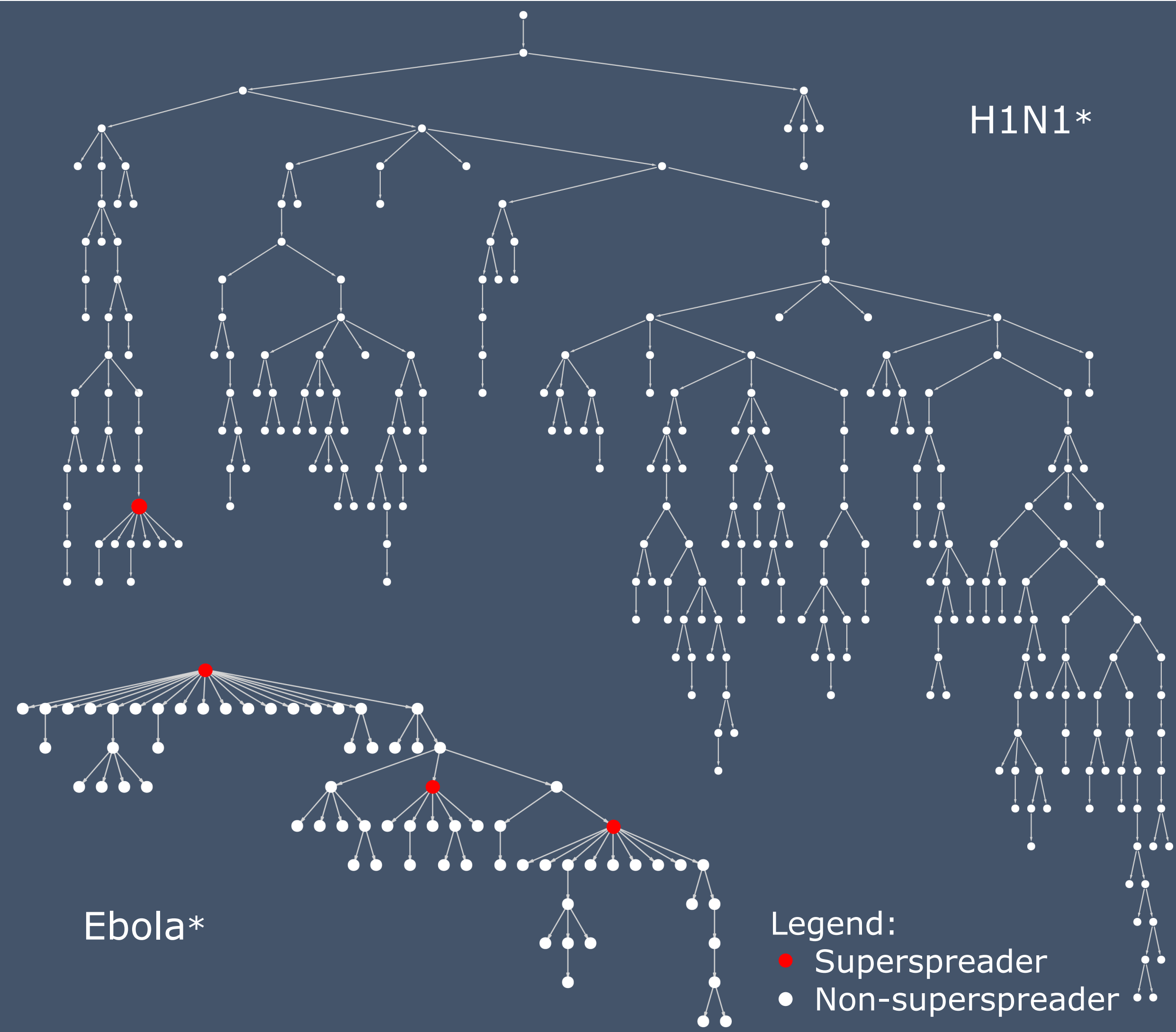


# Who Infected Whom? Creating a Database of Transmission Trees for Comparative Outbreak Analysis

Juliana C. Taube<sup>1</sup>, Paige B. Miller<sup>2</sup>, and John M. Drake<sup>2</sup>

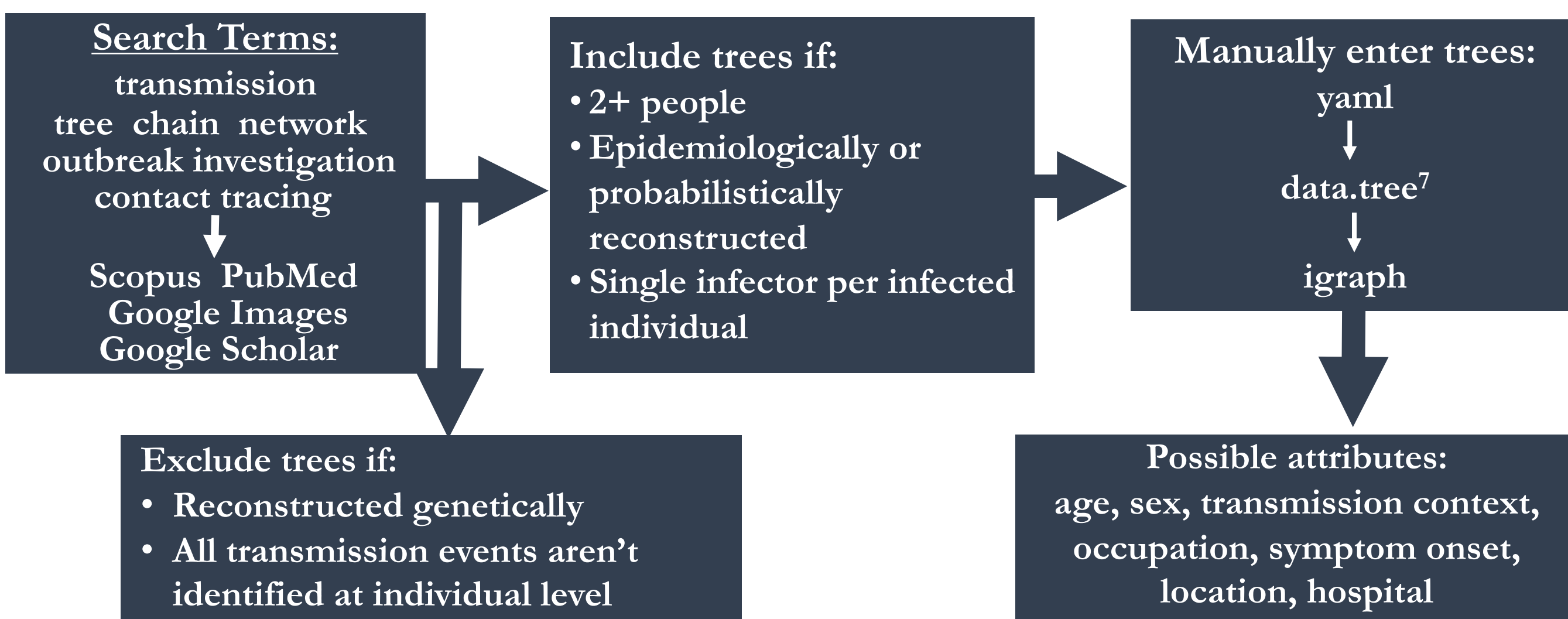
<sup>1</sup> Bowdoin College, Brunswick, ME; <sup>2</sup> Odum School of Ecology, University of Georgia, Athens, GA



## 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 the average reproductive number ( $R_0$ ), the dispersion parameter (variation in  $R_0$ ), pathogen mutation rate, and intervention efficacy, though greater standardization in tree format would make these trees and statistics more comparable. Additionally, a better understanding of predictors of outbreak size<sup>1,2</sup> and the importance of superspreaders to onward transmission across different pathogens<sup>3,4</sup> would help researchers developing transmission tree reconstruction methods<sup>5,6</sup> 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

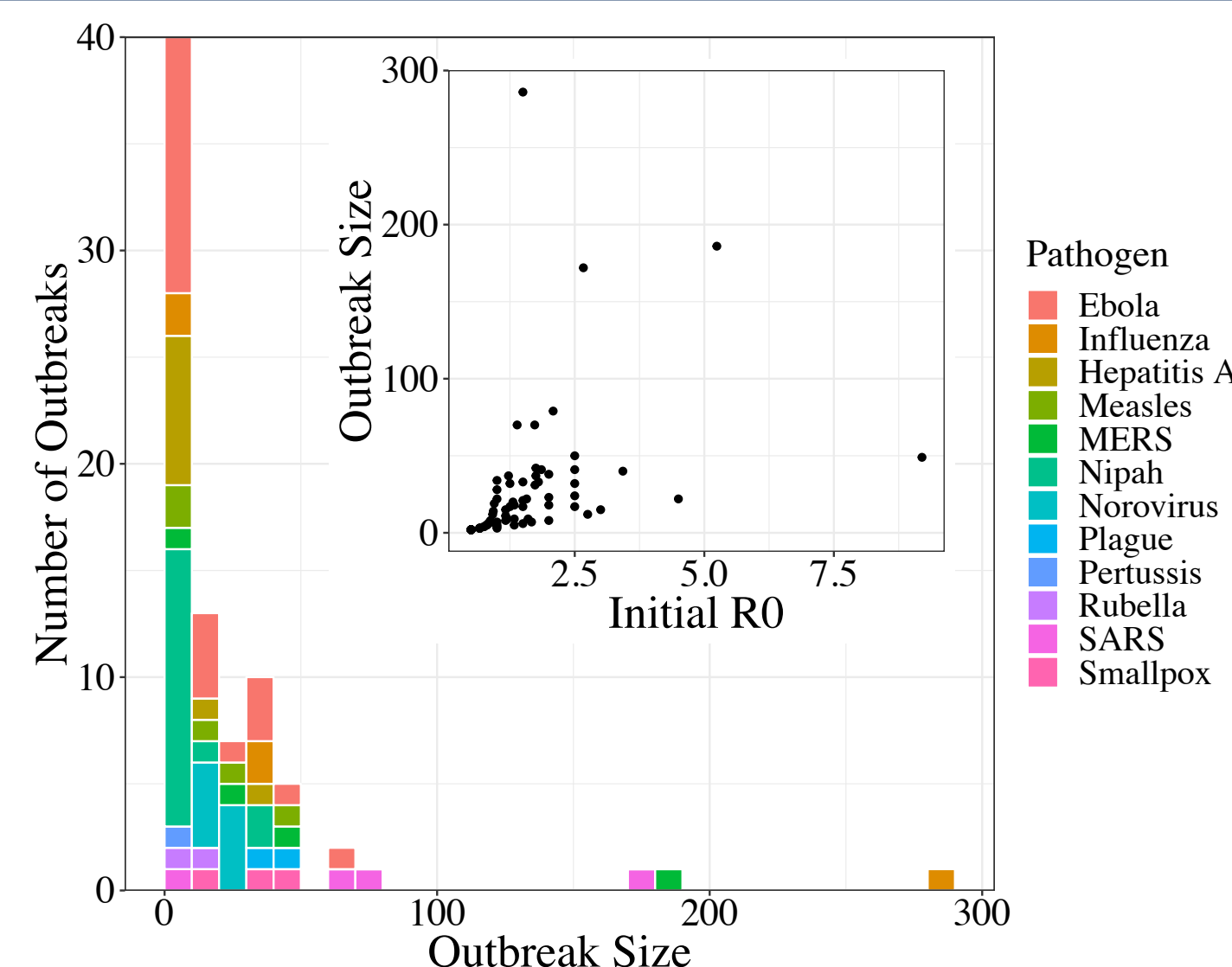


## TREE ANALYSIS

- 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<sup>3</sup>:** cases who transmitted to more individuals than the 99<sup>th</sup> 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 caused by superspreaders 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
Hep. A	9	Pertussis	1
Measles	5	Rubella	2
MERS	4	SARS	4
Nipah	16	Smallpox	3



Outbreak size 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  $R_0$  ( $r = 0.841$ , Spearman).

## REFERENCES

- Meyers et al. (2005) *J Theor Biol*, 232(1).
- Tildesley et al. (2009) *J Theor Biol*, 258(4).
- Lloyd-Smith et al. (2005) *Nature*, 438(7066).
- Stein (2011) *Int J ID*, 15(8).
- Hens et al. (2012) *American J Epi*, 176(3).
- Jombart et al. (2014) *PLoS Comp Bio*, 10(1).
- Glur (2018) <https://cran.r-project.org/web/packages/data.tree/vignettes/data.tree.html>.

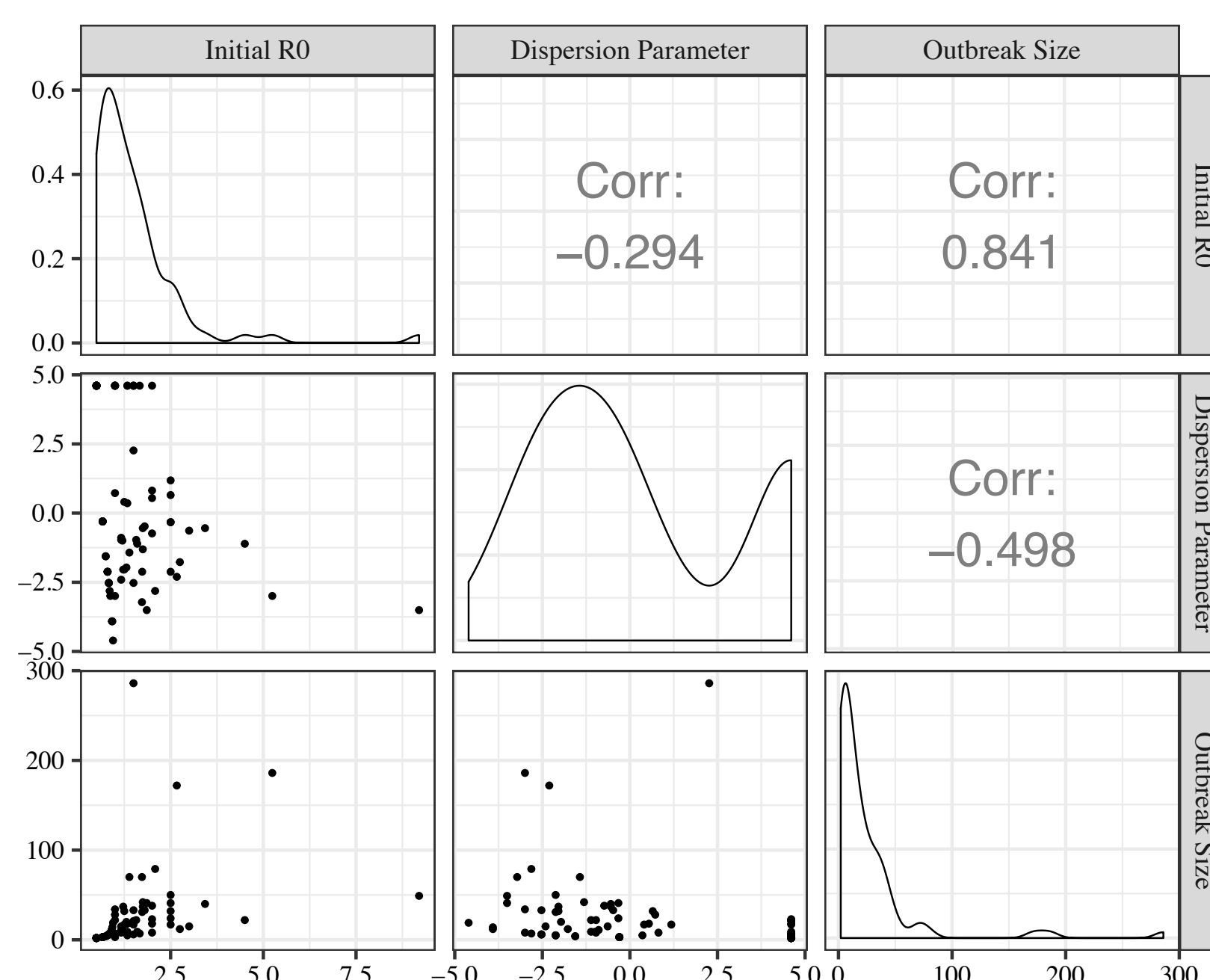
Population Biology of Infectious Diseases REU Site (NIH U54 GM111274) provided funding and the Drake Lab provided helpful comments and assistance.

✉ jtaube@bowdoin.edu



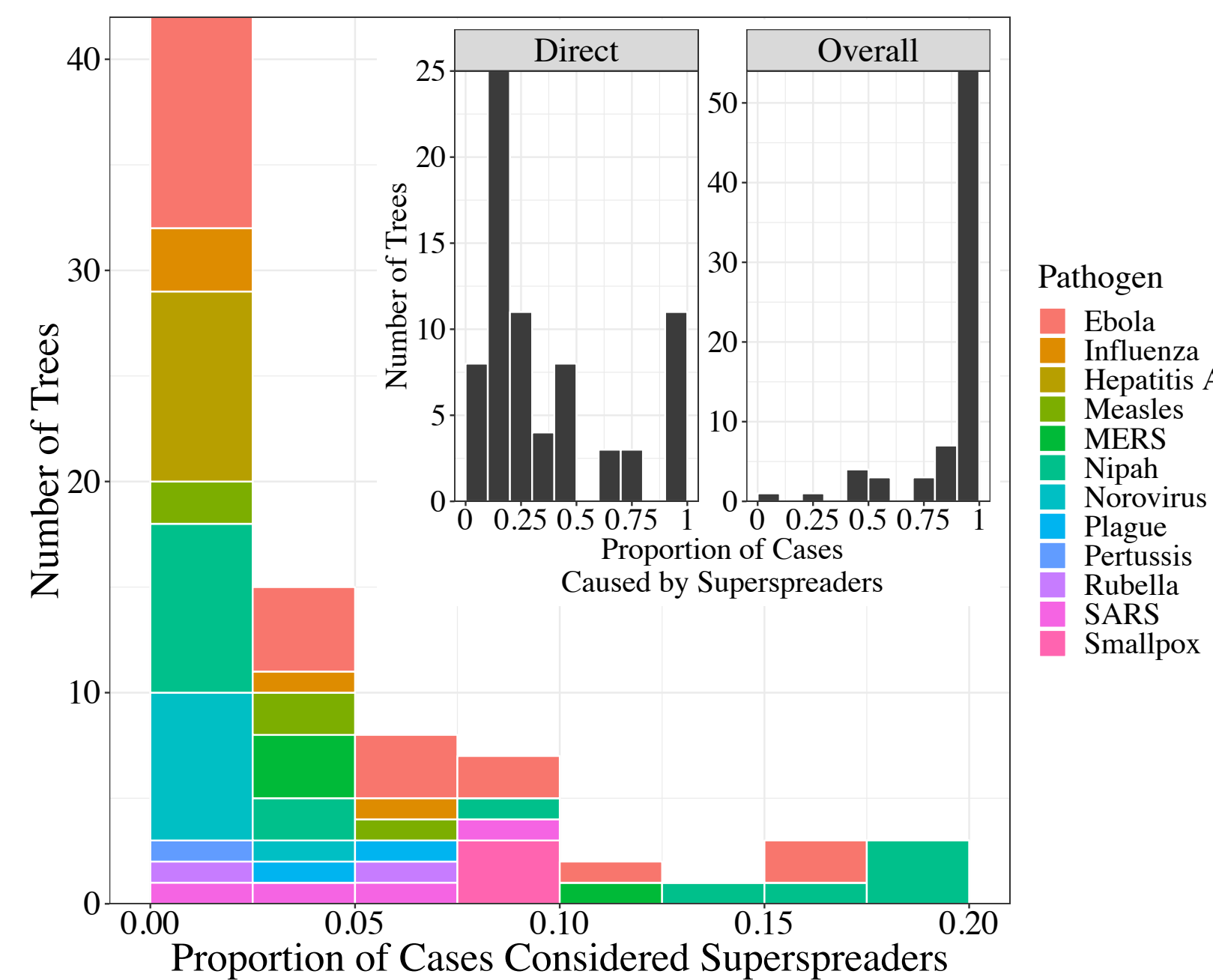
## EXAMPLE QUESTIONS

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



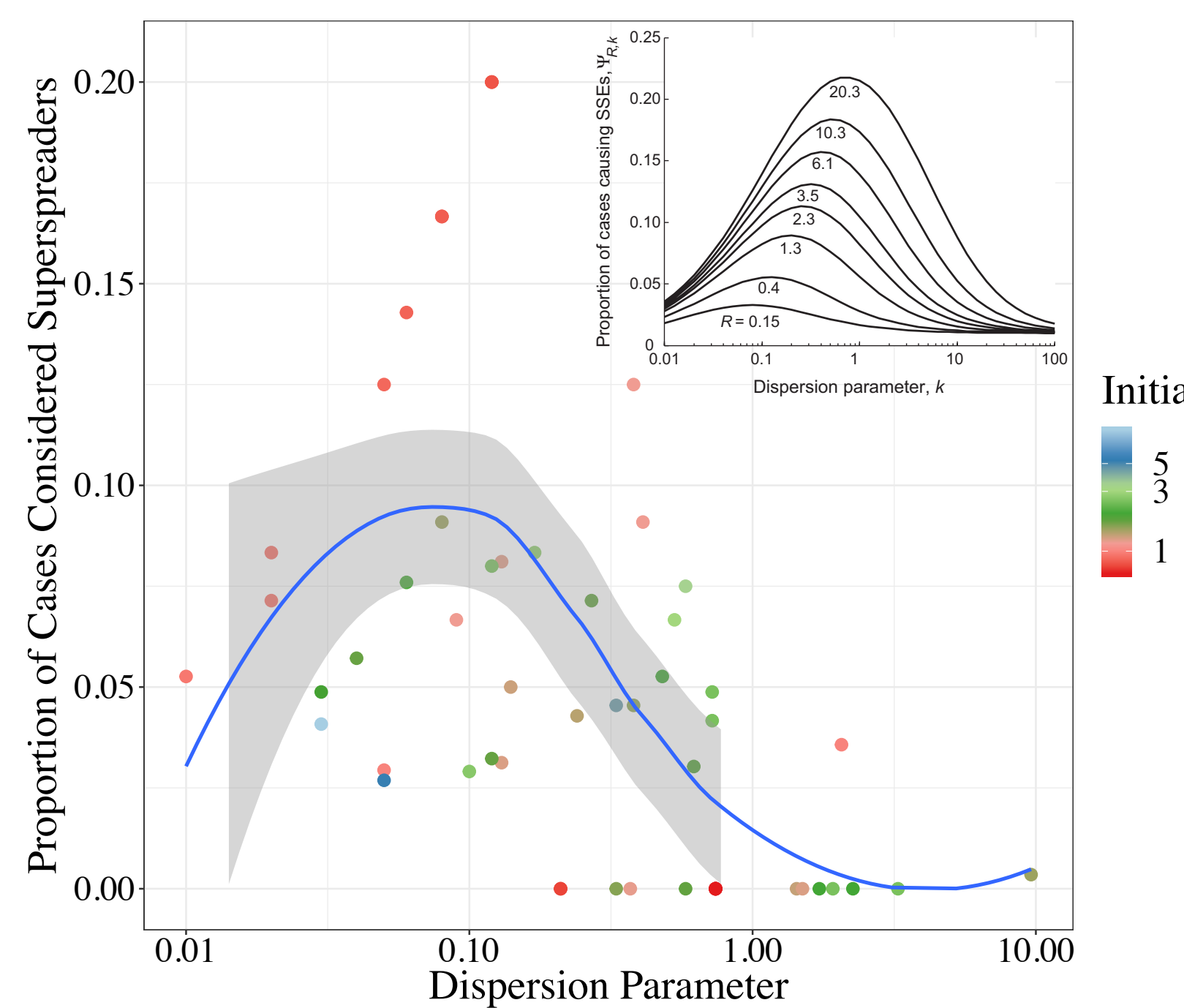
Outbreak size varies with initial  $R_0$  and dispersion parameter, though initial  $R_0$  was a better predictor. Initial  $R_0$  and dispersion parameter showed little correlation. The dispersion parameter is displayed on a natural log scale. Correlation calculated using Spearman's method.

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



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?**



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  $R_0$ s 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.
- 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)?
- Understanding factors associated with increased outbreak size may help predict the extent of outbreaks in the future and lead to more effective preventative measures.