Stephan P. Velsko
Joanne Osburn
Jonathan Allen

Lawrence Livermore National
Laboratory, Global Security
Directorate, Livermore, CA, USA

Received April 16, 2014
Revised July 26, 2014
Accepted August 1, 2014

Research Article

Forensic interpretation of molecular variation on networks of disease transmission and genetic inheritance

This paper describes the inference-on-networks (ION) framework for forensic interpretat ION of molecular typing data in cases involving allegations of infectious microbial transmission, association of disease outbreaks with alleged sources, and identifying familial relationships using mitochondrial or Y chromosomal DNA. The framework is applicable to molecular typing data obtained using any technique, including those based on electrophoretic separations. A key insight is that the networks associated with disease transmission or DNA inheritance can be used to define specific testable relationships and avoid the ambiguity and subjectivity associated with the criteria used for inferring genetic relatedness now in use. We discuss specific applications of the framework to the 2003 severe acute respiratory syndrome (SARS) outbreak in Singapore and the 2001 foot-and-mouth disease virus (FMDV) outbreak in Great Britain.

Keywords:
Forensic / Genetics / Inference / Networks / Transmission
DOI 10.1002/elps.201400205

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Molecular typing data are often used as evidence in investigations of deliberate or negligent transmission of an infectious microbe. A recent example is the 2010 cholera outbreak in Haiti, where DNA sequence data have been cited as evidence supporting the hypothesis that United Nations troops introduced the disease [1]. In such cases a high degree of similarity between the molecular characteristics of microbial isolates sampled from victims and those from the putative source is usually assumed to support the source hypothesis. Similarly, in human immunodeficiency virus or Hepatitis C Virus transmission cases the degree of similarity of the sequences from victim and (putative) source relative to a set of "background" isolates is proffered as evidence favoring or excluding the transmission hypothesis [2, 3].

However, several authors have pointed out that the interpretation of phylogenetic findings as evidence to support a hypothesis about a disease transmission event is subject to many caveats [4, 5]. For example, phylogenetic construction alone cannot assess the probability that two isolates come from a common unidentified source of infection, or are separated by one or more unknown intermediate infected hosts. Obviously, if unknown or unsampled source candidates exist, phylogenetics cannot exclude them. In many investigations it is not possible to identify all potential sources with certainty, or the relevant isolates and their genetic sequences may not be available. This limitation has led to highly precautionary guidelines about the use of microbial phylogenetic evidence in criminal prosecutions and to restrictions on the language of admissible testimony [6]. Thus, how to quantify and express the degree of support that molecular comparisons provide for a source-transmission hypothesis remains a central, yet unresolved question [7].

A closely related application of molecular typing in infectious disease epidemiology is deciding whether an isolate can be associated with a cluster of related cases, i.e. if an observed case of infection "belongs" to a given outbreak. Tenover introduced a set of heuristic criteria in the context of RFLP typing, based on the number of mutational differences among questioned and reference isolates [8]. Epidemiologists using other molecular typing methods typically use some variant of these

Correspondence: Dr. Stephan P. Velsko, Lawrence Livermore National Laboratory, 5508 East Avenue, Livermore, CA 94550, Mail Stop: L-177, USA
E-mail: Velsko2@llnl.gov
Fax: 925-422-4100

Abbreviations: FMDV, foot-and-mouth disease virus; **ION**, inference-on-networks; **SARS**, severe acute respiratory syndrome

Colour Online: See the article online to view Fig. 4 in colour.

"Tenover criteria" to judge whether an infection can be assigned to an ongoing outbreak, or is a sporadic case [9].

However, the arbitrary nature of this approach is unsatisfying: Tenover himself recognized that the interpretation of strain typing results within this framework is a subjective process, based on experience and intuition. In the nearly 20 years since Tenover's seminal paper, the technologies for typing bacteria have evolved substantially, permitting much higher resolution, with the concomitant ability to elucidate more detailed questions about the evolutionary relationships between isolates in an outbreak. But progress toward acquiring a more rigorous answer to Tenover's question has not advanced significantly.

In this paper we outline a framework for performing genetic inference that is based on explicit hypothesis testing of relationships defined on networks of disease transmission and genetic inheritance. The framework provides an analogue to a forensic "match probability"—a quantitative probability estimate for the hypothesis that two microbial sequences are linked by a direct disease transmission event. The framework also addresses in a transparent way whether an isolate "belongs" to a given outbreak, replacing arbitrary qualitative judgments with an explicit probability expression. Such estimates can only be made if the statistical properties of disease transmission networks are taken into account. While approximate, this framework provides an objective way to assess the inferential power of molecular typing results, and increases the rigor and transparency of forensic testimony offered in either a legal or a national security forum.

2 Materials and methods

2.1 The ION framework

The fundamental concept that underlies the inference-on-networks (ION) approach is that genetic lineages are constrained to run along the vertices of a transmission network, and genetic material that is the object of forensic analysis is sampled from nodes in that network. For infectious diseases, the nodes are infected individuals (and the genetic material is that of the infecting organism), while for mitochondrial DNA and Y-chromosomes, the nodes are individuals (viewed as colonies of somatic cells). Figure 1 shows a portion of such a transmission network.

Note that each pair of nodes in a transmission tree like Fig. 1 is connected by an M-step transmission relationship. For example, the two nodes marked with asterisks in Fig. 1 are separated by $M = 7$ steps. We do not distinguish direction of transmission when calculating node-to-node distances. The timing of infections and other contextual information usually indicates the direction of transmission without ambiguity. Under these conditions the assertion that one node is the source of the genetic material found in a second node is equivalent to asserting that the two nodes are separated by $M = 1$ transmission steps.

Regardless of whether we are discussing disease transmission or the inheritance of mtDNA or Y-chromosomes, a transmission event that generates each new node in this network represents a case where a relatively small amount of genetic material is sampled at random from the source node, then transferred to the receiving node, where it creates a new and larger population of sequences. For any pair of isolates sampled from two different nodes in the network, we can define some metric δ that characterizes their degree of similarity. The ION method assumes that we can infer M_0, the number of steps that separate these two nodes, from the observed value of δ by utilizing two empirically derived statistical distributions. The first is P(δ|M), the probability that the sequences of two isolates taken from nodes M steps apart will differ by δ. The second is P(M), the prior probability that two nodes chosen at random from the network will be separated by M transmission steps. Given these distributions it is a straightforward application of Bayes's theorem to show that

$$P(M \leq M_0|\delta) = \left[1 + \frac{\sum_{M>M_0} P(\delta|M)P(M)}{\sum_{M\leq M_0} P(\delta|M)P(M)}\right]^{-1}. \quad (1)$$

When $M_0 = 1$, Eq. (1) provides the probability of a direct transmission relationship as a function of the genetic distance between two isolates. Inferences about other transmission hypotheses (e.g. $M = M_0$ rather than $M \leq M_0$) are easily derived as well. In Section 3 we will apply Eq. (1) to infectious disease outbreaks. The application of the ION framework to inferences involving mtDNA or Y chromosomes is provided in the Supporting Information File SM1. All calculations were performed on a laptop computer using Excel.

The ION approach can be used with a variety of methods for characterizing genetic sequences, as long as the same method is used consistently. In general, we assume that the chosen characterization method can be used to generate

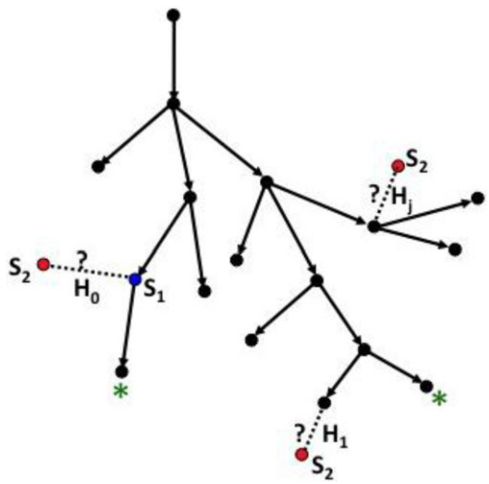

Figure 1. A notional transmission tree. Each node (dot) represents an individual, and S_1 and S_2 are genetic sequences obtained from isolates that come from two nodes of interest (e.g. two infected persons), marked in blue and red, respectively. H_0, H_1, and H_j represent different hypotheses about the source of the genetic material found in the red node. Asterisks mark two nodes separated by seven transmission steps on this tree.

phylogenetic comparisons among isolates, and that **δ** can be assigned to a pair of isolates by summing branch lengths to their common ancestor. Less rigorous genetic distance metrics like simple numbers of mutational differences can also be used when appropriate. However, we assume that horizontal gene transfer has a negligible effect on the population of genotypes over the network size (or duration of the outbreak) of interest. Hence there are obvious limitations to the accuracy of such a simplified approach when attempting to infer genetic relationships on very large "historical" transmission trees.

It is not necessary to know the actual transmission tree or have an extensive set of reference samples drawn from many nodes in the tree to apply the ION framework. In most cases of disease transmission the actual tree that connects nodes associated with an outbreak is not known with certainty, although portions of it may have been inferred from epidemiological studies. In the next sections we will outline some simple empirical methods for estimating the required distributions.

2.2 Estimating P(δ|M)

The most direct way to determine P(**δ**|M) is by comparing the sequences of isolates drawn from pairs of nodes whose transmission relationship is known. For infectious disease outbreaks the selection of such a reference set requires that transmission relationships have been determined with high confidence by epidemiological investigation. Recently several papers have developed methods for integrating genomic sequence data with contact tracing and timing information to infer more accurate transmission trees than can be determined by contact tracing alone [10, 11]. Exhaustive genetic sampling of a large proportion of infected hosts in an outbreak is generally impractical, but these papers show that a small segment of the complete outbreak tree can be studied this way. In any case, techniques borrowed from these somewhat complex and data-demanding tree reconstruction methods can be used to strengthen the selection of a reference set of isolates.

Generally, we expect ION to be most accurate when the P(δ|M) estimated from a carefully studied portion of an outbreak is used to estimate the genetic relatedness between isolates from another, less well characterized portion of the same outbreak where forensic questions are of interest. However, the reference set for ION does not have to encompass an entire connected transmission tree. In fact, only a set of $M = 1$ related pairs is needed. This convenient simplification is implied by the standard theories of genetic change used to construct phylogenetic relationships [12]. 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 distributed as P(δ|$M = 1$) [13]. Distributions consistent with this constraint have a functional form such that if P(δ|$M = 1$) = f(γ), where γ is proportional to the average number of mutations observed between isolates when $M = 1$, then P(δ|M) = f(Mγ). Distributions for discrete random variables such as the Poisson, generalized Poisson, and negative binomial have this property, and can be used to infer P(k|M) from P(k|$M = 1$) when phylogenetic branch lengths can be approximated by the observed number of mutational changes.

While the Poisson model for P(k|M) is attractively simple, especially when there are only a few reference pairs available, distributions with "fatter tails" might be more accurate representations in some cases. Generally, δ is a stochastic variable governed by a probability distribution

$$P(\delta | M; t_1, t_2, \mathcal{N}),$$

where M is the number of transmission steps separating the two nodes, 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, typically unknown stochastic variables in an actual outbreak. $\mathcal{N}$ represents the number of generations of replication that has taken place between infection of node 1 and the transmission event to node 2. Clearly M and $\mathcal{N}$ are roughly proportional on average, although $\mathcal{N}$ itself is a stochastic variable. The ION approach simply assumes that our inferences can be based on empirical approximations to P(δ|M) in which t_1, t_2, and $\mathcal{N}$ have effectively been "averaged out" as nuisance variables [14]. The effect of this averaging is to favor overdispersed models like the negative binomial or generalized Poisson. These distributions have one more parameter than the Poisson, and thus require larger datasets to drive down the relative uncertainty in their parameter values. Therefore, in the face of small datasets we have adopted the practice of using a Poisson model if this hypothesis cannot be rejected by a standard chi-squared test.

A less direct method of estimating P(**δ**|$M = 1$) is to use results from animal passage experiments. Of course, this presupposes that the disease in question has a well-understood laboratory animal model, and that the experiments replicate the important features of the actual host–host transmission process found in nature. It also may be possible to use mutation rates determined by in vitro serial transfer experiments. This approach has been used as the basis for phylogenetic inferences about pathogens in the past [15].

2.3 Estimating P(M)

Like the branch length metric δ, 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 and environmental factors. 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 **P**(M) that any two nodes drawn randomly from the tree will be related by M steps is defined on this ensemble of possible trees.

Imagine a set of outbreaks in which the same number of people (or animals) were infected, but which otherwise evolved independently according to the characteristics of the disease in question. Each outbreak would generate a different transmission tree. An estimate for P(M) can be calculated from each tree by any algorithm that counts the number of steps between each unique pair of nodes on a finite network, then normalizes the resulting histogram by the total number of nodal pairs. For example, the number of paths of length M among the set of nodes can be determined by using a result from graph theory that relates this quantity to the number of unit matrix elements found in successive powers of the adjacency matrix [16]. The observed variations in **P**(M) from tree to tree can be considered sampling errors about some most likely distribution that characterizes trees for outbreaks of that particular disease and that number of nodes. Extensive computer simulations of transmission trees have shown that the variance in P(M) become small for trees larger than about 20 nodes, so that Eq. (1) is rather insensitive to the actual tree. Moreover, for large trees Eq. (1) also becomes independent of the number of nodes. Some ION problems are more conveniently solved by using an ensemble of trees that span a given number of generations, rather than a given number of nodes.

The shape of P(M) does depend on the transmission network connectivity. For example, outbreaks with a large number of "superspreading" events where one infected node generates a large number of secondary infections will differ from those where such events are rare, and this effect can change the calculated posterior probability $P(M \leq M_0|\delta)$, although we have found this effect to be modest in practice.

One result of the insensitivity to actual tree size and branching is that Eq. (1) will give reasonable estimates if we use P(M) distributions derived from actual empirical transmission networks that have been deduced from epidemiological contact tracing. We will show examples of this in Section 3. It is somewhat remarkable that transmission networks observed in independent sections of the same large outbreak give very similar values of $P(M \leq M_0|\delta)$. We have also found that similar results are obtained if we use an analytical functional form for **P**(M) whose parameters have been fit to data from prior outbreaks of the same disease [17].

P(M) could also be derived from simulations of outbreaks on a social contact network that has been developed for epidemiological prediction purposes (for a particular disease). Elaborate models for disease transmission networks have been constructed to investigate outbreak dynamics and the effect of control measures for both human disease transmission and zoonotics in networks of animal hosts [18, 19]. Social contact networks are relatively stable but flexible descriptors of the modes and mechanisms of disease transmission and can easily be stored as reference data. For outbreaks involving animals this may be the only practical method of estimating **P**(M).

2.4 Assigning an isolate to an outbreak

In the IONs framework, known outbreaks of infectious disease are simply regarded as "local" portions of a larger "global" transmission tree that includes (largely unknown) reservoirs and other outbreaks and is extended in geography and time. Thus, deciding if an isolate is part of a given outbreak is equivalent to deciding if it was likely to have been sampled from a node in the local tree. This probability is easily calculated from Eq. (1).

In tree-like networks there is only one path connecting any two nodes [16]. The diameter of a transmission network is defined to be the length (in number of steps) of the longest path found among the set of nodal pairs belonging to that network. It is easy to see that the maximum possible 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, $\mathbf{P}(M) = 0$ for $M > 2G$. Note that in Eq. (1) we find that $P(M \leq 2G|k) = 1$, independent of k, for this reason.

Suppose we have sampled one or more reference isolates from nodes known to be part of a "local" outbreak tree that encompasses G_{loc} generations. Consider a questioned isolate that differs by k mutations from the genetically closest reference isolate (relative to the chosen sequencing or typing scheme). $P(M \leq 2G_{loc}|k)$ is the probability that the questioned isolate was sampled from a node in the outbreak, when we use P(M) for the larger "global" transmission network of which our "local" outbreak was a part.

Typically we do not know the global transmission tree, so it is necessary to use simulations or modeling to infer P(M), assuming some value G_{glo} for the number of generations in the global tree. Fortunately, as long as $G_{glo} \gg G_{loc}$, the precise number of generations used to determine P(M) is not critical. This is illustrated in Fig. 2, which shows how $P(M \leq 5|k)$ stabilizes after $G_{glo} \approx 10$. In addition, simulated trees with a fixed number of generations provide reasonable estimates as shown in Supporting Information Fig. S1.

Many infectious disease transmission networks exhibit both small world behavior and superspreader clusters [20]. In a finite-sized transmission network, increasing the probability of finding nodes that infect large numbers of recipients reduces the probability of observing pairs of nodes connected by a large number of steps and shrinks the right tail of P(M). This increases the likelihood that two randomly selected nodes are related by a smaller number of transmission steps than intuition might suggest.

3 Results and discussion

3.1 The SARS outbreak of 2003

The severe acute respiratory syndrome (SARS) outbreak of 2003 can be used to illustrate the use of our framework in the context of respiratory infection epidemiology. Several papers have discussed the epidemiological linkage among a

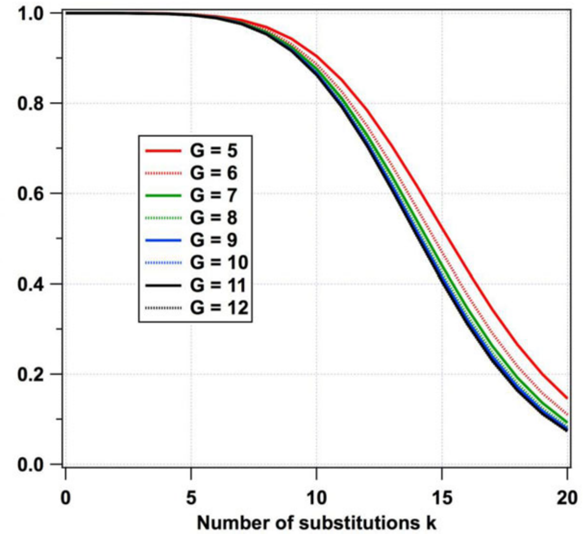

Figure 2. Calculations of P(M ≤ 5|k) for networks with successively larger numbers of generations G. This shows that as long as G > 2M, the precise number of generations used to determine P(M) is not critical.

set of SARS patients associated with the outbreak in Singapore [21–23]. Whole genome sequences of SARS coronavirus isolates were obtained from these patients, and combinations of phylogenetic analysis and contact tracing have been used to generate conflicting putative transmission relationships. This provides a useful, if imperfect dataset for illustrating the methods described in Section 2.

Only a few, if any, linked transmission pairs among SARS patients have been identified with high confidence. However, we can assume a reference set based on four direct transmission pairs identified by contact tracing [20, 21]. The sequence accession numbers, patient (isolate) identifiers, and the cited transmission partners are provided in Supporting Information Table ST1. As indicated in Section 2.3, when there are only a few reference data points, we assume a Poisson distribution unless a simple chi-squared test allows us to reject it. With only four reference pairs, a Poisson distribution is assumed for P(δ|M) with δ = k, the number of substitutional differences between sequences:

$$P(k|M) = \frac{(\gamma\, M)^k}{k!} e^{-\gamma\, M}. \tag{2}$$

The average number of substitutions observed for the four reference isolate pairs provide the estimate γ = 3.0 ± 0.9.

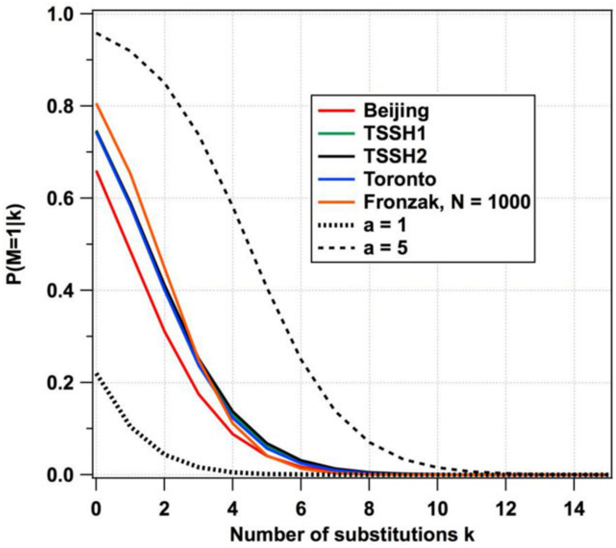

Figure 3. Estimated posterior probability that a pair of SARS isolates arose from direct transmission given that their sequences differ by k substitutions. Solid lines—using the indicated empirical or model P(M) distribution and the Poisson distribution for P(k|M) with γ = 3; Broken lines—calculated with γ = 1 and γ = 5, respectively, with the P(M) for outbreak TSSH1.

To estimate P(M) we turn to the empirical transmission networks that several studies have produced from epidemiological contact tracing. Four of these networks (which are subnetworks of the global SARS transmission network) are listed in Table 1, along with some parameters that describe them [24–26]. Each tree has a different number of nodes (infected patients) and spans a different number of generations. Besides these two characteristics, the detailed form of **P**(M) for each outbreak also depends on the number of "superspreaders" (defined as patients who infect more than five other patients), and the size of the superspreading clusters.

The calculated direct-transmission probability P (M = 1|k) for the various P(M) exemplars is shown in Fig. 3, which demonstrates an important feature of the ION approach to microbial genetic inference that differs from phylogenetic source inference methods. The posterior probability P(M = 1|k) is the apparent empirical probability that isolates from two nodes on the network differing by k substitutions are separated by only one transmission step. Note that an exact match (k = 0) does not imply with certainty that two isolates are related by direct transmission (i.e. P(M = 1|k = 0) ≠ 1 in general). In fact, Fig. 3 implies that

Table 1. Properties of some reported subtrees of the complete SARS transmission tree

Outbreak	Number of nodes	Number of generations	Number of "superspreaders"	Largest superspreadng cluster size	Diameter
TTSH1 [24] (Singapore)	41	4	1	22	7
TTSH2 [24] (Singapore)	36	3	1	21	6
Toronto [25]	72	5	3	16	6
Beijing [26]	69	3	4	33	5

there is a reasonable chance that more than one transmission step separates two isolates even if the sequences are identical (k = 0). (However, it is not true that two isolates separated by a single transmission step are likely to exhibit no mutational differences: $P(k = 1|M = 1) \approx 0.05$). Conversely, as illustrated by the curve for $\gamma = 5$, if the mutation rate is high enough a large mismatch between the two sequences ($k \neq 0$) may still imply a high probability that the isolates are related by direct transmission.

The results in Fig. 3 also demonstrate the basic insensitivity to variation in **P**(M). The curves for TSSH1, TSSH2, and Toronto closely overlap. The Beijing network, with its larger number of superspreaders and concomitant lower ratio of generations to nodes deviates noticeably from the others, but is not qualitatively different. Also shown is a calculation using a theoretical P(M) distribution ("Fronczak," [17]) for a SARS-like transmission network with 1000 nodes, suggesting the relative insensitivity of Eq. (1) to network size.

Jombart et al. have applied a Bayesian approach to transmission network reconstruction to the Singapore SARS data [11]. Supporting Information Table ST1 lists a set of putative transmission pairs predicted from their calculations. (It should be noted that their calculation does not agree with contact tracing findings for one of our reference pairs [11].) The posterior probabilities calculated using our method and shown in Supporting Information Table ST1 are much lower than those quoted in [11] for many of the pairs implying that for those pairs it is much more likely than not that transmission was through at least one intermediate person.

3.2 The UK FMDV outbreak of 2001

Networks of disease transmission often extend over large spatial regions and have long durations. In such situations, subnetworks of infected individuals within cities, herds, flocks, and other social groupings are sometimes considered the infected "nodes" of a more coarsely scaled network. Each node defined this way is itself a transmission network connecting individuals, but this intranode structure is ignored.

Cottam et al. performed an analysis on data from the 2001 foot-and-mouth disease virus (FMDV) outbreak in Great Britain based on such a "rescaled" transmission network consisting of a set of 20 farms [27]. Cottam used a combination of phylogenetic and event-timing data to infer a most likely transmission network then calculated the number of variant nucleotides between pairs of sequences representing farm–farm transmission events based on the inferred network. Within the ION framework this is less ideal than having transmission pairs identified by contact tracing alone because genetic data is thus "counted twice" when we infer P(k|M). However this dataset suffices to illustrate certain points of interest.

Cottam obtained consensus viral sequences from single isolates from each farm. Simple statistical tests indicate that the Poisson hypothesis for the number of nucleotide differences between pairs of sequences cannot be rejected. Therefore we assumed Eq. (2) was valid, and used Cottam's value of $\gamma = 4.3$ for the average number of substitutions per farm-farm transfer.

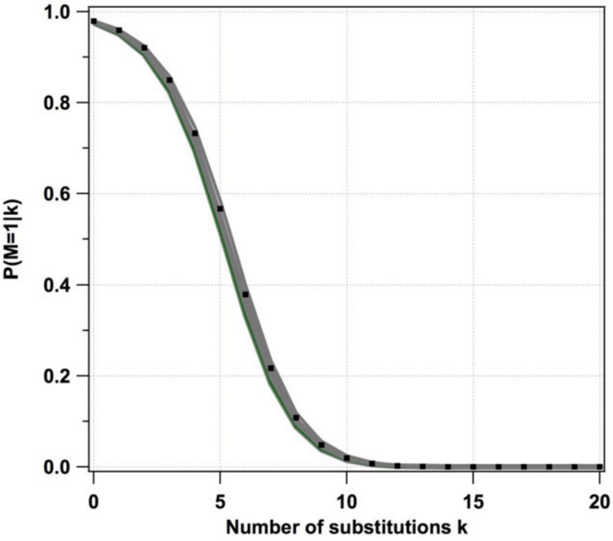

Figure 4. Predicted posterior distribution $P(M = 1|k)$ based on data from [25]. Gray curves are based on 20 randomly generated transmission trees; Black points are based on the tree published in [25].

Rather than using Cottam's transmission tree as an exemplar for estimating P(M), we generated a random set of transmission networks that had the same degree distribution as the network inferred by Cottam. This also avoided some of the circularity that might arise because genetic and epidemiological evidence was already combined in constructing Cottam's tree. Our sample trees ranged in size from 12 to 169 nodes, representing up to six generations of transmission. It should be noted that over 2000 farms were involved in the actual outbreak, but our largest tree size was limited by the computational power of Excel running on a laptop. Each P(M) was used to calculate a separate posterior distribution $P(M = 1|k)$, and the results are shown in Fig. 4, along with the result when Cottam's tree is used as an exemplar of the outbreak. The close similarity of all of the curves shows the basic insensitivity of the posterior distribution to the size of transmission networks when they are generated by similar degree distributions.

Cottam assigned a likelihood to each putative transmission link based on data for the onset and duration of the infection at each node. In Table 2 we compare Cottam's likelihoods with our posterior probability estimates. Although these quantities have different interpretations, it is convenient for discussion to define the cases when both quantities are simultaneously greater or less than 0.5 as "agreement" and cases where one calculation assesses the probability to be less than 0.5 while the other assesses it to be greater as "disagreement." Cases where there is disagreement indicate that timing overlap between outbreaks at the two farms

Table 2. Comparison of probabilities for FMDV farm–farm transmission linkages from the 2001 UK outbreak

Pair	P_{Cottam}	$P(M=1\|k)$	Pair	P_{Cottam}	$P(M=1\|k)$
1-2	0.82	0.92	K-F	0.00	0.57
2-3	0.32	0.57	L-E	0.24	0.38[a)]
3-4	0.21	0.85	F-G	0.00	0.02[b)]
4-5	0.16	0.85	G-I	0.10	0.92
3-A	0.00	0.00[b)]	I-J	0.99	0.92
A-N	0.11	0.05[a)]	M-D	0.29	0.92
4-K	0.00	0.11[a)]	O-C	0.25	0.00[b)]
K-B	0.33	0.11[a)]	O-M	0.00	0.00[b)]
K-L	0.38	0.11[a)]	O-P	0.13	0.00[b)]
K-O	0.14	0.01[b)]	–	–	–

Orange indicates “agreement” while Green indicates “disagreement.”
a) Support is highest for one intermediate link.
b) Support is highest for two intermediate links.

reduces the likelihood of direct transmission, but the genetic sequences are very similar.

Our results provide a high degree of support for the first four links (farms 1–5) in Cottam’s network, which also evidently have very high support from contact tracing [27]. However, there are a significant number of instances where Cottam assessed the opportunity for a transmission event to be low, while $P(M = 1|k)$ supports the hypothesis of transmission. The clearest case of discrepancy involves the direct transmission link between nodes K and F, which receives moderate support from our analysis while the timing discrepancy between outbreaks at the two farms would apparently preclude direct transmission. A possible explanation is that infection of K was caused by contaminated fomites from farm F whose transport to K was delayed, but not stopped by isolation measures. Because FMDV can survive in the environment for long times and remain infective [28], such delayed transmission is not implausible. Note that there is no significance to cases where Cottam’s likelihood is greater than $P(M = 1|k)$ since timing overlap does not necessarily imply transmission. On the other hand, the fact that there are no cases where Cottam finds the likelihood of transmission to be > 0.5 while we find $P(M = 1|k) < 0.5$ does not have an obvious explanation other than chance.

Table 2 also indicates where the calculated posterior probability was highest for $M = 2$ or $M = 3$, implying one or two intermediate nodes between those farms, respectively. Both Cottam and a more recent analysis of the same data in [10] concluded that unknown intermediate nodes were likely to be needed to produce a tree consistent with the combined genetic and epidemiological findings. Both [10] and [27] also point out that when isolates from a single animal are used, there is no guarantee that the sequence is a valid representation of the consensus sequence for an entire herd. Therefore, some of the “intermediate nodes” implied by larger genetic differences might actually be artifacts caused by significant genetic drift within a larger herd that is not taken into account.

Finally, we note that both Cottam and Morelli’s analyses demonstrate that the weight assigned to timing evidence can critically change the most likely tree inferred from tree reconstruction methods. For example, Morelli used only part of the network used by Cottam, which leads to a larger estimate for γ. In addition, his inferred network contains shorter chains than Cottam’s, suggesting a very different degree distribution. This suggests caution in using such trees to generate reference data for ION. Before tree reconstruction methods mature, selecting defensible reference sets will necessarily remain dependent on high quality epidemiological judgments about transmission relationships, or carefully controlled laboratory studies.

4 Concluding remarks

The ION framework allows us to formulate genetic inference problems on transmission networks, where we can be explicit and unambiguous about the hypotheses we are testing. Statistically minded readers will recognize that the P(M) distribution provides the prior probabilities needed to formulate composite hypotheses such as “not related by direct transmission” or “belongs to an outbreak.” This formulation makes it clear that separation of two isolates by a small number of mutations means little unless we know the average rate of change per transmission step, *and* the topology of the underlying transmission network. The potential utility of this approach for assessing the evidential weight of genetic evidence in cases like the 2010 Haiti cholera outbreak, or in human immunodeficiency virus or Hepatitis C Virus transmission cases should be clear.

Practical implementation of ION does require that accurate reference sets of transmission-linked isolates be available, and this is primarily what limits wider application of the framework at present. However, collecting such reference data has close parallels to the collection of population data for mtDNA and Y-STRs and is simply a matter of motivation and resources. Transmission tree data is widely available from the epidemiological literature, and methods for transmission network simulation are widely available. The relative insensitivity to the details of empirical transmission trees suggests that trees from one part of an outbreak can be used to infer relationships in other parts, and that simulated trees based on contact network characteristics can be used as well.

This work was supported by a grant from the Department of Homeland Security during 2008 and 2009. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The authors have declared no conflict of interest.

5 References

[1] Frerichs, R. R., Boncy, J., Barrais, R., Keim, P. S., Piarroux, R., *Proc. Natl. Acad. Sci. USA* 2012, *109*, E3208.

[2] Metzker, M. L., Mindell, D. P., Liu, X. M., Ptak, R., Gibbs, R. A., Hillis, D. M., *Proc. Natl. Acad. Sci. USA* 2002, *99*, 14292–14297.

[3] González-Candelas, F., Bracho, M. A., Wróbel, B., Moya, A., *BMC Biol.* 2013, *11*, 76.

[4] Bernard, E. J., Azad, Y., Vandamme, A. M., Weait, M., Geretti, A. M., *HIV Med.* 2007, *8*, 382–387.

[5] Learn, G. H., Mullins, J. I., in: Leitner, T., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Wolinsky, S., Korber, B. (Eds.), *HIV Sequence Compendium 2003*, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 04–7420, 2003, pp. 22–37.

[6] Pillay, D., *BMJ* 2007, *335*, 460–461.

[7] Bhattacharya, S., *Nature* 2014, *506*, 424–426.

[8] Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H., Swaminathan, B., *J. Clin. Microbiol.* 1995, *33*, 2233–2239.

[9] Foley, S. L., Lynne, A. M., Nayak, R., *Infect. Genet. Evol.* 2009, *9*, 430–440.

[10] Morelli, M. J., Thebaud, G., Chadoeuf, J., King, D. P., Haydon, D. T., Soubeyrand, S., *PLoS Comp. Biol.* 2012, *8*, e1002768.

[11] Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., Ferguson, N., *PLoS Comput. Biol.* 2014, *10*, e1003457.

[12] Tavare, S., *Lectures on Mathematics in the Life Sciences*, Vol. 17, American Mathematical Society, Providence, RI, 1986, pp. 57–86.

[13] Lee, H. Y., Giorgi, E. E., Keele, B. F., Gaschen, B., Athreya, G. S., Salazar-Gonzalez, J. F., Pham, K. T., Goepfert, P. A., Kilby, J. M., Saag, M. S., Delwart, E. L., Busch, M. P., Hahn, B. H., Shaw, G. M., Korber, B. T., Bhattacharya, T., Perelson, A. S., *J. Theor. Biol.* 2009, *261*, 341–360.

[14] Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., *Bayesian Data Analysis*. 2nd edition, Chapman and Hall/CRC, Boca Raton, FL, 2004.

[15] Vogler, A. J., Keys, C. E., Allender, C., Bailey, I., Girard, J., Pearson, T., Smith, K. L., Wagner, D. M., Keim, P., *Mutat. Res.* 2007, *616*, 145–158.

[16] Chartrand, G., *Introductory Graph Theory*, Dover Publications, New York, 1985.

[17] Fronczak, A., Fronczak, P., Holyzt, J. A., *Phys. Rev. E* 2004, *70*, 056110.

[18] Rocha, L. E. C., Liljeros, F., Holme, P., *PLoS Comput. Biol.* 2011, *7*, e1001109.

[19] Ayyalasomayajula, S., *Zoonoses Public Health* 2008, *55*, 497–506.

[20] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., Getz, W. M., *Nature* 2005, *438*, 355–359.

[21] Ruan, Y., Wei, C. L., Ee, L. A., Vega, V. B., Thoreau, H., Yun, S. T., Chia, J.-M., Ng, P., Chiu, K. P., Lim, L., Tao, Z., Peng, C. K., Ean, L. O., Lee, N. M., Sin, L. Y., Ng, L. F. P., Chee, R. E., Stanton, L. W., Long, P. M., Liu, E. T., *Lancet* 2003, *361*, 1779–1785.

[22] Vega, V. B., Ruan, Y., Liu, J., Lee, W. H., Wei, C. L., Se-Thoe, S. Y., Tang, K. F., Zhang, T., Kolatkar, P. R., Ooi, E. E., Ling, A. E., Stanton, L. W., *BMC Infect. Dis.* 2004, *4*, 32.

[23] Liu, J., Lim, S. L., Ruan, Y., Ling, A. E., Ng, L. F. P., Drosten, C., Liu, E. T., Stanton, L. W., Hibberd, M. L., *PLoS Med.* 2005, *2*, e43.

[24] Goh, K.-T., Cutter, J., Heng, B.-H., Ma, S., Koh, B. K. W., Kwok, C., Toh, C.-M., Chew, S.-K., *Ann. Acad. Med. Singapore* 2006, *35*, 301–316.

[25] Varia, M., Wilson, S., Sarwal, S., McGeer, A., Gournis, E., Galanis, E., Henry, B., *CMAJ* 2003, *169*, 285–292.

[26] Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D. P., Zhu, Z., Schuchat, A., *Emerg. Infect. Dis.* 2004, *10*, 256–260.

[27] Cottam, E. M., Thebaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., Haydon, D. T., *Proc. R. Soc. B* 2008, *275*, 887–895.

[28] Bartley, L. M., Donnelly, C. A., Anderson, R. M., *Vet. Rec.* 2002, *151*, 667–669.