A flexible, but computationally intensive parametric approach that can include both selection and drift is to develop an explicit simulator for molecular evolution and propagation on an outbreak tree. The simulator would couple replication with errors and sampling of the resulting genetic population to initiate infection of a new node and subsequent expansion of a new sub-population. One approach to performing such calculations involves using the "quasispecies" propagation equations[23-26], which are a simple set of equations governing the growth and diversification of a population of genetic sequences by replication and mutation. (It is important to distinguish the use of these equations to describe the transient dynamics of viral populations over finite timescales from the use of the infinite time "equilibrium" solution of the equations to describe the genetic structure of viral populations[24]. We do <u>not</u> advocate the latter approach.)

Figure 5 illustrates some results from a model calculation along these lines, described in more detail in Appendix 4. The three simulations were performed with identical transition rate parameters, but very different fitness surfaces. A striking result is the shift of P(δ|M=1) to larger mean values of δ as the fitness surface becomes increasingly "neutral" (The model fitness surfaces are shown in Appendix 4). This dependence of genetic drift rate on the size of the "neutral space" occupied by the sequences was first recognized by Jenkins et.al., who pointed out that this effect is completely lost if the "equilibrium" quasispecies distribution function is used to describe microbial populations[24].

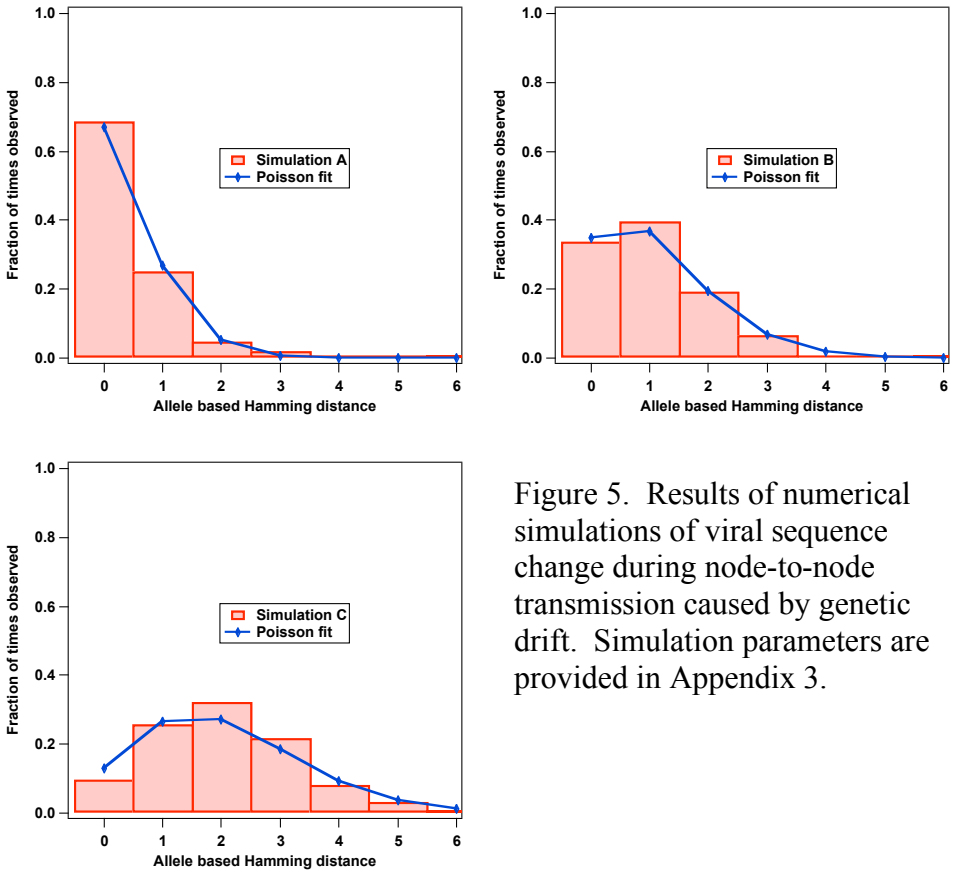


Figure 5. Results of numerical simulations of viral sequence change during node-to-node transmission caused by genetic drift. Simulation parameters are provided in Appendix 3.

The results in Fig. 5 fit reasonably well to Poisson distributions, but other simulations that use different combinations of transition probabilities and fitness surfaces do not. In addition, it is not known what influence the limited number of loci and the somewhat arbitrary nature of the chosen genetic distance metric may have had on the form of these distributions. Neither the potential nor the limitations of this type of simulation has been fully explored at this point. Nonetheless, from a comparison of Fig. 5 with actual data shown in Fig. 3 it appears plausible that such simulations can capture the underlying phenomena in a compelling way[27]. Some of the realistic features of viral evolution that can be incorporated into such simulations are:

(1) Adaptive (non-neutrality) effects (through modifications of the fitness surface.)

(2) Sampling time effects, t_1,t_2, and τ (incorporated through variation in N_{gen}.)

(3) Selective transmission of particular genotypes (by superimposing additional filtering on the basic statistical sampling procedure.)

(4) Multiple infecting virions (sampling two or more genotypes from the initial population)

Currently, our best practical approach to determining P(δ|M) is to fit empirical histogram estimates of P(δ|M=1) to Poisson-like distributions, or to use the convolution based approach. Within the next few years it should be possible to develop statistical models for P(δ|M) that improve upon the simple Poisson or Negative Binomial forms. To assist this effort, it is reasonable to suggest that better experimental data could be available for viruses within the next few years as well. The "deep" sequencing permitted by modern high throughput sequencers has already been applied to elucidating the distribution of genotypes within host virus populations[28]. Thus, we can expect gradual convergence of models and experimental data.

Estimating $\mathcal{P}(M)$

The sampling distribution $\mathcal{P}(M)$ is a function of the size and topology of the transmission network that is relevant to the source attribution question at hand. Here too we can obtain a sense of the form of these distributions by considering data from real outbreaks. As mentioned above, a transmission network determined by contact tracing is considered a random sample from an ensemble of possible trees whose statistical properties reflect the epidemiology of that particular disease. When referring to an estimate of $\mathcal{P}(M)$ obtained from a particular empirical tree T, we will write $\mathcal{P}(M|T)$.

Figure 6 shows $\mathcal{P}(M|T)$ for several geographically or temporally separate SARS outbreaks. Some properties of these outbreaks are provided in Table 2. These local outbreaks were part of the larger global outbreak, and each transmission network represents a sub-tree extracted from the complete SARS transmission network (large parts of which are unknown) by contact tracing[29-31]. In order to calculate $\mathcal{P}(M|T)$ the adjacency matrix of each tree was constructed[13]. Then the number of paths of length M among the set of nodes was determined by using a result from graph theory that relates this quantity to the number of unit matrix elements in successive powers of the adjacency matrix.

Table 2. Properties of some known sub-trees of the complete SARS transmission tree.

Outbreak	Number of nodes	Number of generations	Number of "superspreaders"	Largest superspreadng cluster size
TTSH1[29] (Singapore)	41	4	1	22
TTSH2[29] (Singapore)	36	3	1	21
Toronto[30]	72	5	3	16
Beijing[31]	69	3	4	33

Note that each tree has a different number of nodes N and spans a different number of generations G. Besides these two parameters, the detailed form of $\mathcal{P}(M|T)$ also depends, for example, on the number of "superpreaders" (patients who infect more than 5 other patients), and the size of the superspreading clusters.

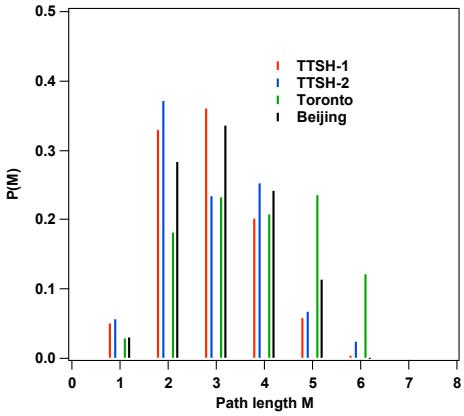


Figure 6. Path length distribution for four transmission trees sampled from the 2003 SARS outbreak.

Since each of the four trees in Table 2 can be considered a random sample from an ensemble of possible trees, the observed variations in $\mathcal{P}(M|T)$ from tree to tree can be considered sampling errors about some average, or most likely, distribution that characterizes SARS outbreaks. Rigorously, since both N and G vary, the appropriate ensemble that contains all four of these trees would be one with no restrictions on number of nodes or the number of generations. Therefore the average of the distributions in Fig. 6 would actually be a *biased* estimate of the ensemble average, since we deliberately restricted the sample set to four trees of a convenient size for manual construction of the adjacency matrix, but not too small to be uninteresting. On the other hand, each tree in Figure 6 could separately be used as an estimate of $\mathcal{P}(M)$ for the ensemble of SARS outbreak trees with the same number of nodes or generations, (or both.)

In most cases, an investigator will not have the actual transmission tree that is relevant to the investigation. However, in many cases, the size of the relevant tree will be known with reasonable accuracy, even though the majority of the nodes and their connections are not known. There are then three possible approaches to deriving $\mathcal{P}(M)$ in any particular case:

(1) Use an empirically determined tree from another outbreak of the same disease where contact tracing has been carried out, and whose size is similar to the one of interest, then use the same procedure used to generate the data shown in Fig. 6. This, of course, can introduce tree "sampling error" as discussed above. However, there is some evidence that the larger the tree, the smaller such errors are. In any case, we will show later that calculations of $P(M=1|\delta)$ using equation (9) are not very sensitive to such errors. It is also possible that a tree from an outbreak of a *different* pathogen could be used as a surrogate if the transmission mechanisms are the same, although this remains to be proven.

(2) Use an analytical functional form for $\mathcal{P}(M)$ whose parameters have been fit to data from prior outbreaks of the same or similar disease. As we will show below, several simple models appear to provide close approximations to empirical data. Choosing the

appropriate model depends on practical considerations. For instance, if the relevant size (expressed as a number of nodes N and/or the number of generations G) of the tree of interest in a source attribution problem is known or can be estimated, the ensemble of interest is one in which N, G or N&G is fixed. In this case we might denote the distribution as $\mathcal{P}(M|N)$, $\mathcal{P}(M|G)$, or P(M|N&G) respectively.

(3) Derive $\mathcal{P}(M)$ from simulations of outbreaks on a social contact network that has been developed for epidemiological prediction purposes (for that disease).

Approach (1) raises no special technical issues, and is a straightforward extension of the network data analysis leading to Figure 6. Approaches (2) and (3) are currently being investigated, and we will now discuss some technical aspects of these last two methods in turn.

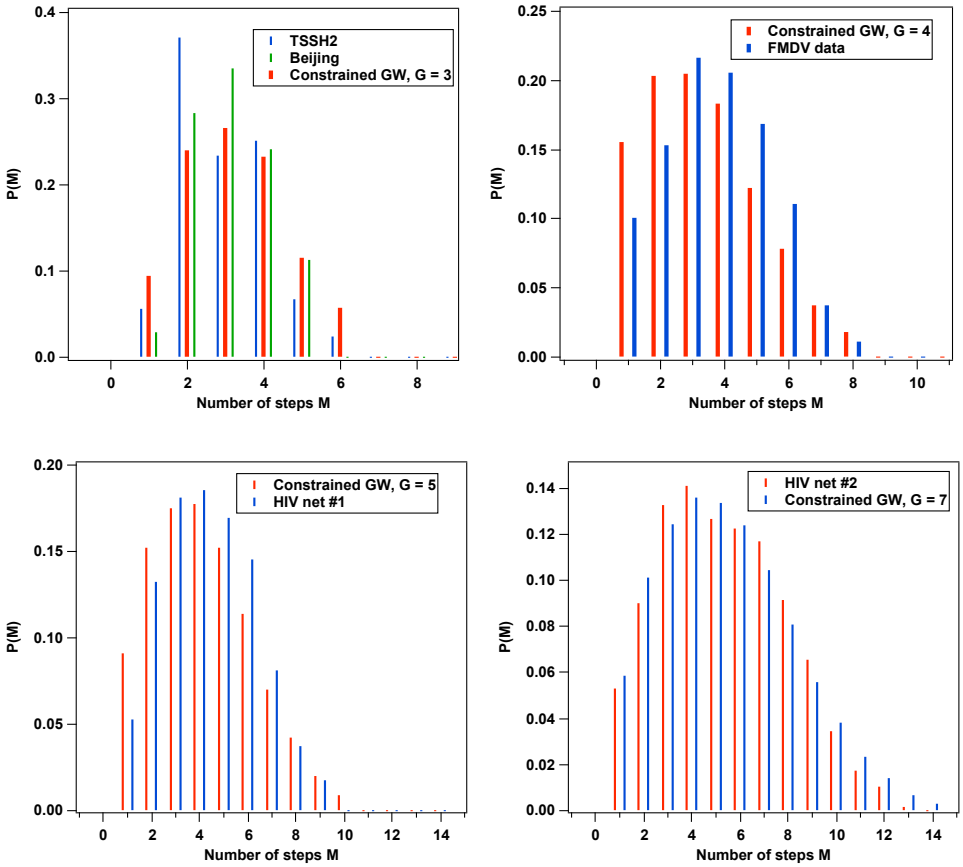


Figure 7. Empirical and calculated $\mathcal{P}(M)$ for various outbreaks using the truncated Galton-Watson branching process model.

Method (2)

A very simple model for a disease transmission tree utilizes the Galton-Watson random branching process, in which each node is assigned a certain probability P(d) for generating d successor (daughter) nodes[32]. The parameters in the model are the probability function $\mathbf{P}(d)$ and either a maximum number of generations over which the

tree is allowed to grow, or a fixed number of nodes. In appendix 5 an analytical expression for $\mathcal{P}(M|G)$ for a general truncated Galton-Watson process is derived. It is important to note that this probability distribution describes the statistical properties of an ensemble of G-constrained trees each of which has a different number of nodes. Using the results of Appendix 5, we have calculated $\mathcal{P}(M|G)$ and compare the results to data from various outbreaks in Figure 7. This includes two of the SARS outbreaks from Table 2, data from the 2003 FMDV outbreak in Great Britain[17], and two HIV transmission trees[33]. The computed distributions are <u>not</u> least-squares fits to the empirical data. Instead, we simply estimated $\mathbf{P}(d)$ from the data, and made small adjustments to align the center the distribution with the data visually. It is clear that this model captures the basic shape of the empirical data and its variation with the number of generations (Note the increasing spread to larger M values as the value of G increases.)

It should be possible to incorporate an explicit fitting algorithm to determine the probability parameters for this model directly from the network adjacency matrix. In addition, the predictive power of the model might be improved by at least one other modification: Real outbreaks are typically terminated through the gradual implementation of infection control measures. This may be modeled by changing $\mathbf{P}(d)$ from generation to generation, until $\mathbf{P}(0) = 1$ for the last generation. For example, during the SARS outbreak, implementation of patient isolation in hospitals essentially eliminated the incidence of superspreading events ($d > 5$) involving health care workers by the second or third generation[29-31].

The Galton-Watson process does not yield simple analytical solutions when the number of nodes is fixed rather than the number of generations. However, an analytical approximation to the path length distribution function for trees defined by a degree distribution $\mathbf{P}(d)$ and a fixed number of nodes has been derived by Fronczak[33]. Figure 8 compares the Fronczak distribution with two of the empirical SARS distributions. As in the case of the Galton-Watson model, we did not perform a least squares fit, but simply adjusted the $\mathbf{P}(d)$ values slightly to place the distribution maximum near that of the data.

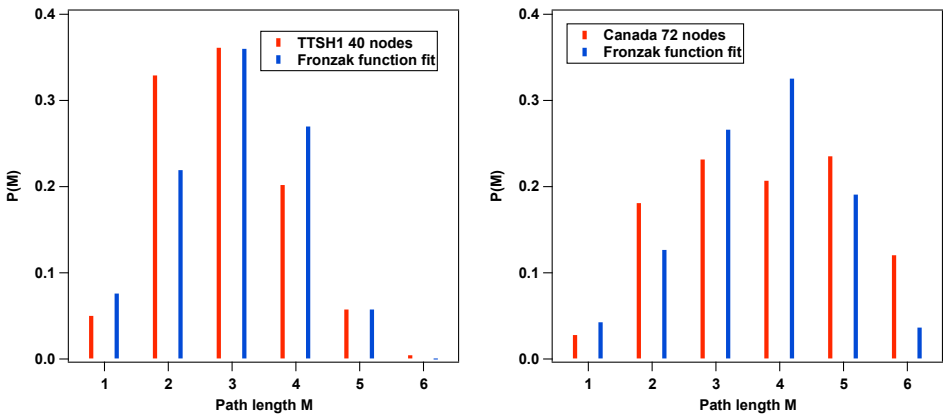


Figure 8. Comparison of the Fronczak distribution with SARS outbreak data.

Other factors besides the size and degree distribution are known to affect the statistical properties of transmission trees[34,35]. These include clustering in the underlying social network, assortive mixing when there are several distinguishable types of nodes, degree correlations between nodes (i.e correlations between the number of infectees generated by a node and the number of infectees generated by its parent node) and community structure (where the social network consists of two or more tightly connected subpopulations linked by a relatively sparser set of contacts.) More elaborate models and functional forms would be required to incorporate these effects.

Method (3)

The most general approach to estimating $\mathcal{P}(M)$ would be to utilize computer simulations of disease transmission on large social networks, as illustrated schematically in Figure 9. A number of elaborate social network models have been constructed to investigate outbreak dynamics and the effect of control measures for many communicable diseases, including zoonotics in networks of animal hosts[10,11,36-46]. Thus, a variety of ready-made models are already available. Social networks are relatively stable in time and can easily be stored as reference data. Moreover, it is often easier to consider collecting field data about the underlying social net, or take advantage of field studies funded through basic epidemiological science programs, than it is to directly gather contact tracing data from an outbreak. Thus, this approach, while as yet unexplored, holds considerable promise as an operational way to determine $\mathcal{P}(M)$ for an outbreak relevant to a forensics case.

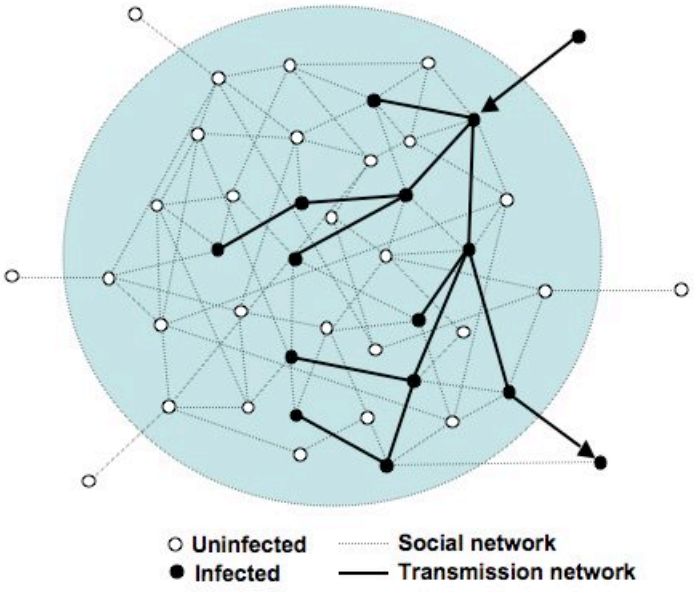


Figure 9. A disease transmission network induced on an underlying social net.

Some general observations about the proposed inferential framework

The likelihood estimator given by equation (9) and the underlying probability distribution functions possess certain compelling general properties in the context of microbial

forensics. These include the ability to quantify the significance of a sequence "*match*" or "*mismatch*" between two isolates; the ability to capture non-intuitive effects of network structure on inferential power, including the "*small world*" effect; the insensitivity of inferences to uncertainties in $\mathcal{P}(M)$; and the concept of *rescaling*, i.e. ability to collapse sub-networks into single nodes and examine transmission inferences on the rescaled network. In this section we will elaborate on each of these properties in turn. In the last part of this section we will discuss the important issue of choosing a comparison metric δ.

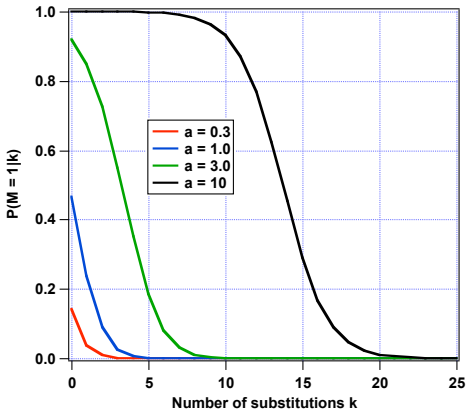


Figure 10. Example calculations of P(M=1|k) vs k. The parameter a is equivalent to γ in equation (11).

Genetic Matching

Figure 10 shows the dependence of P(M=1|δ=k) calculated using the Poisson model equation (11) and a truncated Galton-Watson model for $\mathcal{P}(M)$. For the latter model we have used the same parameters that were used in the FMDV simulation shown in Fig. 7. Figure 10 demonstrates an important feature of microbial genetic inference. First, note that an exact match (δ = 0) does not imply certainty that two isolates are related by direct transmission (i.e. P(M=1|δ = 0) ≠ 1 in general.) Similarly, large mismatch of the consensus sequences (δ ≠ 0) may still imply a high probability that the isolates are related by direct transmission, if the mutation rate is high enough. Clearly, the magnitude of the mutation rate γ is a critical factor for determining the inferential power of such calculations, and the highest power is always achieved when all possible mutations are scored over the entire genome.

Small world effects

Figure 11 shows how the distributions P(M) derived from the Galton-Watson (constrained G) and Fronczak (constrained N) models depend on the number of generations and number of nodes respectively. A striking property is the rather slow dependence of both distributions on network size. This can be understood from some basic properties of random networks.

In network theory a geodesic path is defined to be the shortest path through the network from one node to another[34]. In tree-like networks there is only one geodesic connecting

any two nodes. The diameter of a network is defined to be the length (in number of edges) of the longest geodesic path between any two nodes. In a transmission tree, it is easy to see that the maximum possible geodesic length is 2G, where G is the number of generations spanned by the tree. Thus, in the ensemble of trees defined by a certain number of generations, $\mathcal{P}(M) = 0$ for $M > 2G$. For a large class of infectious viral diseases, typical outbreaks terminate in a relatively few generations, and this effectively confines $\mathcal{P}(M)$ to be non-zero only at relatively small values of M.

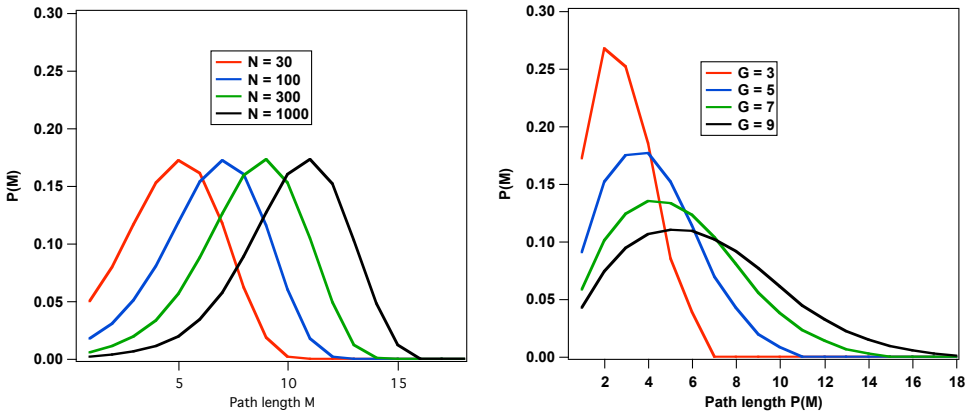


Figure 11. Variation of P(M) with increasing network size. (a) Fronczak model; (b)Truncated Galton-Watson model.

Trees drawn from an ensemble with a fixed number of nodes, but variable numbers of generations, also exhibit effective bounds on their diameters. It is well known that random networks defined by a degree distribution P(d) such that $\langle d \rangle > 1$ exhibit the *small world* scaling property, defined by the relation $D \approx \ln(N)/\ln(\langle d \rangle)$ where N is the number of nodes. This simply arises from the approximate exponential scaling of the number of nodes with the number of generations[34].

Many infectious disease transmission networks exhibit both small world behavior and superspreader clusters[48]. Increasing the number of high d clusters within a transmission network with a fixed number of nodes reduces the probability of observing pairs of nodes connected by a large number of steps. In fact, when the degree distribution of a network has significant probability of having nodes with very large numbers of links, (these are sometimes described as "scale free" networks) the diameter exhibits even slower scaling with node number. For example, Bollobas and Riordan[49] showed that $D \approx \ln(\ln(N))$ for such scale-free networks.

The small world and superspreader effects are important factors in genetic inference because they lead to a high likelihood that two nodes randomly drawn from the tree will have M which is small compared to the maximum possible diameter of the tree. In other words the prior probability that two randomly selected nodes are related by a small number of transmission steps is much higher than one might intuitively believe for a large transmission tree.

Insensitivity to variation in $\mathcal{P}(M)$

Figure 12 shows how P(M=1|k) varies with the number of nodes in the transmission network. For these calculations the Poisson parameter $\gamma = 3$, and $\mathcal{P}(M)$ was calculated using the Froczak model for N = 30, 100, 300, and 1000. The small variation of P(M=1|k) indicates a remarkable insensitivity of the posterior probability to the change in $\mathcal{P}(M)$ as the size of the network increases.

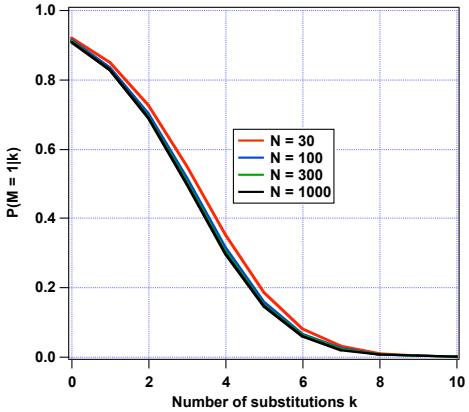


Figure 12. Variation of P(M=1|k) as a function of network size N.

Similar results are obtained when the Galton-Watson distribution is used and G is varied. The apparent lack of sensitivity of the posterior probability distribution P(M=1|k) to the size of the transmission network can be understood by reference to the explicit form that it takes under the assumption of the Poisson form for P(k|M) as shown in equation (12), and the properties of the Galton-Watson and Fronczak forms for $\mathcal{P}(M)$.

$$P(M=1|k) = [\,\sum_M M^k\, e^{-a(M-1)}\, (\mathcal{P}(M)/\mathcal{P}(1))]^{-1} \qquad (12)$$

Regardless of k, the pre-factor $M^k\, e^{-a(M-1)}$ is very small for M >3, while the ratio P(M)/P(1) remains relatively constant for $M \leq 3$ when $N \leq 3000$, making (12) nearly independent of N over the range examined. Functional forms for P(δ|M) that shift probability to higher values of M will increase the sensitivity to changes in $\mathcal{P}(M)$. Nonetheless, it seems plausible that uncertainty in the number N or G will not severely compromise the inferential power of the method in general.

Rescaling – networks of outbreaks

Networks of disease transmission often extend over large spatial regions and have long durations. In such networks, sub-networks of infected individuals within cities, herds, flocks, and other social groupings can often be considered the infected "nodes" in a larger scale network. Each node defined this way is itself a transmission networks connecting individuals, but the intra-node structure is effectively ignored at this scale. This type of re-scaling makes sense when these natural groupings are less well connected than the individuals making up the social groups that define the re-scaled nodes. The connectivity of the re-scaled network may be very different from that of the underlying network of

individuals, so $\mathcal{P}(M)$ will differ as well. It is important to keep in mind that more than one of these rescaled nodes may be contained within a particular city, herd or flock, because several index cases within those social groupings may have been independently infected from distinct sources. Similarly, a node in a re-scaled network may span several geographically distinct regions, if the underlying sub-network does.

As was the case for networks of individuals, transmission hypotheses can be formulated and tested on this type of rescaled network as well. It is necessary to estimate the consensus sequence for the population of pathogens contained in the sub-network of individuals that constitute a re-scaled node. Figure 13 illustrates the case where an individual victim

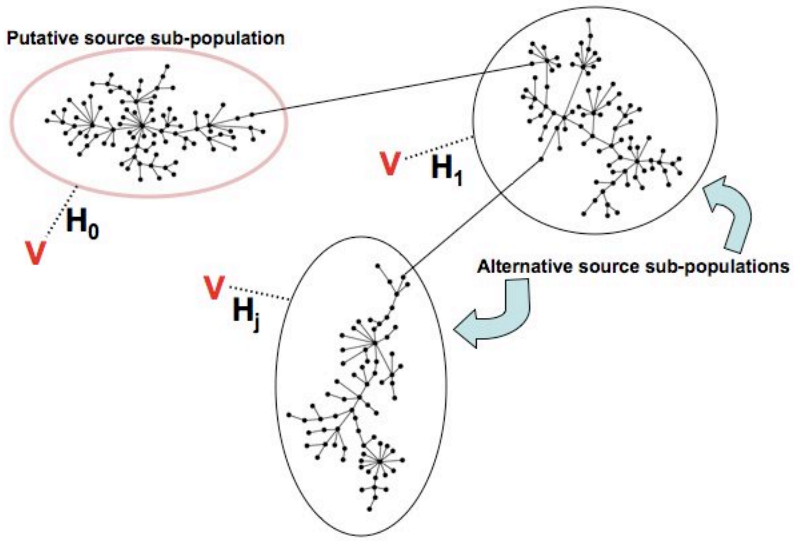


Figure 13. Rescaling of a transmission network.

has been infected with a pathogen associated with an extended outbreak, and there is a question about which city, herd, etc. was the source of the infecting strain. Similar comparisons can be made when the "victim" is a complex entity itself, i.e. another sub-network.

Choosing δ

In the previous sections we have not discussed any specific definitions of the comparison metric δ other than the very simple illustrative choice of the number of substitutions k, or the similarly defined metric used in the simulations of Fig. 5. The choice of δ is, of course, a critical determinant of the shape of P(δ|M) and the value of the posterior probability P(M=1|δ).

A variety of sophisticated metrics can be defined that take into account not only substitutions, but also insertions and deletions and other types of genetic change. Some of these are listed in Table 3. One basic guideline for choosing δ is that the metric should be sensitive to changes across the entire genome to optimize the ability of equation 9 to

resolve sequence differences. Whole genome sequencing will also permit unambiguous identification of recombination and re-assortment events, which can confound simplistic distance measures. Another, related guideline is that genetic differences be scored according to a realistic "biological" model of genetic change. For example, a deletion of n adjacent nucleotides is not simply equivalent to n single nucleotide deletions. Similarly an inverted region should not be scored as a region with a large number of substitutions. Ideally, the value of δ should reflect the number of distinct mutational events that separate two sequences. This may not be unambiguous in some events because there may be more than one possible sequence of events that cause a particular change in the genome. However, we can expect that improved understanding of mutation rates will help decide between alternative evolutionary paths in such cases.

Table 3. Examples of pairwise genetic distance metrics δ.

Metric	Characteristic
Hamming or p distance	Takes into account only substitutions
K_s and K_n	Distance measures based on mutations at synonymous and non-synonymous substitution sites only, resp.
Edit or Levenstein distance	Both substitutions and indels
Likelihood pairwise alignment distance	Based on a model for mutation rates for substitutions and indels
Various phylogenetic branch-length estimators	Usually based on a model for mutation rates for substitutions and indels
Hidden Markov Model (HMM) distance	Measures HMM likelihood that a strain belongs to a given cluster of strains; includes substitutions and indels
Jumping HMM distance	HMM distance that includes gene duplications, transposons, and inversions

A considerable simplification in the calculation of δ arises because, the timescale of most outbreaks of diseases of concern in most actual microbial forensics cases, is short. This has two important effects. First, the total number of genomic changes is generally a modest fraction of the number of loci in the genome, even for viruses. This simplifies corrections for multiple mutations at a given locus, making "infinite sites" approximations reasonable. Second, it is reasonable to approximate the statistical processes that lead to sequence diversification as stationary over the duration of the outbreak[4]. The stationary assumption amounts to assuming that the mutation rates do not change appreciably over the duration of the outbreak. Similarly, there is an implicit assumption that outbreaks of the same disease on the same host type will exhibit similar mutation rates. These related assumptions can only be approximately correct because there is considerable evidence for the variation of mutation rates among pathogen substrains. A case in point is the common appearance of "mutator" or "anti-mutator" strains among certain pathogens where mutation rates can change by an order of magnitude because of mutations in replicase genes. In any case, we can expect the stationary approximation to be most accurate for sequences that are "closely related" to the consensus sequence for the population in question. The effect of non-stationary effects on the practical accuracy of our framework can only be established through

studies of real outbreaks. For now we will simply note that this approximation affects inferences made by phylogenetic construction methods as well.

3. Example applications

SARS outbreak, 2003

The SARS coronavirus outbreak of 2003 was an example of the unexpected emergence of a pathogen that had not been previously characterized. Approximately 8000 people were affected worldwide with all cases assumed to originate with one or a few index cases in Guandong province in China. In Singapore, several hundred people were infected and extensive contact tracing data has been published, along with whole genome sequences for isolates from certain selected nodes (i.e. patients). Although nearly 100 full genome sequences of the SARS coronavirus (SARS CoV) are available from NCBI, there are fewer than 12 pairs of sequences that are associated unambiguously with transmission-linked patients in open literature reporting. Nonetheless, this data is useful for illustrating the application of the framework we have described to a real-world case, and especially for indicating how better data sets may be collected in the future. In the analysis presented here we:

- Use a small group of Singapore isolates as a "training set" for P(k|M), and several empirical transmission trees as a "training set" for $\mathcal{P}$(M).

- Examine the self-consistency of inferences about the training set made on the basis of our inferential framework by applying it to the training set itself.

- Compare the inferential power of our proposed framework to a traditional phylogenetic analysis of the same data.

As our "training set" we will consider a set of isolates described in references 15 and 16, obtained from a small group of patients in Singapore for which full sequences are available, and which have been related by contact tracing.

The transmission relationships among these isolates implied by contact tracing is shown in Figure 14. Contact tracing suggested that Sin2679 could have been the result of infection through an unidentified intermediate patient who had contact with Sin2500, but later phylogenetic analysis was more consistent with transmission through independent contact with patients infected in the Metropol hotel outbreak. The isolate Frankfurt1 was identified as being derived from a doctor who had treated the patient in Singapore from whom Sin2774 had been obtained.

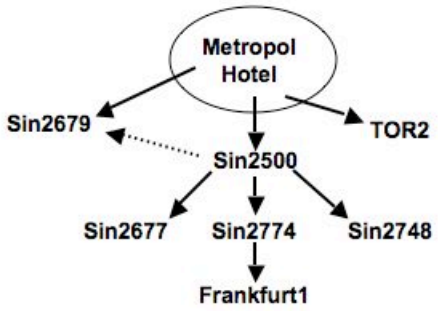


Figure 14. Transmission relationships among fully sequenced isolates from the Singapore SARS outbreak.

The data from references 15 and 16 were used to calibrate the number of substitutions per node-to-node transmission event, and the Poisson formula (11) was used as an estimator for P(k|M). $\mathcal{P}$(M) was derived from averaging the four distributions in Figure 6, which would be appropriate for a network of approximately 50 patients and approximately 4 generations. From this model we calculated P(M=M_0|k) for each pair of isolates as shown in Figure 15.

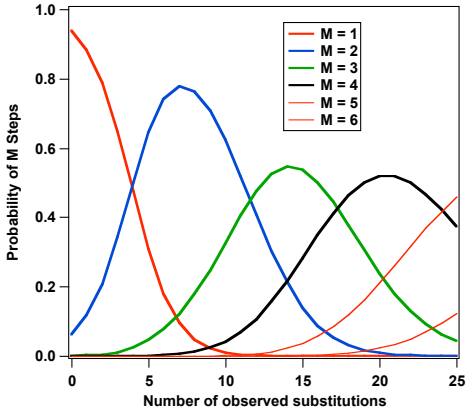


Figure 15. The estimated posterior probability P(M|k) for various values of k, for SARS.

For each pair of isolates, we performed a pairwise sequence alignment and determined the number of substitution differences k. (Indels and other sequence differences were not scored.) For a given k value, the curves in Figure 15 were used to determine the M value for which P(M|k) had the largest value. Figure 15 indicates that P(M=1|k) is the largest for pairs with $k \leq 4$, hence those pairs are most likely to be separated by a single transmission step. Figure 16 shows the pairs which were calculated to be most likely related by M=1, along with the posterior probability value for this relationship. The highest level of support (0.94) was observed for a direct transmission relationship between Sin2774 and Sin2774_P1. Since Sin2774_P1 was an isolate obtained by laboratory passage of the Sin2774 isolate in a vero cell culture, the high level of support is not unexpected. Similarly, the direct transmission relationship between Sin2500 and Sin2748 implied by contact tracing (Fig. 14) is strongly supported (P = 0.88). However, contrary to the contact tracing analysis, Sin2677 is found to be more likely the result of

direct transmission from Sin2748, rather than Sin2500. In addition, only weak support is found for direct transmission from the Sin2774 to Frankfurt1 (probability ≈30%).

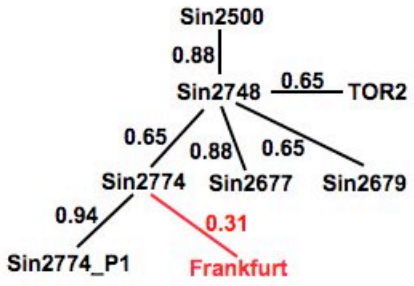


Figure 16. Predicted M=1 transmission relationships among Singapore isolates. Frankfurt1 and Sin2774 were predicted to be at least 2 transmission steps apart.

It is interesting to compare this set of sequences using phylogenetic analysis. Figure 17 is a consensus phylogenetic tree constructed using MrBayes. The maximum likelihood and parsimony phylogenetic construction techniques (using Phylip-3.68) gave essentially the same consensus tree. Included with the isolates of interest are a set of additional Singapore isolates whose transmission relationships are less well characterized. To be consistent with the use of only substitution data for determining transmission relationships we have performed the phylogenetic analysis on sequences from which all regions containing indels have been excised. One of the striking features of the tree in Figure 17 is that the phylogenetic relationships among closely related sequences such as Sin2677, Sin 2748, and Sin2500 are evidently difficult to resolve. In addition, there is at least one case in Figure 17 where contact tracing has identified two isolates with a common source, but the isolates are genetically closer to each other than they are to the putative source isolate. Thus, phylogenic relationships among the isolates also do not generally provide strong support for the relationships determined by contact tracing in this case.

In a standard phylogenetic analysis, the close relationship between Sin2774 and the Frankfurt isolate would be considered evidence in favor of direct transmission. In contrast, our analysis indicates that it is highly probable (70% vs 30%) that there was at least one intermediate transmission step, or that Frankfurt and Sin2774 were infected by a common source. This is a good example of the ambiguity of phylogenetics with respect to source attribution, as discussed in Appendix 1.

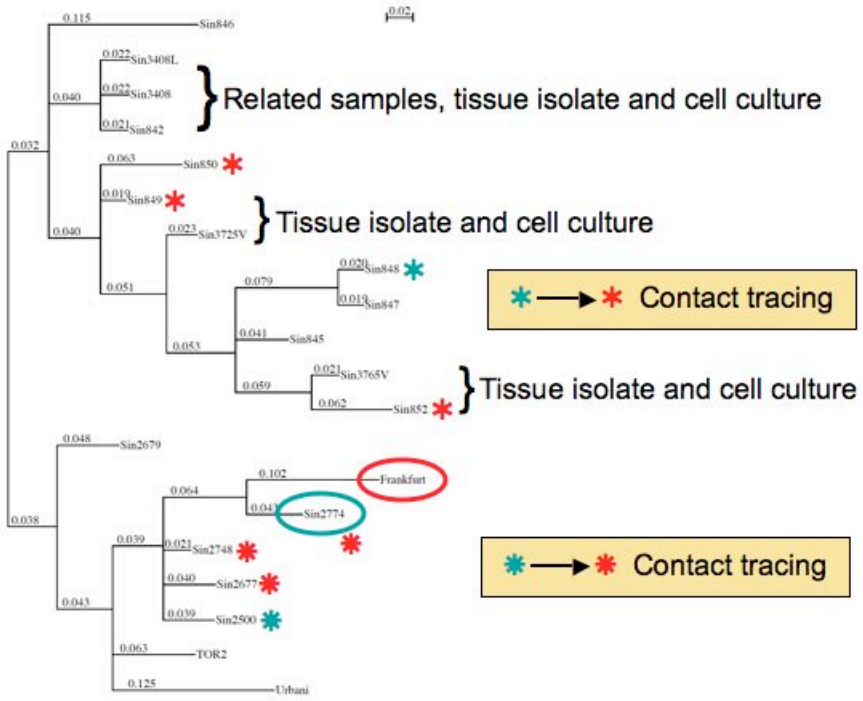


Figure 17. A phylogenetic tree using substitution data from Singapore sequences as determined by Liu. Blue and red asterisks mark source and recipient isolates as determined from contact tracing.

FMDV outbreak, Great Britain, 2001

We performed the same type of analysis on sequences and contact tracing data obtained during the U.K. FMDV outbreak of 2001[17]. We used the Poisson representation of P(k|M) and the $\mathcal{P}(M)$ distribution from Fig. 7, and calculated P(M|k) for pairs of isolates that Cottam, et. al. had identified as direct transmission pairs. (This is a case where we use the $\mathcal{P}(M|T)$ derived from a particular transmission tree to represent $\mathcal{P}(M)$ for the entire class of FMDV outbreaks of that size.) In some cases, the largest posterior probability was obtained for M>1, implying that one or more intermediary transmission steps were involved. The resulting transmission tree is shown in Figure 18. Here inferred unknown transmission nodes are indicated by Xs.

One of the limitations of this analysis is that single isolates were used to determine the sequence associated with a node (herd), so there is no guarantee that the sequence is a valid representation of the consensus sequence for the entire herd. Therefore, some of the "intermediate nodes" implied by our analysis might actually be artifacts caused by significant genetic drift within a larger herd that is not taken into account. Nonetheless, our probabilities are reasonably consistent with the transmission chain constructed by Cottam et. al., who based their analysis on the overlap of the durations of infectious period in each farm. By using data for the onset and duration of the infection on each farm, Cottam calculated a probability for each transmission link. In Table 4 we compare Cottam's probability with ours for selected links with very high or very low probabilities

according to Cottam. There is generally excellent concordance between the two calculations, with one notable exception. The direct transmission link between nodes K and F receives moderate support from our analysis, but Cottam calculates no temporal overlap. One possible explanation is that infection of K was caused by contaminated fomites from farm F whose transport to K was not stopped by isolation measures, but only delayed.

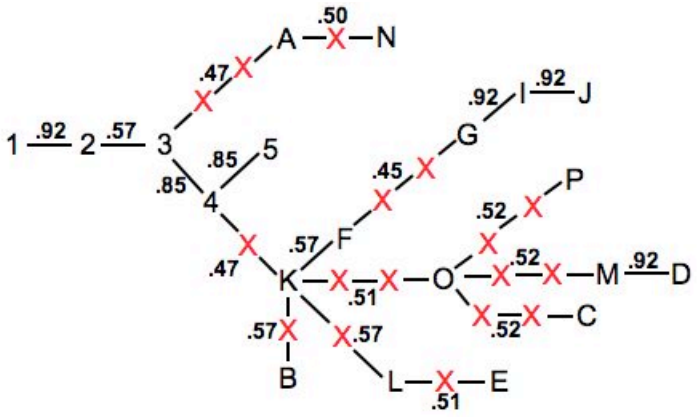


Figure 18. Reconstructed transmission chain for FMDV based on the data of Cottam, et. al. Xs represent implied intermediate nodes, and the posterior probability values are given for each link.

Table 4. Comparison with Cottam, et. al.

Pair	$P(t_1,t_2)$	$P(1\|k)$
1-2	**0.82**	**0.92**
I-J	**0.99**	**0.92**
3-A	**0.00**	**0.00**
4-K	**0.00**	**0.11**
K-F	**0.00**	**0.57**
F-G	**0.00**	**0.02**
O-M	**0.00**	**0.00**

Previous attempts to construct epidemic transmission trees from outbreak data from the 2001 FMDV outbreak invoked arbitrary assumptions about the time or distance relationships between transmission pairs[50,51]. The approach outlined above uses only genetic data and statistical properties of transmission networks. However, as in the case of the SARS outbreak, the paucity of data provides us no independent way to validate the

method, and the reconstructed trees in Figs 16 and 18 simply demonstrate the degree of self-consistency that can be achieved.

4. Extension to other source attribution problems

a. Identifying source reservoirs

Consider a scenario where an outbreak of a viral disease occurs among humans, and it is suspected to be a natural outbreak caused by contact with some localized reservoir. Viral isolates from the initial human victims are obtained along with isolates from the suspected source reservoir. The reservoir isolates could be samples collected directly from a natural host sub-population during the epidemiological investigation, or they might be samples from one or more laboratories, which had collected previously from the natural host sub-population for research purposes. The genetic sequences of the viral isolates are obtained and compared. How confident are we that the source of the human infection was the identified reservoir based on the sequence comparison?

This problem can be addressed by using arguments similar to that leading to equation (9), except that the source node is now a rescaled network. According to Haydon, a reservoir can be defined to be one or more epidemiologically connected populations or environments in which a pathogen can be permanently maintained and from which infection is transmitted to the defined target population[52]. Thus, zoonotic reservoirs are themselves disease transmission networks. Sub-networks that make up the reservoir may be segregated by geographic or social separation (Herd, flock, colony,…) and the testable hypothesis H_0 identifies the source as a certain sub-network from which the infection was derived. (Note that this is *not* necessarily a geographic construct.)

We seldom (if ever) have complete knowledge of the location and nature of all the sub-populations that make up the natural reservoir for a disease, or the corresponding sequences. Sparsely sampled genetic data prevents simple "matching" at the most interesting local scales. Therefore, a statistical inference approach based on population genetics and network theory is necessary.

Each sub-network (sub-reservoir) that is a potential source is characterized by a sub-population size N_S that reflects the current number of infected hosts within that sub-reservoir. We normally have only a few isolates sampled from individual hosts (or perhaps vectors) whose transmission relationship is unknown. If we compare the sequences from the victim and each reservoir isolate, we can calculate a δ value for each comparison, and find δ_{min}, the smallest observed difference among the set. The probability that the victim's infection came from the sub-reservoir in question is given by $P(M \leq D_0 | \delta_{min})$ where D_0 is the diameter of the sub-network.

To calculate this quantity we need an estimate of the total size of the population that could be harboring the pathogen – i.e. all natural hosts of the pathogen over the entire world (the "grand reservoir" with population N_T), as well as an estimate of the size of the subpopulation of hosts in the suspect sub-reservoir N_S. The diameters[48] of the grand reservoir and sub-reservoir are given by:

$$D_T \approx A_T \log_e(N_T)$$

$$D_S \approx A_S \log_e(N_S)$$

A_S and A_T are of order 1 and depend on the network topology; note that the topological properties of the grand and sub-networks might be somewhat different. However, due to the "small world" property, D is not very sensitive to the choice of N.

The probability that the victim's infection came from the sub-reservoir is then:

$$P(M \le D_S|\delta_{min}) = P(\delta_{min}|M \le D_S)\mathcal{P}(M \le D_S)/[P(\delta_{min}|M \le D_S)\mathcal{P}(M \le D_S) + P(\delta_{min}|M \ge D_S)\mathcal{P}(M \ge D_S)]$$

Where the definition of each factor is analogous to that in equation (9). Note that the relevant $\mathcal{P}(M)$ is that for the grand reservoir, and the sums for $M \ge D_S$ go from $D_S + 1$ to D_T.

b. Intentional versus natural outbreaks

Emerging disease outbreaks are often the consequence of changes that increase the contact between humans and reservoirs in the wild that harbor the unknown pathogen. However, outbreaks of such diseases often raise public questions about the possibility that they are due to bio-terrorism. From a network point of view, a large number of infections caused by independent interactions with a natural zoonotic reservoir is easily distinguished from a large number of infections from a common source such as a single laboratory produced culture. Referring to Figure 19 the genetic distance values between pairs of victim isolates in the latter case will be consistent with step separations of $M \le 2$, while victim isolates from independent contacts with a natural network will exhibit high probabilities of $M > 2$. Note that an isolate from the putative reservoir is not needed to draw this conclusion. Obviously, there may be cases where the natural source of an outbreak is a single infected host from the reservoir, and conversely, a terrorist might deliberately mix isolates from many host animals, but simple tests such as this will still be a useful adjunct to standard epidemiological investigations.

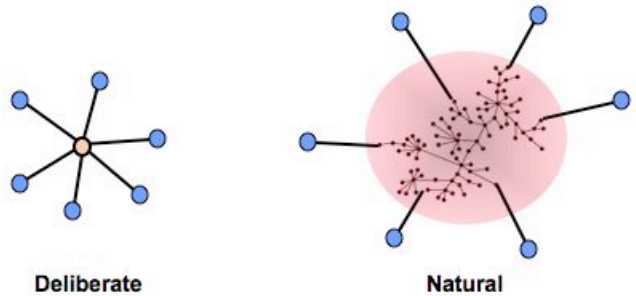


Figure 19. Distinguishing a deliberate biological attack using an isolate obtained from a zoonotic reservoir from multiple infections due to independent contacts with that reservoir.

c. Recombination and re-assortment events

For influenza and certain other viral diseases, re-assortment or recombination events of concern occur in nodes that have been infected from two distinct networks, in particular networks that involve different species. A node in which recombination or re-assortment has taken place may then be the initial node in a new, distinguishable outbreak network. The analysis of such strains is complicated by the need to identify the sub-sequences in the new strain that are likely to belong to each of the original networks (i.e. species.) Once this has been done, the framework can be used on each genomic segment separately, to test the hypothesis that any two particular suspect reservoirs/outbreaks were the sources of the recombining strains. If the most likely suspect reservoirs are unlikely to have a natural route of contact, it might indicate an artificial (i.e. man-made) origin for the event.

5. Foundations for a Microbial DNA Indexing System (MIDIS)

Several authors concerned with microbial forensics have discussed the need for a forensic archive and genetic database of pathogens of concern[53-55]. However, a database by itself is of limited utility in the absence of appropriate algorithms for comparing genetic profiles and assisting the investigator in making decisions about the likelihood that there is a case-relevant relationship between two isolates, or providing information on where additional related samples might be collected for genetic characterization. The framework outlined in this report provides a basis for such algorithms, and can be incorporated into a concept for an electronic case support system analogous to the CODIS system for human DNA[56].

The Microbial DNA Index System (MiDIS) concept shown schematically in Figure 20 is a decision support tool to help investigators narrow the range of possible sources of a bacterial or viral agent that was used in a criminal or bioterror incident. While it is similar in spirit to the CODIS system, its inferential power and utility differ from CODIS in several ways. In particular, MiDIS

- ☞ Provides a systematic way to identify genetically characterized isolates that are most closely related to the attack strain
- ☞ Provides explicit statistical measures to quantify the probability that the attack strain was derived from a suspect source strain under an explicit hypothesized scenario for acquisition and propagation of the agent.
- ☞ Provides guidance for identifying geographical regions or laboratories where additional isolates that may be closely related to the attack strain may be collected.

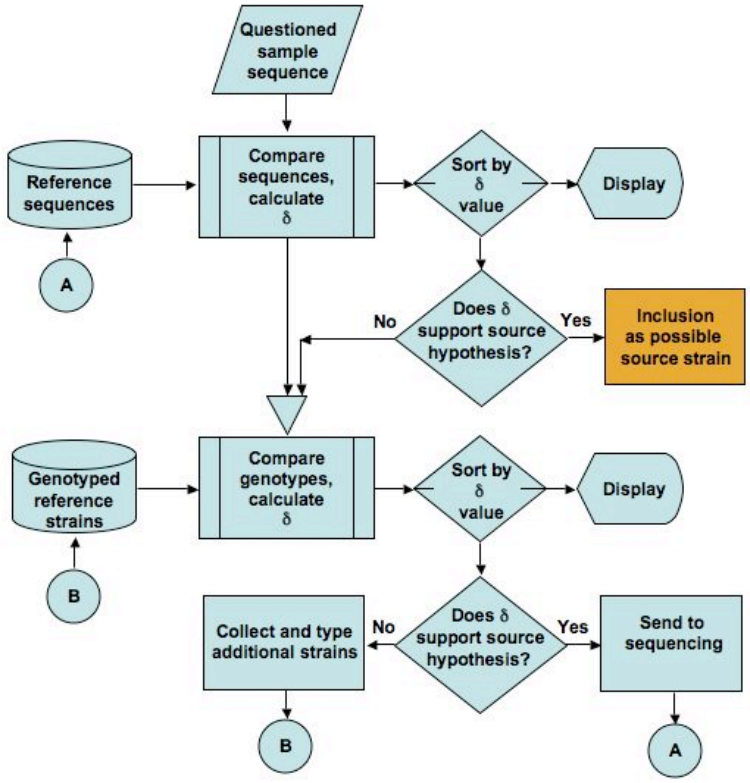


Figure 20. Schematic representation of the MiDIS system

MiDIS does not purport to identify sources. As in the case of CODIS, the output of MiDIS is primarily used to assist investigation; it provides only one part of the information needed to establish attribution. The MiDIS concept as presented here formally implements the probabilistic framework presented in this report.

An important feature of the MiDIS concept is that for each pathogen estimates of $P(\delta|M)$ and $\mathcal{P}(M)$ are generated from data culled from past outbreaks using one or more of the methods described in section 2. Where no data is available, or a novel pathogen is involved, similarity classes with known pathogens are established and used to make the estimates. Comparisons between case and reference samples, or among case samples are ranked by posterior probability metrics such as equations (8), (9) or (10) which provide the investigator with weight-of-evidence estimates for the support given to various hypotheses about the provenance of the pathogen in question.

MiDIS data base standards

The development of MiDIS would necessitate establishing standards for including or rejecting data, or weighting them with respect to certainty. Note that these standards are most critical for transmission network data established by classical contact tracing methods. Eventually it may be possible to influence the field collection of outbreak data by providing recommendations to the CDC and international disease investigation teams

for the number and types of isolates to be collected during a reference outbreak, and the kind of metadata that should accompany a sample. Similarly, MiDIS standards for genotyping could be established to control the genomic regions that are sequenced, the error tolerances, and recommendations concerning the number of clones sampled from a given isolate. Any assumptions used in MiDIS calculations should always be conservative, in the sense of favoring the defense.

6. Strategies for validation

The operational utility of the methods developed in this report, the forensic admissibility of the results generated by their use, the credibility of conclusions drawn from the results to the scientific community and to policymakers, and the value of a MIDIS-like system to guide the investigation of a bio-terror or criminal incident, all depend on an extensive program of validation.

The most fundamental assertion that requires validation is the ability of models for $P(\delta|M)$ and $\mathcal{P}(M)$ that are parameterized on one data set to accurately reproduce the empirical distributions derived from independent data sets obtained from other outbreaks of the same disease. To do this, it will be essential to improve the coordinated collection of coupled epidemiological and genetic data from real outbreaks and increase the number of such studies. In the case of contact tracing, confidence levels associated with transmission events should be reported, so that these uncertainties can be formally folded into the statistical analysis of the data. Large enough data sets should be collected to permit the use of cross-validation and bootstrapping in evaluating empirical distributions. Modern high throughput sequencing technologies should be employed to determine the full genome sequences of all collected viral isolates. Sequencing error should be controlled by the use of high-depth coverage of each sequence.

A much more detailed understanding of the uncertainties introduced by unknown variation in times between infection and transmission, and between infection and sampling must be obtained. Experimental studies using animal models and in-vitro systems may be used to fill gaps that remain when there are only a limited number of field outbreak studies. It should be relatively easy to minimize uncertainties regarding transmission relationships under controlled laboratory conditions. Animal transmission/passage models already exist for a number of airborne infectious diseases, such as the ferret model for influenza[57] and Rhesus Monkey models for SARS[58]. Of course, the best animal models for determining transmission related statistical distributions may be different from those used for vaccine development and other kinds of studies. Important considerations are the levels of viral loads developed during infection, the expression of symptoms conducive to transmission, and similarity between the routes of transmission and infection between humans versus the animal models. In-vitro passage experiments can be used to determine fundamental rate parameters in cases where there is no data from direct host-host transmission experiments.

Additional confidence in inferential methods may be obtained from sensitivity studies using simulations that couple disease propagation with mutational change. In this

context it is interesting to note that these simulations use the same evolutionary models that are used in many phyologenetic construction methods. Simulations are valuable in assessing answers to a number of "what if" questions when no directly applicable data is at hand.

Finally, further study of the statistical and mathematical foundations can provide increased confidence in the validity of the framework we have presented. For example, it remains to develop a theoretical basis for the functional form of $P(\delta|M=1)$. The method of constructing $P(\delta|M)$ for $M>1$ using the convolution principle mentioned in section 2 needs to be tested against empirical data. The remarkable insensitivity of the posterior probability $P(M=1|\delta)$ to $\mathcal{P}(M)$ over a wide range of network size needs to be investigated further to understand the degree of generality of this phenomenon. A formal treatment of the effect of sequencing and epidemiological errors remains to be developed.

7. Concluding remarks

An underlying theme in the framework presented in this report is that the most likely unit of concern in microbial forensics is the *outbreak*. A pathogen used in a bioterror incident or crime will most likely be one whose origin was ultimately a distinct, known historical or contemporaneous outbreak. Most isolates in pathogen collections are sampled from, or associated with, recognized human or animal outbreaks. Chances are that someone trying to obtain a pathogen from a natural source will obtain it during an outbreak. This is because infectious disease outbreaks in humans or domestic animals are generally the most visible, noted, and tracked expansions of pathogen populations. Our framework provides a quantitative method for linking any pathogen isolate to a particular outbreak, a sub-tree of that outbreak, or to an infected node within that outbreak. *It is important to recognize that this is the strongest kind of inference that can be made about the origin of an isolate used in a bioterror incident that is possible from genetic data alone.*

The key to our proposed framework for microbial genetic inference is the recognition that an outbreak at any scale consists of a network of infected nodes, and that the statistical properties of the network determine, to a large extent, the structure of the microbial population within the outbreak. In this framework it is possible to derive "pathogen paternity" equations that permit statistical inferences regarding source hypotheses, and quantitative forensic inferences require explicit consideration of the statistical properties of disease transmission networks. Phylogenetic arguments alone cannot generate quantitative inferences about sources and transmission events.

This deductive framework provides a basis for a CODIS-like database for analyzing microbial evidence, permits quantitative statements to be made regarding the differentiation of natural from intentional events, and can be extended to include inferences about particular animal reservoirs being sources of natural outbreaks. The method can be extended to include tracking concepts based on the genotyping of non-pathogenic (commensal) microbial populations within humans.

Although sufficient coupled genetic-epidemiological data is available to demonstrate applications of this framework to a variety of source attribution problems of forensic interest, rigorous validation will require deliberate and systematic collection of additional data. The framework also helps provide guidance for such an effort. With the spread of "next generation" sequencing technologies, collecting the required volumes of sequence data will be feasible and cost-effective.

References

1. David J. Balding, *Weight-of-evidence for Forensic DNA Profiles*, (John Wiley & Sons, Ltd, West Sussex, England, 2005).

2. John H. Gillespie, *Population Genetics, A Concise Guide*, 2nd Ed., (The Johns Hopkins University Press, Baltimore, MD, 2004). The reader familiar with mtDNA forensics will note the similarity between the picture of microbial evolution presented in this report and published concepts regarding mtDNA evolution. There are clear analogies with regard to "genetic bottlenecks in mtDNA transmission", "star-like gene trees", and population expansion effects on the distribution of pairwise genetic differences.

3. *Statistical Methods in Molecular Evolution*, R. Nielsen, ed. (Springer Science+Business Media, 2005)

4. Wen-Hsiung Li, *Molecular Evolution*, (Sinauer Associates, Inc., Sunderland Massachusetts, 1997).

5. Pastor-Satorras R, and Vespignani A., "Epidemic dynamics in finite size scale-free networks", Phys. Rev. E **65**, 035108 (2002).

6. Newman MEJ, "Spread of epidemic disease on networks", Phys. Rev. E **66**, 016128 (2002).

7. Meyers LA, Newman MEJ, Martin M, and Schrag S., "Applying Network Theory to Epidemics: Control Measures for *Mycoplasma Pneumoniae* Outbreaks", Emerging Infectious Diseases **9**(2) 204-210, (2003).

8. Liljeros F, Edling CR, Amaral LAN, "Sexual networks: implications for the transmission of sexually transmitted infections", Microbes and Infection 2003; **5**:189-196.

9. Masuda N, Konno N, and Aihara K, "Transmission of severe acute respiratory syndrome in dynamical small-world networks", Phys. Rev. E **69**, 031917 (2004).

10. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, and Brunham RC, "Network theory and SARS: predicting outbreak diversity", Journal of Theoretical Biology **232** (2005) 71-81.

11. Meyers LA, Newman MEJ, Pourbohloul B, "Predicting epidemics on directed contact networks", Journal of Theoretical Biology **240** (2006) 400-418.

12. Aparicio JP, and Pascual M, "Building epidemiological models from R_0: an implicit treatment of transmission in networks", Proc. Roy. Soc. B (2007) **274**, 505-512.

13. Chartrand G, *Introductory Graph Theory*, (Dover Publications, New york, 1985).

14. Felsenstein, J. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach, J. Mol. Evol. (1981) **17**:368-376.

15. Vega VB, Ruan Y, Liu J, Lee WH, Wei CL, Se-Thoe SY, Tang KF, Zhang T, Kolatkar PR, Ooi EE, Ling AE, Stanton LW, Long PM, and Liu ET, "Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003", BMC Infectious Diseases 2004; **4**:32.

16.Liu J, Lim SL, Ruan Y, Ling AE, Ng LFP, Drosten C, Liu ET, Stanton LW, and Hibberd ML, "SARS Transmission Pattern in Singapore Reassessed by Viral Sequence Variation Analysis", PLoS Medicine 2005; **2**(2) e43.

17. Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT, "Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus", Proc. Roy. Soc. B (2008) online publication.

18. Trask SA, Derdeyn CA, Fideli U, Chen Y, Meleth S, Kasolo F, Musonda R, Hunter E, Gao F, Allen S, and Hahn BH, "Molecular Epidemiology of Human Immunodeficiency Virus Type 1 Transmission in a Heterosexual Cohort of Discordant Couples in Zambia", Journal of Virology, 2002; **76**(1):397-405.

19. Korber BT, Foley BT, Kuiken CL, Pillai SK, and Sodroski JG, "Numbering Positions in HIV relative to HXB2CG", Los Alamos HIV sequence database, *Human Retroviruses and AIDS*, 1998.

20. Wessa P. (2008), Maximum-likelihood Poisson Distribution Fitting (v1.0.2) in Free Statistics Software (v1.1.23-r3), Office for Research Development and Education, URL http://www.wessa.net/rwasp_fitdistrpoisson.wasp/

21. Uzzell, T, and Corbin, KW, "Fitting Discrete Probability Distributions to Evolutionary Events", Science, (1971); **172**(3988):1089-1096.

22. Gelman A, Carlin JB, Stern HS, and Rubin DB, *Bayesian Data Analysis* 2nd Ed. (Chapman & Hall/CRC, 2004).

23. Bergstrom, CT, McElhany P, Real, LA, Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens, 1999, Proc. Natl. Acad. Sci. USA 96:5095-5100.

24. Jenkins, GM, Worobey, M, Woelk CH, Holmes EC, Evidence for the Non-quasispecies Evolution of RNA Viruses, 2001; Mol. Biol. Evol. 18(6):987-994.

25. Manrubia CM, Arribas M, Lazaro E, Supercritical branching processes and the role of fluctuations under exponential population growth, Journal of Theoretical Biology 225 (2003 497-505. (This reference has a brief discussion of modeling the expansion with

mutation of a population of viruses as a branching network. They perform a model calculation of the fraction of each genotype in the final population after a fixed number of generations, and state that it is strongly peaked around the original "seed" genotype. This type of model is even more explicit than the quasi-species equation approach, which represents a sort of "mean field" approximation to the actual population dynamics.

26. Jain, K, Evolutionary dynamics of the most populated genotype on rugged fitness landscapes, 2007; Physical Review E 76:031922 1-10.

27. A somewhat more restricted approach has been used with some success to describe the distribution of substitutions within the viral population during the early stages of HIV infection. See Keele, et. al., "Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection", PNAS **105**, 7552-7557.

28. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, and Shafer RW, "Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance", Genome Res. 2007; **17**:1195-1201.

29. Goh K-T, Cutter J, Heng B-H, Ma S, Koh BKW, Kwok C, Toh C-M, and Chew S-K, "Epidemiology and Control of SARS in Singapore", Ann. Acad. Med. Singapore 2006; **35**:301-16.

30 Varia M, Wilson S, Sarwal S, McGeer A, Gournis E, Galanis E, and Henry B, "Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada", CMAJ 2003; 169(4):285-92.

31 Shen Z, Ning F, Zhou W, He X, Lin C, Chin DP, Zhu Z, and Schuchat A, "Superspreading SARS Events, Beijing, 2003", Emerging Infectious Diseases 10(2), pp.256-260, (2004).

32. Theodore E. Harris, *The Theory of Branching Processes*, (Prentice-Hall, Inc. Englewood Cliffs, N.J., 1963).

33. Fronczak A, Fronczak P, and Holyzt JA, "Average pathlength in random networks", Phys. Rev. E 70, 056110, (2004).

34. Newman MEJ, The Structure and Function of Complex Networks, SIAM Review 45(2) 167-256 (2003)

35. Albert R, Barabasi A-L, Statistical mechanics of complex networks, Rev. Mod. Phys. 2002; 74:47-97.

36. Ayyalasomayajula S, DeLaurentis DA, Moore GE, and Glickman LT, "A Network Model of H5N1 Avian Influenza Transmission Dynamics in Domestic Cats", Zoonoses and Public Health 55 (2008), 497-506.

37. Riley S, "Large-Scale Spatial-Transmission Models of Infectious Disease", Science **316**, pp. 1298-1301, (2007).

38. Chen H, Chen Y, Larson C, Tseng C, King C-C, Wu T-S J, "Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study of Taiwan SARS Data" National Science Foundation ITR Program presentation (No date available). This presentation contains a description of a 961 patient network dataset obtained from the SARS outbreak in Taiwan.

39. Lowy FD and Miller M, "New methods to investigate infectious disease transmission and pathogenesis – Staphylococcus aureus disease in drug users", The Lancet Infectious Diseases Vol. 2, October 2002, pp. 606-612.

40. Eubank S, Guclu H, Anil Kumar VS, Marathe MV, Srinivasan A, Toroczkal Z, and Wang N, "Modelling disease outbreaks in realistic urban social networks", Nature **429** 180-181, (2004).

41. Webb CR, "Farm animal networks: unraveling the contact structure of the British sheep population", Preventive Veterinary Medicine **68** (2005) 3-17.

42. Robinson SE, Everett MG and Christley RM, "Recent network evolutin increases the potential for large epidemics in the British cattle population", J. R. Soc. Interface (2007) **4**, 669-674.

43. Grassly NC, and Fraser C, "Mathematical models of infectious disease transmission" Nature Reviews Microbiology online publication 13 May 2008.

44. Glass LM, and Glass RJ, "Social contact networks for the spread of pandemic influenza in children and teenagers", BMC Public Health 2008; **8**:61-76.

45. Ohkusa Y, and Sugawara T, "Application of an individual-based model with real data for transportation and location to pandemic influenza", J. Infect. Chemother (2007) **13**:380-389.

46. Lee BY, Bedford VL, Roberts MS, and Carley KM, "Virtual epidemic in a virtual city: simulating the spread of influenza in a US metropolitan area", Translational Research **151**(6):275-287.

47. Corner LA, Pfeiffer DU, Morris RS, "Social-network analysis of Mycobacterium bovis transmission among captive brushtail possums (Trichosurus vulpecula)", Prev Vet Med. 2003 Jun **12**;59(3):147-67.

48. Lloyd-Smith J. O., Schreiber S. J., Kopp P. E. & Getz W. M., "Superspreading and the effect of individual variation on disease emergence", Nature Vol **438**|17 November 2005|doi:10.1038/nature04153

49. Bollobas B, Riordan O, "The diameter of a scale-free random graph", Combinatorica **24**(1) pp. 5-34, (2004).

50. Cottam EM, Haydon DT, Paton DJ, Gloster J, Wilesmith JW, Ferris NP, Hutchings GH, and King DP, "Molecular Epidemiology of the Foot-and-Mouth Disease Virus Outbreak in the United Kingdom in 2001", Journal of Virology 2006; 80(22):11274-11282.

51. Carillo C and Rock D, "Molecular Epidemiology and Forensics of RNA Viruses", in Microbial Forensics R.G. Breeze, B.Budowle, and S. Schutzer eds. (Elsevier, 2005)

52. Haydon, DT, Cleaveland, S, Taylor, LH, and Laurensen, MK, Identifying Reservoirs of Infection: A Conceptual and Practical Challenge. Emerging Infectious Diseases **8**(12), pp1468-1473, (2002).

53. Budowle B, "Defining a New Forensic Discipline: Microbial Forensics", publication #02-12 of the Laboratory Division of the Federal Bureau of Investigation.

54. Budowle B, Murch Randall, Chakraborty R, "Microbial forensics: the next forensic challenge", Int. J. Legal Med. (2005) **119**: 317–330

55. Keim P, et. al. "Microbial Forensics- A Scientific Assessment", American Academy of Microbiology Report, 2003.

56. Baechtel FS, Monson KL, Forsen GE, Budowle B, Kearney JJ., "Tracking the violent criminal offender through DNA typing profiles--a national database system concept" EXS. 1991;**58**:356-60.

57. Herlocher ML, Elias S, Truscon R, Harrison S, Mindell D, Simon C, and. Monto AS, "Ferrets as a Transmission Model for Influenza: Sequence Changes in *HA1* of Type A (H3N2) Virus", The Journal of Infectious Diseases 2001;**184**:542–6

58. Rowe T, Gao G, Hogan RJ, Crystal RG, Voss TG, Grant RL, Bell P, Kobinger GP, Wivel NA, Wilson JM., "Macaque model for severe acute respiratory syndrome", J. Virol. 2004 Oct;**78**(20):11401-4.

Appendix 1.
Limitations of phylogenetic construction for microbial genetic inference

Phylogenetic approaches constitute the primary methods that have been proposed for inferring transmission trees from genetic sequences of isolates sampled from an outbreak. Slatkin and Maddison[a1.1] proposed a method for reconstructing a transmission history from the phylogenetic tree determined by a set of geographically labeled isolates. This method has been refined and modified by Wallace, et. al. to retrospectively reconstruct the historical transmission of H5N1 influenza across the globe[a1.2]. Cottam and co-workers have considered how to meld phylogenetic information with epidemiological records of infection times and the duration of infectious periods in order to infer the transmission history of the 2001 FMDV outbreak in the United Kingdom[a1.3]. Phylogenetic methods have been used to support crimimal prosecution in HIV transmission cases, and the use of phylogenetic trees as evidence for the "close relationship" of HIV isolates has been declared admissible under Daubert[a1.4]. However, in spite of the general acceptance of phylogenetic constructions as inferential tools, all applications of phylogeny to microbial source inference suffer from one or more of the following limitations:

(1) They implicitly assume that all possible infectious nodes have been identified and one or more sequences are available for each node in a contiguous transmission tree so that inference is occurring on a closed set of possible nodes. When this assumption is not true, the inferential power of a tree is restricted to statements that a pair of isolates is genetically closer to each other than to other isolates in the compared set, that the construction provides a measure of genetic similarity between two isolates in the set (e.g. the sum of the branch lengths to the most recent common ancestor,) and that there is an inferred common ancestor sequence for any pair of sequences in the compared set.

(2) An ancestral sequence identified for two isolates cannot be placed in any particular node without some additional information or assumptions, either times associated with transmission events, or evidence excluding the possibility of additional uncharacterized nodes. When the complete set of nodes is not available, an observed phylogenetic relationship may be consistent with many alternative transmission trees. Similarly, several phylogenetic patterns may correspond to the same transmission history. Table A1.1 illustrates alternative transmission relationships that are consistent with a given phylogeny. Conversely, Figure A1.1 is an illustration from Resik, et. al. showing alternative phylogenies that are consistent with the same transmission relationship[a1.5].

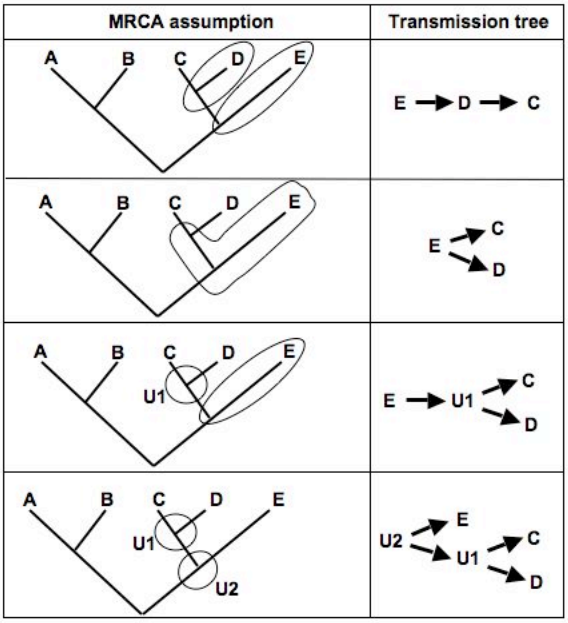


Table A1.1 Alternative transmission trees consistent with a given phylogeny. Encircled regions indicate which node the most recent common ancestor resides in. U1 and U2 correspond to unsampled (possibly unknown) nodes in the transmission tree, or nodes for which genetic information was not available.

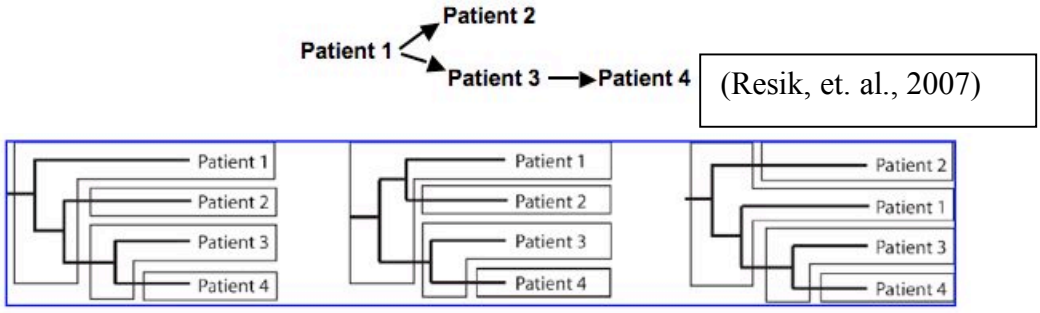


FIG. 1. Tree topologies compatible with the hypothetical transmission chain (Patient 1 → Patient 2; Patient 1 → Patient 3 → Patient 4). The host, in which the viral lineages reside, is superimposed onto the evolutionary history. All three evolutionary histories match the transmission hypothesis depending on the scenario of ancestral diversity and lineage sorting.

(3) Confidence levels are expressed as the ratio of the likelihoods of the most to the next-most probable trees. The likelihoods themselves have no obvious relationship to the probability of the event in question (i.e. the transmission event) and generally provide only a relative measure of confidence that similar data would produce a similar tree.

A common misperception is that phylogenetic constructions based on isolates from *known* viral transmission trees provide support for the use of such constructions for deducing transmission relationships. The classic paper by Leitner on HIV is often cited[a1.6]. In fact, this perception is erroneous, being a clear case of the fallacy of exchanging the conditional. If Φ represents a phylogenetic construction, and $\mathcal{T}$ a transmission tree relating a set of genetic sequences, then comparisons such as Leitner's

provide a measure of $P(\Phi|\mathcal{T})$, *not* $P(\mathcal{T}|\Phi)$. Moreover, it is not widely appreciated that several papers, including Leitner's can be interpreted as demonstrating (perhaps inadvertently) that $P(\Phi|\mathcal{T})$ is, in fact, low. (In Leitner, 14 out of 14 proffered phylogenetic constructions were in error in at least one branch.)

References for Appendix 1

a1.1 Slatkin M, Maddison WP, Genetics. 1990 Sep;126(1):249-60.

a1.2 Wallace RG, Hodac H, Lathrop RH, Fitch WM, Proc Natl Acad Sci U S A. 2007 Mar 13;104(11):4473-8. Epub 2007 Mar 7.

a1.3 Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT, Proc Biol Sci. 2008 Apr 22;275(1637):887-95.

a1.4 Budowle B, Harmon R., Croat Med J. 2005 Aug;46(4):514-21.

a1.5 Resik, et. al., AIDS RESEARCH AND HUMAN RETROVIRUSES **23**(3), 2007, pp. 347–356.

a1.6 Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J., Proc Natl Acad Sci U S A. 1996 Oct 1;93(20):10864-9.

Appendix 2
The complete joint sampling distribution $P(\delta, t_1, t_2, \tau, M)$

In this appendix we derive equation (0) in the main text.

Consider the complete joint sampling distribution that provides the probability of observing a particular genetic distance for a pair of isolates that were sampled at times t_1, t_2, and τ, and separated in the tree by M steps. The distribution is defined over the ensemble of outbreak trees for a particular infectious disease and the ensemble of possible isolate sampling times. It can be written in terms of the conditional probability for observing δ *given* t_1, t_2, τ and M as:

$$P(\delta, t_1, t_2, \tau, M) = P(\delta| t_1, t_2, \tau, M)\, P(t_1, t_2, \tau, M) \qquad (A2.1)$$

$$= P(\delta| t_1, t_2, \tau, M)P(t_1)P(t_2)P(\tau)P(M)$$

where the second equation follows under the assumption that the sampling times and M are presumably independent and uncorrelated.

We can integrate over the times t_1, t_2, and τ (which could be considered "nuisance variables" in our problem) to obtain the joint distribution of δ and M:

$$P(\delta, M) = \iiint P(\delta, t_1, t_2, \tau, M)dt_1dt_2d\tau \qquad (A2.2)$$

Substituting (A.2.1) into (A.2.2) gives:

$$P(\delta, M) = \iiint P(\delta| t_1, t_2, \tau, M)P(t_1)P(t_2)P(\tau)\, dt_1dt_2d\tau\, P(M) \qquad (A2.3)$$

By comparison with equation (0) in the main text we see that:

$$P(\delta|M) = \iiint P(\delta| t_1, t_2, \tau, M)P(t_1)P(t_2)P(\tau)\, dt_1dt_2d\tau \qquad (A2.4)$$

Appendix 3
Fitting the Negative Binomial distribution to substitution data

The negative binomial is often used to fit count data when the variance exceeds that expected for Poisson statistics. It is well known that the negative binomial is the distribution for a sum of random variables each of which is Poisson distributed, but with different rate constants. (It is assumed that the rate constants are distributed according to a gamma distribution.) Thus the negative binomial is one of a class of "overdispersed" counting distributions. To determine the parameters, we fit our data using the R online software (Wessa, P. (2009) Free Statistics Software, Office for Research Development and Education, version 1.1.23-r3, URL http:www.wessa.net/) This fitting routine returns the parameters μ and "size" as recorded below for the four data sets fit in section 2.

Data Set	μ	size
SARS	4.1 ± 1.17	3.2 ± 3.7
FMDV	4.3 ± 0.5	358 ± 3229
HIV env	1.84 ± 0.19	4.3 ± 2.7
HIV gag	2.04 ± 0.47	1.11 ±0.57

There are various forms for the Negative Binomial distribution. One common form is[A3.1]:

$$P(k|\mu,\kappa) = [\Gamma(k + 1/\kappa)/k!\Gamma(1/\kappa)]\ (\kappa\mu/(1 + \kappa\mu)^k\ \ (1/(1+\kappa\mu))^{1/\kappa}$$

Here $\kappa = 1/\text{size}$.

Another version of this distribution is[3.2]:

$$P(k|r,p) = \binom{k + r - 1}{k} p^r(1 - p)^k$$

where r = size + 1 and p = size/(size + μ).

For computations of P(k|M) we use the result derived for sums of independent, equally distributed Negative Binomial random variables derived by Furman[A3.3]. According to that paper, we can simply replace size → M•size in the above forms to obtain the required distribution.

A3.1. Lloyd-Smith JO (2007) Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLoS ONE **2**(2): e180. doi:10.1371/journal.pone.0000180

A3.2. Cook, JD, Notes on the Negative Binomial Distribution, October 8, 2008; available on the Internet at www.johncook.com/negative_binomial.pdf

A3.3. Furman E, "On the convolution of the negative Binomial random variables", Statistics & Probability Letters 2007; **77**:169-172.

Appendix 4
Evolution equation approach to calculating P(δ|M=1)

The calculations leading to Figure 7 in the main text were performed using a discrete form of the "quasi-species equations" which relate the number of microbes with genetic sequence i found in a population to the probabilities of mis-copying (changing genotype i to a different genotype j and vice-versa) and the "fitness" of each genotype measured in terms of an effective reproduction number:

$$N_i(N_{gen}+1) - N_i(N_{gen}) = F_i\, N_i(N_{gen}) - \sum_{j\neq i} Q_{ij}\, N_i(N_{gen}) + \sum_{j\neq i} Q_{ji}\, N_j(N_{gen}) \qquad (A4.1)$$

where

N_i is the number of microbes with genotype i
N_{gen} = the number of generations
F_i = the effective reproduction number (the fitness)
Q_{ij} = the transition matrix containing the probabilities of genotype i changing to j.

A set of coupled "quasi-species equations" is equivalent to (A4.1) with an expression for the total population of microbes $\mathcal{N} = \sum_i N_i$:

$$\mathcal{N}(N_{gen}+1) = (1 + \Phi(N_{gen}))\bullet\mathcal{N}(N_{gen}) \qquad (A4.2)$$

where $\Phi(N_{gen})$ is the "population fitness" defined as:

$$\Phi(N_{gen}) = \sum_i F_i[N_i(N_{gen})/\mathcal{N}(N_{gen})]. \qquad (A4.3)$$

The fraction of the total population with genotype i is simply

$$x_i(N_{gen}) = N_i(N_{gen})/\mathcal{N}(N_{gen}) \qquad (A4.4)$$

The equations are solved subject to an initial condition on the distribution of genotypes in the initial population $\{x_i(N_{gen}=0)\}$.

To generate the data shown in Figure 12 we assumed a 3 locus "genotype" in which each locus has 5 alleles. This generates a 125 (5^3) member sequence space. Each locus was given an identical allele-to-allele transition matrix shown in figure A4.1. Three different profiles for the fitness surface $\{F_i\}$ were used in the three simulations A, B, and C, as shown in Figure A4.2.

Starting with some arbitrary genotype we propagated equations (A4.1) – (A4.4) to generate an initial population of genotypes within that sequence space. This population represented the population of genotypes found in an initially infected node (the "infector"). We then randomly selected a genotype from that population and used it as the initial genotype for a second propagation representing infection of the second node (the "infectee"). The probability of selecting a seed genotype was weighted according to the fraction x_i of each genotype in the infector population. In each case, propagations were performed for 30 generations. This was generally far short of the number of generations needed to achieve the asymptotic "quasispecies population".

	a1	a2	a3	a4	a5
a1	9.70E-01	3.00E-02	0.00E+00	0.00E+00	0.00E+00
a2	3.00E-02	9.40E-01	3.00E-02	0.00E+00	0.00E+00
a3	0.00E+00	3.00E-02	9.40E-01	3.00E-02	0.00E+00
a4	0.00E+00	0.00E+00	3.00E-02	9.40E-01	3.00E-02
a5	0.00E+00	0.00E+00	0.00E+00	3.00E-02	9.70E-01

Figure A4.1. Transition matrix from allele state to allele state at each locus. The genotype to genotype transition matrix Q_{ij} is generated using the function $Q_{ij} = \prod_l \gamma(a(l,I),a(l,J))$ where a(l,J) is the index of the allele state associated with locus l in genotype J.

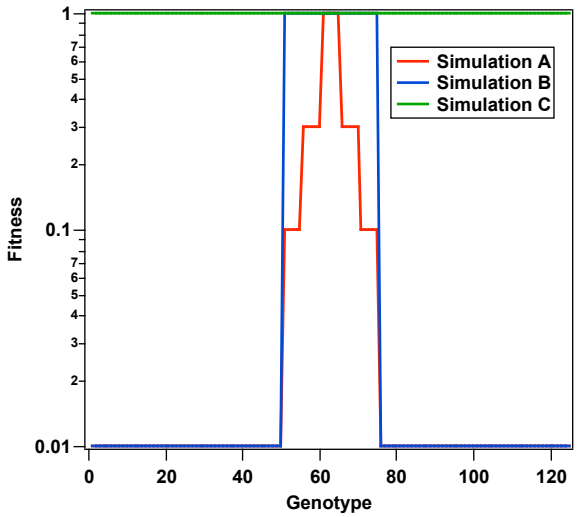


Figure A4.2. Fitness surfaces used in simulations A, B and C. The abscissa represents genotypes 1 – 125 in numerical order, beginning with genotype (1,1,1) and ending with genotype (5,5,5).

Two hundred (200) random samples were generated from the initial population, and propagated to obtain that number of "infectee" populations. For each population so generated, the consensus allele at each locus was generated by adding together all x_i such that genotype i had allele l at that locus and then choosing the allele with the largest sum. The allele based Hamming distance between the consensus sequences of the initial

"infector" population and the 200 "infectee" populations was calculated as the sum of the absolute difference between allele states (1-5) over the three loci:

$$\delta = \Sigma_l |a_{0l} - a_{fl}| \qquad \text{(A4.5)}$$

In this equation, a_{0l} and a_{fl} are the allele states at locus l for the consensus sequences of the infector and infectee populations, respectively.

Consider the set of all genotypes generated by a multiple locus, multiple allele model of a microbial genome. Equations A4.1 represent a mapping from this space of all genotypes $\{g_0\}$ to a smaller subset of consensus genomes $\{g_c\}$ (which are, of course, members of the original set.) Consider the function

$$\mathcal{F}(g_c|g_0,N_{gen},\mathbf{F},\mathbf{Q}) = \{1 \text{ if } g_0 \rightarrow g_c,\ 0 \text{ otherwise}\} \qquad \text{(A4.6)}$$

Here **F** and **Q** represent the fitness surface and transition probability matrix respectively. For the simulation described above, the probability of observing an infectee with a consensus genome g_c, when his infector hosts a population of genotypes with a probability distribution $P(g_0)$ is:

$$P(g_c) = \sum \mathcal{F}(g_c|g_0,N_{gen},\mathbf{F},\mathbf{Q})\, P(g_0) \qquad \text{(A4.7)}$$

where the sum is over all g_0. Since each g_0 maps uniquely to one g_c, we have the formal normalization condition

$$\sum \mathcal{F}(g_c|g_0,N_{gen},\mathbf{F},\mathbf{Q}) = 1 \qquad \text{(A4.8)}$$

Where the sum is over g_c. This ensures that $P(g_c)$ is normalized if $P(g_0)$ is. Thus, if we had an efficient way to calculate $\mathcal{F}(g_c|g_0,N_{gen},\mathbf{F},\mathbf{Q})$ then (A4.7) would provide an analytic solution that could replace our simulation.

In the case of a flat fitness surface (F_i = constant for all i) simulations imply that the consensus sequence of the population generated by any input sequence g_0 will be equal to g_0. It is likely that this can be proven rigorously by consideration of the properties of $\mathcal{F}(g_c|g_0,N_{gen},\mathbf{F},\mathbf{Q})$. When the fitness surface has a peak at g_p, and we allow N_{gen} to become very large, we expect $P(g_c) \approx 1$ for $g_c = g_p$ and negligeable otherwise. This observation was made previously by Jenkins, et. al. [Mol. Biol. Evol. **18**(6):987-994 (2001)].

Appendix 5.
$\mathcal{P}$(M|G) for the truncated Galton-Watson model

The Galton-Watson process is a simple way to generate transmission trees which mimic the form of real disease transmission trees in many ways. For Galton-Watson trees, the numbers of daughters d of each node are statistically independent but obey the same probability distribution P_d. In the absence of a constraint on tree size, this model defines an infinite collection of possible trees some of which may be infinite in extent if the probability parameters meet certain conditions. Here we are primarily concerned with trees constrained to have a fixed number of generations G, so that all trees have a fixed number of nodes. Under the assumption of statistical independence of the nodes, determination of the statistics of the path length M between any two nodes is straightforward and avoids the combinatorial explosion that might be feared of a collection of large trees.

Consider first the average number of nodes in such trees. Averages over a statistical ensemble of trees are indicated by *italics.* The average number of daughters of a given node is:

$$d = \Sigma_d \, d \, P_d \qquad \text{(A5.1)}$$

A node with d daughters has s=d(d-1)/2 sibling pairs. Suppose them to be ordered with older sibling first. The average number of ordered pairs is

$$s = \Sigma_d \, d(d-1)/2 \; P_d \qquad \text{(A5.2)}$$

This statistic is an important determinant of the path-length distribution.

Each tree has a single root node. The number of generations g to a nodes is its path length from the root node. The expected number of nodes at generation g is d^g; so the expected total number of nodes for a tree with a maximum of G generations is

$$N = \Sigma^G_{g=0} \, d^g = \; (1 - d^{G+1}) \, / (1-d) \qquad \text{(A5.3a)}$$

This holds whether $d<1$ (subcritical), or $d>1$ (supercritical). We shall use the 'average for unconstrained trees' N_u for the term $(1-d)^{-1}$ when discussing trees with fixed G.

Two distinct nodes A, B have a unique nearest common ancestor C. If neither A nor B are C, they are elements or descendents of exactly one sibling pair S of C. Let the pair (A,B) be ordered such that either A is C or A is a descendent of the elder sister in S. The path length M is the sum of the lengths a and b of A and B from C. These can be related to the generations of the nodes

$$M = a + b \qquad \text{(A5.4a)}$$

$$a = g(A) - g(C) \tag{A5.4b}$$

$$b = g(B) - g(C) \tag{A5.4c}$$

If the number of generations has a limit G, the average number of paths for a common ancestor of generation g is

$$n_p(M,g) = s\, d^{M-2}\,(M-1) + d^{M} \quad : M \le G-g \tag{A5.5a}$$

$$= s\, d^{M-2}\,[2(G-g)-M+1] \quad : G-g < M \le 2(G-g) \tag{A5.5b}$$

(A5.5a) is equivalent to (6). (A5.5b) results from the constraints $g+a \le G$, $g+b \le G$. The average per tree is the sum over generations g, each with d^g average members:

$$N_p(M) = \Sigma^{G-1}_{g=0}\, n_p(M,g)\, d^g \tag{A5.6}$$

While (A5.5) is simple to evaluate numerically, a closed-form solution exists. The series contains only terms proportional to d^g and $g d^g$, so can be evaluated using (8), although the non-analytic behavior of $n_p(M,g)$, evident in (A5.6), demands patience. The result is

$$N_p(M) = sN_u\{d^{M-2}(M-1) + d^{G+h-1}\,[2dN_u(1-d^{-h})+e]\} + N_u(d^{M}-d^{G+1}) : M \le G \tag{A5.7a}$$

$$N_p(M) = sN_u\{d^{M-2}[2(G-dN_u)-M+1] + d^{G+h-1}(2dN_u+e)\} \quad : G < M \le 2G \tag{A5.7b}$$

$$e(M) = \{1: M \text{ even} ;\ 0: M \text{ odd}\} \ ; \quad h(M) = \text{gif}(M/2) \ ; \ N_u = (1-d)^{-1} \tag{A5.7c}$$

The non-analytic functions e(M), h(M) in (A5.7c) are required due to the kinked nature of the paths: they fit within the constrained number of generations more easily for even M. gif(x) is the greatest integer less than or equal to x.

The probability distribution $\mathcal{P}(M|G)$ can be found by normalizing $N_p(M)$ by its sum, as in (7). An exact closed-form representation should exist, but is unlikely to be simple, or more insightful than (5.7), which gives the shape of $\mathcal{P}$.

$N_p(M)$ has a maximum for $M_m < G$. In the limit of large M, N_u the dominant terms of (A5.7a) are

$$N_p(M) \sim s\, N_u d^{-2}\{\, M\, d^{M} - 2N_u d^{G+2}\,(1-d^{M/2})\} \tag{A5.8}$$

Therefore, the maximum probability occurs at M_m for which

$$(d/dM)\, N_p(M) = sN_u d^{M-2}\,\{M \ln d + 1 + D\, d^{G+2} \ln d\, d^{-M/2}\} = 0 \tag{A5.9}$$

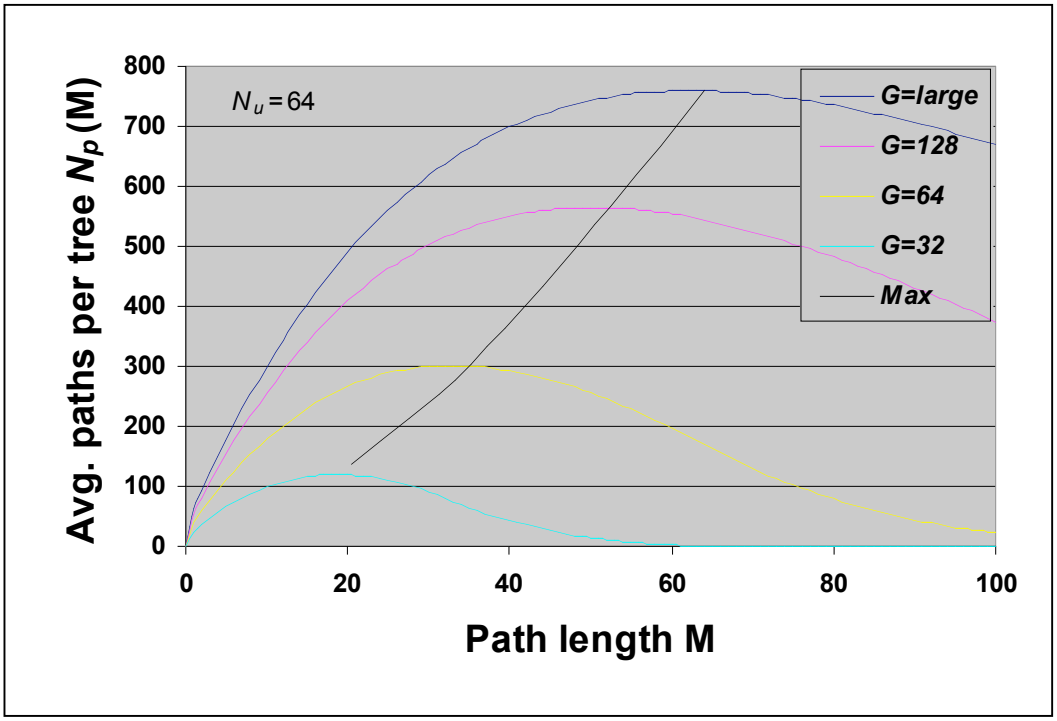
With $d^{-M/2} \sim 1 - \tfrac{1}{2} M \ln d$ and $\ln d \sim d - 1$, this has the approximate solution

$$M_m \sim (1-d^{G+2})/(1-d)(1+\tfrac{1}{2}\, d^{G+2}) = (N+d^{G+1})/(1+\tfrac{1}{2}\, d^{G+2}) \qquad \text{(A5.10a)}$$

$$N = 1+d+\ldots+d^{G} = (1-d^{G+1})/(1-d) \qquad \text{(A5.10b)}$$

Thus, for large subcritical trees, the maximum occurs for the number of nodes N reduced by the factor $(1+\tfrac{1}{2}\, d^{G+2})$.

These results are illustrated in Figure A5.2, where we plot $N_p(M)$ for Poisson trees with N_u=64, constrained by several values of G. Equation (A5.10) is seen to be a good approximation for the positions of the maxima M_m. For example, with G=32 the expected number of nodes is N=26 by (5.10b), but the maximum is in fact somewhat less than the value M_m=20.5 of (5.10a).



We see that the path length distributions are substantially altered by the constraint on number of generations, even, in the subcritical case, for G > $2N_u$ – twice the expected number. (Since d~1, the expected numbers of nodes and generations are similar.) This implies that relatively rare trees with large numbers of nodes contribute substantially to these distributions. Perhaps this should not be surprising given the inherently quadratic dependence of node pairs on nodes, but it is still worth remembering that mean quantities may be misleading.

These examples illustrate a prominent feature of Galton-Watson trees: the path length distribution peaks at a length commensurate with the number of nodes in the tree. They also show that the statistical simplicitly of these trees makes mathematical analysis of them tractable. The computer simulations which we performed were intended simply to

verify analytic results in simple test cases. Indeed, it would be difficult to obtain our results for large trees computationally.

Appendix 6.
Estimating P(δ|M) from empirical or simulated disease transmission data

Imagine that we have a transmission tree and genetic data for each node. Ideally, this data would originate from a well-designed and executed study of a real outbreak. However, it could also come from a simulation. We can define two matrices, the $\{M_{ij}\}$ matrix which has the known path-length distance between each pair of isolates, and the $\{\delta_{ij}\}$ matrix with the corresponding genetic distances. The empirical transmission tree is a single sample of the population of trees, and can be used to provide an estimator for $\mathcal{P}(M)$. In this sense, when we draw a sample ij from the $\{\delta_{ij}\}$ matrix we are basically sampling P(δ,M) for that outbreak (or simulation). Thus, a histogram of δ for all the pairs that had M = 1, is an estimator of the joint probability distribution P(δ,M=1). Similarly, a histogram of δ for all pairs that had M>1 is an estimator for P(δ,M>1).

By the same reasoning, the estimators for $\mathcal{P}(M=1)$ and $\mathcal{P}(M>1)$ are:

$$\mathcal{P}(M=1) = (\#\text{ of matrix elements for which } M_{ij} = 1)/(\text{total \# of matrix elements}) \quad \text{(A6.1)}$$

and

$$\mathcal{P}(M>1) = \#\text{ of matrix elements for which } M_{ij}>1/\text{total \# of matrix elements} \quad \text{(A6.2)}$$

To obtain estimations of the conditional probability distributions that can be used to generate a ROC curve we simply use the relations:

$$P(\delta|M=1) = P(\delta,M=1)/P(M=1) \quad \text{(A6.3)}$$

and

$$P(\delta|M>1) = P(\delta,M>1)/P(M>1) \quad \text{(A6.4)}$$

This procedure provides an exact estimate of P(δ|M) for any M if all the data is "fully connected" i.e. each node for which there is genetic data is part of a fully connected transmission graph. In reality, the empirical data typically obtained from an outbreak (say for the Singapore SARS outbreak) is fragmented, i.e. there are isolates from isolated pairs of direct transmission-related nodes and larger tree fragments, and isolates from a some number of isolated nodes with no known connectivity to the other nodes. We generally also have some larger transmission sub-trees determined by contact tracing, but genetic data for only a few nodes from these trees, without necessarily knowing which nodes on the tree they are from. Currently there is no statistically valid (even approximate) method for "correcting" real outbreak data for this problem in order to directly estimate P(d|M) for arbitrary M. Strictly speaking it is *not* even valid to simply

plot histograms of the δ values obtained from pairs of nodes for which M≠1, and expect to generate an accurate estimate of P(δ|M) when the outbreak data is fragmented. This is because generally there is no way to verify that the existing samples are a representative sample from the true outbreak distributions.

However, it is valid to plot δ values obtained from the M≠1 pairs to get P(δ|M>1), if the set of nodes represents a random sample of possible nodes in the network, and the network is large enough to preclude getting a M=1 pair by accident. Thus, for example, the HIV data set from Trask, et. al. cited in section 2 consists of 63 husband-wife pairs drawn from a larger network of HIV infectees within Zambia. Since these pairs were apparently drawn at random from the larger network, and the network is (evidently) much larger than the sample, the chance of a given wife (husband) being the infector or infectee of another husband (wife) in the sample set is low. Under these circumstances we can construct a reasonable estimate of P(δ|M>1) by pairing each female sample with every other male sample besides her husband.

If simulations of genetic evolution are performed on realistic outbreak transmission trees, the genetic difference data can be used to generate valid estimates of ROC curves and posterior probabilities since all the nodal relationships are known. Of course, the accuracy of these estimates with respect to real outbreaks depends on how accurately the model represents transmission, growth, and mutation of the microbial populations.