LLNL-TR-420850

A Statistical Framework for Microbial Source Attribution: Measuring Uncertainty in Host Transmission Events Inferred from Genetic Data (Part 2 of a 2 Part Report)

Jonathan E. Allen and Stephen P. Velsko
Lawrence Livermore National Laboratory

November 16, 2009

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: Measuring uncertainty in host transmission events inferred from genetic data (Part 2 of a 2 part report)

Jonathan E. Allen and Stephen P. Velsko

November 23, 2009

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

Executive Summary

This report explores the question of whether meaningful conclusions can be drawn regarding the transmission relationship between two microbial samples on the basis of differences observed between the two sample's respective genomes. Unlike similar forensic applications using human DNA, the rapid rate of microbial genome evolution combined with the dynamics of infectious disease require a shift in thinking on what it means for two samples to "match" in support of a forensic hypothesis. Previous outbreaks for SARS-CoV, FMDV and HIV were examined to investigate the question of how microbial sequence data can be used to draw inferences that link two infected individuals by direct transmission. The results are counter intuitive with respect to human DNA forensic applications in that some genetic change rather than exact matching improve confidence in inferring direct transmission links, however, too much genetic change poses challenges, which can weaken confidence in inferred links. High rates of infection coupled with relatively weak selective pressure observed in the SARS-CoV and FMDV data lead to fairly low confidence for direct transmission links. Confidence values for forensic hypotheses increased when testing for the possibility that samples are separated by at most a few intermediate hosts. Moreover, the observed outbreak conditions support the potential to provide high confidence values for hypothesis that exclude direct transmission links.

Transmission inferences are based on the total number of observed or inferred genetic changes separating two sequences rather than uniquely weighing the importance of any one genetic mismatch. Thus, inferences are surprisingly robust in the presence of sequencing errors provided the error rates are randomly distributed across all samples in the reference outbreak database and the novel sequence samples in question. When the number of observed nucleotide mutations are limited due to characteristics of the outbreak or the availability of only partial rather than whole genome sequencing, indel information was shown to have the potential to improve performance but only for select outbreak conditions. In examined HIV transmission cases, extended evolution proved to be the limiting factor in assigning high confidence to transmission links, however, the potential to correct for extended evolution not associated with transmission events is demonstrated. Outbreak specific conditions such as selective pressure (in the form of varying mutation rate), are shown to impact the strength of inference made and a Monte Carlo simulation tool is introduced, which is used to provide upper and lower bounds on the confidence values associated with a forensic hypothesis.

Introduction

The ever decreasing time and costs to sequence microbial isolates provides the opportunity to collect tremendous amounts of genetic data linking current and historic microbial transmission events with their infected hosts. If readily available microbial genome sequence data could then be used to infer host to host transmission links, it would have profound impact on the microbial forensics field. Yet the precise utility of microbial genome sequence data in forensic investigations remains very much an open question. A statistical framework for inferring host to host transmission from genetic sequence data was introduced in part 1 of this report [25]. The goal of the framework is to estimate the probability that two microbial sequence samples are linked by a host transmission event. The basic probability functions that are required to estimate the probabilities of various transmission relationships are derived from known outbreaks of the disease in question. This leads to two potential sources of error. First, there is always uncertainty in the value of parameters derived from empirical data. Secondly, parameters describing the rate of genetic change over the transmission network and certain statistical properties of the transmission network itself are assumed to be representative of the (new) outbreak under investigation. Since every outbreak has unique features, the assumption of the representativeness may be unwarranted. Moreover, there is uncertainty expected to come from inherent limitations in genome sequencing technology as well as knowledge gaps in microbial evolution and epidemiology. Therefore, in this report (Part 2) we investigate the strength and reliability of inferring host transmission links in the presence of a number of potentially confounding factors. The work lays the groundwork for developing accurate microbial genetic inference tools, which use the evolution of microbial genomes as evidence able to pass the rules of admissibility set out in the trilogy of cases initiated by the supreme court ruling on Daubert v. Merrell Dow Pharmaceuticals, Inc. [1]. To provide a tool of value to forensic investigators, the framework is designed to quantify error rates that measure the uncertainty associated with the weight of evidence assigned to a specific forensic hypothesis.

This is the first report to our knowledge that attempts to systemically characterize the utility of microbial genome data for use as forensic evidence in linking two hosts by transmission from the sequenced microbial infections. To ground the work in real world cases, three viruses are examined, each capturing distinct characteristics of past RNA viral outbreaks, Foot and Mouth Disease Virus (FMDV), Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), and Human Immunodeficiency Virus (HIV). These viruses are examined through their spread in previous outbreaks, reported in published literature (described in the context of addressing public health concerns), the FMD 2001 outbreak in Britain, the 2003 SARS outbreak in Singapore and two HIV case studies.

Parameters tied to inferring host transmission links are examined to consider how changes in underlying assumptions may affect the reported confidence values for drawing inferences from sequence data. The errors sources we consider are listed below:

- Uncertainty in mutation rate
- Using whole versus partial genome sequence
- Sequencing error rate
- Delay in sample collection with respect to host infection
- Choice of genetic distance measure
- Constraints on viral evolution

The availability of genome sequence data connected to contact tracing and other epidemiological parameters gives the ability to generate models and predictions based on past outbreak events. In addition, simulation models are constructed to estimate the magnitude of change to outbreak parameters needed to invoke significant changes in the types of inferences made among related sequences.

Simulation tool

A simulation tool was developed to use both observed epidemiological parameters associated with past outbreaks and observed genome sequence evolution rates to examine the precise effects that different outbreak parameters have on inferring transmission links. Moreover, the simulated genomic sequence data generated by the model can be compared with predictions made from past outbreak cases to better understand potential errors in the epidemiology data.

The model takes as input an outbreak transmission tree representing the spread of the virus observed in a past outbreak. For FMDV, the transmission tree includes 20 farms [4] and for SARS 194 human transmission contacts [12]. Sequences associated with each node in the network are generated using a fixed per time unit (e.g. day or year) mutation rate, which can be estimated using time stamp information associated with sequence samples and common assumptions of homogeneous evolution rates [6]. (In most cases, mutation rates published in previous studies are used.) In addition, a function is given describing the timing patterns of the infected host prior to transmission to the target host. The approach is derived from the concept of a SEIR contact model, which describes the host in four states: susceptible (S), exposed (E), infectious (I), and recovered (R) [19].

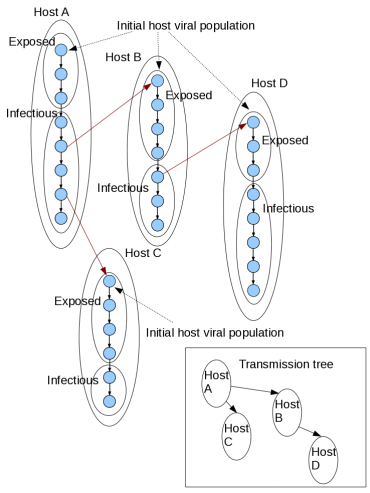

Figure 1: Graphical model for viral intra-host evolution and inter-host transmission. Bottom right box shows the equivalent (simplified) inter-host transmission tree.

Focus is given to using the model parameters, which can be obtained from past outbreak data, which include past exposed (E) and infectious (I) components. Both parameters are uniquely defined by the outbreak type. Figure 1 shows the graphical model for a viral outbreak. Each host node is a pair of subgraphs, representing the exposed and infectious periods respectively. Each node in the subgraph represents a snapshot of the viral population at a discrete time. The viral population is currently represented by a single consensus sequence. The graph has the structure of a linear chain representing the progress of time. The number of nodes in the subgraph is determined by the length of the infection in each of the two phases (exposed and infectious). Inter host transmission is represented by an edge from a node from the infectious subgraph to the initial host viral population node in the exposed subgraph of the receiving host. For the example given in Figure 1, Host A transmits the virus to Host B and Host C at two different time points.

For the SARS 2003 Singapore outbreak, the observable data gives the time between the presentation of symptoms between the transmitting and receiving host. This is reported to be a mean time of 8.3 days with standard deviation of 3.8 [17], which is approximated to be normally distributed. In addition, there are previous efforts to estimate the virus incubation time - the time between the initial infection and the presentation of symptoms. This corresponds to the "E" component of the SEIR model. The time of infection is deduced by subtracting the predicted incubation time from the observed time between the two contact's respective presentation of symptoms. The predicted incubation time

is an estimate derived from observed cases in Singapore 2003, which are fit to a Weibull distribution with scale parameter 5.44 and shape parameter 1.91, which gives a mean incubation time of 4.83 days [15].

In the case of FMD, the exposed (E) host period is described by a gamma distribution with smoothing parameter 1.67 and shape parameter 3.0 [4]. The infectious (I) period begins with the initial symptom presentation and continues to the culling (or isolation) of the farm. This distribution takes the form of a smoothed histogram using a probability density estimator based on the observations taken from a portion of the 2001 UK FMD outbreak [4]. The duration of the outbreak occurs approximately over an 80 day period. SARS simulations are drawn from an approximate time duration of 90 days.

Model comparisons to sequenced outbreak data

A summary statistic is taken to be the average probability of an inferred host transmission link for all observed values of δ e.g. $P(\bar{M}) = \sum_{\delta} P(M|\delta)/|\delta|$ where $M = 1$ is the direct transmission case and $|\delta|$ is a count of the total number of distinct values observed for δ. The posterior probability defined in part 1 of the report is given as:

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

where $\delta = \delta(s_1, s_2)$ reflects the genetic distance between sequences s_1 and s_2. The default measure used unless noted, is the observed number of nucleotide substitutions between the two sequences. (The impact of different genetic distances is discussed later in this report.)

A value of $P(\bar{M}) = 1$ implies complete confidence in asserting a transmission link for all observed values δ associated with all true direct transmission cases and complete confidence in ruling out linkage for all other values of δ, where all values of δ are assumed to be equally likely. A lower and upper bound is taken from the distribution of $P(\bar{M})$ using 99.99% confidence intervals drawn from independent random trials (100 for FMD and 50 for SARS). Figure 2 shows the distribution for the posterior probability of direct transmission as a function of the number of nucleotide differences observed between two sequences. Independent Poisson distributions are used to estimate values for $P(s_1, s_2|M > 1)$ and $P(s_1, s_2|M = 1)$ in Equation 1. A later section (Nonparametric model estimation) will discuss the strengths and weaknesses of using Poisson distributions compared to empirical and other parametric distribution forms. While the slope of the lower and upper bound curves are similar, the observed sequence model shown in black in Figure 2 is right shifted, suggesting a higher expected number of mutations separating two sequences

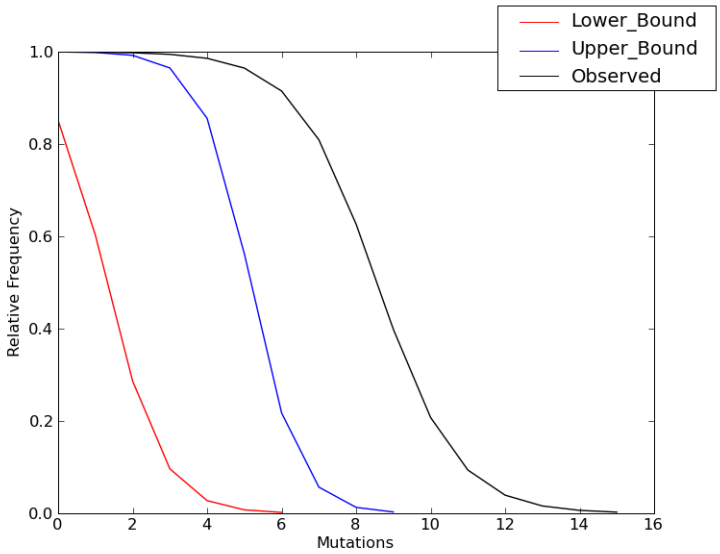

Figure 2: FMD simulated outbreak model versus observed.

linked by direct transmission in the actual 2001 UK outbreak compared to the simulation case. This discrepancy could be explained by a number of factors. The reported direct transmission links from the actual outbreak are in most cases estimates, thus sequence pairs presumed to be related by direct transmission could in fact be separated by one or more farms. In part 1 we showed that about half the putative $M = 1$ linkages were more consistent with $M > 1$. Alternatively epidemiological parameters or sequencing error rates could explain the reduced simulated substitution rates. We will revisit potential explanations later in this report. Despite this discrepancy, the simulation model provides a useful baseline for comparing the relative impact of different genetic and epidemiology parameters on overall forensic hypothesis testing.

Figure 3 shows the equivalent case for Singapore 2003 SARS outbreak. Here, the observed outbreak data (shown in black) more closely matches the model based on the reported epidemiology transmission events, however, the slope of the inference curves differ somewhat with the observed transmission links in some cases having larger genetic distances than expected.

Uncertainty in mutation rates

Obtaining precise measurements of mutation rates is difficult considering the complexity of microbial evolution. In many cases estimates can only be confined to a range of likely

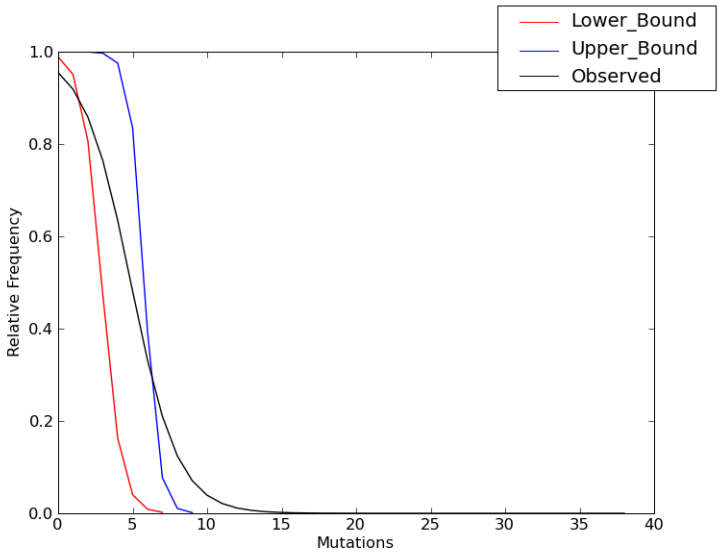

Figure 3: SARS simulated outbreak model versus observed.

values. For the case of FMDV, two inferred mutation rates are examined, which reflect different magnitudes of uncertainty depending on the amount of available sequence used to infer the mutation rate. Figure 4 shows the affect for a range of published mutation rates for FMDV estimated from two different genomic data sets [3, 4]. In one case the average mutation rate is 2.26×10^{-5} with a 95% confidence bounds of $[1.75 \times 10^{-5}, 2.8 \times 10^{-5}]$ and the second range includes a larger mutation rate range $[5.739 \times 10^{-6}, 3.509 \times 10^{-5}]$. The uncertainty reflects a potential 4 fold decrease and up to a 1.55 increase in the mutation rate with 95% confidence bounds around the estimated average mutation rate. Figure 4 shows the probability estimate of direct transmission (y-axis) given the observed genetic distance (x-axis) measured in terms of the number of nucleotide differences. The plots show the lower bound estimate for $P(\overline{M = 1})$ (direct transmission) with 99.99% confidence. The figure shows a degradation of inferential power with the lower mutation rates in inferring direct transmission. However, the strength of excluding direct transmission events remains similar. For example, the probability of direct transmission between virus pairs with more than 7 observed mutations is extremely low regardless of mutation rate.

For the SARS outbreak of 2003, an estimated average mutation rate of 5.7×10^{-6} with a potential range of $[1.48 \times 10^{-6}, 1.04 \times 10^{-5}]$ based on numerous published reports studying the virus mutation rate [28, 24, 2, 27]. On average the SARS-CoV is expected to have slightly less mutations per site per day, than FMDV and therefore based on our model would expect to have slightly weaker direct transmission inferential power. Figure

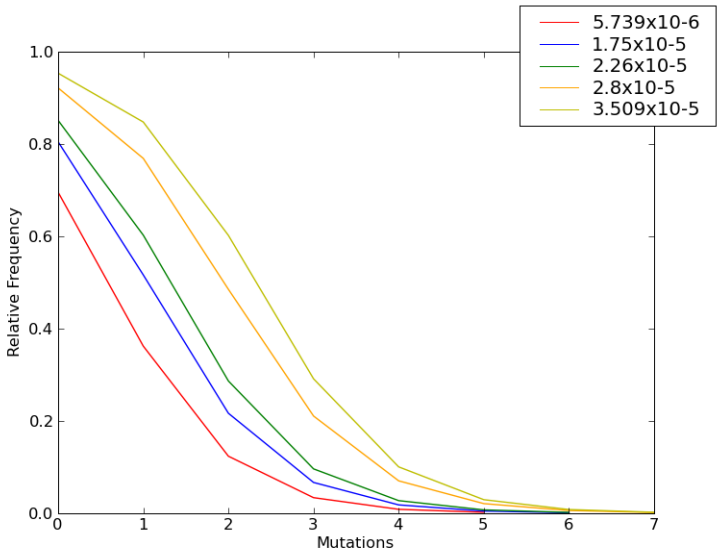

Figure 4: Variation in mutation rates in FMDV. Red=5.739×10^{-6}, Blue=1.75×10^{-5}, Green=2.26×10^{-5}, Orange=2.8×10^{-5} and Yellow = 3.509×10^{5}.

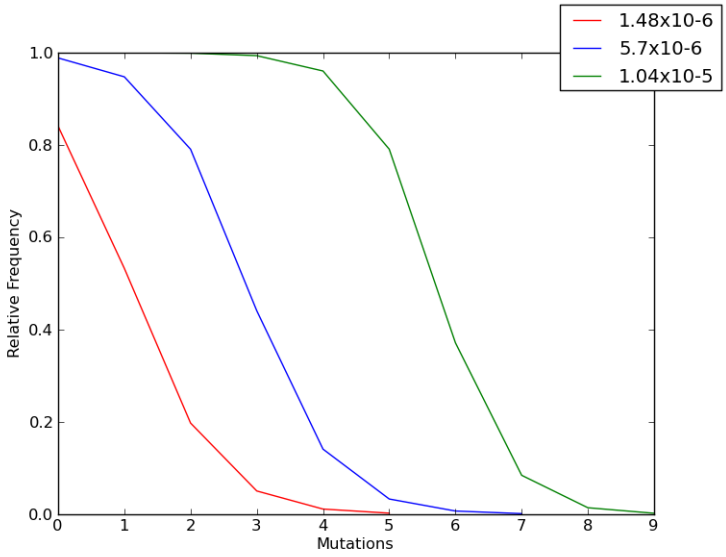

Figure 5: SARS-CoV mutation rates.

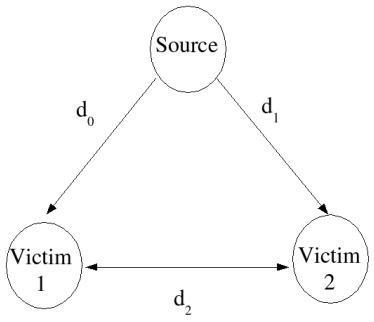

Figure 6: Host transmission triangle principle for transmission inference.

5 shows the posterior probability curve for direct transmission for the range of mutation rates. There appears to be dramatic improvement in the ability to infer direct transmission events under the higher mutation rate assumption. And in all cases, exclusion of direct transmission would be predicted at around 7 observed mutations.

Direct transmission with no mutation weakens genetic inference claims

Figure 6, illustrates the "host transmission triangle" principle for inferring direct transmission from genetic variation. Genetic distance values d_0 and d_1 represent the variation between the source host (node labeled "Source" in Figure 6) and the two respective victims, (labeled "Victim 1" and "Victim 2"). Genetic distance d_2 represents the distance between the two victims (assuming a uni-directional distance measure). Direct transmission inference is only possible when $d_2 > d_0 > 0$ and $d_2 > d_1 > 0$. Assume a simple discrete instantaneous time substitution model with $x = min(d_0, d_1)$ and single genotype population in each host, where the virus is transmitted to each victim within close time proximity followed by sub-sequence parallel host evolution. Assuming each nucleotide mutation is equally likely, the expected number of nucleotide positions where the two victims evolved genome copies with the same shared mutations, distinct from the source, is $x/9L^2$ where L is the length of the genome. Thus, $d_2 = (d_0 + d_1) - x/9L^2$ and under simple assumptions of evolution, d_2 is expected to be greater than both d_0 and d_1.

When a source host transmits the virus to its victim and no mutations are observed to distinguish between the two sequenced samples (taken from source and victim), an increase in uncertainty is introduced when asserting direct transmission. For example, when $d_0 = 0$, then $d_2 = d_1$ and distinguishing which host, "Victim 1" or the "Source" transmitted the virus to "Victim 2" is not possible. Moreover, this ambiguity is propagated to adjacent host nodes in the transmission graph. Figure 7 shows this affect, which contrasts the results for Figure 4 where 0 mutation transmission events are excluded. As

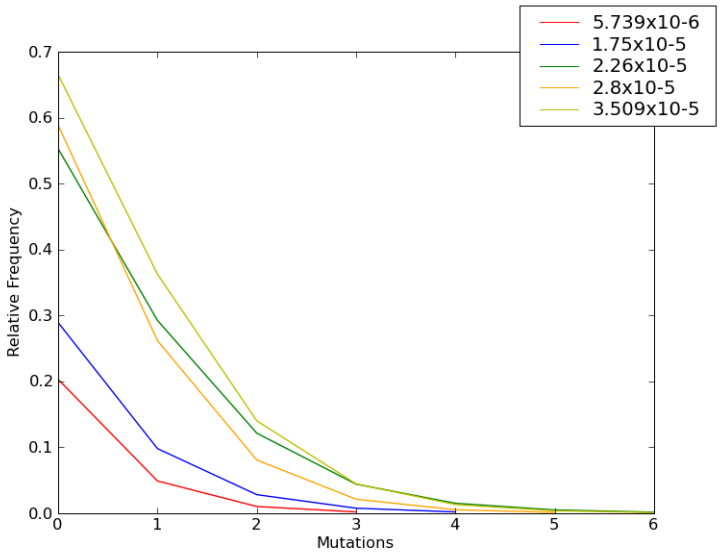

Figure 7: FMDV transmission models with 0 mutation events.

the mutation rate decreases, the probably of a 0 mutation transmission event increases and as a consequence the probability to distinguish direct transmission pairs decreases. It is important to note here that incubation time, in addition to mutation rate, will determine precisely the gradient observed for the probability of 0 mutation transmission events. The results show that mutation rates on the lower end of estimates ($\leq 1.75 \times 10^{-5}$) have substantially weaker direct transmission inferences. The general principle is that when the probability to observe k length transmission chains with 0 mutation events is statistically significant, the ability to positively assert transmission chains with path length $< 2 \times k$ becomes limited.

Preserved, however, is the inferential power to exclude direct transmission as a likely hypothesis. For example, when more than 6 mutations are observed between two sequences the probability of direct transmission is predicted to be very low, a factor, which is very similar to the case where at least 1 mutation per transmission event is assumed (7 mutations).

Weakening transmission linkage hypothesis

To positively assert a host to host transmission link between two sequenced samples under rapidly spreading outbreak conditions that allow for the 0 mutation host transmission events will likely require weakening the hypothesis to allow for the possibility of a few

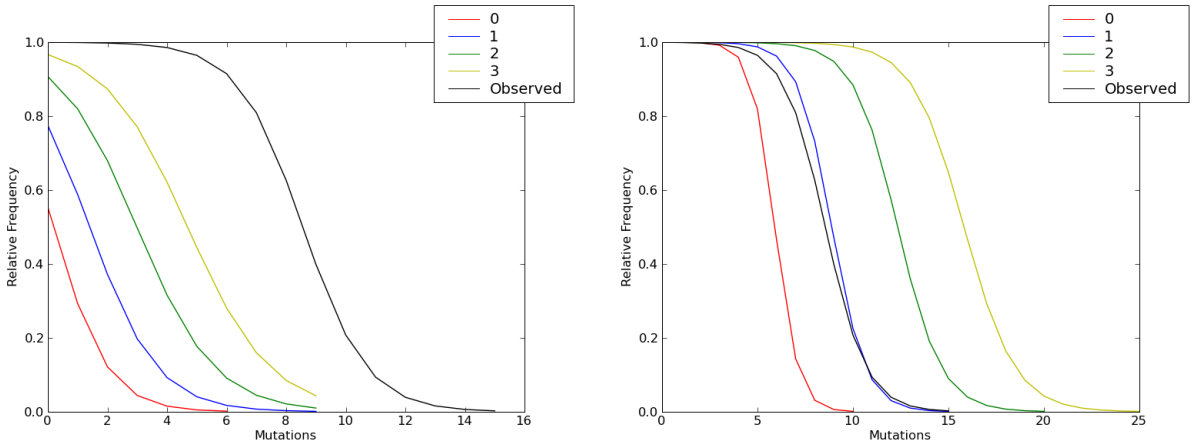

Figure 8: Posterior plots with 0 or more intermediate farms separating two transmission linked farms in FMDV. Left and right panel shows lower and upper bound confidence values respectively.

intermediate hosts. The question becomes one of determining the upper bound on the number of potential intermediate hosts involved in a host transmission chain for which a confident assertion can be made, given the observed number of genetic mutations between the observed sequence samples.

Figure 8 shows the affect of weakening the forensic hypothesis tested from, direct transmission (marked 0 in both figures) to allow for the possibility of up to 3 intermediate farms. The left panel in Figure 8 shows the lower bound confidence values, where the probability jumps from 55% to 78% when the test hypothesis is relaxed to allow for up to 1 intermediate farm. The probability value increases to 97% with up to 3 intermediate farms. Even allowing for the possibility of 3 intermediate farms, the conservative model suggests that no more than 9 substitutions should be observed between the two linked farms. In contrast, using the model derived from the observed outbreak shows non-zero probability of linkage with up to 15 substitutions. The right panel shows the upper bound confidence values for direct linkage. Interestingly, it is the hypothesis, which considers transmission linkage with up to 1 intermediate farm, under the optimistic model, which most closely resembles the observed direct transmission posterior probability plot. The model suggest that even when taking the upper bound confidence values, there is a strong probability that some of the observed sequenced samples are indeed not directly linked, a possibility suggested by Cottam et al. [4]. The results suggest the plausibility that the collection of sequenced samples are linked by at most one intermediate farm, but the potential for some cases where even more intermediate farms are involved remains high.

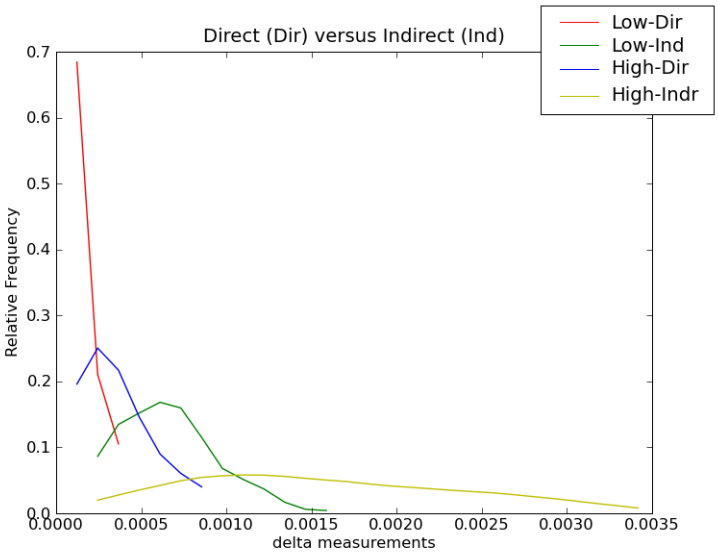

Figure 9: Distribution of δ (delta) values (FMDV) for outbreak cases representing the lower (Low) and upper bound (High) on the distribution $P(\overline{M})$.

Excluding direct linkage between sequence pairs

When presenting evidence at trial it may be desirable to present a conservative estimate for the probability of the hypothesis being true and it is important to note that the definition for the conservative estimate changes for some hypotheses. In inferring direct transmission linkage, "conservative" is defined by using the model with the weakest confidence for inferring direct transmission to provide a lower bound on the measured strength of inference given. Figure 9 points out that under slower evolution scenarios, inferring direct transmission is most difficult. The figure shows the distribution of pairwise genetic distance values (δ) for the two extreme ends of the average strength of inference, $P(\overline{M})$, from independent random trials. (Figure 2 shows the corresponding upper and lower bound distributions for assigning confidence to direct transmission assertions.) The expected number of observed substitutions associated with direct transmission is shown to be higher than the estimate given using the conservative confidence model. Therefore, when changing the hypothesis type to pose a question, which asks whether two samples can be excluded from direct transmission linkage, a conservative confidence value must be derived from the case where the probability of direct transmission is highest. In this scenario, the expected number of mutations associated with direct transmission linkage increases, thus raising the threshold on the number of observed mutations needed to in-

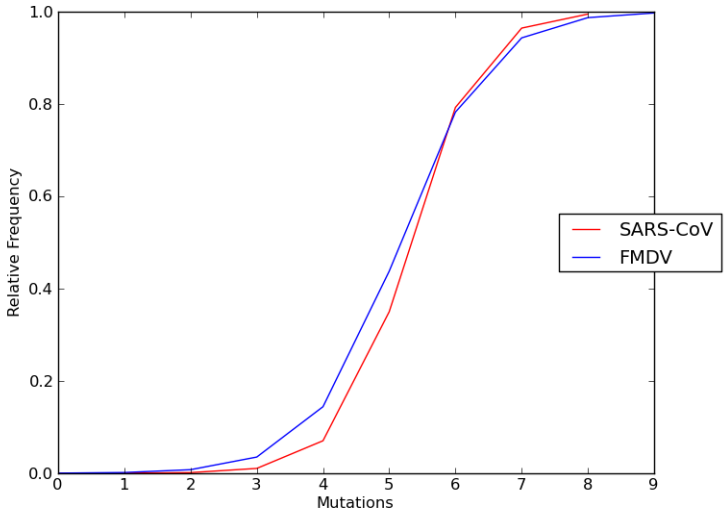

Figure 10: Posterior probability functions for excluding direct transmission linkage in SARS and FMD.

fer two sequences are not linked. This value can be calculated as 1 minus the posterior probability for the upper bound posterior probability for direct transmission. Figure 10 shows the two upper bound distributions used to measure confidence to exclude direct transmission for SARS and FMD. The predicted cutoff value differs by one nucleotide between the two cases (8 and 9). Among the pairwise direct transmission linked cases from the 2001 FMD outbreak, 6 pairs (31%) have between 11 and 14 distinguishing mutations. This apparent contradiction is explained with an examination of the reconstructed phylogeny for the sequenced strains. Figure 11 shows a schematic adapted from a maximum likelihood tree [11] for a portion of the tree containing two sequences related by direct farm to farm transmission. Four farms F, G, I and J are linked by a chain of direct transmission events shown by the arrows. Cottam et al [4] looked at sequenced strains linked by direct transmission and assumed that the inferred most recent common ancestral strain was the donor genome for the recipient farm rather than the sequenced strain taken from the donor farm. The presumption being that the sequenced strain is the product of continued evolution after transmission to the recipient farm. This point is illustrated in Figure 11 where although, farm F is thought to be the source of farm G's infection, there appears to be considerable parallel evolution with the sequenced strain taken from F shown by the branch length of 0.0345. Thus, the sum total branch lengths, (or similarly nucleotide substitutions) separating F and G is relatively large. Using the

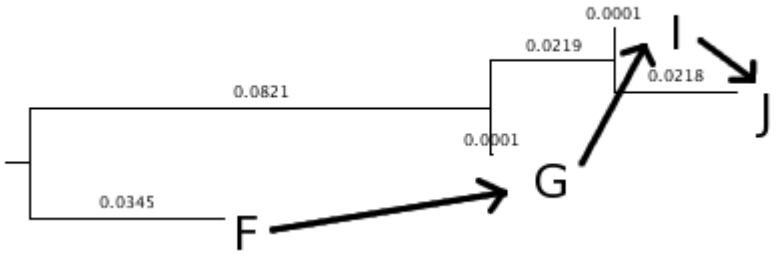

Figure 11: Phylogenetic tree for four FMDV sequences. Arrows mark farm to farm transmission links.

genetic distance from the ancestral strain, however, reduces the observed upper bound on the number of nucleotide changes between two direct transmission linked farms to 8. In the model used to generate the estimate for Figure 10, the sequenced strain is in fact modelled to be the direct ancestor of the sequence taken from the recipient farm, which explains how the simulation tool predicts an upper bound value of 9 for a threshold of high confidence in excluding direct transmission links.

Non-random sampling from the SARS outbreak weakens confidence of direct transmission inference

The optimal strategy for inferring the $P(M = k|\delta)$ transmission model requires a random sampling from the outbreak. Using the 194 person contact tracing graph published from the 2003 SARS outbreak in Singapore [12], provides an opportunity to consider affects of sampling strategy and size. Figure 12 shows random sampling of direct transmission pairs (e.g. k=1) taken from the Singapore outbreak, which span the duration of the outbreak. The results show that with a random sampling strategy, very small sample sizes provide relatively robust lower bounds measuring strength of inference for direct transmission. For example, the difference between a sample size of 25 direct transmission pairs and 50 direct transmission pairs is relatively small. Figure 13 shows the distribution of $P(\overline{M = 1})$ for the different sample sizes, which shows that there is a wider range of possible values at smaller sample sizes with variation leveling off at 25. The key challenge is designing a sampling strategy that maximizes the likelihood of taking a random sample.

The "worst case" affects of a non-random sampling are illustrated with a strategy that picks a single index case, and follows the subsequent spread from the individual for a limited duration. The $P(\overline{M = 1})$ distribution for random and non-random transmission pair sampling is shown in Figure 14. The potential pitfall of this strategy is particularly evident in the SARS outbreak where super spreader nodes persist throughout the outbreak

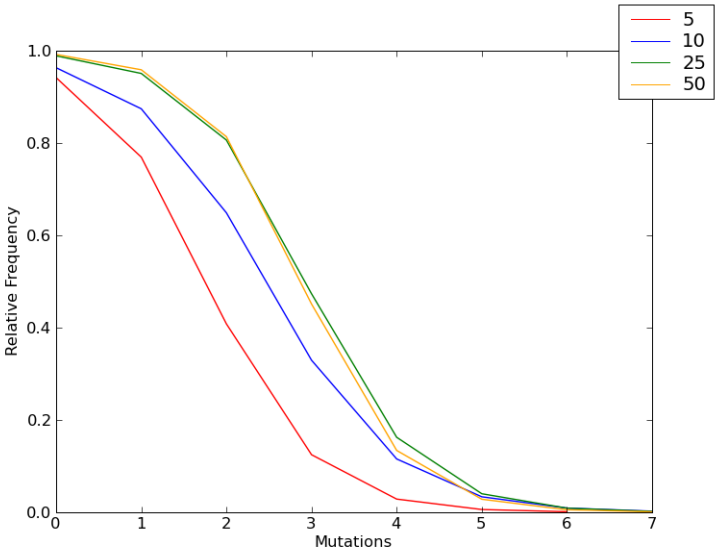

Figure 12: Random sampling for different sample sizes. 5, 10, 25 and 50 randomly sampled transmission pairs in SARS-CoV.

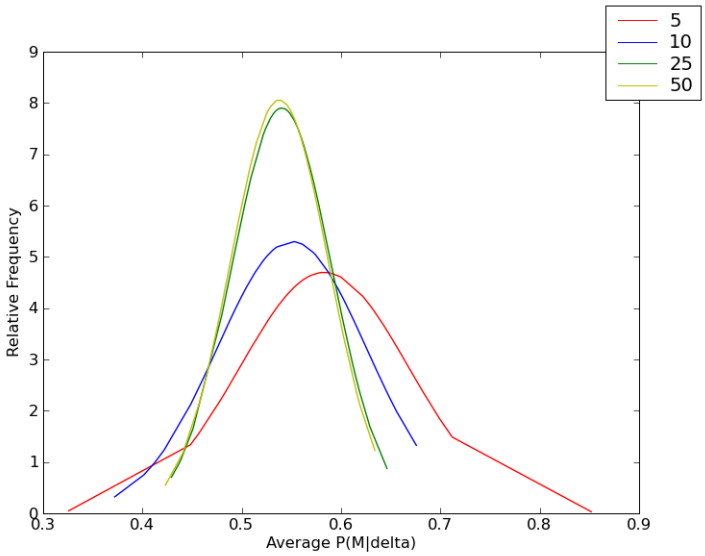

Figure 13: Sampling size distribution for $P(\overline{M=1})$ shows the variance increases as sample size decreases.

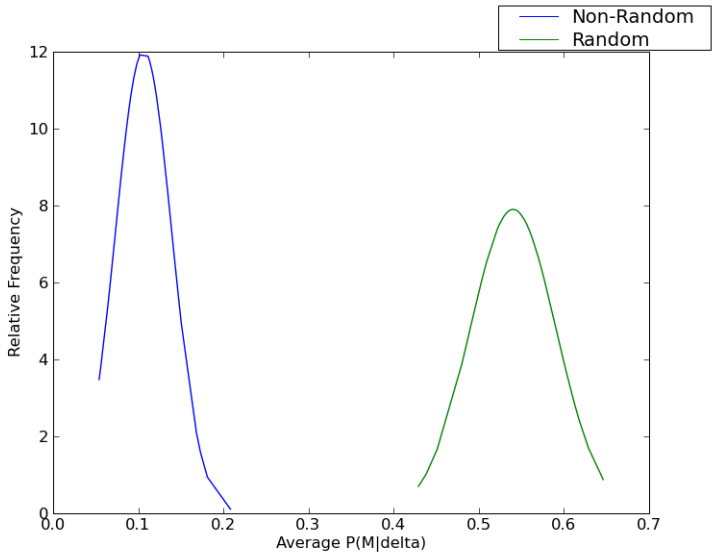

Figure 14: Average distribution for random and non-random sampling.

and such cases may occur in other outbreaks [18]. The non-random sampling strategy collects transmission pair data from the index case for approximately 30 days, which in this case leads to modelling just the relationship between a super-spreader and their contacts, rather than multiple infected individuals from throughout the outbreak. The result is a skewed view of how genetic distance correlates with direct transmission.

Sequencing data minimally requires draft coverage

Figure 15, shows the impact of sequencing error on strength of direct transmission inference for FMDV. Seven different average per nucleotide error rates are shown plus the observed outbreak data is shown for reference (black line in both figures). The left and right panel reflect the lower and upper bound values on the $P(\overline{M=1})$ respectively. The green line in Figure 15 represents the classic error rate associated with 'Finished' quality sequence defined for the human genome sequencing project which defines an error rate of 1 nucleotide in every 10,000 bases [22]. For comparison, 'High_Quality' refers to an error rate of 1 nucleotide every 100,000, which could be possible to achieve with higher read coverage. Higher error rates are given for various levels of draft sequencing including: Draft (2×10^{-4}), raw Sanger reads with an error rate estimate of 3×10^{-4} [26], an intermediate "Low_Quality" error rate of 4×10^{4} and a lower bound 454 read error estimate of 1×10^{-4}[9]. There is virtually no difference between no error and high quality sequence,

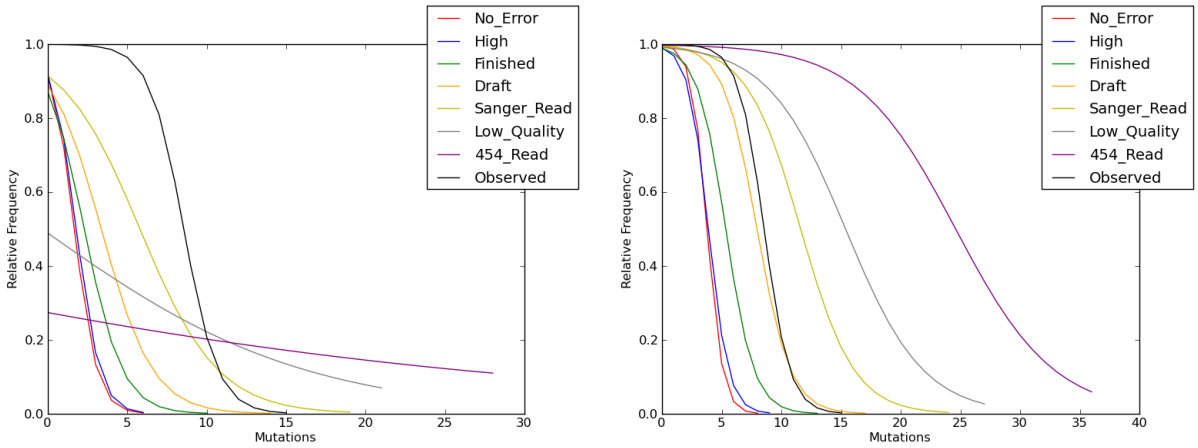

Figure 15: Sequencing error in FMDV for lower bound (left panel) and upper bound (right panel) of $P(M = 1|\delta)$.

but there is a noticeable change in the model with "Finished" error rates, where a higher number of mutations are expect to be associated with direct transmission. Higher error rates found in draft and rough draft sequence show substantial alterations, with increasing numbers of pairwise compared sequences with large numbers of distinct substitutions while still being linked by direct transmission. The lower bound shows that lower quality results in an overall reduced confidence in the model. For the lower bound $P(\overline{M = 1})$ model, the lower error rates appear to yield unusable estimates. The distribution that most closely matches the model derived from observed outbreak data comes from the upper bound $P(\overline{M} = 1)$ model associated with the "Draft" sequencing error rate of 1 in every 5000 bases.

Whole genome versus partial genome input

The question of how much of the genome is needed to infer direct transmission links was examined in the context of the SARS-CoV genome. Inferred mutation rates from time stamped sequenced genomes were used to infer two distinct rates of evolution, a slower rate corresponding to the larger ORF1a and ORF1b regions, and a faster rate corresponding to the Spike protein and other downstream transcribed genes [21]. With the observed rate of mutation being roughly twice the rate of the slower evolving regions, the potential impact of limiting sequence comparison just to the faster evolving regions was examined. Such an approach could provide sufficient information for the relatively "rapid spread" time duration for the SARS outbreak. Figure 16 shows two very similar posterior probability

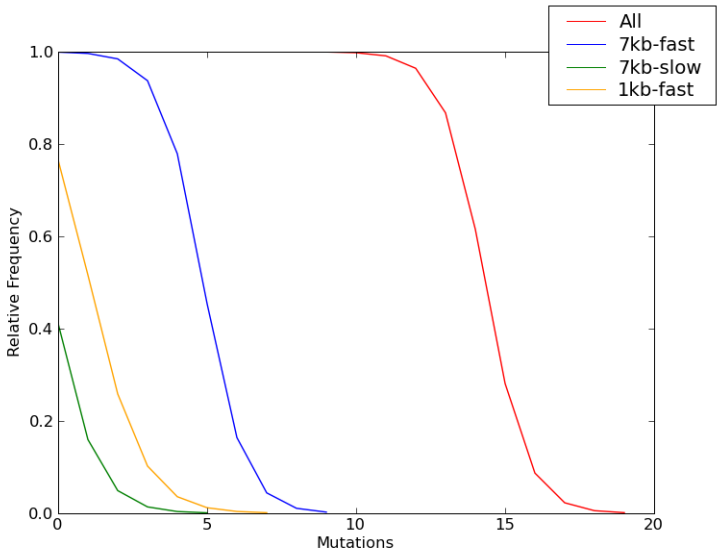

Figure 16: Posterior probability using whole and partial genome sequence.

curves for whole genome and the 7 kilobase (kb) fast evolving region, with the expected "right-shift" of the whole genome, which reflects the fact that more mutations are counted with whole genome comparisons. In contrast, a noticeable decrease in inferential power emerges when taking just 1 kb of the fast evolving region (orange curve), or taking a 7 kb region of the slower evolving region. It is important to note here that the mutation rates used for this estimate are higher than the published reports, which only give an "average" whole genome mutation rate. Thus, the actual mutation rate for the faster evolving region could be closer in practice to the weak inferential power displayed by the example 7 kb "slow" evolving region (green curve). Here, additional time stamped genome sequence data could be useful in calibrating the true rate of evolution.

Even assuming the more "ideal" faster mutation rate, Figure 17, highlights an increase in uncertainty for the 7 kb based model versus the whole genome model. Note that the distribution for the whole genome data is skewed right. This is caused by the fact that more mutations are counted, thus increasing the range of observed mutations and a larger range of mutation counts for which higher confidence inferences are drawn. Nevertheless, it is evident that the width of the distribution is wider for the 7 kb region implying greater variation in the strength of inference that might be drawn from similarly randomly distributed outbreak conditions. This is presumably due to the fact that sampling from the smaller space of candidate distinguishing sequence comparisons, increases the probability for variation in the number of observed mutations important to defining the transmission

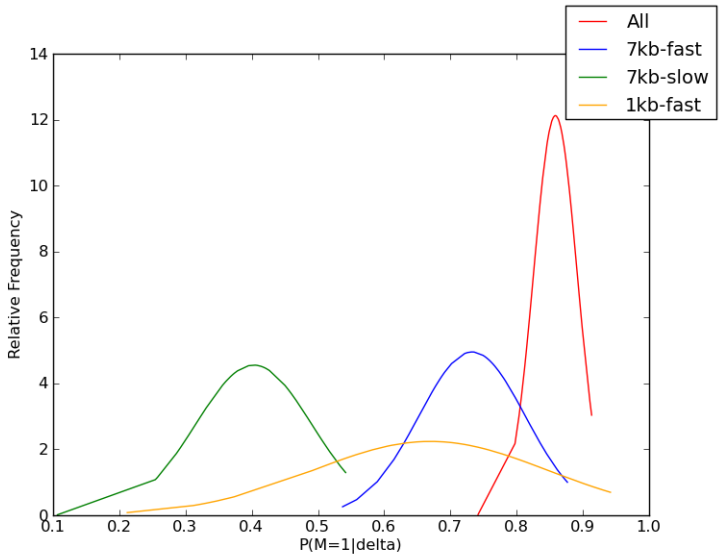

Figure 17: Distribution for $P(\overline{M=1}|\delta)$.

linkage between two sequences.

Differences in genetic distance measures

The importance of different genetic distance measures depends on a number of factors, including the evolutionary time scale: chronic versus acute infection [13]; the viral genome evolution process: strict nucleotide mutation based versus inclusion of short oligimer insertion and deletion (indel) events. To better understand the impact of using indel information for inferring transmission linkage, the SARS-CoV genome was examined where insertions and deletion events are observed within the outbreak.

Figure 18 gives a lower and upper bound of the Area Under the Curve (AUC) value for Receiver Operator Characteristic (ROC) curves drawn from sampling transmission subgraphs in the SARS Singapore 2003 outbreak. The higher the AUC value, the higher the overall true positive to false positive ratio observed in inferring direct transmission pairs. Indel rates were inferred by aligning time stamped sequenced genomes and computing the number of distinct observed pairwise indels and defining a per day indel rate as the average number of observed pairwise indels divided by the number of days separating the two samples. Previous work has shown that over short evolutionary time scales this approach can give reasonably accurate rate estimates[14]. A second parameter modelled the length of the indel assuming a normal distribution. The observed mutation rate (denoted Ob

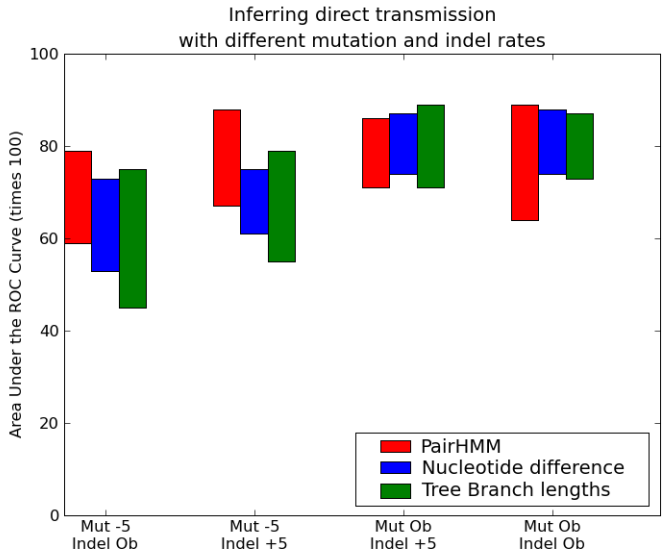

Figure 18: Comparison of genetic distance measures.

in Figure 18) were estimated from the sequence samples using the BEAST software [7]. Using default parameters (right-most column in Figure 18) shows little difference between the three genetic distances: PairHMM, which uses a pair hidden Markov model (HMM) [8] to incorporate both mutation and indel differences in the genetic distance measure; a count of nucleotide differences between two sequences; and branch lengths between two leaves of a maximum likelihood tree inferred using default parameters for the phylogenetic inference program Phylip [10]. The pair HMM genetic distance measure shows an improvement when the mutation rate is reduced by a factor of 5 (two left-most columns). The results show that when less information is available from nucleotide mutation differences, the indel information had a positive affect. A reduction of the inferred mutation rate by 5 brings the mutation rate in line with the published mutation used in Figure 5 (5.7×10^{-6}), which suggests the use of indel information could be of some practical use. However, a more accurate measure of the indel rate could be useful in providing additional support for this observation. It is also important to note that some viruses such as FMDV show little evidence for the presence of indels.

Delays in sample collection for chronic infection (HIV)

Figure 19 shows a HIV transmission tree for infected individuals established from personal interviews with each of the infected individuals [16]. The two digit number gives an

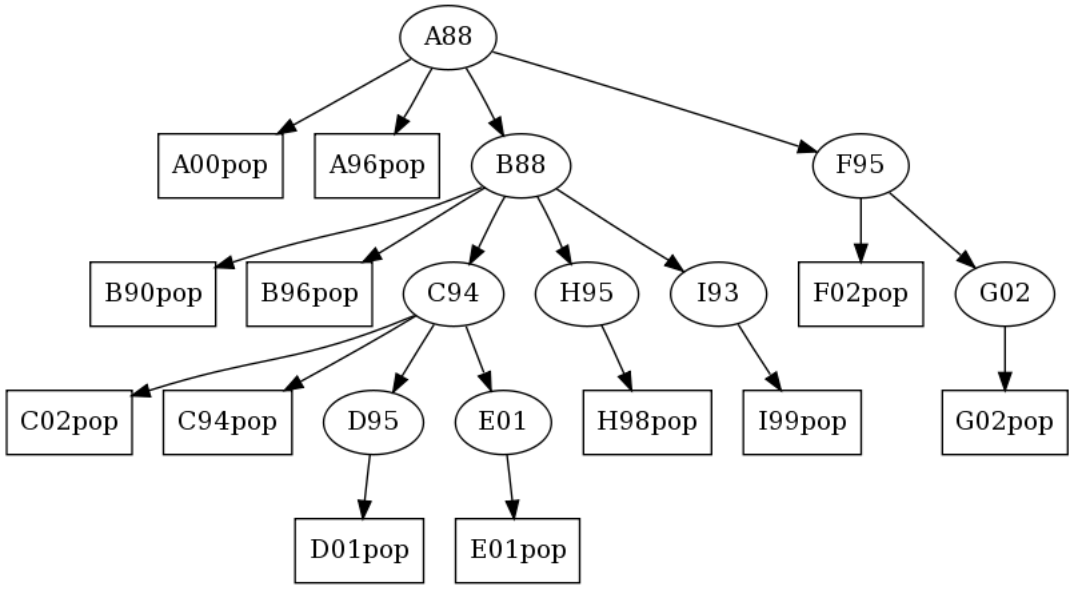

Figure 19: Transmission tree for HIV case study. Letters refer to a unique host (human) identifier. The two digit code refers to the suspected year of infection. Square nodes refer to sequence samples taken from the host.

estimate for the year of initial infection for each individual. Circle nodes reflect the infected host individual and their initial year of infection, square nodes show years in which env gene sequence samples were taken. In this study it is unclear which of the two earliest infected individuals, A or B, were the originators of the infection. Without loss of generality, a direction is assigned since the genetic distance measures employed are time reversible. This example shows the challenge of making transmission inferences from genetic data for cases involving chronic infection. Many of the sequence samples for the direct transmission cases were taken many years after the initial infection. For example, although host A infected host F, sequence samples were taken 8 years after infection for host A (1996) and 7 years after infection for host F, allowing for considerable parallel evolution in each of the hosts, which could lead to substantial sequence divergence between the two samples and therefore weaken inference claims of direct transmission using genetic distance measures. A common approach used to account for extended evolution is to look at the topology of the inferred phylogenetic tree to check the relationship between the most recent common ancestor for the two sequences in question. The limitation to this approach is that the tree construction is not explicitly linked to inferred transmission links and therefore quantifying the certainty associated with the inference claims is not

well defined.

Figure 20 describes in greater detail the transmission links introduced in Figure 19. This illustrates an alternative approach, which uses both the quantitative distance measure, phylogenetic information from the tree, and is linked to the transmission hypotheses of interest. Inferring a direct transmission link between host A and F, suggests that the amount of genetic change associated with direct transmission should measure the change in viral population from host A in 1995 (hexagonal node A95) to host F in 1995 sometime after the virus has adapted to the new host. The difficulty of course is that timely sample collection in most cases is not practical. However, use of inferred transmission timing windows in conjunction with inferred ancestral sequence can be used to link the phylogenetic data with transmission hypotheses.

To illustrate the potential utility of this approach, a simplified version of this idea was implemented, which estimates the amount of genetic change between the inferred initial viral population from the source host, compared to the population after the initial infection of the victim. This is accomplished by mapping the most recent common ancestor for any two leaves in the phylogenetic tree to the estimated point of initial infection of the host. For example, when comparing sequence samples "A96pop" and "F02pop", the most recent common ancestor is mapped to A88 as shown in Figure 21. The measured genetic distance now reflects just the evolution in the suspected source host prior to transmission, with subsequent evolution associated with delays in sample collection ignored. The improvement is seen in the distribution of observed genetic distance values, the summary ROC curve and the derived posterior probability estimate shown in Figure 22. Note that the posterior probability (bottom right panel) in Figure 22 reflects an estimate using probability density functions instead of a Poisson or Negative Binomial to better model the continuous value distributions. Details of this method are given in the next section.

Nonparametric model estimation

Recall that the posterior probability estimate defined by Equation 1 incorporates two independent distance based probability models: $P(\delta|M = 1)$ and $P(\delta|M > 1)$ (where δ is the genetic distance measure and $M = 1$ is the direct transmission hypothesis linking two sequenced samples). Results in this report are based primarily on the classic parametric modelling approach where the data is assumed to be drawn from a known distribution, chosen here to be either a Poisson or a Negative Binomial distribution.

A limitation of the Poisson and Negative Binomial distribution models is the assumption of a discrete random variable, which poses a challenge for modelling continuous value distance measures, such as those derived from phylogenetic tree branch lengths. Moreover, there does not appear to be a clear continuous value parametric analog to the discrete

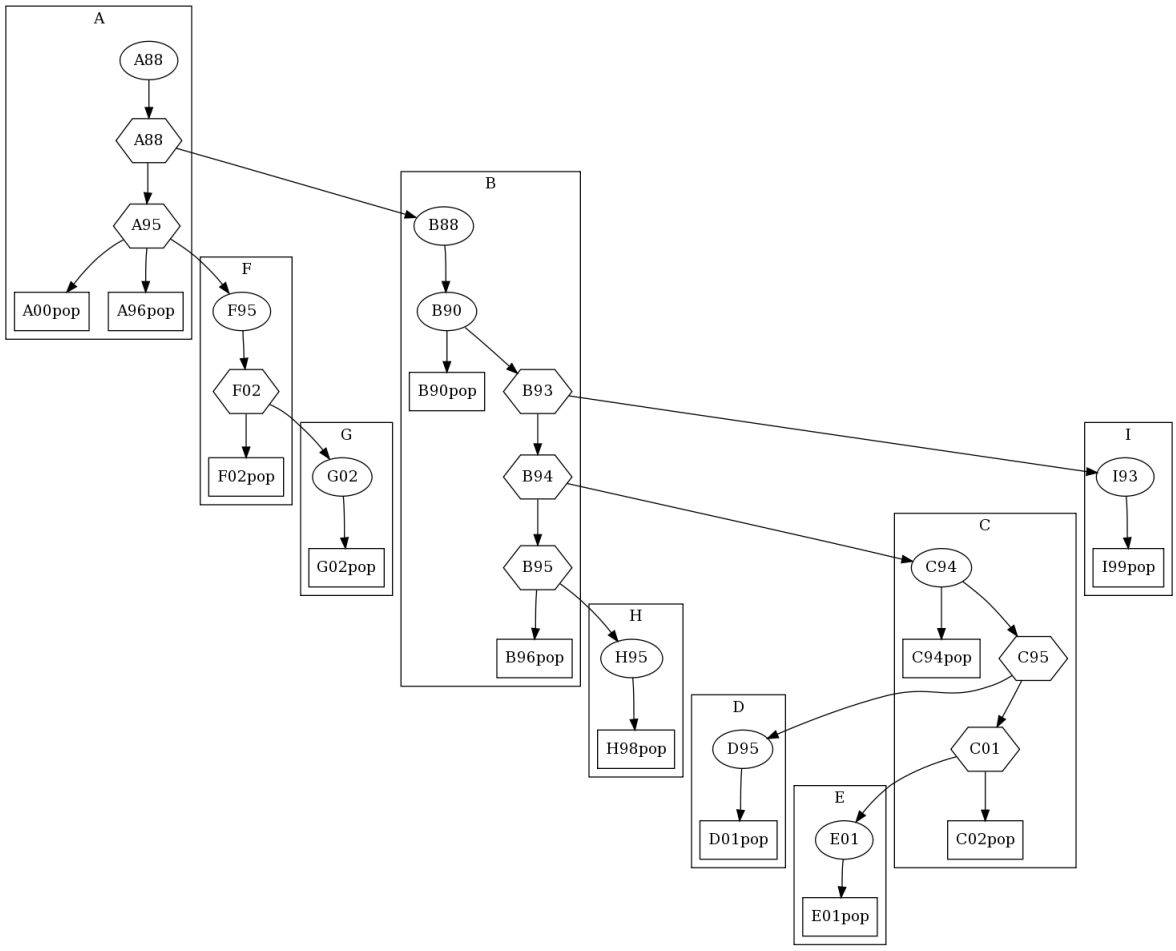

Figure 20: Alternative view of HIV transmission tree. Host nodes contain sub trees, with circle nodes representing initial infection times, rectangular nodes, sequence samples, and hexagonal nodes representing inferred in-host population snapshots.

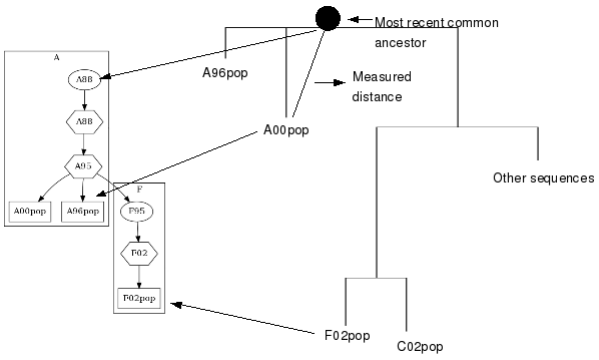

Figure 21: Mapping phylogenetic tree to transmission tree. Right tree shows a portion of a phylogenetic tree, with leaf nodes F02pop and A00pop and their common ancestor node, mapped to the transmission tree on the left.

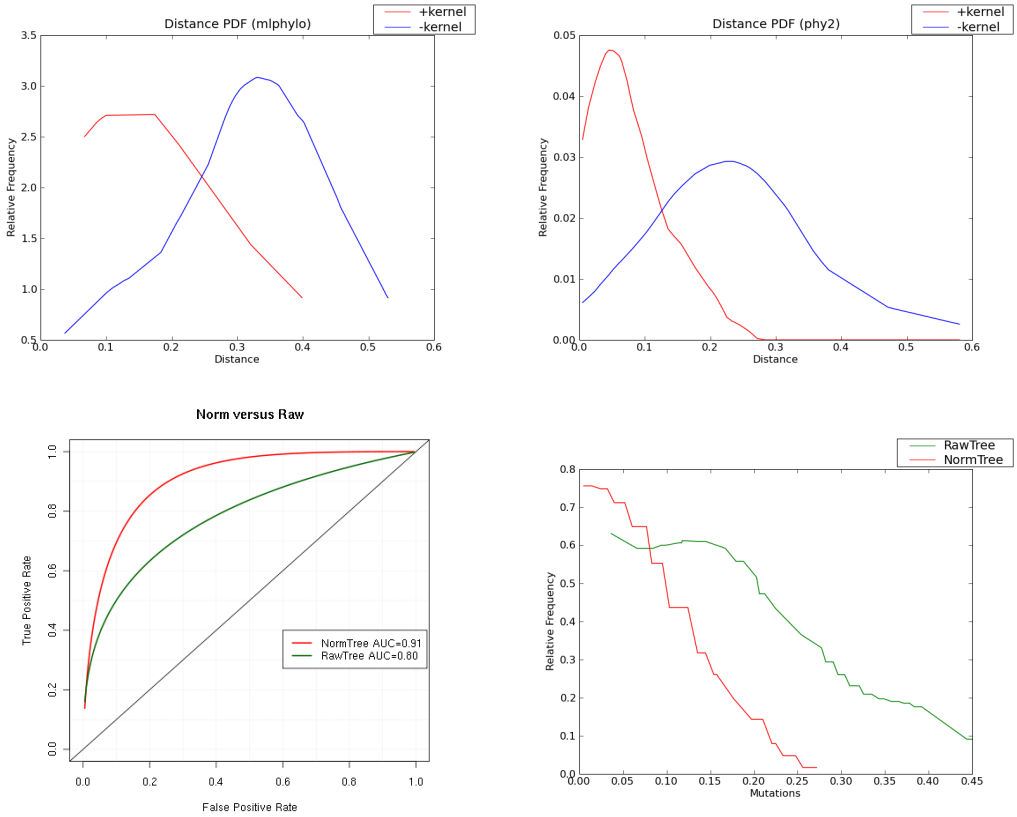

Figure 22: Corrected phylogenetic distance measure compared to traditional branch length distance measure.

value case. To address this issue, a nonparametric approach is considered: a probability density estimate (PDE). The PDE is used to generate the posterior probability plot for the HIV case study in Figure 22, which uses phylogenetic tree branch lengths. The mathematical background for the PDE is given in [20] and the software module is implemented in Python in a freely available software library package [5]. The model estimation procedure can be thought of as an extension of the histogram, which partitions the range of possible values for a random variable into equal size bins of width $2h$. The probability density estimate for a sample point x is given by the number of samples which fall within the same bin as x divided by the total training set size n and $2h$. Since the true probability density for a random variable X is by definition:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

the probability value can be estimated by counting the number of samples in the training set that fall in the interval $[x - h, x + h]$. The naive estimator for $f(x)$ is then defined to be

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w(\frac{x - x_i}{h})$$

where

$$w(t) = \left\{ \begin{array}{ll} \frac{1}{2} & |t| < 1 \\ 0 & |t| \geq 1 \end{array} \right\}$$

To avoid the stepwise behavior exhibited by the function w, a replacement kernel function K is used, with integral 1 ($\int_{-\infty}^{\infty} K(x)dx = 1$), which is symmetric and non-negative (not a requirement in all cases but assumed for now). A standard kernel function, the Epanechnikov kernel, is chosen and defined to be

$$K(t) = \left\{ \begin{array}{ll} \frac{3}{4}(1 - t^2) & |t| < 1 \\ 0 & |t| \geq 1 \end{array} \right\}$$

The one free parameter (once the kernel function is chosen) is h, which determines the "smoothness" of the density estimate. The goal is to calculate a globally optimal value taken to be the minimal mean integrated square error (MISE) between the estimate $\hat{f}(x)$ and the true density $f(x)$. There is a trade off in the optimization between reducing the bias of the estimator by increasing h and reducing the variance of the estimator by decreasing h - both of which contribute to the MISE. In practice, an asymptotic mean integrated square error is minimized, which depends on the number of observations n and the standard deviation σ of the sample data. The error minimization is taken with respect to a reference distribution (normal) assumed to be the true distribution. Under this assumption and using the Epanechnikov kernel, the optimal value for h is $\sigma(\frac{40\sqrt{\pi}}{n})^{\frac{1}{5}}$.

	$\chi^2$for M=1			χ^2 for M>1		
	Pois	Neg Bin	Dens	Pois	Neg B	Dens
SARS-CoV	3.4:4⋆	3.4:4⋆	2.2:4⋆	10696:30	115:27	54:28‡
FMDV	13:12⋆	13:12⋆	4:11⋆	303:20	276:21	16:26⋆
HIV env	141:22	136:22	16:22⋆	690099:69	1870:69	314:69
HIV gag	119:22	118:22	14:22⋆	53836095:64	438:64	223:64

Table 1: Model Fit.⋆=no support against, ‡=moderate support against. Pois=Poisson, Neg Bin=Negative Binomial, Dens=density estimator.

Table 1 shows χ^2 values, which compare the observed frequency of occurrence of δ given $M = 1$ or $M > 1$ to the expected values estimated from the two parametric models (Poisson and Negative Binomial), or the PDE. The table shows support for the parametric direct transmission models (columns 2 and 3) for SARS-CoV and FMDV but no support for selected direct transmission pairs of HIV data samples in 100 paired HIV gag gene sequences and 162 paired HIV env gene sequences [23]. Considerably less support is shown for the indirect $M > 1$ case when using the default estimator for $P(\delta|M > 1)$, which does not rely on *a priori* knowledge of the transmission tree beyond the observed direct transmission links. Here $P(\delta|M > 1)$ is estimated by simply counting the observed frequency of each δ value over all non-linked pairs. Since,

$$P(\delta|M > 1) = \sum_{M=2}^{N} P(\delta|M)P(M)$$

where N is the maximal path length separating two sequences s_1, s_2 for $\delta = \delta(s_1, s_2)$. Thus, there is a better chance at inferring accurate models for each $P(\delta|M)$ for $2 \leq M \leq N$ rather than attempting to model the sum of the constituent distributions. Figure 23 shows χ^2 values for each of the three model estimation methods Poisson, Negative Binomial (Negative_Bin), and Probability Density Estimate (Density) on $P(\delta|M)$for $1 \leq M \leq 6$. The fourth line marked "Significance", shows an upper bound on the χ^2 distribution (P-value=0.001). It is important to note that the degree of support for the null hypothesis (that deviation is attributed to random error) is dependant on the degrees of freedom, which differ for each path length. For example, there is a drop in χ^2 values for $M = 6$, which is due to a drop in the number of observed paths of length 6 in the FMD outbreak tree. The key feature to notice, is the difference between the upper bound χ^2 and each estimator all of which fall below the upper bound. The figure shows a relatively constant difference from the upper bound for all path lengths. These results highlight the potential improvement in model estimation, when having the complete transmission tree available (or possibly even an estimate). It is important to note that while the density estimator

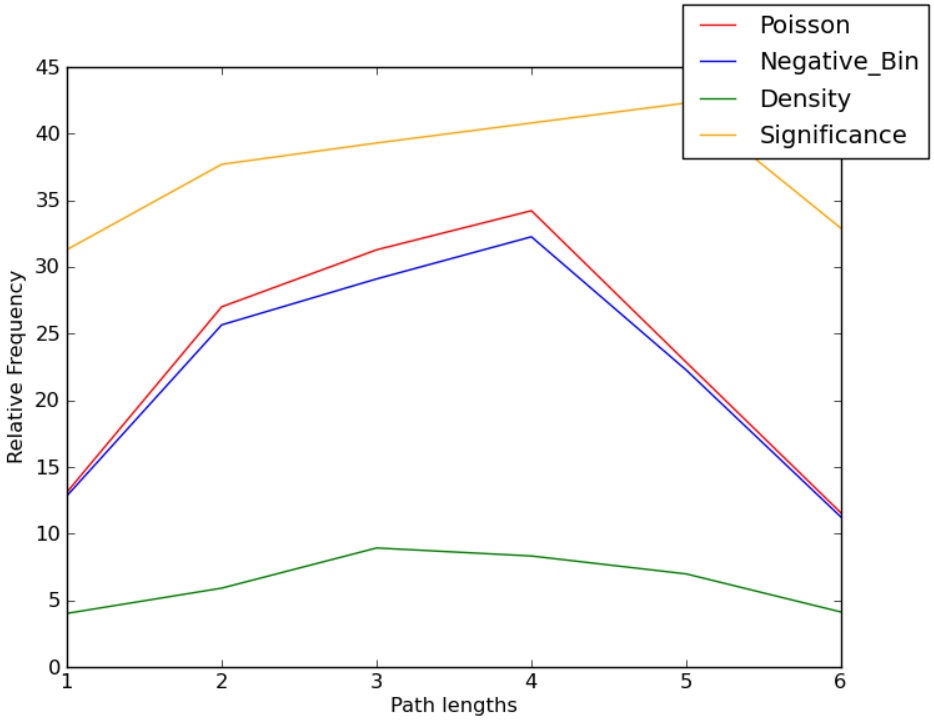

Figure 23: Multiple χ^2 values. Path lengths=# of intermediate hosts +1 between two nodes in the transmission tree.

shows the lowest χ^2 values in all cases, it may be subject to over fitting of the data, making additional cross-validation testing desirable.

Conclusion

Applying the $P(M|\delta)$ framework to observed outbreak data and simulated data revealed a number of findings that point both to the positive potential of using microbial sequence data to infer transmission links as well as potential limitations and areas for improvement. Comparing strength of inference with SARS data and HIV show striking differences between the two infection types. SARS shows a slower evolving, acute, and rapidly spreading infectious disease for which strength of inference differs from the more rapid evolution and extended time period characteristic of reviewed chronic HIV infections. These data sets point to two ends of the extreme, which pose significant but distinct challenges to supporting transmission hypotheses from genetic data. On the one hand, too little evolution on the part of the virus eliminates the potential to infer direct transmission links, an effect which permeates the entire outbreak. The challenge becomes one of estimating the upper bound on the number of intermediate hosts separating two sequenced samples, for which a transmission link hypothesis is likely. On the other extreme, increased rates of evolution increase the overlap between δ values associated with direct and non-direct transmission cases, particularly when the sample collection times vary significantly. Even for the 2001 FMD UK outbreak where the overall duration of the outbreak was relatively short (e.g. several months), the potential for parallel evolution appeared to be a factor that could lower confidence values on inferring direct transmission when not correctly taking this variation into account. Other than ensuring that indels information is used, there are no sequence analysis methods that can get around the problem of too little evolution to strengthen inference of transmission, however, in such instances, in addition to relaxing the hypothesis test, exclusion hypothesis testing should still lead to strong confidence values. For the issue of extended evolution, our preliminary results show the potential in correcting for extended sequence changes not associated with the transmission event and this is an area for targeted improvement.

Hidden variables play an important role to support the hypothesis test by providing confidence bounds on varying conditions unique to the outbreak of interest. Mutation rate, for example, is an important hidden variable, for which obtaining precise estimates is challenging - particularly when sequencing errors must be taken into account. Host incubation time is another important hidden variable that may not be easily ascertained. In the ideal case, with sufficient training data from past outbreaks, these hidden variables

need not be made explicit. However, our results show that the precise strength of direct transmission inference is affected by these underlying biological processes and therefore extra care must be taken to obtain a truly representative sampling from past outbreaks with respect to a number of these outbreak features. For example, the confidence of the direct transmission hypothesis being true for two substitutions separating two FMDV query sequences, can go from $< 20\%$ probability to $> 60\%$ probability, depending on the underlying mutation rate. To guard against imperfections in the sampling process or changes in future outbreaks, estimating values for hidden parameters using the past outbreak data and the simulation tool will provide lower and upper bound estimates on confidence. This serves to quantify confidence in the face of uncertainty with respect to the underlying epidemiology and viral evolution parameters. Even when the value of hidden variables are not in question, random sampling (from either simulated or observed data) appears to be an important procedure to provide a range of confidence values from which a conservative estimate can be chosen. An area targeted for improvement is to more closely integrate the observed sequence data associated with the hidden variables that affect inferring transmission links. This will provide an additional measure of confidence by comparing values assigned to hidden variables with estimates made from other sources.

References

[1] M A Berger. What has a decade of daubert wrought? *Am J Public Health*, 95 Suppl 1:S59–S65, 2005.

[2] The Chinese SARS Molecular Epidemiology Consortium. Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China. *Science*, 303(5664):1666–1669, 2004.

[3] E M Cottam, D T Haydon, D J Paton, J Gloster, J W Wilesmith, N P Ferris, G H Hutchings, and D P King. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the united kingdom in 2001. *Journal of Virology*, 80:11274–11282, 2006.

[4] E M Cottam, G Thebaud, J Wadsworth, J Gloster, L Mansley, D J Paton, D P King, and D T Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B*, 275:887–895, 2008.

[5] Michiel Jan Laurens de Hoon. Statistics for python library extension, 2008.

[6] A J Drummond, G K Nicholls, A G Rodrigo, and W Solomon. Estimating mutation parameters, population history and geneaology simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320, 2002.

[7] A J Drummond and A Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *bMC Evol Biol*, 7:214, 2007.

[8] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1999.

[9] Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W. Shafer, and Niko Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4(5):e1000074, 05 2008.

[10] J Felsenstein. 1993. PHYLIP (phylogeny inference package) version 3.5c. distributed by the author. Department of Genetics, University of Washington, Seattle.

[11] J Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

[12] Kee-Tai Goh, Jeffery Cutter, Bee-Hoon Heng, Stefan Ma, Benjamin KW Koh, Cynthia Kwok, Cheong-Mui Toh, and Suok-Kai Chew. Epidemiology and control of sars in singapore. *Ann Acad Med Singapore*, 35:301–316, 2006.

[13] B T Grenfell, O G Pybus, J R Gog, J L N Wood, J M Daly, J A Mumford, and E C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332, 2004.

[14] I Holmes. Using evolutionary expectation maximization to estimate indel rates. *Bioinformatics*, 21:2294–2300, 2005.

[15] A Y C Kuk and S Ma. The estimation of sars incubation distribution from serial interval data using a convolution likelihood. *Statist Med*, 24:2525–2537, 2005.

[16] P Lemey, I Derdelinckx, A Rambaut, K V Laethem, S Dumont, S Vermeulen, E V Wijngaerden, and A-M Vandamme. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol*, 79:11981–11989, 2005.

[17] M Lipsitch, T Cohen, B Cooper, J M Robins, S Ma, L James, G Gopalakrishna, S K Chew, C C Tan, M H Samore, D Fisman, and M Murray. Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300:1966–1970, 2003.

[18] J O Lloyd-Smith, S J Schreiber, P E Kopp, and W M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438:355–359, 2005.

[19] I B Schwartz and H L Smith. Infinite subharmonic bifurcation in an seir epidemic model. *J Math Biology*, 18:233–253, 1983.

[20] B W Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[21] Konrad Stadler, Vega Masignani, Markus Eickmann, Stephan Becker, Sergio Abrignani, Hans-Dieter Klenk, and Rino Rappouli. Sars – beginning to undertand a new virus. *Nat Rev Micro Bio*, 1:209–218, 2003.

[22] The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[23] S A Trask, C A Derdeyn, U Fideli, Y Chen, S Meleth, F Kasolo, R Musonda, E Hunter, F Gao, S Allen, and B H Hahn. Molecular epidemiology of human immunodeficiency virus type 1 transmission in a heterosexual cohort of discordant couples in zambia. *J Vir*, 76:397–405, 2002.

[24] Vinsensius Vega, Yijun Ruan, Jianjun Liu, Wah Lee, Chia Wei, Su Se-Thoe, Kin Tang, Tao Zhang, Prasanna Kolatkar, Eng Ooi, Ai Ling, Lawrence Stanton, Philip Long, and Edison Liu. Mutational dynamics of the sars coronavirus in cell culture and human populations isolated in 2003. *BMC Infectious Diseases*, 4(1):32, 2004.

[25] S P Velsko, J E Allen, and C Cunningham. A statistical framework for microbial source attribution part 1: Forensic inferences on disease transmission networks. Technical report, LLNL, 2009.

[26] K Eric Wommack, Jaysheel Bhavsar, and Jacques Ravel. Metagenomics: Read Length Matters . *Appl. Environ. Microbiol.*, 74(5):1453–1463, 2008.

[27] Shiou-Hwei Yeh, Hurng-Yi Wang, Ching-Yi Tsai, Chuan-Liang Kao, Jyh-Yuan Yang, Hwan-Wun Liu, Ih-Jen Su, Shih-Feng Tsai, Ding-Shinn Chen, and Pei-Jer Chen. Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: Molecular epidemiology and genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2542–2547, 2004.

[28] Zhongming Zhao, Haipeng Li, Xiaozhuang Wu, Yixi Zhong, Keqin Zhang, Ya-Ping Zhang, Eric Boerwinkle, and Yun-Xin Fu. Moderate mutation rate in the sars coronavirus genome and its implications. *BMC Evolutionary Biology*, 4(1):21, 2004.