Supplementary Information

Host phylogeny can determine parasite host-shift dynamics

Jan Engelstädter & Nicole Z. Fortuna

1 Model extensions

We implemented three model extensions corresponding to three frequently considered events that could occur during host-parasite diversification but are not part of the basic model. These events are coinfection of an already infected host, parasite loss in the process of cospeciation, and parasite speciation within a single host species. These extensions are all part of the cophy package.

1.1 Coinfection

In the basic model, a host that is infected cannot be infected by another parasite. Here, we relax this assumption but still assume that the presence of a parasite may make it harder for new parasites to infect the same host. Specifically, we assume that the probability of successful establishment of a new parasite is multiplied by the term σ^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 how easily coinfections are established and can range from 0 (coinfection impossible, as in the basic model) to 1 (no reduction in establishment probability caused by existing parasites).

1.2 Parasite loss during cospeciation

In the basic model, parasites always cospeciate whenever their hosts speciate. By contrast, the parasite may also be inherited only by one of the two daughter host species but lost in the other, an event that has been referred to as "missing the boat" or "lineage sorting". To implement this event, we define a new parameter δ as the probability that the parasite is lost in one, randomly chosen host lineage during host speciation. $\delta = 0$ thus corresponds to the basic model of perfectly faithful parasite cospeciation.

1.3 Within-host parasite speciation

In the basic model, the parasites only speciate by cospeciation with their hosts or through host shifts. In reality, there may also be speciation events of parasites within their host species (often termed a "duplication"). To incorporate such events into our model, we define a new parameter κ as the rate of within-host-lineage speciation of parasites. This rate is assumed to be constant through time and independent of the number of parasite species already present in a host. When $\kappa = 0$, the basic model is recovered.

2 Supplementary methods

2.1 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 $K = \lambda\hat{N}/(\lambda - \mu)$.

2.2 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).

2.3 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 `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. To calculate the Shannon index, we used the `diversity` function implemented in the R package `vegan` v.2.4-5 (Oksanen et al. 2017).

3 Supplementary results

3.1 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 ($p \ll 0.001$).

3.2 The impact of tree imbalance on infection frequencies

We used the Colless index (Colless 1982; corrected version by Heard 1992), to characterise the imbalance of our host trees. Briefly, this index uses the difference in size (number of terminal descendents) between the two clades branching off from each node, then sums these differences over all nodes, and finally normalises so that values fall between 0 (perfectly symmetrical tree) and 1 (maximally asymmetrical tree).

Figure S3 shows how the Colless indices for our standard set of 100 host trees relates to the frequencies of infected host species. When we used the full trees, no correlation was observed both with and without the phylogenetic distance effect (Fig. S3A and C). When we pruned the host trees to retain only lineages that left extant descendants – reflecting the available situation with most empirically estimated trees – there was a positive correlation in the presence of the phylogenetic distance effect, but not in its absence. In the former case, a linear function explained around 16% of the variance in infection frequencies. This means that the more unbalanced a host tree is, the higher the expected infection frequency. This result is in accord with the Shannon index analysis presented in the main text (Figure 4): the greater the variation in species numbers across subtrees, the greater the Colless index and the smaller the Shannon index of species distributions. However, this result also seems to be at odds with the previous observation that very imbalanced trees should be less conducive to parasite spread than symmetrical ones (Engelstädter & Hurst 2006). This apparent contradiction can be explained by the fact that the trees used in Engelstädter & Hurst (2006) are very extreme, with Colless indices that fall far outside the range of those obtained here. For large trees generated by a birth-death process as we have done here, the Colless index mainly reflects asymmetries in subtree structure and is thus similar to (and indeed inversely correlated with) our Shannon index. However, as is clear from Figure 4, the Shannon index explains a higher fraction of the variability in infection frequencies.

4 Supplementary figures

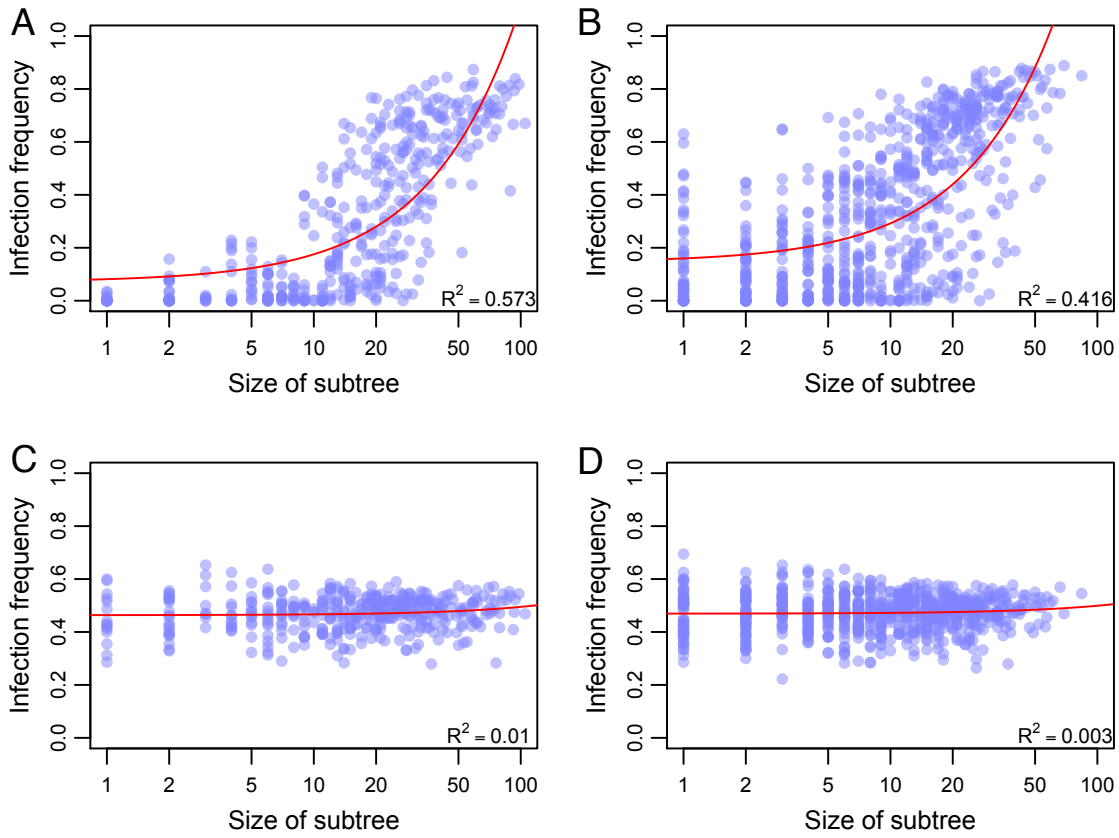

Figure 1: Fraction of infected hosts within host subtrees against the size of these subtrees with (A,B) and without (C,D) the phylogenetic distance effect. Each dot represents the mean infection frequency (across 100 simulations) of a subtree from one of the 100 trees forming the standard host tree set. Partitioning of host trees into subtrees was performed as described in section 2.3, 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.

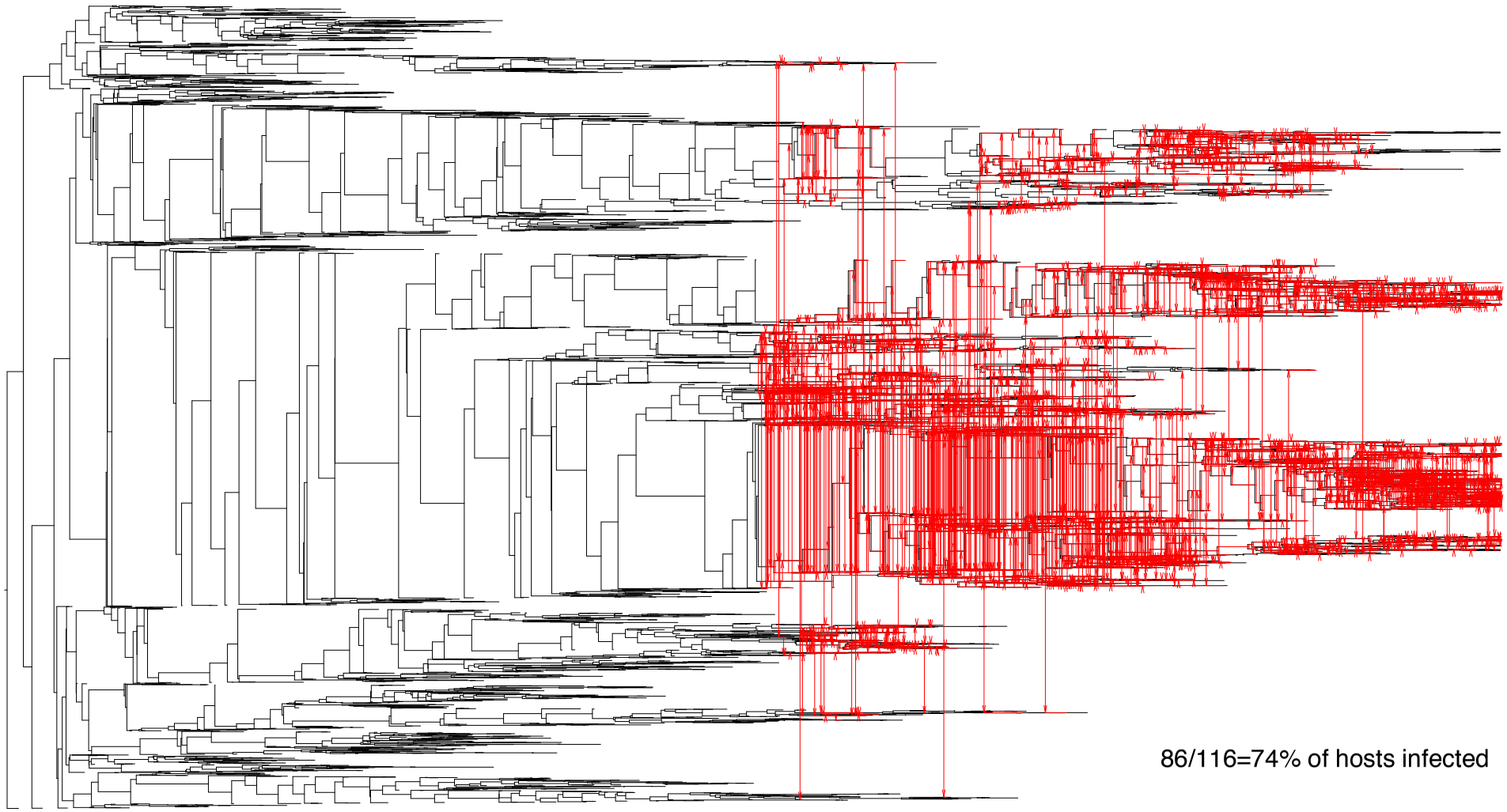

Figure S2A: Example cophylogeny with the standard PDE parameter set, for host tree no.1.

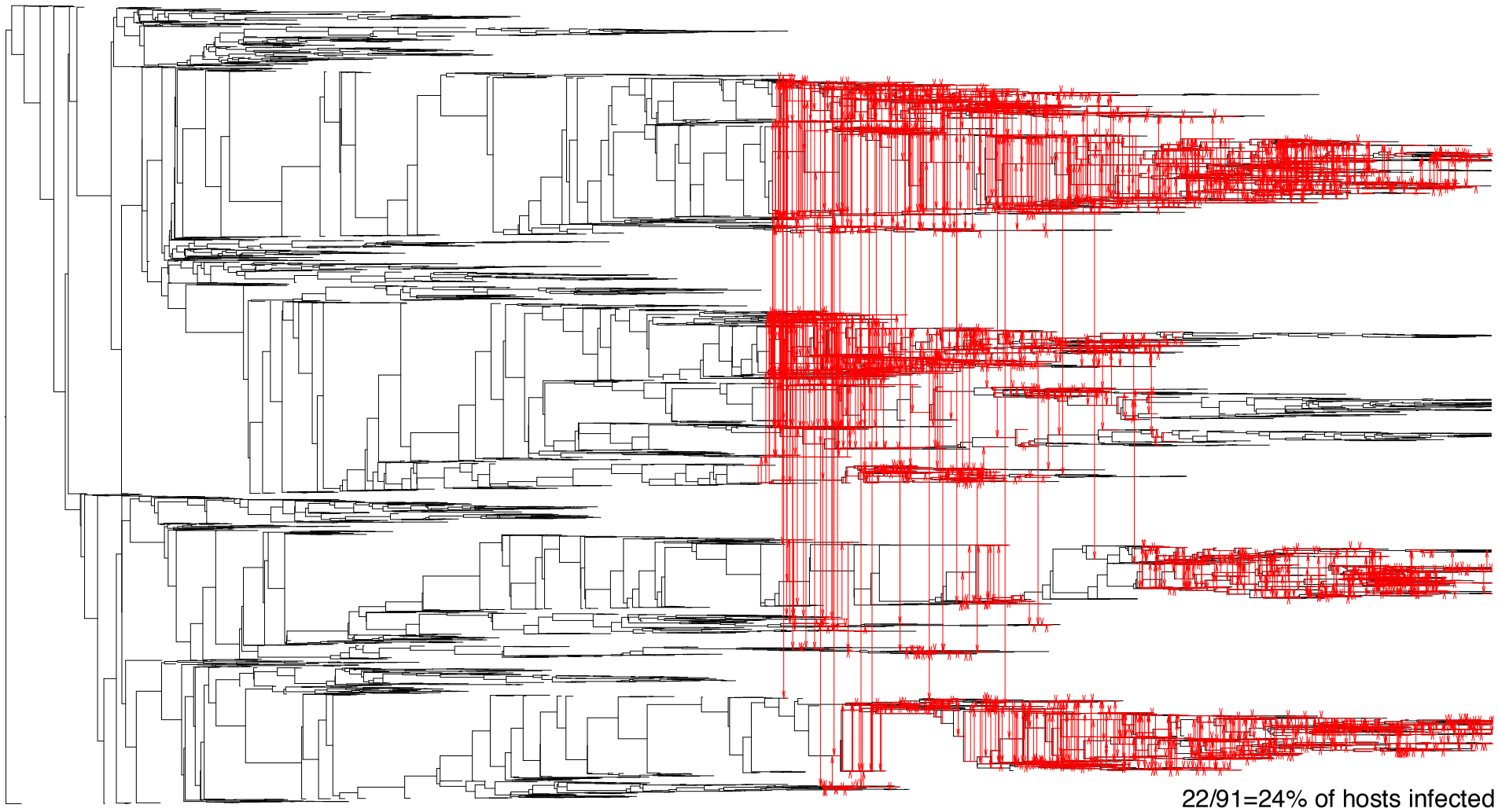

Figure S2B: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.5.

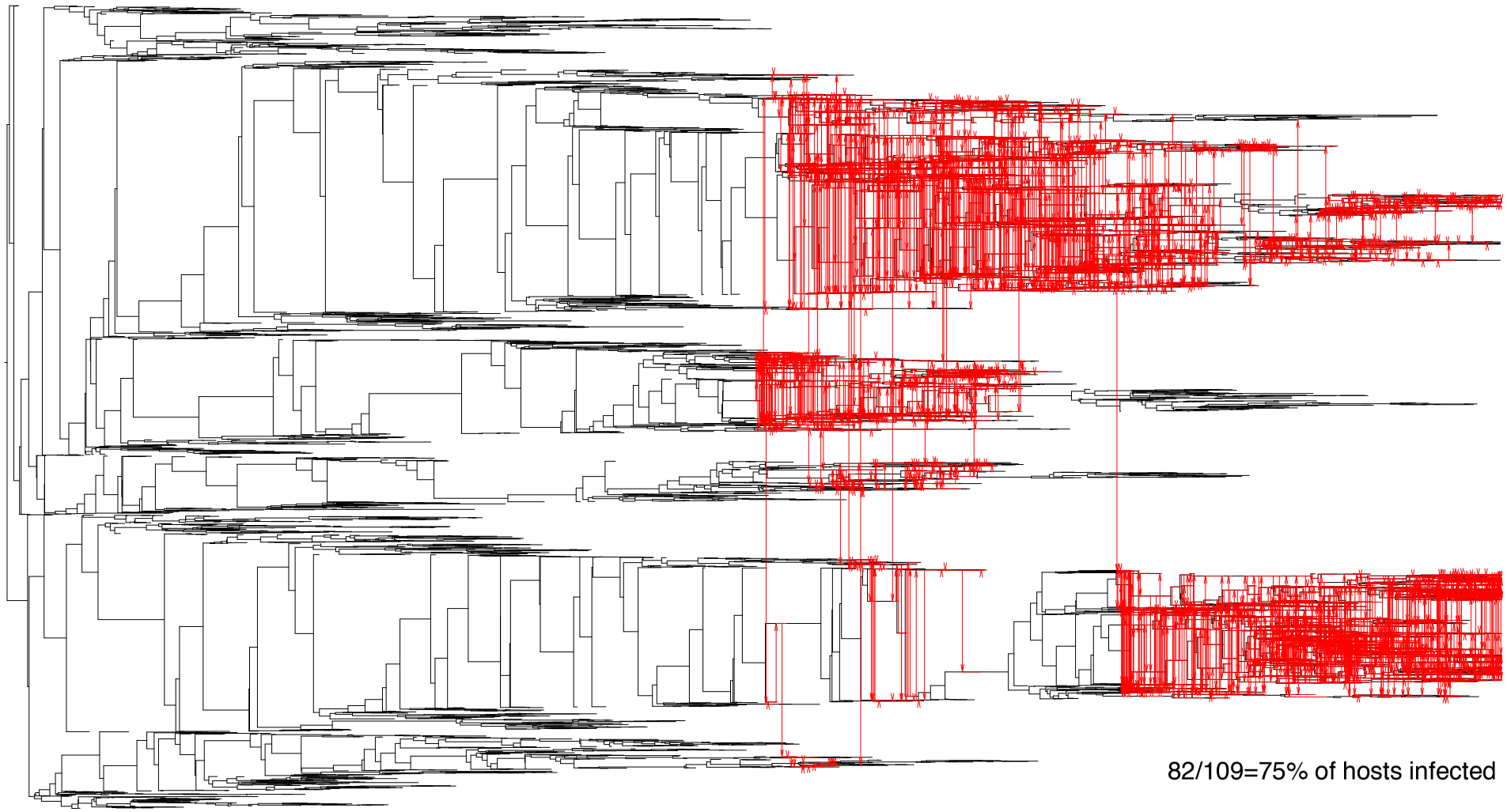

Figure S2C: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.25.

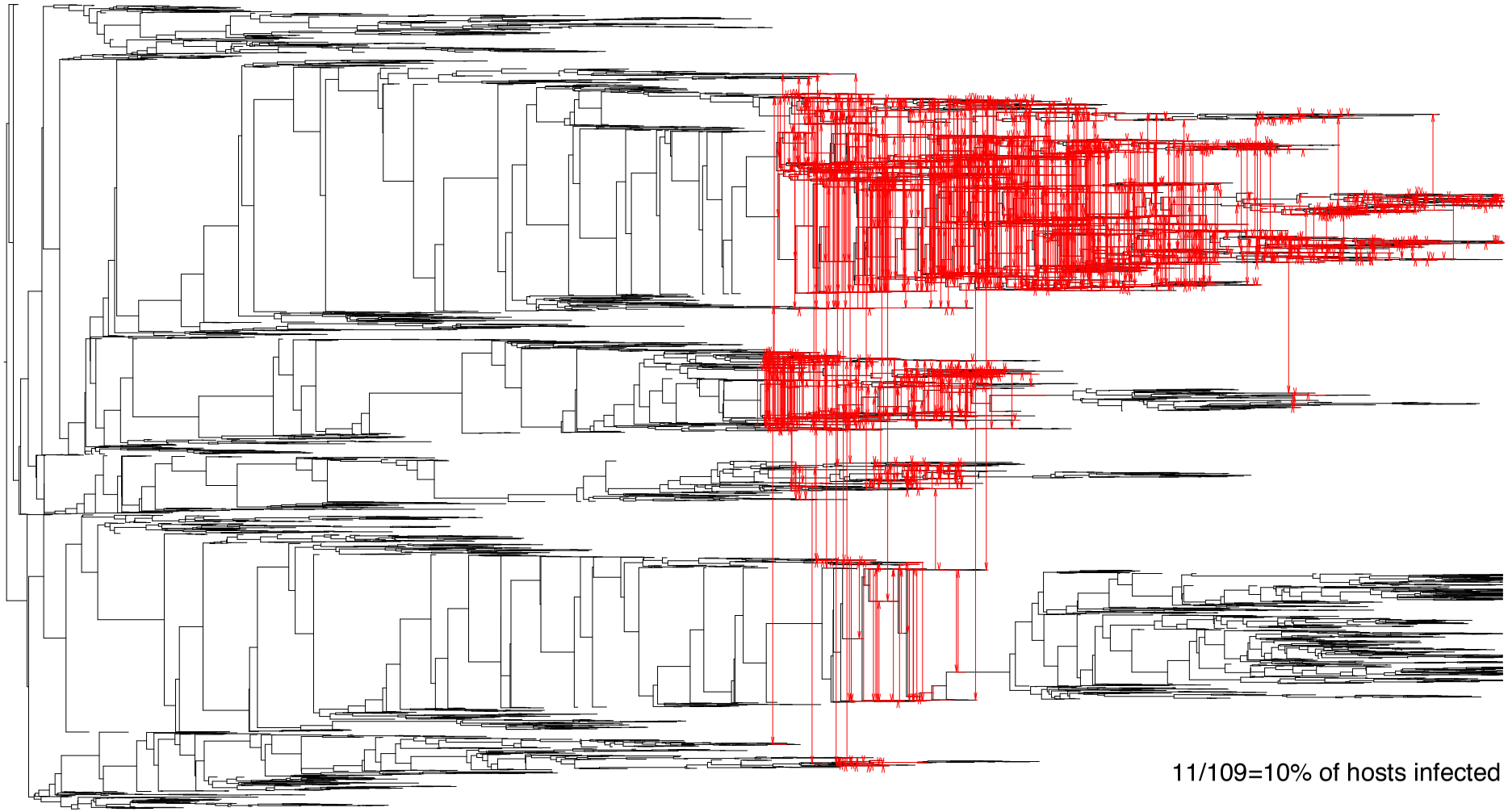

Figure S2D: Example cophylogeny with the standard PDE parameter and host tree set, for host tree no.25.

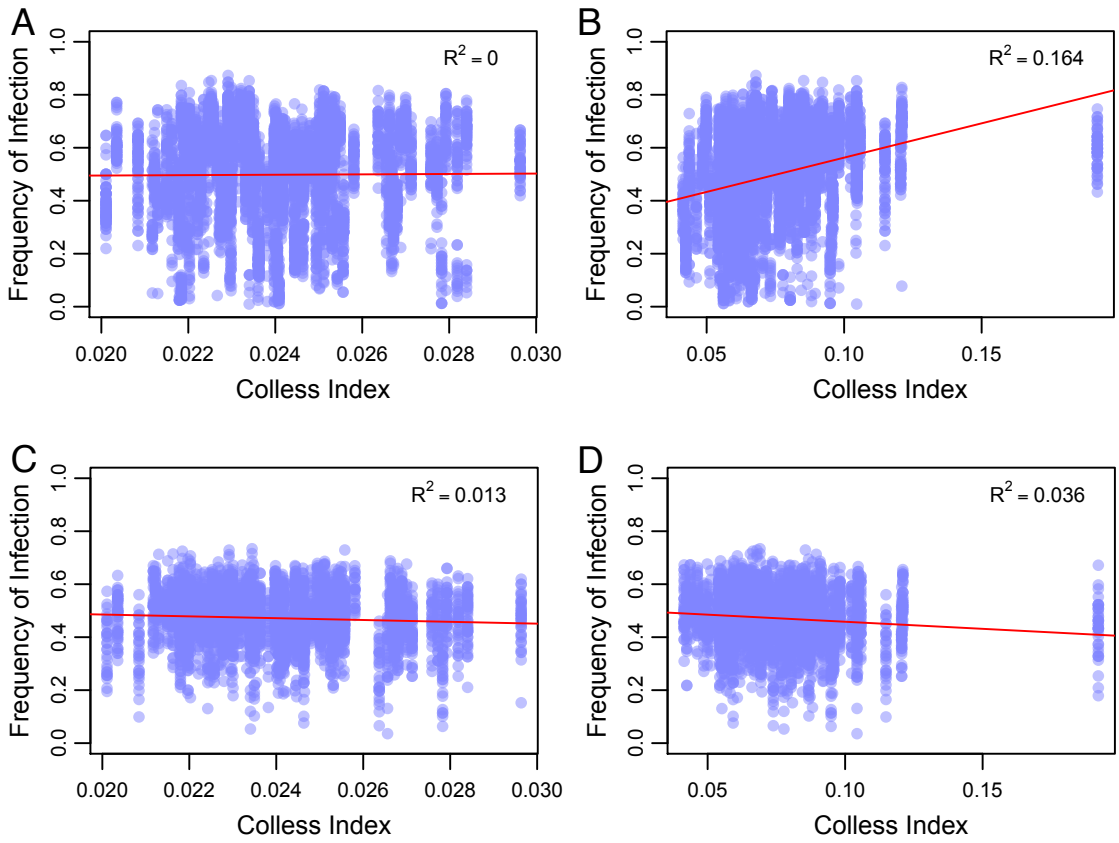

Figure S3: Fraction of infected hosts against the Colless index of host trees with (A,B) and without (C,D) the phylogenetic distance effect. Each dot represents the infection frequency in a single simulation run for those simulations where the parasite did not go extinct. The Colless index was determined either for the full host trees (A, C), or for the pruned trees containing only lineages with extant descendants. Red lines show the fit of a linear regression with R^2 values indicated. All parameters take standard values.

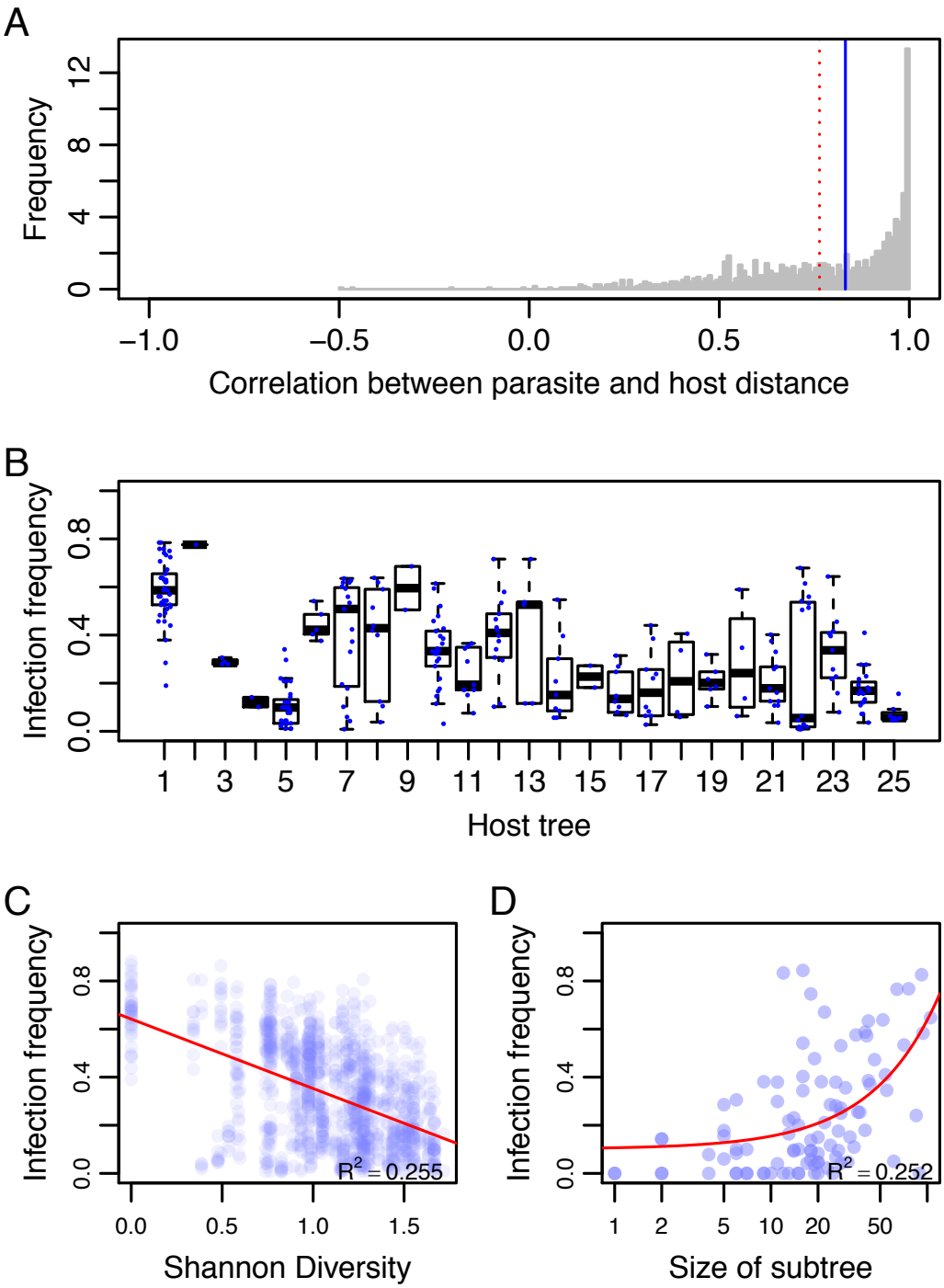

Figure S4: Results for the model with a lower parasite turnover, with parameters $\beta = 0.05$, $\nu = 0.1$ and standard PDE values for the other parameters. The panels show (A) the distribution of the correlation coefficients between parasite and corresponding host phylogenetic distances (as in Fig. 2D), (B) fractions of infected host species across the first 25 host trees (as in Fig. 3), (C) the fraction of infected hosts against the Shannon index of host subtree sizes (as in Fig. 4), and (D) the fraction of infected hosts within subtrees against the size of the subtrees (as in Fig. S3).

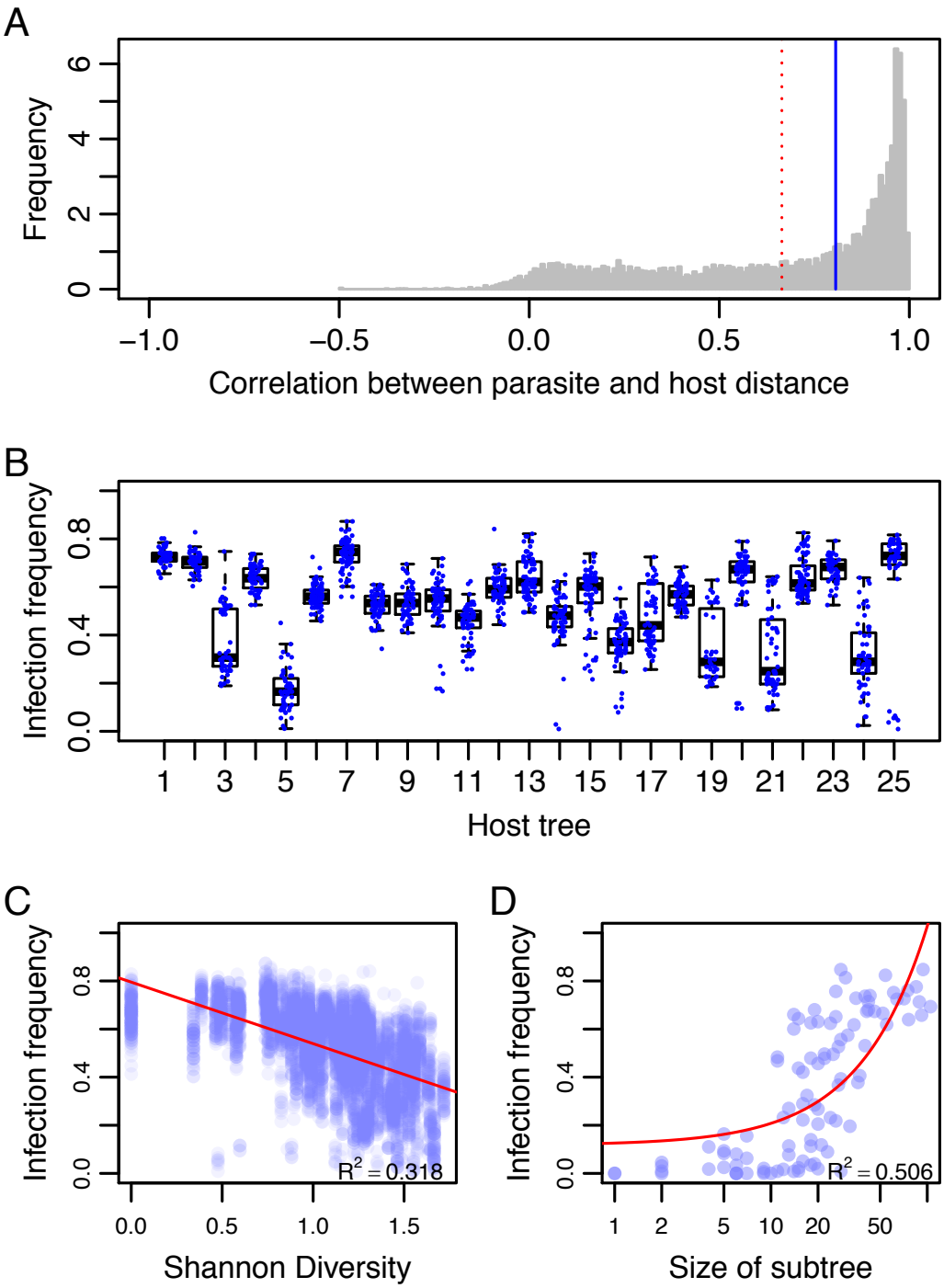

Figure S5: Results for the model with a higher parasite turnover, with parameters $\beta = 1$, $\nu = 2$ and standard PDE values for the other parameters. The panels show (A) the distribution of the correlation coefficients between parasite and corresponding host phylogenetic distances (as in Fig. 2D), (B) fractions of infected host species across the first 25 host trees (as in Fig. 3), (C) the fraction of infected hosts against the Shannon index of host subtree sizes (as in Fig. 4), and (D) the fraction of infected hosts within subtrees against the size of the subtrees (as in Fig. S3).

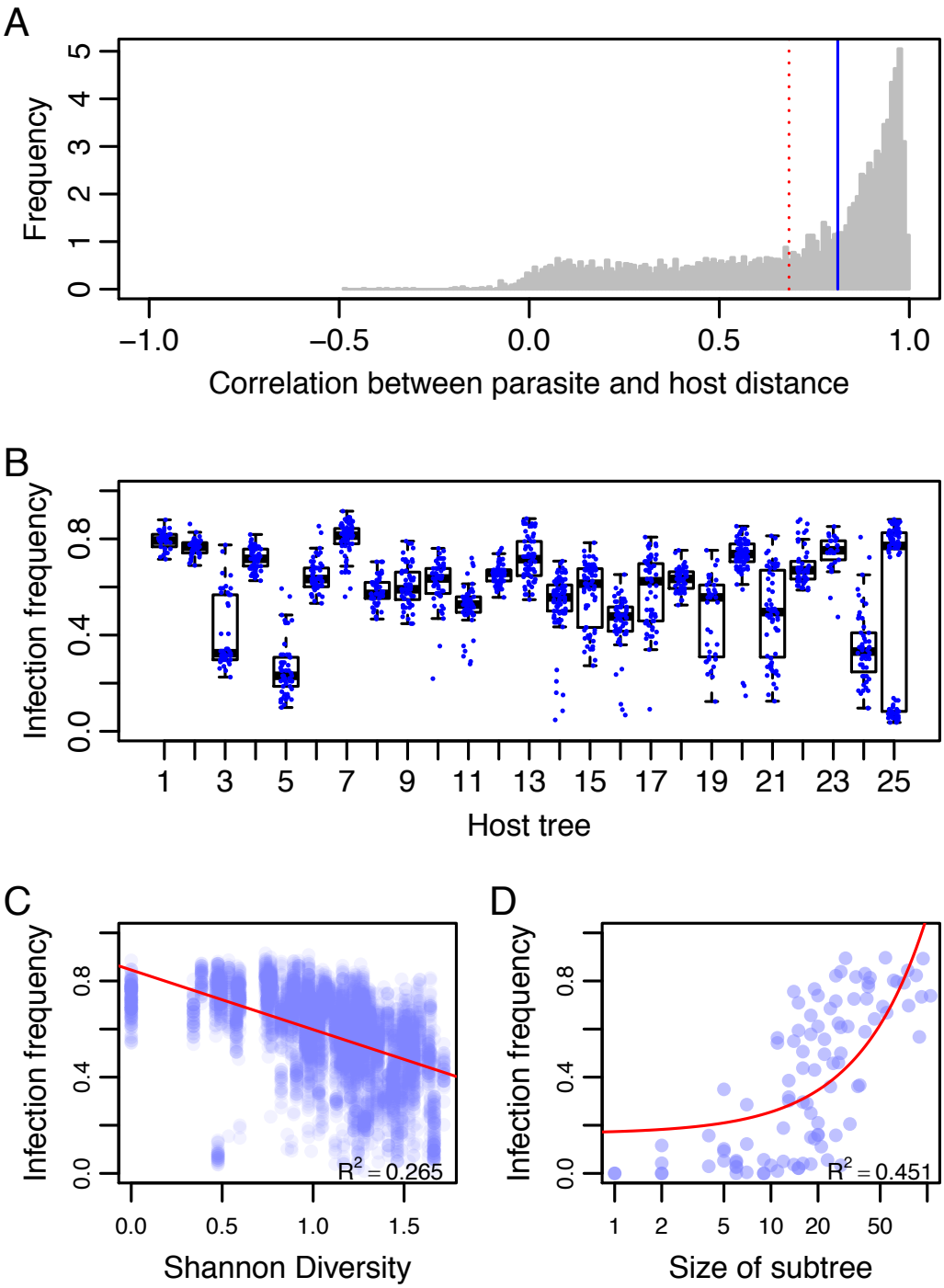

Figure S6: Results for the model extension where coinfections following host-shift events are possible (see section 1.1), with parameter $\sigma = 0.1$ and standard PDE values for the other parameters. The panels show (A) the distribution of the correlation coefficients between parasite and corresponding host phylogenetic distances (as in Fig. 2D), (B) fractions of infected host species across the first 25 host trees (as in Fig. 3), (C) the fraction of infected hosts against the Shannon index of host subtree sizes (as in Fig. 4), and (D) the fraction of infected hosts within subtrees against the size of the subtrees (as in Fig. S3).

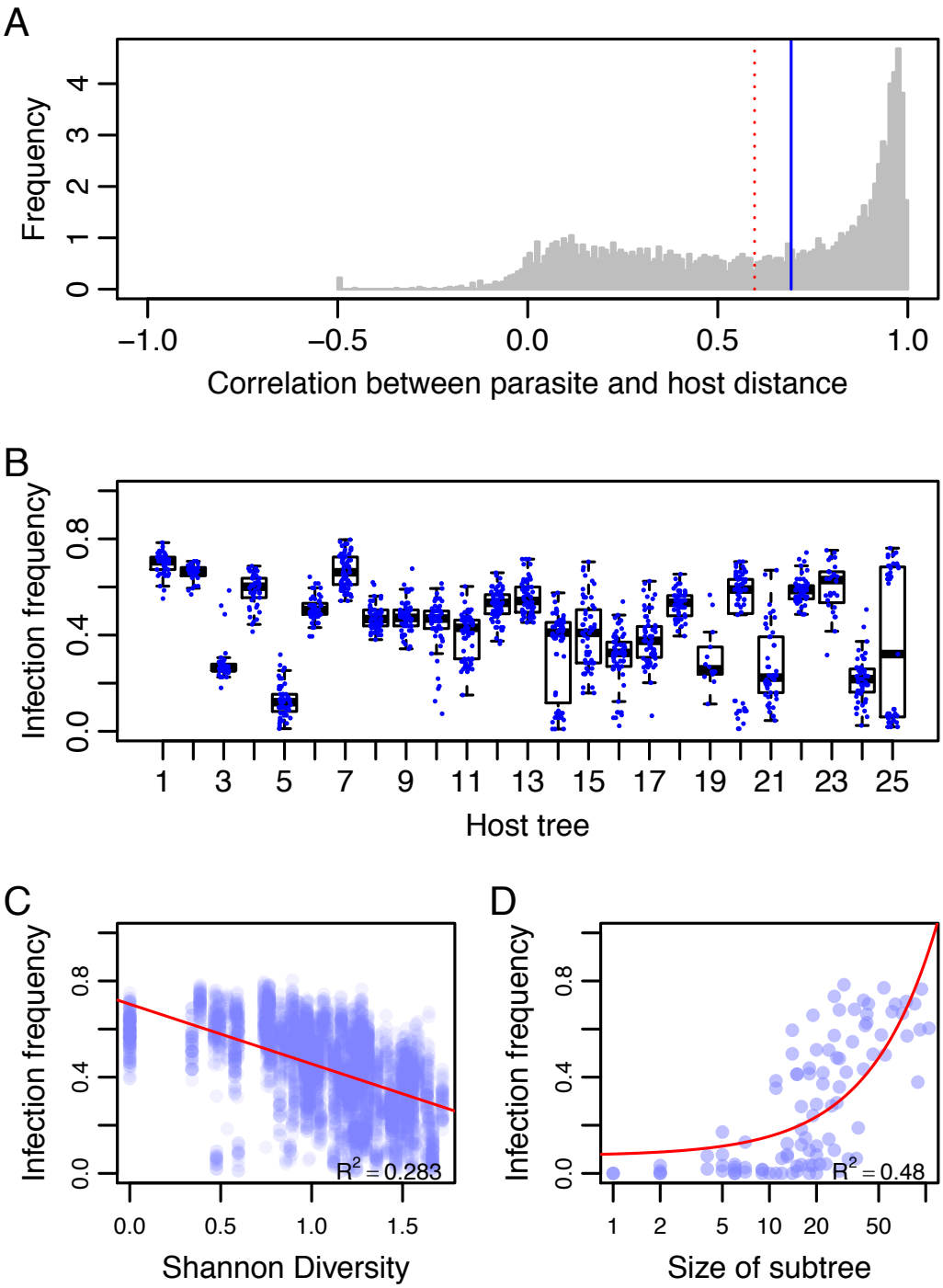

Figure S7: Results for the model extension with parasite loss during host speciation (see section 1.2), with parameter $\delta = 0.5$ and standard PDE values for the other parameters. The panels show (A) the distribution of the correlation coefficients between parasite and corresponding host phylogenetic distances (as in Fig. 2D), (B) fractions of infected host species across the first 25 host trees (as in Fig. 3), (C) the fraction of infected hosts against the Shannon index of host subtree sizes (as in Fig. 4), and (D) the fraction of infected hosts within subtrees against the size of the subtrees (as in Fig. S3).

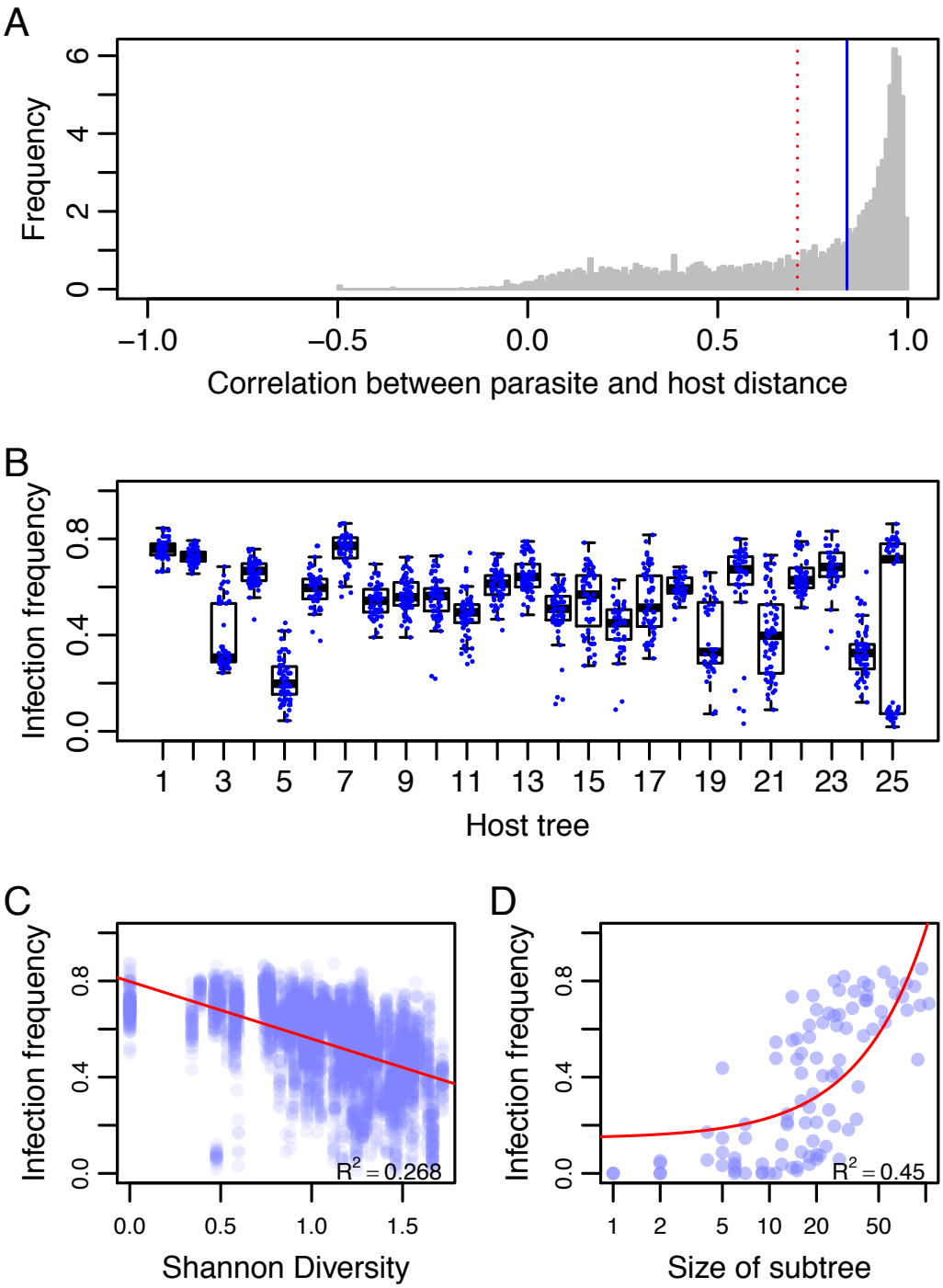

Figure S8: Results for the model extension with within-host parasite speciation (see section 1.3), with parameter $\kappa = 0.1$ and standard PDE values for the other parameters. The panels show (A) the distribution of the correlation coefficients between parasite and corresponding host phylogenetic distances (as in Fig. 2D), (B) fractions of infected host species across the first 25 host trees (as in Fig. 3), (C) the fraction of infected hosts against the Shannon index of host subtree sizes (as in Fig. 4), and (D) the fraction of infected hosts within subtrees against the size of the subtrees (as in Fig. S3).

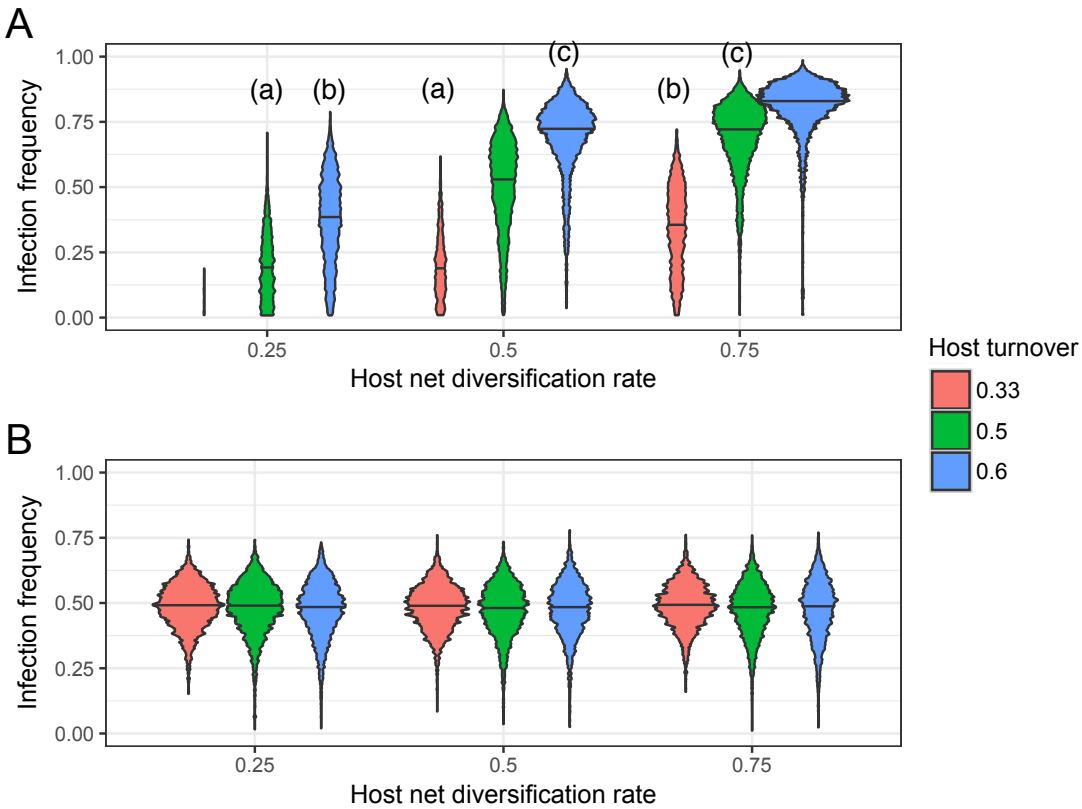

Figure S9: The impact of host net diversification ($\lambda - \mu$) and turnover (μ/λ) on the fraction of infected host species with (A) and without (B) the phylogenetic distance effect. Violins show the distribution of infection frequency, with the total area of each violin being proportional to the number of simulations in which the parasites survived. Letters (a), (b) and (c) indicate parameter combinations with identical values of μ. See section 2.1 for details on simulations and parameter values.