PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data

Finlay Campbell[1]*, **Anne Cori**[1], **Neil Ferguson**[1], **Thibaut Jombart**[1,2,3]*

1 MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom, **2** Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **3** UK Public Health Rapid Support Team, London, United Kingdom

* f.campbell15@imperial.ac.uk (FC); thibautjombart@gmail.com (TJ)

Citation: Campbell F, Cori A, Ferguson N, Jombart T (2019) Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol 15(3): e1006930. https://doi.org/10.1371/journal.pcbi.1006930

Editor: Virginia E. Pitzer, Yale School of Public Health, UNITED STATES

Received: February 7, 2018

Accepted: March 4, 2019

Published: March 29, 2019

Data Availability Statement: All simulations and code for re-creating these simulations are available from github.com/finlaycampbell/PlosComp_ContactModel.

Funding: FC is funded by the Wellcome Trust (https://wellcome.ac.uk). AC is funded by the Medical Research Council Centre for Global Infectious Disease Analysis (https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis). NF is funded by the UK Medical Research Council (https://www.mrc.ac.uk), UK National Institute for Health Research under the Health Protection

Abstract

There exists significant interest in developing statistical and computational tools for inferring 'who infected whom' in an infectious disease outbreak from densely sampled case data, with most recent studies focusing on the analysis of whole genome sequence data. However, genomic data can be poorly informative of transmission events if mutations accumulate too slowly to resolve individual transmission pairs or if there exist multiple pathogens lineages within-host, and there has been little focus on incorporating other types of outbreak data. We present here a methodology that uses contact data for the inference of transmission trees in a statistically rigorous manner, alongside genomic data and temporal data. Contact data is frequently collected in outbreaks of pathogens spread by close contact, including Ebola virus (EBOV), severe acute respiratory syndrome coronavirus (SARS-CoV) and *Mycobacterium tuberculosis* (TB), and routinely used to reconstruct transmission chains. As an improvement over previous, ad-hoc approaches, we developed a probabilistic model that relates a set of contact data to an underlying transmission tree and integrated this in the *outbreaker2* inference framework. By analyzing simulated outbreaks under various contact tracing scenarios, we demonstrate that contact data significantly improves our ability to reconstruct transmission trees, even under realistic limitations on the coverage of the contact tracing effort and the amount of non-infectious mixing between cases. Indeed, contact data is equally or more informative than fully sampled whole genome sequence data in certain scenarios. We then use our method to analyze the early stages of the 2003 SARS outbreak in Singapore and describe the range of transmission scenarios consistent with contact data and genetic sequence in a probabilistic manner for the first time. This simple yet flexible model can easily be incorporated into existing tools for outbreak reconstruction and should permit a better integration of genomic and epidemiological data for inferring transmission chains.

Research Unit initiative (https://www.nihr.ac.uk), National Institute of General Medical Sciences under the Models of Infectious Disease Agent Study initiative (https://www.nigms.nih.gov/Research/specificareas/MIDAS/Pages/default.aspx), and the Bill and Melinda Gates Foundation (http://www.gatesfoundation.org). TJ is funded by the Global Challenges Research Fund (GCRF) project 'RECAP – research capacity building and knowledge generation to support preparedness and response to humanitarian crises and epidemics' managed through RCUK and ESRC (ES/P010873/1), the UK Public Health Rapid Support Team and the National Institute for Health Research - Health Protection Research Unit for Modelling Methodology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Reconstructing the history of transmission events in an infectious disease outbreak provides valuable information for informing infection control policy. Recent years have seen considerable progress in the development of statistical tools for the inference of such transmission trees from outbreak data, with a major focus on whole genome sequence data (WGS). However, complex evolutionary behavior, missing sequences and the limited diversity accumulating along transmission chains limit the power of existing approaches in reconstructing outbreaks. We have developed a methodology that uses information on the contact structures between cases to infer likely transmission links, alongside genomic and temporal data. Such contact data is frequently collected in outbreak settings, for example during Ebola, HIV or Tuberculosis outbreaks, and can be highly informative of the infectious relationships between cases. Using simulations, we show that our contact model effectively incorporates this information and improves the accuracy of outbreak reconstruction even when only a portion of contacts are reported. We then apply our method to the 2003 SARS outbreak in Singapore and describe the range of transmission scenarios consistent with genetic data and contact data for the first time. Our work suggests that, whenever available, contact data should be explicitly incorporated in outbreak reconstruction tools.

Introduction

Inferring chains of transmission in an infectious disease outbreak can provide valuable epidemiological insights into transmission dynamics, which can be used to guide infection control policy. For example, reconstructed outbreaks have been used to identify drivers of ongoing infection [1], characterize heterogeneous infectiousness in a population [2], evaluate the effectiveness of interventions [3] and determine transmission mechanisms [4]. Consequently there has been increased interest in developing statistical and computational tools for inferring such 'transmission trees' from various types of data, including times of symptom onset, contact tracing data, spatial data and, increasingly frequently, pathogen whole genome sequence (WGS) data [5–13].

Most state of the art outbreak reconstruction tools aim at approximating a posterior distribution of likely transmission trees in a Bayesian MCMC framework. Two major approaches have emerged, which can be defined by their treatment of genetic data [14]. The '*pairwise approach*' begins with a model of disease transmission and attaches to this a genetic model that describes the pairwise genetic distance between putative transmission pairs [6–9]. The '*phylogenetic approach*' uses genetic data to infer the unobserved history of coalescent events between sampled pathogen genomes in the form of a phylogenetic tree and infers transmission trees consistent with this phylogeny using epidemiological data. Such methods either use a fixed phylogeny inferred *a priori* [10,15] or jointly infer the phylogeny alongside the transmission tree itself [11–13].

These methodologies differ in their ability to identify unobserved or imported cases, accurately describe evolutionary behavior in the presence of multiple dominant strains within-host or incomplete transmission bottlenecks and accommodate multiple genetic sequences per host. However, a notable similarity between these studies is the fact that they generally only consider temporal and genetic data. Accordingly, such approaches rely heavily on highly informative genetic sequence data for identifying likely transmission pairs, as temporal data is generally consistent with a large number of potential ancestries [16].

However, WGS are not always informative of the transmission route of an epidemic. Firstly, genetic diversity across most outbreaks is low and a significant portion of genetic sequences expected to be identical [17], most prominently if the pathogen genome is small (e.g. human influenza [18]), the mutation rate low (e.g. *Mycobacterium tuberculosis* [19]), or the generation time (delay between primary and secondary infection) short (e.g. *Streptococcus pneumoniae* [20]). In these cases, transmission pairs cannot be accurately identified by genetic data alone, resulting in an overall poorly resolved transmission tree. The informativeness of genetic sequence data is also limited by complex evolutionary behavior. Didelot *et al.* demonstrated that realistic genetic models accounting for within-host diversity, in which several strains coexist inside a host and can be transmitted and sampled, place significant uncertainty around ancestry allocation even when genetic diversity across the outbreak is high, as multiple transmission scenarios are consistent with the genetic data [15]. Pathogens displaying significant within-host diversity include those with long periods of carriage (e.g. *Staphylococcus aureus* [21]) or a propensity for super-infections (*Streptococcus pneumoniae* [22]). WGS is also uninformative of the direction of transmission between donor-recipient pairs if multiple sequences per host are not available [23]. Finally, WGS will generally not be available for all infected individuals, especially in resource poor settings. In the 2014 Ebola outbreak in West Africa, for example, sequences were collected in only 5% of cases [24]. Genetic data is therefore frequently of limited use in reconstructing transmission trees, and inference methods that rely heavily on it will perform poorly in such circumstances.

Integrating other types of outbreak data is therefore necessary for inferring transmission trees in realistic outbreak situations. A frequently collected and highly informative source of data on likely transmission routes is contact data, an integral component of early outbreak response that describes the network of reported contacts with infected individuals. Contact data provided most of the information used to reconstruct transmission chains during Severe Acute Respiratory Syndrome (SARS) [25], Middle East Respiratory Syndrome (MERS) [26] and Ebola [1,27,28] epidemics, and is routinely collected in outbreaks of HIV [29] and Tuberculosis [30]. Contact data can be classified as '*exposure*' data and or '*contact tracing*' data. *Exposure* data describes contacts between a given case and their potential infectors and is an intrinsic part of case definition in diseases with person-to-person transmission. *Contact tracing* data describes contacts between confirmed/probable cases and individuals they could have infected: it is used for active case discovery and rapid isolation and is an integral part of containment strategy. Importantly, both types of contact data potentially contain information on the topology of the transmission tree.

Here, we introduce a model which exploits contact data alongside dates of symptom onset, information on the incubation period (delay between infection and symptom onset) and generation time, and pathogen WGS to reconstruct transmission chains. Our methodology extends the *outbreaker* model introduced by Jombart *et al.* [6] with a contact model that accounts for partial sampling and the presence of non-infectious contacts between cases. As an improvement over other approaches, the integration of a full contact model reduces the reliance on high quality genetic data for accurate inference. We evaluate the performance of this new model and compare the value of the different types of data for inferring who infects whom, using a variety of simulated outbreak scenarios. We then apply our approach to the early stages of the 2003 SARS outbreak in Singapore, integrating the available data on contact structures and genome sequences in a single statistical framework for the first time. The inference tool presented in this study is freely available as the package *outbreaker2* for the R software [31].

Results

Algorithm performance on simulated outbreaks

We tested our new model on simulated outbreaks of two pathogens with well-defined epidemiological and evolutionary parameters, namely EBOV and SARS-CoV [27,32]. As SARS-CoV WGS generally contain greater genetic diversity between transmission pairs and are therefore more informative of transmission events than Ebola WGS [17], we describe contrasting outbreak settings where the added value of incorporating contact data may vary. Outbreaks were simulated using empirical estimates of the generation time distribution, the incubation period distribution and the basic reproduction number R_0 (i.e. the average number of secondary infections caused by an index case in a fully susceptible population [33]). To reflect observed heterogeneities in infectiousness, outbreaks were simulated under strong super-spreading tendencies, where a small number of individuals account for a high number of cases [2,25,34]. Genetic sequence data was simulated using estimates of the genome length and genome wide mutation rate.

To describe contact tracing efforts in various outbreak scenarios, contact data was simulated using two parameters (for a full description of the model, see Methods). Briefly, the probability of a contact being reported is described by ε, the contact reporting coverage. Non-infectious mixing between cases that obscures the topology of the underlying transmission network is described using the non-infectious contact probability λ, defined as the probability of contact occurring between two sampled cases that do not constitute a transmission pair. A useful corollary term to λ is the expected number of non-infectious contacts per person, ψ, as this accounts for the size of the outbreak and describes the amount of non-infectious mixing in terms of numbers of contacts.

We investigated the effect of the coverage of contact tracing efforts and the probability of non-infectious contact on our ability to reconstruct transmission trees using using a grid of values for ε and ψ. The informativeness of different types of outbreak data was determined by reconstructing each outbreak four times, using combinations of times of sampling (T), contact tracing data (C) and genetic sequence data (G): T, TC, TG and TCG. For an example of a simulated transmission network, contact network and reconstructed transmission tree, see S1 Fig.

Transmission tree reconstruction was essentially impossible using only times of sampling, with on average only 9% and 10% of infectors correctly identified in the consensus transmission tree for EBOV and SARs-CoV outbreaks, respectively (Fig 1). Statistical confidence in ancestry allocation as defined by the average Shannon entropy of the posterior distribution of potential infectors for each case, for which a value of 0 indicates complete posterior support for a given ancestry and higher values indicates lower statistical confidence, was also low (S2 Fig). Including genetic data improved both the accuracy of inference and the statistical confidence in these assignments. However, even in the idealized scenario of error free sequencing and WGS for all cases, this data was insufficient for complete outbreak reconstruction under our genetic likelihood, with on average only 29% and 70% of transmission pairs correctly inferred in in EBOV and SARS-CoV outbreaks, respectively.

Incorporating contact tracing data using our new contact model improved the accuracy of transmission tree reconstruction across all simulations, with the magnitude of improvement dependent on the values of ε and ψ (Fig 1). Unsurprisingly, accuracy of inferred ancestries increased with coverage ε, as a greater number infectious contacts were reported, and decreased with the number of non-infectious contacts ψ, as these reduced the proportion of contacts informative of transmission events. In the idealized scenario of complete contact tracing coverage and zero non-infectious contacts, outbreaks were reconstructed with near perfect accuracy, even in the absence of genetic data, with the few incorrectly assigned ancestries

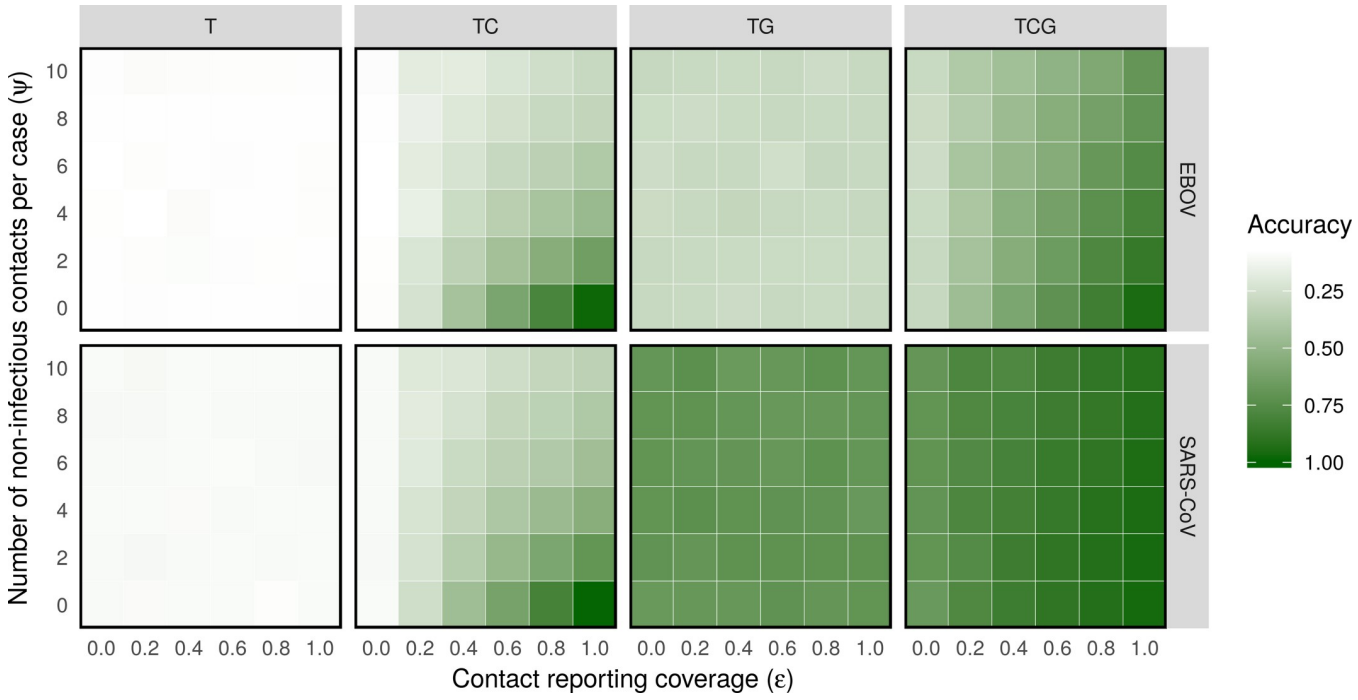

Fig 1. Accuracy of outbreak reconstruction using different types of outbreak data. 100 outbreaks were simulated and reconstructed at each grid point, using different values for the contact reporting coverage ε and number of non-infectious contacts per case ψ. Each outbreak was reconstructed four times, using different combinations of times of sampling (T), contact tracing data (C) and genetic data (G). The color of a grid point represents the average accuracy of outbreak reconstruction.

https://doi.org/10.1371/journal.pcbi.1006930.g001

attributable to misinformative sampling times. Encouragingly, improvements in accuracy persisted in more realistic contact tracing scenarios with partial coverage and large numbers of non-infectious contacts. For example, consider the contact tracing scenario with only 60% coverage and on average two non-infectious contacts per person. When adding this data to the purely temporal *outbreaker* model, the accuracy in reconstructing EBOV outbreaks increased from 9% to 44%. Though more than half of ancestries remained incorrectly assigned, outbreaks were in fact reconstructed with greater accuracy than when using WGS from Ebola cases, for which accuracy was only 28%.

When comparing the informativeness of contact data and genetic data across all simulations, we found that information on contact structures was frequently equally or more informative than fully sampled and error-free genetic sequence data, even under limitations of partial coverage and significant levels of non-infectious contact (Fig 2). For example, contact data with only 40% coverage and 4 non-infectious contacts per person was as informative as fully sampled Ebola genetic data. Similarly, if the reporting coverage was 100%, contact data was as informative as Ebola WGS even when individuals reported 10 non-infectious contacts with other cases on average, meaning that only 17% of reported contacts represented true transmission pairs. Though contact data was generally less informative than SARS-CoV WGS in most scenarios, it still provided comparable increases in accuracy when coverage was high ($\varepsilon > 0.6$) and contact of non-infectious contact low ($\psi < 2$).

As expected, accuracy of outbreak reconstruction was highest when using contact, temporal and genetic data at the same time. Notably, contact data was able to correct a significant portion of ancestries falsely assigned using only temporal data and WGS. For example, incorporating contact data with 80% coverage and 2 non-infectious contacts per person lead to an increase in average accuracy of outbreak reconstruction from 28% to 79% for EBOV outbreaks

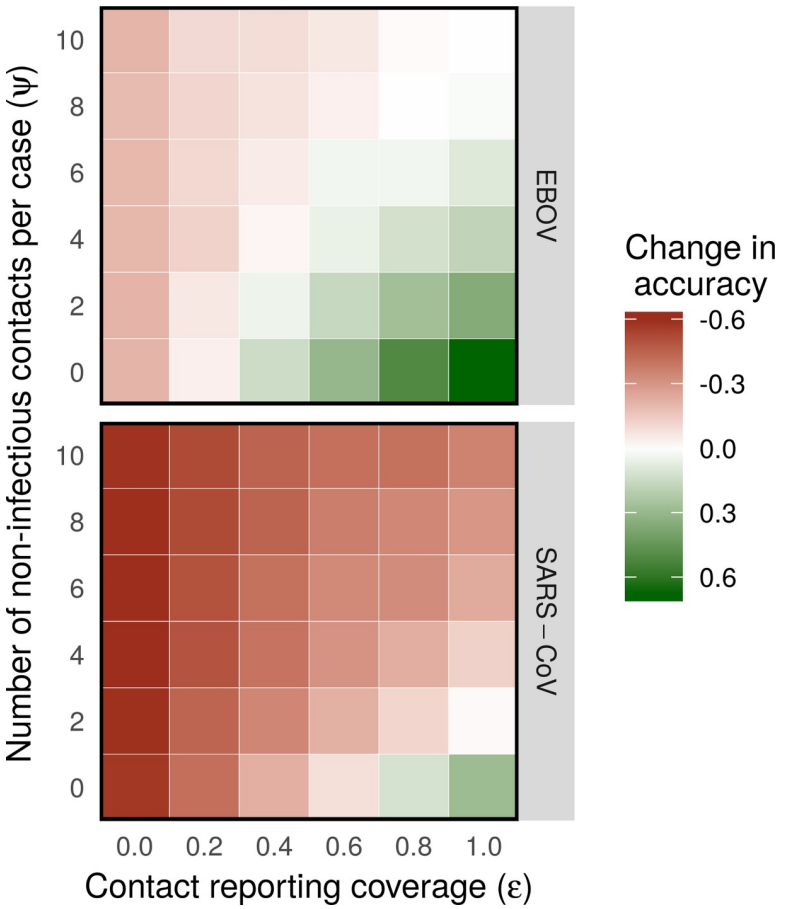

Fig 2. Informativeness of contact data relative to fully sampled genetic data. 100 outbreaks were simulated and reconstructed at each grid point, using different values for the contact reporting coverage ε and number of non-infectious contacts ψ. The color of a grid point represents the difference between accuracy of outbreak reconstruction using times of sampling and contact tracing data and using times of sampling and genetic data.

https://doi.org/10.1371/journal.pcbi.1006930.g002

(Fig 1). Contact data therefore contained significant additional information on likely transmission routes not available from pathogen WGS, which was successfully integrated in our inference framework.

In addition to the transmission tree itself, we inferred the model parameters ε and λ under uninformative priors and observed accurate estimates of the simulated values for both EBOV and SARS-CoV outbreaks (S3 and S4 Figs). When using temporal and contact data, the mean posterior estimates of ε and λ across 100 outbreaks were generally distributed around the true simulated value, and with low variance especially when the coverage ε was high. Only when λ was high were the estimates slightly off-centered from the true value. Including genetic data improved parameter inference across all scenarios, resulting in correctly centered estimates with a reduced variance. ε and λ are therefore identifiable in our contact likelihood and generally well estimated by our inference framework, allowing appropriate probabilistic weighting of contact data in the allocation of ancestries.

2003 SARS outbreak in Singapore

We applied our method to the early stages of the 2003 SARS outbreak in Singapore, for which dates of symptom onset, whole genome sequences and contact information were collected for

the first 13 cases [35,36]. Previous attempts to infer the transmission tree from these data either reconstructed probable lineages by manual inspection [35,36], or entirely discarded information on the six reported contacts between cases [6,37], even though they were all thought to be epidemiologically significant [36].

Using *outbreaker2*, we were able to infer the range of transmission histories consistent with the temporal, genomic and contact data in a probabilistic manner. We analyzed the outbreak several times using different settings; with and without contact data and using different priors on λ (Fig 3). Under the assumption that the reported contacts were very likely to be epidemiologically relevant, by fixing the non-transmission contact rate λ at 1e-4, contact data significantly changed the posterior distribution of ancestries (Fig 3B and 3C). As expected under these assumptions, transmission links in line with reported contacts were better supported. For example, the most likely infector of cases *sin2677* and *sin2774* was *sin2500* when including contact data (Fig 3C), instead of *sin2748* in the default analysis (Fig 3B). Even though these transmission events were less likely under the genetic likelihood, as they implied the accumulation of 2 and 3 mutations, respectively, rather than 1 and 2 mutations, these ancestries were supported by the contact data and were therefore credible under our model. Importantly, the original transmission pathway inferred in the absence of genetic data (*sin2748* infecting *sin2677* and *sin2774*) also remained plausible. Further novel infection routes supported by contact data were *sin849* infecting *sin848*, and *sin848* infecting *sin852*.

However, not all ancestries supported by contact data received significant posterior support. Even though *sin849* was in contact with and therefore a likely infector of *sin848*, *sin847* remained the consensus ancestor of *sin848* with 78% posterior support, as it is separated from *sin848* by only 1 mutation, which is far more favorable under the genetic likelihood compared to the 7 mutations separating *sin849* and *sin848*. Furthermore, though *sin850* and *sin848* had a reported contact, an infectious relationship between the two received no posterior support due to the large number of mutations (10) separating the two. Therefore, while the contact model generally provided support for transmission histories in line with epidemiological observations of contacts, each ancestry allocation was the result of weighing the evidence provided by all three, potentially conflicting, data sources.

Interestingly, incorporating contact data in our analysis affected ancestry allocations not directly referenced in the contact network. For example, *sin848* was suggested as a novel infector of *sin847* with 22% posterior support, though these cases are not linked by a reported contact. This is explained by a change in the inferred infection times (S5 Fig). *sin848* infecting *sin852*, as suggested by the contact data, resulted in an earlier inferred infection time for *sin848*, which in turn made it a plausible infector of *sin847*. A similar change in the inferred infection times of *sin2500* and *sin2748*, driven by the contact data, reversed the directionality of their consensus infectious relationship, even though this directionality was not provided in the contact data. Incorporating the contact model alongside the genetic and temporal model therefore allowed for high level interactions, beyond simply providing support for ancestries indicated in the contact data.

We also analyzed the dataset using a weaker prior on λ (Beta(1, 10)) and an uninformative prior. However, the resulting posterior ancestries were essentially identical to those inferred in the absence of contact data (S6 Fig).

We then reconstructed the outbreak under the assumption that all reported contacts necessarily occurred between direct transmission pairs by fixing λ at a value of 0 (Fig 3D). The posterior distribution of transmission networks therefore spanned the contact network, with 6 of the 12 ancestries remaining fixed. This rigid topology of plausible transmission networks resulted in low variance among the remaining ancestries, producing essentially a single

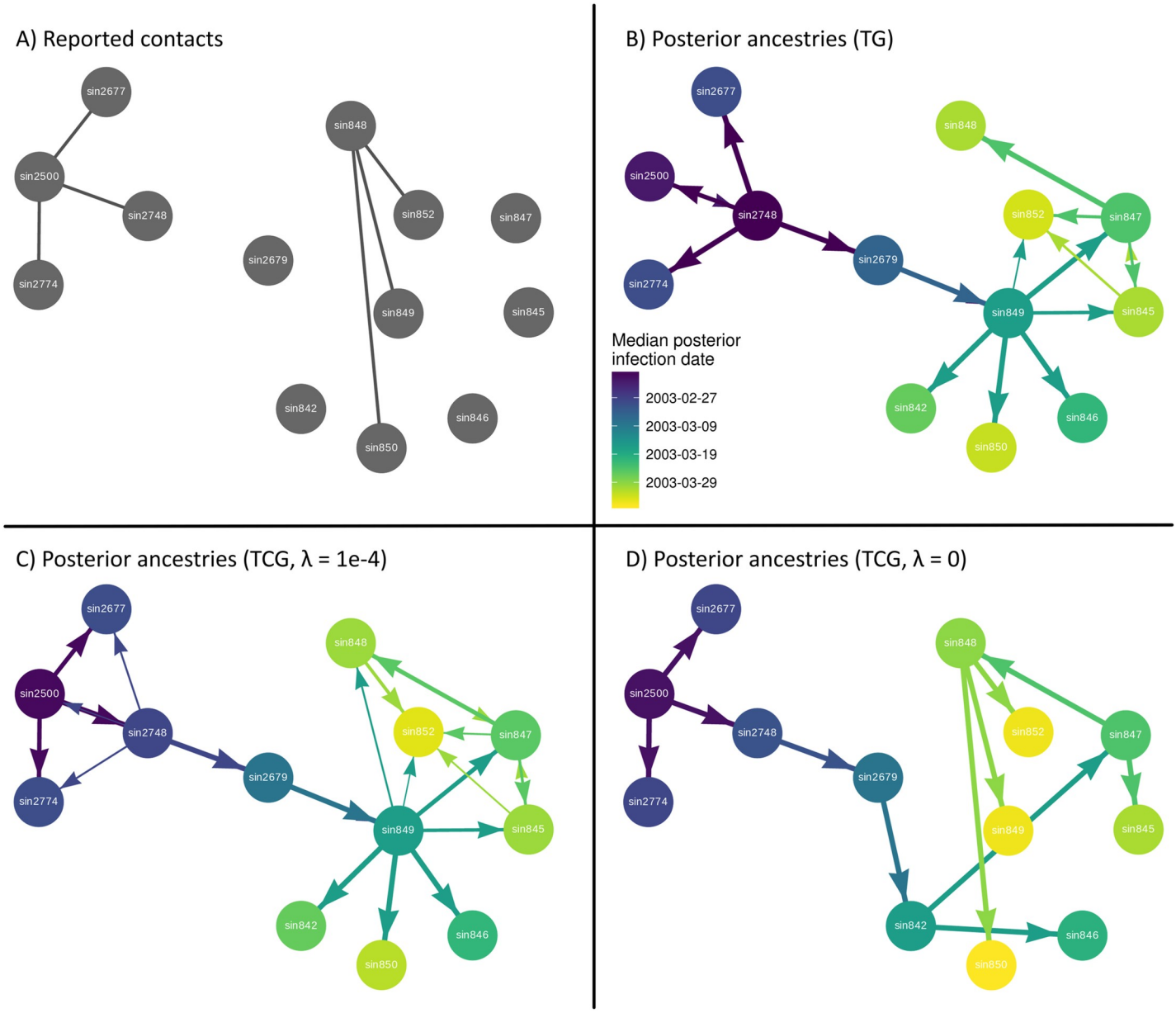

Fig 3. Reconstruction of the 2003 SARS outbreak in Singapore. A) Circles represent individual cases, and edges the epidemiological contacts reported between them. **B)** The outbreak was reconstructed using temporal and genetic data. Arrows represent posterior ancestries between cases, scaled in width by the posterior frequency of that ancestry. Ancestries with a minimum posterior frequency of 0.01 were included. The color of a node corresponds to the median posterior infection time of that case. **C)** The outbreak was reconstructed using temporal, contact and genetic data, and the non-infectious contact probability λ fixed at a value of 1e-4. **D)** The outbreak was reconstructed using temporal, contact and genetic data, and the non-infectious contact probability λ fixed at a value of 0.

https://doi.org/10.1371/journal.pcbi.1006930.g003

posterior tree. Notably, this analysis proposed several new ancestries (*sin2679* to *sin842*, *sin842* to *sin847* and *sin848 to sin850*) rejected with a λ value of 1e-4 and had a substantially lower average log-likelihood (-647.4 compared to -579.2). Therefore, while the assumption that λ was 0 may have been valid, this approach forced the algorithm to accept ancestries highly unlikely under the genetic and temporal likelihoods, thereby preventing a meaningful integration of different data sources.

Discussion

The methodology described here represents, to our knowledge, the first outbreak reconstruction framework integrating contact data alongside the timing of symptom onset, reporting rates and pathogen WGS data. Using simulations, we have shown how contact data can improve epidemiological inference across a range of outbreak settings, including incomplete contact tracing coverage, significant amounts of non-infectious contact and strong super-spreading tendencies. By integrating contact data in the analysis of early stages of the 2003 Singaporean SARS outbreak for the first time, we have illustrated how our approach can work in a realistic outbreak scenario and provide a probabilistic description of plausible transmission routes in the face of conflicting outbreak data. The general applicability of our model, in addition to being implemented in a freely available and well-documented software package, makes *outbreaker2* useful to a broad epidemiological audience.

Our work reduces the reliance of outbreak reconstruction tools on WGS data. This is significant when considering that genetic diversity in many pathogens arises too slowly to resolve a significant portion of transmission pairs by genetic means [17], and that within-host genetic diversity of other pathogens hinders accurate transmission tree reconstruction from genetic data [38]. Furthermore, sequencing pathogen genomes from enough cases in an outbreak to resolve individual transmission events is frequently unrealistic in the face of logistical and financial limitations [24]. In contrast, contact tracing is routinely conducted during outbreak response, and therefore provides a valuable additional window of information on transmission events without placing an additional burden on field epidemiologists. Indeed, given the simulation model and likelihoods used for inference, our work suggests that even incomplete contact tracing data may be more informative than fully sampled, error-free genetic sequence data of some pathogens.

Methodologically, our contact model differs from previous methods for relating contact data to epidemiological processes, with several advantages [39–41]. Soetens *et al*. estimate effective reproduction numbers by assigning transmission links on the basis of contact data, while accounting for right censoring of case counts [39]. However, they assume complete sampling of contacts and cases, and automatically designate confirmed cases with a known contact as transmission pairs. This is equivalent to fixing λ at a value of 0 in our model, which our analysis of the SARS dataset has shown is unsuitable for integrating other types of data in a meaningful manner. Similarly, Hens *et*. *al*. restrict transmission pairs to those supported by reported contacts [40], thereby mis-assigning ancestries if contacts are only partially reported.

Jewell and Roberts establish a more statistically rigorous approach for epidemiological inference from contact data by explicitly modelling the contact process that drives the infectious process in an SINR compartmental framework [41]. Such a mechanistic model natively relates epidemiological processes to a set of observed contact data and has the advantage of potentially accommodating complex contact structures caused by non-random mixing in the future. However, a prospective model of this sort is considerably more complex to develop in a statistically tractable manner and has necessitated the assumption of a single index case, whereas multiple infectious introductions are easily accounted for in our contact likelihood. Furthermore, their approach does not explicitly model under-reporting of contacts, and therefore does not allow valuable prior information on the coverage of the contact tracing effort to inform the analysis. Our approach is therefore applicable to a wider range of realistic outbreak settings.

Incorporating this contact model alongside a temporal and genetic model represents an improvement over previous, ad-hoc methods to data integration, which generally use contact data to exclude transmission links and then explore the remaining transmission tree space

using other data [42,43]. By modelling contact tracing as a probabilistic process in a Bayesian framework, information on the contact tracing effort can also be embedded in the prior to improve the inferential process and more explicitly describe the assumptions underlying it. For example, if most contacts in an outbreak are expected to have been reported, the prior on the contact reporting coverage ε can be shifted to provide greater support for higher values, reducing support for ancestries that lack a contact. ε could even be fixed at a value of 1, meaning a reported contact is *required* for a given transmission pair to be inferred, given the assumption that every contact has been reported. Similarly, as shown for the 2003 SARS outbreak, an informative prior on the non-infectious contact probability λ should generally be used. As most contact tracing efforts are conducted under the belief that non-transmission pairs experience contacts with significantly lower probability than transmission pairs, the prior on λ should provide support for lower values, in turn placing greater weight on reported contacts when assigning ancestries.

Our method also allows conflicting data to be treated in a systematic manner, as demonstrated by the analysis of the the 2003 SARS outbreak, where several ancestries were supported by contact data yet separated by an implausibly large number of mutations. In contrast to existing tools [6,35], *outbreaker2* can evaluate these inconsistencies and determine the distribution of likely transmission trees under multiple data types. While not necessarily improving the accuracy of the inferred transmission tree, our approach better captures the uncertainty around these ancestry assignments given the available data.

However, it is important to note both the intrinsic informational limitations of contact data as well as the methodological limitations of the work presented here. Contact tracing constitutes a significant logistical challenge, as most if not at all infected individuals must be followed up, and suspected cases monitored past the upper end of the incubation period distribution [44–46]. The coverage of contact tracing efforts conducted in low resource settings may therefore be low [47], and consequently poorly informative of the transmission network (Fig 1). Even if a significant proportion of contacts are reported, a high degree of mixing between cases can obscure the topology of the underlying transmission network, for example within hospital wards or classrooms. Contact data alone will therefore not always suffice for complete reconstruction of an outbreak. Nevertheless, the framework presented here allows even minimally informative contact data to be incorporated into transmission tree inference alongside other available data.

Furthermore, the use of strong priors on ε and λ may be required to ensure adequate weighting of contact data, especially in the face of conflicting genetic data as shown in the analysis of the 2003 SARS outbreak. While our framework forces an explicit description of these assumptions, the sensitivity of the algorithm outputs to the prior distributions should be noted and explored adequately.

Our model of epidemiological contacts also makes a number of simplifications, some of which could be improved upon in future work. As the contacts are undated, the model does not consider that they are only indicative of transmission events if they occur during the infectious period of the infector, potentially resulting in overconfident ancestry assignments if contacts frequently occur outside this time period. However, as epidemiologists generally only record meaningful contacts occurring within likely windows of infection, the assumption that recorded contacts represent epidemiologically plausible transmission pairs appears reasonable. As currently implemented, our model also does not account for different weights between contacts, which could be useful for example to stratify different types of sexual intercourse by their risk of HIV transmission [48], or TB contacts by their duration of contact (e.g. household vs. casual). However, it could be easily extended to do so by using separate parameters for the reporting coverage (e.g. $\varepsilon_1, \varepsilon_2, \varepsilon_3$) and non-infectious contact probability (e.g. $\lambda_1, \lambda_2, \lambda_3$) of

each type of contact. Furthermore, the contact model is undirected and treats exposure data and contact tracing data equally, resulting in a loss of information about the potential directionality of the infectious interaction which must instead be inferred from other data. Directionality could be incorporated with relative ease by treating reported contacts as asymmetric (individual *i* contacting individual *j* is distinct from *j* contacting *i*) and relating this to the infector-infectee relationship in the putative transmission tree (*I* infecting *j* is distinct from *j* infecting *I*). However, the current model generally inferred directionality successfully from temporal data simulated under realistic delay distributions (Fig 1).

It should also be noted that the use of fixed generation time and incubation period distributions is poorly suited to epidemic scenarios with highly connected contact networks, for which hazard-based approaches are more suitable [49,50]. However, as demonstrated in Fig 1, contact data is only informative when the contact network itself is fairly sparse (i.e. λ is low). The assumption of fixed generation time and incubation period distributions is therefore suitable for the use cases of our contact model [10,13,51].

Finally, the assumptions underlying the pairwise genetic model should be considered when using *outbreaker2*. The likelihoods of pairwise genetic distances are treated as independent, when in fact they are dependent on the underlying infectious relationships between cases (e.g. the genetic relatedness of case A and its infector B is dependent on the infector of B). Similarly, by considering only genetic distances, our method disregards histories of shared mutations between genomes. These assumptions can result in loss of information and potential misinterpretation of genetic signals, especially when evolutionary histories are complex [38]. In such cases, character-based, phylogenetic models should be considered [10,11].

In conclusion, the work presented here provides a simple yet flexible methodology for integrating contact data with genetic and temporal data in the inference of transmission trees. By allowing contact data to complement and/or substitute genetic data as the primary source of information on infectious relationships between individuals, our work increases both the scope and accuracy of methodologies for outbreak reconstruction.

Methods

Outbreaker model

Our work is an extension of the *outbreaker* model developed by Jombart et al. [16], re-written in a manner to be more extensible This model considers, for each case I ($i = 1, \ldots, N$), the probability of a proposed transmission history given the time of symptom onset t_i and a pathogen genetic sequence s_i (Table 1). Assumptions on the temporal relationship between transmission pairs are given by the generation time distribution w, defined as the distribution of delays between infection of a primary and secondary case, and the incubation period distribution f, defined as the distribution of intervals between infection and symptom onset of a case. w and f are assumed to be known, and not estimated during the inference process.

The unobserved transmission events are modelled using augmented data; case i is infected at time T_i^{inf}, and its most recent sampled ancestor denoted α_i. To allow for unobserved cases, the number of generations separating i and α_i is explicitly modelled and denoted κ_i ($\kappa_i \geq 1$). The proportion of cases that have been sampled is defined by the parameter π and is inferred as part of the estimation procedure. The other estimated parameter is the mutation rate μ, measured per site per generation of infection.

This model is embedded in a Bayesian framework. Denoting D the observed data, A the augmented data and θ the model parameters, the joint posterior distribution of parameters

Table 1. Notation of outbreaker model [6].

Symbol	Type	Description
i	Data	Index of cases
N	Data	Number of cases in the sample
s_i	Data	Sequence of case i
t_i	Data	Collection date of s_i
$c_{i,j}$	Data	Contact status between case i and case j
w	Function	Generation time distribution
f	Function	Incubation period distribution
$d(s_i,s_j)$	Function	Number of mutations between s_i and s_j
$l(s_i,s_j)$	Function	Number of comparable nucleotide positions between s_i and s_j
α_i	Augmented data	Index of the most recent sampled ancestor of case i
κ_i	Augmented data	Number of generations between α_i and i
T_i^{inf}	Augmented data	Date of infection of i
μ	Parameter	Mutation rate, per site and per generation of infection
π	Parameter	Proportion of cases sampled in the outbreak
ε	Parameter	Proportion of contacts reported
λ	Parameter	Probability of non-infectious contact between cases
η	Parameter	Probability of contact between transmission pairs
ζ	Parameter	Probability of false-positive reporting a contact

https://doi.org/10.1371/journal.pcbi.1006930.t001

and augmented data is defined as:

$$P(A,\theta|D) = \frac{P(D,A|\theta)P(\theta)}{P(D)}$$

The first term describes the likelihood of the data, the second term the joint prior (for a complete description of both, see Jombart *et al.* [6]). Briefly, the likelihood is computed as a product of case-specific terms, and can be decomposed into a genetic likelihood Ω^1, a temporal likelihood Ω^2 and a reporting likelihood Ω^3.

The genetic likelihood describes, for a given case i, the probability of observing the genetic distance between sequence s_i and that of its most recent sampled ancestor $s_{\alpha i}$, given the proposed ancestries and parameters:

$$\Omega_i^1 = p(s_i|\alpha_i, s_{\alpha_i}, \kappa_i, \mu)$$

and is defined as:

$$(\kappa_i\mu)^{d(s_i,s_{\alpha_i})}(1-\kappa_i\mu)^{l(s_i,s_{\alpha_i})-d(s_i,s_{\alpha_i})}$$

This calculates the probability of $d(s_i,s_j)$ mutation events occurring at the observed nucleotide positions and no mutations occurring at the remaining positions, while summing over the κ_i generations in which the mutations could have occurred. For a full derivation of this likelihood, see S1 Text. The temporal likelihood describes the probability of observing the time of symptom onset and proposed time of infection:

$$\Omega_i^2 = p(t_i|T_i^{inf})p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i)$$

and is calculated as:

$$f(t_i - T_i^{inf})w^{\kappa_i}(T_i^{inf} - T_{\alpha_i}^{inf})$$

$w^{\kappa} = w^{*}w^{*}\ldots{}^{*}w$, where * is the convolution operator and is applied k times. The first term describes the probability of the imputed time of infection under the incubation period distribution. The second term describes the probability of observing the delay between infection times of the case and its most recent sampled ancestor under the generation time distribution, over the imputed number of generations. The reporting likelihood describes the probability of unobserved intermediate cases:

$$\Omega_i^3 = p(\kappa_i|\pi)$$

and is calculated as:

$$NB(1|\kappa_i - 1, \pi)$$

where NB is the probability mass function of the negative binomial distribution, and describes the probability of not observing κ_i—1 cases given a probability of observation of π.

Contact likelihood

To integrate contact data into *outbreaker*, we developed a method for modelling contact data from transmission trees (Fig 4). The model considers undated, undirected, binary contact data, such that the contact status $c_{i,j}$ is set to 1 if contact is reported between individuals i and j and set to 0 otherwise. The model is hierarchical and describes two processes: the occurrence of contacts and the reporting of contacts. Transmission pairs experience contact with probability η. This formulation accounts for the possibility of transmission occurring without direct contact, for example by indirect environmental contamination as is observed with *Clostridium difficile* [52]. Sampled, infected individuals that do not constitute a transmission pair experience contact with probability λ, the non-infectious contact probability. Contacts that have occurred, either between transmission pairs or non-transmission pairs, are then reported with probability ε, the contact reporting coverage. Contacts that have not occurred are reported with probability ζ, the false positive reporting rate.

We make two assumptions to simplify this model, which can be relaxed in future work if necessary. Firstly, we assume that direct contact is necessary for transmission and set η to 1.

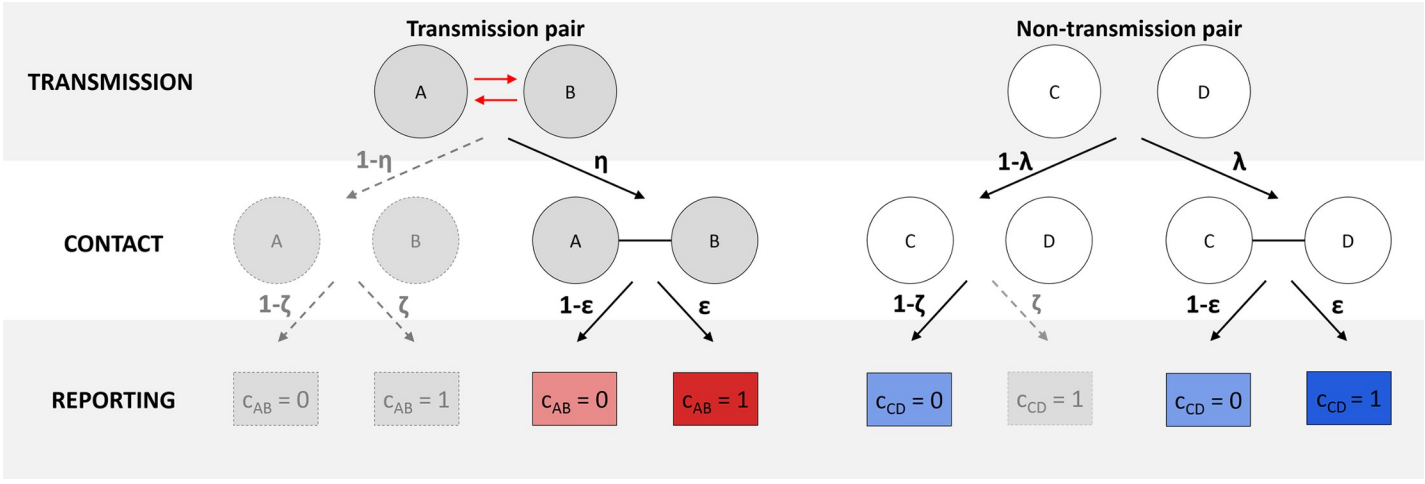

Fig 4. Modelling contact data from transmission trees. Circles represent sampled, infected individuals. $c_{i,j}$ represents the contact status between cases i and j, with 1 indicating a reported contact and 0 the absence of a reported contact. Transmission pairs and non-transmission pairs experience contact with probabilities η and λ, respectively. These contacts are reported with probability ε. False positive reporting of contacts that have not occurred occurs with probability ζ. In the simplified model implemented in *outbreaker2*, as indicated by colored shading and solid outlines, η is assumed to be 1 and ζ assumed to be 0.

https://doi.org/10.1371/journal.pcbi.1006930.g004

Furthermore, we assume that false reporting of contacts that have not occurred is negligible and set ζ to zero. This model allows us to define a contact likelihood Ω^4, describing the probability of observing the contact data C (a symmetrical, binary, NxN adjacency matrix with zeros on its diagonal) given a proposed transmission tree and parameters ε and λ. Formally, for individual i:

$$\Omega_i^4 = \prod_{i=1, j\neq i}^{N} p(c_{i,j}|\alpha_i, \kappa_i, \epsilon, \kappa)$$

Using the contact model described in Fig 4 and the simplifying assumptions made above:

$$p(c_{i,j} = 1|\alpha_i = j, \kappa_i = 1) = \epsilon$$

$$p(c_{i,j} = 0|\alpha_i = j, \kappa_i = 1) = 1 - \epsilon$$

$$p(c_{i,j} = 1|\alpha_i \neq j) = p(c_{i,j} = 1|\alpha_i = j, \kappa_i > 1) = \lambda\epsilon$$

$$p(c_{i,j} = 0|\alpha_i \neq j) = p(c_{i,j} = 0|\alpha_i = j, \kappa_i > 1) = (1 - \lambda) + \lambda(1 - \epsilon)$$

For a mathematical description of the unsimplified model, see S2 Text. The updated joint posterior distribution is therefore proportional to the product of the four likelihood terms and the joint prior:

$$P(A, \theta|D) \propto p(\alpha, \mu, \pi, \epsilon, \lambda) \prod_{i=1}^{N} \Omega_i^1 \Omega_i^2 \Omega_i^3 \Omega_i^4$$

Prior distributions

The prior distributions are assumed independent, such that:

$$p(\alpha, \mu, \pi, \epsilon, \lambda) = p(\alpha)p(\mu)p(\pi)p(\epsilon)p(\lambda)$$

The prior on ancestries p(α) is uniform, and the prior on the mutation rate μ exponentially distributed. π, ε and λ represent probabilities and are assigned Beta distributed priors with user-defined parameters, to allow flexible specification of previous knowledge on the sampling coverage, contact reporting coverage and non-infectious contact probability.

Simulation scenarios

Transmission trees and genetic sequence evolution were simulated using the *simOutbreak* function from the R package *outbreaker*. To describe heterogeneities in infectiousness within a population, well-documented in both EBOV [34] and SARS-CoV [25] outbreaks, and capture consequent 'superspreading' events, in which a small portion of the population accounts for a large number of infections, we described the 'individual reproductive number' R_i, a variable describing the expected number of secondary cases caused by a particular infected individual [2]. Following previous studies by Lloyd-Smith *et al.* [2] and Grassly and Fraser [53], we assumed R_i to be Gamma distributed with a mean of R_0 and a dispersion parameter k, with lower values of k indicating greater heterogeneity in infectiousness. The resulting offspring distribution is a negative binomial [2].

Estimates of the generation time distribution, R_0, mutation rate and genome length were taken from a literature review described by Campbell *et al.* [17] Estimates of the incubation period distribution and dispersion parameter of R_i were drawn from the literature (Table 2).

Generation time distributions and incubation period distributions were described by discretized gamma distributions, generated using the function *DiscrSI* from the R package *EpiEstim* [54].

Contact data was simulated from transmission trees using the model described in Fig 3, using a grid of values for the reporting coverage ($\varepsilon \in [0, 1]$) and the number of non-infectious contacts per person ($\psi \in [0, 10]$, $\lambda \in [0, 0.18]$). For a mathematical description of the relationship between ψ and λ, see S3 Text. At each grid point, 100 outbreaks were simulated, with a single initial infection in a susceptible population of 200 individuals. Simulations were run for 100 days, or until no more infectious individuals remained. The first 60 ancestries of each outbreak were reconstructed four times using the R package *outbreaker2*, using combinations of times of symptom onset (T), contact data (C) and WGS (G): T, TC, TG and TCG. For each analysis, one MCMC chain was run for 10,000 iterations with a thinning frequency of 1/50 and a burn-in of 1,000 iterations. The prior distributions used for ε and λ were uninformative (Beta(1,1)), and default priors used otherwise.

Quantifying accuracy and statistical confidence

The accuracy of outbreak reconstruction was defined as the proportion of correctly assigned ancestries in the consensus transmission tree, itself defined as the tree with the modal posterior infector for each case. The uncertainty associated with an inferred ancestry was quantified using the Shannon entropy of the frequency of posterior ancestors for each case [68]. Given K ancestors of frequency f_K ($k = 1, \ldots, K$), the entropy was defined as:

$$-\sum_{k=1}^{K} f_k log(f_k)$$

Analyzing the 2003 SARS outbreak in Singapore

Thirteen previously published [35,36] and aligned [6] SARS whole genome sequences were obtained for our analysis. Data on epidemiological contacts were described by Vega *et al.* [36]. The same generation time distribution and incubation period distribution used for the analysis of the simulated SARS outbreaks were used (Table 2). As the number of non-transmission contacts was assumed to be low and a total of 6 contacts were reported in an outbreak of 13 cases, the proportion of contacts reported was believed to be about 50%. The prior on ε was therefore chosen as Beta(5, 5). Several priors on the non-transmission contact rate λ were tested; Beta(1, 10), Unif(0, 1), a fixed value of 0 and a fixed value of 1e-4. The priors on the mutation rate μ and proportion of cases sampled π were uninformative. The MCMC chain was run for 1e7 iterations with a thinning frequency of 1/50 and a burn-in of 1,000 iterations.

Table 2. Epidemiological and genetic parameters for EBOV and SARS-CoV.

Parameter	EBOV	SARS-CoV
Mean generation time in days (SD)	14.4 (8.9) [1,55,56]	8.7 (3.6) [57–59]
Mean incubation period (SD)	9.1 (7.3) [55]	6.4 (4.1) [60]
Mean R_i (dispersion)	1.8 (0.18) [34,55]	2.7 (0.16) [2,32]
Mutation rate (per site per day)	0.31×10^{-5} [61–63]	1.14×10^{-5} [36,64,65]
Genome length (bases)	18958 [61,66]	29714 [35,67]

https://doi.org/10.1371/journal.pcbi.1006930.t002

Supporting information

S1 Text. Derivation of genetic likelihood.
(DOCX)

S2 Text. Derivation of un-simplified contact model.
(DOCX)

S3 Text. Derivation of the number of non-infectious contacts per person from the non-infectious contact probability.
(DOCX)

S1 Fig. Example of simulated transmission tree, contact network and reconstructed transmission tree. A) An Ebola-like outbreak of 15 cases was simulated in a susceptible population of 50 susceptible individuals. **B)** A contact network was simulated with a reporting coverage ε of 0.8 and a non-infectious contact probability λ of 0.1. Solid lines represent reported contacts; green lines correspond to transmission pairs, red lines to non-transmission pairs. Dashed green lines represent contacts between transmission pairs that were not reported. **C)** The outbreak was reconstructed using temporal and genomic data, and the consensus transmission tree, describing the modal posterior infector for each case, determined. Green lines correspond to correctly inferred ancestries, red lines to incorrectly inferred ancestries. The accuracy of outbreak reconstruction was 46%. **D)** The outbreak was reconstructed using temporal, genomic and contact data, with an accuracy of 94%.
(TIF)

S2 Fig. Statistical confidence in ancestry assignment using different types of outbreak data. 100 outbreaks were simulated and reconstructed at each grid point, using different values for the contact reporting coverage ε and number of non-infectious contacts per case ψ. Each outbreak was reconstructed four times, using different combinations of times of sampling (T), contact tracing data (C) and genetic data (G). The colour of a grid point represents the average entropy of ancestry assignments and is related to the number of plausible infectors of a given case. Lower average entropy indicates greater statistical confidence in the proposed transmission tree.
(EPS)

S3 Fig. Parameter estimates of the contact reporting coverage ε and non-infectious contact probability λ for simulated EBOV outbreaks. The density plots represent the mean posterior estimates of ε and λ across 100 reconstructed outbreaks. The shading represents the data used during the inference process, namely temporal and contact data only (TC), or temporal, contact and genetic data (TCG). The true, simulated value is indicated by a vertical dashed line.
(EPS)

S4 Fig. Parameter estimates of the contact reporting coverage ε and non-infectious contact probability λ for simulated SARS-CoV outbreaks. The density plots represent the mean posterior estimates of ε and λ across 100 reconstructed outbreaks. The colour of the plot represents the data used during the inference process, namely temporal and contact data only (TC), or temporal, contact and genetic data (TCG). The true, simulated value is indicated by a vertical dashed line.
(EPS)

S5 Fig. Infection time estimates for the 2003 SARS outbreak in Singapore. The violin plots indicate the posterior distribution of infection times for the 13 cases in the outbreak. The black dots represent times of symptom onset. The colour of the violin plot indicates the settings used

to reconstruct the outbreak, namely using temporal and genetic data only (TG), or temporal, contact and genetic data (TCG). The prior used for the non-transmission contact probability λ is indicated in brackets.
(EPS)

S6 Fig. Posterior ancestries for the 2003 SARS outbreak in Singapore under different prior distributions for non-infectious contact probability λ. Columns represent sampled cases in the outbreak, rows represent potential sampled infectors. The size of each circle represents the posterior frequency of a given infector-infectee pair.
(TIF)

Acknowledgments

We are thankful to Github (https://github.com), CRAN (http://cran.r-project.org) and the wider R community for providing great resources for the development and hosting of *outbreaker2*.

Author Contributions

Conceptualization: Finlay Campbell, Anne Cori, Neil Ferguson, Thibaut Jombart.

Formal analysis: Finlay Campbell, Thibaut Jombart.

Funding acquisition: Neil Ferguson.

Investigation: Finlay Campbell, Thibaut Jombart.

Methodology: Finlay Campbell, Anne Cori, Neil Ferguson, Thibaut Jombart.

Software: Finlay Campbell, Thibaut Jombart.

Supervision: Anne Cori, Neil Ferguson, Thibaut Jombart.

Validation: Finlay Campbell, Thibaut Jombart.

Visualization: Finlay Campbell, Thibaut Jombart.

Writing – original draft: Finlay Campbell, Thibaut Jombart.

Writing – review & editing: Finlay Campbell, Anne Cori, Neil Ferguson, Thibaut Jombart.

References

1. Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N 'faly, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis. 2015; 15: 320–326. https://doi.org/10.1016/S1473-3099(14)71075-8 PMID: 25619149
2. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. Nature Publishing Group; 2005; 438: 355–359.
3. Ferguson NM, Donnelly CA, Anderson RM. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature. 2001; 413: 542–548. https://doi.org/10.1038/35097116 PMID: 11586365
4. Spada E, Sagliocca L, Sourdis J, Garbuglia AR, Poggi V, De Fusco C, et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. J Clin Microbiol. 2004; 42: 4230–4236. https://doi.org/10.1128/JCM.42.9.4230-4236.2004 PMID: 15365016
5. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010; 11: 31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069
6. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol. 2014; 10: e1003457. https://doi.org/10.1371/journal.pcbi.1003457 PMID: 24465202

7. Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc Biol Sci. 2014; 281: 20133251. https://doi.org/10.1098/rspb.2013.3251 PMID: 24619442
8. Lau MSY, Marion G, Streftaris G, Gibson G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. PLoS Comput Biol. 2015; 11: e1004633. https://doi.org/10.1371/journal.pcbi.1004633 PMID: 26599399
9. Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat. 2016; 10: 395–417. PMID: 27042253
10. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol. 2017; https://doi.org/10.1093/molbev/msw275 PMID: 28100788
11. De Maio N, Wu C-H, Wilson DJ. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. PLoS Comput Biol. 2016; 12: e1005130. https://doi.org/10.1371/journal.pcbi.1005130 PMID: 27681228
12. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. PLoS Comput Biol. 2015; 11: e1004613. https://doi.org/10.1371/journal.pcbi.1004613 PMID: 26717515
13. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol. 2017; 13: e1005495. https://doi.org/10.1371/journal.pcbi.1005495 PMID: 28545083
14. Hall MD, Woolhouse MEJ, Rambaut A. Using genomics data to reconstruct transmission trees during disease outbreaks. Rev Sci Tech. 2016; 35: 287–296. https://doi.org/10.20506/rst.35.1.2433 PMID: 27217184
15. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol Biol Evol. 2014; 31: 1869–1879. https://doi.org/10.1093/molbev/msu121 PMID: 24714079
16. Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, et al. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proc Biol Sci. 2003; 270: 121–127. https://doi.org/10.1098/rspb.2002.2191 PMID: 12590749
17. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog. 2018; https://doi.org/10.1371/journal.ppat.1006885 PMID: 29420641
18. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009; 459: 1122–1125. https://doi.org/10.1038/nature08182 PMID: 19516283
19. Dye C, Williams BG. Criteria for the control of drug-resistant tuberculosis. Proc Natl Acad Sci U S A. 2000; 97: 8180–8185. https://doi.org/10.1073/pnas.140102797 PMID: 10859359
20. Crum NF, Wallace MR, Lamb CR, Conlin AMS, Amundson DE, Olson PE, et al. Halting a pneumococcal pneumonia outbreak among United States Marine Corps trainees. Am J Prev Med. 2003; 25: 107–111.
21. Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA. Within-host diversity of MRSA antimicrobial resistances. J Antimicrob Chemother. 2015; 70: 2191–2198. https://doi.org/10.1093/jac/dkv119 PMID: 25957384
22. Stegemann S, Dahlberg S, Kröger A, Gereke M, Bruder D, Henriques-Normark B, et al. Increased susceptibility for superinfection with Streptococcus pneumoniae during influenza virus infection is not caused by TLR7-mediated lymphopenia. PLoS One. 2009; 4: e4840. https://doi.org/10.1371/journal.pone.0004840 PMID: 19290047
23. Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Mol Biol Evol. 2017; https://doi.org/10.1093/molbev/msx304 PMID: 29186559
24. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017; 544: 309–315. https://doi.org/10.1038/nature22040 PMID: 28405027
25. Shen Z, Ning F, Zhou W, He X, Lin C, Chin DP, et al. Superspreading SARS events, Beijing, 2003. Emerg Infect Dis. 2004; 10: 256–260. https://doi.org/10.3201/eid1002.030732 PMID: 15030693
26. Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DAT, et al. Hospital outbreak of Middle East respiratory syndrome coronavirus. N Engl J Med. 2013; 369: 407–416. https://doi.org/10.1056/NEJMoa1306742 PMID: 23782161

27. Who Ebola Response Team. Ebola Virus Disease in West Africa—The First 9 Months of the Epidemic and Forward Projections. N Engl J Med. 2014; 371: 1481–1495. https://doi.org/10.1056/NEJMoa1411100 PMID: 25244186

28. International Ebola Response Team, Agua-Agum J, Ariyarajah A, Aylward B, Bawo L, Bilivogui P, et al. Exposure Patterns Driving Ebola Transmission in West Africa: A Retrospective Observational Study. PLoS Med. 2016; 13: e1002170. https://doi.org/10.1371/journal.pmed.1002170 PMID: 27846234

29. Broeckaert L, Haworth-Brockman M. You may have come into contact with . . .": HIV Contact Tracing in Canada. Prevention. 2014; Available: http://www.catie.ca/en/pif/fall-2014/you-may-have-come-contact-hiv-contact-tracing-canada

30. National Tuberculosis Controllers Association, Centers for Disease Control and Prevention (CDC). Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. MMWR Recomm Rep. 2005; 54: 1–47.

31. R Development Core Team R. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2011. https://doi.org/10.1007/978-3-540-74686-7

32. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science. 2003; 300: 1961–1966. https://doi.org/10.1126/science.1086478 PMID: 12766206

33. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. A: mathematical, ... rspa.royalsocietypublishing.org; 1927; Available: http://rspa.royalsocietypublishing.org/content/royprsa/115/772/700.full.pdf

34. Althaus CL. Ebola superspreading. Lancet Infect Dis. 2015; 15: 507–508. https://doi.org/10.1016/S1473-3099(15)70135-0 PMID: 25932579

35. Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, Su STY, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet. 2003; 361: 1779–1785. PMID: 12781537

36. Vega VB, Ruan Y, Liu J, Lee WH, Wei CL, Se-Thoe SY, et al. Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. BMC Infect Dis. 2004; 4: 32. https://doi.org/10.1186/1471-2334-4-32 PMID: 15347429

37. Liu J, Lim SL, Ruan Y, Ling AE, Ng LFP, Drosten C, et al. SARS transmission pattern in Singapore reassessed by viral sequence variation analysis. PLoS Med. 2005; 2: e43. https://doi.org/10.1371/journal.pmed.0020043 PMID: 15736999

38. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput Biol. 2014; 10: e1003549. https://doi.org/10.1371/journal.pcbi.1003549 PMID: 24675511

39. Soetens L, Klinkenberg D, Swaan C, Hahné S, Wallinga J. Real-time Estimation of Epidemiologic Parameters from Contact Tracing Data During an Emerging Infectious Disease Outbreak. Epidemiology. 2018; 29: 230–236. https://doi.org/10.1097/EDE.0000000000000776 PMID: 29087987

40. Hens N, Calatayud L, Kurkela S, Tamme T, Wallinga J. Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. Am J Epidemiol. 2012; 176: 196–203. https://doi.org/10.1093/aje/kws006 PMID: 22791742

41. Jewell CP, Roberts GO. Enhancing Bayesian risk prediction for epidemics using contact tracing. Biostatistics. 2012; 13: 567–579. https://doi.org/10.1093/biostatistics/kxs012 PMID: 22674466

42. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011; 364: 730–739. https://doi.org/10.1056/NEJMoa1003176 PMID: 21345102

43. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. 2013; 13: 110. https://doi.org/10.1186/1471-2334-13-110 PMID: 23446317

44. Cori A, Donnelly CA, Dorigatti I, Ferguson NM, Fraser C, Garske T, et al. Key data for outbreak evaluation: building on the Ebola experience. Philos Trans R Soc Lond B Biol Sci. 2017; 372. https://doi.org/10.1098/rstb.2016.0371 PMID: 28396480

45. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci U S A. 2004; 101: 6146–6151. https://doi.org/10.1073/pnas.0307506101 PMID: 15071187

46. World Health Organization, Others. Contact tracing during an outbreak of Ebola virus disease. World Health Organization; 2014.

47. Dixon MG, Taylor MM, Dee J, Hakim A, Cantey P, Lim T, et al. Contact Tracing Activities during the Ebola Virus Disease Epidemic in Kindia and Faranah, Guinea, 2014. Emerg Infect Dis. 2015; 21: 2022–2028. https://doi.org/10.3201//eid2111.150684 PMID: 26488116

48. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. AIDS. 2014; 28: 1509–1519. https://doi.org/10.1097/QAD.0000000000000298 PMID: 24809629

49. Cauchemez S, Ferguson NM. Methods to infer transmission risk factors in complex outbreak data. J R Soc Interface. 2012; 9: 456–469. https://doi.org/10.1098/rsif.2011.0379 PMID: 21831890

50. Kenah E. Contact intervals, survival analysis of epidemic data, and estimation of R(0). Biostatistics. 2011; 12: 548–566. https://doi.org/10.1093/biostatistics/kxq068 PMID: 21071607

51. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004; 160: 509–516. https://doi.org/10.1093/aje/kwh255 PMID: 15353409

52. Jou J, Ebrahim J, Shofer FS, Hamilton KW, Stern J, Han JH, et al. Environmental transmission of Clostridium difficile: association between hospital room size and C. difficile Infection. Infect Control Hosp Epidemiol. 2015; 36: 564–568. https://doi.org/10.1017/ice.2015.18 PMID: 25652311

53. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol. 2008; 6: 477–487. https://doi.org/10.1038/nrmicro1845 PMID: 18533288

54. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013; 178: 1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437

55. WHO Ebola Response Team. Ebola Virus Disease in West Africa—The First 9 Months of the Epidemic and Forward Projections. The England New Journal of Medicine. 2014; 371: 1481–1495.

56. WHO Ebola Response Team. West African Ebola Epidemic after One Year—Slowing but Not Yet under Control. N Engl J Med. 2015; 372: 584–587. https://doi.org/10.1056/NEJMc1414992 PMID: 25539446

57. Lipsitch M. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. Science. 2003; 300: 1966–1970. https://doi.org/10.1126/science.1086616 PMID: 12766207

58. Reynolds MG, Anh BH, Thu VH, Montgomery JM, Bausch DG, Shah JJ, et al. Factors associated with nosocomial SARS-CoV transmission among healthcare workers in Hanoi, Vietnam, 2003. BMC Public Health. 2006; 6: 207. https://doi.org/10.1186/1471-2458-6-207 PMID: 16907978

59. Varia M, Wilson S, Sarwal S, McGeer A, Gournis E, Galanis E, et al. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. CMAJ. 2003; 169: 285–292. PMID: 12925421

60. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet. 2003; 361: 1761–1766. https://doi.org/10.1016/S0140-6736(03)13410-1 PMID: 12781533

61. Hoenen T, Groseth A, Feldmann F, Marzi A, Ebihara H, Kobinger G, et al. Complete Genome Sequences of Three Ebola Virus Isolates from the 2014 Outbreak in West Africa. Genome Announc. 2014; 2: 647–648.

62. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014; 345: 1369–1372. https://doi.org/10.1126/science.1259657 PMID: 25214632

63. Tong Y-G, Shi W-F, Di Liu, Qian J, Liang L, Bo X-C, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015; https://doi.org/10.1038/nature14490 PMID: 25970247

64. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol. 2004; 4: 21. https://doi.org/10.1186/1471-2148-4-21 PMID: 15222897

65. Wu S-F, Du C-J, Wan P, Chen T-G, Li J-Q, Li D, et al. The genome comparison of SARS-CoV and other coronaviruses. Yi chuan = Hereditas/Zhongguo yi chuan xue hui bian ji. 2003; 25: 373–382.

66. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N 'faly, et al. Emergence of Zaire Ebola Virus Disease in Guinea—Preliminary Report. N Engl J Med. 2014; 1–8.

67. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science. 2003; 300: 1394–1399. https://doi.org/10.1126/science.1085952 PMID: 12730500

68. Shannon CE. The mathematical theory of communication. 1963. MD Comput. 1997; 14: 306–317. PMID: 9230594