Who Infected Whom? Creating a Database of Transmission Trees for Comparative Outbreak Analysis

Juliana Taube[1], Paige B. Miller[2], and John M. Drake[2]

[1] Bowdoin College, Brunswick, ME; [2] Odum School of Ecology, University of Georgia, Athens, GA

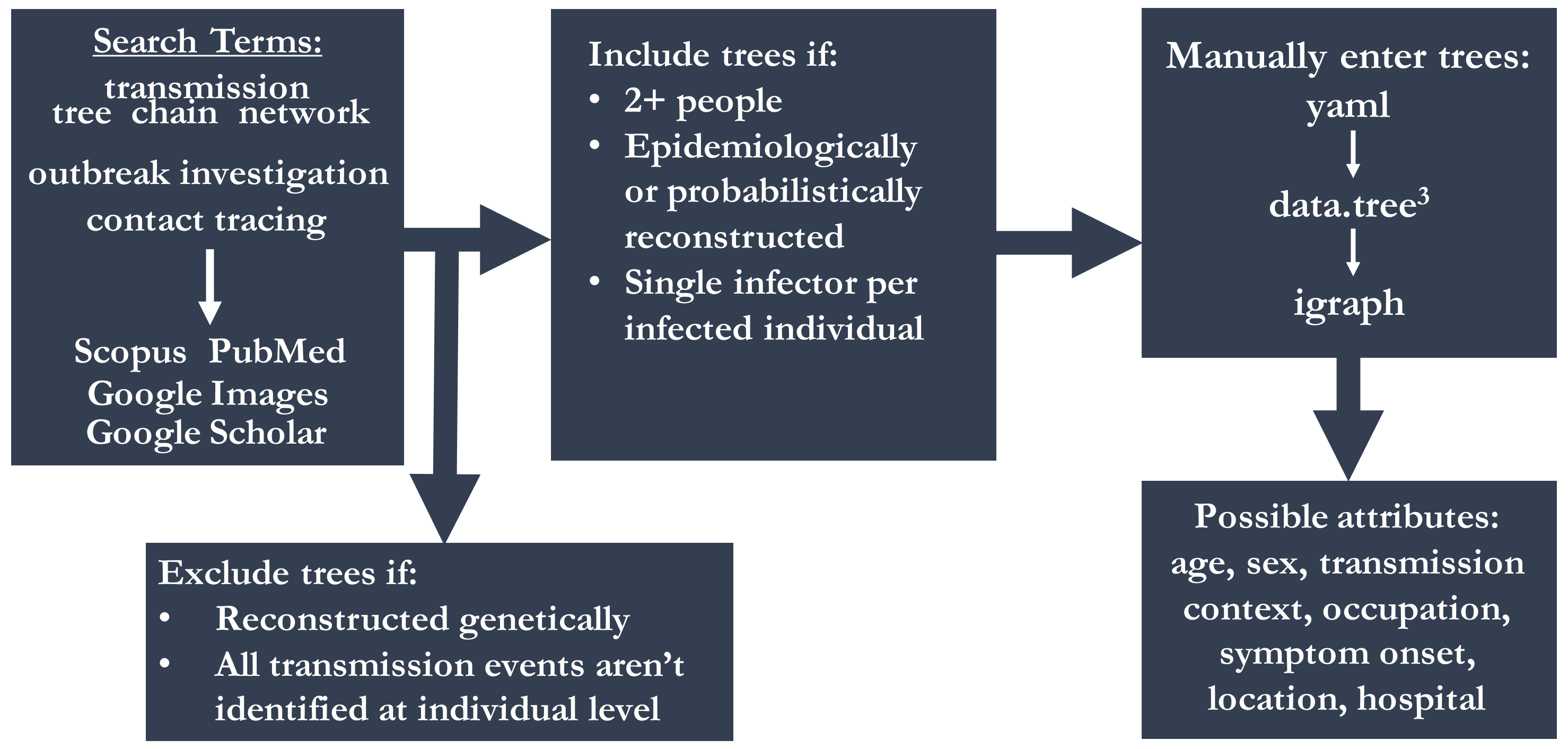
ABSTRACT

Transmission trees contain valuable details about who infected whom in infectious disease outbreaks. Here, we created a database with 81 published, standardized transmission trees consisting of 12 directly-transmitted pathogens (mostly viruses). We also demonstrated how the database could be used to help answer research questions in infectious disease epidemiology. First, we analyzed overall and pathogen-specific patterns between tree parameters (R_0 and variation in secondary infections). We found that outbreak size is nonlinearly associated with R_0 and the dispersion parameter, but emphasize that pathogen-specific patterns and intervention efforts may alter theoretical relationships between these variables. Second, we examined how superspreader contribution to onward transmission, either directly or through their tree descendants, varies across pathogens. Superspreaders were responsible for most cases via their descendants and the number of superspreaders varied across pathogens. Additional database exploration matched theory[6] about how the proportion of superspreaders increases at intermediate levels of dispersion, an idea that should be further explored. We hope that our database will assist both theoretical and applied infectious disease epidemiology research.

INTRODUCTION

When an infectious disease outbreak occurs, epidemiologists undergo time and money-intensive investigations to determine how the outbreak started and the patterns of disease transmission. They often store this information in a transmission tree, where individuals are represented by nodes, and disease transmission by branches. From these trees, one can calculate key statistics like R_0, the dispersion parameter (variation in R_0), pathogen mutation rate, and intervention efficacy, though greater standardization in tree format would make the trees and statistics more comparable. Additionally, a better understanding of predictors of outbreak size[7,9] and the importance of superspreaders to onward transmission across different pathogens[6,8] would help researchers developing transmission tree reconstruction methods[4,5] and inform public health intervention efforts. This project is a first attempt at standardizing and compiling transmission trees into an open-access database that can be a resource for future research.

DATABASE CONSTRUCTION



Search Terms:
transmission
tree chain network
outbreak investigation
contact tracing
Scopus PubMed
Google Images
Google Scholar

Include trees if:
- 2+ people
- Epidemiologically or probabilistically reconstructed
- Single infector per infected individual

Exclude trees if:
- Reconstructed genetically
- All transmission events aren't identified at individual level

Manually enter trees:
yaml
data.tree[3]
igraph

Possible attributes:
age, sex, transmission context, occupation, symptom onset, location, hospital

TREE ANALYSIS

- Average individual reproductive number (R_0): average number of secondary infections across all individuals in the tree
- Initial R_0: average number of secondary infections in the first two generations of the outbreak, the second generation may have zero individuals
- Secondary infections were assumed to follow a negative binomial distribution ($\mathbb{P}(X=k) = \binom{k+r-1}{k} p^k (1-p)^r$) and dispersion parameters were estimated using maximum likelihood methods (MASS package, fitdistr, no initial values given)
- Superspreaders[6]: cases who transmitted to more individuals than the 99th percentile of a Poisson(R_0) ($\mathbb{P}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$)
- Cases caused directly by superspreaders were defined as the individuals personally infected by the superspreader. Overall cases were defined as all individuals for whom their common infection "ancestor" was the superspreader. They were infected as a result of the superspreader infecting others, even if several generations later.

DATABASE SUMMARY

Total Pathogens	Total Trees
12	81

Pathogen	Number of Trees	Pathogen	Number of Trees
Ebola	22	Norovirus	8
Influenza	5	Plague	2
Hepatitis A	9	Pertussis	1
Measles	5	Rubella	2
MERS	4	SARS	4
Nipah	16	Smallpox	3

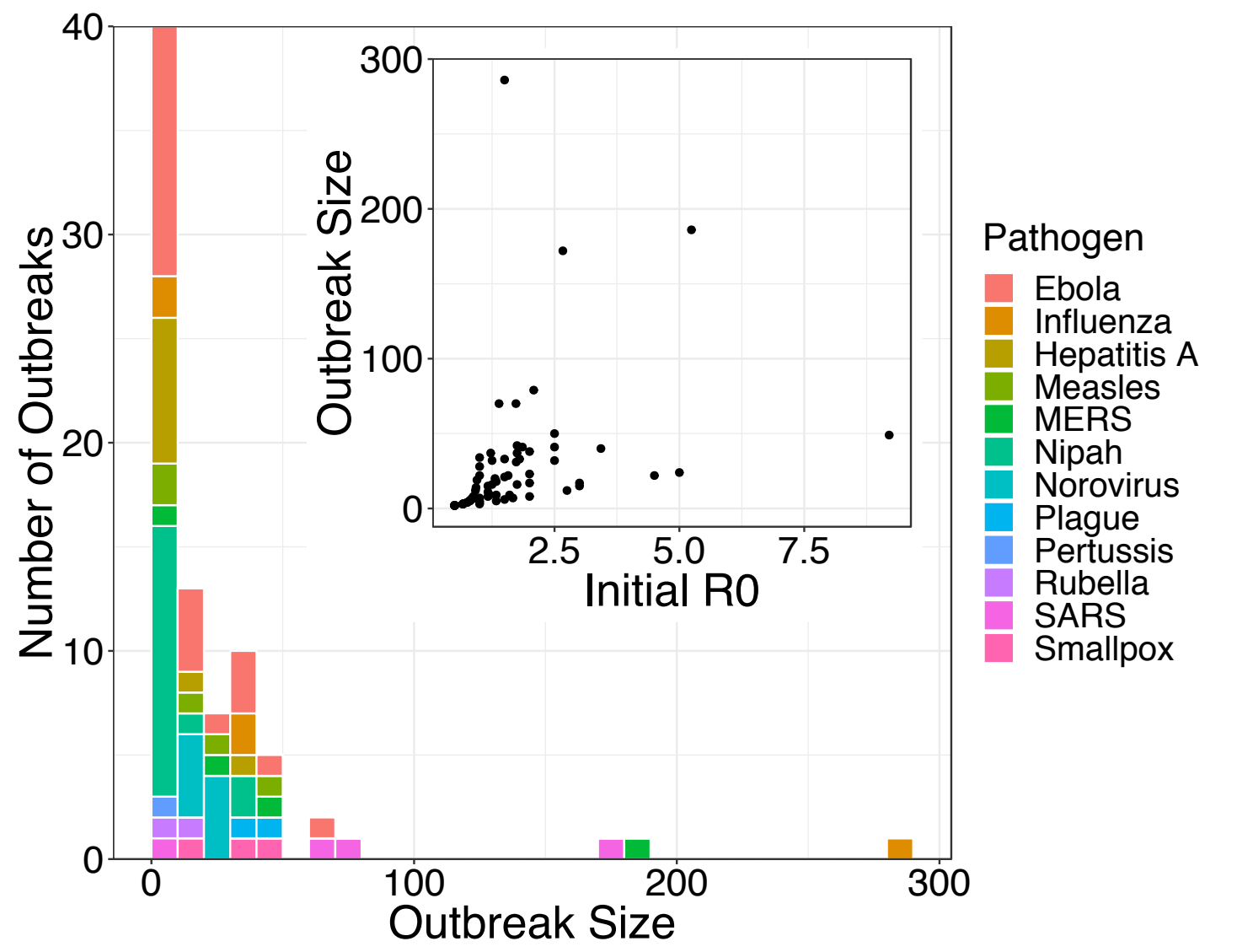
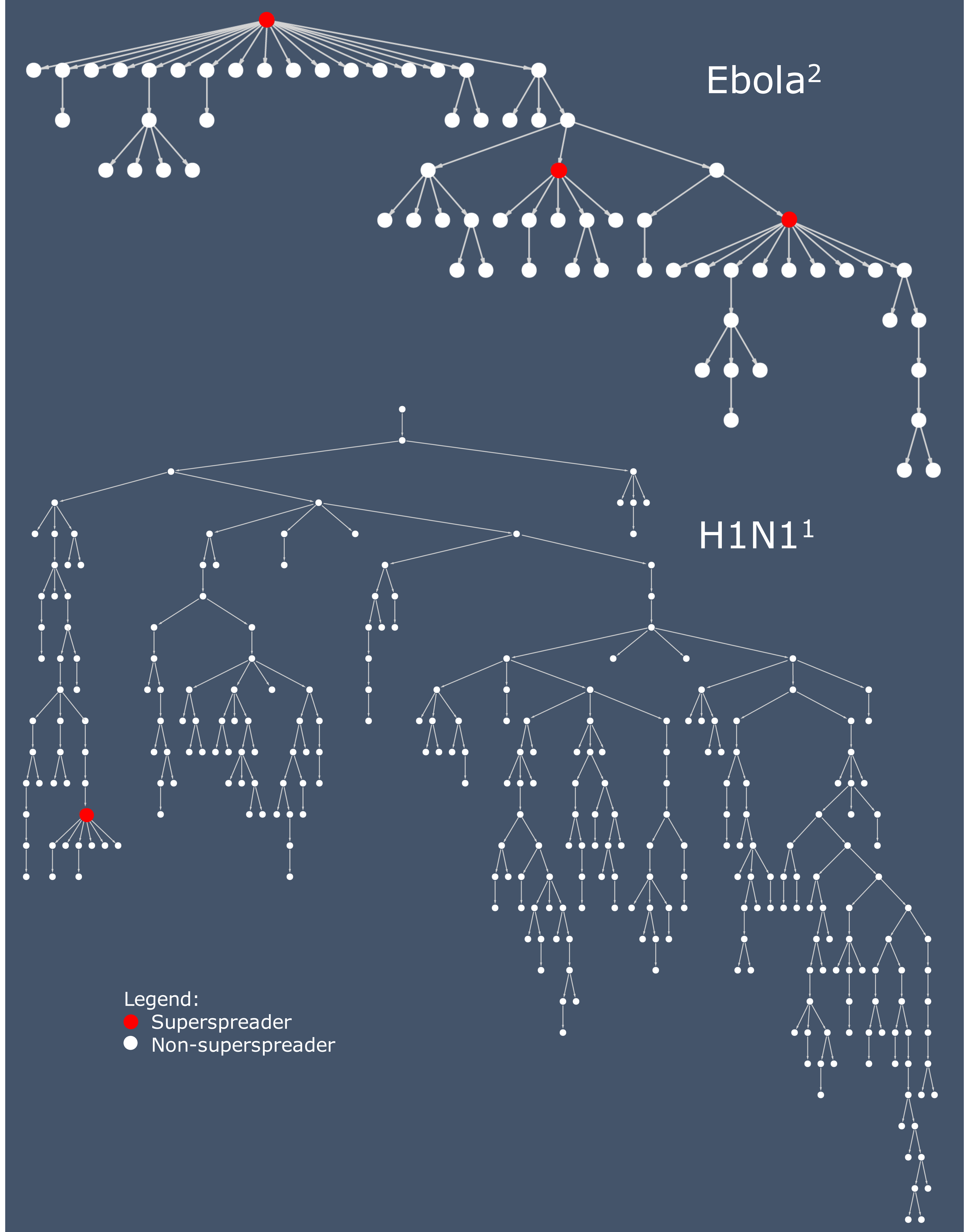


Figure 1. **Outbreak size distribution ranged from 2 to 286 and was skewed towards smaller trees.** Three trees had more than 100 cases: outbreaks of SARS, MERS, and influenza. Inset: Outbreak size is correlated with R_0 (r = 0.838, Spearman).



EXAMPLE QUESTIONS

Question 1: What is the relationship between individual variation in number of secondary infections, R_0, and total outbreak size?

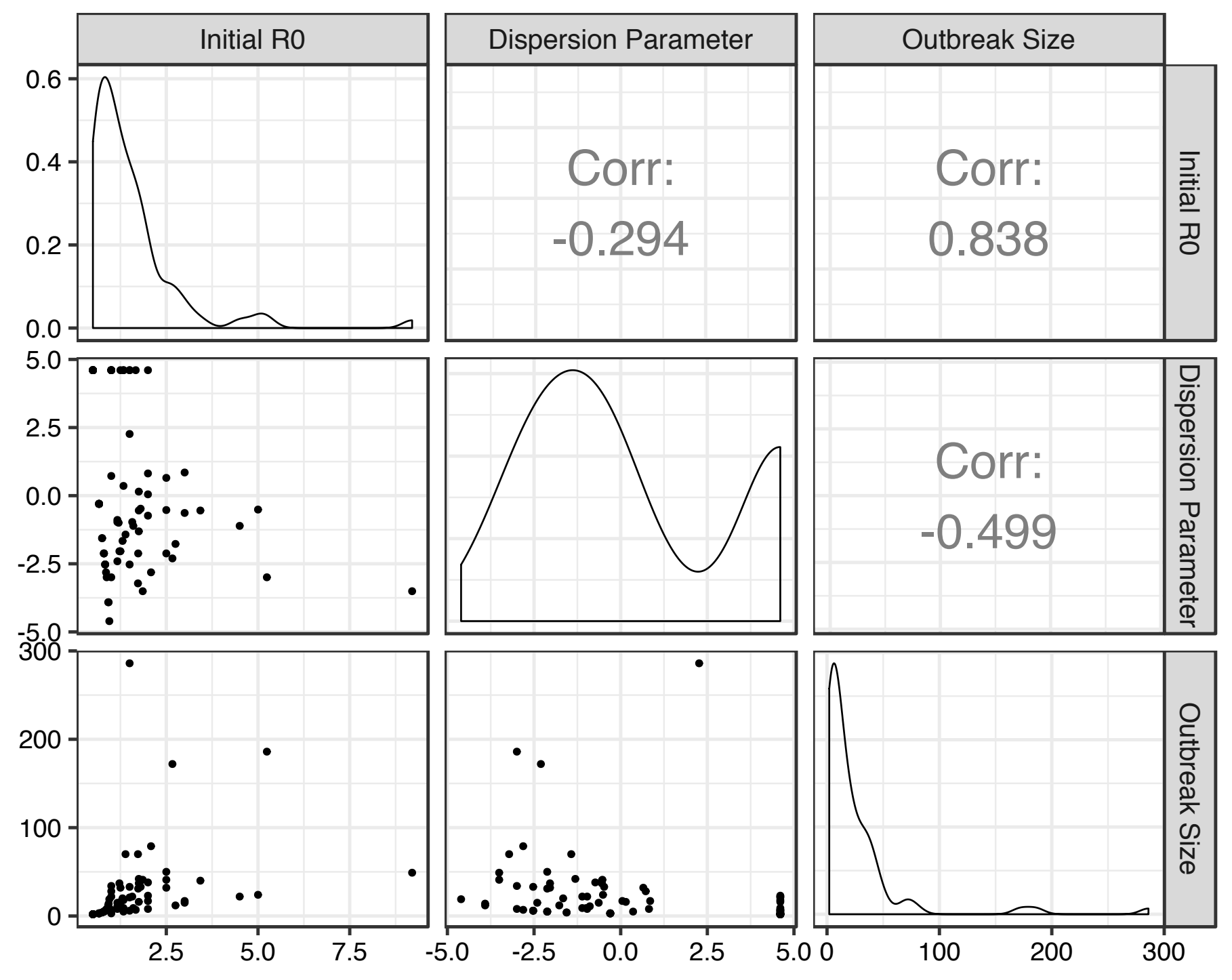


Figure 2. **Outbreak size varies with initial R_0 and dispersion parameter, though initial R_0 was a better predictor.** Initial R_0 and dispersion parameter showed little correlation. The dispersion parameter is displayed on a natural log scale. Correlation calculated using Spearman's method.

Question 2: What is the quantitative contribution of superspreading to outbreak size?

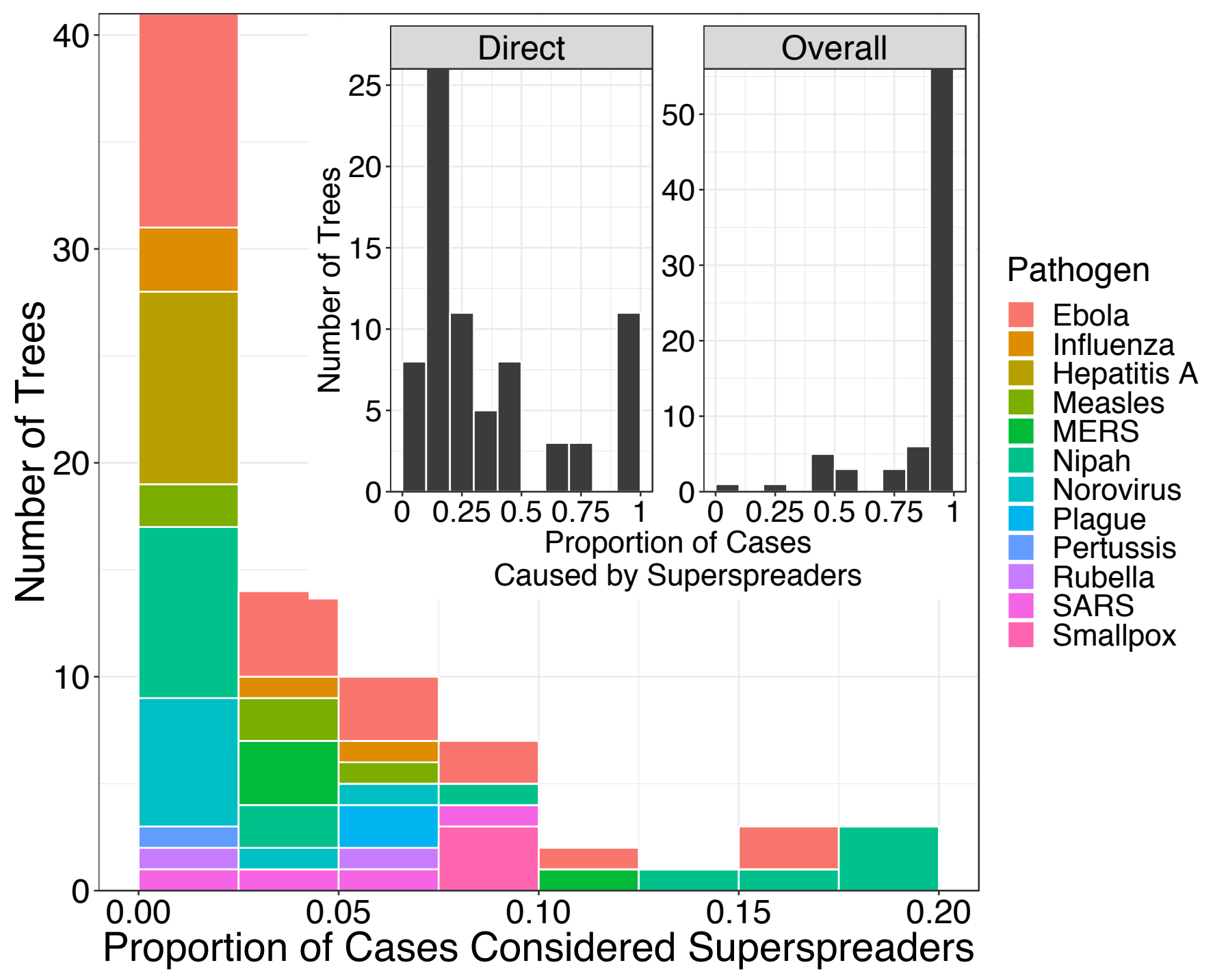


Figure 3. **Superspreaders were rare but epidemiologically important.** Nipah virus and Ebola virus outbreaks had the highest proportion of cases considered superspreaders, while the absolute number of superspreaders was higher in other, larger trees, such as those caused by SARS and MERS. Inset: Superspreaders often directly caused only 10 to 20% of infections, but descendants of superspreaders often accounted for more than 90% of cases overall.

Question 3: What determines the frequency of superspreading events?

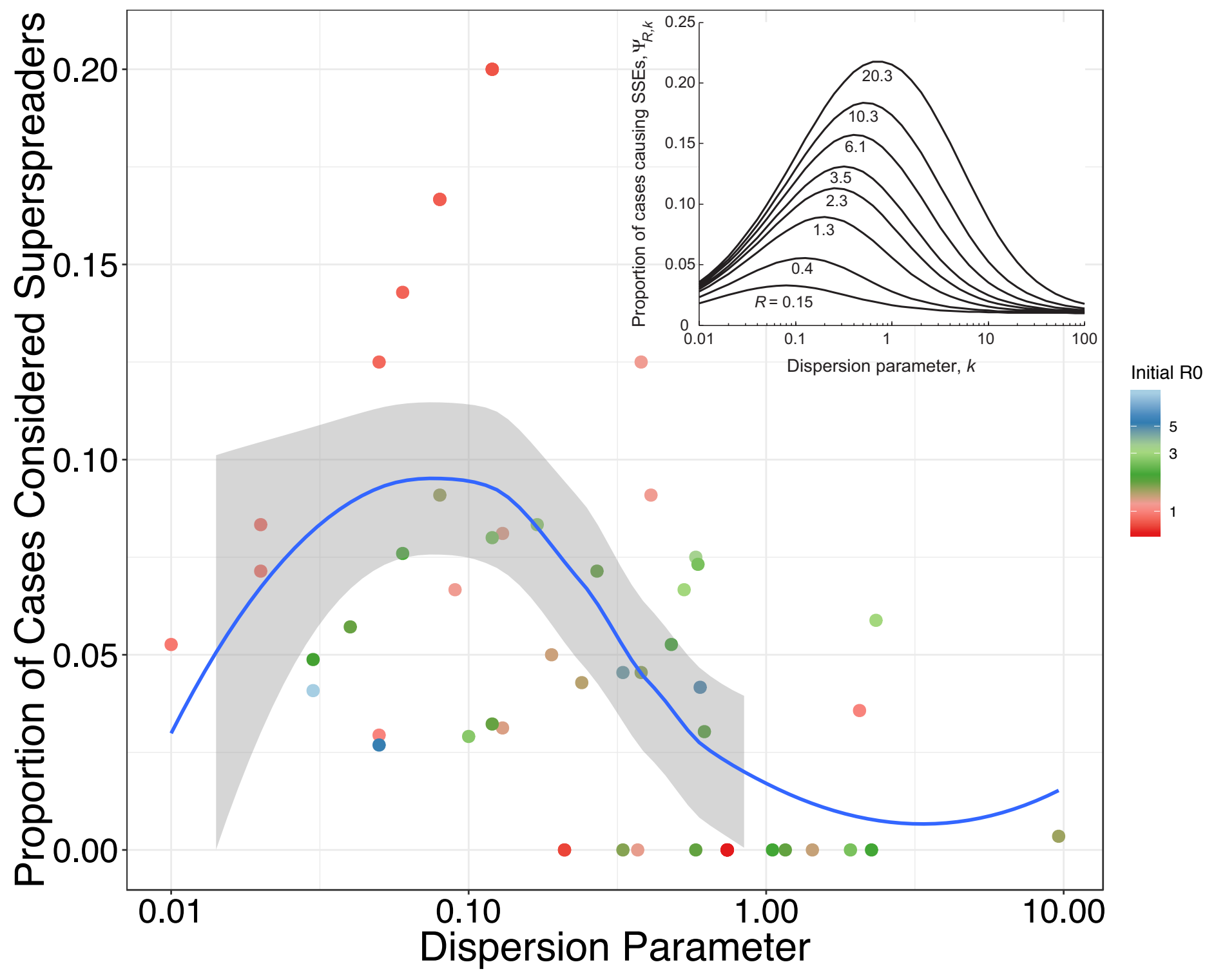


Figure 4. **Agreement between tree database and theoretical prediction about intermediate levels of dispersion leading to a higher proportion of superspreading events.** Intermediate dispersion parameters and low initial R_0s led to the greatest proportion of cases considered superspreaders. Curve represents a LOESS regression. The same maximum likelihood methods were used to determine the dispersion parameters in both figures. Trees with dispersion parameter > 10 excluded. Inset: Figure 3b from Lloyd-Smith et al. (2015).

CONCLUSIONS

- Transmission trees contain valuable information about specific pathogen outbreaks, which is costly to collect. Our database standardizes tree format, allowing for greater comparative analyses.
- Understanding factors associated with increased outbreak size may help predict the extent of outbreaks in the future and lead to more effective preventative measures.
- The impact of superspreading was quantified using a new statistic which we called overall effect, which suggested that superspreaders are perhaps more important than previously realized.
- This database provides the information to test theoretical hypotheses about disease transmission and inspire new ideas, such as:
  - How sensitive is outbreak size and length to superspreader introduction timing?
  - Does knowing the transmission tree of a disease allow us to predict the mode of transmission or type of pathogen (bacterial or viral)?

ACKNOWLEDGEMENTS

Population Biology of Infectious Diseases REU Site (NIH U54 GM111274) provided funding and the Drake Lab provided helpful comments and assistance.

REFERENCES

1. Cauchemez et al. (2011) *PNAS*, **108**(7). 2. Faye et al. (2015) *Lancet ID*, **15**(3). 3. Glur (2018) https://cran.r-project.org/web/packages/data.tree/vignettes/data.tree.html. 4. Hens et al. (2012) *American J Epi*, **176**(3). 5. Jombart et al. (2014) *PLoS Comp Bio*, **10**(1). 6. Lloyd-Smith et al. (2005) *Nature*, **438**(7066). 7. Meyers et al. (2005) *J Theor Biol*, **232**(1). 8. Stein (2011) *Int J ID*, **15**(8). 9. Tildesley et al. (2009) *J Theor Biol*, **258**(4). 10. Tree Sources: https://bit.ly/2GdLLUq