Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks

Rolf J. F. Ypma,^{*,†} W. Marijn van Ballegooijen,^{*} and Jacco Wallinga^{*}

^{*}Centre for Infectious Disease Control, National Institute for Public Health and the Environment, 3721 MA Bilthoven, The Netherlands, [†]Julius Center for Health Research and Primary Care, University Medical Center, 3584 CG Utrecht, The Netherlands

ABSTRACT Transmission events are the fundamental building blocks of the dynamics of any infectious disease. Much about the epidemiology of a disease can be learned when these individual transmission events are known or can be estimated. Such estimations are difficult and generally feasible only when detailed epidemiological data are available. The genealogy estimated from genetic sequences of sampled pathogens is another rich source of information on transmission history. Optimal inference of transmission events calls for the combination of genetic data and epidemiological data into one joint analysis. A key difficulty is that the transmission tree, which describes the transmission events between infected hosts, differs from the phylogenetic tree, which describes the ancestral relationships between pathogens sampled from these hosts. The trees differ both in timing of the internal nodes and in topology. These differences become more pronounced when a higher fraction of infected hosts is sampled. We show how the phylogenetic tree of sampled pathogens is related to the transmission tree of an outbreak of an infectious disease, by the within-host dynamics of pathogens. We provide a statistical framework to infer key epidemiological and mutational parameters by simultaneously estimating the phylogenetic tree and the transmission tree. We test the approach using simulations and illustrate its use on an outbreak of foot-and-mouth disease. The approach unifies existing methods in the emerging field of phylodynamics with transmission tree reconstruction methods that are used in infectious disease epidemiology.

ESTIMATING who infected whom for an outbreak of an infectious disease can provide valuable insights. Estimated transmission trees have been used to evaluate effectiveness of intervention measures (Ferguson *et al.* 2001; Keeling *et al.* 2003; Wallinga and Teunis 2004; Heijne *et al.* 2009), to quantify superspreading (Lloyd-Smith *et al.* 2005), to estimate key parameters (Haydon *et al.* 2003; Heijne *et al.* 2012; Hens *et al.* 2012), and to identify mechanisms of transmission (Spada *et al.* 2004; Ypma *et al.* 2013). Transmission trees can be statistically reconstructed using epidemiological data from outbreak investigations, such as time of symptom onset, geographical location, and social ties; these data generally must be very detailed to allow for accurate reconstructions.

For many pathogens, in particular RNA viruses, evolutionary processes occur on the same timescale as epidemiological processes (Holmes *et al.* 1995; Pybus and Rambaut 2009). This makes it possible to draw conclusions about epidemiology from genetic analysis. The field that infers epidemiological characteristics from genetic sequences by simultaneously considering host dynamics and pathogen genetics has been dubbed “phylodynamics” (Grenfell *et al.* 2004). In practical applications researchers have considered a specific epidemiological model dependent on the phylogenetic tree inferred from sequence data, simultaneously estimating mutational and epidemiological parameters. This allowed them to answer questions on relative population sizes and dates of introduction of pathogens. Initially the epidemiological models used were classical models from population genetics, such as the Wright–Fisher model (Pybus *et al.* 2001). Recently more realistic epidemiological models such as the Susceptible–Infected–Recovered (SIR) model (Volz *et al.* 2009; Rasmussen *et al.* 2011) and birth–death model (Stadler 2009) have been suggested. In these methods, the sampled hosts are thought of as the leaves of the phylogenetic tree, while the internal nodes

Copyright © 2013 by the Genetics Society of America

doi: 10.1534/genetics.113.154856

Manuscript received June 27, 2013; accepted for publication September 5, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.154856/-/DC1>.

Corresponding author: Rolf Ypma, National Institute of Public Health and the Environment, Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven, the Netherlands, +31302747054. Email: rolf.ypma@rivm.nl

coincide with transmission times to unobserved hosts. The phylogenetic tree is thus equated with the partially observed transmission tree.

Although the transmission tree and the phylogenetic tree of an outbreak may appear as two incarnations of the same tree, they are in fact different in interpretation and in local characteristics. The phylogenetic tree represents the clonal ancestry of sampled pathogen its leaves are sampled pathogens, and its internal nodes are most recent common ancestors of the sampled and transmitted pathogens (Figure 1) (Pybus and Rambaut 2009). As a pair of lineages corresponding to two transmitted pathogens can coalesce together before coalescing to the lineage sampled from the infecting host, the topology of the two trees need not be the same (Figure 1A). The difference between phylogenetic trees and transmission trees is closely related to the difference between phylogenetic trees and species trees; in the latter context this phenomenon is known as "incomplete lineage sorting" (Rosenberg and Nordborg 2002; Maddison and Knowles 2006). While the timing of nodes in the transmission tree corresponds to transmission times, the timing of internal nodes of the phylogenetic tree corresponds to coalescent events that take place prior to transmission. The absolute difference in branch length between the two trees depends on the epidemiological generation interval and within-host dynamics. As the branch length of the phylogenetic tree decreases with sampling rate of hosts, the relative difference between the partially observed transmission tree and the phylogenetic tree is largest when the sampling rate is high (Figure 1).

Recently, methods that focus on including genetic information in transmission tree reconstructions have been proposed (Morelli *et al.* 2012; Ypma *et al.* 2012). However, these approaches either ignore the phylogenetic tree or assume that internal nodes of the phylogenetic trees coincide with transmission events. As these approaches require a high sampling fraction, the within-host genetic diversity can contribute to the genetic diversity observed between the sampled sequences. Because this contribution is ignored in the analyses, it can lead to incorrect inference of the transmission tree and biased estimates of parameters.

Here, we present a consistent way to use pathogen genetic sequence data in transmission tree reconstruction, by simultaneously estimating the phylogenetic tree of the pathogens. This requires inclusion of a model of within-host pathogen dynamics. In this article we describe the likelihood framework for such a joint estimation of the transmission tree and the phylogenetic tree. We investigate the performance and robustness of the methodology using simulations of an influenza outbreak in a confined setting. In these simulations, we also investigate the sensitivity of the outcome to unsampled or unobserved hosts. Finally, we illustrate the use of the method by applying it to a previously published data set on an outbreak of foot-and-mouth disease (FMD).

Methods

Joint likelihood of the transmission tree, phylogenetic tree, and within-host dynamics

We focus on outbreaks for which nearly all cases are observed, host characteristics such as time of symptom onset are known, and sequences are obtained from pathogens sampled from a proportion of the hosts. The transmission events are not observed. From these data we simultaneously estimate both the transmission tree and the phylogenetic tree, by writing down the likelihood for any pair of these trees. Throughout, we assume that the first infected host, or index case, introduced infection and all other infected hosts are infected by another host in this outbreak. We assume every host is infected at most once.

Define a transmission tree T as the set of all transmissions between infected hosts, including transmission times (Wallinga and Teunis 2004). The transmission times could be observed or unknown, in which case we estimate them simultaneously. The phylogenetic tree P is the usual dichotomous tree, with timed internal nodes. The function $W(t, h)$ gives a measure of within-host genetic diversity, as the product of the pathogen generation time and the within-host effective pathogen population size in host h at time t . $W(t, h)$ can be thought of as being proportional to the total number of virus particles in host h at time t . Let θ be the epidemiological parameters, μ the mutational parameters, and the data D consist of genetic sequences D_G and epidemiological data D_E .

The probability of the transmission tree T , phylogenetic tree P , within-host dynamics W , and parameters θ and μ can be found by first applying Bayes' theorem and then the chain rule of probability,

$$\begin{aligned} p(T, \theta, W, P, \mu | D_E, D_G) &\propto p(D_E, D_G | T, \theta, W, P, \mu) \\ &\quad \times \pi(T, \theta, W, P, \mu) \\ &= p(D_E | T, \theta, W, P, \mu) \\ &\quad \times p(D_G | D_E, T, \theta, W, P, \mu) \\ &\quad \times p(P | T, \theta, W, \mu) \pi(T, \theta, W, \mu), \end{aligned}$$

where π denotes the prior probability. We can further simplify this equation by using conditional dependencies. In particular, if we know the full transmission tree, epidemiological parameters, and within-host dynamics, the phylogenetic tree and mutational parameters give no further information on the probability of the epidemiological data:

$$p(D_E | T, \theta, W, P, \mu) = p(D_E | T, \theta, W).$$

Likewise, when the ancestry of sampled pathogens and the mutational parameters are known, the epidemiological data and parameters give no further information on the sequence data:

$$p(D_G | D_E, T, \theta, W, P, \mu) = p(D_G | P, \mu).$$

Also, the mutational and epidemiological parameters give no further information on the ancestry when the transmission

Figure 1 Schematic for viral dynamics. Throughout the figure, time progresses from left to right. Hosts are depicted as gray pods, virus particles as blue dots, and sampled virus particles as red dots. (A) The timing of coalescence of viral lineages depends on within-host viral dynamics. Virus (blue) numbers within hosts (gray) rapidly increase at onset of infection and decrease near the end of the infection, influencing coalescent rates. A possible ancestry between sampled viruses (red) is given in black. Although the initial host infects the latter two, the sampled viruses from these latter two are more closely related, as they coalesce with each other before coalescing with the virus sampled from the initial host. (B) When viruses are sampled from only a few hosts in a large outbreak, the timing of the coalescence of the sampled viruses is nearly identical to the timing of transmission immediately following the coalescence. The timing of coalescent events is mainly governed by interhost infection dynamics, and

the phylogenetic tree derived from the sequences (blue) is very similar to the one derived when internal node times are equated with transmission times (red). (C) When viruses are sampled from all hosts in an outbreak, coalescent times and transmission times are very different. The phylogenetic tree derived when approximating coalescent times by transmission times (red) is very different from the actual phylogenetic tree (blue).

tree and within-host dynamics are known (we assume that there is no information on selection pressures):

$$p(P|T, \theta, W, \mu) = p(P|T, W).$$

We thus get

$$p(T, \theta, W, P, \mu | D_E, D_G) \propto p(D_E | T, \theta, W) p(D_G | P, \mu) p(P | T, W) \pi(T, \theta, W, \mu). \quad (1)$$

The amount of prior information differs per application; prior information on parameters and within-host dynamics might be available due to previous studies and prior information on the transmission tree might be available through contact tracing.

We point out the similarity of Equation 1) to previous methodologies. The first term on the right-hand side is, up to the inclusion of the within-host model W , identical to the likelihood equations found in transmission tree reconstruction methods (Wallinga and Teunis 2004). The second and third terms together resemble the likelihood equations found in many phylodynamic approaches (Pybus and Rambaut 2009; Volz *et al.* 2013), with the difference that the epidemiological model in such approaches is replaced here by the combination of the transmission tree and the within-host model.

Numerical implementation: Sampling from the probability distribution

Equation 1 defines (up to a constant) a probability distribution on the space of transmission trees, phylogenetic trees, and parameters. We can sample from this distribution using Markov chain Monte Carlo (MCMC) methods. The initial state consists of a transmission tree, phylogenetic trees, and parameters with probability larger than 0. We construct the initial state as follows. We first ensure that every host has a time of infection by assigning one if the time was unobserved. We then construct a transmission tree by assigning an infector to all infected hosts, apart from the

host infected first. The infecting hosts are randomly taken from the set of hosts that are infectious at the time of infection of the infected host. We construct a phylogenetic tree that is consistent with this tree and take initial values for the parameters from their prior distributions.

In each iteration of the MCMC, the trees and parameters are updated. An iteration consists of five different steps:

- For a host infected in the outbreak, pick a new infector (in this step we also alter the phylogenetic tree, to make sure that it is compatible with the proposed transmission tree);
- consider a new root for the transmission tree, switching infection times between the current and proposed root;
- update infection times;
- per host, update the phylogenetic tree contained in that host; and
- update parameter values.

See supporting information, [File S1](#), for technical details. Below, we illustrate this general concept using simulated and real data.

Testing the approach on simulated data sets

To illustrate the method and assess robustness to missing data, we apply it to simulated data on an influenza outbreak in a confined school-based population of 200 individuals. Such confined outbreaks are amenable to analysis as, due to the low number of cases, individual infections can be tracked using only epidemiological data (Cauchemez *et al.* 2011; Hens *et al.* 2012).

Each simulation starts with 1 infected and 199 susceptible individuals, half of them children; only outbreaks with at least 50 cases were used for further analysis. We use a latent period of 2 days, and a gamma-distributed infectious period with a mean of 3 days and a variance of 2. During their infectious period, individuals exert an equal and constant force of infection on all susceptible individuals; *i.e.*, we assume

homogeneous mixing, such that the expected number of infections caused by an infectious individual is 0 for adults and 2 for children at the start of the outbreak. We assume that symptom onset coincides with the start of the infectious period. See [File S1](#) for details.

We take a simple analytically tractable function, a quartic kernel, for the product of pathogen generation time and effective population size,

$$W(t, h) = 1 + 1000 \left(1 - \left(\frac{2(t - t_h)}{T_h} - 1 \right)^2 \right)^2,$$

where T_h is the time between infection and recovery of host h , and $(t - t_h)$ is the time since infection t_h of h . Note that the model assumes an effective size of 1 at transmission, ensuring only one strain can infect a host. We take a substitution rate of 0.003 substitutions/site/year (Jenkins *et al.* 2002).

For each case, we record the time of symptom onset, the recovery time, a sequence sampled 1 day after symptom onset, and whether the case is a child or adult. These data are then used for inference, as described below.

Inference

We estimate the transmission tree and parameters θ and μ using the proposed likelihood Equation 1. To this end, we need to specify the three components of the likelihood.

Likelihood component for the transmission tree: We assume that the incubation period, defined as the time between infection and developing symptoms, is gamma distributed with mean 2 and variance 1. The likelihood of the transmission tree then becomes the product over all infected hosts of the infectiousness of the infecting host relative to the total infectiousness of all hosts times the probability distribution s for the length of the incubation period,

$$p(D_E | T, \theta, W) = \prod_{x \in H^-} s(o_x - t_x) \frac{I_{t_x}(v(x) | \theta, D_E)}{\sum_{y \in H} I_{t_x}(y | \theta, D_E)},$$

where H is the set of all infected hosts, H^- is the set of all infected hosts minus the index case, t_x and o_x are the start of the infection and the start of the infectious period (respectively) of host x , $v(x)$ is the infector of host x , and $I_t(a | \theta, D_E)$ is the infectiousness of host a at time t . Here we take infectiousness to be the rate at which a host infects any other host. The infectiousness $I_t(a | \theta, D_E) = \theta$ if $t > s_a$ and a has not recovered at time t , 0 otherwise. Note that the infectiousness does not depend on the infected host, as we assume homogeneous mixing and we do not want to *a priori* assume a difference in infectiousness between adults and children.

Likelihood component for the phylogenetic tree: The likelihood of the phylogenetic tree can be obtained from coalescent theory for haploid organisms. Going backward in

time, two events that alter the number of lineages of the phylogenetic tree contained within one host can take place. First, the number can decrease by one due to a coalescent event. Second, the number can increase by one due to an incoming lineage: a transmission event from this host to another, akin to a new sample in standard coalescent models. The first case is described by the likelihood that the lineages did not coalesce for a time, and finally coalesced; the second case is described by the likelihood that the lineages did not coalesce. Using forward time we get

$$p(P | T, W) = \prod_{x \in H} \prod_{[\tau_1, \tau_2] \in C_x} W(\tau_1, x)^{-1_{\text{coal}}} e^{-\binom{n_\tau}{2} \int_{\tau_1}^{\tau_2} (1/W(t, x)) dt},$$

where H is the set of all infected hosts and C_x is the set of all intervals $\tau = [\tau_1, \tau_2]$, where the number n_τ of viral lineages within host x with sampled offspring is constant and >1 . The indicator function 1_{coal} is 1 if the interval starts with a coalescent, 0 otherwise. Here, we assume W to be fully known.

Likelihood component for the mutational parameters: We take the simplest feasible substitution model; all mutations are equally likely and happen with rate μ . We therefore get

$$p(D_G | P, \mu) = \prod_{\text{bases}} \sum_{\{A, C, G, T\}^N} \prod_{\text{edges}} (1 - e^{-\mu t})^{1_{\text{mut}}} + (e^{-\mu t})^{(1 - 1_{\text{mut}})},$$

where the first product is over all base pairs, the sum is over all possible assignments of each of the nucleotides to each of the N internal nodes, the second product is over all branches of the phylogenetic tree, t denotes branch length, and the indicator denotes whether a mutation occurred on that branch. We omit the Jukes–Cantor correction because time-scales are very short and selection pressures absent, and hence the probability of multiple mutations at the same locus is negligible (see [File S1](#)). We can avoid having to sum over all 4^N possible states using Felsenstein’s pruning algorithm (Felsenstein 1981). Inclusion of other, perhaps more realistic, substitution models would be straightforward.

Evaluating performance

We examine how well the transmission tree can be reconstructed by evaluating the probability assigned to the actual transmission events. We estimate the probability that host j infected host i by the proportion of sampled trees in which j infected i . We assess false positives by evaluating, for each infected host, whether the infector assigned the highest probability is the actual infector. We further evaluate how well the method estimates the substitution rate and whether the method is able to find the reduced infectiousness of adults. The latter is done by counting the fraction of infections caused by adults among transmission events estimated

Figure 2 Accuracy of estimating transmission trees using genetic sequences of pathogens, for different simulation scenarios. Solid lines give the average percentage of infected hosts for which the actual infector has been assigned a probability of at least the level indicated on the x-axis. Dashed lines give the average percentage of infected hosts for which the infector has been incorrectly identified, with a probability at least the level indicated on the x-axis. (A) Results when 100% (black), 50% (blue), or 0% (pink) of infected hosts have been sampled. When fewer hosts are sampled, only a few infectors can be identified at a high probability level. (B) Results when all (black) or 80% (turquoise) of hosts are observed. When fewer hosts are observed, fewer infectors are identified correctly, and incorrect inferences are made even at high probability levels. (C) Results when substitution rate is 3×10^{-3} (black), an increased 1×10^{-2} (yellow), or a decreased 1×10^{-3} substitutions/site/year (green). At higher substitution rates the inference is more accurate. (D) Results when coalescent events are allowed to differ from transmission events (black) or when coalescent events are incorrectly assumed to coincide with transmission events (orange). The incorrect assumption leads to incorrect estimations even at a high probability level.

at probability at least 0.9. As we are interested in the parameters and transmission tree, the phylogenetic tree can be considered a nuisance parameter, and we do not further investigate its estimation.

To assess robustness of the estimation procedure, we simulate data, estimate parameters and evaluate performance for seven different simulation scenarios. In the baseline scenario, all data are available. In the second and third scenario, we examine the impact of incomplete sampling by randomly discarding 0 or 50% of sequences. In a fourth scenario, we examine the impact of unobserved hosts by randomly discarding 20% of infected hosts. To keep the number of observed hosts comparable, we start these simulations with 20% more susceptibles. In a fifth scenario, we examine sensitivity to an increased substitution rate of 0.01 substitutions/site/year. In a sixth scenario, we examine sensitivity to a decreased substitution rate of 0.001 substitutions/site/year. In a seventh scenario, we examine the impact of a misspecified within-host model, by setting $W(t, h) = 1$ in the analysis. Specifying this within-host model is equivalent to making the

Table 1 Accuracy of estimating transmission trees from genetic sequences of pathogens

	Average probability assigned to actual transmission events
All data	50
50% sampled	31
0% sampled	24
80% observed	44
High substitution rate	77
Low substitution rate	37
Coalescent at transmission	48

incorrect assumption that coalescent events coincide with transmission events.

Application to an outbreak of foot-and-mouth disease

To illustrate the use of the method in practice, we reanalyze data on an outbreak of FMD in 2001 in Durham County, England (Cottam *et al.* 2006, 2008; Morelli *et al.* 2012). Here, for a cluster of 12 infected farms, spatial information, date of symptom onset, culling date, and one full genome sequence is known for all 12 farms in the cluster. Both the phylogenetic tree and transmission tree have been estimated before for this outbreak separately, giving inconsistent results (Cottam *et al.* 2008). We apply our method to illustrate how to estimate both trees simultaneously, using the same epidemiological model and substitution model as were used in a previous study (Morelli *et al.* 2012). We assume an exponentially increasing function for W (see [File S1](#) for details), as at time of culling the disease is still spreading within the farms.

Results

We first test the proposed method by estimating the transmission tree from the simulated data sets. Figure 2 shows how well individual transmissions can be estimated, and Table 1 shows the average probability assigned to the actual transmissions. When all data are available, on average half of the transmissions in the reconstructed transmission tree are correct (Figure 2); this number is almost one if we restrict ourselves to transmission events assigned a high probability. Few transmissions can be correctly estimated when the percentage of infected hosts sampled decreases. When instead the percentage of hosts observed decreases, transmissions can be estimated correctly quite often, but the percentage of incorrectly estimated transmissions increases strongly. More transmission events can be estimated when the substitution rate is higher and fewer transmission events can be estimated when the substitution rate is lower. Nearly as many transmission events can be correctly estimated when we assume that coalescent events coincide with transmission events as when we know the correct within-host model. However, the percentage of transmissions that is incorrectly estimated increases.

We now turn to estimation of mutational parameters from the simulated data sets. The substitution rate can be

Figure 3 Robustness of estimates of genetic and epidemiological parameters under various simulation scenarios. (A) Distribution of point estimates of the substitution rates for 100 simulations, for three simulation scenarios. Actual value is 0.003 (black line). Estimates are accurate when all information is available (gray) and when 50% of hosts are sampled (blue), although the latter leads to a broader distribution. Assuming that coalescent events coincide with transmission events (pink), however, leads to a large overestimation, since the total branch length of the phylogenetic tree is underestimated. Mean estimates are 3.3×10^{-3} , 3.3×10^{-3} , and 6.7×10^{-3} , respectively. (B) Point estimate (black) and 95% confidence interval (gray) of the fraction of infections due to adults, where actual value is 0. Shown are 100 sorted estimates, for each of seven scenarios

(complete data, missing sequences, unobserved hosts, altered substitution rate, and incorrect within-host model). Estimates are away from the actual value of 0 when only 80% of hosts are observed or when coalescent times are equated with transmission times. The width of the confidence interval depends largely on the amount of information available; e.g., when less genetic information is available due to incomplete sampling; the point estimates are accurate, but the confidence interval can become very broad.

estimated well even when data are missing (Figure 3A); the mean estimated rate was 0.0033, and the actual value of 0.003 was contained in the 95% credibility interval for 90% of the simulations. Assuming that coalescent events coincide with transmission leads to a large overestimation (mean 0.0067, coverage 15%); this is because the total branch length of the phylogenetic tree is underestimated, leading to an overestimate of the substitution rate (Figure 1).

Next, we focus on performance in estimating the relative infectiousness of adults from the simulated data sets. Figure 3B gives, for each of 100 simulations under each of the seven scenarios, the point estimate and confidence interval for p , the fraction of infections caused by adults. Most estimates of p are close to the correct value of 0, except for the scenarios where only 80% of hosts are observed (mean $P = 0.05$), or when transmission times are equated with coalescent times (mean $P = 0.11$). More importantly, the confidence interval for p becomes larger when there is more uncertainty about the transmission tree, most notable under the two scenarios with decreased sampling rates.

Having tested the method, we apply it to infer transmission trees from epidemiological and genetic sequence data collected during an FMD outbreak in Durham County, England, in 2001. Figure 4A illustrates a typical sample from the MCMC, consisting of a transmission chain connecting the infected farms and a phylogenetic tree connecting the sampled sequences. The results are broadly similar to those obtained from previous analyses on the same data set (File S1). The mean latency period was estimated at 7.8 (95% credibility interval, CI: 4.9, 12) days (Figure 4), which is in agreement with a typical latency period of 5 days (95% CI: 1–12 days) of FMD virus (Keeling *et al.* 2001; Gibbens and Wilesmith 2002; Charleston *et al.* 2011). Ignoring the within-host genetic diversity would have led to an unrealistically large estimate for the latency period of 24 (95% CI: 17,

35) days (Morelli *et al.* 2012). The substitution rate was estimated at 1.1×10^{-2} (95% CI: 8.7×10^{-3} , 1.5×10^{-2}) substitutions per site per year, which is higher than a typical value of 7.7×10^{-3} based on genetic data only (Cottam *et al.* 2008). We found the value to be sensitive to the precise specification of the within-host dynamics. See File S1 for further results and comparison to previous estimates. Together, these results confirm that the method is able to estimate plausible parameter values, while reconstructing the transmission tree and a consistent phylogenetic tree of the outbreak.

Discussion

We have shown how the within-host dynamics of pathogens relate the phylogenetic tree of sampled pathogens to the transmission tree of an outbreak of an infectious disease. We use this relationship to estimate key parameters by combining genetic data on the pathogen with epidemiological data. The advantage of this estimation procedure is that estimates are more accurate, and estimation is feasible even when hosts are unsampled or unobserved.

The method is able to correctly estimate epidemiological and mutational parameters and can infer individual transmission events. Although the methods perform best when all infected hosts have been observed and sampled, simulations show that even when 20% of infected hosts are unobserved the transmission tree can still be estimated reasonably well. Note, however, that estimation of some epidemiological parameters might become biased. For example, in the FMD example, the slight overestimation of both duration of the latency period and substitution rate could well be due to unobserved farms (Morelli *et al.* 2012). As expected, the accuracy of the estimates increases with substitution rate. The method therefore seems most suited to analyzing data

Figure 4 Results from the analysis on the foot-and-mouth disease data sets. (A) A typical transmission tree sampled from the MCMC. Shown are infected farms (labeled pods), their latent periods (gray) and infectious periods (green), samples viruses (red), and the phylogenetic tree connecting these viruses (black). The phylogenetic tree is contained within the transmission tree; due to the exponentially increasing within-host effective pathogen population size assumed, most coalecscents occur early during an infection. (B) Posterior distribution for the mean latency period β_1 . Solid black line gives the median, and dashed lines give the 2.5th and 97.5th percentile. Blue line gives a previous estimate from the literature, and green line gives the estimate derived from the same data set in

a previous study that ignored within-host genetic diversity. The estimate (solid black) is higher than we would expect from the literature (blue). The overestimation could be due to unobserved infected farms. Not allowing for within-host genetic diversity gives an overestimation (green). (C) Posterior distribution for the substitution rate μ . Solid black line gives the median, dashed lines give the 2.5th and 97.5th percentile, and blue line gives a previous estimate from the literature. The higher estimate we obtained could be due to an overly simplified within-host model.

sets on RNA viruses. Analysis for DNA viruses or bacteria might still be feasible if the times between infection are large enough, as sufficient genetic diversity might still accumulate over the course of the outbreak.

To correctly estimate both the transmission tree and the phylogenetic tree, knowledge on the within-host effective pathogen population size and pathogen generation time is needed. This knowledge will, however, not be available in general. On the short timescales we are considering, we expect it will be challenging to estimate the shape of the time-varying population size from the data. Conversely, the impact of a misspecification of the within-host effective pathogen population size is minor. Even the extreme case of taking a constant population size of 1 (*i.e.*, equating transmission times with coalescent times) leads to reasonable estimation of the transmission tree. We see this in our application to FMD; although the substitution rate is overestimated, the estimated epidemiological parameters agree very well with previous estimates. It is possible that within-host dynamics are easier to estimate for chronic infections. However, for chronic infections, the epidemiological data are usually less informative of the transmission tree.

In proposing this methodology, we extend previously proposed methods that aim to reconstruct transmission trees using both epidemiological and genetic data. The field was pioneered by Cottam *et al.* (2008), who considered a sequential estimation procedure for the phylogenetic tree and the transmission tree, an approach that leads to loss of information contained in the phylogenetic tree. More recent approaches considered both data types simultaneously (Jombart *et al.* 2011; Morelli *et al.* 2012; Ypma *et al.*

2012), but implicitly or explicitly associated sequences with hosts, rather than with individual pathogens within the host. This means that transmission times coincide with coalescent times, an approximation that is inaccurate when the sampling fraction is high. Our simulations show that this leads to suboptimal inference of the transmission tree and to a large overestimation of the substitution rate.

Transmission tree reconstruction methods focus on individual transmission events, whereas many published phylogenetical approaches (Pybus and Rambaut 2009) typically focus on finding general characteristics of an outbreak, *e.g.*, epidemiological parameters, from sampled sequences. To estimate individual transmissions, detailed data are needed; we expect that the number of outbreaks that are studied in such detail will increase. More importantly, transmission tree reconstruction allows for an understanding of the outbreak at a high level of detail; for example, hypotheses regarding the transmission mechanism can be tested using the reconstructed transmission tree (Ypma *et al.* 2013).

Reconstructing large outbreaks at the detailed level of individual transmissions is feasible only when highly informative data are available. These could take the form of detailed epidemiological data on who infected whom, informative genetic data, *i.e.*, a large number of sampled sequences exhibiting high genetic diversity, or a combination of both. The method we have presented uses both data types to estimate simultaneously the transmission tree and the phylogenetic tree, acknowledging the fact that these two are, in fact, different. With the decreasing cost of sequencing technologies it is likely that more and more of such detailed

molecular epidemiological data sets will become available for a range of pathogens, a trend we have already seen in the past few years (Bataille *et al.* 2011; Gardy *et al.* 2011; Harris *et al.* 2012). We therefore expect that the usefulness of this combination of two historically separated fields will become even more apparent in years to come.

Literature Cited

- Bataille, A., F. Van Der Meer, A. Stegeman, and G. Koch, 2011 Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathog.* 7: e1002094.
- Cauchemez, S., A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff *et al.*, 2011 Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl. Acad. Sci. USA* 108: 2825–2830.
- Charleston, B., B. M. Bankowski, S. Gubbins, M. E. Chase-Topping, D. Schley *et al.*, 2011 Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science* 332: 726–729.
- Cottam, E. M., D. T. Haydon, D. J. Paton, J. Gloster, J. W. Wilesmith *et al.*, 2006 Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *J. Virol.* 80: 11274–11282.
- Cottam, E. M., G. Thebaud, J. Wadsworth, J. Gloster, L. Mansley *et al.*, 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Biol. Sci.* 275: 887–895.
- Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.
- Ferguson, N. M., C. A. Donnelly, and R. M. Anderson, 2001 Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* 413: 542–548.
- Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah *et al.*, 2011 Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* 364: 730–739.
- Gibbens, J. C., and J. W. Wilesmith, 2002 Temporal and geographical distribution of cases of foot-and-mouth disease during the early weeks of the 2001 epidemic in Great Britain. *Vet. Rec.* 151: 407–412.
- Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303: 327–332.
- Harris, S. R., E. J. Cartwright, M. E. Torok, M. T. Holden, N. M. Brown *et al.*, 2012 Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* 13: 130–136.
- Haydon, D. T., M. Chase-Topping, D. J. Shaw, L. Matthews, J. K. Friar *et al.*, 2003 The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. Biol. Sci.* 270: 121–127.
- Heijne, J. C., P. Teunis, G. Morroy, C. Wijkman, S. Oostveen *et al.*, 2009 Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerg. Infect. Dis.* 15: 24–30.
- Heijne, J. C., M. Rondy, L. Verhoef, J. Wallinga, M. Kretzschmar *et al.*, 2012 Quantifying transmission of norovirus during an outbreak. *Epidemiology* 23: 277–284.
- Hens, N., L. Calatayud, S. Kurkela, T. Tamme, and J. Wallinga, 2012 Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. *Am. J. Epidemiol.* 176: 196–203.
- Holmes, E. C., S. Nee, A. Rambaut, G. P. Garnett, and P. H. Harvey, 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 349: 33–40.
- Jenkins, G. M., A. Rambaut, O. G. Pybus, and E. C. Holmes, 2002 Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* 54: 156–165.
- Jombart, T., R. M. Eggo, P. J. Dodd, and F. Balloux, 2011 Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 106: 383–390.
- Keeling, M. J., M. E. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping *et al.*, 2001 Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294: 813–817.
- Keeling, M. J., M. E. Woolhouse, R. M. May, G. Davies, and B. T. Grenfell, 2003 Modelling vaccination strategies against foot-and-mouth disease. *Nature* 421: 136–142.
- Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz, 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* 438: 355–359.
- Maddison, W. P., and L. L. Knowles, 2006 Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55: 21–30.
- Morelli, J. M., G. Thebaud, J. Chadoeuf, D. P. King, D. T. Haydon *et al.*, 2012 A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLOS Comput. Biol.* 8: e1002768.
- Pybus, O. G., and A. Rambaut, 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10: 540–550.
- Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes *et al.*, 2001 The epidemic behavior of the hepatitis C virus. *Science* 292: 2323–2325.
- Rasmussen, D. A., O. Ratmann, and K. Koelle, 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLOS Comput. Biol.* 7: e1002136.
- Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3: 380–390.
- Spada, E., L. Sagliocca, J. Sourdis, A. R. Garbuglia, V. Poggi *et al.*, 2004 Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J. Clin. Microbiol.* 42: 4230–4236.
- Stadler, T., 2009 On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* 261: 58–66.
- Volz, E. M., S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. Frost, 2009 Phylodynamics of infectious disease epidemics. *Genetics* 183: 1421–1430.
- Volz, E. M., K. Koelle, and T. Bedford, 2013 Viral phylodynamics. *PLOS Comput. Biol.* 9: e1002947.
- Wallinga, J., and P. Teunis, 2004 Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* 160: 509–516.
- Ypma, R. J., A. M. Bataille, A. Stegeman, G. Koch, J. Wallinga *et al.*, 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. Biol. Sci.* 279: 444–450.
- Ypma, R. J., M. Jonges, A. M. Bataille, A. Stegeman, G. Koch *et al.*, 2013 Genetic data provide evidence for wind-mediated spread of highly pathogenic avian influenza. *J. Infect. Dis.* 207: 730–735.

Communicating editor: M. A. Beaumont

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.154856/-/DC1>

Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks

Rolf J. F. Ypma, W. Marijn van Ballegooijen, and Jacco Wallinga

File S1

Supporting information

Relating phylogenetic trees to transmission trees of infectious disease outbreaks

R.J.F. Ypma, W.M. van Ballegooijen, J. Wallinga

Contents

1	Test of estimation procedure on simulated data.	2
1.1	Simulating outbreaks	2
1.2	Simulated data	3
2	Application of the estimation procedure to data on foot-and-mouth disease.	3
2.1	Data	3
2.2	Model	3
2.3	Results	4
3	Sampling from the joint posterior distribution using MCMC	6
3.1	Initial state	6
3.2	Update phylogenetic tree	6
3.3	Update transmission tree	7
3.4	Update infection times	8
3.5	Update epidemiological parameters	8
3.6	Update mutational parameters	8
3.7	Convergence	9
3.8	Estimation	9
	References	10

1 Test of estimation procedure on simulated data.

1.1 Simulating outbreaks

To illustrate the method and assess robustness to missing data, we apply it to simulated data based on an influenza outbreak in a confined school-based population. A common question in such a setting is whether one group of individuals (e.g. children) is more infectious than another (e.g. adults). To emphasize the difference, we performed simulations where only children were infectious.

We start with a population of 200 hosts, of which one is infected. Each individual belongs to one of two groups (children and adults). The outbreak is generated using a simple stochastic SIR model with homogeneous mixing in continuous time. For each infected host x an infectious period of length l_x is drawn from a gamma distribution with a mean of three days and a variance of two days [2]. The infectiousness of host x is then

$$B_t(x) = \begin{cases} b_x & \text{if } t_x + 2 < t < t_x + 2 + l_x \\ 0 & \text{else} \end{cases}$$

where $b_x = \frac{2}{3 \times 200}$ if x is a child (i.e. $R_0 = 2$), 0 otherwise. In the inference procedure described in the main text, the function $I_t(a)$ is used to estimate the infectiousness.

Using the (known) transmission tree, we generate sequences as follows; we assume each infected individual is sampled one day after symptom onset ($t_x + 2 + 1$). We construct a phylogenetic tree using these timepoints as tips, by backwards simulation of the tree using $p(P|T, W)$ as specified in the main text. We then generate point mutations on this tree using a simple molecular clock; the number of mutations per edge of the phylogenetic tree is Poisson distributed with expected value $L \times \mu \times t = 10000 \times 0.003 \times t$ for the baseline scenario, with $L = 10000$ the number of base pairs and $\mu = 0.003$ the substitution rate. Note that the approximation of a Poisson distribution is valid here due to the short timescales. In fact, the probability of a recurrent mutation occurring is approximately equal to the probability any nucleotide mutates twice in 5 days, times the number of nucleotides, which is $(0.003 \times \frac{5}{365})^2 \times 10000 \approx 1.7 \times 10^{-5}$.

A quantity of interest is the genetic distance between the sequences sampled from an infected host and its infector, as this is ultimately the data we are working with. The evolutionary time separating the samples taken from a host and its infector is roughly between 3 and 11 days, so the expected number of mutations separating the sequences is between $10000 \times 0.003 \times \frac{3}{365} = 0.25$ and $10000 \times 0.003 \times \frac{11}{365} = 0.9$, which seems plausible for an RNA-virus [7, 1, 3].

We investigated a total of seven scenarios. To improve comparability between results from the different scenarios, we re-used the same simulated outbreaks for the different scenarios. This re-usage eliminates variation resulting from the stochastic nature of the simulations. For example, the same phylogenetic tree was used in different scenarios to generate sequences based on different substitution rates. The exception is formed by the scenario where 20% of cases were discarded. As we wanted to keep the number of infected individuals comparable, separate simulations were performed for this scenario, which started with an increased susceptible population of $200/(1 - 0.2) = 250$.

1.2 Simulated data

For each case, we assume we know the time of symptom onset, the recovery time, a sequence sampled one day after symptom onset, and whether the case is a child or adult. The likelihood equations used in the MCMC are given in the main text. For efficient calculation of the likelihood $p(D_G|P, \mu)$ of the phylogenetic tree and mutation rate given the sequences we use Felsenstein's pruning algorithm [5]. We use a Uniform(0,1000) prior for all parameters.

2 Application of the estimation procedure to data on foot-and-mouth disease.

2.1 Data

In 2001, a large epidemic of foot-and-mouth disease (FMD) occurred in the United Kingdom. A subset of 15 farms of this large epidemic, the so-called 'Darlington cluster' has been extensively studied [3, 4, 8]. Three of the farms are not epidemiologically linked to the other 12, and subsequently dropped from the analysis [8]. The remaining 12 farms are labelled $\mathcal{F} = \{C, D, E, F, G, H, I, J, K, L, M, O\}$. For each of the farms $i \in \mathcal{F}$ we have

- T_i^{obs} , the date of detection of the virus,
- D_i^{obs} , the estimated age of infection at date of infection, as assessed by a visual exam of the clinical state of lesions found,
- T_i^{end} , the date of culling,
- $\mathbf{x}_i$, the spatial location (as latitude/longitude),
- S_i^{obs} , an 8000 bp DNA sequence sampled at T_i^{obs} .

All these data are freely available from the open-access publication by Morelli et al. [8].

2.2 Model

The likelihood component for the transmission tree is based on an epidemiological model used by Morelli et al. [8]. In particular, after infection farms enter a latent period, whose duration L_i is gamma distributed with expectation β_1 and variance β_2^2 . After the latent period, the farms enter an infectious period, as infected animals develop lesions. After a certain time D_i , the farm is detected as being infected. The data contain an estimate of D_i , which is assumed to be gamma distributed with mean D_i^{obs} and variance $D_i^{obs}/4$. Infectiousness of farms is assumed to stay constant during the infectious period, but to decrease exponentially with distance, with a mean transmission distance of $2\alpha_2$. A vague prior is assumed for all epidemiological parameters; an exponential distribution with

mean 100. Letting the epidemiological data D_E consist of the four vectors containing the farm-level epidemiological data, $D_E = (T^{obs}, D^{obs}, T^{end}, \mathbf{x})$, we get

$$p(D_E|T, \theta, W) = \prod_{i \in \mathcal{F}^-} \frac{e^{-\frac{|\mathbf{x}_i - \mathbf{x}_{v(i)}|}{\alpha_2}}}{2\pi\alpha_2^2} \mathbb{1}_{t_{v(i)} + L_{v(i)} < t_i < T_{v(i)}^{end}} \prod_{i \in \mathcal{F}} g(L_i, \beta_1, \beta_2^2) g(D_i, D_i^{obs}, \frac{D_i^{obs}}{4})$$

where $\mathcal{F}^-$ is the set of all farms except the index case, t_i is the time of infection of farm i , $v(i)$ is the infector of farm i , and $g(x, m, w)$ is the probability density function of the gamma distribution with mean m and variation w .

In this application, the hosts we consider are actually farms. These farms themselves contain a large number of animals. Per farm, several animals were infected [4]. We can thus view the host (i.e. farm) to be strongly compartmentalized, resulting in most coalescents occurring just after infection. We therefore take the within-host pathogen effective population size times pathogen generation time W at time t in host h to be exponentially increasing

$$W(t, h) = e^{r(t-t_h)}$$

if h is infected at time t , 0 otherwise. We set the growth rate r at 1.02.

With this model, we get

$$\begin{aligned} L(P|T, W) &= \prod_{x \in \mathcal{F}} \prod_{[\tau_1, \tau_2] \in C_x} W(\tau_1, x)^{-\mathbb{1}_{coal}} e^{-\binom{n_\tau}{2} \int_{\tau_1}^{\tau_2} \frac{1}{W(t, x)} dt} \\ &= \prod_{x \in \mathcal{F}} \prod_{[\tau_1, \tau_2] \in C_x} e^{-(r(\tau_1 - t_x) \mathbb{1}_{coal})} e^{-\binom{n_\tau}{2} (\frac{1}{r} e^{-r(\tau_1 - t_x)} - \frac{1}{r} e^{-r(\tau_2 - t_x)})} \end{aligned}$$

where τ_1 and τ_2 are the start and end of coalescent intervals.

Following [8] we use a one parameter substitution model, yielding

$$p(D_G|P, \mu) = \prod_{bases} \sum_{\{A, C, G, T\}^N} \prod_{edges} (1 - e^{-\mu t})^{\mathbb{1}_{mut}} + (e^{-\mu t})^{(1 - \mathbb{1}_{mut})}$$

For each locus, this equation sums over all 4^N possible states at the internal nodes, multiplying the transition probabilities over all edges. The probability of remaining in the same state during time t in this very simple model is $e^{-\mu t}$: inserting more elaborate substitution models here would be straightforward. Not all 4^N states have to be considered as we can traverse the tree from the leaves up using Felsenstein's pruning algorithm [5], which makes use of the conditional independencies inherent in tree structures. The prior for the substitution rate μ is uniformly distributed on (0,1000).

2.3 Results

Posterior distributions for the transmission tree and epidemiological parameters (α_2, β_2) are given in figure S1. To allow for comparison to previous studies, we also show the two transmission trees estimated before [4, 8], and the point estimates previously obtained for the parameters [8].

Figure S1. Posterior distribution for the transmission tree, mean transmission distance $2\alpha_2$ and standard deviation of latency duration β_2 . (Top) Posterior distributions for the transmission tree, with infector farm on the x-axis and infected farm on the y-axis. The size of the blobs corresponds to the posterior probability assigned to this pair. Estimates are given for the proposed method (left) and taken from previous studies (right) by Cottam et al. (pink) and Morelli et al. (green)[4, 8] which ignore within-host dynamics. The transmission tree cannot be established with high certainty for this outbreak, possibly due to unobserved infected farms. Allowing for within-host dynamics captures this uncertainty, in contrast to previous analyses (top right) which yield trees mainly containing pairs with posterior probability close to one. (Bottom) Posterior distributions for the mean transmission distance $2\alpha_2$ (left) and standard deviation of latency duration β_2 (right). Lines are point estimate (black solid), 95% credibility interval (black striped) and previous estimates obtained by Morelli et al. (green)

3 Sampling from the joint posterior distribution using MCMC

3.1 Initial state

As an initial state, a random permissible state was chosen, as follows. For each host x but the first, we chose a random infector $v(x)$ from the set of hosts infectious at the time of infection of x . We then constructed a phylogenetic tree consistent with this transmission tree (thus ignoring sequence data). Initial values for parameters were randomly sampled from their prior distributions.

3.2 Update phylogenetic tree

We update the phylogenetic tree P per host. We choose a host x at random, and update both the topology of the part of P contained within x , and the timing of the internal nodes contained in x . The number of internal nodes n_{i_x} of P contained in x is equal to the number of pathogen lineages that coalesce within x minus one, which is equal to the number of sequences n_{s_x} sampled from x plus the number of hosts infected by x that have a sequence minus one. More formally, let $T(x)$ be the smallest subtree of T such that

- $x \in T(x)$,
- for any node $i \in T$ other than x , if its infector $v(i) \in T(x)$, then $i \in T(x)$.

So $T(x)$ is the subtree that consists of x and the hosts directly or indirectly infected by x . Let $l(T)$ be 1 if T contains at least one host which has at least one sampled sequence, 0 otherwise. Then the number of coalescent events c_x that happen within host x is

$$c_x = \max \left\{ 0, n_{s_x} - 1 + \sum_{y \in T} \mathbb{1}_{v(y)=x} l(T(y)) \right\}$$

If $c_x = 0$, we do nothing. Otherwise, we propose a phylogenetic tree P^* which differs from P only in the nodes contained in x . Let $f(x)$ be the likelihood component relating to the coalescent structure in host x :

$$f(x) = \prod_{[\tau_1, \tau_2] \in C_x} W(\tau_1, x)^{-1_{coal}} e^{-\binom{n_\tau}{2} \int_{\tau_1}^{\tau_2} \frac{1}{W(t, x)} dt}$$

where C_x is the set of all intervals $\tau = [\tau_1, \tau_2]$ where the number n_τ of viral lineages within host x with sampled offspring is constant and larger than 1. We sample from this joint distribution on the node times by iteratively drawing a coalescent interval, and selecting a random pair of lineages to coalesce. Note that an extra term $\binom{n_\tau}{2}$ is used in drawing the coalescent intervals, as these are exponentially distributed with rate $\frac{\binom{n_\tau}{2}}{W(t, x)}$. At coalescence, each of the $\binom{n_\tau}{2}$ pairs is chosen with equal probability, which means the term $\binom{n_\tau}{2}$ drops out again. If the sample should have multiple lineages at time of

infection, it is redrawn. Note that a slightly more efficient sampling might be obtainable by choosing non-random pairs of lineages such that $p(D_G|P^*, \mu)$ is optimised. We accept P^* with probability

$$\begin{aligned} p(\text{accept}) &= \min\left\{1, \frac{p(D_E|T, \theta, W)p(D_G|P^*, \mu)p(P^*|T, W)\pi(T, \theta, W, \mu)}{p(D_E|T, \theta, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta, W, \mu)} \frac{q(P|P^*)}{q(P^*|P)}\right\} \\ &= \min\left\{1, \frac{p(D_G|P^*, \mu)f^*(x)}{p(D_G|P, \mu)f(x)} \frac{\alpha(n_{i_x})f(x)}{\alpha(n_{i_x})f^*(x)}\right\} \\ &= \min\left\{1, \frac{p(D_G|P^*, \mu)}{p(D_G|P, \mu)}\right\} \end{aligned}$$

where $\frac{q(P|P^*)}{q(P^*|P)}$ is the Metropolis-Hastings ratio.

3.3 Update transmission tree

We randomly choose a host x . If x is not the index case, let $v(x)$ be its infector. We propose a new infector $v^*(x)$ randomly from the set of hosts that are

- infectious at time of infection t_x ,
- not $v(x)$,
- not x (although this is usually already ensured by the first condition).

We could then try to construct the new transmission tree T^* generated by accepting $v^*(x)$ as the new infector of x . Note that T^* contains no cycles, as hosts are never infectious before being infected.

However, the phylogenetic tree will in general no longer correspond to T^* . More precisely, if both $l(T(x)) = 1$ and $l(T - T(x)) = 1$, then there must be an edge of the phylogenetic tree with one node contained in a host in $T - T(x)$, and one in $T(x)$. Let y be the host in which the older of these two nodes is contained. If all hosts have sequences, the nodes are contained in $y = v(x)$ and x . In general, this edge causes P to be inconsistent with the new tree T^* . We therefore simultaneously propose a new tree P^* that is consistent with T^* , by removing the older of the two nodes of this edge from y , and relocating it to the first host y^* in the infection chain leading up from $v^*(x)$ (i.e. $\{v^*(x), v(v^*(x)), v(v(v^*(x)))\}$) that contains at least one pathogen lineage. This is a particular form of a pruning and regrafting operator [6], also see figure S2. If all hosts are sequenced, $y^* = v^*(x)$. If no such edge exists, we take $f(y)$ and $f(y^*)$ below to be 1. We then update the phylogenetic trees contained in y and y^* as described above. The proposed pair of trees (T^*, P^*) is accepted with probability

$$\begin{aligned} p(\text{accept}) &= \min\left\{1, \frac{p(D_E|T^*, \theta, W)p(D_G|P^*, \mu)p(P^*|T^*, W)\pi(T^*, \theta, W, \mu)}{p(D_E|T, \theta, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta, W, \mu)} \frac{q(T, P|T^*, P^*)}{q(T^*, P^*|T, P)}\right\} \\ &= \min\left\{1, \frac{s(o_x - t_x) \frac{I_x(v^*(x)|\theta, D_E)}{\sum_{y \in H} I_x(y|\theta, D_E)} p(D_G|P^*, \mu) f^*(y) f^*(y^*) \pi(T^*)}{s(o_x - t_x) \frac{I_x(v(x)|\theta, D_E)}{\sum_{y \in H} I_x(y|\theta, D_E)} p(D_G|P, \mu) f(y) f(y^*) \pi(T)} \frac{f(y) f(y^*)}{f^*(y) f^*(y^*)}\right\} \\ &= \min\left\{1, \frac{I_x(v^*(x)|\theta, D_E) p(D_G|P^*, \mu)}{I_x(v(x)|\theta, D_E) p(D_G|P, \mu)}\right\} \end{aligned}$$

as we take $\pi(T)$ equal for all trees.

If x is the index case, we let $v^*(x)$ be the first host it infects, and we switch infection times for the two hosts [8]. We then follow the procedure above.

Figure S2. Example of a proposal transmission tree/phylogenetic tree (T^*, P^*) pair. A new infecting host $v^*(x)$ is proposed for host x . A new phylogenetic tree is then also proposed, by removing the branch from $v(x)$ and drawing a new phylogenetic subtree for $v^*(x)$ which includes the branch.

3.4 Update infection times

If not all infection times are observed, pick a host x at random and propose a new infection time t_x^* , sampled from the latency period distribution s . This creates the proposal transmission tree T^* . If $v(x)$ is not infectious at t_x^* , or T^* is inconsistent with P , reject. If not, accept with probability

$$\begin{aligned}
 p(\text{accept}) &= \min\left\{1, \frac{p(D_E|T^*, \theta, W)p(D_G|P, \mu)p(P|T^*, W)\pi(T^*, \theta, W, \mu)}{p(D_E|T, \theta, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta, W, \mu)} \frac{q(T|T^*)}{q(T^*|T)}\right\} \\
 &= \min\left\{1, \frac{s(o_x - t_x^*) \frac{I_x^*(v(x)|\theta, D_E)}{\sum_{y \in H} I_x^*(y|\theta, D_E)} f^*(x) f^*(v(x)) \pi(T^*)}{s(o_x - t_x) \frac{I_x(v(x)|\theta, D_E)}{\sum_{y \in H} I_x(y|\theta, D_E)} f(x) f(v(x)) \pi(T)} \frac{s(o_x - t_x)}{s(o_x - t_x^*)}\right\} \\
 &= \min\left\{1, \frac{\frac{I_x^*(v(x)|\theta, D_E)}{\sum_{y \in H} I_x^*(y|\theta, D_E)} f^*(x) f^*(v(x))}{\frac{I_x(v(x)|\theta, D_E)}{\sum_{y \in H} I_x(y|\theta, D_E)} f(x) f(v(x))}\right\}
 \end{aligned}$$

3.5 Update epidemiological parameters

We create a proposal θ^* by choosing i from $1, \dots, |\theta|$ and adding $Y \sim \text{Normal}(0, \sigma_i)$ to θ_i . The value of σ_i is calibrated in the burn-in period. We accept with probability

$$\begin{aligned}
 p(\text{accept}) &= \min\left\{1, \frac{p(D_E|T, \theta^*, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta^*, W, \mu)}{p(D_E|T, \theta, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta, W, \mu)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right\} \\
 &= \min\left\{1, \frac{s(o_x - t_x) \frac{I_{t_x}(v(x)|\theta^*, D_E)}{\sum_{y \in H} I_{t_x}(y|\theta^*, D_E)} \pi(\theta^*)}{s(o_x - t_x) \frac{I_{t_x}(v(x)|\theta, D_E)}{\sum_{y \in H} I_{t_x}(y|\theta, D_E)} \pi(\theta)}\right\}
 \end{aligned}$$

3.6 Update mutational parameters

We create a proposal μ^* by choosing i from $1, \dots, |\mu|$ and adding $Y \sim \text{Normal}(0, \delta_i)$ to μ_i . The value of δ_i is calibrated in the burn-in period. We accept with probability

$$\begin{aligned}
 p(\text{accept}) &= \min\left\{1, \frac{p(D_E|T, \theta, W)p(D_G|P, \mu^*)p(P|T, W)\pi(T, \theta, W, \mu^*)}{p(D_E|T, \theta, W)p(D_G|P, \mu)p(P|T, W)\pi(T, \theta, W, \mu)} \frac{q(\theta|\mu^*)}{q(\mu^*|\mu)}\right\} \\
 &= \min\left\{1, \frac{p(D_G|P, \mu^*)\pi(\mu^*)}{p(D_G|P, \mu)\pi(\mu)}\right\}
 \end{aligned}$$

3.7 Convergence

For all analyses, we ran an initial burn-in period of 10^5 iterations. Convergence of parameters was checked by eye. The variance of the proposal distributions of the parameters was adjusted during this time, such that the acceptance probability was around 0.5. The chain was then run for 3×10^5 iterations, and sampled every 200th iteration. This yields a chain of 1500 samples from the posterior distribution. Increasing these values gave no observable difference in results.

3.8 Estimation

Point estimates of parameters were obtained as the median of the posterior distribution sampled from the Markov chain. 95% credibility interval boundaries were computed as the 2.5th and 97.5th percentile.

References

- [1] A. Bataille, F. van der Meer, A. Stegeman, and G. Koch. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathogens*, 7(6), 2011.
- [2] S. Cauchemez, F. Carrat, C. Viboud, A. Valleron, and P. Boëlle. A bayesian mcmc approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.
- [3] E. Cottam, D. Haydon, D. Paton, J. Gloster, J. Wilesmith, N. Ferris, G. Hutchings, and D. King. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the united kingdom in 2001. *Journal of Virology*, 80(22):11274–11282, 2006.
- [4] E. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. Paton, D. King, and D. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, 2008.
- [5] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.
- [6] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Sunderland, MA, 2004.
- [7] J. Liu, S. Lim, Y. Ruan, A. Ling, L. Ng, C. Drosten, E. Liu, L. Stanton, and M. Hibberd. Sars transmission pattern in singapore reassessed by viral sequence variation analysis. *PLoS Medicine*, 2:0162–0168, 2005.
- [8] M. Morelli, G. Thébaud, J. Chadœuf, D. King, D. Haydon, and S. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computational Biology*, 8(11), 2012.