Code to reproduce figures in 'An open-access database of infectious disease transmission trees to explore superspreader epidemiology'

Juliana C. Taube* Paige B. Miller† John M. Drake‡

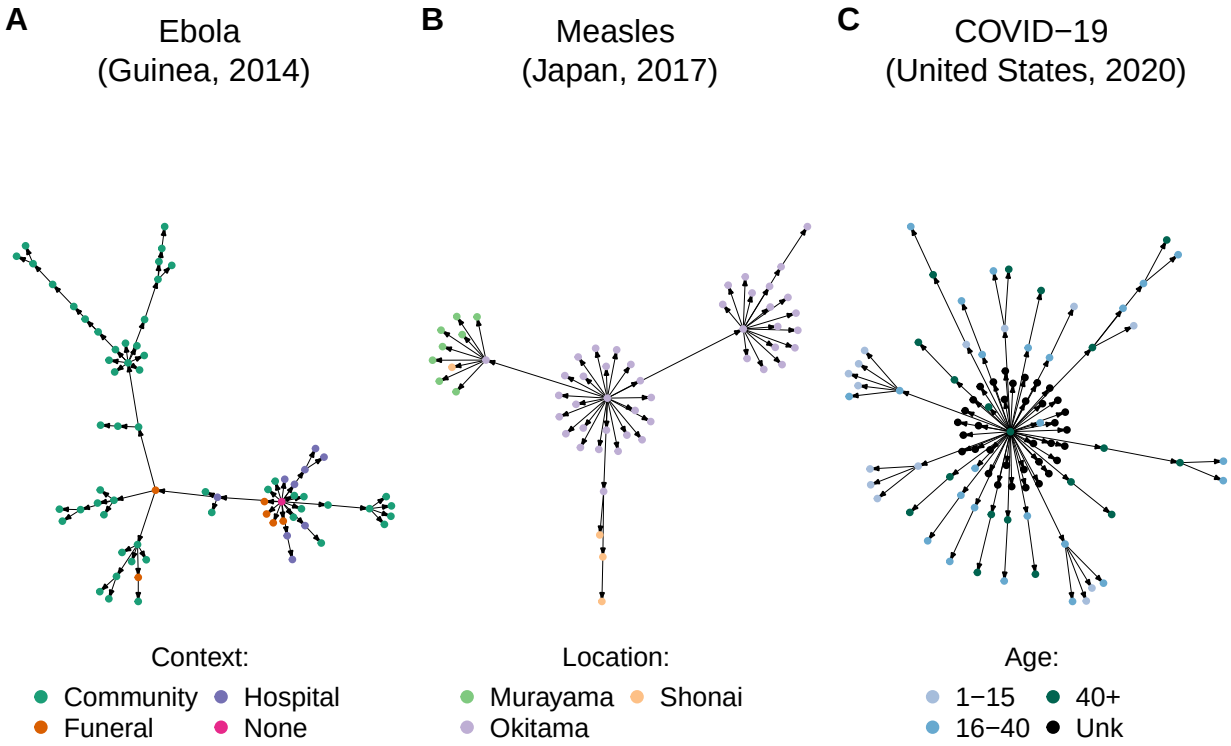


Figure 1. We compiled infectious disease transmission trees from the literature along with reported attribute information. Shown here are example trees in the database. (A) Ebola spread in different contexts (Faye et al., 2015). (B) Measles spread in different locations (Komabayashi et al., 2018). (C) COVID-19 spread among age classes (https://coronavirus.ohio.gov/wps/portal/gov/covid-19/resources/news-releases-news-you-can-use/covid-19-update-08-04-20). Primary sources for transmission trees are available in `OutbreakTrees` and listed in the Supplemental Material. `OutbreakTrees` may be accessed online at http://outbreaktrees.ecology.uga.edu.

*Department of Mathematics, Bowdoin College, Brunswick, ME, USA, taubejc@gmail.com
†Odum School of Ecology, University of Georgia, Athens, GA, USA, paigemiller554@gmail.com
‡Odum School of Ecology, University of Georgia, Athens, GA, USA, jdrake@uga.edu

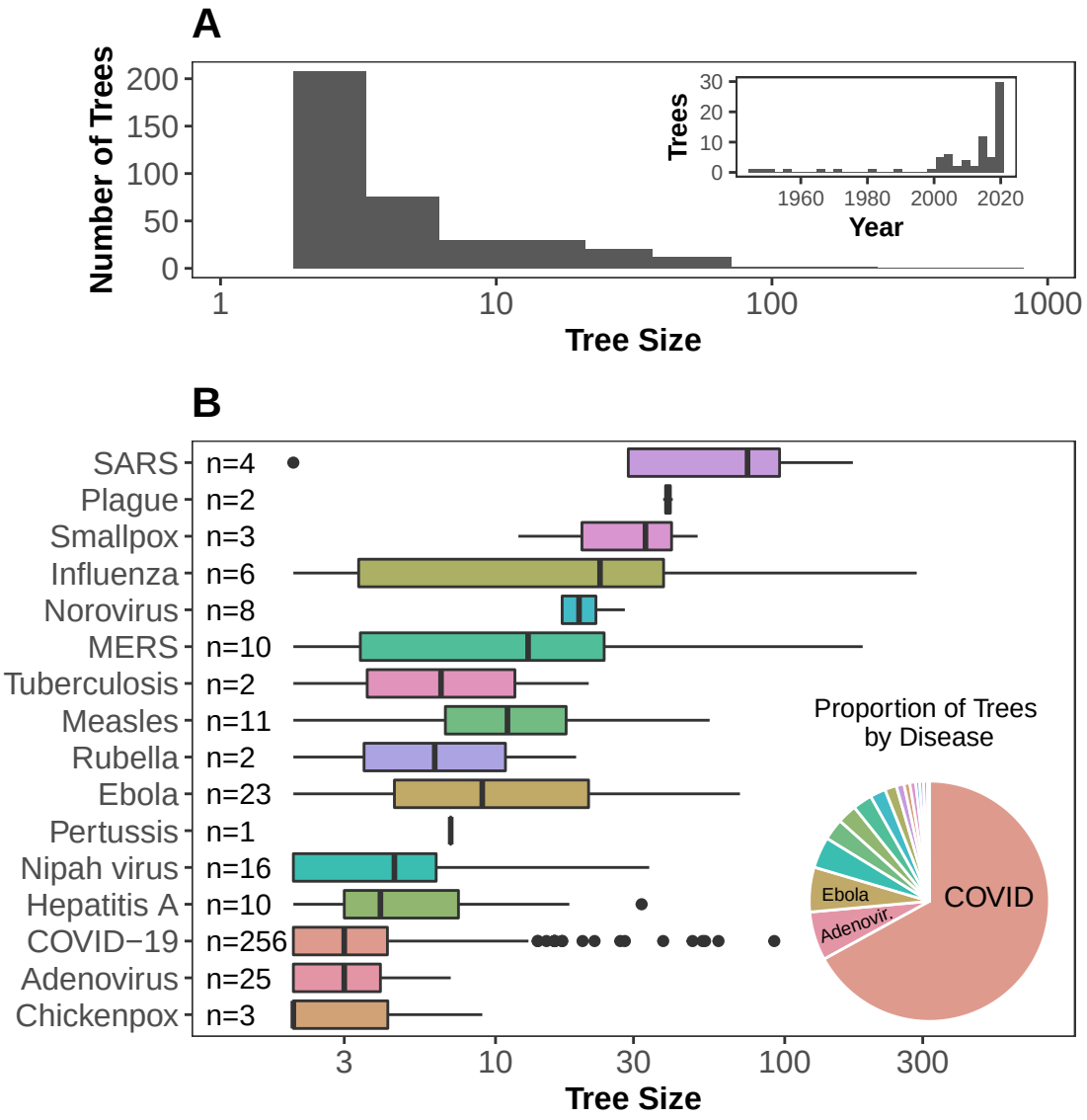


Figure 2. Characteristics of transmission trees in `OutbreakTrees`. (A) Tree size varies from 2 to 286 with a median of 3 and most trees represent outbreaks taking place in the past 20 years (only trees with 10 or more cases shown in date plot due to large number of small COVID-19 trees from 2020). (B) 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.

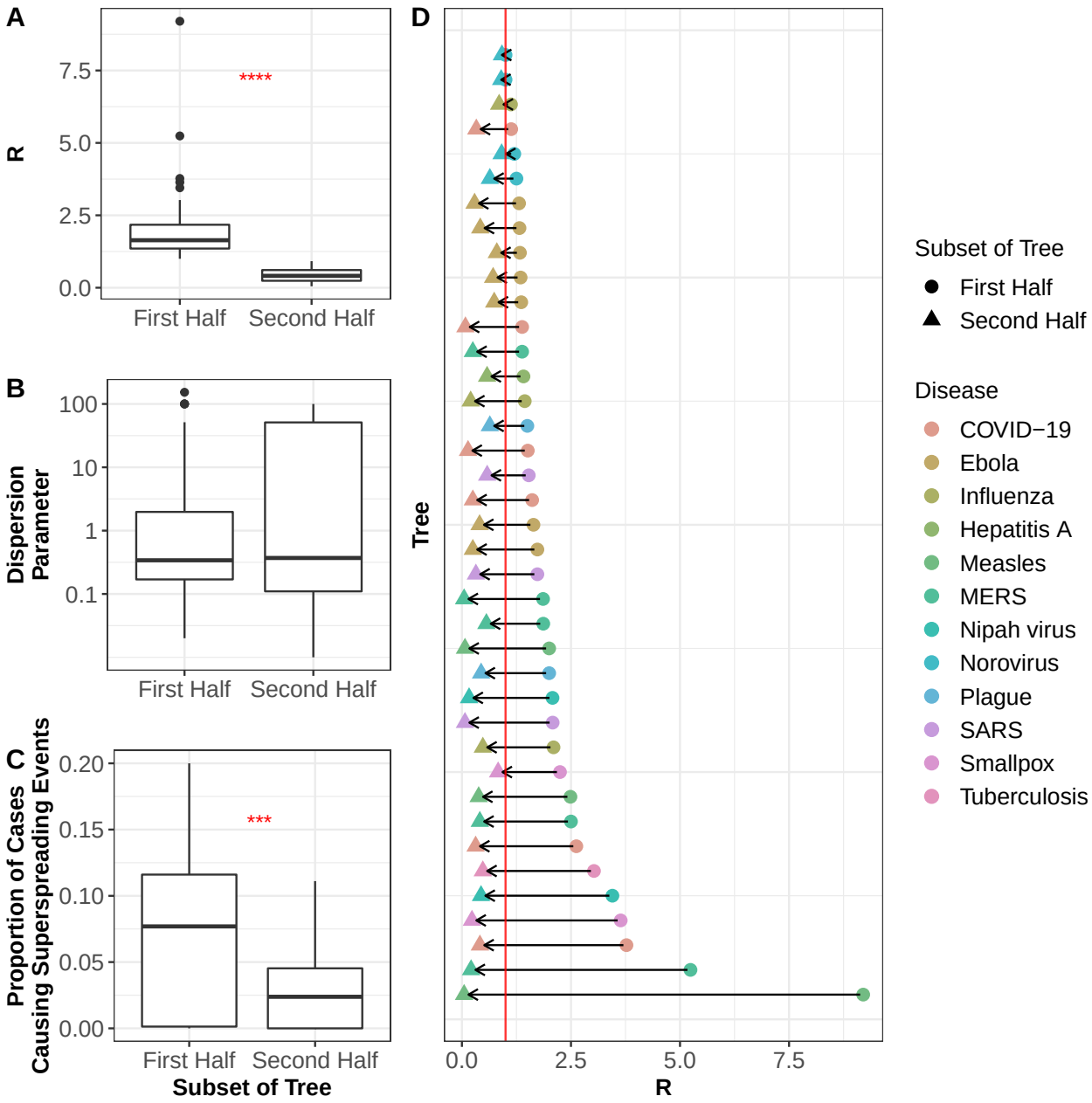


Figure 3. The time dependence of R, k, and the proportion of cases causing superspreading events. (A) R decreased significantly between the first and second halves of transmission trees. (B) k did not differ significantly between the first and second halves of transmission trees. Y-axis is on a log_{10} 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 (*: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ***: $p <= 0.0001$) 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.

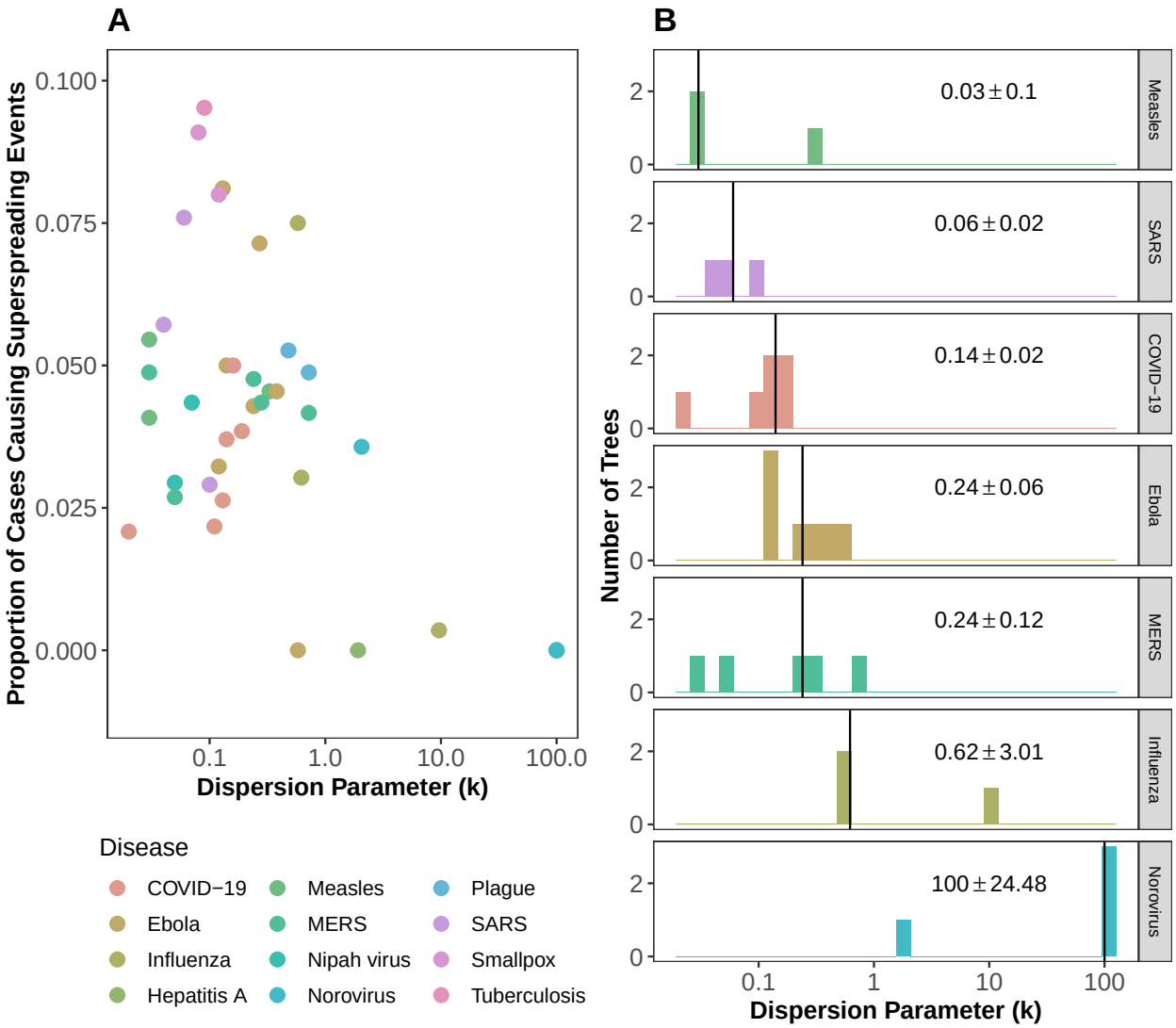


Figure 4. The importance and expected frequency of superspreading across diseases. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters, as predicted by theory (Lloyd-Smith et al., 2005). (B) Dispersion parameter (k) of a negative binomial 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 Fig and S5 Fig 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.

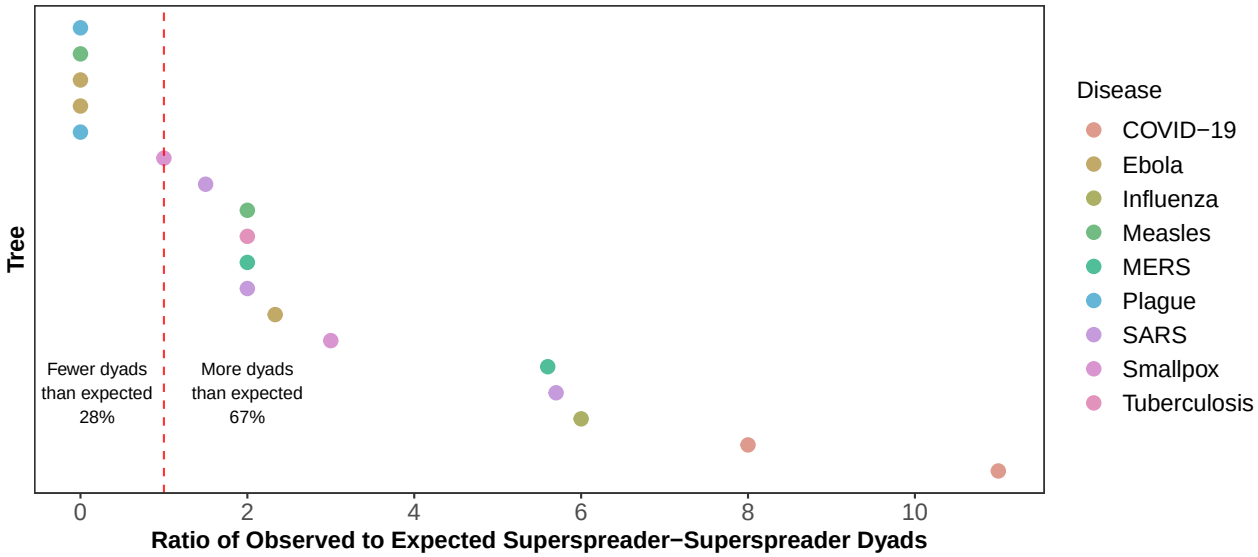
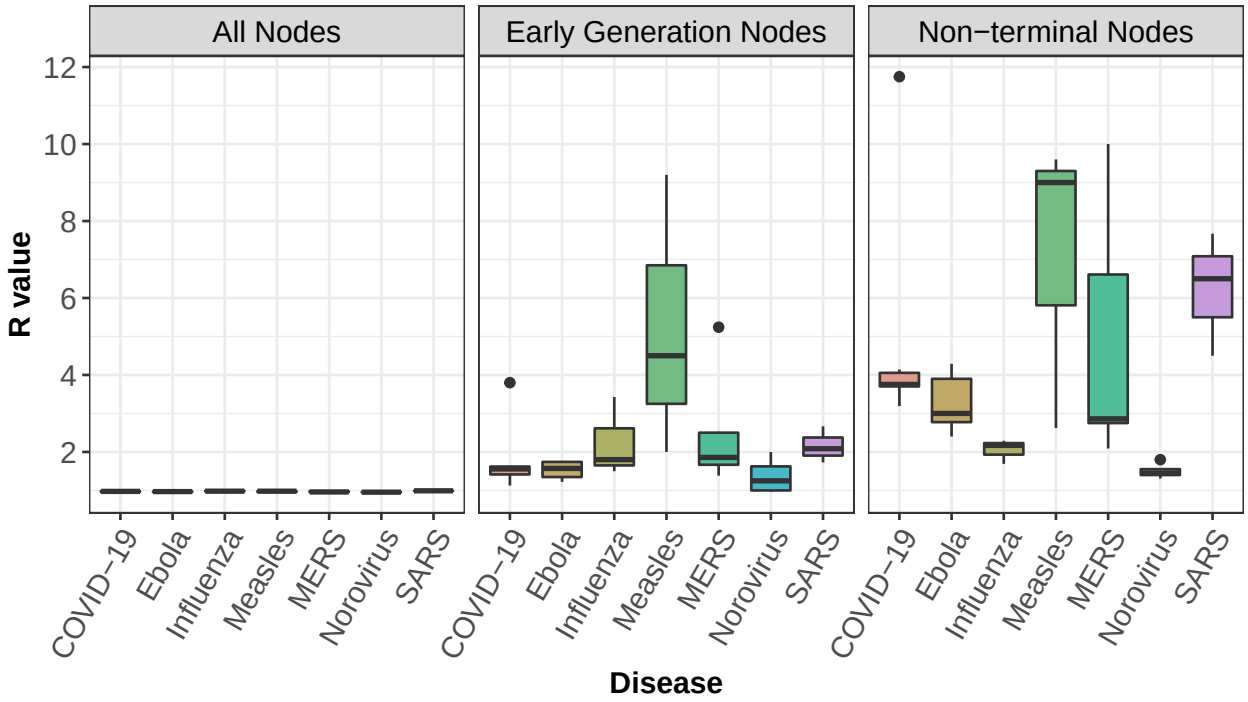
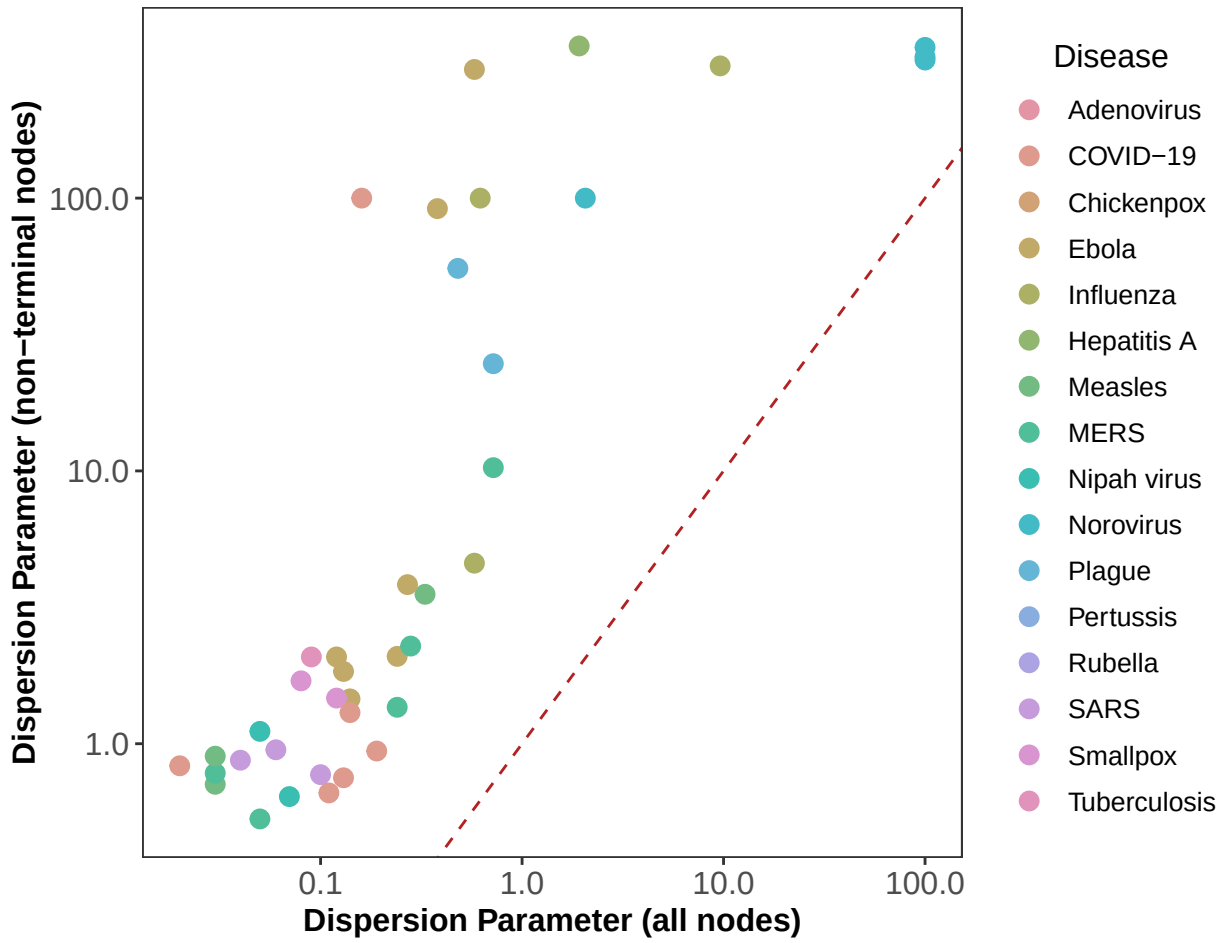


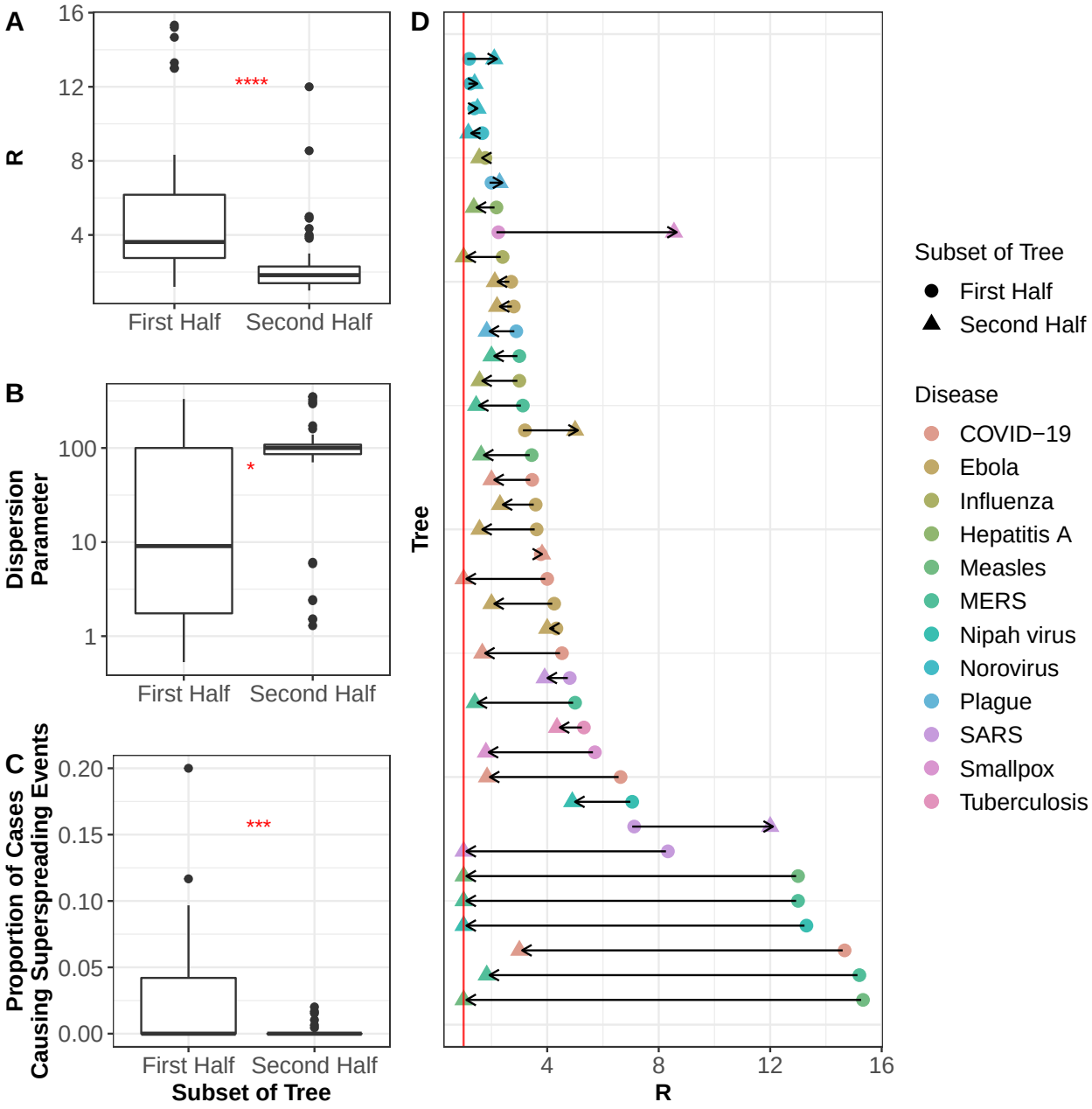
Figure 5. In two thirds of transmission trees, superspreaders infect superspreaders more often than would be expected by chance. The expected number of superspreader-superspreader dyads was calculated by $s(s-1)/(S-t)$ for each tree, where s is the number of superspreaders in the tree, t is the number 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 one 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.



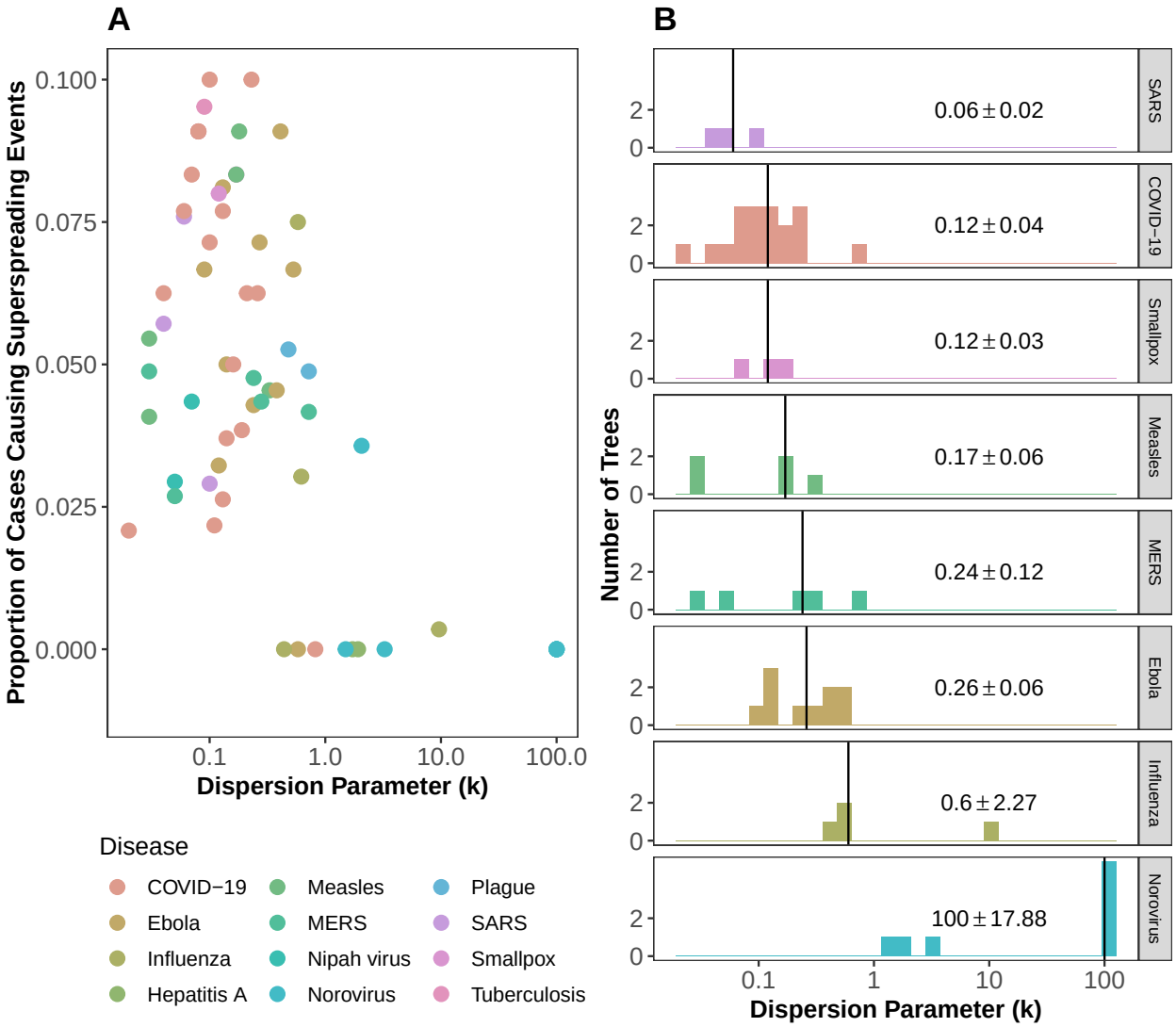
S1 Figure. R values for each disease varied depending on calculation method. R values tended to be highest when calculated over non-terminal nodes, and lowest when calculated over all nodes, with estimates based on early generation nodes (root and first generation nodes) falling somewhere in between. Non-terminal 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 (Li et al., 2020; Locatelli et al., 2021; Zhao et al., 2020; Read et al., 2021; Chowell et al., 2004; Legrand et al., 2007; Lau et al., 2017; WHO, 2009; Nishiura et al., 2017; Guerra et al., 2017; O'Dea et al., 2014; Zelner et al., 2020; Sukhrie et al., 2012), except for MERS and SARS which had low literature R estimates (Kucharski et al., 2015; Chowell et al., 2015; Cauchemez et al., 2014; Lloyd-Smith et al., 2005). 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.



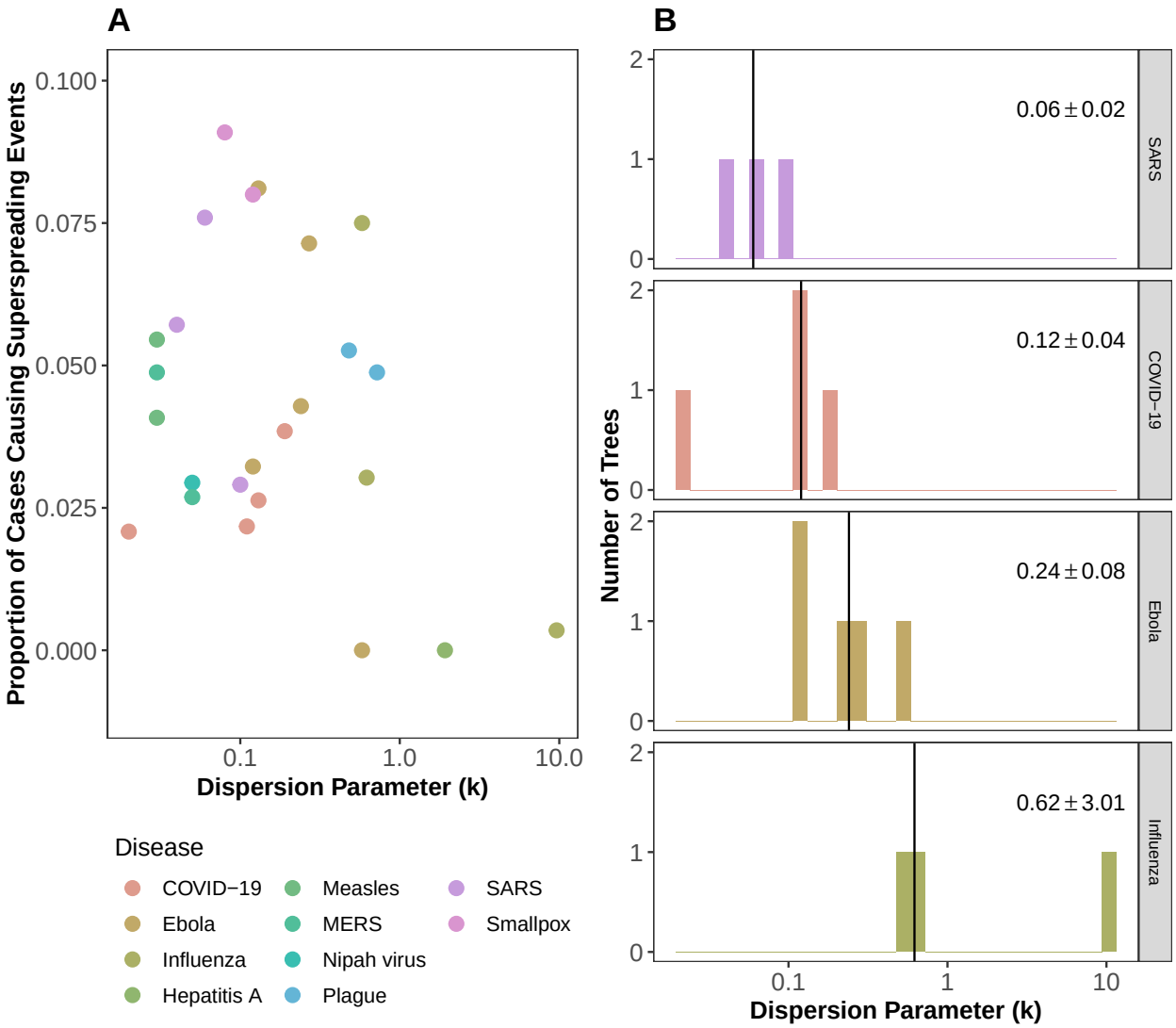
S2 Figure. Dispersion parameters were consistently higher when calculated over only non-terminal nodes versus all nodes in a tree. Dispersion parameter calculated over all nodes is on x-axis on log_{10} scale; dispersion parameter calculated over all non-terminal nodes is on y-axis on log_{10} scale. 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.



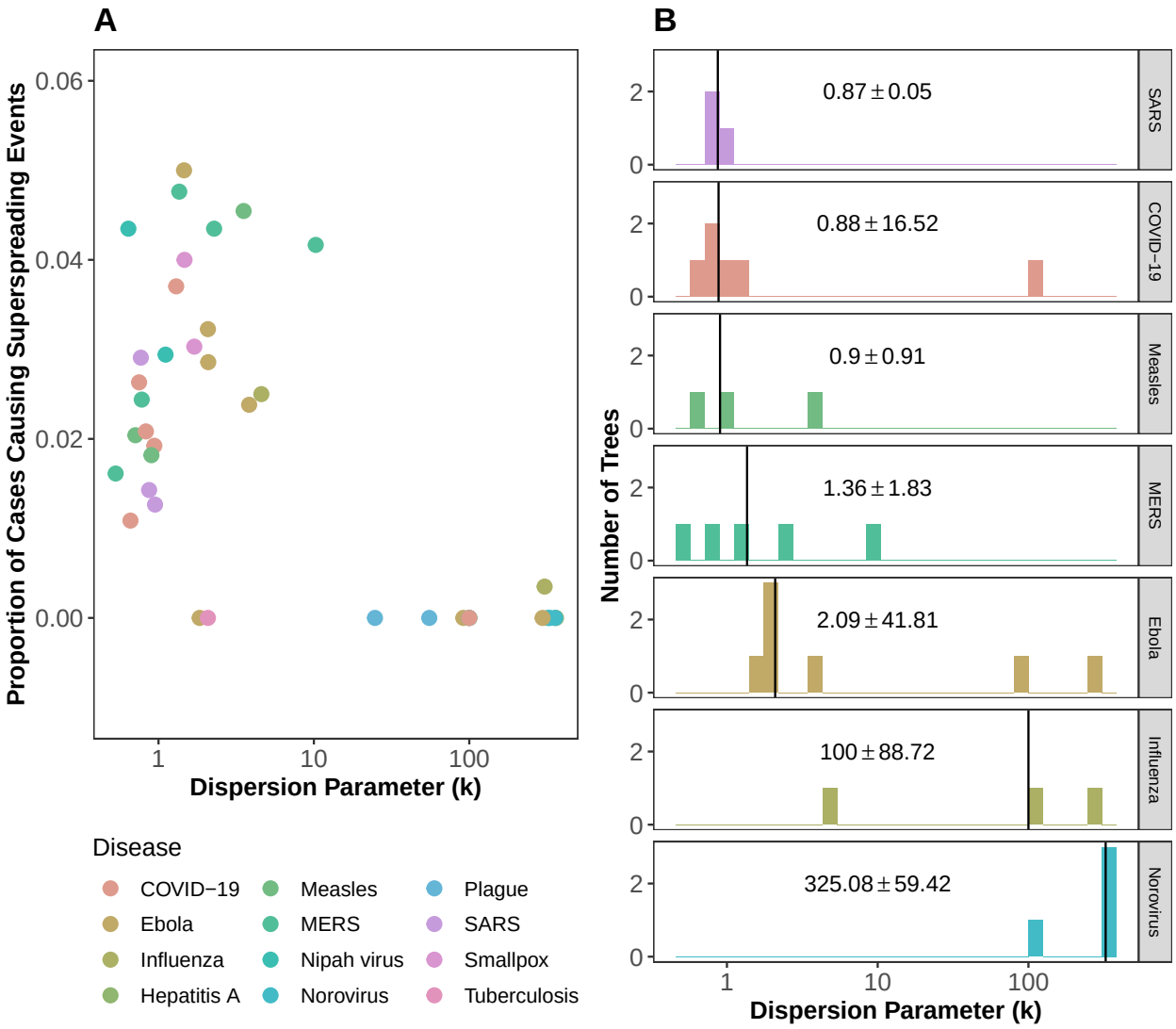
S3 Figure. The time dependence of R, k, and the proportion of cases causing superspreading events assuming trees are incomplete. (A) R decreased significantly between the first and second halves of transmission trees. (B) k increased significantly between the first and second halves of transmission trees. Seven of 39 trees had non-optimizable 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 (*: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ****: $p <= 0.0001$) 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.



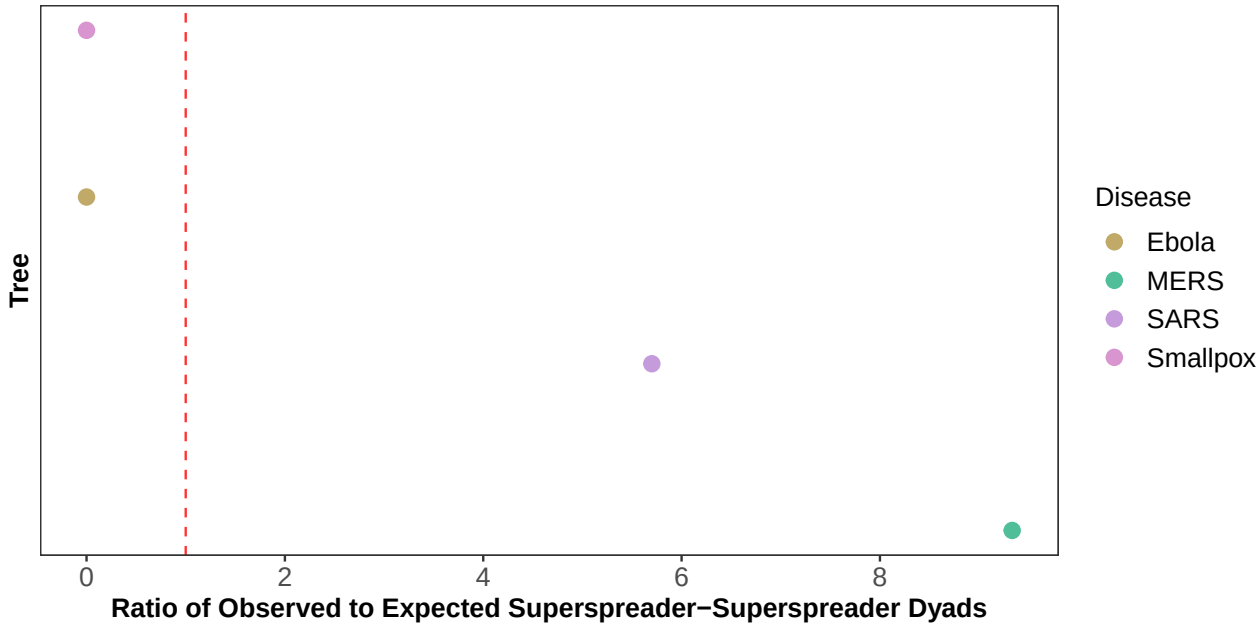
S4 Figure. Proportion of cases causing superspreading events and dispersion parameter estimates do not differ considerably with cutoff of 10 or more cases. (A) The highest proportion of cases causing superspreading events is observed at intermediate dispersion parameters, as predicted by theory (Lloyd-Smith et al, 2005). (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.



S5 Figure. 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 (Lloyd-Smith et al, 2005). (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.



S6 Figure. Peak proportion of cases causing superspreading events is observed at a higher dispersion parameter (~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 (Lloyd-Smith et al., 2005). (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 overdispersed. 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.



S7 Figure. 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 $s(s-1)/(S-t)$ 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.