PLOS BIOLOGY

METHODS AND RESOURCES

An open-access database of infectious disease transmission trees to explore superspreader epidemiology

Juliana C. Taube[1¤*], **Paige B. Miller**[2], **John M. Drake**[2,3]

1 Department of Mathematics, Bowdoin College, Brunswick, Maine, United States of America, **2** Odum School of Ecology, University of Georgia, Athens, Georgia, United States of America, **3** Center for the Ecology of Infectious Diseases, University of Georgia, Athens, Georgia, United States of America

¤ Current address: Department of Biology, Georgetown University, Washington, D.C., United States of America
* taubejc@gmail.com

Citation: Taube JC, Miller PB, Drake JM (2022) An open-access database of infectious disease transmission trees to explore superspreader epidemiology. PLoS Biol 20(6): e3001685. https://doi.org/10.1371/journal.pbio.3001685

Academic Editor: Steven Riley, Imperial College London, UNITED KINGDOM

Received: March 7, 2022

Accepted: May 23, 2022

Published: June 22, 2022

Data Availability Statement: All tree data are available at https://outbreaktrees.ecology.uga.edu/. Intermediate data to reproduce these analyses are also available on Dryad at https://doi.org/10.5061/dryad.nk98sf7w7.

Funding: JCT and JMD were supported by the Population Biology of Infectious Diseases REU Site, National Science Foundation grant DBI-1659683 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1659683). PBM was supported by National Science Foundation grant DGE-1545433 (https://www.nsf.gov/awardsearch/

Abstract

Historically, emerging and reemerging infectious diseases have caused large, deadly, and expensive multinational outbreaks. Often outbreak investigations aim to identify who infected whom by reconstructing the outbreak transmission tree, which visualizes transmission between individuals as a network with nodes representing individuals and branches representing transmission from person to person. We compiled a database, called OutbreakTrees, of 382 published, standardized transmission trees consisting of 16 directly transmitted diseases ranging in size from 2 to 286 cases. For each tree and disease, we calculated several key statistics, such as tree size, average number of secondary infections, the dispersion parameter, and the proportion of cases considered superspreaders, and examined how these statistics varied over the course of each outbreak and under different assumptions about the completeness of outbreak investigations. We demonstrated the potential utility of the database through 2 short analyses addressing questions about superspreader epidemiology for a variety of diseases, including Coronavirus Disease 2019 (COVID-19). First, we found that our transmission trees were consistent with theory predicting that intermediate dispersion parameters give rise to the highest proportion of cases causing superspreading events. Additionally, we investigated patterns in how superspreaders are infected. Across trees with more than 1 superspreader, we found preliminary support for the theory that superspreaders generate other superspreaders. In sum, our findings put the role of superspreading in COVID-19 transmission in perspective with that of other diseases and suggest an approach to further research regarding the generation of superspreaders. These data have been made openly available to encourage reuse and further scientific inquiry.

Introduction

In the past 20 years, emerging and reemerging infectious diseases have caused large, deadly, and expensive multinational outbreaks of SARS-CoV (Severe Acute Respiratory Syndrome

showAward?AWD_ID=1545433&HistoricalAwards=false). 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.

Abbreviations: COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome; SARS-CoV, Severe Acute Respiratory Syndrome Coronavirus 1.

(SARS)), Zika, Ebola, measles, and SARS-CoV-2 (Coronavirus Disease 2019 (COVID-19)). During outbreaks, public health officials conduct routine investigations to identify who infected whom and reconstruct the transmission tree. Transmission trees visualize transmission between cases as directed networks with nodes representing individuals and edges representing transmission from person to person. Transmission trees are typically reassembled by case-finding, contact-tracing, and detailed epidemiological interviews, followed sometimes by genome sequencing and/or probabilistic reconstruction, where the probability that one case infected another is calculated for each pair of cases [1,2]. These investigations are costly but valuable because transmission trees are information rich, including details about the settings of transmission and variation in number of secondary infections.

When published, transmission trees are shown and described in a variety of formats that makes them difficult to compare across outbreaks, let alone pathogens. Some are presented graphically using a number of different symbols and colors, or are buried in the text, making connections hard to piece together. The primary goal of this project was to create a standardized database of transmission trees that is easily accessible and analyzable. We hope that the OutbreakTrees database allows scientists and public health officials to take further advantage of outbreak investigations and their findings.

One phenomenon that is apparent in transmission trees is superspreading, which is important to the propagation patterns of several infectious diseases [3]. Lloyd-Smith and colleagues [3] quantitatively defined superspreaders as cases that cause more secondary infections than the 99th percentile of a Poisson(R_0) distribution, where R_0 is the basic reproductive number, or average number of secondary infections per case. Lloyd-Smith and colleagues [3] also conceptualized the offspring distribution (i.e., the number of infections caused by each infected individual) as a negative binomial distribution with dispersion parameter k and mean R. Large values of k denote little variation in number of secondary infections caused by each case, while small values of k ($k<1$) correspond to high heterogeneity in the offspring distribution. It was hypothesized that intermediate dispersion parameters between 0.1 and 1, depending on R, would give rise to the highest proportion of cases causing superspreading events [3].

Lloyd-Smith and colleagues' theory on superspreading assumes stability of R and k over the course of an outbreak. In reality, most outbreaks are subject to control measures. These control measures, as well as changes in behavior, can reduce disease transmission and disperse the offspring distribution, thus leading to shifts in R and k from their pre-control values, as explored by [3]. Given information on the timing of control measures, parameter values can be compared before and after controls were imposed. In the absence of this information, we propose that a comparison of parameter values in the first versus second half of a transmission tree indicates the effect of control measures and behavior changes on a given transmission tree.

While previous work has characterized the biological and social factors that give rise to superspreading events [4], how superspreaders are generated (i.e., who spreads to superspreaders) is poorly understood. In 2020, Beldomenico [5] suggested that the generation of superspreaders may be linked to biological patterns in initial viral dosage: If individuals with unusually high viral shedding cause those they infect to also have high viral shedding, then cases infected by superspreaders may be disproportionately likely to be superspreaders themselves. Another possibility is that superspreaders may be more likely to engage in riskier behavior (such as attending large gatherings or not taking precautionary measures) making them more likely to infect others with similar behavior. This behavioral heterogeneity may be a larger contributor to superspreader generation than biological heterogeneity [6]. We investigate this issue using transmission tree data, hypothesizing that superspreaders will be more likely to be infected by other superspreaders than non-superspreading cases.

Methods

Data

Transmission trees were collected by searching Google Scholar, Scopus, PubMed, and Google Images for published literature containing graphs of transmission trees or written accounts of transmission events. We used the following terms to find papers containing transmission tree information: "transmission AND (tree OR network OR chain) AND (outbreak OR disease)," "outbreak investigation," "contact tracing," "case report," and "transmission tree outbreak reconstruction." We also used the bibliographies of other papers (e.g., [3]) to find more references. With the emergence of COVID-19, we expanded our search for transmission trees to include news articles and preprints (e.g., medRxiv.org). For COVID-19, many of the trees were identified with an online database [7]. If trees could not be collected from a public source or if trees did not identify single infectors for each infectee, we contacted the authors of identified documents for further clarification or additional information. We also compiled readily available node attributes reported in the tree source. Attributes available for each tree varied but included age, sex, context of transmission, date of symptom onset, occupation, quarantine status, survival status, location, hospital, ward of hospital or care facility, symptomatic status, duration of exposure to infected individual, whether the edge was probabilistically reconstructed, relationship between individuals, serial interval, immunization status, source of edge (if tree was constructed from 2 sources), and strain or genomic sequence. Articles in languages other than English were translated using Google Translate software.

Examples of trees contained in our published database OutbreakTrees are shown in Fig 1.

Inclusion criteria

For consistency, we required that trees meet the following criteria for inclusion in the database:

- Trees must have contained 2 or more individuals; case studies of isolated infected individuals were excluded.
- Trees must represent outbreaks of directly transmitted infectious diseases in humans; trees describing sexually transmitted, foodborne, vector-borne, or waterborne diseases, as well as diseases in nonhumans (e.g., outbreaks among farm animals [11,12]) were excluded.
- Trees were constructed through epidemiological or probabilistic methods; trees constructed using only genomic or phylogenetic methods were excluded.
- Trees had to report a single infector per infectee (i.e., trees that had multiple possible infectors for any case were excluded). However, if tree topology was unaffected by randomly assigning ambiguous infectors, we included the tree and omitted specific attribute data for the assigned infector.
- Trees were presented as completed investigations in the publication; we excluded trees presented as under ongoing investigation at the time of reporting.

Data entry

Trees were manually encoded as data.tree [13] objects using relevant information from each source and converted to igraph [14] objects for manipulation and accession. Any assumptions made in entering the tree are listed with the tree in the database (e.g., if an infector is assumed due to nodes obscuring branches or a case of an ambiguous infector assignment). All scripts to

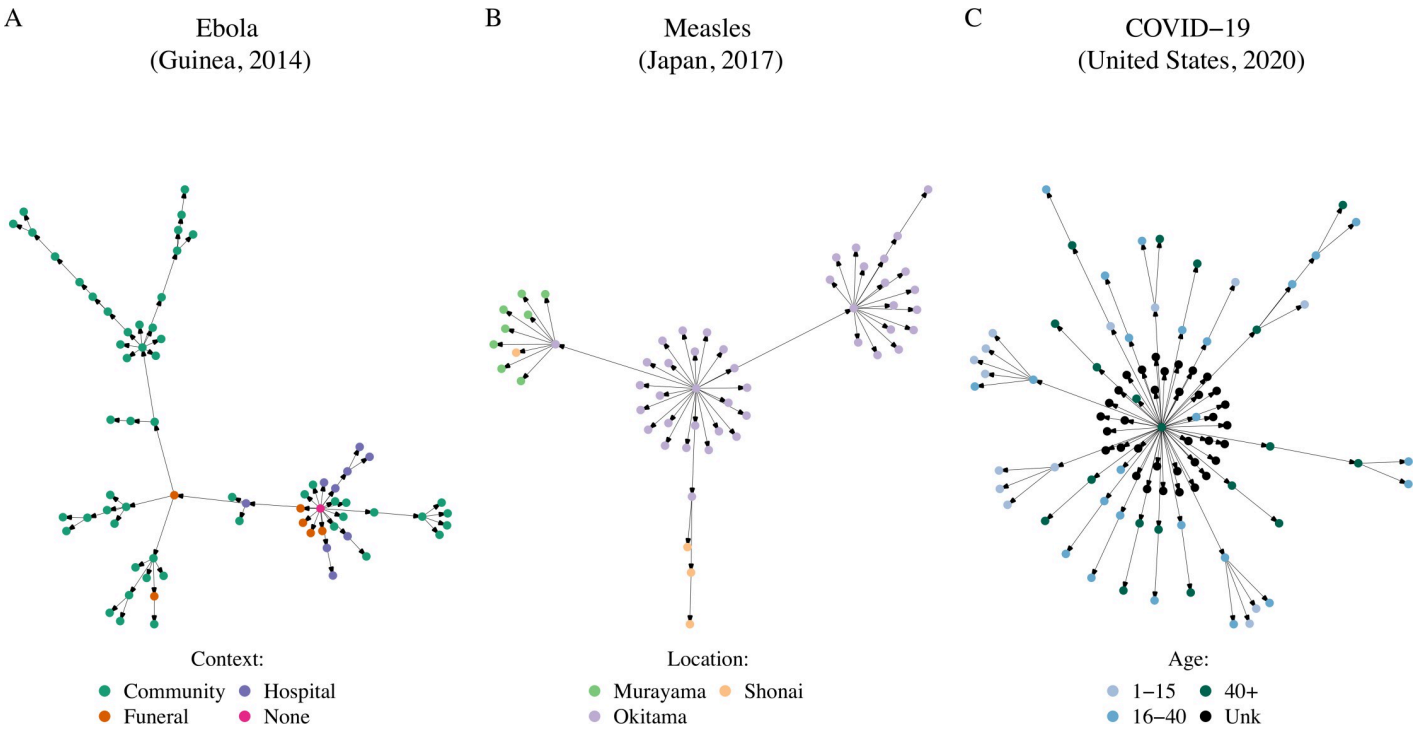

Fig 1. We compiled infectious disease transmission trees from the literature along with reported attribute information. Shown here are example trees in the database. (A) Ebola spread in different contexts [8]. (B) Measles spread in different locations [9]. (C) COVID-19 spread among age classes [10]. Primary sources for transmission trees are available in OutbreakTrees and listed in the Supporting information. OutbreakTrees may be accessed online at http://outbreaktrees.ecology.uga.edu. COVID-19, Coronavirus Disease 2019.

https://doi.org/10.1371/journal.pbio.3001685.g001

compile trees and analyze data are available at http://github.com/DrakeLab/taube-transmission-trees, and tree sources are listed in S1 File. The database is available online at http://outbreaktrees.ecology.uga.edu.

Data analysis

We demonstrated how OutbreakTrees can be used to address questions about the time dependence of epidemiological parameters and the role of superspreading in infectious disease transmission through 3 different analyses using trees with 20 or more cases and 2 or more generations of spread. We calculated key statistics under 2 contrasting assumptions about outbreak investigation completeness, explained in the Sensitivity analyses section below.

Parameter time dependence. Shifts in human behavior or disease control efforts can cause changes in key epidemiological parameters as outbreaks progress [3]. While information on intervention timing was not readily available, we explored how R, k, and the proportion of cases causing superspreading events varied over time by comparing these values in the first versus second halves of each tree. Excluding the last generation of the tree (composed solely of terminal nodes), we divided each tree into first and second halves by generation. Middle generation nodes were randomly assigned to either the first or second half of the tree. We repeated this process 10 times to account for random variation in the assignment of middle generation nodes and took the mean parameter values over the 10 repetitions. Differences were tested for significance using the Wilcoxon rank test. If population control efforts or human behavior changed transmission dynamics partway through the tree, we expected to see decreases in R, k, and the proportion of cases causing superspreading events between the first and second halves of a tree [3].

Superspreading events across diseases. To evaluate how common superspreading is among different diseases, we focused on 2 tree statistics: (1) the proportion of cases causing superspreading events and (2) the dispersion parameter, *k*. The proportion of cases causing superspreading events was calculated by dividing the number of superspreaders in a tree by the total number of nodes in the tree, where the number of superspreaders was estimated using the Lloyd-Smith and colleagues [3] definition. The dispersion parameter was calculated using maximum likelihood estimation with the fitdistr function from the mass package in R [15] assuming secondary infections followed a negative binomial distribution. Small dispersion parameters indicate more heterogeneous offspring distributions with fewer individuals accounting for the majority of transmission compared with large dispersion parameters. We performed sensitivity analyses for cutoffs of trees with 10 and 30 or more cases.

Generation of superspreaders. Next, we investigated patterns in the individuals who infected superspreaders. We calculated the ratio of observed to expected superspreader-superspreader dyads. Superspreader-superspreader dyads occur when 1 superspreader infects another. To determine the expected number of dyads per tree, we calculated the probability that a given edge connects 2 superspreaders. Denoting the number of superspreaders by *s*, number of terminal nodes (nodes that do not cause onward transmission) by *t*, and tree size by *S*, the probability that a node at 1 end of the edge is a superspreader is $\frac{s}{S-t}$, or p_1. Conditional on this first node being a superspreader, the probability that the node on the other end of the edge is a superspreader is $\frac{s-1}{S-t-1}$, or p_2. Then, the joint probability of an edge with superspreaders at either end (a dyad) is $p_1 \cdot p_2$. Given that there are *S-t-1* edges leading to nonterminal nodes in a tree, the expected number of dyads is $(S-t-1)\cdot p_1 \cdot p_2 = \frac{(S-t-1)s(s-1)}{(S-t)(S-t-1)}$ which simplifies to $\frac{s(s-1)}{S-t}$. Thus, we expect to see $\frac{s(s-1)}{S-t}$ superspreader-superspreader dyads per tree. If generation of superspreaders is not random but tied to characteristics of the infector, we would expect to see large ratios of observed to expected superspreader-superspreader dyads.

Sensitivity analyses for tree completeness. We made the assumption that trees in the database depicted complete epidemics, e.g., that all transmission events were documented and that terminal nodes did not transmit disease, yet we know that not all trees in the database are complete (see Limitations section). Recognizing that this is an extreme assumption, we conducted sensitivity analyses of the opposite extreme: Assuming all trees were incomplete, i.e., terminal nodes did transmit disease but these transmission events went unreported. In reality, the database is composed of both types of trees, complete and incomplete, as well as trees somewhere in between (e.g., last generation terminal nodes are not reliable but terminal nodes in earlier generations may be reliable), though we cannot identify which trees fall into which categories. Assuming that trees were complete, we calculated *R*, *k*, and the superspreading cutoff over all nodes in the tree, whereas under the assumption of incompleteness, we calculated *R*, *k*, and the superspreading cutoff by excluding the out-degree (zero) of all terminal nodes in any generation from the offspring distribution. We expect that *R* and *k* estimates will be higher and proportion of cases causing superspreading events estimates lower when we calculate these parameters over only nonterminal nodes than when calculated over all nodes in a tree. Results from our repeated analyses under this alternative set of assumptions can be found in the Supporting information (S1–S3, S6 and S7 Figs).

Results and discussion

Database summary statistics

Currently, OutbreakTrees includes 382 trees describing 16 directly transmitted infectious diseases (see Fig 1 for examples), most of which are caused by viruses (Fig 2). COVID-19 trees comprise 256, or approximately 67%, of the trees in the database. Trees range in size from 2 to 286 individuals; half are composed of 3 cases or fewer. This database contains data for outbreaks that took place in the years 1946 through 2020. The most common node attributes for trees include context of transmission (work, school, family, etc.), date of symptom onset, sex,

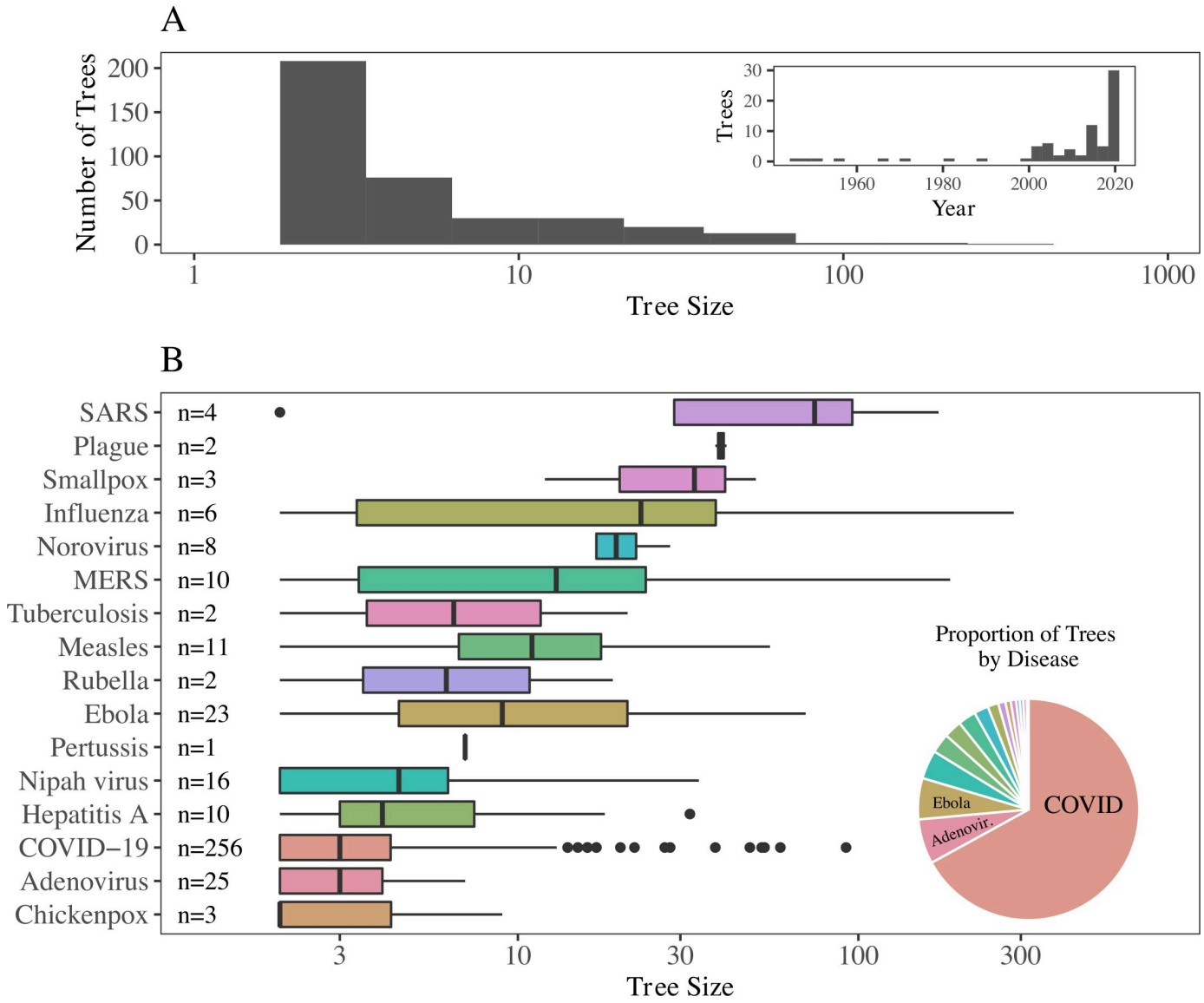

Fig 2. Characteristics of transmission trees in OutbreakTrees. (A) Tree size varies from 2 to 286 with a median of 3 and most trees represent outbreaks taking place in the past 20 years (only trees with 10 or more cases shown in date plot due to large number of small COVID-19 trees from 2020). (B) The largest trees are from H1N1 and SARS outbreaks, while the highest proportion of trees in the database are from outbreaks of COVID-19, followed by adenovirus and Ebola. Tree size axes in both plots are shown on a log_{10} scale to better illustrate variation in medium-sized trees. All trees are used in this analysis. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.

https://doi.org/10.1371/journal.pbio.3001685.g002

Table 1. List of most common attributes for individuals in trees.

Attribute	Database code	Number of trees
Transmission context	cont	301
Symptom onset	onset	137
Sex	sex	86
Age	age	69
Location	loc	56
Quarantine status	quar	36
Occupation	occp	34
Survival	surv	20

https://doi.org/10.1371/journal.pbio.3001685.t001

and age (Table 1). Due to imperfect investigation or recall, specific attributes are not available for every node in every individual tree (S1 Table).

Analyses

For the following analyses, we use a subset of trees in the database to ensure sufficient sample size for statistical analysis [16]. Specifically, estimates of R, the dispersion parameter k, the threshold number of secondary infections to be considered a superspreader, and the proportion of cases causing superspreading events for each tree are limited to trees with 20 or more cases and at least 2 generations of spread. There were 39 trees in our database that fit these criteria. The differences in R and k values depending on our assumptions of tree completeness are shown in S1 and S2 Figs. Note that when we calculate R assuming all cases are reported and the infection has died out, then R is necessarily <1 (S1 Fig). Applying the Lloyd-Smith and colleagues [3] definition of superspreading with $R\approx1$, the superspreading threshold is always more than 4 secondary infections. When we instead assume that a transmission tree is incomplete (i.e., not all cases are reported) and exclude terminal nontransmitting nodes from our calculation of R, we observe higher R values, and consequently higher superspreading cutoffs that show greater variation across diseases (S1 Fig).

Parameter time dependence. We found a significant decrease in R ($p\leq0.0001$, Wilcoxon rank test) and the proportion of cases causing superspreading events ($p\leq0.01$, Wilcoxon rank test) between the first and second halves of transmission trees with 20 or more nodes and 2 or more generations of spread assuming tree completeness (Fig 3A and 3C). The dispersion parameter did not change significantly between the first and second halves of these transmission trees (Fig 3B, Wilcoxon rank test). While all but 3 trees had $R>1$ in the first half of the tree, all trees had $R<1$ in the second half of the tree (Fig 3D). Under the assumption of incomplete trees, all 3 parameters changed significantly between the first and second halves of the trees (S3 Fig); R decreased ($p\leq0.0001$, Wilcoxon rank test), k increased ($p\leq0.01$, Wilcoxon rank test), and the proportion of cases causing superspreading events decreased ($p\leq0.001$, Wilcoxon rank test). The observed decreases in R may be the result of control measures or behavior changes in the affected populations, or could be caused by reporting biases where case follow-up is more robust in earlier generations. Similarly, the decreases in proportion of cases causing superspreading events could be due to control measures, but also superspreaders may be more likely to be identified in earlier generations if superspreading events spur outbreak investigations which may only trace transmission so far back in time. The increase in k under an assumption of tree incompleteness contradicts our expectation but may be due to the truncation of the offspring distribution to a minimum of 1 secondary infection when terminal nodes are dropped from our calculations. This truncation may disproportionately affect the

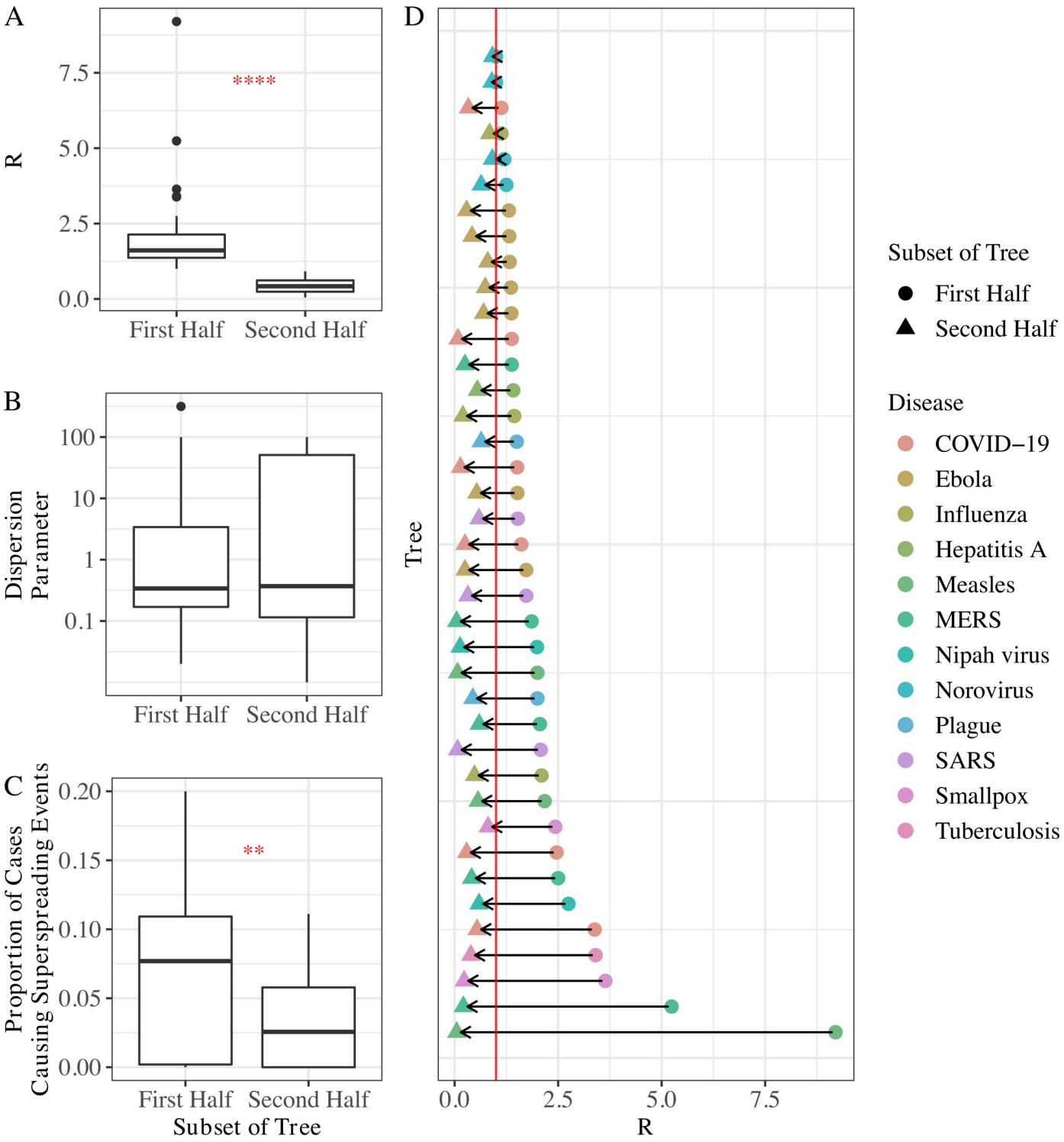

Fig 3. The time dependence of *R*, *k*, and the proportion of cases causing superspreading events. (A) *R* decreased significantly between the first and second halves of transmission trees. (B) *k* did not differ significantly between the first and second halves of transmission trees. Y-axis is on a log_{10} scale for visual aid. (C) The proportion of cases causing superspreading events decreased significantly between the first and second halves of transmission trees. (D) Decrease in *R* shown for each tree by disease. *R* was below 1 in the second half of all trees; red line denotes $R = 1$. The Wilcoxon rank test was used for all significance tests (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$), and results are shown in red stars. Trees were assumed to be complete and only trees with 20 or more cases and at least 2 generations of spread were used in these analyses. Results assuming tree incompleteness are shown in S3 Fig. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.

https://doi.org/10.1371/journal.pbio.3001685.g003

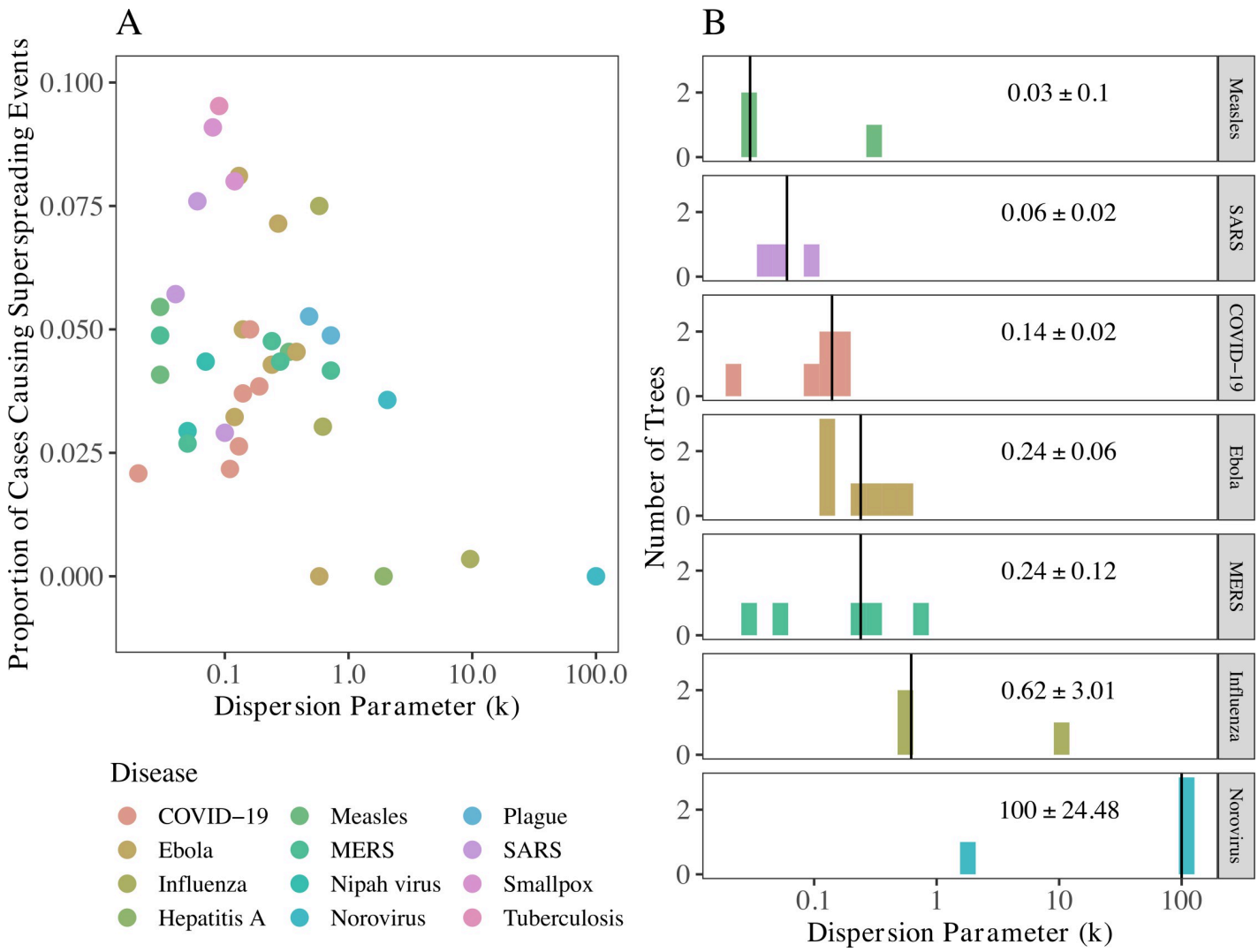

Fig 4. The importance and expected frequency of superspreading across diseases. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters, as predicted by theory [3]. (B) Dispersion parameter (k) of a negative binomial distribution fit to the offspring distribution of trees by disease (for diseases with at least 3 trees). Lower dispersion parameters are indicative of greater variation in number of secondary infections. Vertical line and value printed in each facet shows the median k and standard error for each disease. X-axes are on a log_{10} scale in both plots for visual aid. Trees were assumed to be complete and only trees with 20 or more cases and at least 2 generations of spread were used in these analyses. Other size cutoffs are shown in S4 and S5 Figs and results assuming tree incompleteness are shown in S6 Fig. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.

https://doi.org/10.1371/journal.pbio.3001685.g004

second half of a tree with many terminal nodes, decreasing the heterogeneity in the number of secondary infections, and increasing k. This analysis informs the following 2 analyses by indicating how frequently our trees may be capturing disease spread after interventions are imposed or behavior changes take place.

Superspreading characteristics across diseases. Consistent with theory proposed by [3], intermediate dispersion parameters gave rise to the highest proportion of cases causing superspreading events (Fig 4A). COVID-19 trees had a median dispersion parameter ($k = 0.14$) (Fig 4B) between that of SARS (0.06) and Middle East Respiratory Syndrome (MERS) (0.24). Six diseases had overdispersed offspring distributions (median $k<1$): measles, SARS, COVID-19, Ebola, MERS, and influenza. Norovirus was the only disease with median $k>1$. Dispersion

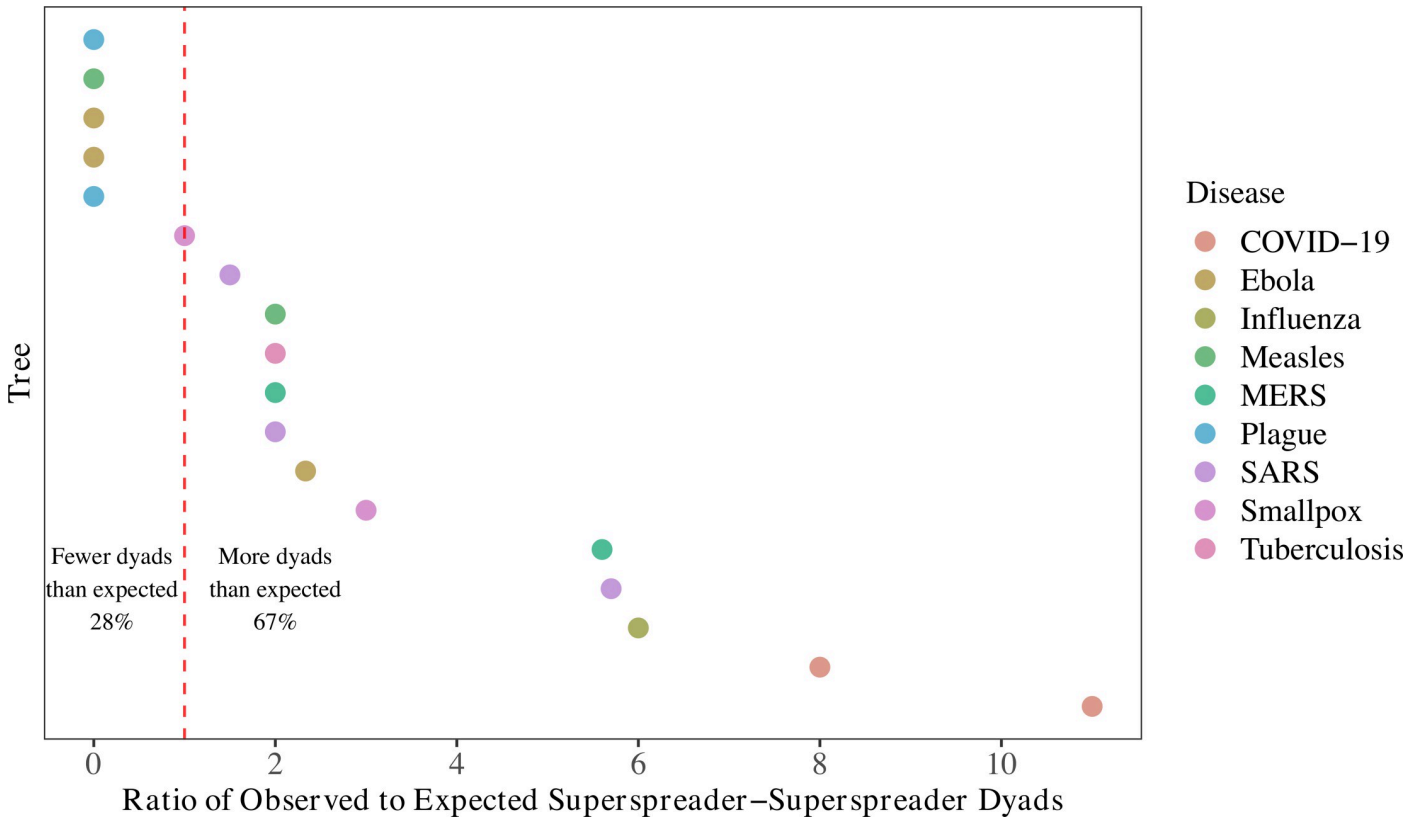

Fig 5. In two-thirds of transmission trees, superspreaders infect superspreaders more often than would be expected by chance. The expected number of superspreader-superspreader dyads was calculated by $\frac{s(s-1)}{S-t}$ for each tree, where s is the number of superspreaders in the tree, t is the number of terminal nodes (nodes that do not cause onward transmission), and S is tree size. Ratios larger than 1 indicate more superspreader-superspreader dyads were observed than would be expected by chance. This analysis was limited to trees with more than 1 superspreader, 20 or more cases, and 2 or more generations of spread. We assumed tree completeness here, but results assuming incompleteness are shown in S7 Fig. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.

https://doi.org/10.1371/journal.pbio.3001685.g005

parameter estimates calculated over all nodes tend to be lower than (or at the lower end) of values/ranges in the literature, while estimates calculated excluding all terminal nodes (shown in S6 Fig) tend to be higher than (or at the higher end) of values/ranges in the literature [3,17–31]. Given that our assumptions about tree completeness lie at opposite extremes, we expect the true outbreak dispersion parameters to fall between these extremes, which aligns well with the literature. The most notable exceptions are influenza, which is not typically associated with superspreading (though our median dispersion parameter estimate was less than 1), and norovirus, for which we could not find a previously published dispersion estimate. As observed with some of the large standard errors of k, and covered extensively in [16], these estimates are imprecise, especially when based on smaller trees. However, we observe little change in median dispersion parameter estimates or the relationship between dispersion parameter and proportion of cases causing superspreading events when we restrict the analysis to trees with at least 2 generations of spread and 10 or more cases (S4 Fig) or 30 or more cases (S5 Fig). Lack of follow-up in outbreak investigations may result in underreporting of onward transmission, affecting tree offspring distributions, and consequently, estimates of k.

Generation of superspreaders. The ratio of observed to expected superspreader-superspreader dyads, calculated by enumerating superspreader-superspreader pairs divided by all possible nonterminal infector–infectee pairs, was greater than 1 for 12 of 18 trees, indicating

that superspreaders infected other superspreaders more than would be expected by chance in two-thirds of eligible trees (Fig 5). Notably, both COVID-19 trees under consideration had large ratios of observed to expected superspreader-superspreader dyads. (Recall that we expect $\frac{s(s-1)}{S-t}$ dyads in a tree of size S with s superspreaders and t terminal nodes.) Despite most trees in our sample being small—29 of 39 trees have less than 50 cases—our observation of a large number of dyads suggests that this transmission pattern must be common. If we instead assume tree incompleteness, only 4 trees have enough superspreaders to compare ratios of observed to expected dyads (S7 Fig). Though additional information regarding the contexts in which superspreaders are infected would be required to understand these patterns, these results suggest some nonrandomness in generation of superspreaders providing preliminary support for our hypothesis that superspreaders infect other superspreaders.

Limitations of OutbreakTrees

While OutbreakTrees has allowed us to investigate questions about the nature of superspreading, the database has several limitations. First, trees in the database do not constitute a random nor necessarily representative sample of directly transmitted infectious disease outbreaks. For example, we omitted nearly 100 reported transmission events and trees due to lack of single infector identification, which limits the generalizability of our findings. Furthermore, as shown by Lloyd-Smith and colleagues [3], diseases with larger variation in offspring distributions have a greater chance of extinction. Early superspreading events may prevent extinction by increasing the size from which the outbreak grows and making infection propagation more likely [32]. The probability of detecting an outbreak may also be higher if there is a superspreading event because public health officials are more likely to investigate a cluster than an isolated case. Thus, the trees represented in our database are prone to both selection bias, in which outbreaks are noticed, and publication bias, in which outbreaks are published in an accessible format.

Second, although trees are meant to be complete representations of clusters (see Inclusion criteria), they are typically a subset from a larger chain of transmission events. For example, Ebola was likely only introduced once in the 2014 outbreak in West Africa, yet we have several separate trees because the transmission events could not all be connected. Moreover, outbreak investigations may miss cases, sometimes in random or consistent ways. For example, secondary cases with ambiguous infectors may be more readily attributed to superspreaders than their actual infectors, making it look like superspreaders accounted for more cases than they actually did. Or, as an outbreak continues, later cases may not be investigated in the same depth as earlier generations, underrepresenting the number of secondary infections produced by cases in later generations.

Third, control measures or behavior changes can alter parameters of disease spread in the middle of an outbreak. Due to limited available data, we have not included the timing of these events in the database, but they have the potential to affect every outbreak. For example, interventions may reduce the number and disperse the distribution of secondary infections caused by each individual. The scope of the database also does not include details about how each tree was constructed for publication. Reconstruction methods may be biased in different ways; methods focused on symptomatic cases may miss asymptomatic cases and transmission events. We were mindful of these biases and sought to examine how several key parameters change over the generations in our trees. These limitations should be kept in mind by others using the database for different purposes.

Usage notes

We have constructed the database so that other research groups may take advantage of this new resource, but we acknowledge that care and understanding of the limitations are required

for responsible analyses. Thus, we provide these recommendations for future users to encourage appropriate use and generalizable conclusions. We opted to include small trees in the database for the sake of completeness and to allow for the possibility of minor outbreak analysis in the future (e.g., [33]), but suggest that these smaller trees be excluded if users are seeking to calculate epidemiological quantities (as we did with a size cutoff of 20 individuals in our analyses). We also urge caution in viewing trees as absolute or complete. Several trees in the database are the result of probabilistic reconstruction, and so may represent only one possible way in which transmission may have occurred. Lack of ongoing transmission at the terminal nodes of a tree may be real but also could be due to lack of follow-up or investigation. While conclusions drawn from the database may be biased, they are no more biased than the original inferences drawn from the individual trees which compose the database. With these suggestions in mind, we hope that OutbreakTrees can be used to properly address new questions in the future.

Conclusions

In summary, we developed OutbreakTrees, an open-access database of infectious disease transmission trees, for research and public health officials. We illustrated how this database can be used to explore questions surrounding superspreader epidemiology, and we calculated a few important parameters for COVID-19 and examined their time dependence. In particular, we estimated the dispersion parameter from transmission trees and the value for COVID-19 was in between that of SARS and MERS. Additionally, our analysis provided tentative support for the theory that superspreaders generate other superspreaders. The development and release of OutbreakTrees highlights the benefits of data sharing and offers a new resource for epidemiologic research.

Supporting information

S1 File. Sources for transmission trees in OutbreakTrees which were used for this analysis.
(PDF)

S1 Fig. *R* values for each disease varied depending on calculation method. *R* values tended to be highest when calculated over nonterminal nodes and lowest when calculated over all nodes, with estimates based on early generation nodes (root and first generation nodes) falling somewhere in between. Nonterminal node estimates tended to be at the high end of literature values and early generation estimates at the low end, with estimates calculated over all nodes typically far below literature values [20,29,34–44], except for MERS and SARS which had low literature *R* estimates [3,21,30,45]. Analysis was limited to trees with 20 or more cases and at least 2 generations of spread and diseases with at least 3 trees that meet these criteria. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S2 Fig. Dispersion parameters were consistently higher when calculated over only nonterminal nodes versus all nodes in a tree. Dispersion parameter calculated over all nodes is on x-axis on *log*10 scale; dispersion parameter calculated over all nonterminal nodes is on y-axis on *log*10 scale. Dashed red line is y = x. Analysis was limited to trees with 20 or more cases and at least 2 generations of spread. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S3 Fig. The time dependence of *R*, *k*, and the proportion of cases causing superspreading events assuming trees are incomplete. (A) *R* decreased significantly between the first and second halves of transmission trees. (B) *k* increased significantly between the first and second halves of transmission trees. Seven of 39 trees had nonoptimizable degree distributions for the second half of the tree in each of 10 repetitions; these trees are excluded from this analysis and the boxplot. Y-axis is on a log_{10} scale for visual aid. (C) The proportion of cases causing superspreading events decreased significantly between the first and second halves of transmission trees. (D) While, on average, *R* decreased between first and second halves of trees, some trees had higher values of *R* in the second half of the tree than the first. Red line denotes $R = 1$. The Wilcoxon rank test was used for all significance tests (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$) and results are shown in red stars. Only trees with 20 or more cases and at least 2 generations of spread were used in these analyses. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S4 Fig. Proportion of cases causing superspreading events and dispersion parameter estimates do not differ considerably with cutoff of 10 or more cases. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters, as predicted by theory [3]. (B) Dispersion parameter (*k*) of a negative binomial distribution fit to the offspring distribution of trees by disease (for diseases with at least 3 trees). Lower dispersion parameters are indicative of greater variation in number of secondary infections. Vertical line and value printed in each facet shows the median *k* and standard error for each disease. X-axes are on a log_{10} scale in both plots for visual aid. Only trees with 10 or more cases and at least 2 generations of spread were used in these analyses, and trees were assumed to be complete. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S5 Fig. Proportion of cases causing superspreading events and dispersion parameter estimates do not differ considerably with cutoff of 30 or more cases, though fewer diseases are eligible for median dispersion parameter analysis. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters, as predicted by theory [3]. (B) Dispersion parameter (*k*) of a negative binomial distribution fit to the offspring distribution of trees by disease (for diseases with at least 3 trees). Lower dispersion parameters are indicative of greater variation in number of secondary infections. Vertical line and value printed in each facet shows the median *k* and standard error for each disease. X-axes are on a log_{10} scale in both plots for visual aid. Only trees with 30 or more cases and at least 2 generations of spread were used in these analyses, and trees were assumed to be complete. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S6 Fig. Peak proportion of cases causing superspreading events is observed at a higher dispersion parameter ($\approx$1), and dispersion parameter estimates are an order of magnitude higher when terminal nodes are excluded from dispersion parameter and *R* calculations than when terminal nodes are included. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters near 1, as opposed to the

range of 0.2 to 0.6, as predicted by theory for higher values of R [3]. (B) Dispersion parameter (k) of a negative binomial distribution fit to the offspring distribution of trees by disease (for diseases with at least 3 trees). Lower dispersion parameters are indicative of greater variation in number of secondary infections. SARS now has the lowest median dispersion parameter of 0.87, mildly overdispersed. MERS, Ebola, and influenza would no longer be considered over-dispersed. Vertical line and value printed in each facet shows the median k and standard error for each disease. X-axes are on a log_{10} scale in both plots for visual aid. Only trees with 20 or more cases and at least 2 generations of spread were used in these analyses. Terminal nodes were excluded from offspring distributions, i.e., trees were assumed to be incomplete. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. COVID-19, Coronavirus Disease 2019; MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S7 Fig. There are too few trees with 2 or more superspreaders to examine superspreader dyads when R is calculated excluding terminal nodes. The expected number of superspreader-superspreader dyads was calculated by $\frac{s(s-1)}{S-t}$ for each tree, where s is the number of superspreaders in the tree, t is the number of terminal nodes, and S is tree size. Ratios larger than 1 indicate more superspreader-superspreader dyads observed than would be expected by chance. This analysis was limited to trees with more than 1 superspreader, 20 or more cases, and 2 or more generations of spread. The data to reproduce this figure can be found at https://doi.org/10.5061/dryad.nk98sf7w7. MERS, Middle East Respiratory Syndrome; SARS, Severe Acute Respiratory Syndrome.
(PDF)

S1 Table. The mean proportion of nodes with complete attribute information when that attribute was listed as available for a given tree. Analysis was limited to 5 most common attributes in the database and trees with 20 or more cases and 2 or more generations of spread.
(PDF)

Acknowledgments

We thank E. Marty for assistance developing the online interface to OutbreakTrees.

Author Contributions

Conceptualization: John M. Drake.

Data curation: Juliana C. Taube.

Formal analysis: Juliana C. Taube, Paige B. Miller.

Funding acquisition: John M. Drake.

Investigation: Juliana C. Taube.

Methodology: Juliana C. Taube.

Project administration: John M. Drake.

Software: Paige B. Miller.

Supervision: Paige B. Miller, John M. Drake.

Validation: Paige B. Miller, John M. Drake.

Visualization: Juliana C. Taube, Paige B. Miller.

Writing – original draft: Juliana C. Taube, Paige B. Miller.

Writing – review & editing: Juliana C. Taube, Paige B. Miller, John M. Drake.

References

1. Teunis P, Heijne JC, Sukhrie F, van Eijkeren J, Koopmans M, Kretzschmar M. Infectious disease transmission as a forensic problem: Who infected whom? J R Soc Interface. 2013; 10(81):20120955. https://doi.org/10.1098/rsif.2012.0955 PMID: 23389896
2. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics. 2018; 19(Suppl 11)(363):17–24. https://doi.org/10.1186/s12859-018-2330-z PMID: 30343663
3. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438:355–9. https://doi.org/10.1038/nature04153 PMID: 16292310
4. Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, et al. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. PLoS Biol. 2020; 18(11):e3000897. https://doi.org/10.1371/journal.pbio.3000897 PMID: 33180773
5. Beldomenico PM. Do superspreaders generate new superspreaders? A hypothesis to explain the propagation pattern of COVID-19. Int J Infect Dis. 2020; 96:461–3. https://doi.org/10.1016/j.ijid.2020.05.025 PMID: 32422375
6. Susswein Z, Bansal S. Characterizing superspreading of SARS-CoV-2: From mechanism to measurement. medRxiv. 2020.
7. Leclerc QJ, Fuller NM, Knight LE, Funk S, Knight GM, Group CCW, et al. What settings have been linked to SARS-CoV-2 transmission clusters? [version 2; peer review: 2 approved]. Wellcome Open Res. 2020; 5(83).
8. Faye O, Boëlle PY, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: An observational study. Lancet Infect Dis. 2015; 15(3):320–6. https://doi.org/10.1016/S1473-3099(14)71075-8 PMID: 25619149
9. Komabayashi K, Seto J, Tanaka S, Suzuki Y, Ikeda T, Onuki N, et al. The largest measles outbreak, including 38 modified measles and 22 typical measles cases, Yamagata, Japan, 2017 in its elimination era. Jpn J Infect Dis. 2018; 71(6):413–8. https://doi.org/10.7883/yoken.JJID.2018.083 PMID: 29962488
10. Ohio Department of Health. COVID-19 Update: Masks in Schools, Rapid Testing, Community Spread and Spread from Faith-Based Settings, Dr. Amy Acton; 2020. Available from: https://coronavirus.ohio.gov/resources/news-releases-news-you-can-use/covid-19-update-08-04-20.
11. Vergne T, Fournié G, Markovich MP, Ypma RJ, Katz R, Shkoda I, et al. Transmission tree of the highly pathogenic avian influenza (H5N1) epidemic in Israel, 2015. Vet Res. 2016; 47(109).
12. 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(1637):887–95. https://doi.org/10.1098/rspb.2007.1442 PMID: 18230598
13. Glur C. data.tree: General Purpose Hierarchical Data Structure; 2019. Available from: https://CRAN.R-project.org/package=data.tree.
14. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. Complex Systems 1695; 2006.
15. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/.
16. Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE. 2007; 2(2):e180. https://doi.org/10.1371/journal.pone.0000180 PMID: 17299582
17. Becker AD, Birger RB, Teillant A, Gastanaduy PA, Wallace GS, Grenfell BT. Estimating enhanced pre-vaccination measles transmission hotspots in the context of cross-scale dynamics. Proc Natl Acad Sci U S A. 2016; 113(51):14595–600. https://doi.org/10.1073/pnas.1604976113 PMID: 27872300
18. Ackley SF, Hacker JK, Enanoria WT, Worden L, Blumberg S, Porco TC, et al. Genotype-specific measles transmissibility: A branching process analysis. Clin Infect Dis. 2018; 66(8):1270–5. https://doi.org/10.1093/cid/cix974 PMID: 29228134

19. Blumberg S, Worden L, Enanoria W, Ackley S, Deiner M, Liu F, et al. Assessing measles transmission in the United States following a large outbreak in California. PLOS Currents. 2015;7. https://doi.org/10.1371/currents.outbreaks.b497624d7043b1aecfbfd3dfda3e344a PMID: 26052471

20. Nishiura H, Mizumoto K, Asai Y. Assessing the transmission dynamics of measles in Japan, 2016. Epidemics. 2017; 20:67–72. https://doi.org/10.1016/j.epidem.2017.03.005 PMID: 28359662

21. Chowell G, Abdirizak F, Lee S, Lee J, Jung E, Nishiura H, et al. Transmission characteristics of MERS and SARS in the healthcare setting: A comparative study. BMC Med. 2015; 13(210). https://doi.org/10.1186/s12916-015-0450-0 PMID: 26336062

22. Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu W, et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nat Commun. 2020; 11(5006). https://doi.org/10.1038/s41467-020-18836-4 PMID: 33024095

23. Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res. 2020; 5(67). https://doi.org/10.12688/wellcomeopenres.15842.3 PMID: 32685698

24. Adam DC, Wu P, Wong JY, Lau EH, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nat Med. 2020; 26:1714–9. https://doi.org/10.1038/s41591-020-1092-0 PMID: 32943787

25. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. Lancet Infect Dis. 2020; 20(8):911–9. https://doi.org/10.1016/S1473-3099(20)30287-5 PMID: 32353347

26. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Ann Intern Med. 2020; 172(9):577–82. https://doi.org/10.7326/M20-0504 PMID: 32150748

27. Ajelli M, Parlamento S, Bome D, Kebbi A, Atzori A, Frasson C, et al. The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: Epidemiology and impact of interventions. BMC Med. 2015; 13(281). https://doi.org/10.1186/s12916-015-0524-z PMID: 26607790

28. Althaus CL. Ebola superspreading. Lancet Infect Dis. 2015; 15(5):507–8. https://doi.org/10.1016/S1473-3099(15)70135-0 PMID: 25932579

29. Lau MS, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. Proc Natl Acad Sci U S A. 2017; 114(9):2337–42. https://doi.org/10.1073/pnas.1614595114 PMID: 28193880

30. Kucharski A, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eur Secur. 2015; 20(25):21167. https://doi.org/10.2807/1560-7917.es2015.20.25.21167 PMID: 26132768

31. Brugger J, Althaus CL. Transmission of and susceptibility to seasonal influenza in Switzerland from 2003 to 2015. Epidemics. 2020; 30:100373. https://doi.org/10.1016/j.epidem.2019.100373 PMID: 31635972

32. Meyers LA, Pourbohloul B, Newman ME, Skowronski DM, Brunham RC. Network theory and SARS: Predicting outbreak diversity. J Theor Biol. 2005; 232(1):71–81. https://doi.org/10.1016/j.jtbi.2004.07.026 PMID: 15498594

33. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. J Theor Biol. 2012; 294:48–55. https://doi.org/10.1016/j.jtbi.2011.10.039 PMID: 22079419

34. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 2020; 382(13):1199–207. https://doi.org/10.1056/NEJMoa2001316 PMID: 31995857

35. Locatelli I, Trächsel B, Rousson V. Estimating the basic reproduction number for COVID-19 in Western Europe. PLoS ONE. 2021; 16(3):e0248731. https://doi.org/10.1371/journal.pone.0248731 PMID: 33730041

36. Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis. 2020; 92:214–7. https://doi.org/10.1016/j.ijid.2020.01.050 PMID: 32007643

37. Read JM, Bridgen JR, Cummings DA, Ho A, Jewell CP. Novel coronavirus 2019-nCoV (COVID-19): Early estimation of epidemiological parameters and epidemic size estimates. Philos Trans R Soc Lond B Biol Sci. 1829; 2021(376):20200265.

38. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. J Theor Biol. 2004; 229(1):119–26. https://doi.org/10.1016/j.jtbi.2004.03.006 PMID: 15178190

39. Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. Understanding the dynamics of Ebola epidemics. Epidemiol Infect. 2007; 135(4):610–21. https://doi.org/10.1017/S0950268806007217 PMID: 16999875

40. World Health Organization. Transmission dynamics and impact of pandemic influenza A (H1N1) 2009 virus. Wkly Epidemiol Rec. 2009; 84(46):481–4. PMID: 19928298

41. Guerra FM, Bolotin S, Lim G, Heffernan J, Deeks SL, Li Y, et al. The basic reproduction number (R0) of measles: A systematic review. Lancet Infect Dis. 2017; 17(12):e420–8. https://doi.org/10.1016/S1473-3099(17)30307-9 PMID: 28757186

42. O'Dea EB, Pepin KM, Lopman BA, Wilke CO. Fitting outbreak models to data from many small norovirus outbreaks. Epidemics. 2014; 6:18–29. https://doi.org/10.1016/j.epidem.2013.12.002 PMID: 24593918

43. Zelner J, Adams C, Havumaki J, Lopman B. Understanding the importance of contact heterogeneity and variable infectiousness in the dynamics of a large norovirus outbreak. Clin Infect Dis. 2020; 70 (3):493–500. https://doi.org/10.1093/cid/ciz220 PMID: 30901030

44. Sukhrie FH, Teunis P, Vennema H, Copra C, Thijs Beersma MF, Bogerman J, et al. Nosocomial transmission of norovirus is mainly caused by symptomatic cases. Clin Infect Dis. 2012; 54(7):931–7. https://doi.org/10.1093/cid/cir971 PMID: 22291099

45. 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. 2014; 14(1):50–6. https://doi.org/10.1016/S1473-3099(13)70304-9 PMID: 24239323