

Appendix: Host communities shape pathogen outbreak dynamics through a phylogenetic dilution effect.

Marjolein E.M. Toorians^{1,*}, Isabel M. Smallegange², and T. Jonathan Davies^{1,3,4}

¹Department of Botany, Biodiversity Research Centre, University of British Columbia, 2212 Main Mall, Vancouver, BC V6T 1Z4, Canada

²School of Natural Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

³African Centre for DNA Barcoding, University of Johannesburg, Johannesburg 2092, South Africa

⁴Department Forest & Conservation Sciences, University of British Columbia, 2212 Main Mall, Vancouver, BC V6T 1Z4, Canada

*Corresponding author: marjolein.toorians@botany.ubc.ca

May 5, 2023

1 Model validation

1.1 Community simulations

Host communities were simulated using a Yule-process with the *pbtree()* function from the *Picante* library in R. This simulates trees with varying tree depth and an assigned number of species (S) with no death. The birth rate of the species (b) determines the nodes in the tree, the split of a species into two, or a speciation event (Figure 1). Figure 2 shows how the birth rate of species, b , affects the PD of the tree comprising the simulated community, and shows that increasing b reduces PD and increases R_0 .

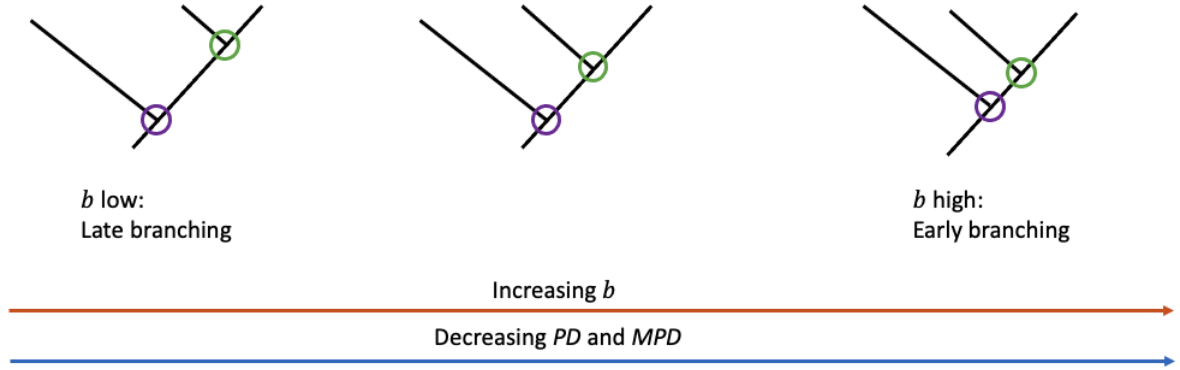


Figure 1: Illustration showing the function of species birth rate parameter b in speciation for the simulation trees. Increasing b increases branching time between speciation events, therefore the species (tips of the tree) are more closely related. In these simulations we allow for the tree depth to vary. The purple circle represents the first speciation event, the green circle the second speciation event in this 3-species tree.

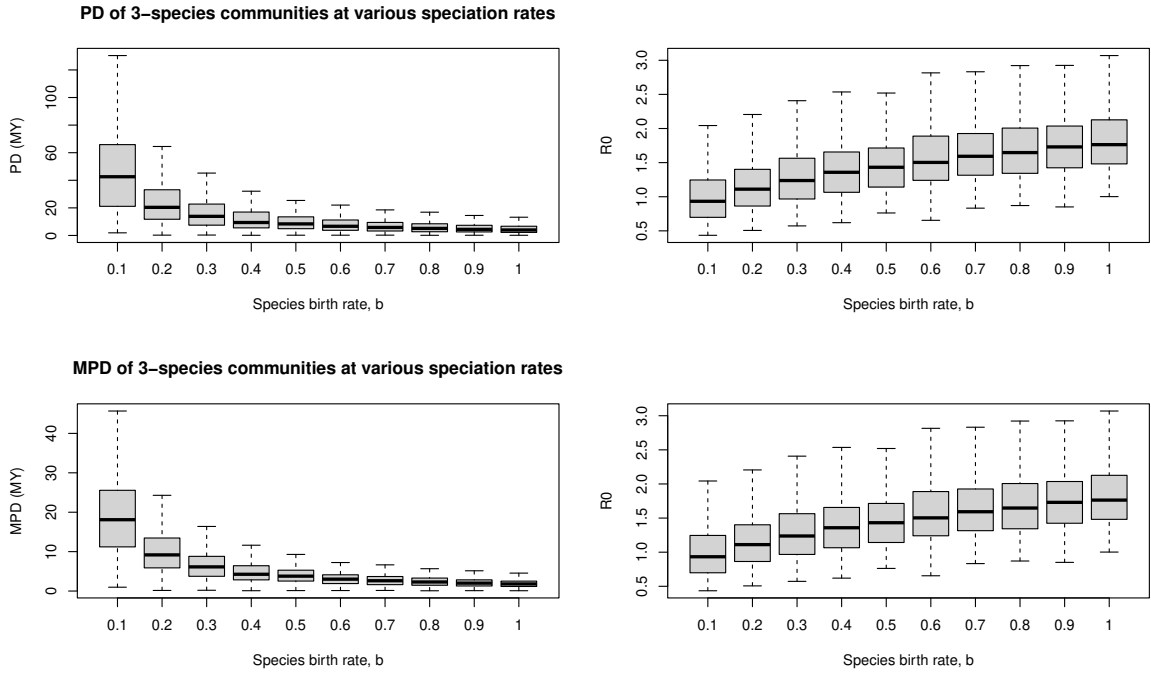


Figure 2: Boxplots showing host tree phylogenetic diversity (upper left), Mean Pairwise phylogenetic Distance (lower left) and pathogen R_0 (right panels) for three-taxon cases, simulated using different values for species birth rate, b . Data was created from simulations using $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$, $\lambda = 1$. 100 communities were simulated per species birth rate value, b .

1.2 The effect of species richness and birth rate on R_0 for various phylogenetic metrics

1.2.1 Scenario 1

Supplementary Figures 3,4 and 5 show the relationship between the different phylogenetic metrics and R_0 at different values of species birth rate, b and species richness for the Scenario 1. In general, increasing the species birth rate, b , increases R_0 .

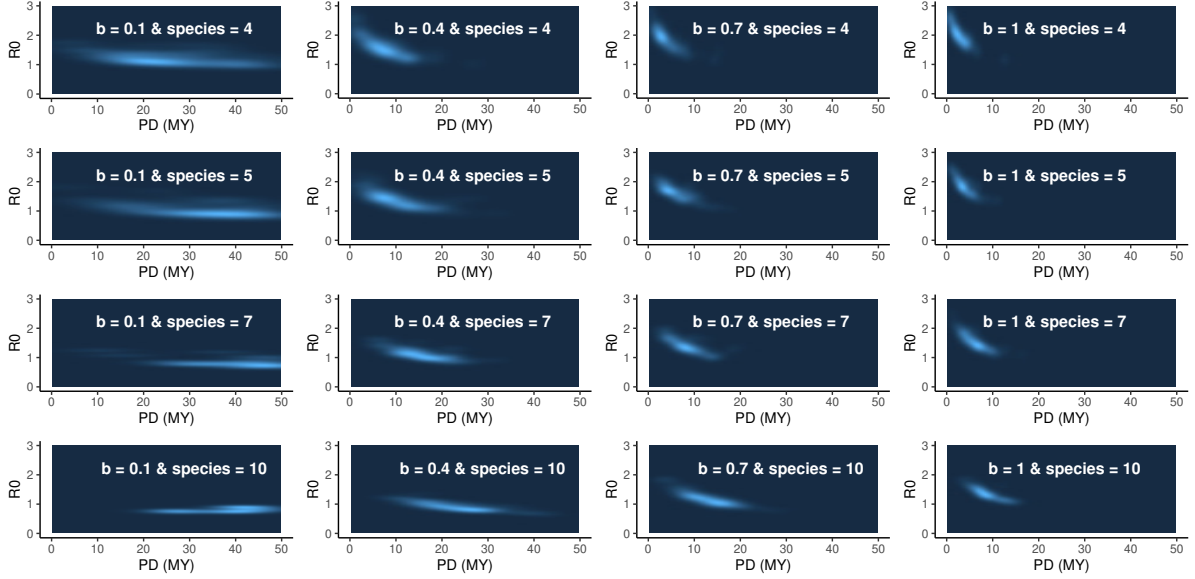


Figure 3: Scenario 1: Density plots of R_0 and Phylogenetic Diversity (PD) of simulated trees at different levels of species in the community, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$.

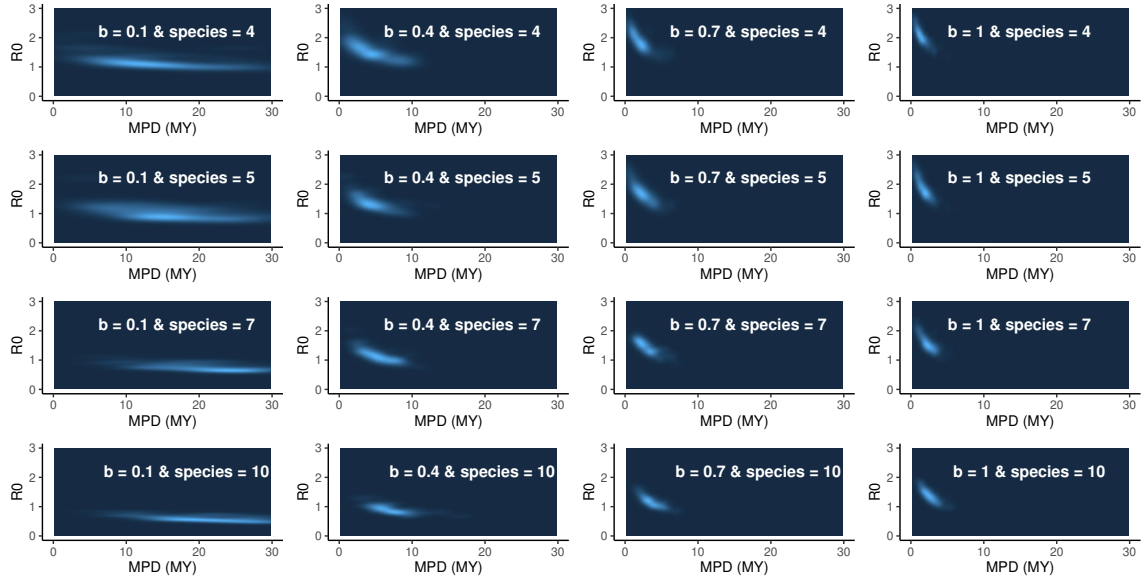


Figure 4: Scenario 1: Density plots of R_0 and Mean Pairwise Distances (MPD) of the host species in the simulated trees at different levels of species in the community, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$.

1.2.2 Scenario 2a

Similar to the preceding graphs, but here the establishment probability of the pathogen is dependent on the primary host, as in Eqn (12). Here, we show how R_0 scales with PD , MPD and Ic at different values of species and b for simulations in which the primary host is tightly nested within the host phylogeny (host with lowest MPD).

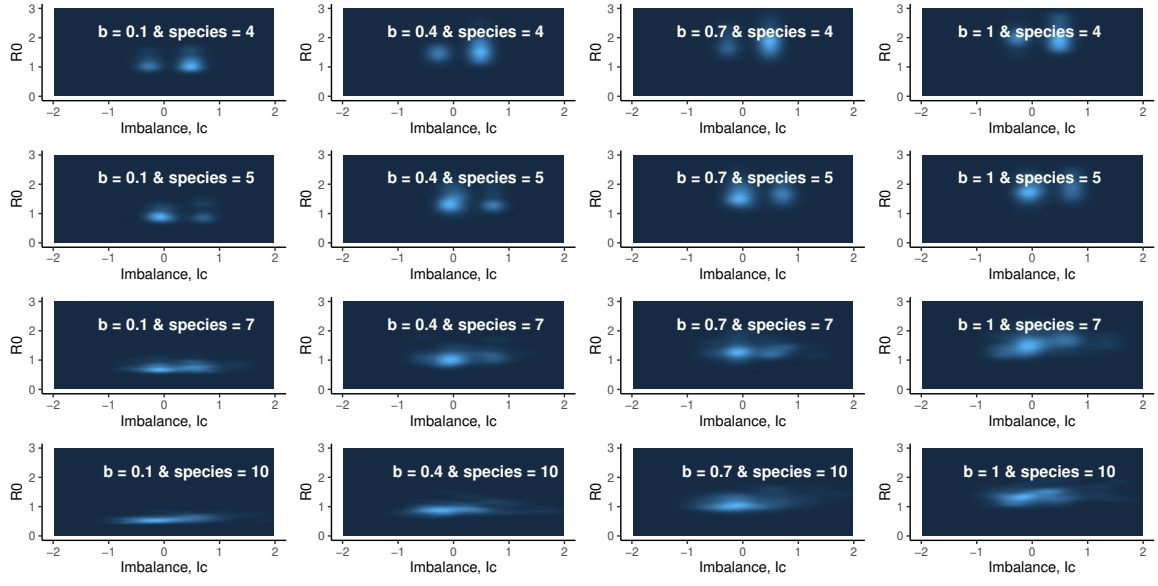


Figure 5: Scenario 1: Density plots of R_0 and tree Imbalance of simulated host phylogeny at different levels of species in the community, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$.

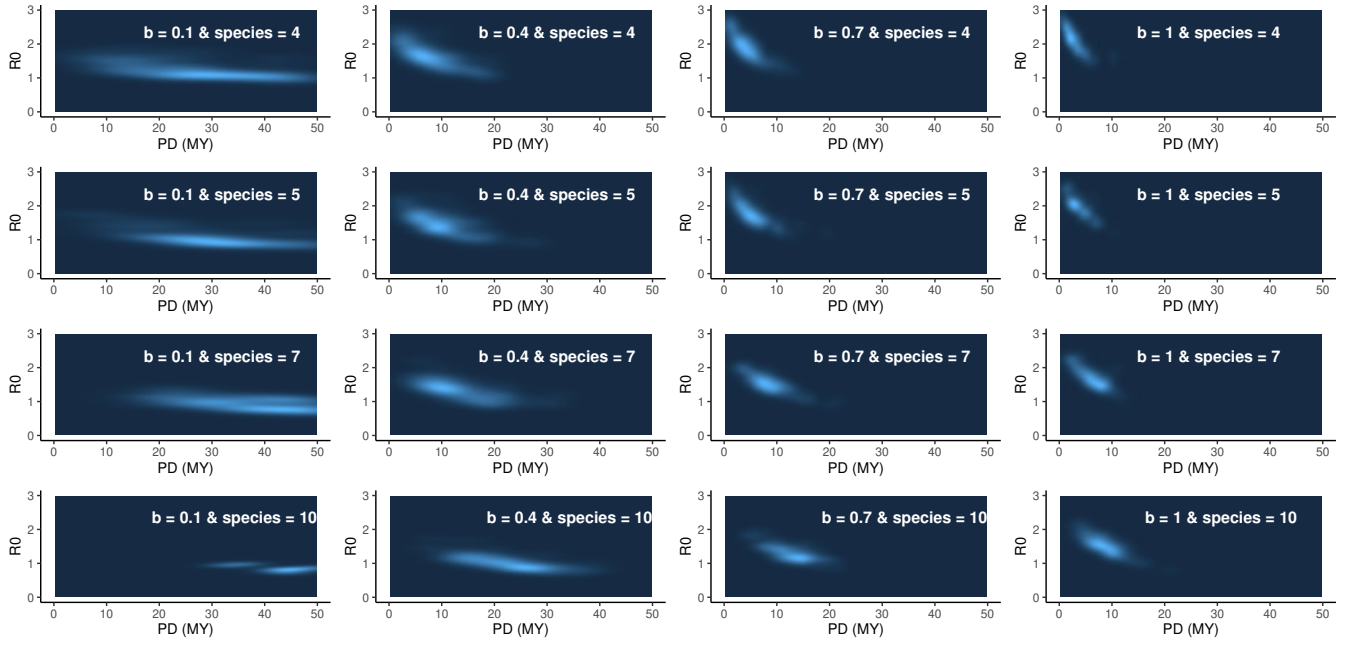


Figure 6: Scenario 2a: Density plots of R_0 and Phylogenetic Diversity (PD) of simulated trees with different host richness, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is tightly nested within the host phylogeny, defined as the species in the community with with smallest MPD .

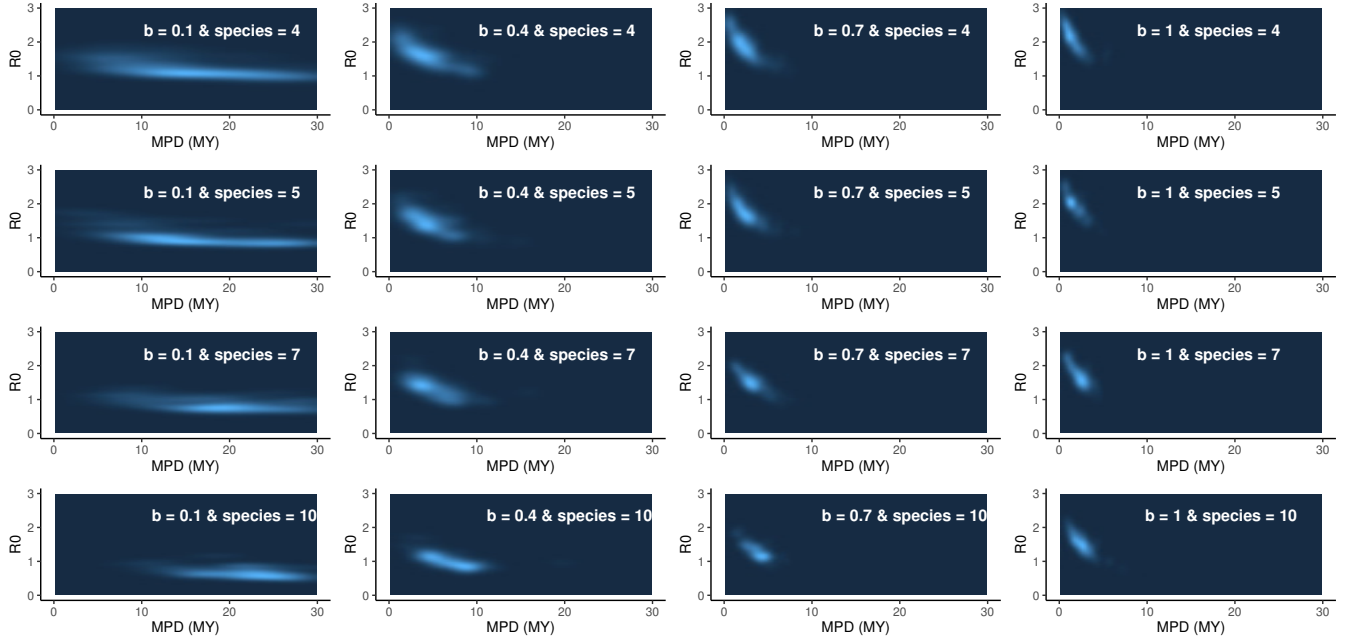


Figure 7: Scenario 2a: Density plots of R_0 and Mean Pairwise Distance (MPD) of simulated trees with different host richness, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is tightly nested within the host phylogeny, defined as the species in the community with the smallest MPD .

1.2.3 Scenario 2b

Similar to Scenario 2a, here, we show how R_0 scales with PD , MPD and Ic at different values of species richness and b for simulations in which the primary host is phylogenetically distant from secondary hosts (host with highest MPD).

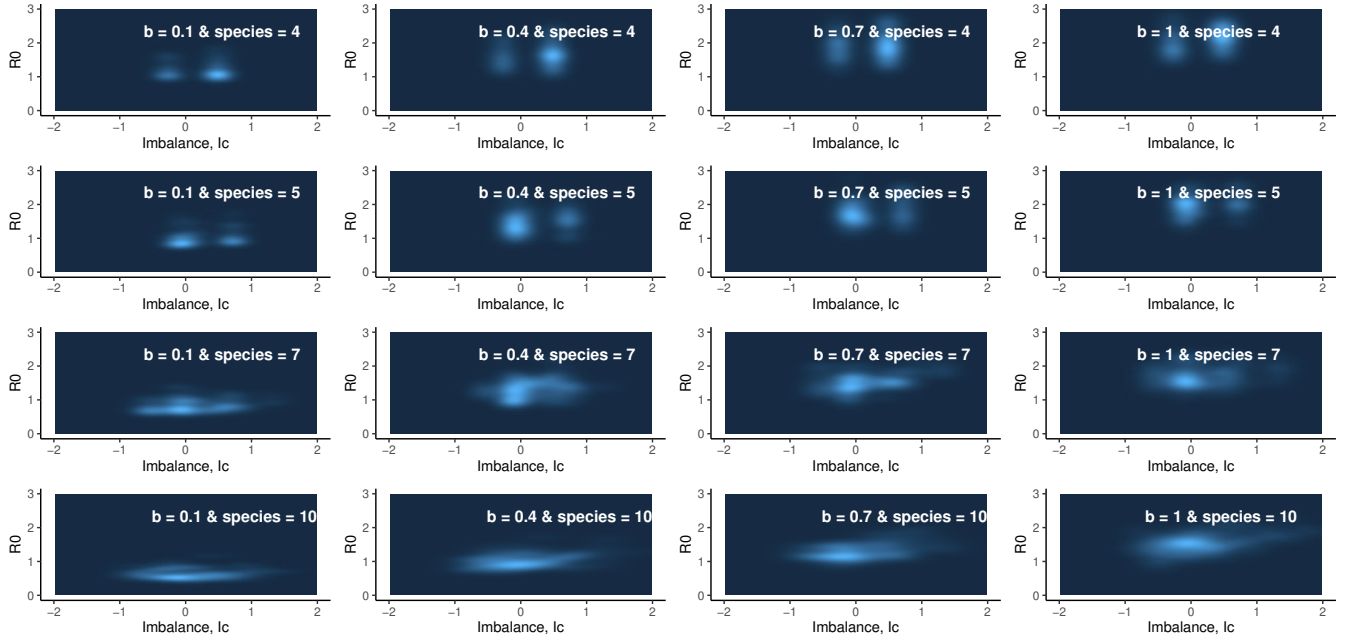


Figure 8: Scenario 2a: Density plots of R_0 and Imbalance (I_c) of simulated host phylogeny with different host richness, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is tightly nested within the host phylogeny, defined as the species in the community with with smallest MPD .

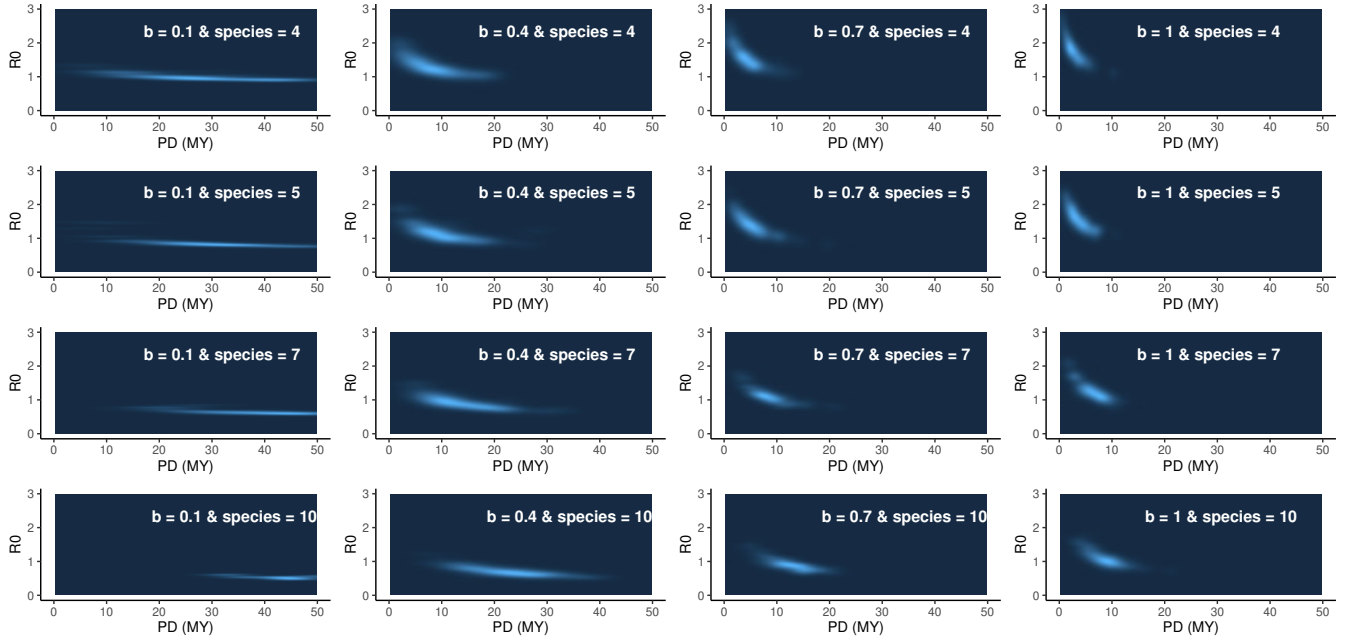


Figure 9: Scenario 2b: Density plots of R_0 and Phylogenetic Diversity (PD) of simulated trees with different host richness, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is the furthest phylogenetically distant from the secondary hosts, defined as the species in the community with with largest MPD .

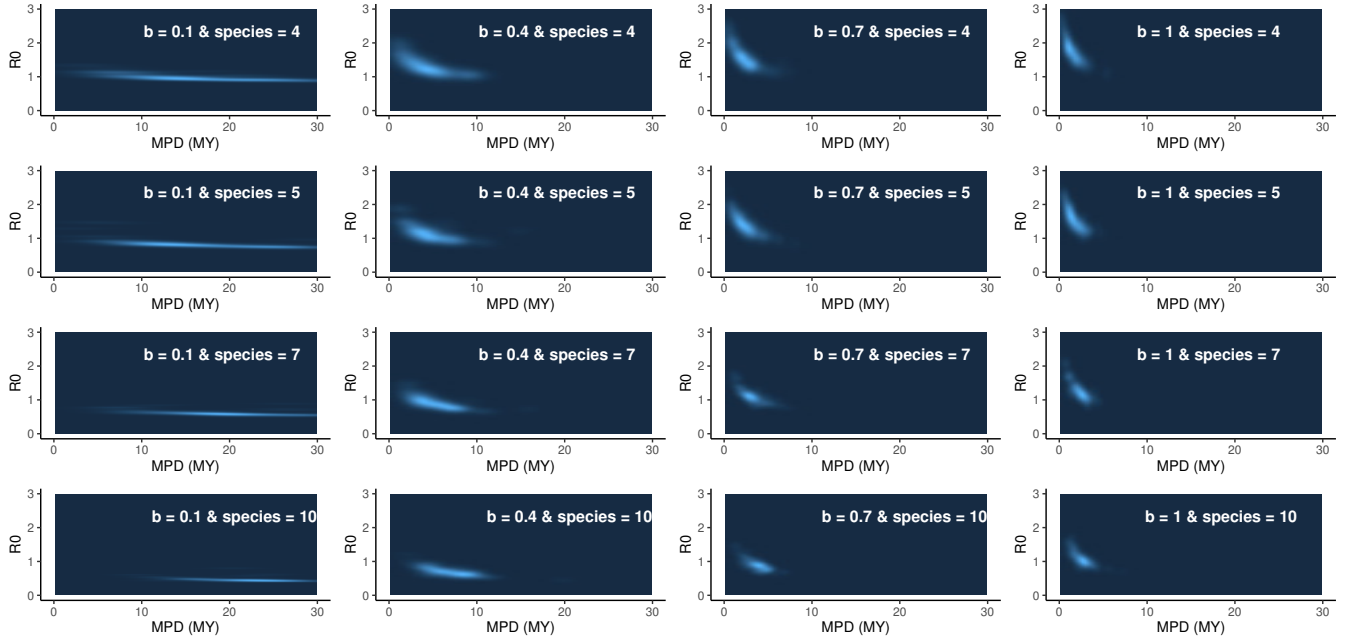


Figure 10: Scenario 2b: Density plots of R_0 and Mean Pairwise Distances (MPD) of the host species in the simulated trees at different levels of species in the community, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is the furthest phylogenetically distant from the secondary hosts, defined as the species in the community with with largest MPD .

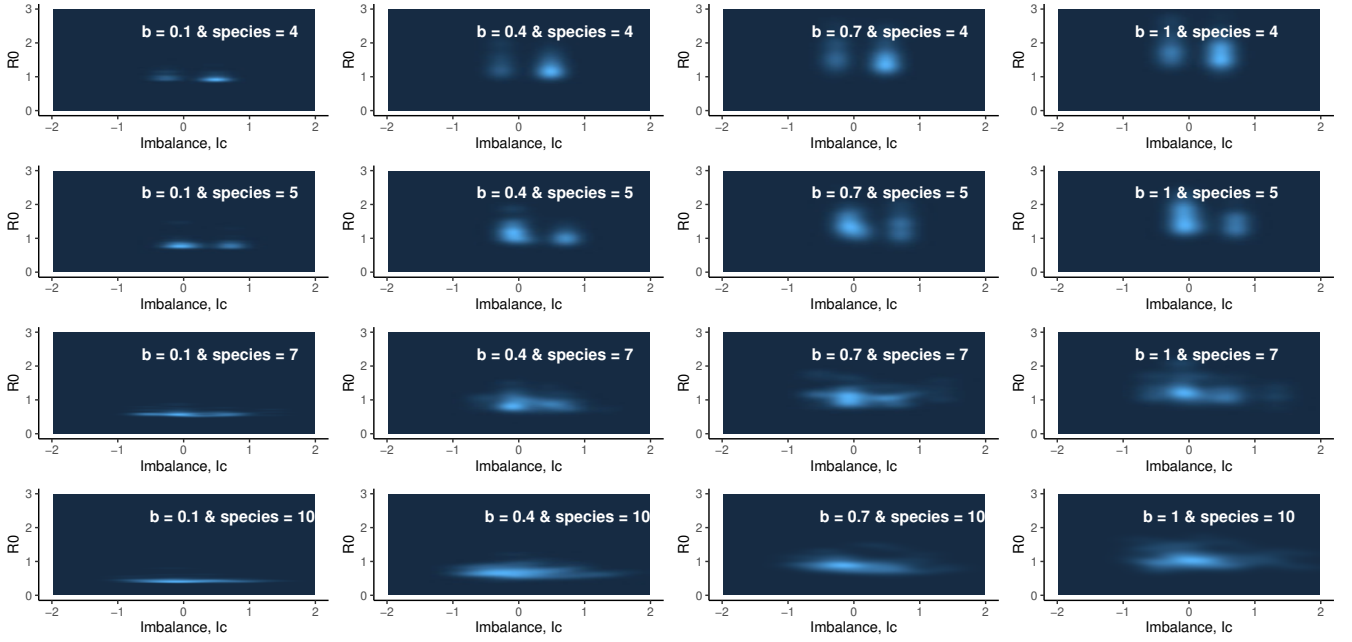


Figure 11: Scenario 2b: Density plots of R_0 and tree Imbalance (Ic) of simulated host phylogeny at different levels of species in the community, S , and at various species birth rates, b . Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. The primary host is the furthest phylogenetically distant from the secondary hosts, defined as the species in the community with with largest MPD .

2 PCA

2.1 Scenario 1

Using simulations, we show that R_0 varies with both the richness of hosts and the topology of the phylogenetic tree that connects them (Figure 12, left panel). The Principal Component Analysis (PCA) shows that 88% of variation falls along two independent axes: PCA1 - 63.73%- with weights predominantly reflecting MPD and PD , and PCA2 - 24.63%- which captures variation in tree imbalance (Figure 12, left panel). Increasing the number of species in the tree causes increasing variation between biodiversity metrics and R_0 (colours in Figure 12).

2.2 Scenario 2

The PCA for scenario 2 with multiple hosts (where the primary host is randomly assigned in each simulation) reveals that 86% of the variation is distributed over the first two axes (Figure 12, right panel). ED is closely aligned with PD and MPD .

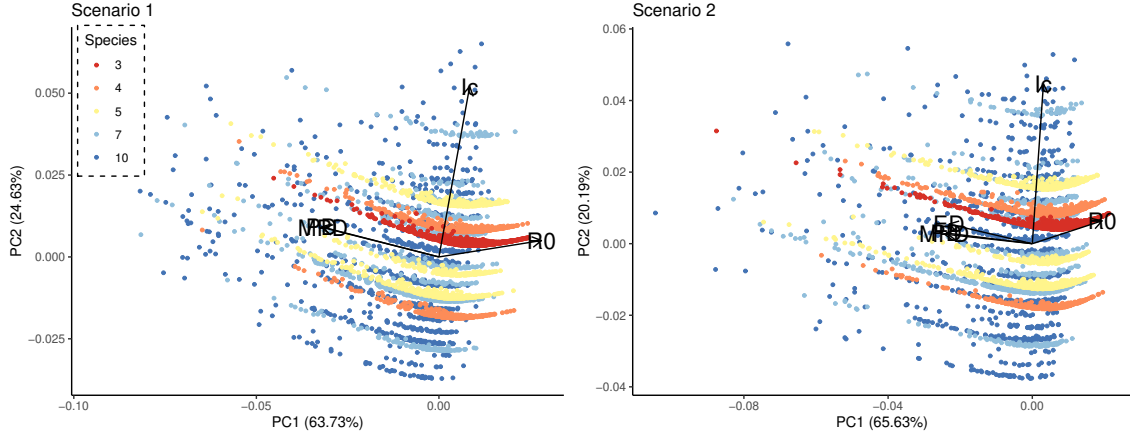


Figure 12: Principal Component Analysis for Scenarios 1 (left) and 2 (right) with $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. Colours indicate simulated communities with different host richness ($S=3,4,5,7,10$).

3 Alternative contact rate structures

To explore more realistic contact rates, we generated an alternative set of simulations with interspecific contact rates lower than intraspecific contact rates. Here, we use *Kronecker's* δ function for the contact rate:

$$\kappa(N) = \kappa_0(\delta_{ij} + (1 - \delta_{ij})\alpha) \cdot f(N) \quad (1)$$

when $i = j$, then $\delta = 1$, if $i \neq j$, $\delta = 0$. κ_0 stands for the intraspecific contacts, and α represents the interspecific contacts (note that here δ is Kronecker's δ , not the natural mortality rate used in the *SI* model in the main text). The function $f(N)$ determines the nature of the transmission. In density-dependent systems, $f(N) = 1$, where in frequency-dependent systems, such as used in the model in the main text, $f(N) = \frac{1}{N}$. Here, we explore two other contact structures for both scenarios:

3.1 Interspecific = 0.1 x Intraspecific contact rate

In this model, we assume that the interspecific contacts are 10 times less likely than intraspecific contacts. Here, the interspecific contact rates are set to 0.1, and the intraspecific contact rates are set to 1. Now, R_0 no longer drops below 1, thus, the pathogen always causes an epidemic. The pathogen mostly spreads within the primary reservoir, and there is less spillover. This is shown in Figure 13. Multivariate model (a) for Scenario 1 is $R_0 = 2.61 - 0.13\log(MPD) + 0.01Ic - 26S$, with an $R^2 = 0.966$. For Scenario 2, model (a) is $R_0 =$

48 $2.39 - 0.06\log(MPD) + 0.01Ic - 0.33S$ with $R^2 = 0.955$, and model (b) $R_0 = 2.39 - 0.07\log(ED) + 0.01Ic - 0.34S$
 49 with $R^2 = 0.964$.

50

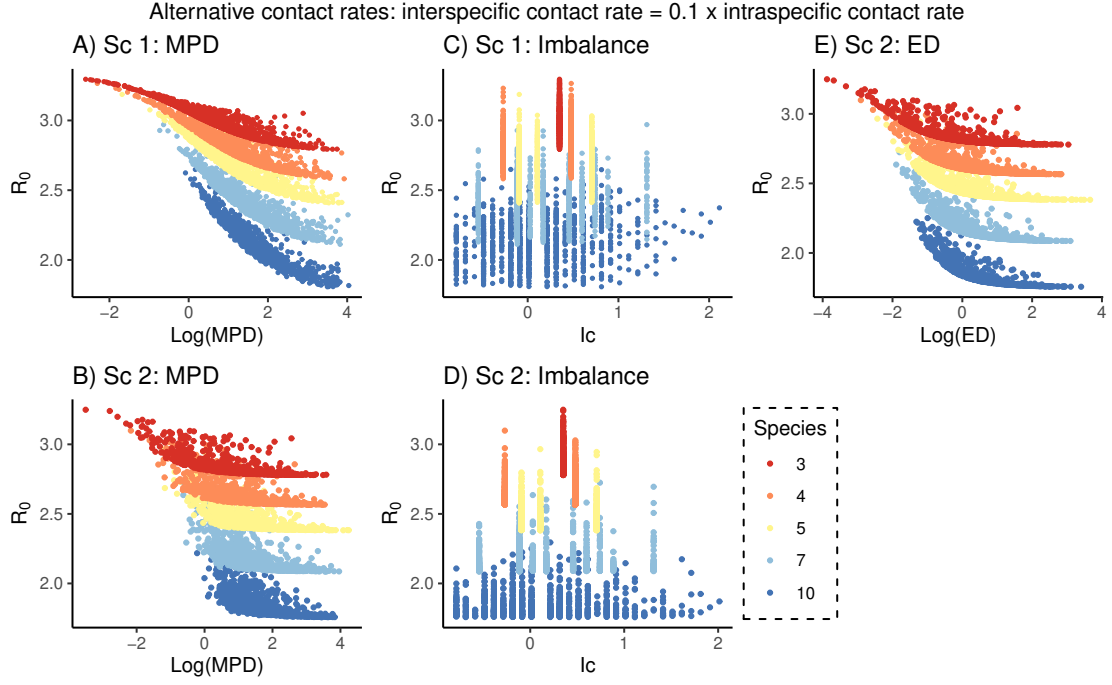


Figure 13: A,B: Mean Pairwise Distance (MPD) of the community is inversely related to R_0 in Scenario 1 (A) and Scenario 2 (B). C,D: Imbalance of the phylogenetic tree representing the host species community shows high variation in its effect on R_0 , with increasing variation when species richness, S , is increased. E: Evolutionary Distinctiveness for the randomly assigned primary host in Scenario 2. Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. Colours indicate simulated communities with different host richness ($S=3,4,5,7,10$). Interspecific contact rates, κ_{ij} , are 0.1, while intraspecific contact rates, κ_{ii} , are 1.

3.2 Interspecific contact rates drawn from a uniform distribution

Here, interspecific contact rates, κ_{ij} , were drawn from $U[0.1, 0.5]$. This allows greater interspecific contact rates than in the previous simulations, reducing R_0 , as is shown in Figure 14. Multivariate model (a) for Scenario 1 is $R_0 = 1.30 - 0.43\log(MPD) + 0.03Ic - 0.12S$, with an $R^2 = 0.83$. For Scenario 2, model (a) is $R_0 = 1.40 - 0.44\log(MPD) + 0.01Ic - 0.16S$ with $R^2 = 0.80$, and model (b) $R_0 = 1.40 - 0.46\log(ED) + 0.02Ic - 0.22S$ with $R^2=0.91$.

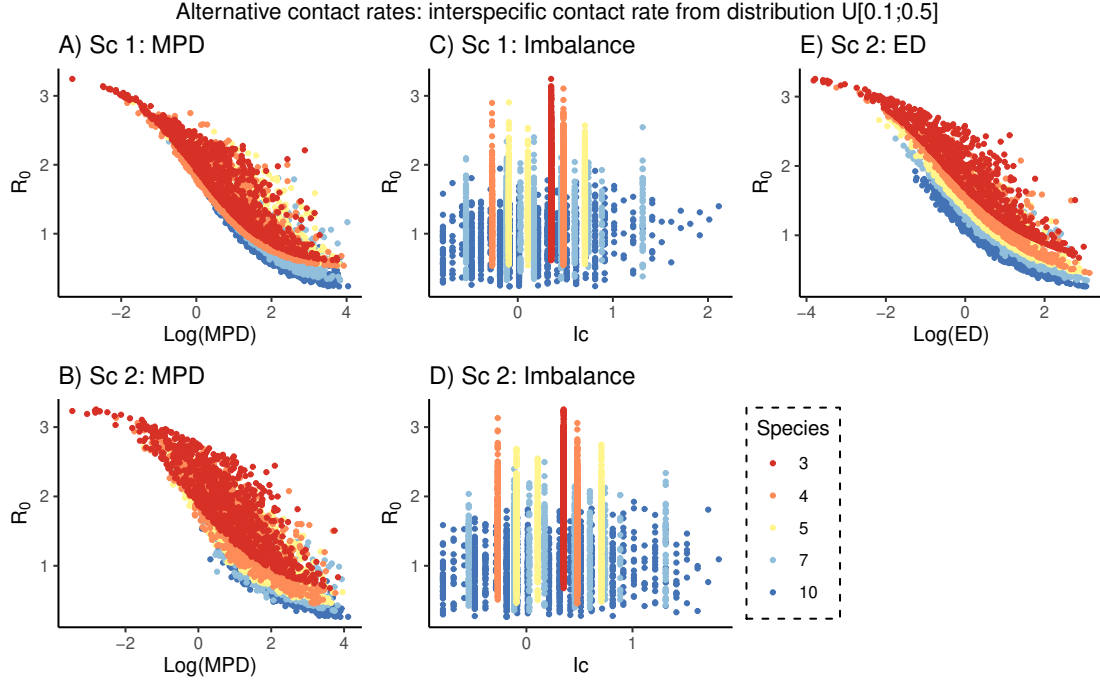


Figure 14: A,B: Mean Pairwise Distance (MPD) of the community is inversely related to R_0 in Scenario 1 (A) and Scenario 2 (B). C,D: Imbalance of the phylogenetic tree representing the host species community shows high variation in its effect on R_0 , with increasing variation when species richness, S , is increased. E: Evolutionary Distinctiveness for the randomly assigned primary host in Scenario 2. Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. Colours indicate simulated communities with different host richness ($S=3,4,5,7,10$). κ_{ij} , were drawn from a uniform distribution ranging from 0.1-0.5. Intraspecific contact rates, κ_{ii} , are 1.

4 Negative exponential probability of successful transmission

Although the original model in the main text describes an inverse relationship of transmission potential to phylogenetic distance between hosts (Lively, 2010; Poullain and Nuismer, 2012), it is yet unknown what this relationship may look like, and it can differ between host species and pathogen-types (Gilbert et al., 2012). This alternative relationship is based on empirical work by Gilbert et al. (2012), and shows an exponential decrease in successful transmission with an increase of phylogenetic distance, PPd . The regressions change as shown in table 1.

Variable	Scenario 1		Scenario 2	
	Estimate effect on R_0	R^2	Estimate effect on R_0	R^2
log(PD)	-0.42	0.75	-0.44	0.71
log(MPD)	-0.43	0.61	-0.45	0.58
Ic	0.30	0.05	0.21	0.02
S	-0.14	0.41	-0.14	0.41
log(ED)	-	-	-0.44	0.55
Multivariate (a)	1.19-0.33log(MPD)-0.22S+0.02Ic		-	-
Multivariate (b)	1.10-0.35log(MPD)-0.24S+0.01Ic		1.10-0.37log(ED)-0.30S+0.02Ic	0.83

Table 1: Regressions for SI - model using negative exponential function. Uni- and multivariate regressions, describing the variation in R_0 using various phylogenetic metrics for simulations of Scenario 1 and 2, using $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$ in simulated communities with a host species richness of 3, 4, 5, 7 and 10 species. Table shows the standardized effects size of each metric and its explanatory power for R_0 . All parameter estimations were significant. Model (a) in Sc 1 had an AIC of 51.1, and in Sc 2 model (a) AIC = 1672 and for model (b) AIC = -280.

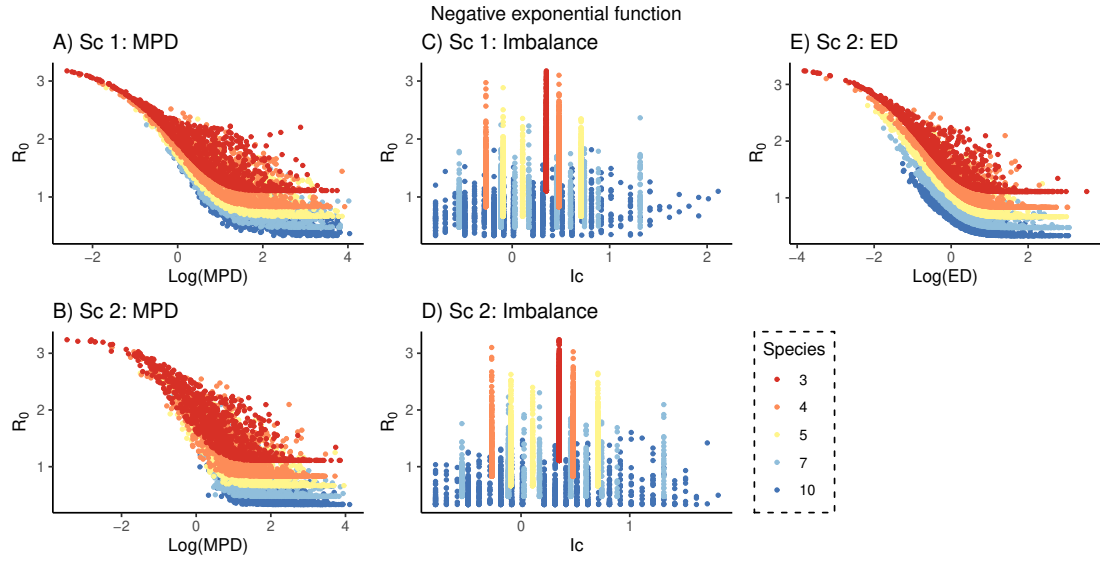


Figure 15: Negative exponential relationship for c and PPd . A,B: Mean Pairwise Distance (MPD) of the community is inversely related to R_0 in Scenario 1 (A) and Scenario 2 (B). C,D: Imbalance of the phylogenetic tree representing the host species community shows high variation in its effect on R_0 , with increasing variation when species richness, S , is increased. E: Evolutionary Distinctiveness for the randomly assigned primary host in Scenario 2. Data from simulations with $\psi = 1$, $\delta = 0.1$, $\mu = 0.2$ and $\lambda = 1$. Colours indicate simulated communities with different host richness ($S=3,4,5,7,10$). Intra- and interspecific contact rates, κ_{ij} , are 1.

Pathogen group	Sc 1: Model (a) R^2	Sc 2: Model (a) R^2	Sc 2: Model (b) R^2
Bacteria	0.87	0.82	0.85
Funghi	0.78	0.74	0.75
Oomycetes	0.94	0.88	0.93
Insects	0.86	0.81	0.84
Mites	0.93	0.87	0.92
Mollusks	0.9619	0.90	0.9624
Nematodes	0.90	0.84	0.88
Viruses	0.49	0.44	0.43
Plants	0.93	0.87	0.92

Table 2: Multivariate regressions of models (a) and (b) for all the pathogen groups from Gilbert et al. (2012). Mollusks is the only pathogen group that shows the same improved R^2 in model (b) compared to model (a) as in our inverse-model. All parameter estimates were significant.

5 Application to Gilbert *et al.* (2012) functions

Here we show the results of the model parameterized by Gilbert et al. (2012) functions of various pathogen groups. The figures for viruses and bacteria are shown in the main text. Using the general function $\text{logit}(c_{ij}) = b_0 - b_1 * \log_{10}(PPd_{ij} + 1)$, we used for funghi: $b_0 = 4.40$, $b_1 = -3.32$ ($n = 95$); For oomycetes: $b_0 = 2.08$, $b_1 = -2.67$ ($n = 32$); For insects: $b_0 = 3.24$, $b_1 = -2.70$ ($n = 637$); For mites: $b_0 = 1.96$, $b_1 = -2.17$ ($n = 87$); For mollusks: $b_0 = -0.47$, $b_1 = -1.14$ ($n = 37$); For nematodes: $b_0 = 2.72$, $b_1 = -2.62$ ($n = 70$); For plants: $b_0 = 2.00$, $b_1 = -2.28$ ($n = 53$). Table 2 shows the results for the univariate and multivariate regressions for each of the pathogen groups. Mollusks are the only group that show an increase in R^2 in model (b) for scenario 2, as in our inverse model example in the main text. Overall viruses had the lowest explanatory power, which we discuss in the main text.

References

- Gilbert, G. S., Magarey, R., Suiter, K., and Webb, C. O. (2012). Evolutionary tools for phytosanitary risk analysis: Phylogenetic signal as a predictor of host range of plant pests and pathogens. *Evolutionary Applications*, 5(8):869–878.
- Lively, C. M. (2010). The effect of host genetic diversity on disease spread. *American Naturalist*, 175(6):17–20.
- Poullain, V. and Nuismer, S. L. (2012). Infection genetics and the likelihood of host shifts in coevolving host-parasite interactions. *American Naturalist*, 180(5):618–628.

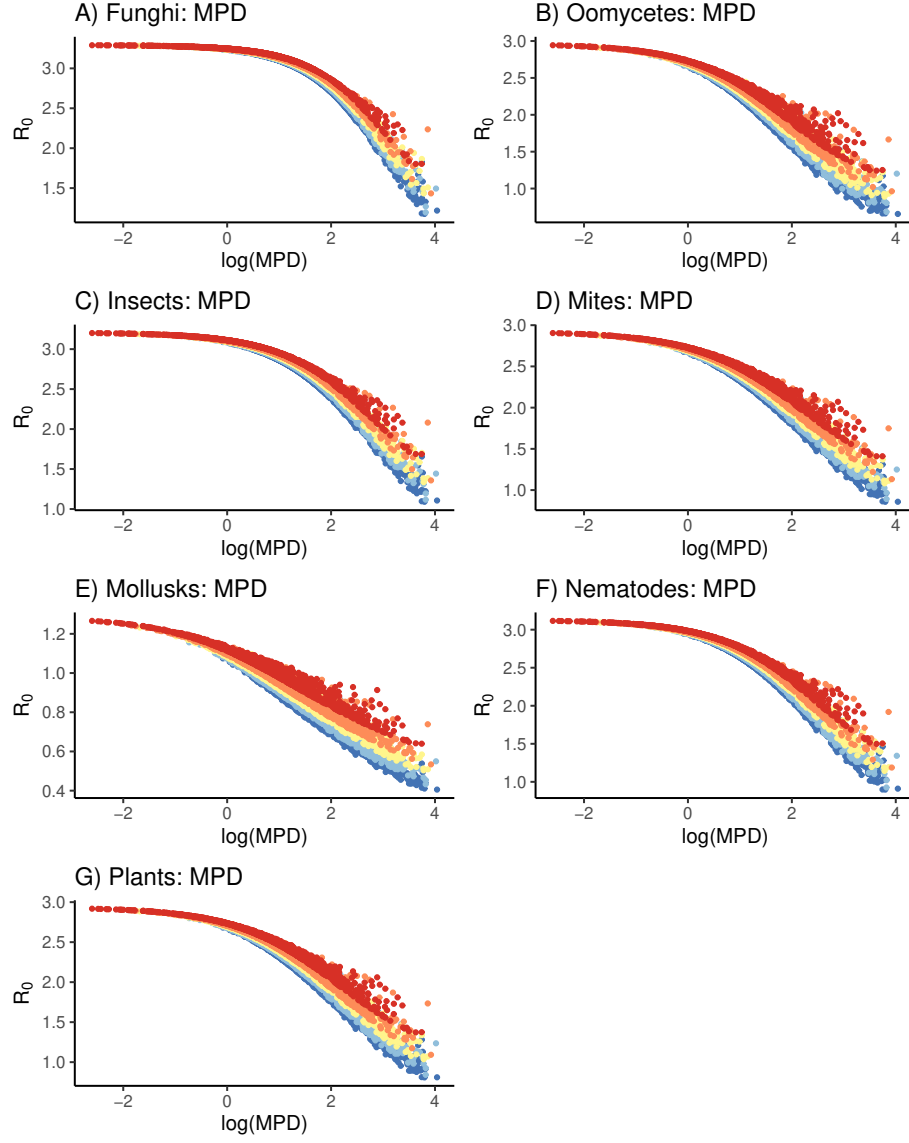


Figure 16: Mean Pairwise Distance (MPD) of the host community is inversely related to R_0 . For different pathogen types, data used from Gilbert et al. (2012): A: $\text{logit}(c_{ij}) = 4.40 - 3.32 * \log_{10}(PPd_{ij} + 1)$, $n = 95$. B: $\text{logit}(c_{ij}) = 2.08 - 2.67 * \log_{10}(PPd_{ij} + 1)$, $n = 32$. C: $\text{logit}(c_{ij}) = 3.24 - 2.70 * \log_{10}(PPd_{ij} + 1)$, $n = 637$. D: $\text{logit}(c_{ij}) = 1.96 - 2.17 * \log_{10}(PPd_{ij} + 1)$, $n = 87$. E: $\text{logit}(c_{ij}) = -0.47 - 1.14 * \log_{10}(PPd_{ij} + 1)$, $n = 37$. F: $\text{logit}(c_{ij}) = 2.72 - 2.62 * \log_{10}(PPd_{ij} + 1)$, $n = 70$. G: $\text{logit}(c_{ij}) = 2.00 - 2.28 * \log_{10}(PPd_{ij} + 1)$, $n = 53$.

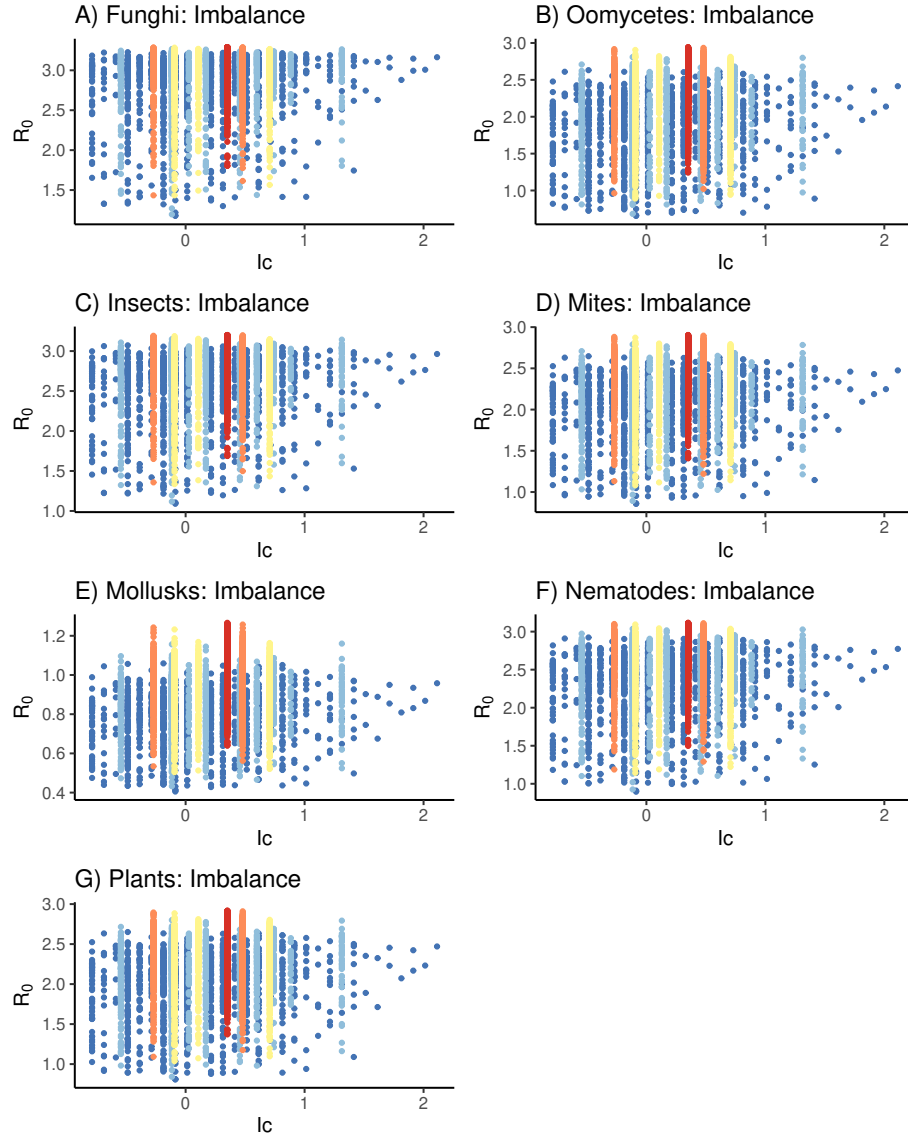


Figure 17: Tree Imbalance (I_c) of the host community is inversely in relationship to R_0 . For different pathogen types, data used from Gilbert et al. (2012): A: $\text{logit}(c_{ij}) = 4.40 - 3.32 * \log_{10}(PPd_{ij} + 1)$, $n = 95$. B: $\text{logit}(c_{ij}) = 2.08 - 2.67 * \log_{10}(PPd_{ij} + 1)$, $n = 32$. C: $\text{logit}(c_{ij}) = 3.24 - 2.70 * \log_{10}(PPd_{ij} + 1)$, $n = 637$. D: $\text{logit}(c_{ij}) = 1.96 - 2.17 * \log_{10}(PPd_{ij} + 1)$, $n = 87$. E: $\text{logit}(c_{ij}) = -0.47 - 1.14 * \log_{10}(PPd_{ij} + 1)$, $n = 37$. F: $\text{logit}(c_{ij}) = 2.72 - 2.62 * \log_{10}(PPd_{ij} + 1)$, $n = 70$. G: $\text{logit}(c_{ij}) = 2.00 - 2.28 * \log_{10}(PPd_{ij} + 1)$, $n = 53$.