PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bayesian reconstruction of transmission within outbreaks using genomic variants

Nicola De Maio[1]*, Colin J. Worby[2], Daniel J. Wilson[1,3☯], Nicole Stoesser[1☯]

1 Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, **2** Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, United States of America, **3** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

☯ These authors contributed equally to this work.
* demaio@ebi.ac.uk

OPEN ACCESS

Citation: De Maio N, Worby CJ, Wilson DJ, Stoesser N (2018) Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput Biol 14(4): e1006117. https://doi.org/10.1371/journal.pcbi.1006117

Editor: Katia Koelle, Duke University, UNITED STATES

Received: November 9, 2017

Accepted: April 3, 2018

Published: April 18, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Daniel J. Wilson was supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust (www.wellcome.ac.uk/) and the Royal Society (https://royalsociety.org/grant101237/Z/13/Z). CJW was supported by the Bill and Melinda Gates Foundation (https://www.gatesfoundation.org/grantOPP1091919). NS is currently funded through a Public Health England (https://www.gov.uk/government/organisations/

Abstract

Pathogen genome sequencing can reveal details of transmission histories and is a powerful tool in the fight against infectious disease. In particular, within-host pathogen genomic variants identified through heterozygous nucleotide base calls are a potential source of information to identify linked cases and infer direction and time of transmission. However, using such data effectively to model disease transmission presents a number of challenges, including differentiating genuine variants from those observed due to sequencing error, as well as the specification of a realistic model for within-host pathogen population dynamics. Here we propose a new Bayesian approach to transmission inference, BadTrIP (BAyesian epiDemiological TRansmission Inference from Polymorphisms), that explicitly models evolution of pathogen populations in an outbreak, transmission (including transmission bottlenecks), and sequencing error. BadTrIP enables the inference of host-to-host transmission from pathogen sequencing data and epidemiological data. By assuming that genomic variants are unlinked, our method does not require the computationally intensive and unreliable reconstruction of individual haplotypes. Using simulations we show that BadTrIP is robust in most scenarios and can accurately infer transmission events by efficiently combining information from genetic and epidemiological sources; thanks to its realistic model of pathogen evolution and the inclusion of epidemiological data, BadTrIP is also more accurate than existing approaches. BadTrIP is distributed as an open source package (https://bitbucket.org/nicofmay/badtrip) for the phylogenetic software BEAST2. We apply our method to reconstruct transmission history at the early stages of the 2014 Ebola outbreak, showcasing the power of within-host genomic variants to reconstruct transmission events.

Author summary

We present a new tool to reconstruct transmission events within outbreaks. Our approach makes use of pathogen genetic information, notably genetic variants at low frequency within host that are usually discarded, and combines it with epidemiological information of host exposure to infection. This leads to accurate reconstruction of transmission even in cases where abundant within-host pathogen genetic variation and weak transmission

public-health-england)/University of Oxford (http://www.ox.ac.uk/) Clinical Lectureship. N.D.M. was supported by the Oxford Martin School (http://www.oxfordmartin.ox.ac.uk/). 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.

bottlenecks (multiple pathogen units colonising a new host at transmission) would otherwise make inference difficult due to the transmission history differing from the pathogen evolution history inferred from pathogen isolets. Also, the use of within-host pathogen genomic variants increases the resolution of the reconstruction of the transmission tree even in scenarios with limited within-outbreak pathogen genetic diversity: within-host pathogen populations that appear identical at the level of consensus sequences can be discriminated using within-host variants. Our Bayesian approach provides a measure of the confidence in different possible transmission histories, and is published as open source software. We show with simulations and with an analysis of the beginning of the 2014 Ebola outbreak that our approach is applicable in many scenarios, improves our understanding of transmission dynamics, and will contribute to finding and limiting sources and routes of transmission, and therefore preventing the spread of infectious disease.

Introduction

Understanding transmission is important for devising effective policies and measures that limit the spread of infectious diseases. In recent years, affordable whole genome sequencing has provided unprecedented detail on the relatedness of pathogen samples [1–4]. Consequently, accurately inferring transmission between hosts is becoming more feasible. However, this requires robust statistical approaches that make use of the full extent of genetic and epidemiological data available. Here, we present a new approach that makes use of within-host genetic variation and epidemiological data to infer transmission.

A number of approaches have been developed that reconstruct transmission from genetic data. The number of substitutions between samples from different hosts can be used to rule out transmission [5–7], or the phylogenetic tree of the pathogen samples can be used as a proxy for the transmission history [8, 9]. However, while the phylogenetic signal can be very informative of transmission, it can also be misleading [10, 11], due to within-host variation that can generate discrepancies between the phylogenetic and epidemiological relatedness of hosts, and can bias estimates of infection times [12, 13].

In recent years a number of methods have been proposed explicitly modelling both the transmission process and within-host pathogen genetic evolution to infer transmission events [11, 13–28]. Some of these methods use epidemiological data and genetic sequences from pathogen samples, and ignore within-host evolution and other causes of phylogenetic discordance with transmission history [14–19, 21–23]. Methods that explicitly model pathogen evolution within hosts and within an outbreak [13, 20, 24, 25, 27] generally assume, among other things, that samples provide individual and reliable pathogen haplotypes. This is often true for bacteria that are sampled and cultured before being sequenced, but it is mostly false for viruses and bacteria that are sequenced directly from samples without culturing. In fact, in these cases the sequencing process delivers reads coming from the different pathogen haplotypes that constitute the within-host pathogen population, and it is often very hard (if not impossible) to reconstruct complete haplotypes from these reads. In such cases, within-sample genetic variation is often neglected, and a single haplotype (which we call the consensus sequence of the sample) is built. While this procedure might lead to errors (and maybe biases), it also certainly discards a very informative part of the available genetic data, because within-sample genetic variants can be very informative of epidemiological distance, direction of transmission, time from infection and transmission bottleneck intensity (see [29–32] and Fig 1). Furthermore, it is generally assumed that the pathogen does not recombine, so that a single phylogeny describes the

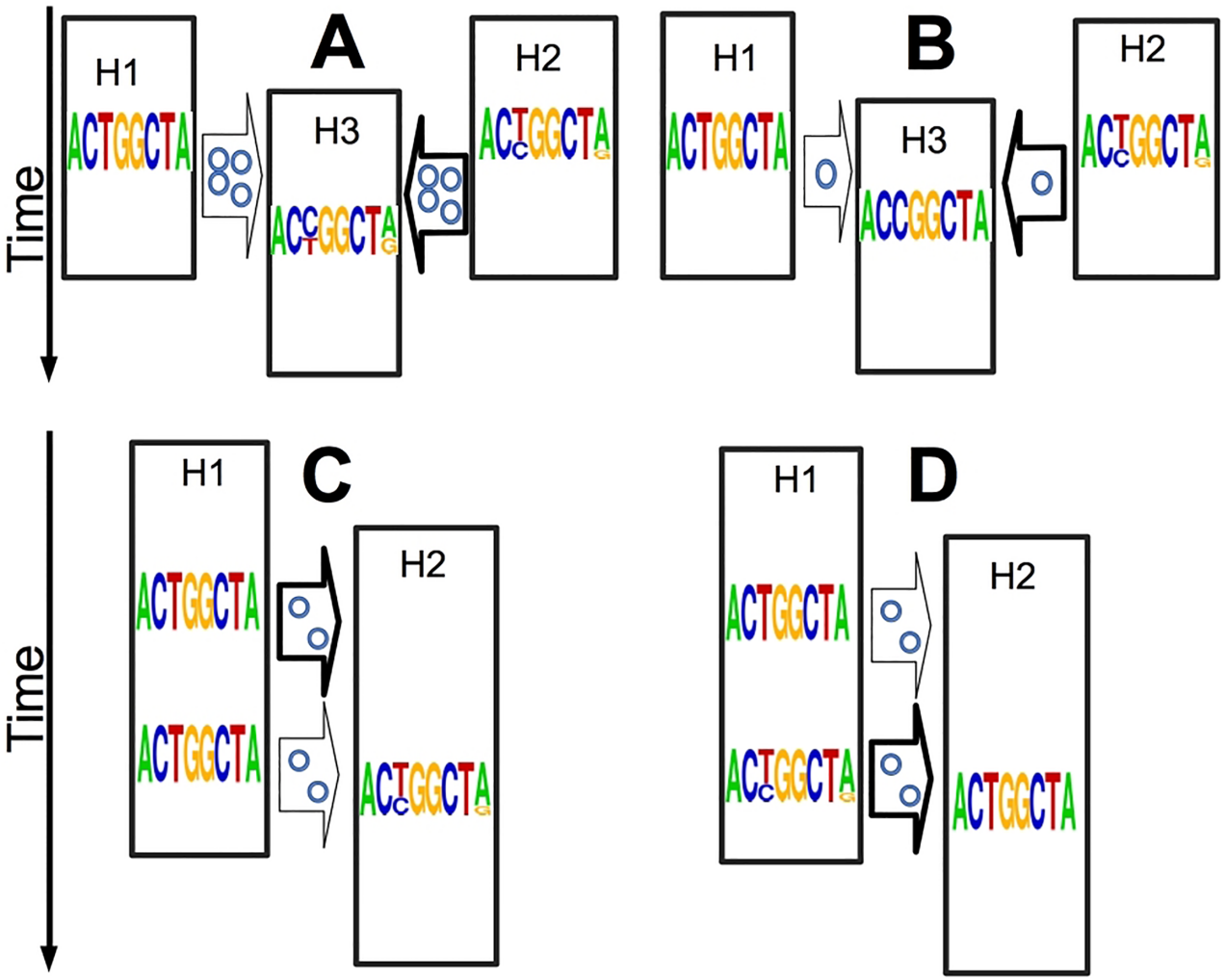

Fig 1. Examples of informativeness of within-host genetic variants. Here we show how within-host within-sample genetic variants can be useful without requiring pathogen haplotypes. Each string of letters (a frequency sequence logo [34, 35]) represents the collective genome of the pathogen at a certain point in time, as could be observed through deep sequencing. Multiple letters in the same column represent a genetic variant, with letter size representing allelic abundance. Time is on the Y axis, hosts are represented as black rectangles (a host is only active in the outbreak for the portion of vertical axis it occupies), and plausible transmission events as arrows. The posterior probability of different transmission events is represented by the arrow thickness. The number of little circles within arrows represents the inoculum size (transmission bottleneck). **A**) Shared genetic variants hint to epidemiological relatedness: the two top hosts (H1 and H2) are both possible infectors of the central host (H3), but H2 shares two genetic variants with H3, making it a likely infector of H3. Furthermore, the presence of shared genetic variants suggests a large transmission inoculum (a weak transmission bottleneck). **B**) A genetic variant of the same type of a substitution can hint to an infector: as before, but now H3 has a substitution (at third genome position, from T to C), which means that its within-host population is non-polymorphic at this position, but with a different nucleotide than the index case. This substitution is between the two nucleotides present at the same position in H2 (where this position is a genetic variant), consistent with H2 being the infector of H3. Also, this time the absence of shared genetic variants is indicative of a small transmission inoculum (a strong transmission bottleneck). **C**-**D**) The number of new genetic variants is informative of the age of an infection (but possibly also of the history of the pathogen population size within the host): in **C** the presence of non-shared variants in H2 suggests that the infection is older, while in **D** their absence suggests that the infection is younger.

https://doi.org/10.1371/journal.pcbi.1006117.g001

evolutionary history of the whole genome, but this assumption does not fit highly recombinant pathogens such as HIV [33]. For these reasons, a few approaches have recently been proposed that use within-host genetic variants to reconstruct transmission [30, 32].

Here, we propose a new Bayesian approach called BadTrIP (BAyesian epiDemiological TRansmission Inference from Polymorphisms) that not only uses within-sample genetic

variants (from possibly multiple samples per host) to reconstruct transmission (including directionality and time of infection), but also combines this information with epidemiological data and an explicit model of within-host pathogen population evolution and transmission. We use the phylogenetic models with polymorphisms PoMo [36–38] to model population evolution along branches of the transmission tree; thanks to this, our transmission tree and phylogenetic tree are the same entity, and within-host evolution and recombination (resulting from a single primary infection, not multiple infections) do not create discrepancies that make statistical inference hard and computationally demanding [24, 25, 27]. We also explicitly model transmission bottlenecks, with one parameter defining the intensity of the bottleneck, and therefore the number of pathogen particles that establish a new population at transmission. Another feature of our approach is that we assume that different genomic positions are unlinked, an assumption also made by other methods using within-host variants [30, 32]; most coalescent-based methods assume instead no recombination at all. Because of our assumption of no linkage, we expect our approach to work well when recombination is strong enough to break linkage between genetic variants in the same host, or when the evolutionary rate is slow so that very few new mutations originate with each new transmission.

BadTrIP is implemented as an open-source package for the Bayesian phylogenetic software BEAST2 [39], and as such, it can be freely installed, used, and modified. We compare the performance of BadTrIP, of the shared variants-based clustering (SVC) method of [30], and of the coalescent-based method SCOTTI [13] on simulated data and on a real dataset from the early stages of the 2014 Ebola outbreak [40]. These applications show that BadTrIP has high accuracy to reconstruct transmission thanks to its explicit model of population evolution, the use of within-host genetic variants, and the inclusion of epidemiological data, and can provide important information to understand and limit the spread of infectious disease.

In the rest of the manuscript, we refer to a "host" as any entity that can contain and transmit a pathogen. Typically a host is a human within a community or nosocomial outbreak, or patients, but the concept of host can also be generalised for example to farms within a livestock outbreak. We will refer to the collection of all pathogens of the type under consideration within an individual host at a certain time as a "pathogen population" (for example all Ebola virions within an infected host, excluding non-Ebola pathogens and Ebola virions from other hosts). We will call a "pathogen unit" a single pathogen individual within a population, for example an individual bacterial cell or an individual virion. We call a pathogen population "polymorphic" at a particular genome position if pathogen units with different nucleotides at that position are present in the population; in this case, we also call the considered genome position a "genetic variant".

Results

Modelling within-host evolution, transmission, and sequencing

Methods to reconstruct transmission that account for within-host evolution usually have to deal with the complex task of modelling and inferring the discrepancies between the transmission tree and the pathogen phylogenetic trees [13, 20, 24, 25, 27]. We avoid this complication by adopting and adapting a substitution model, PoMo [36–38], that describes population evolution along the branches of a species (or population) tree. In this model, a virtual population, similar to a Moran model [41] without selection and with fixed population size, evolves by accumulating random changes in nucleotide frequencies (genetic drift, eventually resulting in the fixation of polymorphic sites), and new mutations resulting in new polymorphic sites. Different genome positions are modelled as completely unlinked.

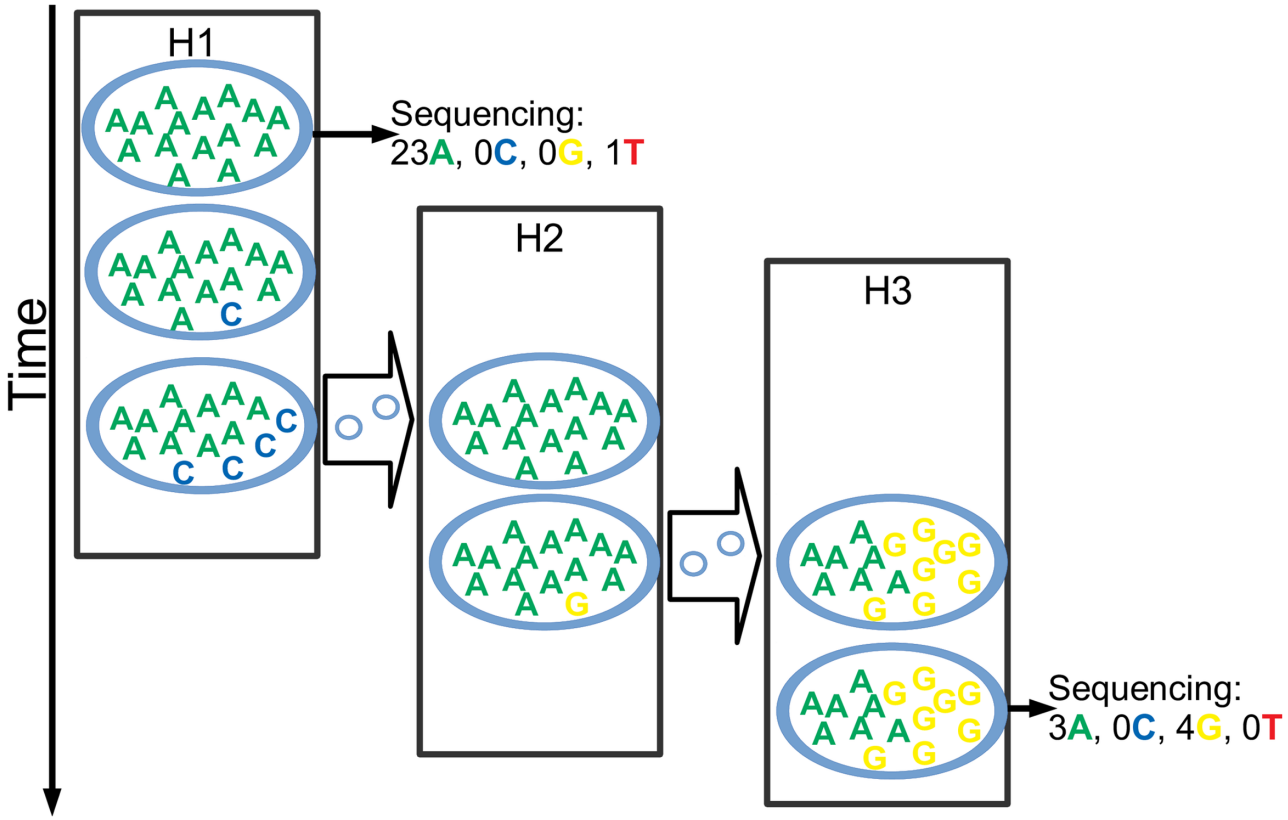

Fig 2. Graphical representation of the transmission, evolution and sequencing model. Here we describe some key aspects of our model. The figure depicts a possible evolutionary outcome for one position of the pathogen genome and the given transmission history. There are three hosts in this outbreak, represented by the black rectangles: H1 infects H2, which in turn infects H3. Time is on the vertical axis, and transmission events are represented by the thick arrows between hosts. Within each host, while it is colonised, the pathogen population consists of 15 units, each of which can have one of the four nucleotides at the considered position and at any time. For example, H1 starts off with all 15 pathogen units having an A, but during infection one of them mutates to C, and through genetic drift when H1 infects H2 it has 4 C's and 11 A's. While instantaneously only small changes can occur (one pathogen unit changing its nucleotide), along a time interval any number of changes can occur. As H2 is infected by H1, H2 is colonised by a copy of the pathogen population of H1, but the transmission bottleneck in this case causes one of the nucleotides to be lost, so that H2 is founded by a homogenous population of A's. Within H2 again a mutation occurs and now a G is present in the pathogen population, but when H3 is colonised by H2 both nucleotides survive the transmission bottleneck, so H3 starts off with a polymorphic population. In the figure, H1 and H3 both have samples extracted and sequenced once, while H2 is not sampled at all. The sequencing process can result in any coverage (24 for H1 and 7 for H3 at the considered position). Furthermore, the observed nucleotide frequencies don't necessarily exactly match the real nucleotides frequencies due to the randomness of read sampling, and because sequencing error can cause absent nucleotides to be observed at very low frequencies.

https://doi.org/10.1371/journal.pcbi.1006117.g002

The adoption of such a population genetic model within a transmission tree structure means that the phylogenetic tree and the transmission tree are now the same entity, and that each point of the tree represents the state of the pathogen population at a certain time within a host (Fig 2). Each bifurcation in the tree represents a transmission event, where the pathogen population splits in two groups: one remaining in the current host, and a small sub-population colonising a new host. We use a population bottleneck at time of transmission for the colonising branch to better model the transmission process.

Our method uses two sources of information: epidemiological and genetic data. Epidemiological data is in the form of dates: the times when genetic samples are collected (it is possible to give any number of samples ≥ 0 for any host, even no sample at all) and a time interval for each host describing when it can contribute to the outbreak. Each host can only be infected, be sampled, and can infect other hosts within its time interval [13]. Genetic data from each sample is in the form of nucleotide counts: for each position of the genome, for a certain sample,

the model expects the number of times each of the four nucleotides is observed in the reads (for example: 59 As, 0 Cs, 12 Gs, 1 Ts). We assume that reads are sampled with replacement from the pathogen population according to the (hidden) true nucleotide frequencies, and we model the sequencing error. This in particular means that sites without any sequencing coverage, or with very low coverage, are also allowed, and that differently from similar approaches (i.e. [30, 32]) we don't require the specification of a minimum genetic variant frequency threshold.

While in our model we make the strong assumption that sites are completely unlinked, we test the performance of our approach with simulations in which we explicitly model within-host recombination events and we assume that a limited number of individuals in the pathogen population is sequenced. We even simulate scenarios in the total absence of recombination (complete linkage) to measure the robustness of our method. We simulate a broad range of scenarios: different transmission bottleneck severities (weak vs. strong), different amounts of genetic information, different recombination and mutation rates, different sequencing coverage levels, different sequencing error rates, and different virtual population sizes. We give further details on the model used and the simulations in the Materials and Methods section.

Accuracy of inference on simulated data

To test the accuracy of our new method BadTrIP in inferring transmission events, and to compare it to previous methods [13, 30], we simulated pathogen evolution within outbreaks and sample sequencing, and we used different methods to reconstruct the transmission history from sequencing and epidemiological data. To simulate pathogen evolution, first we simulated an outbreak using SEEDY [42] (we used a fixed population of 15 hosts, one initial case, and a basic reproduction number of 1.43, see Materials and methods); then, we translated the transmission history into a population history, and simulated within-population pathogen coalescent, recombination and mutation with fastsimcoal2 [43]. Throughout the simulations each host in the outbreak is sampled exactly once. We measure the accuracy of a method as the frequency with which the correct transmission source of each host is inferred to be the most likely a posteriori. We also give a measure of how well calibrated [44] methods are by counting how often the correct source is in the 95% posterior credible set, defined as the minimum set of sources with cumulative probability $\geq$ 95% such that all sources in the set have higher posterior probability than all sources outside of it.

BadTrIP shows elevated accuracy in detecting the correct source of transmission (between 50% and 90%) and calibration (between 80% and 100%), in particular compared to the SVC approach (accuracy between 20% and 45% and calibration between 45% and 95%), see Fig 3. This shows that the use of epidemiological data and an explicit model of evolution can help to reconstruct transmission. Using alternative statistics for accuracy and calibration leads to similar patterns (Fig F in S1 Text). BadTrIP also shows more accuracy than the coalescent approach SCOTTI (accuracy between 25% and 70%). The latter method appears very conservative in this application (calibration between 95% and 100%). SCOTTI uses the same epidemiological information as BadTrIP, but a different format of genetic data and a different model of genetic evolution. In fact, like most coalescent-based approaches, SCOTTI requires a full haplotype to be given for each sample; in these simulations we used the consensus sequence of a sample as its haplotype for SCOTTI, discarding within-sample genetic variation. The fact that SCOTTI has strictly less genetic information available than BadTrIP can explain why generally it has less accuracy and is more conservative, however it is not the only factor at play, another being recombination. For example in the scenario with 1x coverage BadTrIP seems to have higher accuracy than SCOTTI, despite the two methods having the same

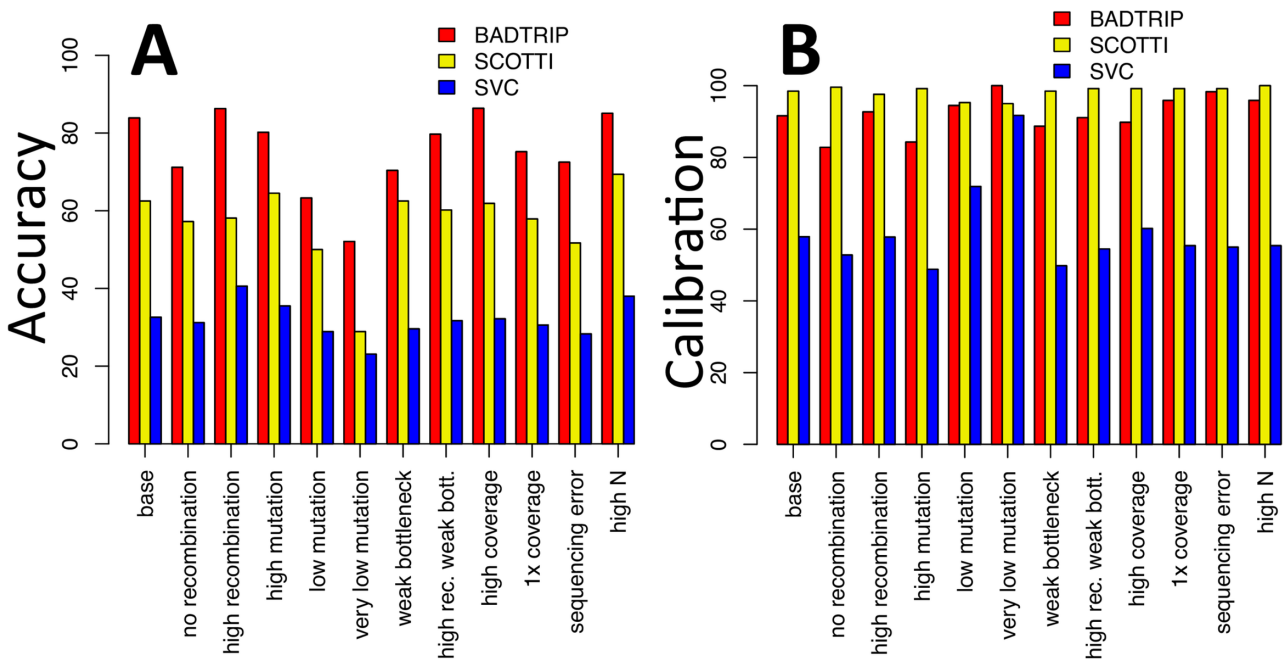

Fig 3. Accuracy and calibration of BadTrIP on simulated data. A) We represent accuracy as the frequency with which the correct simulated transmission event is more likely a posteriori than the alternatives. **B**) Calibration is the frequency with which the correct transmission event is in the 95% posterior credible set (the minimum set of sources with cumulative probability $\geq$ 95% such that all sources in the set have higher posterior probability than all sources outside of it). Bars represent percentages (from 0, worst, to 100, best) for BadTrIP (red), SCOTTI [13] (yellow) and the shared variants-based clustering (SVC) approach [30] (blue). On the x axis are different simulation scenarios with the first one, "base", being the basic simulation scenario with 10–15 cases per outbreak, about 300–500 SNPs among all hosts, recombination 10 times stronger than mutation, complete bottleneck (no transmission of within-host genetic variants), read coverage of 40x, PoMo virtual population size of 15, actual pathogen population size of 1000, and genome size of 5 kb. All other scenarios are obtained from the base one changing one or two parameters: in "no recombination" the recombination rate is set to 0; in "high recombination" the recombination rate is 10 times higher; in "high mutation" the mutation rate is 10 times higher resulting in 2000–3000 SNPs per outbreak; in "low mutation" the mutation rate is 10 times lower resulting in 30–50 SNPs per outbreak; in "very low mutation" the mutation rate is 1000 times lower, resulting in 0–1 SNPs per outbreak; in "weak bottleneck" at transmission 5 pathogen units from the infector colonised the infected host, instead of just 1; in "high rec. weak bott." both the recombination rate is 10 times higher and the founding population at transmission is made of 5 pathogen particles; in "high coverage" read coverage in sequencing is 100x instead of 40x; in "1x coverage" read coverage in sequencing is 1x instead of 40x; in "sequencing error" 0.2% of read bases are randomly modified to simulate sequencing error, coverage is reduced to 20x, and genome size is reduced to 1kb; in "high N" the PoMo virtual population size is 25 instead of 15.

https://doi.org/10.1371/journal.pcbi.1006117.g003

information available: this can be explained with the fact the SCOTTI wrongly assumes that there is no recombination. Similarly, the simulations suggest that the accuracy gap between SCOTTI and BadTrIP reduces with no recombination, and increase at high recombination: this fits well with the fact BadTrIP assumes no linkage between genomic positions, while SCOTTI assumes complete linkage (no recombination). While these results are very suggestive and fit with our expectations, we also have to warn that for each individual scenario we have 10–20 simulated outbreaks, so while the general patterns are clear, the specific patterns of each scenario are subject to considerable uncertainty.

Comparing the base scenario with the one with almost no mutation, we see that BadTrIP accuracy drops from about 80% to about 50%; this drop hints to the contribution given by genetic data to the inference of transmission. Also, since in the latter scenario almost no genetic information is available, it also suggests what is the contribution of epidemiological information alone. Calibration of BadTrIP seems to increase as mutation rate decreases, one probable contributing factor being that as mutation rate decreases the effect of genetic linkage on the pathogen evolutionary dynamics decreases (neither method models genetic linkage), or possibly as a result of the increased uncertainty on the evolutionary process. The complete

absence of recombination seems to negatively affect calibration in BadTrIP, but the difference is not dramatic (from about 90% to about 80%) suggesting that even in the worst case scenario of complete absence of recombination BadTrIP can still provide meaningful inference and posterior distributions. Accuracy of all methods seems to decrease with decreasing mutation rate, as is expected because of the reduced genetic information. However, very high mutation rates (to the point that about half the genome, of length 5kb, is polymorphic within the outbreak) do not seem to improve inference, probably because of saturation.

Accuracy of BadTrIP seems higher (around 10% difference) in the presence of a strong bottleneck (small inoculum) than a weak bottleneck (large inoculum), while calibration seems almost unaffected; this probably happens because, with strong bottlenecks, polymorphisms are unlikely shared between hosts, and so polymorphisms leading to substitutions (see Fig 1B) become more informative for identifying infectors. An increase in coverage (from 40x to 100x) does not seem to bring improvement in accuracy or calibration to BadTrIP; on the other hand, when a single uniform colony is sequenced (which is equivalent to reducing coverage to 1x, and therefore removing information on within-host genetic variation), accuracy seems moderately reduced ($\approx$ 10%) but not calibration. Introducing sequencing error (0.2% of mis-called bases, slightly more than what typical for high-throughput DNA sequencing [45]) accompanied by reduced coverage (20x) and genome length (1kb) still seems to result in elevated accuracy (72.5%) and calibration (97.5%). Increasing the PoMo virtual population size (from 15 to 25, while the actual simulated population size remains 1000) showed negligible effects on the inference.

BadTrIP also infers the time of infection. Calibration seems to increase with recombination, and to decrease with mutation (Fig 4), probably again an effect of our assumption of no linkage. Also, very high mutation rates seem to reduce the error in time inference, as do high coverage and virtual population size.

The running time of BadTrIP is affected by the number of genetic variants present in the alignment and by the number of hosts present in the outbreak (Fig A in S1 Text). The number of variants affect the number of likelihoods that need to be calculated at each MCMC step, while the number of hosts affects the size of the transmission/population tree (so both the computational and statistical complexity of BadTrIP). However, the time required to complete an analysis is not always a linear function of these two quantities: at low mutation rates BadTrIP requires similar times for different outbreak sizes. The reason is probably that with less data there is more uncertainty (in particular in the posterior distribution of the mutation rate), and so it takes longer to explore the the parameter space effectively. Overall, it takes a few hours to completely investigate an outbreak of moderate size (one or two dozen hosts) with BadTrIP.

Analysis of the early 2014 Ebola outbreak in Sierra Leone

To demonstrate the applicability of BadTrIP and the advantage of using a model that combines epidemiological and within-sample genetic variation data, we use BadTrIP to infer transmission within the early cases of the 2014 Ebola outbreak in Sierra Leone. We use data published by Gire and colleagues [40] and previously analysed with the SVC method by Worby and colleagues [30]. One of the factors that make this dataset important to this study is the presence of within-host variants shared by multiple hosts, with one genetic variant that was even shared by eleven hosts [40]. While classical approaches based on consensus sequences would struggle to accommodate such data, in particular due to their assumption of strong transmission bottleneck that would not allow the transmission of variants, BadTrIP can accommodate such features, and such shared genomic variants are expected to increase the resolution of our

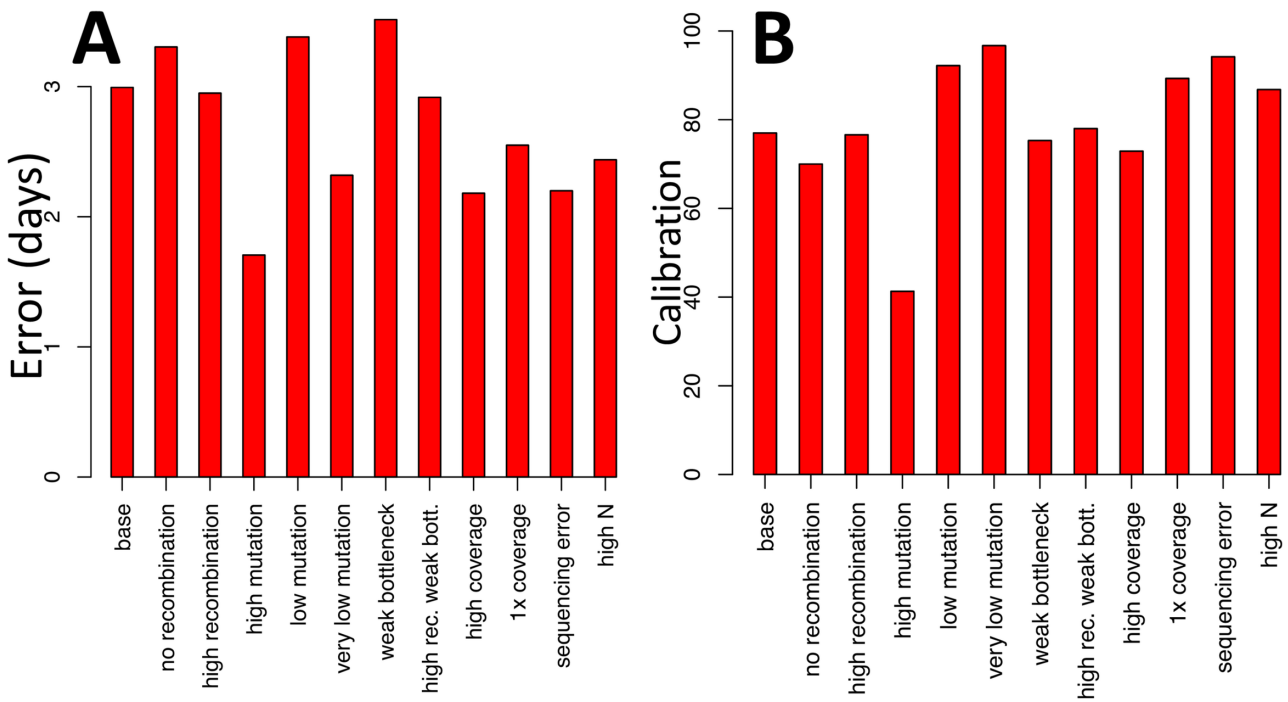

Fig 4. Error and calibration of BadTrIP inferring infection time from simulated data. A) Error (root mean square error) of the inferred median times of infection with BadTrIP. The time unit is days, with a simulated transmission rate of 0.1 per day, and a recovery rate of 0.07 per day (mean duration of infection ≈ 14.3 days). **B**) Calibration (the percentage with which the true time of infection is within the inferred 95% credible interval) for the time of infection with BadTrIP. Simulation scenarios are as in Fig 3.

https://doi.org/10.1371/journal.pcbi.1006117.g004

transmission history inference. We investigate a collection of 62 samples with associated time and location of sampling. As observed by previous researchers, the number of substitutions (and more generally the number of SNPs) within this partial outbreak is very limited, and as such we expect to see a lot of uncertainty in the inference [30]; furthermore, all the samples were collected over a time interval of two months, and we assume transmission from a host to be possible from three weeks prior to three weeks following the sample collection, so the epidemiological data are also not very informative. Indeed, we see that most of the cases are inferred by BadTrIP to have a flat distribution of possible infectors, with highest per-infectee values generally under 30% posterior probability (Fig 5). However, we also see that BadTrIP identifies some pairs of infector-infectee with very high posterior probabilities (Fig B in S1 Text). These pairs not only generally fit with the geographical epidemiological data, with most transmission with posterior probability $> 50\%$ happening within chiefdoms (with two exceptions discussed later), but also with the SVC inference [30]. Of these, transmission from EM119 to G3770 was inferred by Worby and colleagues [30] using consensus sequence genetic distance, while transmission from EM096 to G3679, from G3826 to G3827, from G3820 to G3838, from EM110 to G3809, and from G3729 to G3795 was inferred with the help of shared within-host genetic variants. All highly likely transmission pairs in [30] are also inferred by BadTrIP, but there are some highly likely transmission events inferred by BadTrIP that were not detected by SVC. For example, transmission from G3834 to G3817 is inferred by BadTrIP and is supported by a 3% frequency variant within G3834 that becomes fixed in G3817; however, such a variant fixation, attributable to the transmission dynamics described in Fig 1B, is not informative in the SVC method [30] and was further ignored due to the imposition of a 5% variant frequency

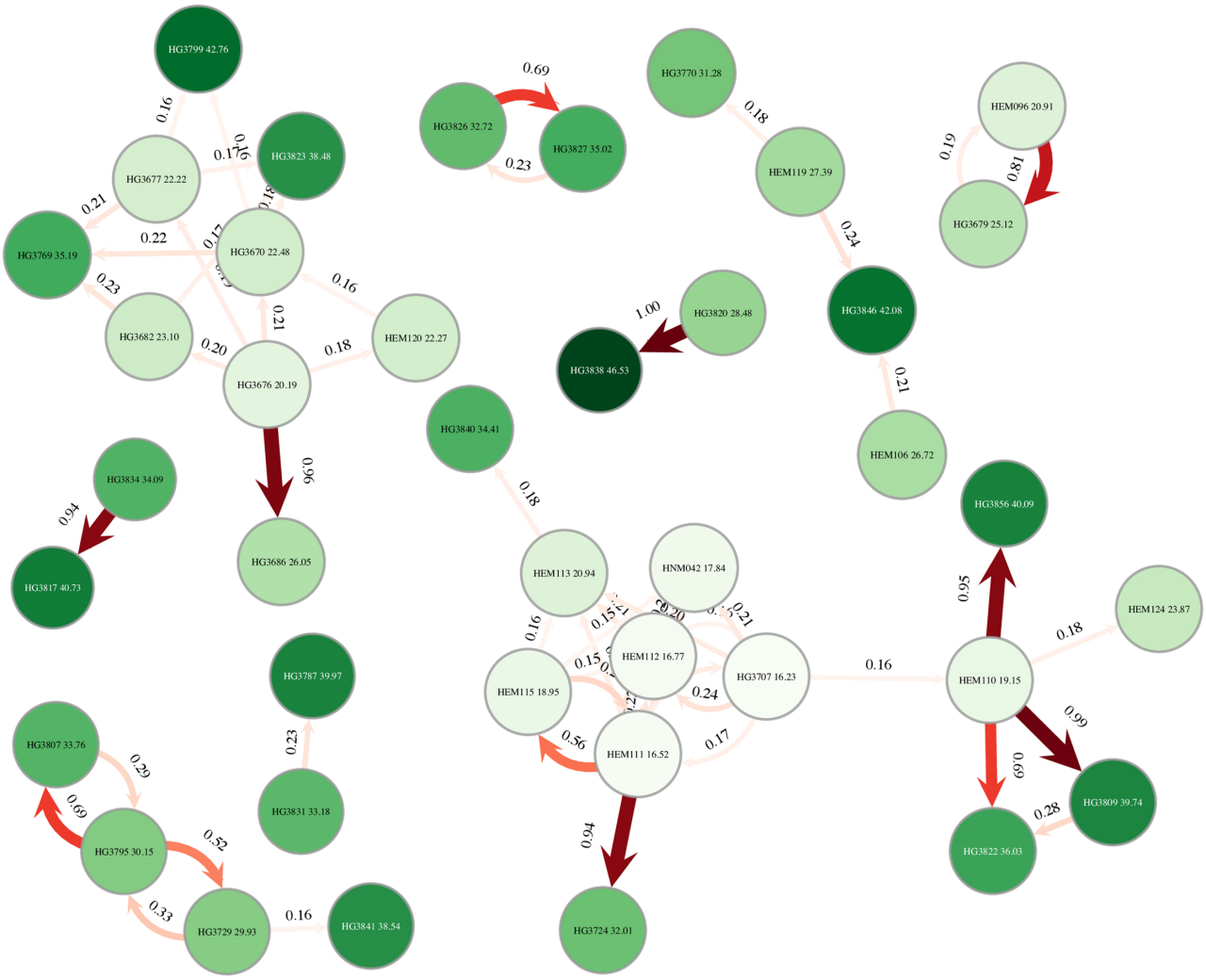

Fig 5. Inference of transmission in the early 2014 Ebola outbreak in Sierra Leone. A) Transmission events with posterior probability higher than 15% as inferred by BadTrIP. Circles represent hosts, while arrows are transmission events between hosts. Only hosts connected to any other host are represented. The numbers next to arrows represent their posterior probability (between 0.0 and 1.0), as does their shade of red (from pale to dark red) and arrow thickness. Numbers within circles represent the inferred (posterior median) time of infection of the respective host, as also does the shade of green (from pale to dark green) of the circle. Time is expressed in days from the date of the first availability of the first host.

https://doi.org/10.1371/journal.pcbi.1006117.g005

threshold that we could avoid thanks to our explicit model of sequence evolution and sequencing error. Other cases similar to the latter are the inferred transmissions from EM110 to G3856, from EM110 to G3822, and from EM111 to G3724.

Cross-chiefdom transmissions inferred by BadTrIP with elevated posterior distributions are from EM110 in the chiefdom of Jawie, district of Kailahun, to G3856 in the chiefdom of Nongowa, district of Kenema; and from G3834 in the chiefdom of Kpeje to G3817 in the chiefdom of Jawie, both in the district of Kailahun. Neither of them had a high probability in [30], but they are both supported by low-frequency variants becoming fixed in the recipient.

Our inference of the sequencing error rate ϵ is extremely low ($2 \cdot 10^{-7} < \epsilon < 7 \cdot 10^{-7}$) consistent with the thorough filtering steps adopted by Gire and colleagues [40] prior to within-host variant calling.

Discussion

Methods to infer transmission histories within outbreaks are important to determine the causes of transmission, and to limit and prevent future outbreaks. Genomic pathogen data from an outbreak reveals in detail the genetic relatedness of pathogens from different cases. Most methods to infer transmission from pathogen genetic data require full haplotypes, but it is often not possible to reconstruct haplotypes due to pathogen recombination and short or inaccurate reads. This leads in many cases to discarding information regarding within-sample genetic diversity, and only use a sample consensus.

In recent years two methods have been proposed to infer transmission from genetic distances between samples and shared within-sample variants [30, 32]. Here we presented BadTrIP, a Bayesian approach to transmission inference that makes use of within-sample variants and allows inference of transmission direction and time. Compared to other similar methods [30, 32], our approach has the advantage of implementing an explicit model of pathogen population evolution, transmission and sequencing, of allowing the inclusion of epidemiological data (sampling times and host exposure times), of not requiring minimum thresholds for within-host variant frequencies, of accounting for sequencing errors, and of being implemented as part of an open source phylogenetic package (BEAST2 [39]). These aspects can result in more applicability, but also, as we have seen in our simulations, in greater accuracy. Compared to existing methods based on the coalescent (e.g. [13, 24, 25, 28]) BadTrIP does not require the reconstruction of haplotypes and consensus sequences, but instead uses data of within-sample genetic variability, therefore having access to important information that can reveal otherwise cryptic transmission events.

Using simulations, we show that our approach achieves higher accuracy and calibration than SVC [30], has more accuracy than the coalescent-based method SCOTTI [13] used on consensus sequences of pathogen population genetic data, and can reliably identify likely transmission histories. The comparison between BadTrIP and SCOTTI is particularly interesting, because it shows us that reducing the genetic data of a within-host pathogen population to a single consensus sequence leads not only to the loss of within-host genetic diversity information, but can also lead to errors by ignoring recombination and weak transmission bottlenecks. Also, using a dataset of the early 2014 Ebola outbreak in Sierra Leone, and making use of information of within-sample variation and an explicit population evolution model, BadTrIP could infer previously unidentified likely transmission events, including transmissions between different geographic locations.

BadTrIP infers transmission from both epidemiological time data and pathogen genetic data. In most circumstances, both types of data are extremely useful, and we see in our simulations that removing genetic data information leads to a loss of $\approx$ 30% accuracy, and similarly the epidemiological data is expected to provide $\approx$ 40% accuracy (the baseline accuracy without any data is expected to be around 10% in our simulations). However, the contribution of the two types of data will be extremely dependent on the particular context at hand. As we showed in our simulations, BadTrIP can account for uninformative genetic data, with which it still provides meaningful inference. Our approach can however also account for uninformative epidemiological data: in the absence of exact dates, the user can specify arbitrarily large exposure intervals, allowing hosts to be infected any time by any host; as with the lack of genetic data, in this case we would also expect a significant drop in the accuracy of our method.

Despite these results, BadTrIP also has limitations, for example its model of genetic linkage. By assuming that all sites are unlinked, our model could be poorly calibrated in cases where there is no within-host recombination but high within-host mutation, causing strong correlations between inherited variants that are not expected in our model. However, we show in our

simulations that our method is robust in a large variety of scenarios, including in the absence of recombination and with reads coming from few pathogen units. Another limitation is that our approach is generally not fast enough to deal with very large datasets, and, at the current stage, application is recommended to outbreaks with fewer than 100 cases. Also, BadTrIP is only applicable to the case where all hosts in the outbreak have been observed. In fact, our current implementation does not allow to infer the number of non-observed hosts (hosts for which there is no sample or epidemiological data). However, BadTrIP does allow to model non-sampled hosts with epidemiological data, or a fixed number of non-observed hosts (such hosts could be given uninformative epidemiological data, such as exposure intervals without ends). The assumption that all cases are observed or sampled is very common among transmission inference methods [11, 14–20, 23–26, 28], but it limits their applicability. Extending our method to infer the presence of possible non-sampled and non-observed intermediate hosts would be relatively straightforward and would increase the method's applicability, but it would also lead to a significant increase in the statistical complexity and computational demand (but see [13, 27]).

Another scenario that is not accounted for in our model is multiple infections of the same host (one host being infected by multiple sources, or by the same source multiple times). This scenario can be relatively frequent in many viruses, for example HIV [46], but it is very hard to model in our context as it would require the use of a population network (see e.g. [47]) instead of a population phylogeny, which would make likelihood calculation more computationally demanding. Another similarly looking and equally concerning problem is sample contamination. We recommend sequencing data to be tested for possible contaminations and multiple infections using methods such as PHYLOSCANNER [48] prior to being investigated with BadTRiP. In our Ebola dataset we found no obvious pattern of mixed infection or contamination (like an excess of similar frequency SNPs in one sample). However, none of these approaches would detect multiple infections from closely related cases. BadTrIP uses a very simple model of sequencing error, only accounting for the two most common nucleotides at a given position and sample. This sequencing error model would probably have sub-optimal performance when sequencing error rate is high (e.g. with Nanopore sequencing technologies) and coverage is high or mutation rate is elevated. In these circumstances, a more realistic and computationally demanding model of sequencing error might be preferable. Similarly, our model of evolution only allows 2 alleles for one genome position in one host at one time. If mutation rates are so high that more than 2 alleles are frequently present simultaneously in the same host, time and position, then our model could have sub-optimal performance. However, our approach can still account for the more common scenario where a site has more than 2 alleles but not all in the same host: for example if at a certain position host 1 has a fixed A, host 2 has a polymorphism with A and C, and host 3 has a polymorphism with C and G.

BadTrIP does not account for selective pressure, which could sometimes cause errors, for example by creating homoplasies due to the same mutation appearing multiple times in different hosts, or by the same polymorphism being maintained by balancing selection. However, our approach weighs information from both fixed substitutions and polymorphic variants, so the same mutation appearing in different genetic backgrounds will not be as nearly as misleading as for the SVC method (which gives much more weight to shared variants than to genetic distances). We assume that within-host population sizes are constant after an initial expansion. Size fluctuations in all hosts are unlikely to cause problems, as the PoMo drift rate would in this case represent the average drift rate in hosts. On the other hand, if fluctuations only happen in certain hosts, so that different hosts have different average drift rates, it might have averse effects on the estimate of infection times.

As our model is implemented in BEAST2, it is possible to specify a broad range of models of genomic variation in substitution rates which could at least partly account for the effects of selection. An additional feature that could be added to BadTrIP is indel evolution. For example, by assuming an infinite sites mutation model, indel data could be reduced to 0–1 states, and a PoMo matrix with two alleles instead of 4 nucleotides could be used. This approach could be useful to complement SNP data, but would only work at relatively low indel rates.

Finally, it is possible that errors in the bioinformatic processing of reads, for example mapping errors, cause the identification of the same spurious genetic variants in multiple hosts. We therefore encourage the investigations of genetic variants shared by many hosts to assess their biological plausibility. In the future we will work to solve some of the limitations of BadTrIP, in particular to reduce its computational demand and to model non-sampled non-observed hosts.

In conclusion, we have presented a new method that addresses the urgent need for software to efficiently and accurately analyse genomic and epidemiological data, in particular taking advantage of within-sample genetic variants to identify transmission pairs and reconstruct direction and time of infection. BadTrIP can be used in a broad range of outbreaks, and will be important for devising effective strategies to fight the spread of infectious disease.

Materials and methods

Model of transmission

We model each host as a deme $d \in D$ that can be colonised by a pathogen population, with total number of hosts-demes being n_D. Each deme d is associated with an exposure interval limited by an introduction time $i_d \in (-\infty, +\infty]$ and a removal time $r_d \in [-\infty, +\infty)$, with $r_d < i_d$ (we consider time backward as typical in coalescent theory); the host only contributes to the outbreak within this interval, which is determined by the epidemiological data. In the least informative scenario where no information on host d exposure is provided, it is assumed that d is exposed for the whole outbreak ($i_d = +\infty$ and $r_d = -\infty$). We will denote as $\boldsymbol{X}$ the collection of exposure times.

Each host-deme starts off as non-colonised and is colonised (infected) at some time t_d between i_d and the time that the first sample is collected from d (if no sample is collected from d, then we require only $t_d > r_d$). Also, unless d is the first host to be infected in the outbreak, d is infected by another host in the outbreak $I_d \neq d$, such that $r_{I_d} < t_d < t_{I_d}$, that is, d is infected after I_d is infected, but before I_d reaches its removal time. If d is indeed the first case of the outbreak, then I_d is assigned the $\emptyset$ (we assume $\emptyset \notin D$). We assume for simplicity that transmission between any pair of hosts and at any time is equally likely, as long as it is consistent with the epidemiological data. A transmission event of host d at time t_d is inconsistent with the epidemiological data if t_d is outside the exposure interval of d or its infector I_d, or if d is sampled, infects another host, before t_d. Given the epidemiological data, some infector-infectee pairs are a priori more likely than others, depending on the length of time that a transmission between them is allowed.

Each host is also provided with a (possibly empty) set of samples, S_d. Each sample s consists of a sampling time t_s and genetic data G_s. Each sample s in S_d has to be collected after d is infected ($t_s < t_d$) and before d is removed ($t_s > r_d$). Assuming that the genome is L bases long, then the genetic data G_s of every sample s has to be in the form of a list of L quadruples, with for example the quadruple for genome position i being $G_{si} = (a_i, c_i, g_i, t_i)$, the four positive natural values being the numbers of A's, C's, G's and T's observed at position i in the sample. If there is no read mapping to position i in sample s, then its quadruple is simply $G_{si} = (0, 0, 0, 0)$. We denote the set of all sequencing data as G.

All hosts share a common parameter B (with real positive values) describing the intensity of the transmission bottlenecks associated with transmission events. Generally, the value of B can be inferred jointly with other model parameters, however its interpretation in terms of the size of the transmission inoculum is not straightforward. T denotes the transmission-population tree consisting of all sampling times, all infection times and all infectors of each host, and $\boldsymbol{\mu}$ denotes the pathogen evolution model (described below). An example of tree T and of model parameters is given in Fig C in S1 Text.

We aim to sample from the following joint posterior distribution with a Monte Carlo Markov Chain approach:

$$P(T, \boldsymbol{\mu}, B|G, \boldsymbol{X}) \propto P(G|T, \boldsymbol{\mu}, B)P(T|\boldsymbol{X})P(\boldsymbol{\mu})P(B). \quad (1)$$

$P(\boldsymbol{\mu})$ and $P(B)$ are the prior probabilities for respectively the substitution model and the bottleneck size, which can be chosen arbitrarily by the user. We ignore the prior for the transmission tree $P(T|\boldsymbol{X})$ as in [13]. $P(G|T, \boldsymbol{\mu}, B)$ is the likelihood of the sequences given the genealogy and substitution model, and is calculated as described below, using an adaptation of [36–38] to transmission trees. So once we calculate the likelihood $P(G|T, \boldsymbol{\mu}, B)$, we can use Eq 1 with an MCMC to infer a posterior distribution of infection times, infectors, bottleneck size and substitution model parameters.

Model of pathogen evolution

Here, we make use of a phylogenetic model for population evolution, PoMo [36–38], to model mutation and drift in the within-host pathogen populations; also, we extend the model to include transmission bottlenecks and sequencing errors. Sequence evolution is usually modelled along phylogenetic trees, which can differ from the transmission tree [13]. However, PoMo describes evolution along species (or population) trees, and the population tree of a pathogen within an outbreak corresponds to the transmission tree T described in the previous section. If we consider the pathogen community within a host d as a population, we see that this population exists from time of infection t_d, when it originates from a split with the population of its infector I_d. So, transmission events corresponds to timed splits in the population tree, similar to the bifurcations of a species tree. However, one difference is that the split is asymmetrical, as we assume that the pathogen population size is not affected at t_d in I_d, but at the start of the branch leading to d it undergoes a bottleneck of intensity B. All events in the tree are timed in real time (e.g., days) with some values fixed (for samples) and some values inferred in the MCMC (infection times).

We use a procedure very similar to the Felsenstein pruning algorithm [49] to calculate the likelihood of the genetic data over the tree. First of all, the substitution process along the branches of the transmission-population tree is not a simple DNA substitution process, but is similar to a 4-allelic Moran model [41] with mutation. We assume we have a continuous-time Markov process along each branch of the tree, where the state space is not made by the four nucleotides, as is typical, but by all 1- and 2-allelic states possible for a population of N units. Typical values of N that we use here are 15 or 25, that is, we describe evolution of a large within-host pathogen population (possibly with billions of units) with a small virtual within-host population of N units. Such an approximation generally leads to reasonably good results as long as we rescale the mutation rates between the real and the virtual population [36–38]. N here is not estimated, but is fixed by the user. Lower values of N are expected to reduce the computational demand of the method, but can result in lower accuracy. The states of our Markov process always include the four fixed states, where only one nucleotide is present in the population. In addition, they also include six groups of polymorphic states, where two

nucleotides are present in the virtual population at the same site at the same time. Each group corresponds to one of the six unordered pairs of nucleotides ($\{A, C\}$, $\{A, G\}$, $\{A, T\}$, $\{C, G\}$, $\{C, T\}$, $\{G, T\}$) and contains $N-1$ states: if the two nucleotides present in the population are n_1 and n_2, then such $N-1$ states are the ones in which the population contains i times nucleotide n_1 and $N-i$ times nucleotide n_2, for $0 < i < N$. So in total our state space is of size $4 + 6(N-1)$. Our substitution rate matrix is sparse, in that we only allow one unit in the virtual population to change at the time. So, from a fixed state with nucleotide n_1, a instantaneous move is only possible to one of the three states with $N-1$ times nucleotide n_1 and one time any other nucleotide n_2 different from n_1. Such moves correspond to mutation events, and we represent their rates as μ_{n_1,n_2}. Instead, if we are already in a polymorphic state with i times nucleotide n_1 and $N-i$ times nucleotide n_2, we only allow nucleotide counts to instantaneously change by one, so an instantaneous move is only possible to the state with $i+1$ times nucleotide n_1 and $N-i-1$ times nucleotide n_2, or to state $i-1$ times nucleotide n_1 and $N+1-i$ times nucleotide n_2 (one of these two latter states might be a fixed state). The instantaneous rate at which such changes happen is $\frac{i(N-i)}{N^2}R$ which corresponds to the rate of genetic drift; here R scales the rate of drift in the virtual population in units of real time; the rate of drift in the virtual population also depends on N, and it represents the rate of drift in a real pathogen population, which in turn depends on the pathogen effective population size, the pathogen generation time, and the time unit. All other non-diagonal substitution rates are set to 0. All these states and rates constitute the substitution process $\boldsymbol{E}$. The rate matrix is further described in Fig D in S1 Text. Our model only allows 2 alleles to be present in one host at one time at one position. This can be unrealistic where mutation rates are extremely high, or selection favours several variants at the same site.

The likelihood of T is calculated starting from the hosts in the outbreaks who don't infect others (the leaves of the transmission tree). For such leaves, the likelihood is first calculated from the latest sample (if no sample is present, then the likelihood of such leaf at time of their transmission is 1 for every state). Given any state of our substitution process with nucleotides n_1 and n_2 with respectively abundances i and $N-i$ in the virtual population (here for generality i can also be 0), given a sample and site at which the nucleotides with the highest coverage are x_1 with coverage c_1, and x_2 with coverage c_2 (we ignore the nucleotides with lower counts for numerical stability, and in case of a tie random nucleotides are selected from the tying ones), then the likelihood of this state at this sample and site is approximated as:

$$\begin{aligned} &P(c_1, x_1, c_2, x_2 | i, n_1, N-i, n_2, \epsilon) = \\ &= (I_{x_1=n_1}(\frac{i(1-\epsilon)}{N} + \frac{(N-i)\epsilon}{3N}) + I_{x_1=n_2}(\frac{(N-i)(1-\epsilon)}{N} + \frac{i\epsilon}{3N}) + I_{x_1\neq n_1, x_1\neq n_2} * \frac{\epsilon}{3})^{c_1} \cdot \\ &\cdot (I_{x_2=n_1}(\frac{i(1-\epsilon)}{N} + \frac{(N-i)\epsilon}{3N}) + I_{x_2=n_2}(\frac{(N-i)(1-\epsilon)}{N} + \frac{i\epsilon}{3N}) + I_{x_2\neq n_1, x_2\neq n_2} * \frac{\epsilon}{3})^{c_2} \cdot \\ &\cdot \begin{pmatrix} c_1 + c_2 \\ c_1 \end{pmatrix} \end{aligned} \tag{2}$$

Where ϵ is a parameter describing the sequencing error rate. Here, due to sequencing errors and to random sampling of reads from the pathogen population, the observed alleles c_1 and c_2 are allowed be different from the alleles n_1 and n_2 in the virtual population. We assume that each read has the same probability to represent any of the individuals in the virtual population, and that there is a probability ϵ that the considered position of the read is a sequencing error (in which case any of the three wrong nucleotides is equally likely to be on the read).

$(\frac{i(1-\epsilon)}{N} + \frac{(N-i)\epsilon}{3N})$ is the probability to see a n_1 nucleotide: the first part is the probability that the read comes from an individual in the virtual population with nucleotide n_1 at the given position and that no sequencing error happened; the right end part represents the probability that the virtual individual had a different nucleotide but there was a sequencing error. ϵ can be estimated with the other model parameters as we do with the real data and with the simulations including sequencing error. For all other simulations we set $\epsilon = 0$. This sequencing model assumes that there are at most 2 alleles in the reads data for one sample at one position. If more than 2 alleles are observed, then only the counts from the 2 most common alleles are retained.

Along branches of T, the likelihood is updated using the matrix exponential of $\boldsymbol{E}$. At bifurcations (corresponding either to internal samples or transmission events) the likelihood is also updated according to the classical pruning algorithm, but at transmission events an extra step is added. A new drift-only substitution matrix $\boldsymbol{E}_D$ is defined by setting the mutation rates in $\boldsymbol{E}$ to 0. Then, we describe a bottleneck as a branch of length B along which the population evolves under drift alone, that is, under $\boldsymbol{E}_D$. The length B does not count toward the branch lengths in real time, so that changing the intensity of the bottleneck does not affect the timing of the events in T. Under this model, a more intense bottleneck, corresponding to a small transmission inoculum, will be represented by a longer bottleneck branch, so a larger B. If we have a transmission event from I_d to d at time t_d, we first calculate the likelihood within population I_d up to right before time t_d (likelihoods are updated backward in time), then within population d up to right before time t_d, then we update the likelihood within d using the bottleneck branch, and finally we multiply the two likelihoods in d and I_d to obtain the likelihood in I_d right after t_d (again backward in time). This backward-in-time likelihood update process is terminated after the transmission event of the index case, and before its bottleneck we assume state equilibrium frequencies. We now describe an example of likelihood calculation in Fig E in S1 Text.

MCMC proposals

We use typical BEAST2 scalar proposals for B, ϵ and $\boldsymbol{E}$, which, given a constant s and a random uniform real number $0 < u < 1$, propose to scale the given parameter by a factor of $s + (1/s - s)u$; the reciprocal of this factor is the Hastings ratio of the proposal. We also define below five new operators (proposals) for updating our transmission-population tree.

- The first operator picks a random host d uniformly, then picks its new transmission time t_d uniformly within the time interval allowed by i_d, r_d, the first sampling time of d (if any is present), the first time d infects another host (if any), and the exposure interval of the infector of d, I_d. This operator does not modify any other parameter, not even I_d. The Hastings ratio is 1.
- The second operator picks a random non-index case d and, without modifying its infection time t_d, picks a random new infector I_d among the ones compatible with infection time t_d. The Hastings ratio is 1.
- The third operator is similar to the second, but first picks a new infection time t'_d for d among those allowed by i_d, r_d, first sample time of d and the first time d infects another host (but not based on the current infector I_d), and then picks a new infector I'_d of d uniformly among those compatible with t'_d (if any is present, otherwise the proposal is rejected). The Hastings ratio is calculated taking the number of possible infectors of d compatible with the

new infection time t'_d, and dividing it by the number of possible infectors of d compatible with the old infection time t_d.

- The fourth operator swaps infector-infectee. First, a random non-tip host (a host with some infectees) d is uniformly chosen; we call its first infectee c. Given infection times t_c and t_d, if the swap is legal (d has no samples collected before t_c, and t_d is within the exposure interval of c) then I_d (possibly $\emptyset$) becomes the infector I'_c of c at time $t'_c = t_d$, and c becomes the infector I'_d of d at time $t'_d = t_c$. The Hastings ratio is 1.
- The last operator picks a random case d uniformly, selects a new infection time t'_d as in operator three, then picks a random new infector I'_d uniformly within the set of infectors compatible with t'_d and within the epidemiological upper neighbourhood of d (the grandparent I_{I_d}, its infectees, and the infectees of I_d different from d); if no compatible infector is found, the move is rejected. The Hastings ratio is calculated like in operator three, but counting only compatible infectors within epidemiological upper neighbourhoods.

We will now give a very informal intuition of why the above proposals make an irreducible MCMC. We will focus on the transmission history, and not on B, ϵ or $\boldsymbol{E}$, but the extension is trivial. We will discuss intuitively how it is possible to move from any given tree T to a specific tree $\tilde{T}$ (the tree we use as a starting point of the MCMC). As proposals are reversible, this is sufficient to have irreducibility. $\tilde{T}$ is defined as the tree where the host with the earliest introduction time is the index case; each non-index host d in $\tilde{T}$ is infected by the host I_d with the earliest introduction time i_{I_d} among those with an exposure overlap to d; in $\tilde{T}$ infection time t_d of any host d is set to i_d (for a more formal proof an infinitesimal interval after i_d might be considered). Starting from T, we first approach $\tilde{T}$ by moving host h, the index case in $\tilde{T}$, up from its starting position in T by using repeatedly operator four. At each step, before applying operator four, we use operator one to move t_h up to make sure it is the first infectee of its infector, and that it is infected before the first sample of its infector is collected. Repeating these two steps long enough, h is guaranteed to become the root, at which point we can apply operator one to make sure its infection time is the same as in $\tilde{T}$. We then proceed to apply a similar strategy iteratively on all other hosts in order based on their introduction time (from earlier to latest). We stop when we obtain $\tilde{T}$.

Simulations of pathogen evolution

To test the accuracy of our new method BadTrIP in inferring transmission events, and to compare it with the SVC method [30], we simulated pathogen evolution within outbreaks and sample sequencing, and we used different methods to reconstruct the transmission history from sequencing and epidemiological data. To simulate pathogen evolution, first we simulated an outbreak using SEEDY [42] with a host population of 15 hosts and an infection rate of 0.1 per day, a recovery rate 0.07 per day, and conditionally accepting only outbreaks that achieve a minimum total of 10 infected cases. Given these parameters, SEEDY will start at time 0 with one infected individual in the community of 15 hosts. Each day every infected host has a 0.1 chance of infecting any other host, and a 0.07 chance of recovering (recovered hosts are no more infectious or infectable). If the outbreak runs out of infected hosts before a total of 10 hosts are infected, the simulation is repeated. We then took the outbreak simulated by SEEDY and translated the transmission history into a population history, assuming a within-host pathogen population size of 1000 and using fastsimcoal2 [43] to simulate pathogen coalescent, recombination and mutation with scenario-dependent parameters. fastSimCoal2 is an approximate coalescent simulator implementing the sequential Markov coalescent model [50, 51]

with cross-over recombination. This model describes viral recombination more appropriately than bacterial recombination, for which a coalescent simulation software modeling gene conversion is preferable [52, 53]. The use of a coalescent simulator with recombination is also the main difference with the simulations made by [30], where within-host recombination was not allowed. Within each infection we assume that the population size is constantly 1000 individuals, but at the time of transmission we assume an instantaneous population bottleneck (founding population size either 1 or 5 individuals depending on the scenario). At the time of a transmission (simulated by SEEDY) the whole infectee population is, backward in time, merged with the infector population. We observed that some times, in particular at high recombination rates, fastSimCoal can crash: if this happens we simply repeat the fastSimCoal2 simulation with a different seed. Throughout all simulations each host was sampled exactly once.

We define a basic group of simulations (called "base"), and nine variants, in each of which one or two aspects of the base group of simulations is modified. In "base" we simulated about 300–500 SNPs (counting also variants present at very low frequency in just one host) or 45 substitutions per outbreak (which might be typical for HIV but high for many other pathogens), recombination rate 10 times higher than the mutation rate, complete bottlenecks (no transmission of within-host genetic variants), homogeneous read coverage of 40x, no sequencing error, PoMo virtual population size of 15, all equal mutation rates, and genome size of 5 kb. The eleven variant settings are:

- **no recombination**—the recombination rate is set to 0.
- **high recombination**—the recombination rate is increased 10-fold.
- **high mutation**—the mutation rate is 10-fold higher resulting in 2000–3000 SNPs and about 385 substitutions per outbreak.
- **low mutation**—the mutation rate is 10-fold lower resulting in 30–50 SNPs and about 4–5 substitutions per outbreak.
- **very low mutation**—the mutation rate is 1000-fold lower, resulting in 0–1 SNPs and 0 substitutions per outbreak.
- **weak bottleneck**—at transmission, 5 pathogen particles from the infector colonise the infected host, instead of just 1.
- **high recombination and weak bottleneck**—the recombination rate is 10-fold higher and the founding population at transmission is made of 5 pathogen particles.
- **high coverage**—read coverage is higher (100x instead of 40x).
- **1x coverage**—read coverage is extremely low (1x instead of 40x).
- **sequencing error**—read coverage is lower (20x instead of 40x), genome size is reduced (1kb instead of 5kb) and read bases are randomly modified to simulate sequencing error (0.2% of bases in reads are wrong).
- **high N**—the PoMo virtual population size is 25 instead of 15 (this only affects the BADTRIP inference and not the simulation itself).

We ran 10 replicates for all scenarios, and 20 for "base", "weak bottleneck" and "no recombination" (some scenarios are more computationally demanding due to the effect of recombination on coalescent simulations and of genetic diversity on transmission inference). For each repeat in each scenario we ran a completely different simulation with different seeds resulting

in different transmission and coalescent histories, even when outbreak or coalescent parameters do not change across scenarios. We ran the BadTrIP MCMC for $5 \cdot 10^5$ steps for each replicate, sampled from the posterior every 100 steps and with a 20% burn-in. We specified in BadTrIP the true simulated sampling time and removal time of each host, while we specified as introduction time of each host its infection time minus one quarter of the mean duration of infection (so that the true infection time is contained within the exposure time of the host). For SCOTTI we used the same epidemiological data and options as for BadTrIP, except that we ran the MCMC for $2 \cdot 10^6$ steps for each replicate. We did not allow unobserved cases in SCOTTI. We measured accuracy as the frequency with which the correct transmission source of each host is inferred by a method to be the most likely a posteriori. We also measured calibration as how often the correct transmission source is the the 95% posterior credible set (the minimum set of sources with cumulative probability $\geq$ 95% such that all sources in the set have higher posterior probability than all sources outside of it).

We also used the SVC method [30] to infer transmission from simulated data. This method consists of selecting, for each host *d*, the set of possible infectors as those cases with most shared variants with *d*, or, if *d* does not share variants with other hosts, the cases with the smallest consensus genetic distance from *d*. If a single possible infector is found, it is assigned 100% posterior probability, otherwise if multiple possible infectors are found they are assigned the same posterior probability. For example, if 4 cases all have 2 shared genetic variants with *d*, and all other cases have fewer than 2, than each of those 4 cases is assigned a posterior probability of 25% of infecting *d*. This is very different from BadTrIP, which always weighs the information from shared variants, genetic distances, and epidemiological data simultaneously from all cases. So, the 4 cases sharing 2 genetic variants with *d* can have very different posterior probabilities in BadTrIP of being infectors of *d*, depending on the other data. For example, if one of these 4 cases has very high genetic distance from *d*, or epidemiological data incompatible with a transmission to *d*, BadTrIP would infer very low (or null) probability of it being the infector of *d*.

The 2014 Sierra Leone Ebola dataset

We use sequencing and epidemiological data published by Gire and colleagues [40] and analysed by Worby and colleagues [30]. In particular, we use information from sampling dates, nucleotide frequencies and sequencing coverage. We specify the introduction date (removal date) of each host as its sampling date minus (plus) 21 days. This means that we allow each host to be infected at most 21 days before it being sampled, and to infect others at most 21 days after being sampled. We ran the BadTrIP MCMC until an effective sample size of 1000 was reached for each parameter and for the posterior probability (requiring $\approx$ 3.5 million MCMC steps). to reduce the computational time required we subsampled the reads from each sample to obtain a per-base coverage of at most 100.

Software availability

BadTrIP is distributed as an open source package for the Bayesian phylogenetic software BEAST2 [39]. It can be downloaded from https://bitbucket.org/nicofmay/badtrip/ or via the BEAUti interface [54] of BEAST2.

Supporting information

S1 Text. The supplementary text containing supplementary figures.
(PDF)

S1 Data. Contains the xml script to replicate our Ebola analysis with BadTrIP, an R script to run the SVC approach, and a python script to replicate our simulations.
(ZIP)

Author Contributions

Conceptualization: Nicola De Maio.

Data curation: Nicola De Maio, Colin J. Worby.

Formal analysis: Nicola De Maio.

Investigation: Nicola De Maio.

Methodology: Nicola De Maio.

Project administration: Nicola De Maio.

Resources: Nicola De Maio, Colin J. Worby.

Software: Nicola De Maio.

Supervision: Daniel J. Wilson, Nicole Stoesser.

Validation: Nicola De Maio.

Visualization: Nicola De Maio.

Writing – original draft: Nicola De Maio.

Writing – review & editing: Nicola De Maio, Colin J. Worby, Daniel J. Wilson, Nicole Stoesser.

References

1. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics. 2012; 13(9):601–612. https://doi.org/10.1038/nrg3226 PMID: 22868263
2. Wilson DJ. Insights from genomics into bacterial pathogen populations. PLoS Pathog. 2012; 8(9): e1002874. https://doi.org/10.1371/journal.ppat.1002874 PMID: 22969423
3. Köser CU, Ellington MJ, Cartwright E, Gillespie SH, Brown NM, Farrington M, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012; 8(8):e1002824. https://doi.org/10.1371/journal.ppat.1002824 PMID: 22876174
4. Le V, Diep BA. Selected insights from application of whole-genome sequencing for outbreak investigations. Current opinion in critical care. 2013; 19(5):432–439. https://doi.org/10.1097/MCC.0b013e3283636b8c PMID: 23856896
5. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. New England Journal of Medicine. 2013; 369 (13):1195–1205. https://doi.org/10.1056/NEJMoa1216064 PMID: 24066741
6. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet infectious diseases. 2013; 13(2):137–146. https://doi.org/10.1016/S1473-3099(12)70277-3 PMID: 23158499
7. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. The Lancet Respiratory Medicine. 2014; 2(4):285–292. https://doi.org/10.1016/S2213-2600(14)70027-X PMID: 24717625
8. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences. 1996; 93(20):10864–10869. https://doi.org/10.1073/pnas.93.20.10864
9. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327(5964):469–474. https://doi.org/10.1126/science.1182395 PMID: 20093474

10. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics. 2009; 10(8):540–550. https://doi.org/10.1038/nrg2583 PMID: 19564871
11. 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
12. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T. Timing and order of transmission events is not directly reflected in a pathogen phylogeny. Molecular biology and evolution. 2014; 31(9):2472–2482. https://doi.org/10.1093/molbev/msu179 PMID: 24874208
13. De Maio N, Wu CH, Wilson DJ. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology. 2016; 12(9):e1005130. https://doi.org/10.1371/journal.pcbi.1005130 PMID: 27681228
14. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proceedings of the Royal Society of London B: Biological Sciences. 2008; 275(1637):887–895. https://doi.org/10.1098/rspb.2007.1442
15. Aldrin M, Lyngstad T, Kristoffersen A, Storvik B, Borgan Ø, Jansen P. Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. Journal of The Royal Society Interface. 2011; 8(62):1346–1356. https://doi.org/10.1098/rsif.2010.0737
16. Jombart T, Eggo R, Dodd P, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity. 2011; 106(2):383–390. https://doi.org/10.1038/hdy.2010.78 PMID: 20551981
17. Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR Jr, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nature genetics. 2011; 43(12):1275–1280. https://doi.org/10.1038/ng.997 PMID: 22081229
18. Morelli MJ, Thebaud G, Chadoeuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol. 2012 11; 8(11):e1002768. https://doi.org/10.1371/journal.pcbi.1002768 PMID: 23166481
19. Ypma R, Bataille A, Stegeman A, Koch G, Wallinga J, Van Ballegooijen W. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings of the Royal Society of London B: Biological Sciences. 2012; 279(1728):444–450. https://doi.org/10.1098/rspb.2011.0913
20. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013; 195(3):1055–1062. https://doi.org/10.1534/genetics.113.154856 PMID: 24037268
21. Volz EM, Frost SD. Inferring the source of transmission with phylogenetic data. PLoS Comput Biol. 2013;. https://doi.org/10.1371/journal.pcbi.1003397
22. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology. 2014; 10(1). https://doi.org/10.1371/journal.pcbi.1003457 PMID: 24465202
23. 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. Proceedings of the Royal Society of London B: Biological Sciences. 2014; 281(1782):20133251. https://doi.org/10.1098/rspb.2013.3251
24. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Molecular biology and evolution. 2014; 31(7):1869–1879. https://doi.org/10.1093/molbev/msu121 PMID: 24714079
25. 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(12):e1004613. https://doi.org/10.1371/journal.pcbi.1004613 PMID: 26717515
26. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proceedings of the National Academy of Sciences. 2016; p. 201522930.
27. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular biology and evolution. 2017; 34(4):997–1007. https://doi.org/10.1093/molbev/msw275 PMID: 28100788
28. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS computational biology. 2017; 13(5):e1005495. https://doi.org/10.1371/journal.pcbi.1005495 PMID: 28545083

29. Bull RA, Eden JS, Luciani F, McElroy K, Rawlinson WD, White PA. Contribution of intra-and interhost dynamics to norovirus evolution. Journal of virology. 2012; 86(6):3219–3229. https://doi.org/10.1128/JVI.06712-11 PMID: 22205753

30. Worby CJ, Lipsitch M, Hanage WP. Shared genomic variants: identification of transmission routes using pathogen deep sequence data. American Journal of Epidemiology. 2015;.

31. Leonard AS, Weissman D, Greenbaum B, Ghedin E, Koelle K. Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. Journal of Virology. 2017; p.

32. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 2017; p.

33. Shriner D, Rodrigo AG, Nickle DC, Mullins JI. Pervasive genomic recombination of HIV-1 in vivo. Genetics. 2004; 167(4):1573–1583. https://doi.org/10.1534/genetics.103.023382 PMID: 15342499

34. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990; 18:6097–6100. https://doi.org/10.1093/nar/18.20.6097 PMID: 2172928

35. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14(6):1188–1190. https://doi.org/10.1101/gr.849004 PMID: 15173120

36. De Maio N, Schlötterer C, Kosiol C. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Molecular biology and evolution. 2013; 30(10):2249–2262. https://doi.org/10.1093/molbev/mst131 PMID: 23906727

37. De Maio N, Schrempf D, Kosiol C. PoMo: an allele frequency-based approach for species tree estimation. Systematic biology. 2015; 64(6):1018–1031. https://doi.org/10.1093/sysbio/syv048 PMID: 26209413

38. Schrempf D, Minh BQ, De Maio N, von Haeseler A, Kosiol C. Reversible polymorphism-aware phylogenetic models and their application to tree inference. Journal of theoretical biology. 2016; 407:362–370. https://doi.org/10.1016/j.jtbi.2016.07.042 PMID: 27480613

39. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014; 10(4):e1003537. https://doi.org/10.1371/journal.pcbi.1003537 PMID: 24722319

40. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014; 345(6202):1369–1372. https://doi.org/10.1126/science.1259657 PMID: 25214632

41. Moran PAP; Cambridge Univ Press. Random processes in genetics. Math Proc Cambridge. 1958; 54:60–71. https://doi.org/10.1017/S0305004100033193

42. Worby CJ, Read TD. 'SEEDY'(Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. PloS one. 2015; 10(6):e0129745. https://doi.org/10.1371/journal.pone.0129745 PMID: 26075402

43. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS genetics. 2013; 9(10):e1003905. https://doi.org/10.1371/journal.pgen.1003905 PMID: 24204310

44. Dawid AP. The well-calibrated Bayesian. Journal of the American Statistical Association. 1982; 77 (379):605–610. https://doi.org/10.1080/01621459.1982.10477856

45. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012; 30(5):434–439. https://doi.org/10.1038/nbt.2198 PMID: 22522955

46. Romero-Severson EO, Bulla I, Hengartner N, Bártolo I, Abecasis A, Azevedo-Pereira JM, et al. Donor-Recipient identification in para-and poly-phyletic trees under alternative HIV-1 transmission hypotheses using approximate Bayesian computation. Genetics. 2017; 207(3):1089–1101. https://doi.org/10.1534/genetics.117.300284 PMID: 28912340

47. Yu Y, Than C, Degnan JH, Nakhleh L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology. 2011; 60(2):138–149. PMID: 21248369

48. Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Analysing Within-and Between-Host Pathogen Genetic Diversity to Identify Transmission, Multiple Infection, Recombination and Contamination. bioRxiv. 2017; p. 157768.

49. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17(6):368–376. https://doi.org/10.1007/BF01734359 PMID: 7288891

50. McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2005; 360(1459):1387–1393. https://doi.org/10.1098/rstb.2005.1673 PMID: 16048782

51. Marjoram P, Wall JD. Fast "coalescent" simulation. BMC genetics. 2006; 7(1):16. https://doi.org/10.1186/1471-2156-7-16 PMID: 16539698

52. Brown T, Didelot X, Wilson DJ, De Maio N. SimBac: simulation of whole bacterial genomes with homologous recombination. Microbial genomics. 2016; 2(1). https://doi.org/10.1099/mgen.0.000044 PMID: 27713837

53. De Maio N, Wilson DJ. The bacterial sequential Markov coalescent. Genetics. 2017; 206(1):333–343. https://doi.org/10.1534/genetics.116.198796 PMID: 28258183

54. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969–1973. https://doi.org/10.1093/molbev/mss075 PMID: 22367748