PLOS | PATHOGENS

RESEARCH ARTICLE

When are pathogen genome sequences informative of transmission events?

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

1 MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom, **2** Centre for Preventive Medicine, Department of Epidemiology and Disease Surveillance, Animal Health Trust, Suffolk, United Kingdom

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

Citation: Campbell F, Strang C, Ferguson N, Cori A, Jombart T (2018) When are pathogen genome sequences informative of transmission events? PLoS Pathog 14(2): e1006885. https://doi.org/10.1371/journal.ppat.1006885

Editor: Colin Parrish, Cornell University, UNITED STATES

Received: June 7, 2017

Accepted: January 18, 2018

Published: February 8, 2018

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

Funding: FC is funded by the Wellcome Trust (https://wellcome.ac.uk). NF is funded by the UK Medical Research Council (https://www.mrc.ac.uk); UK National Institute for Health Research under the Health Protection 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/

Abstract

Recent years have seen the development of numerous methodologies for reconstructing transmission trees in infectious disease outbreaks from densely sampled whole genome sequence data. However, a fundamental and as of yet poorly addressed limitation of such approaches is the requirement for genetic diversity to arise on epidemiological timescales. Specifically, the position of infected individuals in a transmission tree can only be resolved by genetic data if mutations have accumulated between the sampled pathogen genomes. To quantify and compare the useful genetic diversity expected from genetic data in different pathogen outbreaks, we introduce here the concept of 'transmission divergence', defined as the number of mutations separating whole genome sequences sampled from transmission pairs. Using parameter values obtained by literature review, we simulate outbreak scenarios alongside sequence evolution using two models described in the literature to describe transmission divergence of ten major outbreak-causing pathogens. We find that while mean values vary significantly between the pathogens considered, their transmission divergence is generally very low, with many outbreaks characterised by large numbers of genetically identical transmission pairs. We describe the impact of transmission divergence on our ability to reconstruct outbreaks using two outbreak reconstruction tools, the R packages *outbreaker* and *phybreak*, and demonstrate that, in agreement with previous observations, genetic sequence data of rapidly evolving pathogens such as RNA viruses can provide valuable information on individual transmission events. Conversely, sequence data of pathogens with lower mean transmission divergence, including *Streptococcus pneumoniae*, *Shigella sonnei* and *Clostridium difficile*, provide little to no information about individual transmission events. Our results highlight the informational limitations of genetic sequence data in certain outbreak scenarios, and demonstrate the need to expand the toolkit of outbreak reconstruction tools to integrate other types of epidemiological data.

Author summary

The increasing availability of genetic sequence data has sparked an interest in using pathogen whole genome sequences to reconstruct the history of individual transmission events

MIDAS/Pages/default.aspx); Bill and Melinda Gates Foundation (http://www.gatesfoundation.org). AC is funded by the Medical Research Council Centre for Outbreak Analysis and Modelling (https://www.imperial.ac.uk/mrc-outbreaks). TJ is funded by the Medical Research Council Centre for Outbreak Analysis and Modelling (https://www.imperial.ac.uk/mrc-outbreaks); National Institute for Health Research—Health Protection Research Unit for Modelling Methodology (https://www.imperial.ac.uk/hpru-modelling). 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.

in an infectious disease outbreak. However, such methodologies rely on pathogen genomes mutating rapidly enough to discriminate between infected individuals, an assumption that remains to be investigated. To determine pathogen outbreaks for which genetic data is expected to be informative of transmission events, we introduce here the concept of 'transmission divergence', defined as the number of mutations separating pathogen genome sequences sampled from transmission pairs. We characterise transmission divergence of ten major outbreak causing pathogens using simulations and find significant variation between diseases, with viral outbreaks generally exhibiting higher transmission divergence than bacterial ones. We reconstruct these outbreaks using the R-packages *outbreaker* and *phybreak* and find that genetic sequence data, though useful for rapidly evolving pathogens, provides little to no information about outbreaks with low transmission divergence, such as *Streptococcus pneumoniae* and *Shigella sonnei*. Our results demonstrate the need to incorporate other sources of outbreak data, such as contact tracing data and spatial location data, into outbreak reconstruction tools.

Introduction

Understanding transmission dynamics in the early stages of an infectious disease outbreak is essential for informing effective control policy. Valuable insights can be gained by the reconstruction of the transmission tree, which describes the history of infectious events at the resolution of individual cases [1–4]. Recent years have seen significant progress in the development of statistical and computational tools for inferring such trees [5–15], with a major emphasis placed on the analysis of whole genome sequence (WGS) data, now routinely collected in many outbreak scenarios [16].

Two approaches to the inference of transmission trees from WGS have emerged. One begins with an underlying transmission model, attaching to this a model of sequence evolution that relates observed genetic relationships between pathogens to unobserved epidemiological relationships (*i.e.* transmission pairs) between infected individuals. A simple implementation involves ruling out direct transmission events between individuals separated by more than a fixed threshold of substitutions [17–19]. More sophisticated methods have specified models of sequence evolution as components in a joint likelihood, formalising expected genetic relationships in a probabilistic manner [5–9,20]. The other approach considers outbreak reconstruction from a phylogenetic perspective, inferring unobserved historical relationships between pathogen samples to capture more complex evolutionary dynamics. WGS data is used to reconstruct phylogenetic trees which are either treated as data upon which transmission histories are overlaid [10–12], or jointly inferred alongside the transmission tree itself [13–15]. Given the unprecedented level of detail of WGS data and the epidemiological insights it has provided in real-life scenarios [21–23], genetic analysis is clearly an indispensable tool for outbreak reconstruction.

However, a fundamental and so far largely unaddressed limitation of WGS data in informing outbreak reconstruction is the requirement for genetic diversity to accumulate on epidemiological timescales. The scope of outbreak scenarios for which such requirements are met has, to our knowledge, never been described. Specifically, at least one mutation must accumulate in the time between sampling of two individuals in a given transmission pair (i.e., an infector and a secondary case) in order for their position within the transmission tree to be distinguishable by genetic means. This represents a limit in the resolution of the data itself, independent of the methodology considered. Though groups of genetically identical pathogens may be identified

as a cluster of infections, finer reconstruction of the transmission events would be impossible based on genetic data alone. Such limitations may be even more problematic in methods relying on accurately estimated phylogenetic trees for inferring transmission events [11,14].

The impact of limited genetic diversity on the reconstruction of disease outbreaks remains to be investigated. While this impact undoubtedly varies across different methods, the intrinsic informativeness of genetic data with respect to the underlying transmission tree can be evaluated. The genetic diversity accumulating along transmission chains depends on various genomic and epidemiological factors. To quantify this diversity, we introduce the concept of '*transmission divergence*', which we define as the number of mutations accumulating between pathogen WGS sampled from transmission pairs.

Transmission divergence can be estimated empirically from a transmission tree by determining the number of mutations separating pathogen samples of known transmission pairs. However, accurately reconstructed transmission trees with corresponding genetic sequence data are generally not available for most pathogens. We present here a simulation based approach for estimating the transmission divergence of different pathogens using parameters available in the literature, namely the length of the pathogen genome (L), its overall mutation rate (M), its generation time distribution (W) (*i.e.* the distribution of delays between primary and secondary infections [24]) and its basic reproduction number R_0 (i.e. the average number of secondary infections caused by an index case in a fully susceptible population [25]). Specifically, we model transmission trees alongside sequence evolution and extract the number of mutations separating individual transmission pairs. Intuitively, greater transmission divergence should enable better reconstruction of these transmission trees, although the nature of this relationship remains to be described.

To explore this issue, we compare the transmission divergence of ten major outbreak-causing pathogens, namely *Zaire ebolavirus* (EBOV), SARS coronavirus (SARS-CoV), MERS coronavirus (MERS-CoV), pandemic influenza A (H1N1), Methicillin-Resistant *Staphylococcus aureus* (MRSA), *Klebsiella pneumoniae (K. pneumoniae)*, *Streptococcus pneumoniae (S. pneumoniae)*, *Shigella sonnei* (*S. sonnei*), *Mycobacterium tuberculosis* (*M. tuberculosis*), and *Clostridium difficile* (*C. difficile*). We first conduct a literature review to obtain estimates of W, M, L and R_0 for each pathogen and then estimate transmission divergence using simulations. To compare estimates of transmission divergence under different models, we use two approaches described in the literature, namely the *outbreaker* model by Jombart *et al.* [5] and the *phybreak* model by Klinkenberg *et al.* [26] These differ significantly in their model of sequence evolution, with the prior considering a single dominant pathogen strain within each host and the latter modelling the additional complexities of multiples lineages coexisting and coalescing within host. Finally, we illustrate the impact of transmission divergence on our ability to infer transmission trees, using the *outbreaker* and *phybreak* inference algorithms for the R software [27].

Results

Transmission divergence

We conducted a literature review to obtain, for each pathogen, estimates of W (S1 Table), M (S2 Table), L (S3 Table) and R_0 (S4 Table), and used these to simulate outbreaks under the *outbreaker* and *phybreak* models (Table 1). Simulated outbreaks varied in size from 30 to 99 infected individuals, with a median size of 63 and 62 cases for *outbreaker* and *phybreak* simulations, respectively. To describe the distribution of transmission divergence values for each pathogen, we calculated the number of mutations separating individual transmission pairs (Fig 1A, S5 Table). As expected from the mutational models of *outbreaker* and *phybreak*,

Table 1. Epidemiological and genomic parameters for ten major outbreak causing pathogens.

Pathogen	Generation time (SD) (in days)	Mutation rate (per site per day)	Genome length (base pairs)	Basic reproduction number R_0	References
EBOV	14.4 (8.9)	0.31×10^{-5}	18958	1.8	[28–34]
MERS-CoV	10.7 (6.0)	0.25×10^{-5}	30115	1.2	[35–40]
SARS-CoV	8.7 (3.6)	1.14×10^{-5}	29714	2.7	[41–49]
Influenza A (H1N1)	3.0 (1.5)	1.19×10^{-5}	13155	1.5	[50–56]
MRSA	15.6 (10.0)	5.21×10^{-9}	2842618	1.3	[57–68]
K. pneumoniae	62.7 (24.0)	6.30×10^{-9}	5305677	2.0	[69–78]
S. pneumoniae	6.6 (1.8)	5.44×10^{-9}	2126652	1.4	[79–88]
M. tuberculosis	324.4 (384.5)	0.24×10^{-9}	4411621	1.8	[11,12,18,89–94]
S. sonnei	8.5 (3.0)	1.64×10^{-9}	4825265	1.1	[95–100]
C. difficile	28.4 (14.9)	0.88×10^{-9}	4290252	1.5	[23,101–105]

https://doi.org/10.1371/journal.ppat.1006885.t001

transmission divergence appears to follow a mixed Poisson distribution, with the mixing distribution of the Poisson rate determined by variation in the generation-, sampling- and coalescent times.

Transmission divergence simulated under the two models differed significantly, with *phybreak* consistently estimating higher values than *outbreaker* (S5 Table). This discrepancy ranged from 1.56 times higher on average for *S. pneumoniae* to 1.84 times higher for *M. tuberculosis*, with significantly longer tailed distributions especially for *K. pneumoniae* and SARS-CoV. On the other hand, *outbreaker* and *phybreak* agreed on the relative amount of transmission divergence between pathogens, both assigning larger mean transmission divergence to viral pathogens than bacterial pathogens, with the exception of *K. pneumoniae*.

Notable across both models was the fact that transmission divergence was generally low. Pathogens such as *S. sonnei*, *S. pneumoniae* and *C. difficile* were essentially never separated by more than one mutation even when accounting for within-host diversity, suggesting that little to no genetic diversity is to be expected over the course of such outbreaks. Even rapidly mutating viral pathogens such as EBOV and MERS were generally separated by no more than five mutations under both models, and in the absence of significant within-host diversity the most common number of mutations separating such transmission pairs was indeed zero. In fact, *outbreaker* estimated a mean value below one for eight of the ten pathogens considered. In contrast, two pathogens that demonstrated significantly higher transmission divergence were *K. pneumoniae* and SARS-CoV, which accumulated as many as 15 mutations between individual transmission pairs and were rarely separated by fewer than two.

We also quantified genetic diversity by the number unique of sequences as a proportion of total sequences (Fig 1B, S5 Table). This value is closely related to the zero term in the transmission divergence distribution, however notable observations include that over 90% of sequences in *S. sonnei* and *S. pneumoniae* outbreaks were identical under both models of sequence evolution, and on average 30% to 50% of sequences in MERS-CoV and EBOV outbreaks were identical depending on the model of within-host diversity. Few genetically identical cases were observed in SARS-CoV and *K. pneumoniae* outbreaks.

Comparison with empirical results

As the proportion of unique sequences in an outbreak can be determined without knowledge of the transmission tree, we used this metric to compare our predictions with empirical estimates from studies collecting WGS in an outbreak setting (Fig 1B, S6 Table). The proportion of unique sequences observed in *M. tuberculosis*, Influenza A, MERS-CoV and SARS-CoV

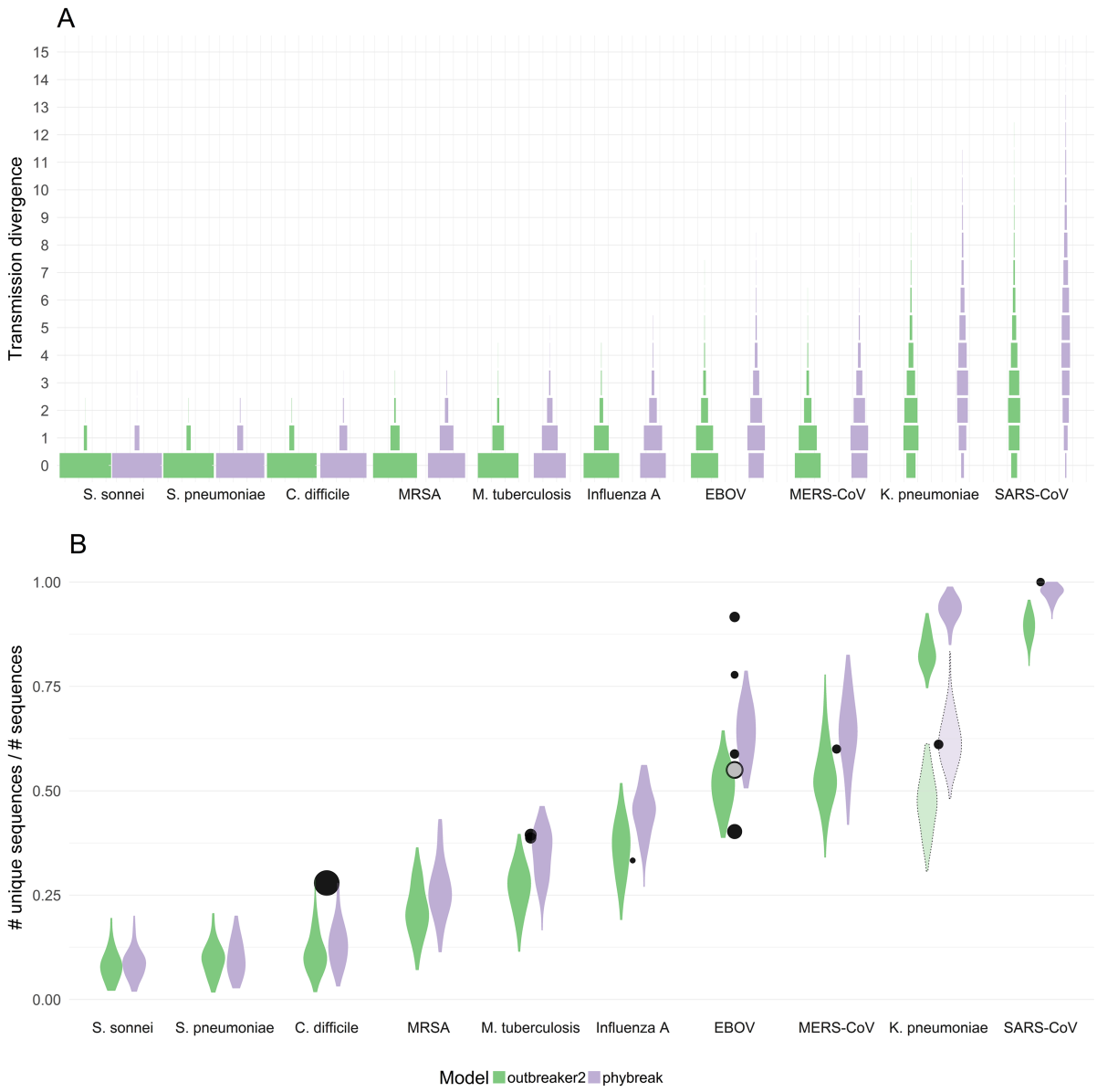

Fig 1. Distributions of simulated transmission divergence values for different pathogens using the *outbreaker* and *phybreak* models. A) Transmission divergence is defined as the number of mutations separating pathogen WGS sampled from transmission pairs. Horizontal bars indicate the proportion of transmission pairs separated by that number of mutations, across 100 outbreak simulations per pathogen. Outbreaks were simulated using both the *outbreaker* and *phybreak* models. **B)** For each simulated outbreak, we calculated the proportion of sequences that were unique. Black circles represent empirical observations of the proportion of unique sequences for a given outbreak (S6 Table), scaled by the size of the outbreak. The grey circle in the EBOV column represents the weighted mean across the four outbreaks. The violin plots with the dotted outlines in the *K. pneumoniae* column were generated using the empirical serial interval of 25.8 days observed over the course of the outbreak described by Snitkin *et al.* [106], which differs significantly from the value of 62.7 days in our literature review.

https://doi.org/10.1371/journal.ppat.1006885.g001

outbreaks were well predicted by one or both evolutionary models. The *phybreak* model better predicted the observed genetic diversity for both *M. tuberculosis* outbreaks and the SARS-CoV outbreak, the latter of which fell outside the prediction interval of the *outbreaker* model, whereas the diversity observed in the Influenza A and MERS-CoV outbreaks was similarly supported by both models.

The mean proportion of unique sequences observed across four EBOV outbreaks was also in good agreement with our simulations, with slightly greater support by the *outbreaker* model. However, diversity between these outbreaks was more variable than expected, ranging from 0.40 to 0.92, with one value falling outside both prediction intervals and two values only expected under either the *phybreak* or *outbreaker* model. Furthermore, though the genetic diversity observed across 333 cases of *C. difficile* infection in Oxfordshire, UK [17] fell just within the prediction interval of our simulations, this result was unlikely under both evolutionary models, especially given the large sample size of the study.

The greatest disagreement with our predictions was observed for a *K. pneumoniae* outbreak described by Snitkin *et al.* [106], for which 7 out of 18 WGS were identical, while our simulations predicted nearly all cases to be genetically distinguishable. However, the average serial interval over this outbreak was only 25.8 days (S6 Table), which was unusually short compared to the average value of 62.7 days from our literature review. When repeating our simulations using the realised serial interval, the observed genetic diversity was well predicted by the *phybreak* model (Fig 2B).

Impact on outbreak reconstruction

To quantify how these results affect the inference of transmission trees, we analysed the simulated outbreaks using the *outbreaker* and *phybreak* inference algorithms, applying the same models used for outbreak simulation in their reconstruction. We reconstructed each outbreak with and without WGS data, and quantified the accuracy in outbreak reconstruction, as well as the statistical confidence in ancestry assignments using the posterior entropy (S1 Fig, S2 Fig). To describe the informativeness of the genetic data alone, and minimise confounding epidemiological differences between pathogens (e.g. temporal data will be more informative with lower R_0, as there exist longer chains with a greater temporal signal), we calculated the absolute change in accuracy of outbreak reconstruction upon incorporating WGS data and related this to the mean transmission divergence of the outbreak. As the accuracy of reconstruction in the absence of WGS data was very low (S1 Fig), there was similar, considerable scope for improvements in accuracy for all pathogens.

Unsurprisingly, higher average transmission divergence across an outbreak led to greater improvements in the accuracy of outbreak reconstruction of both *outbreaker* and *phybreak* simulations (Fig 2A). However, the nature of this relationship differed between the two models. Using *outbreaker*, a sharp contrast was observed between outbreaks with low mean transmission divergence (*C. difficile*, *S. pneumoniae*), for which WGS provided essentially no additional information, and outbreaks exhibiting the largest mean transmission divergence (*K. pneumoniae*, SARS-CoV), for which WGS improved nearly every incorrect ancestry assignment. The effect of mean transmission divergence on the accuracy of outbreak reconstruction was strongly nonlinear, with the greatest improvement in accuracy obtained between 0 and 1 mutations on average between transmission pairs.

Under the *phybreak* model this relationship was less pronounced, with increases in mean transmission divergence resulting in markedly lower improvements in accuracy (Fig 2A). This was most evident in SARS-CoV and *K. pneumoniae* outbreaks, for which improvements in accuracy were lower than in the *outbreaker* simulations even though the average transmission divergence was nearly two times higher, with a significant number of ancestries remaining incorrectly assigned (S1 Fig). At high values, mean transmission divergence was also poorly predictive of increases in accuracy, which were identical between SARS-CoV and *K. pneumoniae* in spite of significantly different average transmission divergence (4.83 and 3.40, respectively).

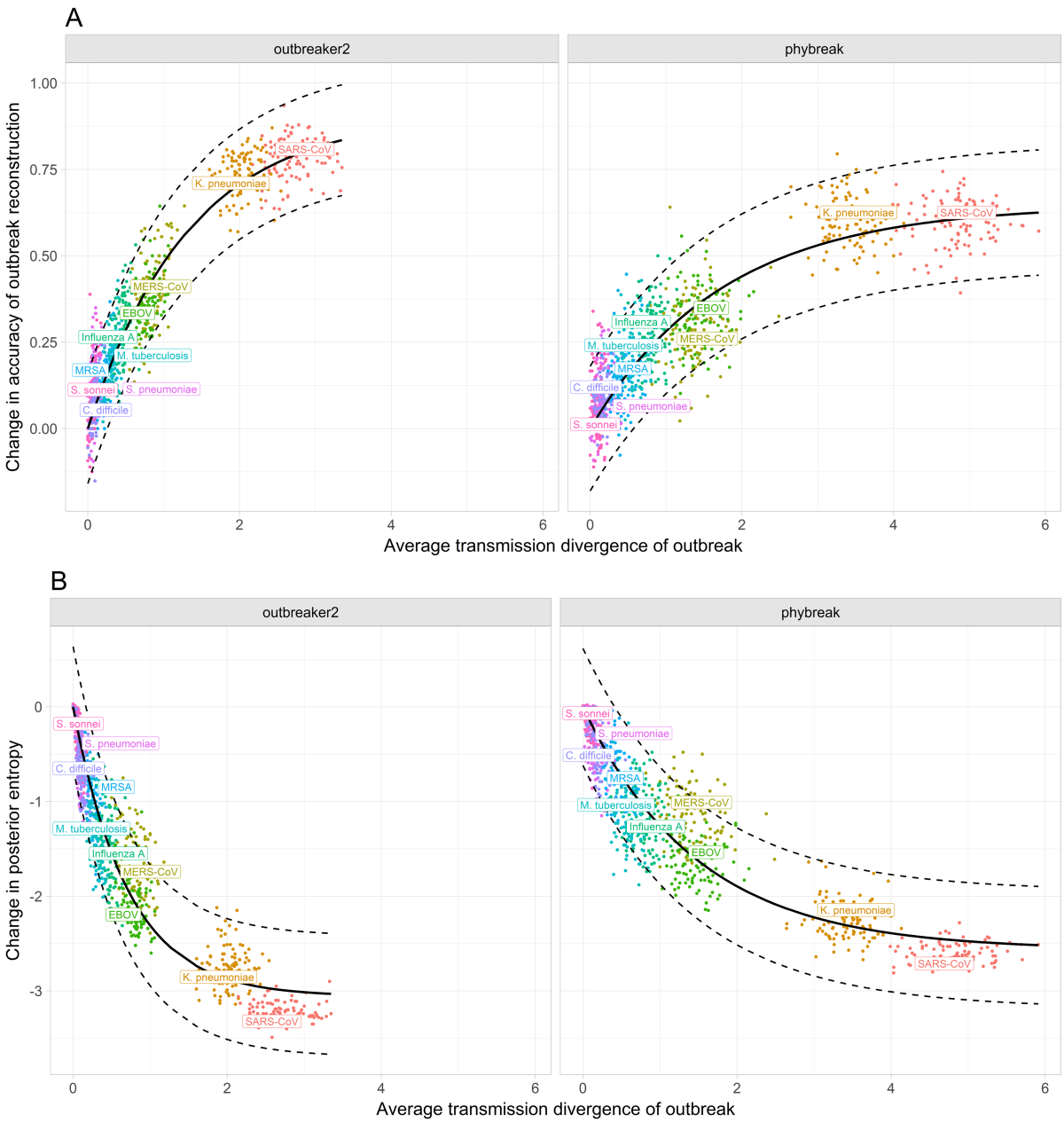

Fig 2. Impact of transmission divergence on outbreak reconstruction. Transmission divergence is defined as the number of mutations separating pathogen WGS sampled from transmission pairs. **A) Change in accuracy of outbreak reconstruction.** Accuracy of outbreak reconstruction is defined as the proportion of correctly assigned ancestries in the consensus transmission tree, itself defined as the tree with the most frequent posterior infector for each infectee. Coloured points represent individual simulated outbreaks. The solid black line represents the fitted relationship of the form $i—i^* \exp(-a^* K)$, where K is the transmission divergence and a and i the fitting variables. Dotted black lines represent the corresponding 95% prediction interval. **B) Change in posterior entropy.** Posterior entropy is related to the number of plausible posterior infectors for a given case. Lower average entropy indicates greater statistical confidence in the proposed transmission tree. The solid black line represents the fitted relationship of the form $i^* \exp(-a^* K)—i$, where K is the transmission divergence and a and i the fitting variables.

https://doi.org/10.1371/journal.ppat.1006885.g002

A similar trend was observed when considering the statistical confidence in ancestry assignments (Fig 2B). Under the *outbreaker* model, higher average transmission divergence strongly reduced posterior entropy, resulting in essentially complete support for a single transmission

tree at high values. In contrast, the consideration of within-host diversity by the *phybreak* model left outstanding uncertainty around ancestry assignments even when transmission pairs were resolved by a large number of mutations (S2 Fig).

We also related the informativeness of WGS data to the proportion of sequences that were unique in an outbreak, and found a nearly linear relationship for both *outbreaker* and *phybreak* reconstructions (Fig 3A). Once again, the slope of this relationship was significantly steeper in the *outbreaker* model, to the extent that the proportion of unique sequences was a near perfect predictor of the informativeness of WGS data, and outbreaks essentially perfectly reconstructed if all sequences were genetically distinct. This was not the case with the *phybreak* reconstructions, which assigned incorrect ancestries even when all sequences were unique. However, the proportion of unique cases was still a good predictor of the increase in accuracy of outbreak reconstruction, and successfully identified *K. pneumoniae* and SARS-CoV as having similarly informative WGS, where mean transmission divergence as a metric had placed them far apart.

The change in posterior entropy was also linearly correlated with the proportion of unique ancestries (Fig 3B), with an outbreak of genetically distinct cases sufficient for *outbreaker* to converge on a single posterior transmission tree. In contrast, there remained considerable uncertainty around *phybreak* reconstructions even in a fully genetically resolved outbreak.

Discussion

This paper has introduced the concept of 'transmission divergence' as a measure of the informativeness of WGS data for reconstructing transmission chains during an infectious disease outbreak. We estimated transmission divergence for ten major outbreak causing pathogens with a simulated based approach, using two distinct models of sequence evolution for comparison. We then demonstrated how the mean transmission divergence across an outbreak affects our ability to infer transmission histories.

Though average transmission divergence varied significantly amongst the diseases studied, it was generally very low, with a modal value of zero for most pathogens under both evolutionary models. Our results suggest that a large fraction or even a majority of cases will be genetically indistinguishable in many epidemic scenarios, including outbreaks of rapidly evolving RNA viruses such as EBOV and MERS.

These results were generally well supported by empirical observations of genetic diversity. Our simulations accurately predicted *C. difficile*, Influenza A and *M. tuberculosis* cases to be genetically identical a majority of the time, EBOV and MERS-CoV cases to display greater diversity yet still be frequently identical, and SARS-CoV cases to be largely genetically distinct. Though the proportion of unique sequences observed across 333 *C. difficile* cases was unexpectedly high, this metric assumes that each case is related to another case by a direct transmission event. This appears unrealistic when considering that 120 patients (36%) had no recorded epidemiological contact with their genetically related case [17]. A number of these pairs were probably separated by unobserved transmission events, for example by asymptomatic carriers, increasing the observed genetic diversity between supposed transmission pairs. The true proportion of unique sequences likely agrees better with our model predictions. Furthermore, though our original estimates for *K. pneumoniae* disagreed with empirical observations, this discrepancy was largely due to an unusually short serial interval compared to previously reported values in the literature. Once accounted for, the reported diversity agreed with our predictions.

Therefore, even though predicting the specific number of unique sequences observed in an outbreak is challenging due to confounding factors such as unobserved cases and stochastic

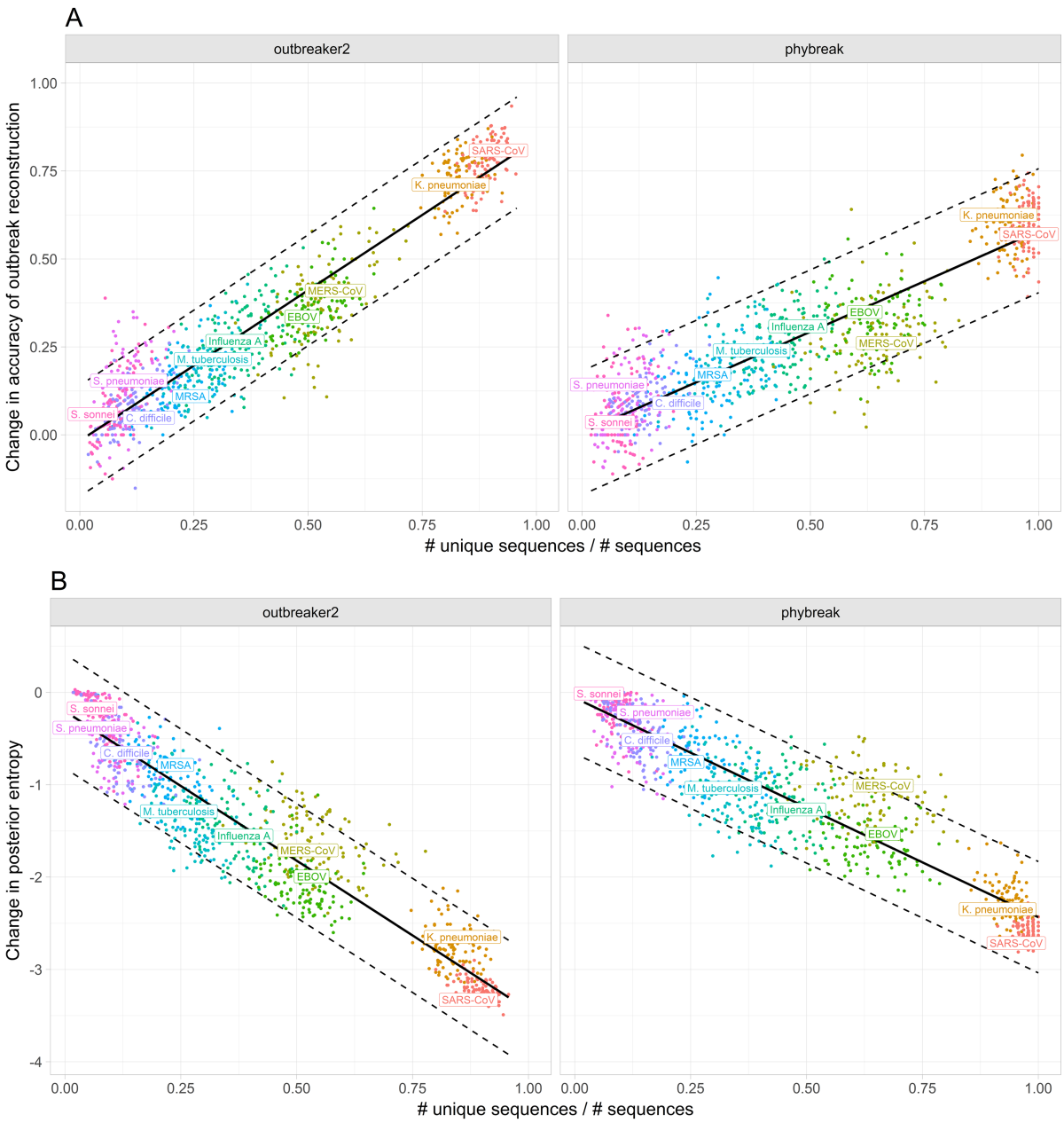

Fig 3. Impact of the proportion of unique sequences on outbreak reconstruction. A) Change in accuracy of outbreak reconstruction. Accuracy of outbreak reconstruction is defined as the proportion of correctly assigned ancestries in the consensus transmission tree, itself defined as the tree with the most frequent posterior infector for each infectee. Coloured points represent individual simulated outbreaks. The solid black line represents the fitted linear model, the dotted black lines the 95% prediction interval. **B) Change in posterior entropy.** Posterior entropy is related to the number of plausible posterior infectors for a given case. Lower average entropy indicates greater statistical confidence in the proposed transmission tree. The solid black line represents the fitted linear model, the dotted black lines the 95% prediction interval.

https://doi.org/10.1371/journal.ppat.1006885.g003

variations in epidemiological context, our predictions using two fairly simple models of sequence evolution largely agreed with the data. This suggests that our results represent broadly useful predictors of the extent to which cases in a transmission cluster are genetically resolvable. The wider conclusion that a significant proportion of cases are expected to be genetically identical for a number of different pathogens is certainly well supported.

It is tempting to suggest that the data from *M. tuberculosis* and SARS-CoV outbreaks lend greater support to the *phybreak* model, thereby indicating significant levels of within-host genetic diversity among these pathogens. While this explanation is feasible, the significant variation in unique sequences across four EBOV outbreaks demonstrates the sensitivity of such individual observations to stochastic effects. In the absence of greater amounts of empirical data, any such conclusions are only weakly supported.

The limited genetic diversity as predicted by our simulations had a considerable impact on our ability to reconstruct outbreaks, and clearly identified transmission divergence as a limiting factor in the utility of WGS data for many pathogens in an outbreak setting. These informational limitations were further compounded by within-host genetic diversity, which significantly reduced our ability to reconstruct outbreaks even when mean transmission divergence was high, in agreement with previous studies [11,107]. Combined, these results demonstrate that WGS data will often be insufficient to fully resolve transmission chains, and reveal the need to incorporate other sources of information into transmission inference frameworks. Promising avenues include an analysis of deep sequencing data as an alternative to WGS, which may reveal additional within-host variation informative of likely transmission events, as well as a methodological approach to inferring transmission routes from contact data.

Our results do not imply that WGS is of no use for inferring transmission routes as a whole. For example, Didelot *et al.* used *C. difficile* WGS to identify distinct transmission chains caused by separate introductions to the same ward, vastly reducing the number of plausible transmission links given only epidemiological data [23]. However, samples within the transmission chains were genetically identical and remained unresolved in the absence of additional data. Low transmission divergence therefore represents a hard limit to the resolution of various reconstruction methods using WGS as primary source of information, regardless of the underlying genetic model. This may especially impact approaches relying on previously constructed phylogenetic trees [12,14], which are known to skew infection time estimates in the presence of multiple genetically identical sequences, as described by Hall *et al.* [15]

We also showed that greater transmission divergence generally improved the inference of transmission histories, however only to an extent. Beyond the first discriminatory mutation, diversity between transmission pairs seemed to provide limited additional information, as demonstrated by the fact that the proportion of genetically unique sequences, rather than the average transmission divergence, best predicted the informativeness of WGS data. Though this relationship was weaker in the *phybreak* model, as within-host diversity increases the number of plausible transmission trees even when all cases are genetically distinct, this linear relationship held across both models.

It is important to note that other epidemiological factors beyond genetic diversity will impact the accuracy of outbreak reconstruction, such as R_0, heterogeneities in infectiousness, the generation time distribution and the sampling time distribution. To account for these effects, our study focused on the improvement in the accuracy of reconstructed transmission chains, compared to a baseline without WGS data while keeping these other factors constant. Importantly, in spite of considerable variation in R_0 and generation time distributions (Table 1), the robustness of the correlations presented in Figs 2 and 3 suggests that our measures of genetic diversity have captured a central determinant of the utility of WGS data in reconstructing outbreaks.

This study made several assumptions which might be relaxed in further work. Firstly, we assumed that the sampling time distribution is equivalent to the generation time distribution. While this assumption was largely driven by the lack of available data on sampling delay distributions, this could result in biases. For instance, our approach would underestimate transmission divergence in the presence of systematic and substantial lags between transmission times

and sampling (S3 Fig). It is worth noting, however, that additional mutations accumulating in a lineage after onward transmission has ceased would increase the overall genetic distance from this lineage to all other isolates equally, without providing additional information about the underlying epidemiological relationships between hosts. Therefore, it is unclear how this additional diversity would translate in terms of improving outbreak reconstruction, and we believe the approach used in this study should capture the diversity informative of the transmission network.

Secondly, outbreaks were simulated and reconstructed under idealised scenarios, in that all cases were observed, WGS were available for all cases, and the same parameter values used for simulation and inference. Most importantly, we assumed error-free sequencing. In reality, when considering that transmission divergence is generally on the order of single mutations, individual sequencing errors can heavily bias the topology of the inferred transmission tree. Our estimates of the informativeness of WGS in inferring individual transmission links are therefore likely to be optimistic.

Finally, both the *outbreaker* and the *phybreak* model assumed a complete bottleneck at transmission, with a single strain being transmitted. Allowing for an incomplete bottleneck greatly increase the complexity of the problem, as two strains in a given host may have diverged several infectious generations ago and passed through multiple bottleneck together. This issue also opens up a number of questions about optimal sampling and sequencing strategies, the exact magnitude of the genetic diversity bottleneck at transmission, and the more fundamental mechanisms permitting the coexistence of multiple strains within a host. Further work should be dedicated to investigating the impact of these issues on the use of WGS for outbreak reconstruction [107,108].

The advent of WGS data has initiated a revolution in modern infectious disease epidemiology, shedding new light into disease dynamics and evolution at a variety of scales [109,110]. At a local scale, these data have opened up exciting perspectives for improving our understanding of the person-to-person transmission process [5,6,13,20]. This work suggests that, while useful, the analysis of WGS alone will struggle to reconstruct transmission trees accurately for a large number of pathogens, in particular bacterial ones. Integrating other types of outbreak data, such as locations of patients, community structure, or contact tracing data, therefore represents a promising alternative strategy for outbreak reconstruction.

Materials and methods

Epidemiological and genomic parameters

We conducted a literature review using OvidEmbase to attain values for the generation time distribution, basic reproduction number, mutation rate and genome length of the pathogens under consideration. The searches were performed between 1st June 2015 and 3rd March 2017 and limited to publications in English. Common name variants for each pathogen were included in the searches as follows:

1. Ebola OR Ebola virus
2. MERS or MERS-CoV OR Middle East respiratory syndrome
3. SARS OR SARS-CoV OR Severe acute respiratory syndrome
4. (Influenza OR flu) AND (H1N1 OR A(H1N1) OR pandemic)
5. *Staphylococcus aureus* OR MRSA OR Methicillin resistant *Staphylococcus aureus*
6. *Klebsiella pneumoniae*

7. *Streptococcus pneumoniae*
8. *Mycobacterium tuberculosis*
9. *Shigella sonnei*
10. *Clostridium difficile*

For the generation time distribution, we used the search terms 'generation time OR serial interval OR generation interval'. Estimates of the serial interval were used as a proxy for the generation time. We summarised the mean and standard deviation of these distributions by calculating the arithmetic mean weighted by the sample size of the study. We then generated discretized gamma distributions of the generation time distribution, using the function *DiscrSI* from the R package *EpiEstim* [111].

Studies describing the mutation rate were identified using the search terms 'mutation rate OR substitution rate OR spontaneous mutation'. The arithmetic mean was calculated to summarise findings (Table 1). A discrepancy of two orders of magnitude between mutation rate estimates for *Clostridium difficile* was resolved by choosing the short-term molecular clock estimate, derived from serial pairs of isolates in a hospital outbreak [23], over a long-term estimate using historical phylogenetic analysis [112]. Mutation rates were converted to units of mutations per site per day.

Core genome length estimates were retrieved from complete genome assemblies in the GenBank repository [113], and the rounded arithmetic mean used as a summary value.

Studies estimating R_0 were identified using the search terms: 'basic reproduction number OR basic reproductive number'. Only studies explicitly inferring the *basic* reproduction number, defined as the expected number of secondary infections caused by an index case in a wholly susceptible population, were selected. The arithmetic mean was used as a summary value.

Estimating transmission divergence from simulated outbreaks

We define the transmission divergence K as the number of mutations separating pathogen WGS sampled from transmission pairs. We estimated the distribution of values for K by simulating transmission events alongside sequence evolution under two different models. The first is the *outbreaker* model, described in full by Jombart *et al.* [5] Briefly, the infectiousness of a case at a given time since infection is described by the generation time distribution W scaled by the basic reproduction number R_0. The time of sampling is drawn from the sampling time distribution S. Mutations accumulate in the time between infection of a primary and secondary case, at a daily rate given by the product of the genome length L and the mutation rate M. Within-host pathogen diversity is considered negligible, such that the same strain is both onwardly transmitted and sampled, and the bottleneck at transmission is assumed complete.

The *phybreak* model is described by Klinkenberg *et al.* [26], and differs from the *outbreaker* model primarily in its model of sequence evolution. Instead of modelling mutations as independent events between individual transmission pairs, *phybreak* accounts for patterns of shared evolution and within-host diversity by simulating phylogenetic 'mini-trees' within each case, which are combined according to the transmission tree. The coalescent events within-host are simulated under a linearly growing pathogen population size, assuming a complete bottleneck at transmission. Mutations accumulate along the branches as a Poisson process with a mean value of the mutation rate M. As with *outbreaker*, infection times and sampling times are drawn from the generation time distribution W and sampling time distribution S, respectively. The number of contacts is Poisson distributed with a mean of R_0, resulting in transmission if these occurred with previously uninfected cases.

Simulation settings

For both *outbreaker* and *phybreak*, we simulated outbreaks with 100 susceptible hosts and a single initial infection using parameter values for M, W, L and R_0 obtained by literature review. The sampling time distribution was assumed to be the equivalent to the generation time distribution, and external imports of infection were not considered. Within-host evolution was modelled in *phybreak* with an effective pathogen population size increasing at a daily rate of 1. Simulations were run for 500 days or until no more infectious individuals remained, except *M. tuberculosis* simulations which were run for 500 weeks due to the longer generation time. For each pathogen, we generated 100 *outbreaker* simulations and 100 *phybreak simulations* with a minimum size of 30 infected individuals. The distributions of transmission divergence values were extracted by determining the number of mutations separating each transmission pair.

Outbreak reconstruction

We reconstructed *outbreaker* and *phybreak* simulations using the transmission tree inference algorithms in the *outbreaker2* and *phybreak* packages, respectively, which implement the same models used for simulation described above. The generation time and sampling time distributions used for simulations were also used for inference, and the assumed rate of within-host population growth in *phybreak* fixed at the simulated value. *outbreaker* MCMC chains were run for 100,000 iterations with a thinning frequency of 1/200, and *phybreak* MCMC chains for 10,000 iterations with a thinning frequency of 1/20. The burn-in period for both analyses was 1,000 iterations. To assess the improvement in transmission tree reconstruction due to genetic data alone, two types of analyses were performed for each simulated dataset, the first one using only sampling times, and the second one using both sampling times and WGS data. As the phybreak algorithm requires WGS data to be provided, all cases were assigned identical genomes to imitate the absence of genetic information.

For each simulation, we quantified the accuracy of outbreak reconstruction as the proportion of correctly inferred ancestries in the consensus transmission tree, defined as the tree with the modal posterior infector for each sampled case. Cycles were resolved using Edmond's algorithm [114]. The change in accuracy was defined as the absolute difference in accuracy upon inclusion of WGS data.

To quantify the statistical confidence in ancestry assignments contained in the posterior distribution, we calculated the entropy of posterior ancestries for each case [115]. Given K ancestors of frequency f_K ($k = 1, \ldots, K$), the entropy is defined as:

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

An entropy value of 0 therefore indicates complete posterior support for a given ancestry, with higher values indicating a larger number of plausible transmission scenarios.

Supporting information

S1 Fig. Accuracy of outbreak reconstruction with and without WGS. 100 outbreaks were simulated and reconstructed for each pathogen, using both the *outbreaker* and *phybreak* model. Accuracy of outbreak reconstruction is defined as the proportion of correctly assigned ancestries in the consensus transmission tree, itself defined as the tree with the most frequent posterior infector for each infectee.
(TIF)

S2 Fig. Posterior entropy with and without WGS. 100 outbreaks were simulated and reconstructed for each pathogen, using both the *outbreaker* and *phybreak* model. Posterior entropy is related to the number of plausible posterior infectors for a given case, with lower average entropy indicating greater statistical confidence in the proposed transmission tree.
(TIF)

S3 Fig. Quantifying the time for mutations to accumulate between pathogen genomes sampled from a transmission pair under the *outbreaker* model. Individual *i* infects individual *j*. Infection and sampling times are indicated by circles and diamonds, respectively. The generation time $W_{i,j}$ is defined as the intervals between infection of *i* and the secondary case *j*, and is drawn from the distribution W. S_i denotes the time to sampling of individual *i*, and is drawn from the distribution *S*. The time for discriminatory mutations to occur between pathogen genomes sampled from *i* and *j* is denoted $O_{i,j}$, and is represented by red lines.
A. If sampling of *i* occurs after onwards infection:
$O_{i,j} = S_i — W_{i,j} + S_j$
$E(O_{i,j}) = 2^*E(S) — E(W)$
If the difference between the expected generation time and expected time to sampling is negligible:
$E(O_{i,j}) \approx E(W)$
B. If sampling of *i* occurs before onwards infection:
$O_{i,j} = W_{i,j} — S_i + S_j$
$E(O_{i,j}) = E(W)$
The time for mutations to occur is well approximated by the generation time if the delay between sampling and onwards infection is small. If sampling consistently occurs long after onwards infection, the time for mutations to occur will be underestimated.
(TIF)

S1 Table. Generation time distributions.
(DOCX)

S2 Table. Mutation rates.
(DOCX)

S3 Table. Genome lengths.
(DOCX)

S4 Table. Basic reproduction number R0.
(DOCX)

S5 Table. Transmission divergence and its effect on outbreak reconstruction for different pathogens.
(DOCX)

S6 Table. Proportion of unique WGS collected in an outbreak setting.
(DOCX)

Author Contributions

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

Data curation: Finlay Campbell, Camilla Strang, Thibaut Jombart.

Formal analysis: Finlay Campbell, Thibaut Jombart.

Funding acquisition: Neil Ferguson.

Investigation: Finlay Campbell, Camilla Strang, Thibaut Jombart.

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

Project administration: Neil Ferguson, Anne Cori, Thibaut Jombart.

Software: Finlay Campbell.

Supervision: Neil Ferguson, Anne Cori, Thibaut Jombart.

Validation: Finlay Campbell, Neil Ferguson.

Visualization: Finlay Campbell, Thibaut Jombart.

Writing – original draft: Finlay Campbell.

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

References

1. 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
2. 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
3. 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
4. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438: 355–359. https://doi.org/10.1038/nature04153 PMID: 16292310
5. 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. https://doi.org/10.1371/journal.pcbi.1003457 PMID: 24465202
6. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol. 2012;8. https://doi.org/10.1371/journal.pcbi.1002768 PMID: 23166481
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. 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
9. 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
10. 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. Proc Biol Sci. 2008; 275: 887–895. https://doi.org/10.1098/rspb.2007.1442 PMID: 18230598
11. 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
12. 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

13. Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013; 195: 1055–1062. https://doi.org/10.1534/genetics.113.154856 PMID: 24037268

14. 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

15. 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

16. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl J m., Laurent F, et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013; 18: 20380. PMID: 23369389

17. 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. N Engl J Med. 2013; 369: 1195–1205. https://doi.org/10.1056/NEJMoa1216064 PMID: 24066741

18. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013; 13: 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3 PMID: 23158499

19. 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. 2014/4; 2: 285–292. https://doi.org/10.1016/S2213-2600(14)70027-X PMID: 24717625

20. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data [Internet]. Proceedings of the Royal Society B: Biological Sciences. 2012. pp. 444–450. https://doi.org/10.1098/rspb.2011.0913 PMID: 21733899

21. Hatherell H-A, Didelot X, Pollock SL, Tang P, Crisan A, Johnston JC, et al. Declaring a tuberculosis outbreak over with genomic epidemiology. Microbial Genomics. Microbiology Society; 2016;2. https://doi.org/10.1099/mgen.0.000060 PMID: 28348853

22. Chung The H, Karkey A, Pham Thanh D, Boinett CJ, Cain AK, Ellington M, et al. A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of Klebsiella pneumoniae. EMBO Mol Med. 2015; 7: 227–239. https://doi.org/10.15252/emmm.201404767 PMID: 25712531

23. Didelot X, Eyre DW, Cule M, Ip CLC, Ansari MA, Griffiths D, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol. 2012; 13: R118. https://doi.org/10.1186/gb-2012-13-12-r118 PMID: 23259504

24. Fine PEM. The interval between successive cases of an infectious disease. Am J Epidemiol. 2003; 158: 1039–1047. PMID: 14630599

25. 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

26. 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

27. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available: https://www.R-project.org/

28. 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 Medecine. 2014; 371: 1481–1495.

29. 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

30. 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

31. 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.

32. 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

33. 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

34. 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.

35. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, et al. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. Lancet Infect Dis. Elsevier; 2014; 14: 50–56. https://doi.org/10.1016/S1473-3099(13)70304-9 PMID: 24239323

36. 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

37. Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. Family cluster of Middle East respiratory syndrome coronavirus infections. N Engl J Med. 2013; 368: 2487–2494. https://doi.org/10.1056/NEJMoa1303729 PMID: 23718156

38. Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. Lancet. 2013; 382: 1993–2002. https://doi.org/10.1016/S0140-6736(13)61887-5 PMID: 24055451

39. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. MBio. 2014;5. https://doi.org/10.1128/mBio.01062-13 PMID: 24549846

40. de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, et al. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol. 2013; 87: 7790–7792. https://doi.org/10.1128/JVI.01244-13 PMID: 23678167

41. 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

42. 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

43. 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

44. 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

45. 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

46. 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.

47. 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

48. 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

49. 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

50. Levy JW, Cowling BJ, Simmerman JM, Olsen SJ, Fang VJ, Suntarattiwong P, et al. The serial intervals of seasonal and pandemic influenza viruses in households in Bangkok, Thailand. Am J Epidemiol. 2013; 177: 1443–1451. https://doi.org/10.1093/aje/kws402 PMID: 23629874

51. Hahné S, Donker T, Meijer A, Timen A, Van Steenbergen J, Osterhaus A, et al. Epidemiology and control of influenza A (H1N1) v in the Netherlands: the first 115 cases. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2009; 14: 2335–2346.

52. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, Kent CK, et al. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. N Engl J Med. 2009; 361: 2619–2627. https://doi.org/10.1056/NEJMoa0905498 PMID: 20042753

53. 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

54. Rambaut A, Holmes E. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr. 2009; 1: RRN1003.

55. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. J Virol. 2008; 82: 596–601. https://doi.org/10.1128/JVI.02005-07 PMID: 17942553

56. Boëlle P-Y, Ansart S, Cori A, Valleron A-J. Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: a review. Influenza Other Respi Viruses. 2011; 5: 306–316.

57. Patel M, Thomas HC, Room J, Wilson Y, Kearns A, Gray J. Successful control of nosocomial transmission of the USA300 clone of community-acquired meticillin-resistant Staphylococcus aureus in a UK paediatric burns centre. J Hosp Infect. 2013/8; 84: 319–322. https://doi.org/10.1016/j.jhin.2013.04.013 PMID: 23711818

58. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, et al. Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak [Internet]. New England Journal of Medicine. 2012. pp. 2267–2275. https://doi.org/10.1056/NEJMoa1109910 PMID: 22693998

59. Lee H, Kim ES, Choi C, Seo H, Shin M, Bok JH, et al. Outbreak among healthy newborns due to a new variant of USA300-related meticillin-resistant Staphylococcus aureus. J Hosp Infect. 2014; 87: 145–151. https://doi.org/10.1016/j.jhin.2014.04.003 PMID: 24856113

60. Uhlemann A- C, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MTG, et al. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. Proc Natl Acad Sci U S A. 2014; 111: 6738–6743. https://doi.org/10.1073/pnas.1401006111 PMID: 24753569

61. Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013; 23: 653–664. https://doi.org/10.1101/gr.147710.112 PMID: 23299977

62. McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. Proc Natl Acad Sci U S A. 2012; 109: 9107–9112. https://doi.org/10.1073/pnas.1202869109 PMID: 22586109

63. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327: 469–474. https://doi.org/10.1126/science.1182395 PMID: 20093474

64. Harris SR, Cartwright EJP, Török ME, Holden MTG, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis. 2013; 13: 130–136. https://doi.org/10.1016/S1473-3099(12)70268-2 PMID: 23158674

65. Holden MTG, Feil EJ, Lindsay JA, Peacock SJ, Day NPJ, Enright MC, et al. Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A. 2004; 101: 9786–9791. https://doi.org/10.1073/pnas.0402521101 PMID: 15213324

66. Holt DC, Holden MTG, Tong SYC, Castillo-Ramirez S, Clarke L, Quail MA, et al. A very early-branching Staphylococcus aureus lineage lacking the carotenoid pigment staphyloxanthin. Genome Biol Evol. 2011; 3: 881–895. https://doi.org/10.1093/gbe/evr078 PMID: 21813488

67. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. Lancet. 2006; 367: 731–739. https://doi.org/10.1016/S0140-6736(06)68231-7 PMID: 16517273

68. Wang X, Panchanathan S, Chowell G. A data-driven mathematical model of CA-MRSA transmission among age groups: evaluating the effect of control interventions. PLoS Comput Biol. 2013; 9: e1003328. https://doi.org/10.1371/journal.pcbi.1003328 PMID: 24277998

69. Van't Veen A, Van Der Zee A, Nelson J, Speelberg B, Kluytmans JAJW, Buiting AGM. Outbreak of infection with a multiresistant Klebsiella pneumoniae strain associated with contaminated roll boards in operating rooms. J Clin Microbiol. 2005; 43: 4961–4967. https://doi.org/10.1128/JCM.43.10.4961-4967.2005 PMID: 16207948

70. Kassis-Chikhani N, Decré D, Gautier V, Burghoffer B, Saliba F, Mathieu D, et al. First outbreak of multidrug-resistant Klebsiella pneumoniae carrying blaVIM-1 and blaSHV-5 in a French university hospital. J Antimicrob Chemother. 2006; 57: 142–145. https://doi.org/10.1093/jac/dki389 PMID: 16284103

71. Macrae MB, Shannon KP, Rayner DM, Kaiser AM, Hoffman PN, French GL. A simultaneous outbreak on a neonatal unit of two strains of multiply antibiotic resistant Klebsiella pneumoniae controllable only by ward closure. J Hosp Infect. 2001; 49: 183–192. https://doi.org/10.1053/jhin.2001.1066 PMID: 11716635

72. Mathers AJ, Stoesser N, Sheppard AE, Pankhurst L, Giess A, Yeh AJ, et al. Klebsiella pneumoniae carbapenemase (KPC)-producing K. pneumoniae at a single institution: insights into endemicity from whole-genome sequencing. Antimicrob Agents Chemother. 2015; 59: 1656–1663. https://doi.org/10.1128/AAC.04292-14 PMID: 25561339

73. Karkey A, Thanh DP, Boinett CJ, Cain AK, Ellington M, Baker KS, et al. A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of Klebsiella pneumoniae. EMBO Mol Med. EMBO Press; 2015; 7: 227–239. https://doi.org/10.15252/emmm.201404767 PMID: 25712531

74. De Champs C, Rich C, Chandezon P, Chanal C, Sirot D, Forestier C. Factors associated with antimicrobial resistance among clinical isolates of Klebsiella pneumoniae: 1-year survey in a French university hospital. Eur J Clin Microbiol Infect Dis. 2004; 23: 456–462. https://doi.org/10.1007/s10096-004-1144-2 PMID: 15148654

75. Liu P, Li P, Jiang X, Bi D, Xie Y, Tai C, et al. Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. J Bacteriol. 2012; 194: 1841–1842. https://doi.org/10.1128/JB.00043-12 PMID: 22408243

76. Fookes M, Yu J, De Majumdar S, Thomson N, Schneiders T. Genome Sequence of Klebsiella pneumoniae Ecl8, a Reference Strain for Targeted Genetic Manipulation. Genome Announc. 2013;1. https://doi.org/10.1128/genomeA.00027-12 PMID: 23405357

77. Shin SH, Kim S, Kim JY, Lee S, Um Y, Oh M-K, et al. Complete genome sequence of the 2, 3-butanediol-producing Klebsiella pneumoniae strain KCTC 2242. J Bacteriol. Am Soc Microbiol; 2012; 194: 2736–2737.

78. Sypsa V, Psichogiou M, Bouzala G-A, Hadjihannas L, Hatzakis A, Daikos GL. Transmission dynamics of carbapenemase-producing Klebsiella pneumoniae and anticipated impact of infection control strategies in a surgical unit. PLoS One. 2012; 7: e41068. https://doi.org/10.1371/journal.pone.0041068 PMID: 22859965

79. 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.

80. Thomas HL, Gajraj R, Slack MPE, Sheppard C, Hawkey P, Gossain S, et al. An explosive outbreak of Streptococcus pneumoniae serotype-8 infection in a highly vaccinated residential care home, England, summer 2012. Epidemiol Infect. 2015; 143: 1957–1963. https://doi.org/10.1017/S0950268814002490 PMID: 25298247

81. Kuroki T, Ishida M, Suzuki M, Furukawa I, Ohya H, Watanabe Y, et al. Outbreak of Streptococcus pneumoniae serotype 3 pneumonia in extremely elderly people in a nursing home unit in Kanagawa, Japan, 2013. J Am Geriatr Soc. 2014; 62: 1197–1198. https://doi.org/10.1111/jgs.12863 PMID: 24925560

82. Stevens KE, Sebert ME. Frequent beneficial mutations during single-colony serial transfer of Streptococcus pneumoniae. PLoS Genet. 2011; 7: e1002232. https://doi.org/10.1371/journal.pgen.1002232 PMID: 21876679

83. Henderson-Begg SK, Livermore DM, Hall LMC. Effect of subinhibitory concentrations of antibiotics on mutation frequency in Streptococcus pneumoniae. J Antimicrob Chemother. 2006; 57: 849–854. https://doi.org/10.1093/jac/dkl064 PMID: 16531433

84. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011; 331: 430–434. https://doi.org/10.1126/science.1198545 PMID: 21273480

85. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science. 2001; 293: 498–506. https://doi.org/10.1126/science.1061217 PMID: 11463916

86. Camilli R, Bonnal RJP, Del Grosso M, Iacono M, Corti G, Rizzi E, et al. Complete genome sequence of a serotype 11A, ST62 Streptococcus pneumoniae invasive isolate. BMC Microbiol. 2011; 11: 25. https://doi.org/10.1186/1471-2180-11-25 PMID: 21284853

87. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, et al. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol. 2010; 11: R107. https://doi.org/10.1186/gb-2010-11-10-r107 PMID: 21034474

88. Hoti F, Erästö P, Leino T, Auranen K. Outbreaks of Streptococcus pneumoniae carriage in day care cohorts in Finland—implications for elimination of transmission. BMC Infect Dis. BioMed Central; 2009; 9: 102. https://doi.org/10.1186/1471-2334-9-102 PMID: 19558701

89. ten Asbroek AH, Borgdorff MW, Nagelkerke NJ, Sebek MM, Devillé W, van Embden JD, et al. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis. 1999; 3: 414–420. PMID: 10331731

90. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013; 10: e1001387. https://doi.org/10.1371/journal.pmed.1001387 PMID: 23424287

91. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015;4. https://doi.org/10.7554/eLife.05166 PMID: 25732036

92. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998; 393: 537–544. https://doi.org/10.1038/31159 PMID: 9634230

93. Camus J-C, Pryor MJ, Médigue C, Cole ST. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology. 2002; 148: 2967–2973. https://doi.org/10.1099/00221287-148-10-2967 PMID: 12368430

94. 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

95. Makintubee S, Mallonee J, Istre GR. Shigellosis outbreak associated with swimming. Am J Public Health. 1987; 77: 166–168. PMID: 3541651

96. World Health Organization. Guidelines for the control of shigellosis, including epidemics due to Shigella dysenteriae type 1. Geneva: World Health Organization; 2005; Available: http://apps.who.int/iris/handle/10665/43252

97. European Centre for Disease Prevention and Control. Systematic review on the incubation and infectiousness/shedding period of communicable diseases in children. 2016;Stockholm.

98. Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, et al. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012; 44: 1056–1059. https://doi.org/10.1038/ng.2369 PMID: 22863732

99. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 2005; 33: 6445–6458. https://doi.org/10.1093/nar/gki954 PMID: 16275786

100. Joh RI, Hoekstra RM, Barzilay EJ, Bowen A, Mintz ED, Weiss H, et al. Dynamics of shigellosis epidemics: estimating individual-level transmission and reporting rates from national epidemiologic data sets. Am J Epidemiol. 2013; 178: 1319–1326. https://doi.org/10.1093/aje/kwt122 PMID: 24008913

101. Pépin J, Gonzales M, Valiquette L. Risk of secondary cases of Clostridium difficile infection among household contacts of index cases. J Infect. 2012; 64: 387–390. https://doi.org/10.1016/j.jinf.2011.12.011 PMID: 22227466

102. Sarah Walker A, Eyre DW, Wyllie DH, Dingle KE, Harding RM, O'Connor L, et al. Characterisation of Clostridium difficile Hospital Ward–Based Transmission Using Extensive Epidemiological Data and Molecular Typing. PLoS Med. Public Library of Science; 2012; 9: e1001172.

103. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, et al. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet. 2006; 38: 779–786. https://doi.org/10.1038/ng1830 PMID: 16804543

104. Lanzas C, Dubberke ER, Lu Z, Reske KA, Gröhn YT. Epidemiological model for Clostridium difficile transmission in healthcare settings. Infect Control Hosp Epidemiol. 2011; 32: 553–561. https://doi.org/10.1086/660013 PMID: 21558767

105. Norén T, Akerlund T, Bäck E, Sjöberg L, Persson I, Alriksson I, et al. Molecular epidemiology of hospital-associated and community-acquired Clostridium difficile infection in a Swedish county. J Clin Microbiol. 2004; 42: 3635–3643. https://doi.org/10.1128/JCM.42.8.3635-3643.2004 PMID: 15297509

106. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing [Internet]. Science Translational Medicine. 2012. pp. 148ra116–148ra116. https://doi.org/10.1126/scitranslmed.3004129 PMID: 22914622

107. 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

108. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 2016; 14: 150–162. https://doi.org/10.1038/nrmicro.2015.13 PMID: 26806595

109. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009; 10: 540–550. https://doi.org/10.1038/nrg2583 PMID: 19564871

110. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. Syst Biol. 2017; 66: e47–e65. https://doi.org/10.1093/sysbio/syw054 PMID: 28173504

111. 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

112. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, et al. Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc Natl Acad Sci U S A. 2010; 107: 7527–7532. https://doi.org/10.1073/pnas.0914322107 PMID: 20368420

113. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016; 44: D67–72. https://doi.org/10.1093/nar/gkv1276 PMID: 26590407

114. Gibbons A. Algorithmic Graph Theory. Cambridge University Press; 1985.

115. Shannon C. A mathematical theory of communication, Bell System Technical Journal 27: 379–423 and 623–656. Mathematical Reviews (MathSciNet): MR10, 133e. 1948;