Supplementary Information

Host phylogeny can determine parasite host-shift dynamics

Jan Engelstädter & Nicole Z. Fortuna

October 9, 2017

1 Supplementary methods

1.1 Coinfection

In the basic model, a host that is infected cannot be infected by another parasite. Here, we relax this assumption but we still assume that the presence of a parasite will make it harder for new parasites to infect the same host. This is achieved by assuming that the probability of successful establishment of a new parasite is multiplied by the term

$$(1-\sigma)^n,$$

where n is the number of pre-existing parasites on the host branch to which a new parasite is about to switch. The parameter σ is a measure for the strength of competitive exclusion and can take values ranging from 0 (no reduction in establishment probability caused by existing parasites) to 1 (coinfection impossible, corresponding to basic model).

1.2 Correlation between host and parasite genetic distance

We used the correlation between host and parasite phylogenetic distances as a measure to quantify the distribution of parasites within the clade of host species (see Figures 2 and S1). For this measure, we first computed the matrices of phylogenetic distances (i.e., the total branch length connecting any two species) between all extant host species, and the corresponding matrix for all extant parasite species. We then calculated Pearson's product-moment correlation coefficient between the phylogenetic distances of all pairs of parasite species and the phylogenetic distances of the corresponding pairs of host species that the parasites infect. Note that this statistic is only defined when there are at least three parasite species and that in this case, the correlation coefficient is always either 1 (when the parasite tree and the tree of associated host species have the same topology) or $-1/2$ (when they do not).

1.3 Host subtree analyses

For any given host tree, we first computed again the pairwise phylogenetic distance matrix. We then performed a hierarchical cluster analysis on this distance matrix using the `hclust` (R package `stats`) function with standard settings. Next, we applied the `cutree` (R package `stats`) to this clustering in order to split the tree into subtrees with specified heights. Using this partitioning into subtrees, we determined for each subtree the frequency of hosts that are infected by the parasites. We also calculated the Shannon index of the distribution of host species among the subtrees. This was done using the formula $H = -\sum_{i=1}^{n} p_i \ln p_i$, where n is the number of subtrees and p_i the number of host species in subtree i divided by the total number of species in the tree.

1.4 Impact of host net diversification and turnover

In addition to the standard set of host trees and a set with increasing carrying capacity, we also analysed eight additional sets of host trees that varied in their patterns and rates of diversification. For these sets, we varied the speciation rate λ and the extinction rate μ in a way that produced (together with the standard set) all nine combinations of three different net diversification rates ($D = \lambda - \mu$) and turnover rates ($T = \mu/\lambda$):

		T		
		0.333	0.5	0.6
D	0.25	λ = 0.375, μ = 0.125	λ = 0.5, μ = 0.25	λ = 0.625, μ = 0.375
	0.5	λ = 0.75, μ = 0.25	λ = 1.0, μ = 0.5	λ = 1.25, μ = 0.75
	0.75	λ = 1.125, μ = 0.375	λ = 1.5, μ = 0.75	λ = 1.875, μ = 1.125

Each of these nine sets consists of 100 randomly generated trees, all initialised with a single species and simulated for 100 time units. For each set, the carrying capacity parameter K was adjusted so that the expected equilibrium number of species would always be $\hat{N}$ = 100. This was achieved using the formula $K = \lambda\hat{N}/(\lambda - \mu)$.

2 Random effect model results on host tree impact

It is clear visually that under the phylogenetic distance effect, individual host trees exert a major influence on the distribution of parasite infection frequencies (see Figure 3A). To lend some statistical support to this claim, we fitted a linear random effect models to our simulation data. The response variable is the fraction of infected host species at the end of the simulations and two random effects are considered: the host tree and, nested within the host tree, the first infected host branch from which the parasite spread was initiated. The model thus has the form

$$f_{ijk} = \bar{f} + T_i + B_{ij} + R_{ijk}, \tag{1}$$

where p_{ijk} is the fraction of infected host species in a given simulation run, $\bar{f}$ is the mean infection frequency across all simulations, T_i is the effect of tree i on this frequency, B_{ij} is the effect of host branch j within tree i on this frequency, and R_{ijk} is the effect of the individual simulation run k on tree i starting from branch j. In principle, i can take values from 1 to 100, j can take values from 1 to 10 and k can take values from 1 to 10 as well (see Methods). However, since the parasites did not survive in all of these 10,000 simulations, not all combinations of i, j and k yield valid data points. (E.g., with the standard set of host trees and the standard PDE parameter set, the number of simulations where the parasites survived is 5398.)

We fitted this model in R using the function lmer (package lme4 version 1.1-13; Bates et al. 2015), with standard settings. For the simulations run under the standard PDE parameter set, host trees were found to explain 57% of the total variance in infection frequencies (0.018 out of 0.032), whereas the initial host branch did not explain any of the variance. By contrast, with the no-PDE parameter set, host trees explained only 32% of the total variance (0.003 out of 0.01) whilst the initial host branch again did not explain any of the variance in this model. It might be surprising at first that a large fraction of the variance in infection frequencies is explained by host trees even in the complete absence of the phylogenetic distance effect. However, this observation is explained by the fact that the dynamics of parasite spread are influenced by the number of host species through time and that this varies with each host tree. With both the standard PDE and no-PDE parameters, the full model does not provide a better fit than a model without initial branch as a random effect (chi-square test: $p \approx 1$), but the full model fits the data significantly better than a model without any random effects ($p \ll 0.001$).

3 Supplementary figures

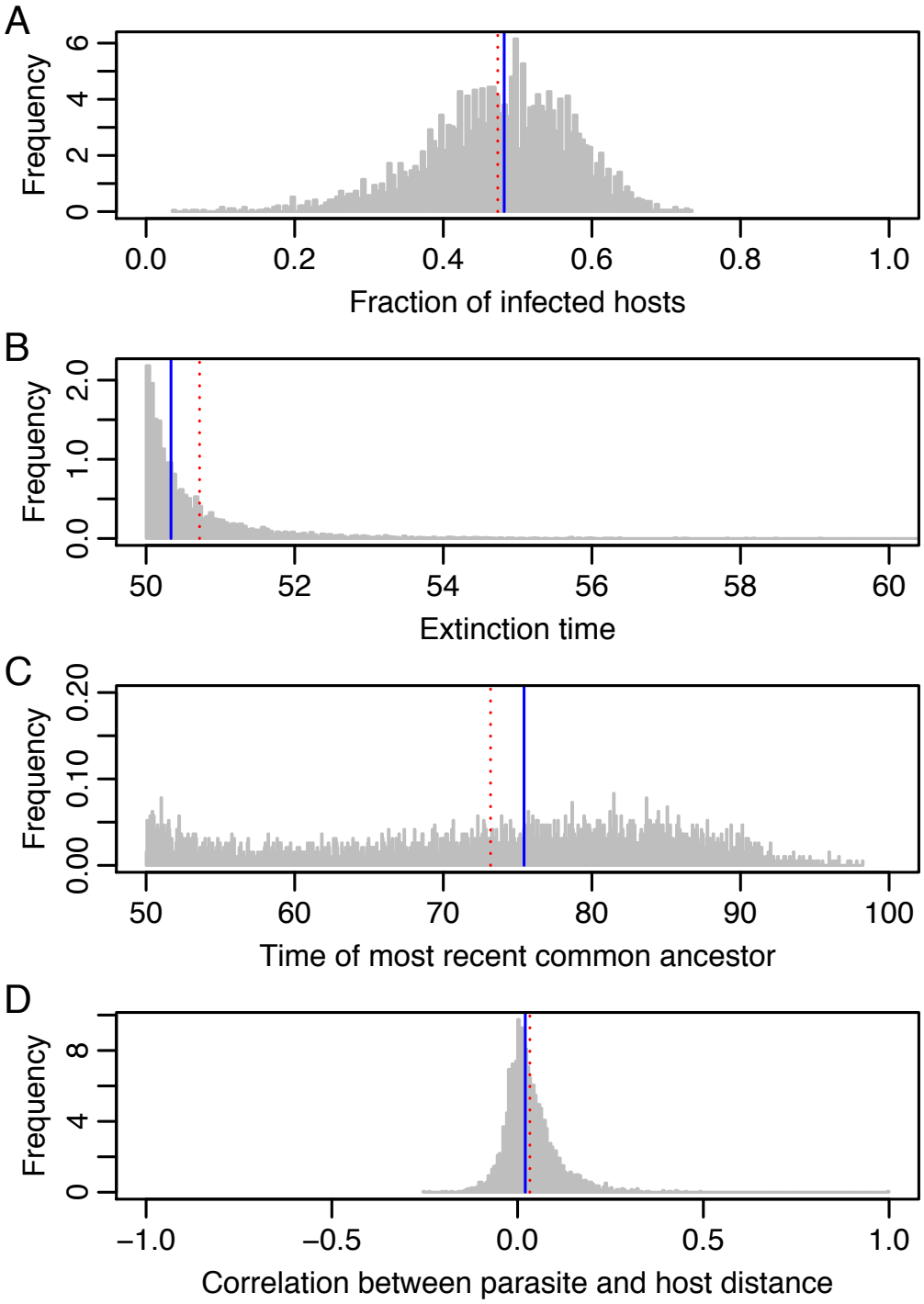

Figure S1: Summary statistics for simulations in the absence of the phylogenetic distance effect. In these simulations, the standard no-PDE parameter and host tree set was used. Panel (A) shows the distribution of the fraction of infected host species across the 10,000 simulations, contingent on parasite survival. Panel (B) shows the distribution of parasite extinction times when the parasite did not survive, following its introduction at time 50. Panel (C) shows the distribution of the time of the most recent common ancestor of all surviving parasite species. In panel (D), the distribution of the correlation between parasite and host phylogenetic distances is shown. In all plots, the solid blue line indicates the median and the dashed red line the mean of the distributions.

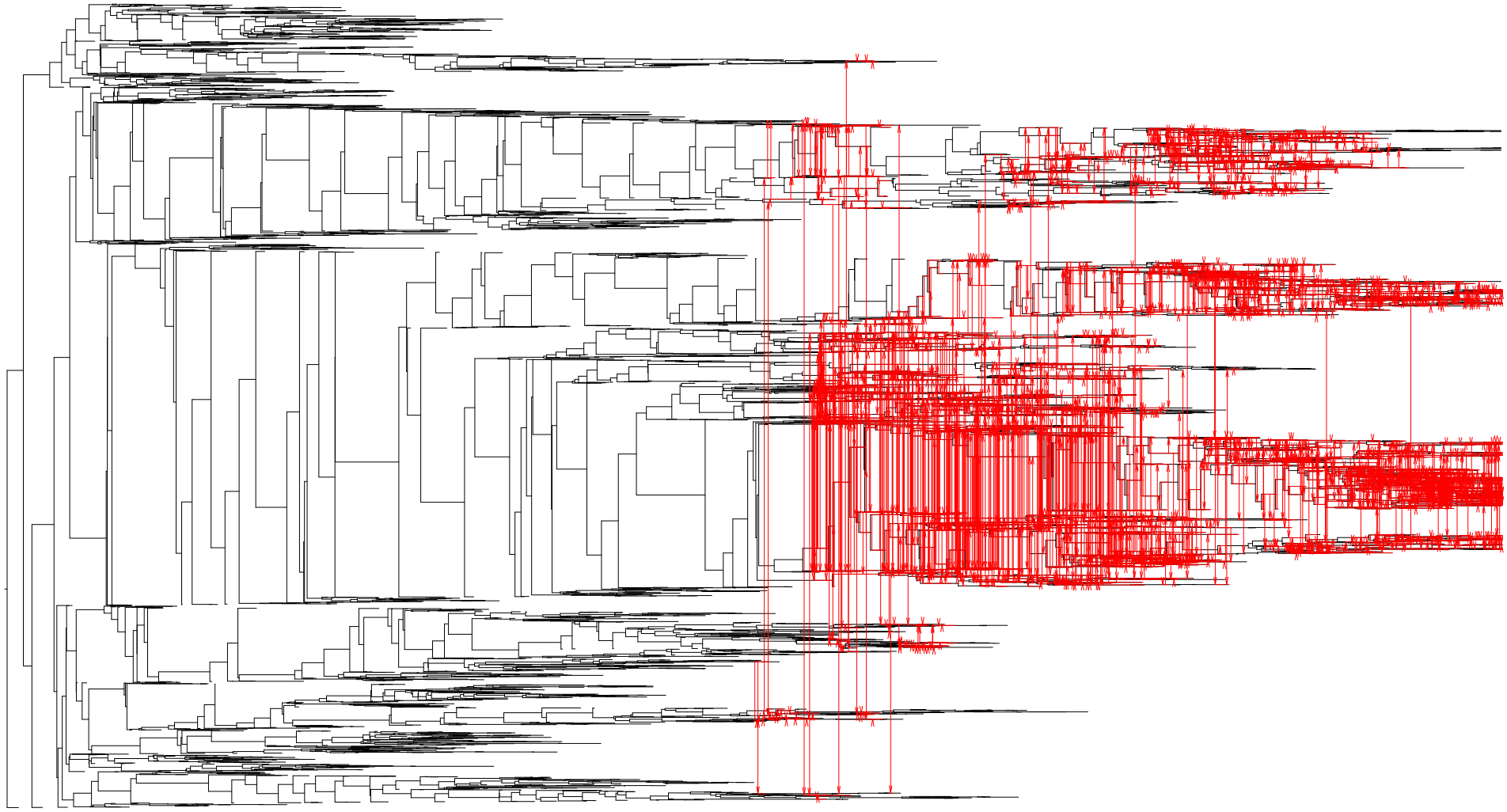

Figure S2A: Example cophylogeny with the standard PDE parameter set, for host tree no.1.

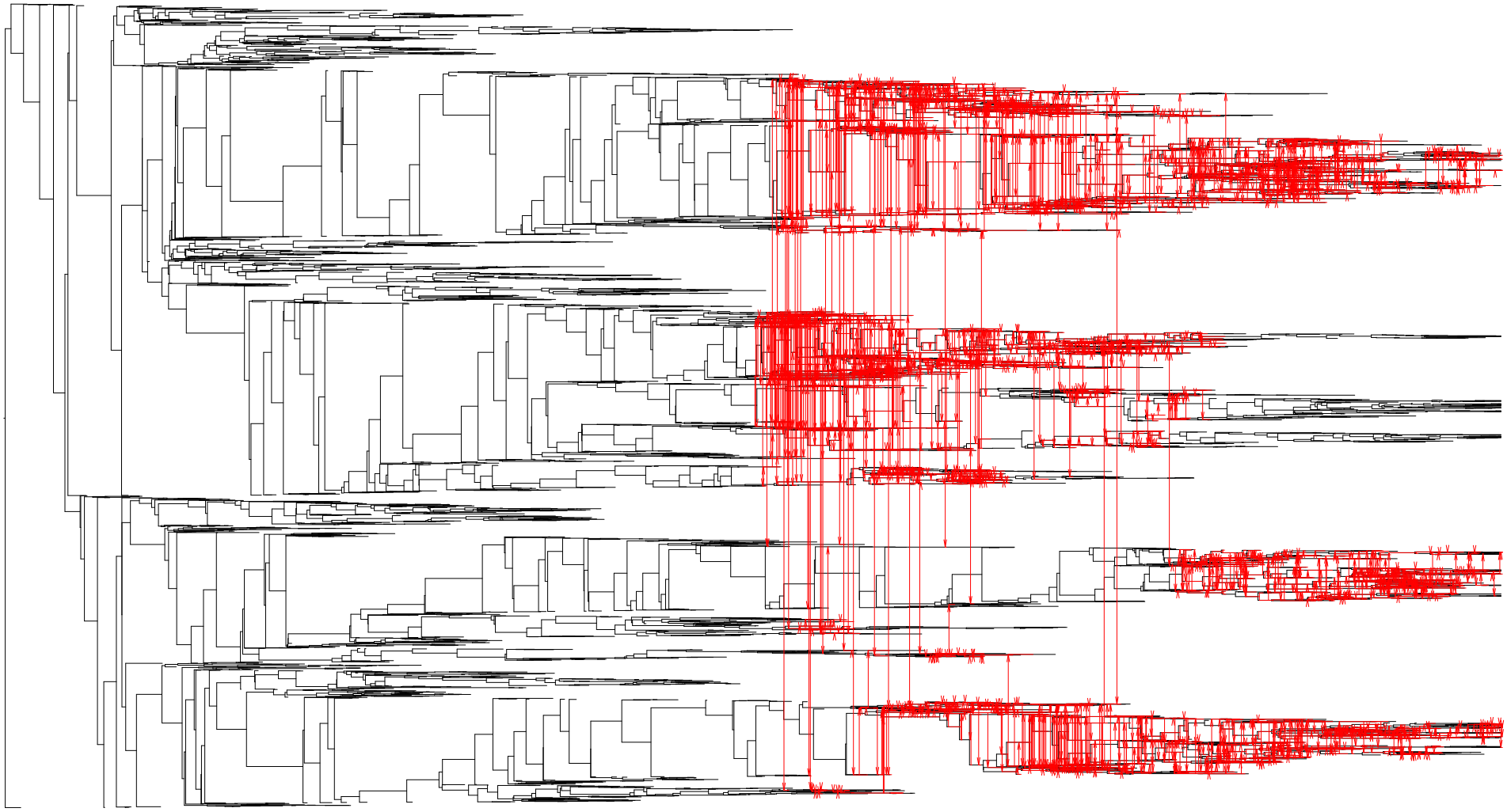

Figure S2B: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.5.

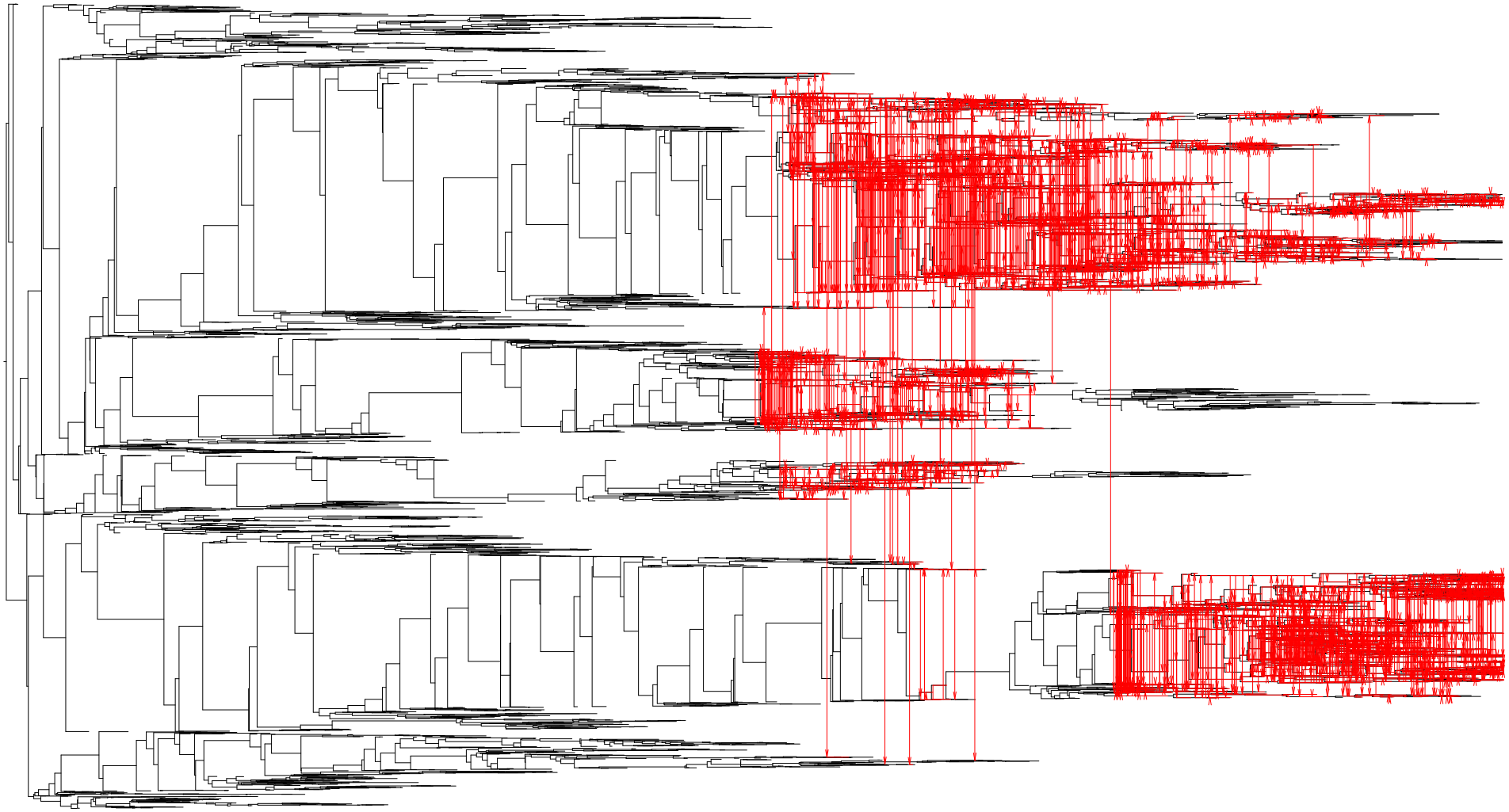

Figure S2C: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.25.

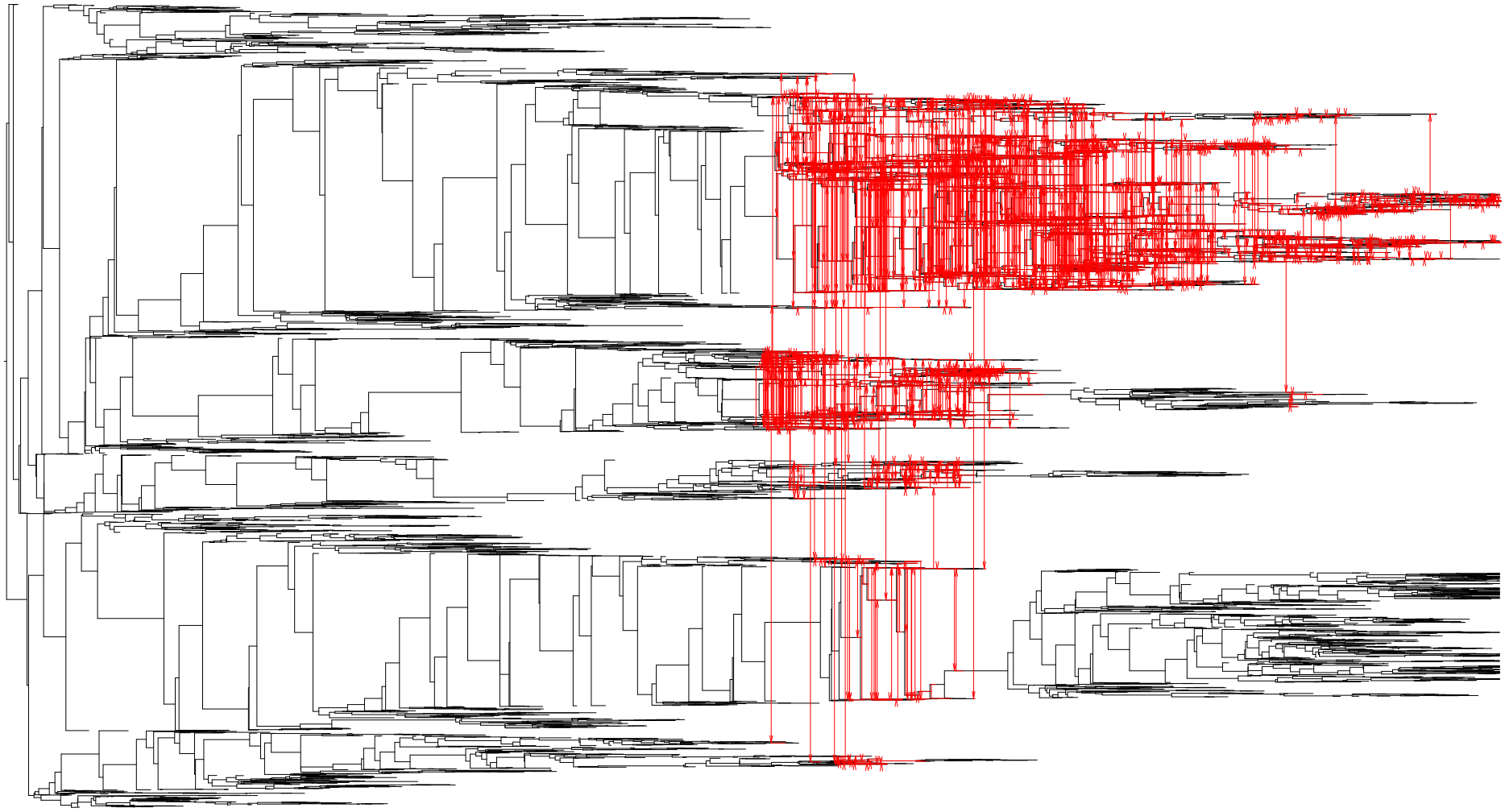

Figure S2D: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.25.

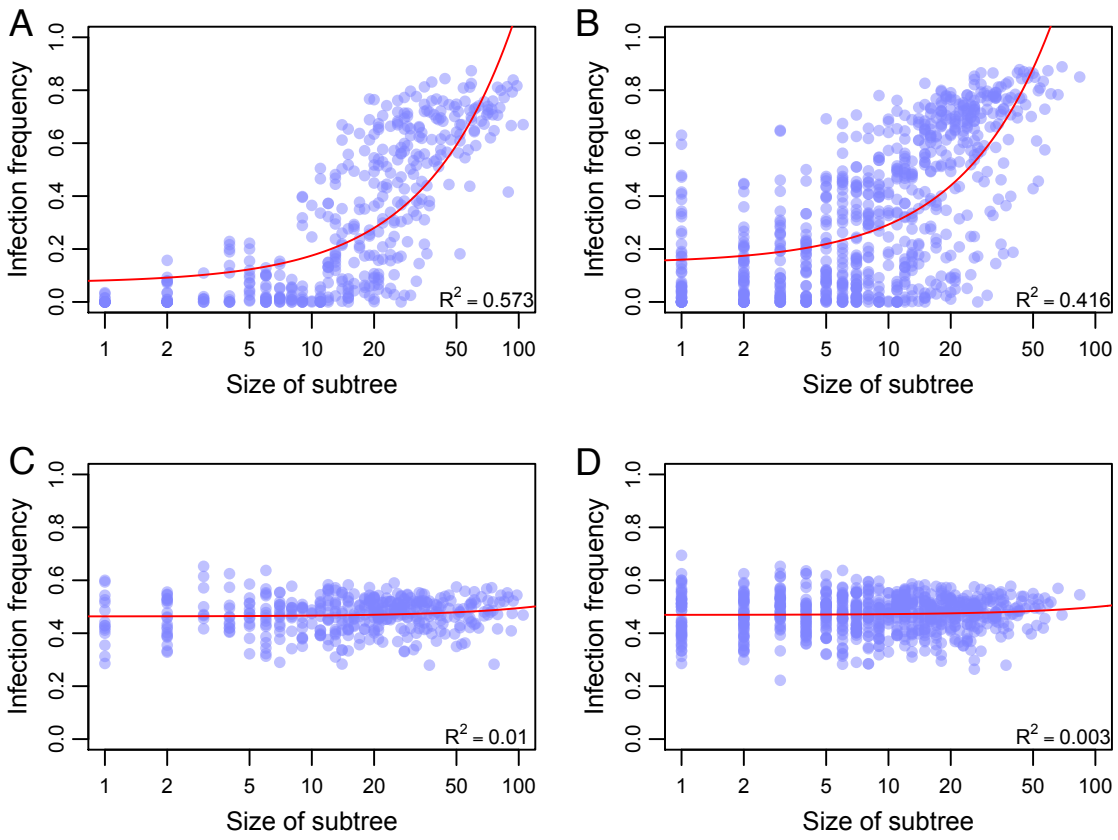

Figure S3: Fraction of infected hosts within host subtrees against the size of these subtrees with (A,B) and without (C,D) the phylogenetic distance effect. Each dot represents the mean infection frequency (across 100 simulations) of a subtree from one of the 100 trees forming the standard host tree set. Partitioning of host trees into subtrees was performed as described in section 1.3, with the height parameter set to either 100 (plots A and C, corresponding to few large subtrees) or 50 (plots B and D, corresponding to more but smaller subtrees). Red lines show the fit of a linear regression with R^2 values indicated. All parameters take standard values.

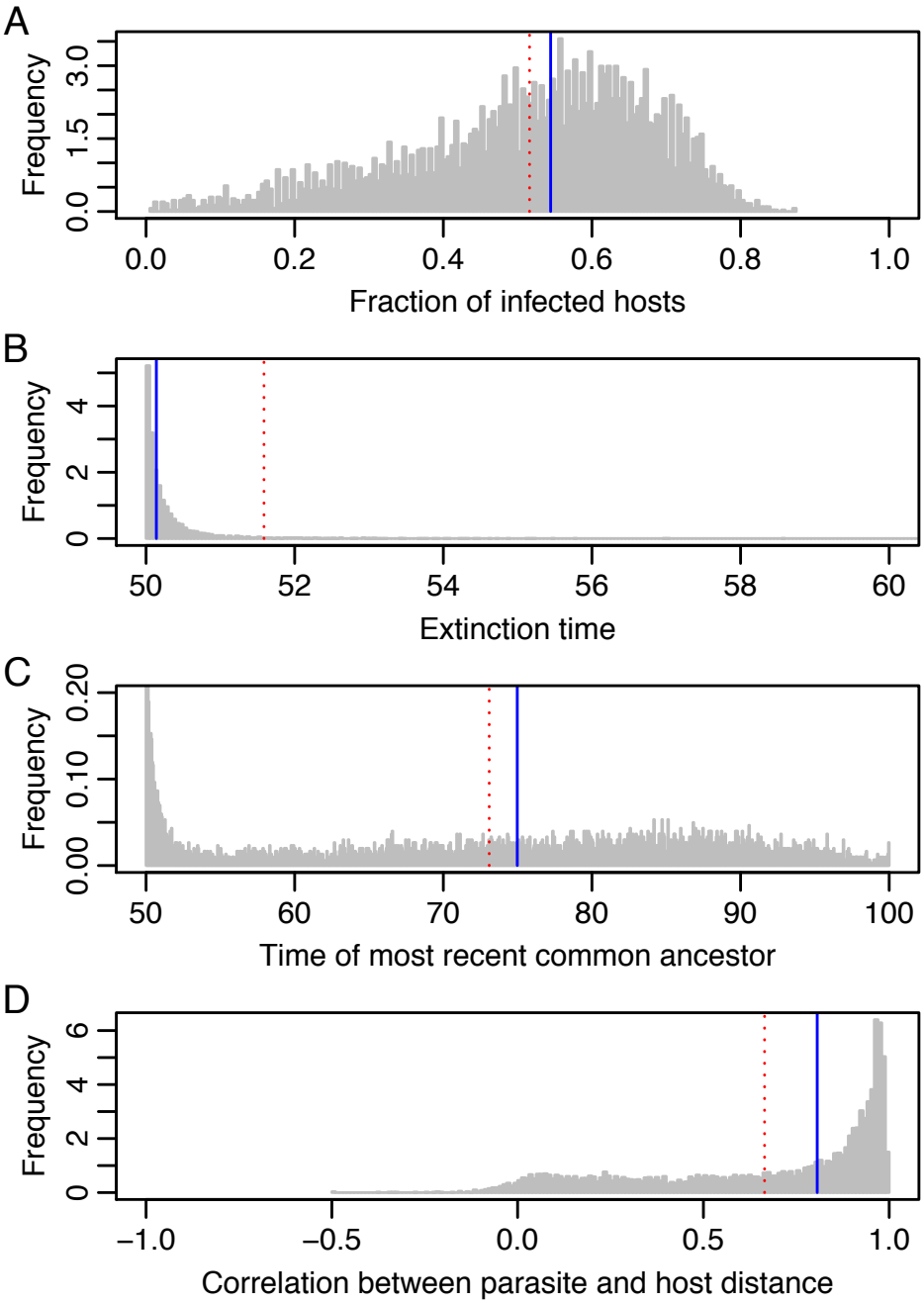

Figure S4: Summary statistics for simulations in the presence of the phylogenetic distance effect. This figure is the same as Figure 2 in the main text except that parasite transmission and extinction rates are doubled relative to the standard parameter set. Panel (A) shows the distribution of the fraction of infected host species across the 10,000 simulations, contingent on parasite survival. Panel (B) shows the distribution of parasite extinction times when the parasite did not survive, following its introduction at time 50. Panel (C) shows the distribution of the time of the most recent common ancestor of all surviving parasite species. In panel (D), the distribution of the correlation between parasite and host phylogenetic distances is shown. In all plots, the solid blue line indicates the median and the dashed red line the mean of the distributions. All parameters take standard values except $\nu = 2$ and $\beta = 1$.

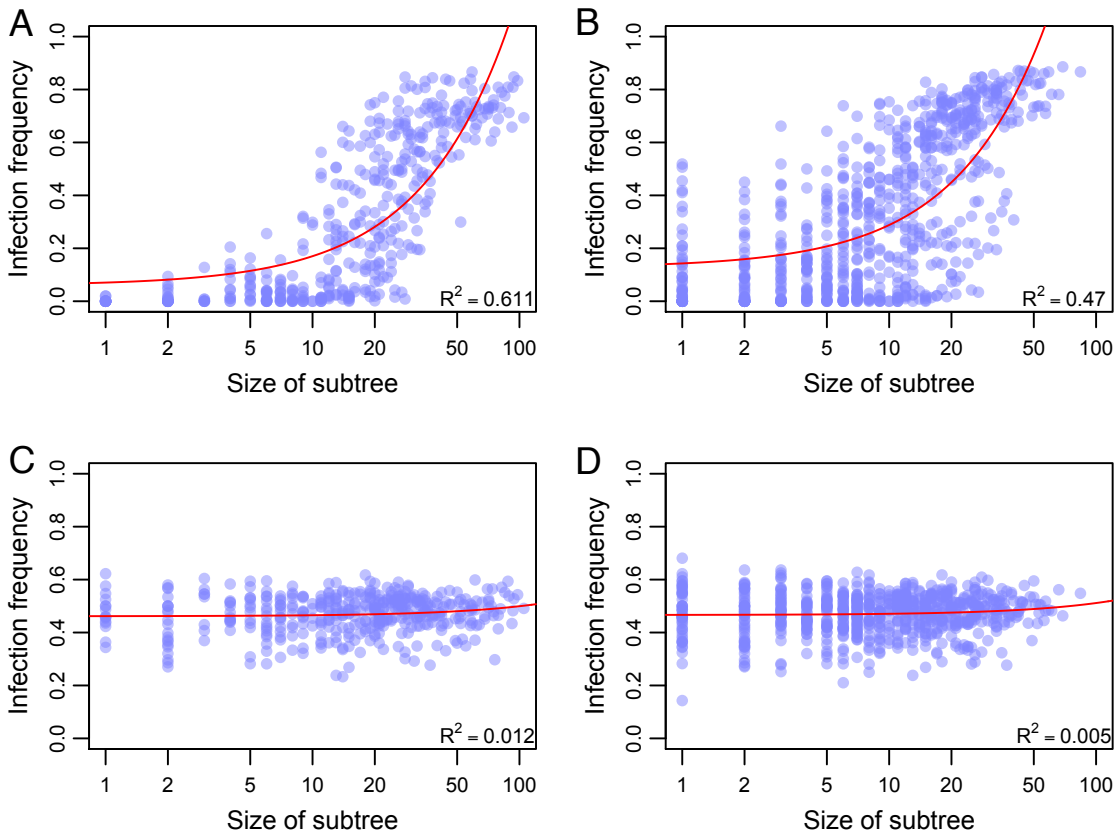

Figure S5: Fraction of infected hosts within host subtrees against the size of these subtrees with (A,B) and without (C,D) the phylogenetic distance effect. This figure is the same as Figure S3 except that parasite transmission and extinction are twice that of the standard parameter set. Each dot represents the mean infection frequency (across 100 simulations) of a subtree from one of the 100 trees forming the standard host tree set. Partitioning of host trees into subtrees was performed as described in section 1.3, with the height parameter set to either 100 (plots A and C, corresponding to few large subtrees) or 50 (plots B and D, corresponding to more but smaller subtrees). Red lines show the fit of a linear regression with R^2 values indicated. All parameters take standard values except $\nu = 2$ and (A,B) $\beta = 1$, (C,D) $\beta = 0.04$.

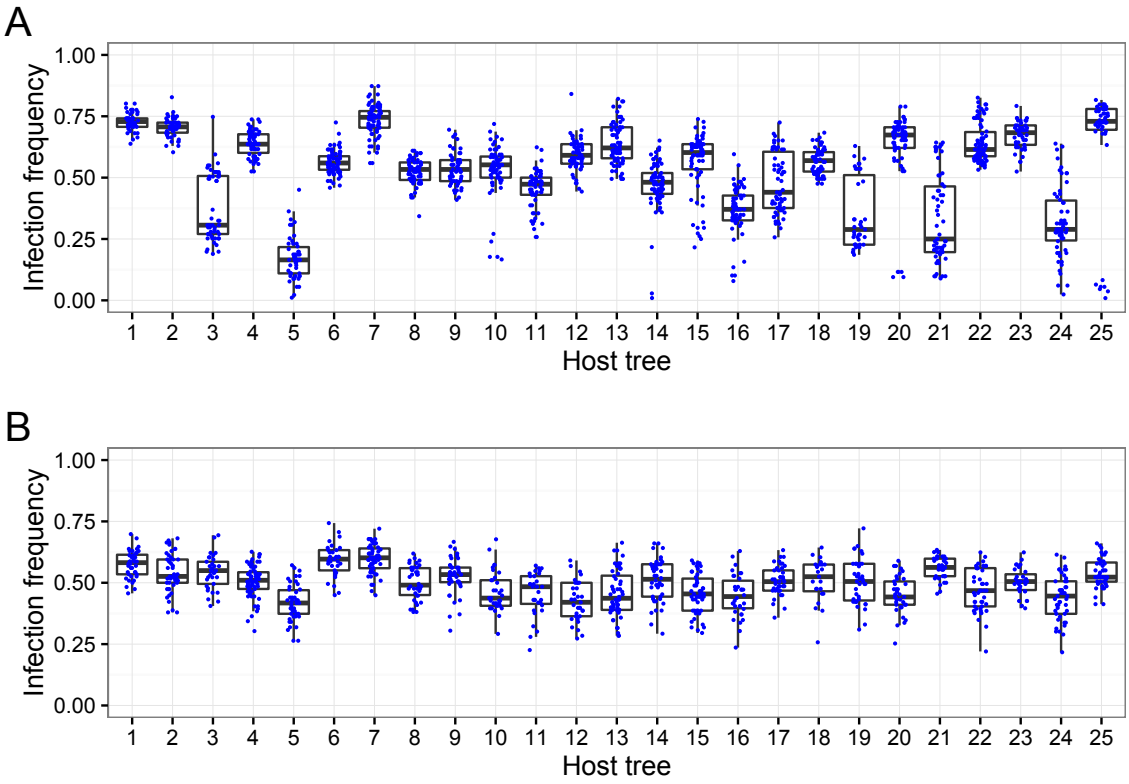

Figure S6: Distributions of infection frequencies with (A) and without (B) the phylogenetic distance effect on the first 25 host trees. This figure is the same as Figure 3 in the main text except that parasite transmission and extinction rates are doubled relative to the standard parameter set. Each dot shows the fraction of infected host species at the end of a simulation run. Simulations in which the parasites did not survive until the end of the simulation are not shown. Boxes show the interquartile range with the horizontal line indicating the median and whiskers indicating the distance from the box to the largest value no further than 1.5 times the interquartile range. All parameters take the standard values except $\nu = 2$ and (A) $\beta = 1$, (B) $\beta = 0.04$.

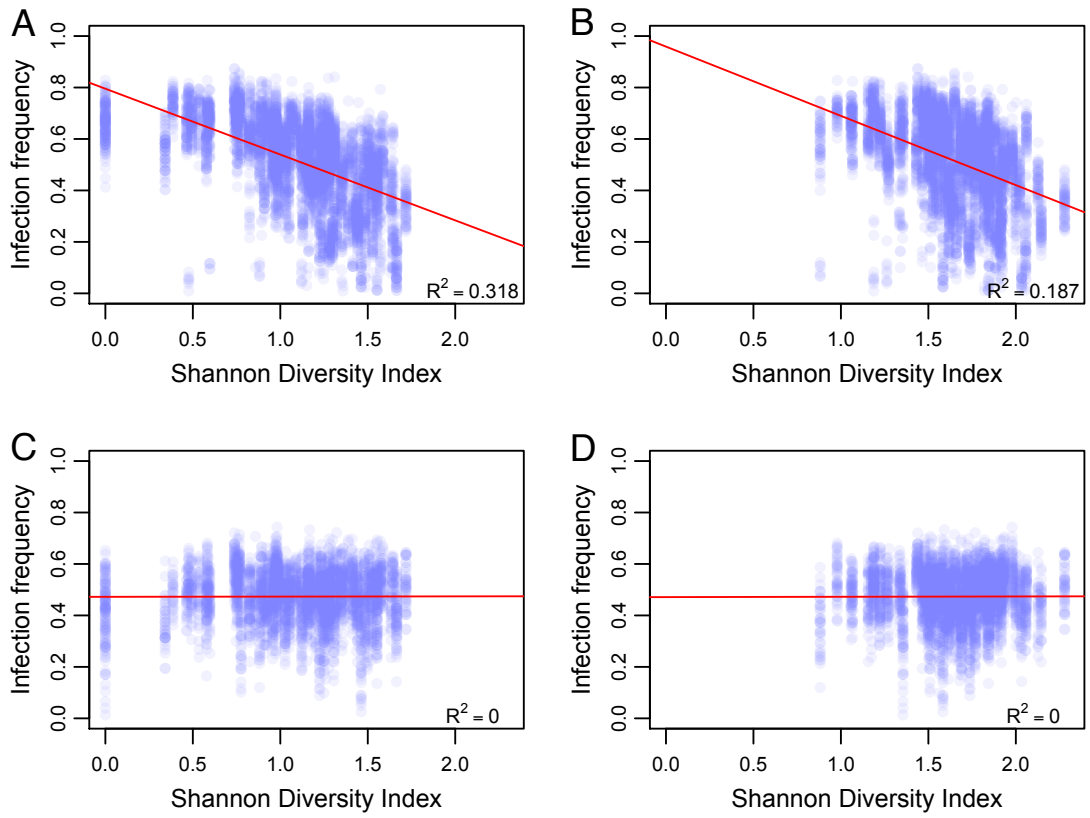

Figure S7: Fraction of infected hosts at the end of simulations against the Shannon index size of host species distribution within the respective host tree, with (A,B) and without (C,D) the phylogenetic distance effect. This figure is the same as Figure 4 in the main text except that parasite transmission and extinction rates are doubled relative to the standard parameter set. Each dot represents the outcome of a single simulation; simulations in which the parasites became extinct were discarded. Partitioning of host trees into subtrees (or clades) and calculating the Shannon index was performed as described in section 1.3, with the height parameter set to either 100 (plots A and C, corresponding to few large subtrees) or 50 (plots B and D, corresponding to more but smaller subtrees). Red lines show the fit of a linear regression with R^2 values indicated. All parameters take the standard values except $\nu = 2$ and (A,B) $\beta = 1$, (C,D) $\beta = 0.04$.

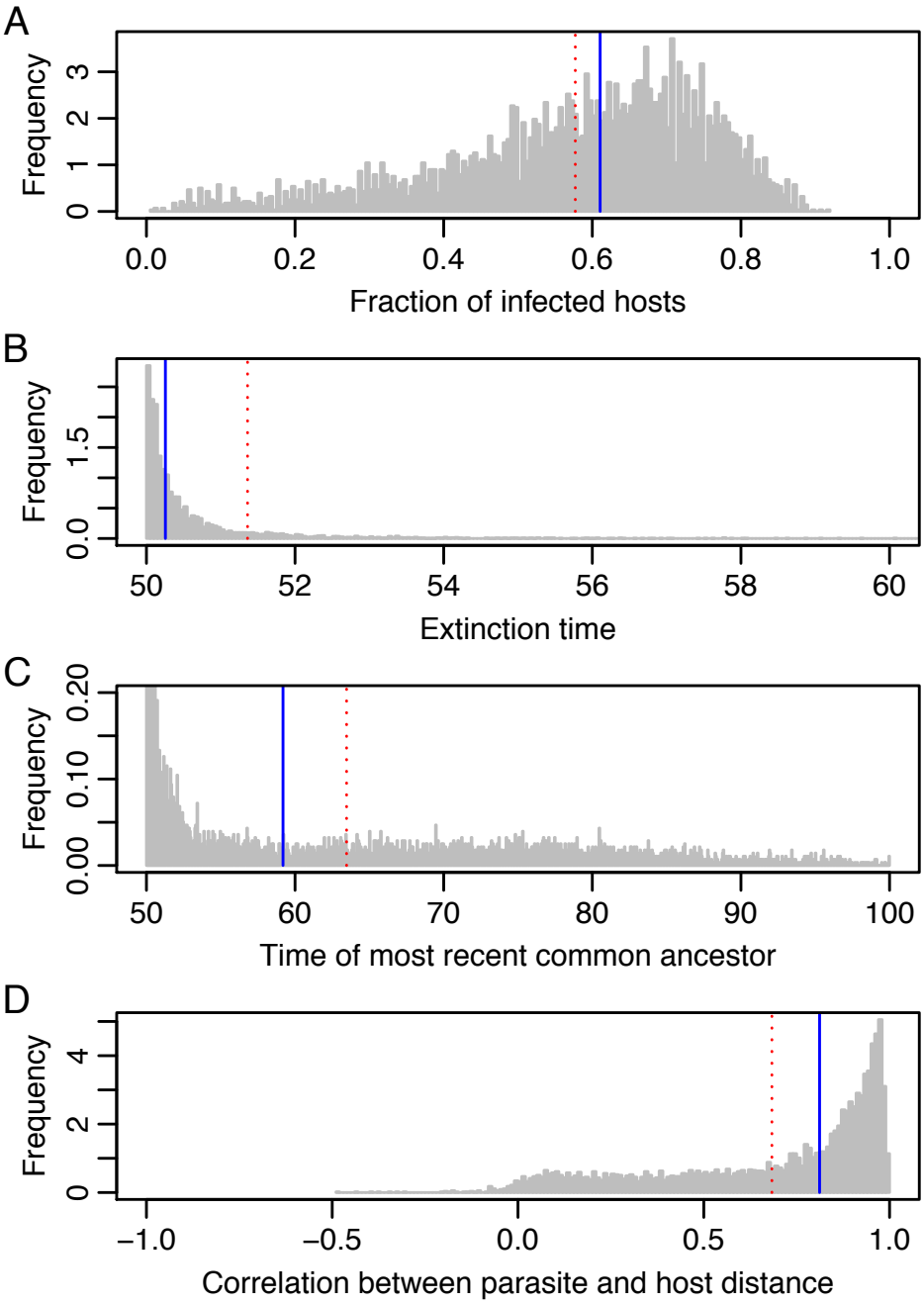

Figure S8: Summary statistics for simulations in the presence of the phylogenetic distance effect. This figure is the same as Figure 2 in the main text except that coinfections are possible (see section 1.1). Panel (A) shows the distribution of the fraction of infected host species across the 10,000 simulations, contingent on parasite survival. Panel (B) shows the distribution of parasite extinction times when the parasite did not survive, following its introduction at time 50. Panel (C) shows the distribution of the time of the most recent common ancestor of all surviving parasite species. In panel (D), the distribution of the correlation between parasite and host phylogenetic distances is shown. In all plots, the solid blue line indicates the median and the dashed red line the mean of the distributions. All parameters take standard values except $\sigma = 0.1$.

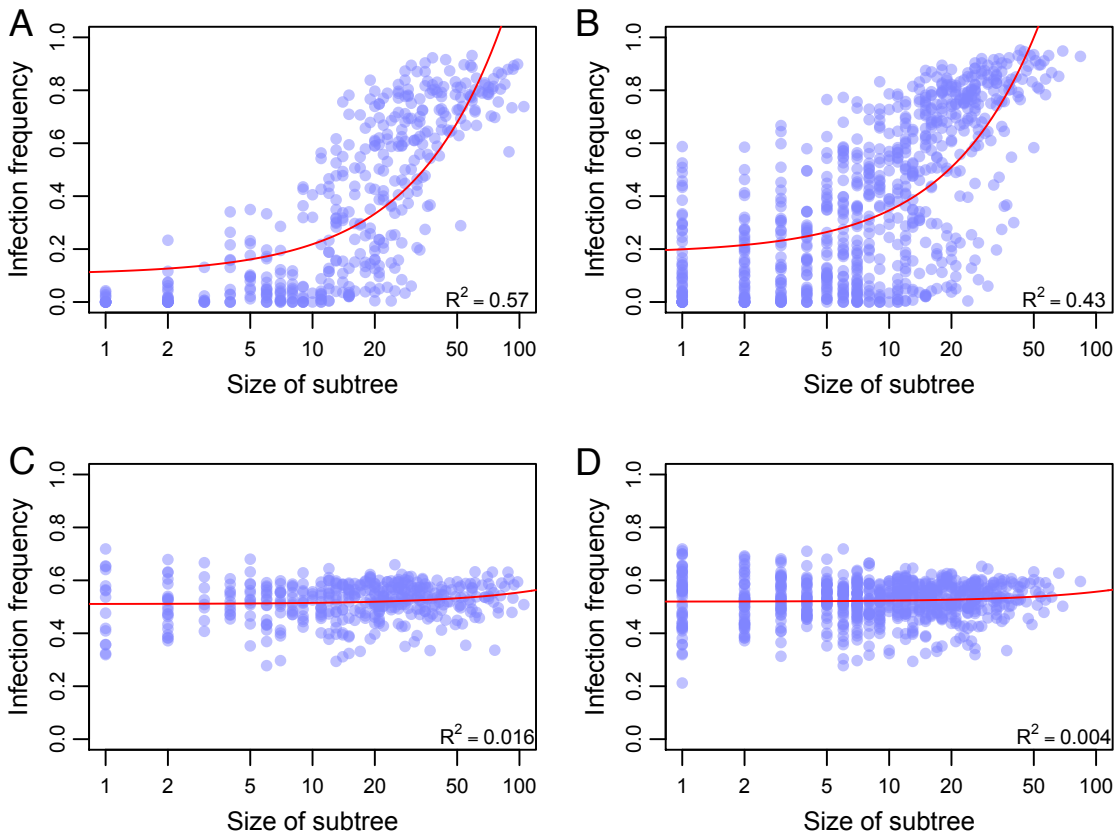

Figure S9: Fraction of infected hosts within host subtrees against the size of these subtrees with (A,B) and without (C,D) the phylogenetic distance effect. This figure is the same as Figure S3 except that coinfections are possible (see section 1.1). Each dot represents the mean infection frequency (across 100 simulations) of a subtree from one of the 100 trees forming the standard host tree set. Partitioning of host trees into subtrees was performed as described in section 1.3, with the height parameter set to either 100 (plots A and C, corresponding to few large subtrees) or 50 (plots B and D, corresponding to more but smaller subtrees). Red lines show the fit of a linear regression with R^2 values indicated. All parameters take standard values except $\sigma = 0.1$.

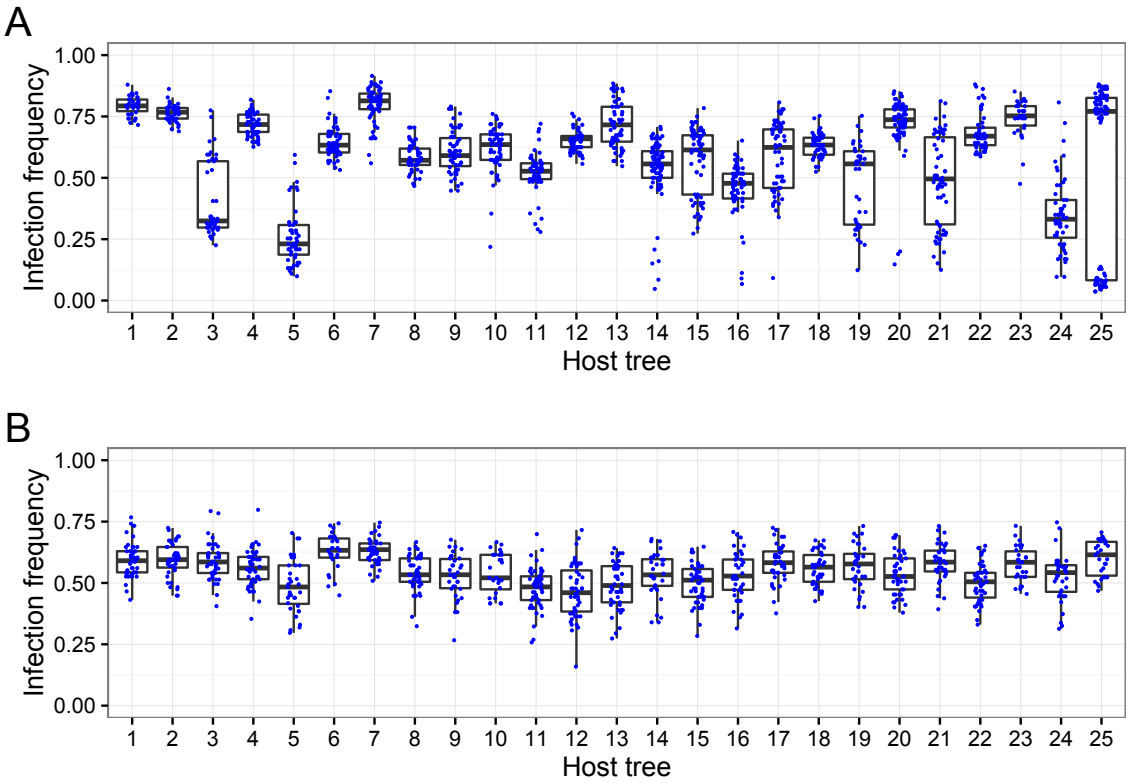

Figure S10: Distributions of infection frequencies with (A) and without (B) the phylogenetic distance effect on the first 25 host trees. This figure is the same as Figure 3 in the main text except that coinfections are possible (see section 1.1). Each dot shows the fraction of infected host species at the end of a simulation run. Simulations in which the parasites did not survive until the end of the simulation are not shown. Boxes show the interquartile range with the horizontal line indicating the median and whiskers indicating the distance from the box to the largest value no further than 1.5 times the interquartile range. All parameters take the standard values except $\sigma = 0.1$.

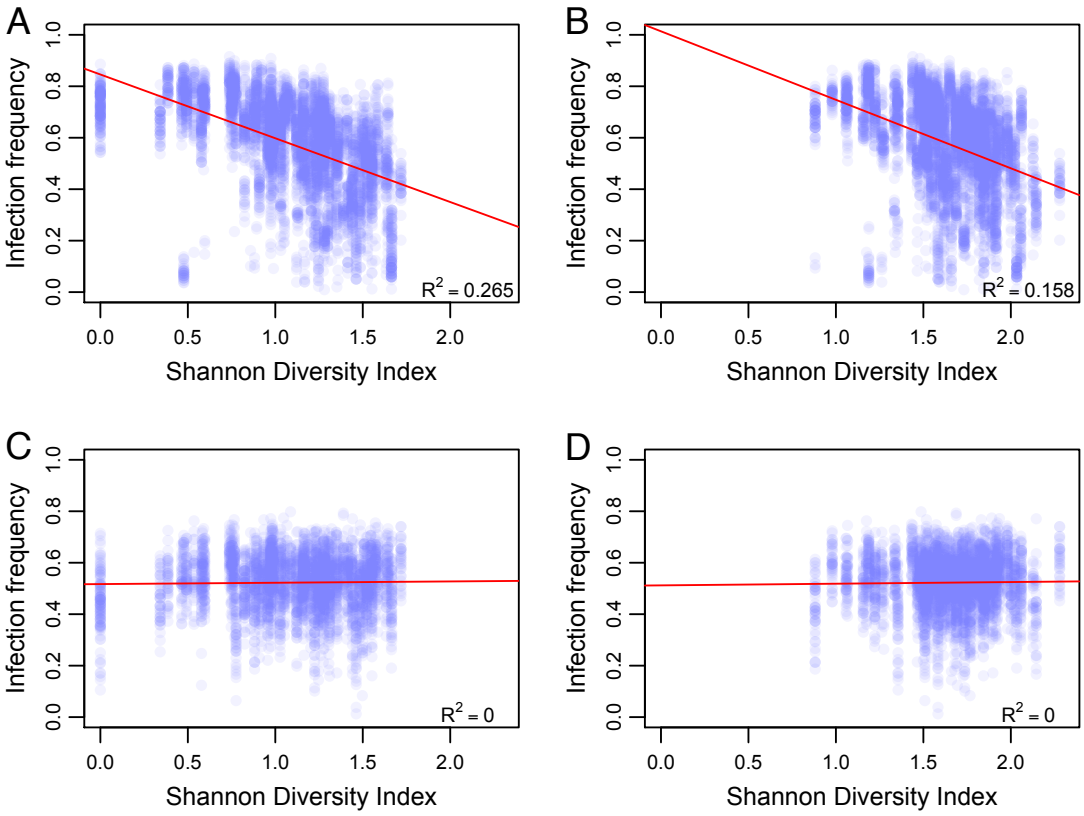

Figure S11: Fraction of infected hosts at the end of simulations against the Shannon index size of host species distribution within the respective host tree, with (A,B) and without (C,D) the phylogenetic distance effect. This figure is the same as Figure 4 in the main text except that coinfections are possible (see section 1.1). Each dot represents the outcome of a single simulation; simulations in which the parasites became extinct were discarded. Partitioning of host trees into subtrees (or clades) and calculating the Shannon index was performed as described in section 1.3, with the height parameter set to either 100 (plots A and C, corresponding to few large subtrees) or 50 (plots B and D, corresponding to more but smaller subtrees). Red lines show the fit of a linear regression with R^2 values indicated. All parameters take standard values except $\sigma = 0.1$.

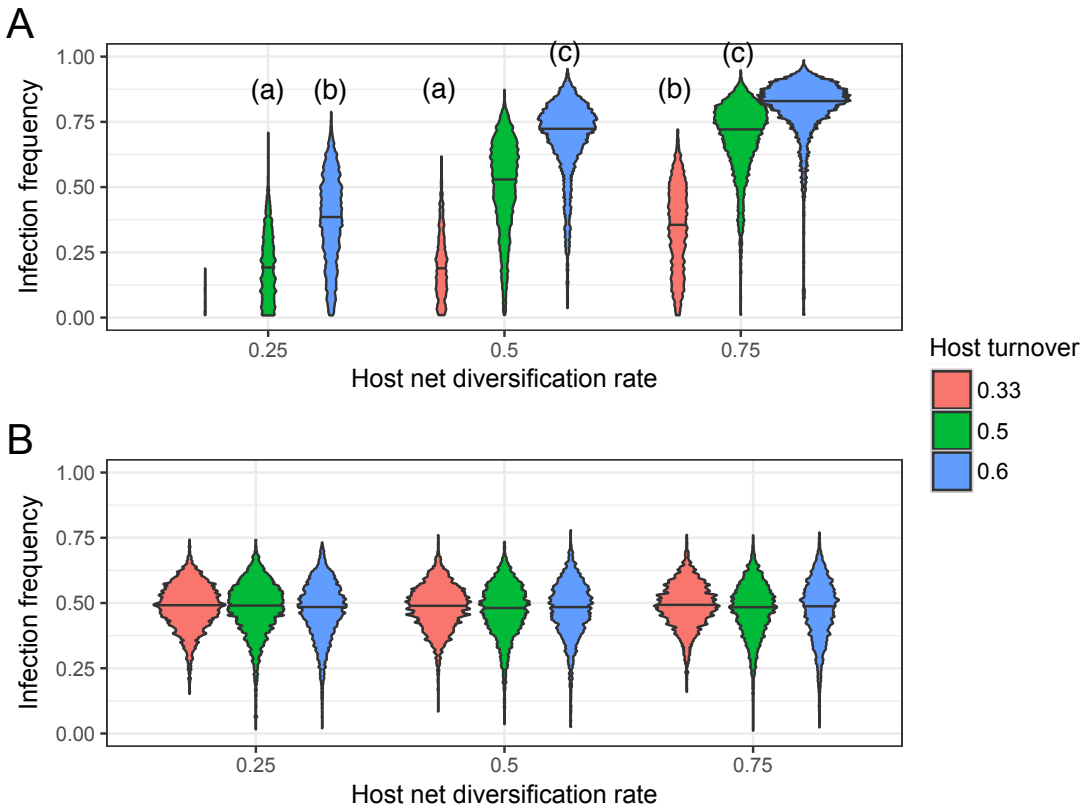

Figure S12: The impact of host net diversification ($\lambda - \mu$) and turnover (μ/λ) on the fraction of infected host species with (A) and without (B) the phylogenetic distance effect. Violins show the distribution of infection frequency, with the total area of each violin being proportional to the number of simulations in which the parasites survived. Letters (a), (b) and (c) indicate parameter combinations with identical values of μ. See section 1.4 for details on simulations and parameter values.