The dynamics of preferential host switching: host phylogeny as a key predictor of parasite prevalence and distribution

Jan Engelstädter* & Nicole Z. Fortuna

School of Biological Sciences, The University of Queensland, Brisbane, Australia

* Corresponding author.
Address: The University of Queensland, School of Biological Sciences, Brisbane, QLD 4072, Australia
Phone: +61 7 336 57959
Email: j.engelstaedter@uq.edu.au

Keywords: host-shifts, host switches, codiversification, host-parasite coevolution, emerging diseases, mathematical model, host range

Authorship: JE conceived and designed the project, JE and NZF implemented the model, analysed the simulation results and wrote the paper.

Abstract

New parasites commonly arise through host-shifts, where parasites from one host species jump to and become established in a new host species. There is much evidence that the probability of host-shifts decreases with increasing phylogenetic distance between donor and recipient hosts, but the consequences of such preferential host switching remain little explored. We develop a mathematical model to investigate the dynamics of parasite host-shifts in the presence of this phylogenetic distance effect. Host trees evolve under a stochastic birth-death process and parasites co-evolve concurrently on those trees, undergoing host-shifts, co-speciation and extinction. Our model indicates that host trees have a major influence on these dynamics. This applies both to individual trees that evolved under the same stochastic process and to sets of trees that evolved with different macroevolutionary parameters. We predict that trees consisting of a few large clades of host species and those with fast species turnover should harbour more parasites than trees with many small clades and those that diversify more slowly. Within trees, large clades should exhibit a higher infection frequency than small clades. We discuss our results in the light of recent cophylogenetic studies in a wide range of host-parasite systems, including the intracellular bacterium *Wolbachia*.

41 # Introduction

42 Parasitism represents one of the most successful modes of life. Humans harbour
43 more than 1400 species of parasites (Taylor *et al.* 2001), which extrapolates to an
44 enormous total number of parasites across all host species. Where do all these
45 parasites come from? Some parasites may already have been present in their host
46 species' ancestor and maintained ever since. This scenario of 'cospeciation' has
47 been described in some mutualists but appears to be rare in parasites (de Vienne *et*
48 *al.* 2013). Other parasites may originate from organisms that are either free-living, or
49 non-parasitic symbionts (Crook 2014; Hurst 2016). Finally, some parasites may have
50 switched from another host species to their present-day host. Such host-shifts have
51 been widely documented. The majority of human pathogens originate through host-
52 shifts, including HIV and malaria (Wolfe *et al.* 2007). Host-shifts are also the
53 predominant cause of new host-parasite associations for *Wolbachia* endosymbionts
54 and their arthropod hosts (Werren *et al.* 1995), rabies viruses in bats (Streicker et al.
55 2010), lentiviruses in primates (Sharp *et al.* 2000), oomycetes in Asteraceae (Choi &
56 Thines 2015), and malaria in birds (Ricklefs *et al.* 2014).

57 Establishing a sustainable relationship with a new host species represents a
58 considerable challenge to parasites. While many opportunities for host-switches
59 exist, most attempts are unsuccessful and lead to mere 'spill-over' infections, i.e.
60 infections with no or short transmission chains (Taylor *et al.* 2001; Wood *et al.* 2012).
61 Examples of such spillovers in humans include rabies, Hendra, and Ebola viruses.
62 Successful host-shifts are difficult because the parasite must be able to enter,
63 proliferate within, and transmit efficiently between, members of a new host species
64 that they are not adapted to. These requirements mean that all else being equal,

shifts to new hosts that are similar to the original host with respect to relevant traits should be easier than shifts to hosts that are very different from the original one. Given that this similarity will be positively correlated with phylogenetic relatedness between host species, we can predict that host-shifts to closely related new hosts should be more common than host-shifts to distantly related hosts (Charleston & Robertson 2002; Engelstädter & Hurst 2006; Longdon *et al.* 2014). We will refer to this expectation as the 'phylogenetic distance effect'.

There are two lines of evidence for the phylogenetic distance effect. First, a number of transfection experiments have been conducted in which parasites from one species were exposed to a range of hosts from different species. For example, Longdon *et al*. (2011) demonstrated that for three sigma viruses endogenous to different species of *Drosophila*, phylogenetic distance between the donor and recipient host species was negatively correlated with the viruses' ability to replicate within the recipient host. Similarly, for male-killing *Spiroplasma* bacteria in ladybird beetles, Tinsley & Majerus (2007) reported that as the distance between the original host and a new host increased, the ability of the parasite to kill male offspring (the phenotype driving the infection) was reduced. Other systems in which experimental evidence for the phylogenetic distance effect has been obtained include nematodes infecting *Drosophila* flies (Perlman & Jaenike 2003), feather-lice infecting pigeons and doves (Clayton *et al.* 2003), and plant-fungal systems (Gilbert & Webb 2007; de Vienne *et al.* 2009). (De Vienne *et al*. (2009) also showed that the phylogenetic distance between a native and a new parasite was a good predictor of infection success as well.) Strong evidence for the phylogenetic distance effect from 25

publications reporting the success or failure of *Wolbachia* transfection experiments is reviewed in Russell *et al*. (2009).

Second, different phylogenetic methods have been used to investigate whether host-shifts occur preferentially between related host species. Much early work comparing host and parasite phylogenetic trees focused on reconciling those trees and identifying the degree of cospeciation. However, Charleston & Robertson (2002) showed that the observation that closely related lentiviruses tend to infect closely related primate hosts is best explained not by codivergence but by preferential host-switching between related hosts (because the viruses only spread relatively recently on the primate tree). Studies of rabies viruses infecting various bat species confirmed the presence of the phylogenetic distance effect (Streicker *et al.* 2010) and further demonstrated that while species range overlap was the best predictor of spillover events, phylogenetic distance was the best predictor of host-shift events (Faria *et al.* 2013). Clark & Clegg (2017), studying the distribution of malaria among south-Melanesian birds, found that despite ample opportunity for host-switching due to vector-borne transmission, similar parasites were restricted to similar hosts. Some studies have also provided evidence that *Wolbachia* endosymbionts switch preferentially between related arthropod host species (Baldo *et al.* 2008; Russell *et al.* 2009; see also Discussion). In summary, the experimental and comparative work indicates that although not ubiquitous (e.g., Stahlhut *et al.* 2010; Longdon *et al.* 2015), the phylogenetic distance effect is an important determinant of host-shifts in many systems.

Most of the previous theoretical work on host-shifts has focused on reconciling host and parasite phylogenetic trees, identifying host-shift vs. cospeciation events, and

inferring parameters underlying these processes (older literature reviewed in de Vienne *et al.* 2013; newer work includes Baudet *et al.* 2015; Wieseke *et al.* 2015; Drinkwater & Charleston 2016; Alcala *et al.* 2017). Mathematically speaking, these are very hard problems and most of the developed algorithms are computationally expensive. It is therefore not surprising that the phylogenetic distance effect is usually not considered in these methods, despite the widely recognised fact that preferential host switching may be misinterpreted as cospeciation (de Vienne *et al.* 2007). Exceptions include a study where data from RNA virus-mammal associations were used to test two different models describing the decline in host-shift success with increasing phylogenetic distance between host species (Cuthill & Charleston 2013), and a study in which the host-shift dynamics of protozoan parasites in new world monkeys were inferred (Waxman *et al.* 2014). In contrast to the development of inference methods for host-parasite cospeciation and host-shifts, little work has been done to explore the consequences of the phylogenetic distance effect for the dynamics of parasites spread between host species the expected patterns of parasite distribution. In simulations of parasite host switching, Engelstädter and Hurst (2006) demonstrated that the 'shape' of a host clade strongly influences parasite prevalence and distributions within host clades. However, their model (like the model by de Vienne *et al.* 2007) only considered a few idealised host trees (e.g., either completely symmetrical or ladder-like), and they (like Cuthill & Charleston 2013; Waxman *et al.* 2014) assumed that host switching occurred only at the tips of the trees.

Here, we present the results of a stochastic model in which a clade of host species evolves under a birth-death process and a clade of parasites spreads concurrently

136 on this host tree through both cospeciation events and host-shifts (either preferential
137 or random). Through extensive computer simulations we investigate how often the
138 parasites can invade a naïve host tree, how many hosts will become infected and
139 how the parasites are distributed across host species. Our model predicts that both
140 individual host phylogenies and the macroevolutionary processes underlying these
141 phylogenies have a major influence on host-shift dynamics when the phylogenetic
142 distance effect is important.

143 Methods

144 *Mathematical model*

145 We considered a stochastic model of host-parasite co-diversification, illustrated in
146 Figure 1. Host trees emerge from a single ancestor according to a density-dependent
147 birth-death process. Hosts go extinct at a constant rate μ and speciate at a baseline
148 rate λ that is multiplied by the term $(1-N/K)$, resulting in a decreasing speciation rate
149 as the number of host species N approaches the carrying capacity K.

150 Each parasite species is associated with a single host species. Parasites go extinct
151 at a constant rate ν and always co-speciate whenever their hosts speciate. Host-
152 shifts represent an alternative, independent mode of parasite speciation in which one
153 lineage remains associated with the original host and a new lineage arises that is
154 associated with a new host species. Host-shifts occur at a baseline rate $\beta(N-1)$ per
155 parasite lineage. Potential new hosts are chosen randomly but not all host-shifts are
156 successful. First, host-shifts are unsuccessful if the new host is already infected (but
157 see below for an extension of the model where this assumption is relaxed). Second,
158 parasites may not become established if the new host is phylogenetically too distant

159 from the original one. Specifically, we assume a parasite establishment probability,
160 exp($-\gamma D_{ij}$). (The same relationship but using a different notation was used by
161 Engelstädter & Hurst 2006; Cuthill & Charleston 2013). Here, the parameter
162 γ determines how fast the establishment probability declines with increasing
163 phylogenetic distance D_{ij} between the donor host species *i* and the new host species
164 *j* (i.e., D_{ij} is the total length of branches connecting the two species with their most
165 recent common ancestor). When $\gamma=0$, all host-shifts are successful (no phylogenetic
166 distance effect) but with larger values of γ, host species that are distantly related to
167 the original host are increasingly unlikely to become infected.
168 In addition to this basic model, we also investigated three model extensions that
169 incorporate 1) coinfection of multiple parasites in one host species, 2) parasite loss
170 during cospeciation, and 3) within-host speciation of parasites. For details we refer to
171 the Supplementary Information (SI), section 1.

172 *Model implementation*

173 We analysed our model using computer simulations. Time proceeds in small steps
174 ($\Delta t=10^{-4}$) in which the different events (host speciation, host extinction etc.) take
175 place with probabilities given by their rates multiplied by Δt. Since host evolution is
176 not affected by the parasites in our model, we first simulated the host trees and then
177 simulated parasite diversification on those host trees.
178 The routines to simulate the cophylogenetic process were implemented in the
179 programming language R (R Core Team 2017). We bundled these routines, along
180 with other functions for simulation, subsequent analyses and plotting of
181 cophylogenetic trees, into a new R-package named 'cophy'. This package depends
182 on the R-packages ape v4.1 (Paradis *et al.* 2004), parallel v3.3.2 (R Core Team

2017), foreach v1.4.3 (Revolution Analytics & Weston 2015b), and doParallel v1.0.10 (Revolution Analytics & Weston 2015a). We used the R-packages devtools v1.13.2 (Wickham & Chang 2017) and roxygen2 v6.0.1 (Wickham *et al.* 2017) to generate our package. The cophy package will be made available on CRAN upon publication of this article. For data analysis, we also used lme4 v1.1-12 (Bates *et al.* 2015) and vegan v2.4-5 (Oksanen *et al.* 2017).

Simulations

We started by simulating different sets of host trees, each containing 100 trees that were initialised with a single species and evolved for 100 time units. Only trees that survived this time span were retained. For an initial standard set of trees, we chose a speciation rate of λ=1, an extinction rate of μ=0.5 and a carrying capacity of K=200, yielding an expected equilibrium tree size of N=100 species. Using this set as a baseline, we created three series of similar sets with 1) the same speciation and extinction rate but with N increasing from 30 to 200, 2) the same equilibrium clade size and net diversification rate ($\lambda-\mu$=0.5), but extinction rate μ increasing from 0.1 to 0.9, and 3) eight other sets with the same equilibrium clade size but different net diversification and turnover rates (see SI section 2.1 for details).

To simulate parasite diversification on those host trees, we introduced a single parasite species at time t=50 on a given host tree and simulated until the parasite went extinct or the present (t=100) was reached. For each host tree, we randomly chose ten branches on which the first parasite species arrived and performed ten replicate simulations for each of these initial branches. Thus, for each set of host trees we performed a total of 100×10×10=10,000 simulations.

We focused on two parameter sets for parasite evolution. First, we used a parameter combination with which the phylogenetic distance effect is present: β=0.5, γ=0.06 and ν=1. Second, as a control, we used a parameter combination with which the phylogenetic distance effect is absent: β=0.02, γ=0 and ν=1. We refer to these two standard parameter combinations as the standard PDE and no-PDE parameters, respectively. The parameters were chosen so that both the probability of parasite establishment and the observed frequency of infected hosts at the end of the simulation are roughly the same (around 0.5; see Results). In order to test whether our results are robust with respect to the choice of parameters, we also performed simulations with two other PDE / no-PDE parameter combinations that are characterised by either a lower or a higher turnover rate in parasite diversification. Finally, we also performed the same simulations for our three model extensions (SI section 1).

Analyses of results

For each simulation we obtained some basic statistics, including the fraction of simulations in which the parasites established a surviving infection on the host trees, the distribution of the number of host and parasite species and the frequency of infected hosts at the end of the simulation (contingent on parasite survival). For parasite trees that did not leave any surviving species we obtained the time of extinction, and for those which did we obtained the time of the most recent common ancestor of all extant species. As a simple statistic describing the distribution of parasites within the host phylogeny we used the correlation coefficient between host and parasite phylogenetic distances (see SI, section 2.2). We also investigated the

229 frequency of infected host species within different clades of the host tree (see SI
230 section 2.3).

231 Results

232 *Patterns of parasite spread and distributions*

233 We first focused on understanding the host-shift dynamics under the phylogenetic
234 distance effect on a standard set of host trees simulated under the same birth-death
235 process. Figure 2 compares some basic summary statistics for simulations in
236 presence vs. absence of the phylogenetic distance effect (standard PDE vs. no-PDE
237 parameters). By choice of parameters, the final mean frequency of infected hosts for
238 simulations with surviving parasites was similar in both cases (Fig. 2A). However, the
239 variance in infection frequencies was greater with the phylogenetic distance effect
240 than without (see also below). If the parasites went extinct, this usually occurred
241 early during the simulations in both scenarios (Fig. 2B). The most recent common
242 ancestor of all surviving parasites lived later on average with than without the
243 phylogenetic distance effect (Fig. 2C), reflecting higher parasite turnover in the latter
244 case.

245 In Figure 2D, we plot the distribution of correlation coefficients between phylogenetic
246 distances between pairs of parasite species and the phylogenetic distances between
247 their associated host species. In the presence of the phylogenetic distance effect,
248 this distribution shows a strong positive trend: >98% of simulations where the
249 parasites survived exhibited a positive correlation, with a median of 0.807. Thus,
250 closely related parasites tend to be found in closely related host species and *vice*
251 *versa*. This is not primarily a consequence of co-speciation events but of the

252 phylogenetic distance effect. In the absence of the phylogenetic distance effect, the
253 host-parasite phylogenetic correlation coefficients are distributed around zero. The
254 median of this distribution is still positive (0.021), which is explained by recent co-
255 speciation events, but the distribution is very distinct from the one observed in the
256 presence of the phylogenetic distance effect.

257 We can also ask how parasites are distributed within different host clades when the
258 phylogenetic distance effect is important. Parasites will shift predominantly within
259 host clades but rarely between different clades in this case. One might therefore
260 expect that all else being equal, larger host clades should on average harbour more
261 parasites than smaller clades. Figure S1 shows that this expectation is fulfilled both
262 when host trees are split into a few large and into many small clades (Fig. S1A and
263 B). In the absence of the phylogenetic distance effect, host clade size has no effect
264 on the fraction of hosts that are infected within those clades (Fig. S1C and D).

265 *Host trees are important in determining parasite spread*

266 Figure 3A shows that in the presence of the phylogenetic distance effect, the
267 distribution of the fraction of infected host species observed at the end of the
268 simulations differs according to host tree. A random effects model confirms the visual
269 impression that much of the variation in the fraction of infected host species
270 observed at the end of the simulations is due to the specific host tree on which the
271 parasites spread (see SI, section 2). By contrast, in the absence of the phylogenetic
272 distance effect, the observed mean infection frequencies are much more
273 homogeneous across host trees (Figure 3B), with a lower fraction of variance
274 explained by host trees (SI section 2).

To obtain some intuition for the importance of host trees in shaping host-shift dynamics, consider the example co-phylogenies shown in Figures 3C and S2, corresponding to host trees number 1, 5 and 25. With host tree #1 (Fig. 3C and S2A), most of the extant host species form one large, relatively recently formed clade of species. A second, smaller clade is still closely related to the first one. This means that for most host species there is an abundance of closely related host species, which enables the parasites to readily undergo host switches and thus reach a high frequency. Host tree #5 (Figs. 3C and S2B) shows the opposite extreme: the host tree consists of several clades that are only distantly related to each other. Parasite spread and survival within those clades is difficult because these clades are small, and switches between clades are unlikely. Combined, this explains the low infection frequencies observed on this tree. Host tree #25 (Fig. S2C, D) contains a large clade of closely related host species in which the parasites can thrive. If the parasites are successful in infecting this large clade, they can reach a high frequency of infected host species (Fig. S2C). However, this clade is very isolated from the other clades and connected to the rest of the tree by a long branch. Therefore, in many cases the parasites fail to reach this clade and are confined to the other, much smaller clades (Fig. S2D). As a consequence, we observe a bimodal distribution of infection frequencies for this tree.

To formalise some of the above intuitive explanations for variation in parasite abundance across host trees, we calculated for each host tree the Shannon index for the distribution of host species among different host clades (see SI section 1.3). This Shannon index is greater the more host clades there are and the more evenly species are distributed among those clades. Figure 4 shows that the Shannon index

is negatively correlated with the fraction of infected host species, indicating that host trees whose species are clustered in a with few large clades are most conducive to parasite spread. In line with these results, we also found that tree imbalance, as measured by the Colless index (Colless 1982; Heard 1992), has a similar effect but explains less of the variance in infection frequencies than the Shannon index of clade sizes (see SI section 3.1; Fig. S3).

Robustness to parasite parameters and model assumptions

We repeated all simulations with a higher parasite transmission rate (β=1) and a higher extinction rate (ν=2). Figures S4 and S5 show that our results are very robust to this change in parameters. We also re-ran our simulations relaxing the assumption that no coinfections can occur, that parasites can be lost during host speciation or that they can speciate within a host linage; again, this did not qualitatively affect our results (Figures S6 to S8).

Host tree size

We next asked how the equilibrium size of the host trees – determined by the carrying capacity K – affects the dynamics of parasite spread. In the absence of the phylogenetic distance effect, increasing host tree size results in both an increasing probability of parasite survival and an increasing number of infected hosts at the end of simulations where parasites do survive (Figure 5). Both of these results are straightforward in the light of standard epidemiological models with density-dependent transmission in well-mixed host populations (Keeling & Rohani 2008). In the presence of the phylogenetic distance effect, there is a comparatively modest increase in the parasite survival probability with increasing host tree size, and no change in the infection frequency. This is because from any given infected host

species, the number of uninfected hosts that can be reached through host-shifts will generally be limited by the phylogenetic distance effect rather than the total size of the tree.

Dynamics of host diversification

The results presented above all assumed that host trees evolved under the same birth-death process, with a speciation rate of λ=1 and an extinction rate of μ=0.5. In order to explore the impact of host diversification on parasite spread, we generated sets of host trees with increasing values of λ and μ while keeping the difference $\lambda-\mu$ constant. This means that for all sets of host trees generated, the host trees will initially grow at the same net diversification rate but when they reach their carrying capacity, the rate at which new host species are born and go extinct increases (both occurring at rate μ).

Figure 6A shows that in the presence of the phylogenetic distance effect, the host tree sets generated in this way vary strongly in both the parasite survival probability and the fraction of infected host species. When host trees evolve with very low speciation and extinction rates, the parasites almost always become extinct, and if they survive they reach only a very low infection frequency. This is because branches are very long in such host trees, resulting in large phylogenetic distances between host species that are difficult to overcome by the parasites. When λ and μ are high, there will be much turnover in host species and genetic distances will become short so that parasite spread is facilitated, resulting in a high fraction of simulations where parasites survive and reach high infection frequencies.

In the absence of the phylogenetic distance effect, mean infection frequencies are not affected by λ and μ (Figure 6B). However, the probability of parasite survival decreases slightly with increasing λ and μ. This is because host species numbers vary more through time with high than with low host speciation and extinction rates (results not shown), producing correspondingly strong stochastic variation in infection rates. As a result, when λ and μ are high, stochastic parasite extinction is more likely than when λ and μ are low.

Finally, we explored whether host net diversification rate ($\lambda-\mu$) or species turnover (μ/λ) had any impact on the dynamics of parasite spread beyond the impact of the rate of speciation and extinction in the steady state discussed above. We generated eight additional sets of host trees with different combinations of values for $\lambda-\mu$ and μ/λ (see SI section 1.4). Under the phylogenetic distance effect, the parasite survival rate and the fraction of infected hosts increases with both net diversification rate and host species turnover on these trees (Figure S9A). However, the results are always very similar with identical host extinction rates, suggesting that early host tree evolution was not important. In the absence of the phylogenetic distance effect, different host tree sets only differ mainly in the fraction of simulations where the parasites survived (Figure S9B), presumably again due to different degrees of stochastic fluctuations in host tree size.

Discussion

Using a mathematical model, we have investigated how the phylogenetic distance effect (preferential host-shifts between closely related species) impacts the prevalence and distribution of parasites across host species. Our model makes a

369 number of predictions: all else being equal, 1) host trees in which most species are
370 found in a few large clades should harbour more parasites than those consisting of
371 many small clades, 2) host trees characterised by high species turnover (including
372 rapid adaptive radiations) should harbour more parasites than host trees that are
373 evolutionarily more inert, and 3) small and isolated clades within trees should
374 harbour fewer parasites than large clades. These predictions can be tested without
375 any cophylogenetic analyses and indeed, without any knowledge about phylogenetic
376 relationships between the parasites. In contrast to previous models where parasites
377 only switch between extant host species (Engelstädter & Hurst 2006; de Vienne *et*
378 *al.* 2007; Cuthill & Charleston 2013; Waxman *et al.* 2014), in our model parasite and
379 host diversification occurs concurrently and potentially on similar time scales.
380 The power of our predictions depends on how strong the phylogenetic distance effect
381 is, both in absolute terms and relative to other effects. The phylogenetic distance
382 effect emerges from the fact that related species tend to be physiologically and
383 immunologically similar, thus increasing the chances that a parasite can successfully
384 replicate in a new host. However, relevant host traits such as the presence or
385 absence of certain cell surface receptors may also evolve repeatedly during host
386 diversification. This can give rise to 'clade effects' in which a host clade that is only
387 distantly related to a donor host may nevertheless have a high propensity to be
388 recipients of a parasite (Longdon *et al.* 2011; Waxman *et al.* 2014). Moreover, the
389 probability of host-shifts will depend not only on similarity between host species, but
390 also on opportunities for parasites from one species to encounter hosts from another
391 species. This means that both geographical range overlap and ecological
392 interactions between donor and potential recipient host species may be important

393 determinants of host-shifts. These factors may obscure the phylogenetic distance
394 effect.
395 Little is known about the relative importance of (phylo)genetic vs. ecological factors
396 for host-shifts, but it appears that this varies widely across systems. On the one
397 hand, several pathogens (e.g., influenza viruses and *Mycobacterium tuberculosis*)
398 have shifted between humans and domesticated animals such as cattle or fowl –
399 species that are only distantly related to humans but have close physical contact
400 (Smith *et al.* 2009; Ren *et al.* 2016). On the other hand, several studies have
401 reported evidence for a strong phylogenetic distance effect. For example, in
402 microalgae-virus associations in the open sea where no ecological barriers to host-
403 shifts should exist, there was a clear signal for the phylogenetic distance effect
404 (Bellec *et al.* 2014). In a study of rabies in bats, host genetic distance was identified
405 as a key factor for host-shifts whereas ecological factors (range overlap and
406 similarities in roost structures) had no predictive power (Faria *et al.* 2013).
407 The case of *Wolbachia,* an intracellular bacterium infecting nematodes and
408 arthropods (Werren *et al.* 2008), indicates that even for a single parasite there may
409 be considerable variation in the relative importance of different factors affecting host-
410 shift rates. For example, *Wolbachia* underwent preferential host-shifts to related
411 species within the spider genus *Agelenopsis* (Baldo et al. 2008). By contrast, in
412 mushroom-associated dipterans, ecological similarity (mycophagous vs. non-
413 mycophagous) appeared to be an important determinant of *Wolbachia* host-shifts
414 whereas host phylogeny and sympatry did not appear to play a major role (Stahlhut
415 *et al.* 2010). In bees, neither phylogenetic relatedness between hosts nor ecological
416 interactions (kleptoparasitism) predicted *Wolbachia* host-shifts (Gerth *et al.* 2013).

417 Among different orders of arthropods, our prediction that larger clades should have
418 higher infection levels than smaller clades is not supported in *Wolbachia* (Weinert *et*
419 *al.* 2015), perhaps indicating that at least at this level the phylogenetic distance effect
420 is not important. Overall, the *Wolbachia*-arthropod system is characterised by
421 complex patterns of codiversification that differ between *Wolbachia* strains and host
422 taxa and that we are only beginning to understand (e.g., Gerth *et al.* 2014; Bailly-
423 Bechet *et al.* 2017).
424 In order to keep our model as simple as possible we made several assumptions.
425 Most importantly, we assumed that each parasite species is strictly associated with a
426 single host species only. This assumption will be met in parasites that are highly
427 specialised on their hosts or that are vertically transmitted, so that transmission
428 between host individuals belonging to different species is very limited. For parasites
429 infecting multiple hosts, we expect that the phylogenetic distance effect should be
430 less pronounced and our results therefore less applicable. For parasite speciation,
431 we assumed barring host-shifts, parasites speciate if and only if their hosts speciate.
432 Both parasite loss during host speciation and parasite speciation within a host could
433 be incorporated into our model (which already allows for multiple parasites per host),
434 but we do not expect this to affect our results qualitatively. Host-shifts were modelled
435 as density-dependent transmission events, i.e. the more host species there are
436 within the host phylogeny, the greater the rate of host-shifts for a parasite. Given that
437 tree size was roughly constant and not affected by the parasites in our model, we
438 again believe that the assumption of density-dependent (as opposed to frequency-
439 dependent) transmission is not crucial to our results. Finally, we assumed an
440 exponential decline in host-shift rates with increasing phylogenetic distance between

hosts. This is arguably the simplest function one can assume for this relationship. A sigmoidal relationship has also been proposed (Engelstädter & Hurst 2006) and in a study of RNA viruses in mammals was found to explain the data better than the exponential function (Cuthill & Charleston 2013), but it remains to be seen how general this result is.

In conclusion, we have developed a model of host-parasite codiversification that should be most suitable for parasites that are host-specific and undergo preferential host-shifts according to the phylogenetic distance effect. Our model provides a novel framework to understand host-shift dynamics across large numbers of host species and over long evolutionary time periods. This framework has enabled the generation of several testable predictions regarding the distribution and frequency of parasites, highlighting the importance of host phylogeny in shaping the process of codiversification.

Acknowledgments

We thank Sylvain Charlat, Ben Longdon, Daniel Ortiz-Barrientos and Tanja Stadler for helpful discussions and Sylvain Charlat, Ben Longdon, Nathan Medd, Damien de Vienne and Lucy Weinert for insightful comments on our manuscript. NF acknowledges funding from an Australian Postgraduate Award and a Global Change Scholars Award from The University of Queensland.

462 # References

463 1.

464 Alcala, N., Jenkins, T., Christe, P. & Vuilleumier, S. (2017). Host shift and
465 cospeciation rate estimation from co-phylogenies. *Ecol Lett*, 20, 1014-1024.

466 2.

467 Bailly-Bechet, M., Martins-Simoes, P., Szollosi, G.J., Mialdea, G., Sagot, M.F. &
468 Charlat, S. (2017). How Long Does Wolbachia Remain on Board? *Mol Biol Evol*,
469 34, 1183-1193.

470 3.

471 Baldo, L., Ayoub, N.A., Hayashi, C.Y., Russell, J.A., Stahlhut, J.K. & Werren, J.H.
472 (2008). Insight into the routes of *Wolbachia* invasion: high levels of horizontal
473 transfer in the spider genus *Agelenopsis* revealed by *Wolbachia* strain and
474 mitochondrial DNA diversity. *Mol. Ecol.*, 17, 557-569.

475 4.

476 Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effect
477 Models Using lme4. *Journal of Statistical Software*, 67, 1-48.

478 5.

479 Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C. *et al.*
480 (2015). Cophylogeny reconstruction via an approximate Bayesian computation.
481 *Syst Biol*, 64, 416-431.

482 6.

483 Bellec, L., Clerissi, C., Edern, R., Foulon, E., Simon, N., Grimsley, N. *et al.* (2014).
484 Cophylogenetic interactions between marine viruses and eukaryotic
485 picophytoplankton. *BMC Evol Biol*, 14, 59.

7.

Charleston, M.A. & Robertson, D.L. (2002). Preferential Host Switching by Primate Lentiviruses Can Account for Phylogenetic Similarity with the Primate Phylogeny. *Syst. Biol.*, 51 528-535.

8.

Choi, Y.J. & Thines, M. (2015). Host Jumps and Radiation, Not Co-Divergence Drives Diversification of Obligate Pathogens. A Case Study in Downy Mildews and Asteraceae. *PLoS One*, 10, e0133655.

9.

Clark, N.J. & Clegg, S.M. (2017). Integrating phylogenetic and ecological distances reveals new insights into parasite host specificity. *Mol Ecol*, 26, 3074-3086.

10.

Clayton, D.H., Bush, S.E., Goates, B.M. & Johnson, K.P. (2003). Host defense reinforces host-parasite cospeciation. *PNAS*, 100, 15694-15699.

11.

Colless, D.H. (1982). Book Review of "Phylogenetics: The Theory and Practice of Phylogenetic Systematics", by E. O. Wiley. *Syst Zool*, 31, 100-104.

12.

Crook, M. (2014). The dauer hypothesis and the evolution of parasitism: 20 years on and still going strong. *Int J Parasitol*, 44, 1-8.

13.

Cuthill, J.H. & Charleston, M.A. (2013). A Simple Model Explains the Dynamics of Preferential Host Switching among Mammal Rna Viruses. *Evolution*, 67, 980-990.

14.

510 de Vienne, D.M., Giraud, T. & Shykoff, J.A. (2007). When can host shifts produce
511 congruent host and parasite phylogenies? A simulation approach. *J. Evol. Biol.*,
512 20, 1428-1438.
513 15.
514 de Vienne, D.M., Hood, M.E. & Giraud, T. (2009). Phylogenetic determinants of
515 potential host shifts in fungal pathogens. *J Evol Biol*, 22, 2532-2541.
516 16.
517 de Vienne, D.M., Refregier, G., Lopez-Villavicencio, M., Tellier, A., Hood, M.E. &
518 Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing,
519 evidence from natural associations and relation to coevolution. *The New*
520 *phytologist*, 198, 347-385.
521 17.
522 Drinkwater, B. & Charleston, M.A. (2016). Towards sub-quadratic time and space
523 complexity solutions for the dated tree reconciliation problem. *Algorithms Mol Biol*,
524 11, 15.
525 18.
526 Engelstädter, J. & Hurst, G.D.D. (2006). The dynamics of parasite incidence across
527 host species. *Evol. Ecol.*, 20, 603-616.
528 19.
529 Faria, N.R., Suchard, M.A., Rambaut, A., Streicker, D.G. & Lemey, P. (2013).
530 Simultaneously reconstructing viral cross-species transmission history and
531 identifying the underlying constraints. *Philosophical transactions of the Royal*
532 *Society of London. Series B, Biological sciences*, 368, 20120196.
533 20.

534 Gerth, M., Gansauge, M.T., Weigert, A. & Bleidorn, C. (2014). Phylogenomic
535 analyses uncover origin and spread of the Wolbachia pandemic. *Nat Commun*, 5,
536 5117.
537 21.
538 Gerth, M., Rothe, J. & Bleidorn, C. (2013). Tracing horizontal Wolbachia movements
539 among bees (Anthophila): a combined approach using multilocus sequence typing
540 data and host phylogeny. *Mol Ecol*, 22, 6149-6162.
541 22.
542 Gilbert, G.S. & Webb, C.O. (2007). Phylogenetic signal in plant pathogen-host range.
543 *Proc Natl Acad Sci U S A*, 104, 4979-4983.
544 23.
545 Heard, S.B. (1992). Patterns in Tree Balance among Cladistic, Phenetic, and
546 Randomly Generated Phylogenetic Trees. *Evolution*, 46, 1818-1826.
547 24.
548 Hurst, C.J. (2016). *The Rasputin effect : when commensals and symbionts become*
549 *parasitic*. Springer.
550 25.
551 Keeling, M.J. & Rohani, P. (2008). *Modeling infectious diseases in humans and*
552 *animals*. Princeton University Press, Princeton.
553 26.
554 Longdon, B., Brockhurst, M.A., Russell, C.A., Welch, J.J. & Jiggins, F.M. (2014). The
555 evolution and genetics of virus host shifts. *PLoS Pathog*, 10, e1004395.
556 27.

557 Longdon, B., Hadfield, J.D., Day, J.P., Smith, S.C., McGonigle, J.E., Cogni, R. *et al.*
558 (2015). The causes and consequences of changes in virulence following pathogen
559 host shifts. *PLoS Pathog*, 11, e1004728.
560 28.
561 Longdon, B., Hadfield, J.D., Webster, C.L., Obbard, D.J. & Jiggins, F.M. (2011). Host
562 phylogeny determines viral persistence and replication in novel hosts. *PLoS*
563 *Pathog*, 7, e1002260.
564 29.
565 Oksanen, J., Blanchet, F.G., Michael Friendly, M., Kindt, R., Legendre, P., McGlinn,
566 D. *et al.* (2017). vegan: Community Ecology Package.
567 30.
568 Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and
569 evolution in R language. *Bioinformatics*, 20, 289-290.
570 31.
571 Perlman, S.J. & Jaenike, J. (2003). Infection success in novel hosts: an experimental
572 and phylogenetic study of Drosophila -parasitic nematodes. *Evolution*, 57, 544-
573 557.
574 32.
575 R Core Team (2017). R: A language and environment for statistical computing. R
576 Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-
577 project.org/.
578 33.

579 Ren, H., Jin, Y., Hu, M., Zhou, J., Song, T., Huang, Z. *et al.* (2016). Ecological
580 dynamics of influenza A viruses: cross-species transmission and global migration.
581 *Scientific reports*, 6, 36839.
582 34.
583 Revolution Analytics & Weston, S. (2015a). doParallel: Foreach Parallel Adaptor for
584 the 'parallel' Package. R package version 1.0.10. https://CRAN.R-
585 project.org/package=doParallel.
586 35.
587 Revolution Analytics & Weston, S. (2015b). foreach: Provides Foreach Looping
588 Construct for R. R package version 1.4.3. https://CRAN.R-
589 project.org/package=foreach.
590 36.
591 Ricklefs, R.E., Outlaw, D.C., Svensson-Coelho, M., Medeiros, M.C., Ellis, V.A. &
592 Latta, S. (2014). Species formation by host shifting in avian malaria parasites.
593 *Proc Natl Acad Sci U S A*, 111, 14816-14821.
594 37.
595 Russell, J.A., Goldman-Huertas, B., Moreau, C.S., Baldo, L., Stahlhut, J.K., Werren,
596 J.H. *et al.* (2009). Specialization and Geographic Isolation among Wolbachia
597 Symbionts from Ants and Lycaenid Butterflies. *Evolution*, 63, 624-640.
598 38.
599 Sharp, P.M., Bailes, E., Gao, F., Beer, B.E., Hirsch, V.M. & Hahn, B.H. (2000).
600 Origins and evolution of AIDS viruses: estimating the time-scale. *Biochem Soc*
601 *Trans*, 28, 275-282.
602 39.

603 Smith, N.H., Hewinson, R.G., Kremer, K., Brosch, R. & Gordon, S.V. (2009). Myths
604 and misconceptions: the origin and evolution of Mycobacterium tuberculosis.
605 *Nature reviews. Microbiology*, 7, 537-544.
606 40.
607 Stahlhut, J.K., Desjardins, C.A., Clark, M.E., Baldo, L., Russell, J.A., Werren, J.H. *et*
608 *al.* (2010). The mushroom habitat as an ecological arena for global exchange of
609 Wolbachia. *Mol. Ecol.*, 19, 1940-1952.
610 41.
611 Streicker, D.G., Turmelle, A.S., Vonhof, M.J., Kuzmin, I.V., McCracken, G.F. &
612 Rupprecht, C.E. (2010). Host phylogeny constrains cross-species emergence and
613 establishment of rabies virus in bats. *Science*, 329, 676-679.
614 42.
615 Taylor, L.H., Latham, S.M. & Woolhouse, M.E.J. (2001). Risk factors for human
616 disease emergence. *Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci.*, 356, 983-989.
617 43.
618 Tinsley, M.C. & Majerus, M.E.N. (2007). Small steps or giant leaps for male-killers?
619 Phylogenetic constraints to male-killer host shifts. *Bmc Evolutionary Biology*, 7.
620 44.
621 Waxman, D., Weinert, L.A. & Welch, J.J. (2014). Inferring host range dynamics from
622 comparative data: the protozoan parasites of new world monkeys. *The American*
623 *naturalist*, 184, 65-74.
624 45.

625 Weinert, L.A., Araujo-Jnr, E.V., Ahmed, M.Z. & Welch, J.J. (2015). The incidence of
626 bacterial endosymbionts in terrestrial arthropods. *Proceedings. Biological*
627 *sciences / The Royal Society*, 282, 20150249.
628 46.
629 Werren, J.H., Baldo, L. & Clark, M.E. (2008). *Wolbachia*: master manipulators of
630 invertebrate biology. *Nat. Rev. Microbiol.*, 6, 741-751.
631 47.
632 Werren, J.H., Zhang, W. & Guo, L.R. (1995). Evolution and phylogeny of *Wolbachia*:
633 Reproductive parasites of arthropods. *Proc. R. Soc. Lond. B*, 261, 55-63.
634 48.
635 Wickham, H. & Chang, W. (2017). devtools: Tools to Make Developing R Packages
636 Easier.
637 49.
638 Wickham, H., Danenberg, P. & Eugster, M. (2017). roxygen2: In-Line Documentation
639 for R.
640 50.
641 Wieseke, N., Hartmann, T., Bernt, M. & Middendorf, M. (2015). Cophylogenetic
642 Reconciliation with ILP. *Ieee Acm T Comput Bi*, 12, 1227-1235.
643 51.
644 Wolfe, N.D., Dunavan, C.P. & Diamond, J. (2007). Origins of major human infectious
645 diseases. *Nature*, 447, 279-283.
646 52.
647 Wood, J.L., Leach, M., Waldman, L., Macgregor, H., Fooks, A.R., Jones, K.E. *et al.*
648 (2012). A framework for the study of zoonotic disease emergence and its drivers:

spillover of bat pathogens as a case study. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367, 2881-2892.

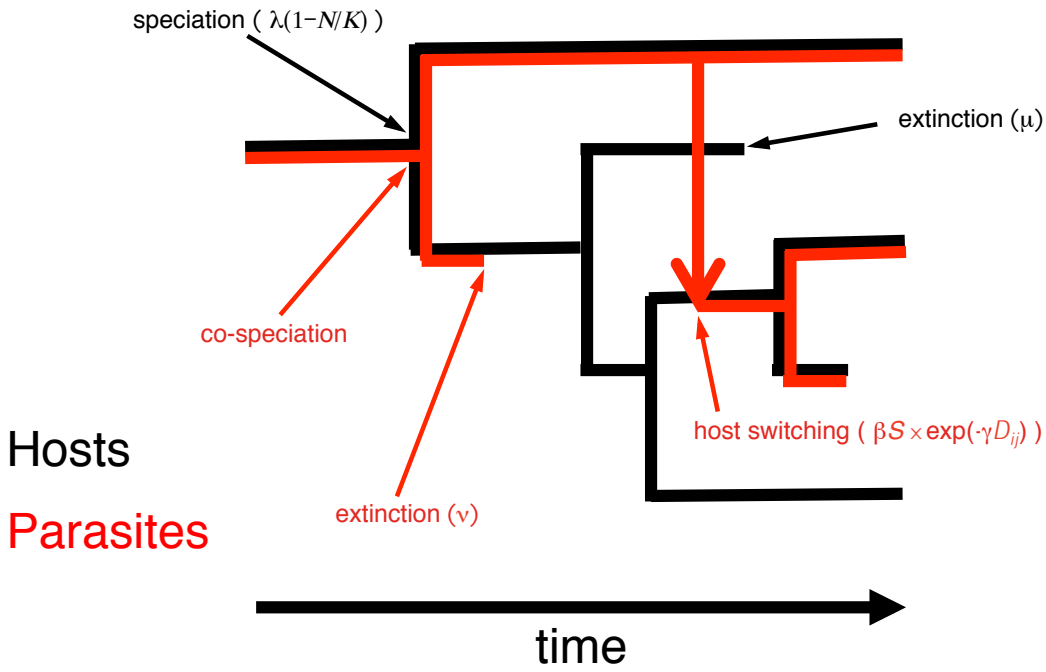

Figure 1. Illustration of the model; see Methods section for details.

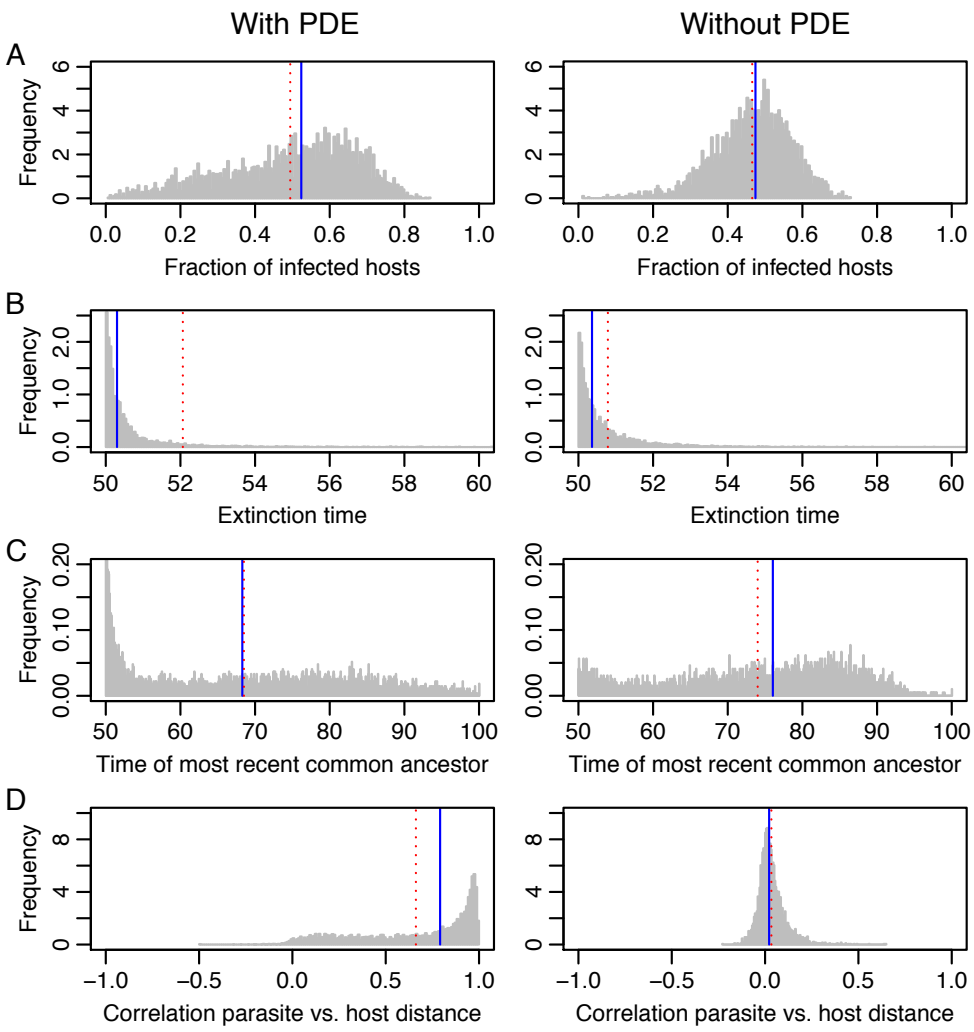

Figure 2. Summary statistics for simulations in the presence and absence of the phylogenetic distance effect, with the standard host tree set and standard PDE vs. no-PDE parameters. Panel (A) shows the distribution of the fraction of infected host species across the 10,000 simulations, contingent on 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 (where time=100 is the present). 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.

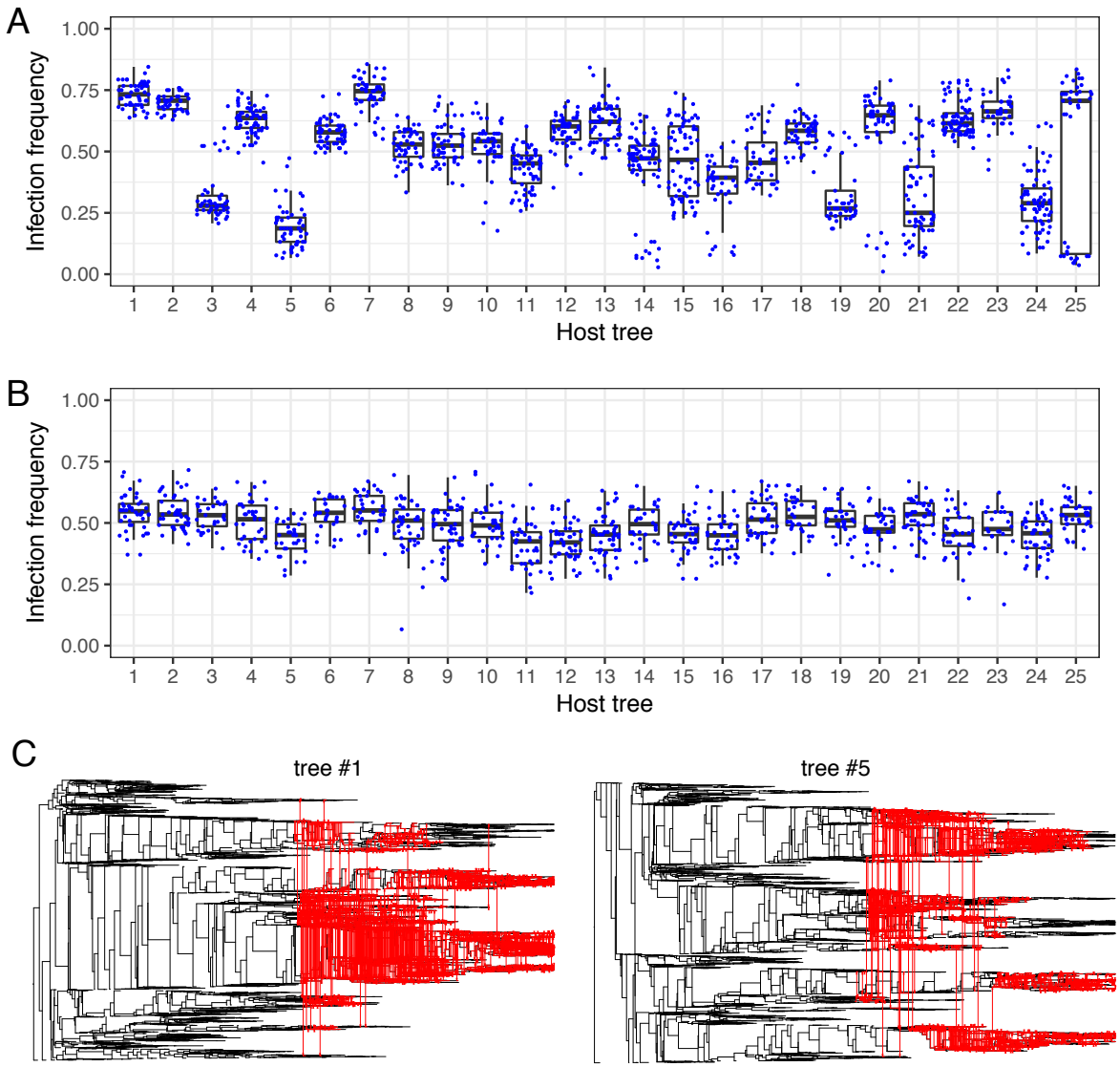

Figure 3. Distributions of infection frequencies with (A) and without (B) the phylogenetic distance effect on the first 25 host trees. Each dot shows the fraction of infected host species at the end of a simulation run. Simulations in which the parasites did not survive until the end of the simulation 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. Panel (C) shows examples host-shift dynamics for two of the host trees in presence of the phylogenetic distance effect, yielding final infection frequencies of 74% and 24%, respectively. For larger trees and more examples, see Figure S2.

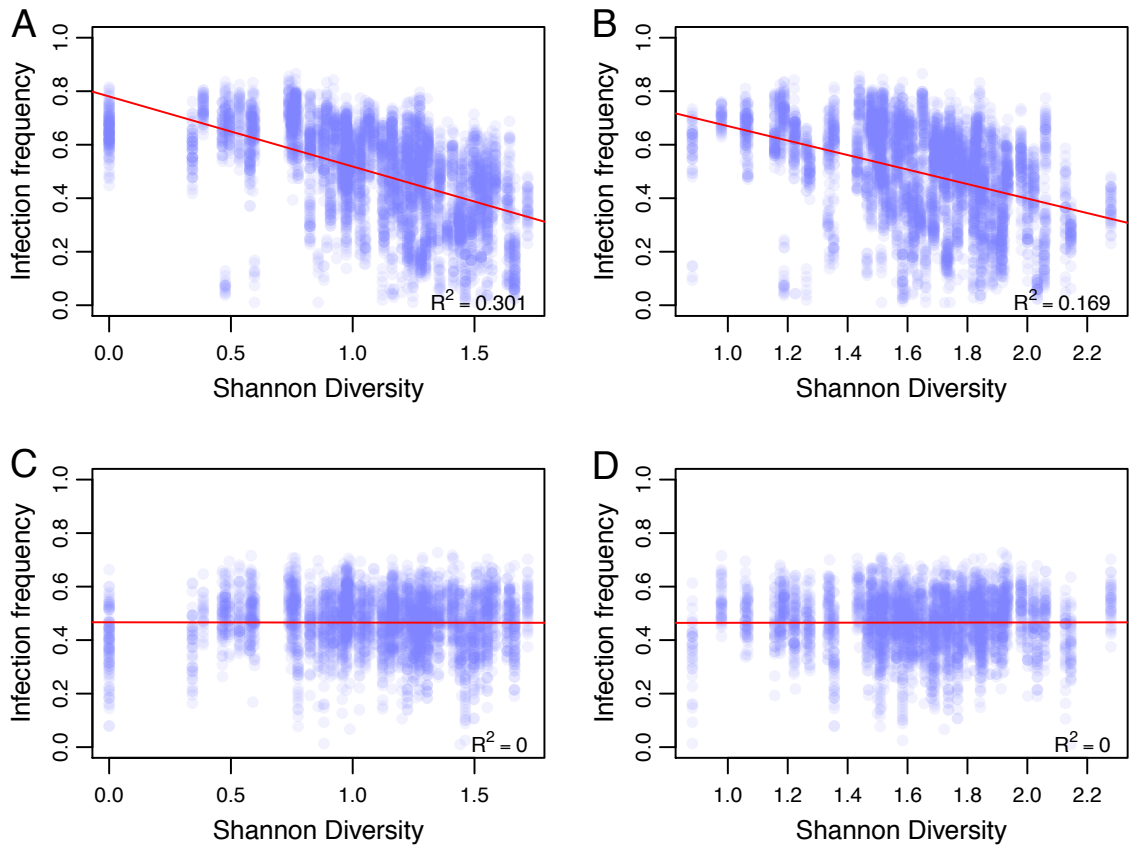

Figure 4. Fraction of infected hosts at the end of simulations against the Shannon index of host species distribution within the respective host tree, with (A,B) or without (C,D) the phylogenetic distance effect. Each dot represents the outcome of a single simulation; simulations in which the parasites became extinct were discarded. Partitioning of host trees into subtrees (or clades) and calculating the Shannon index was performed as described in SI 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.

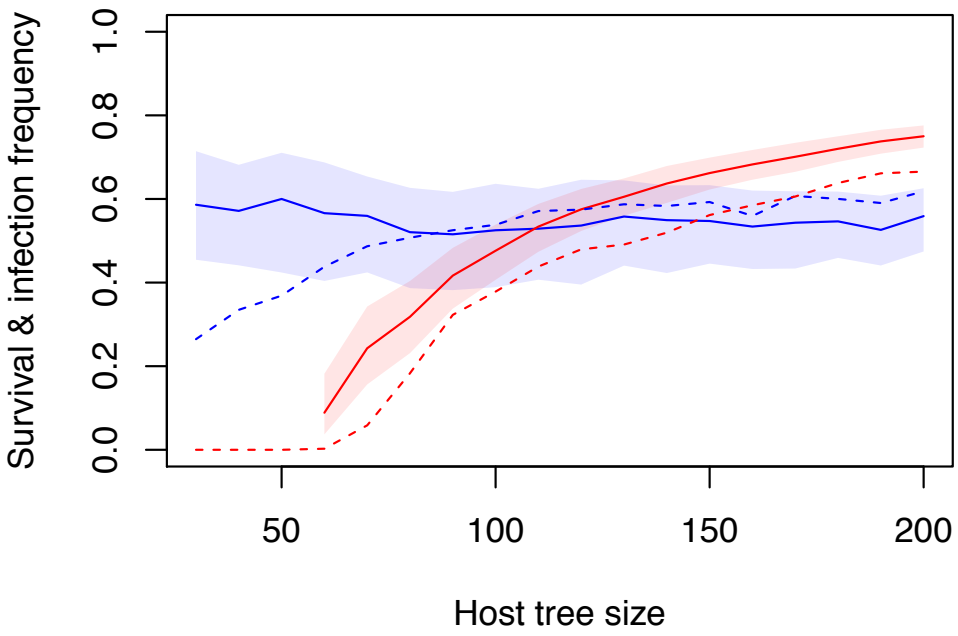

Figure 5. Influence of the equilibrium host tree size on parasite survival rates and infection frequencies in presence (blue) and absence (red) of the phylogenetic distance effect. Dashed lines show the fraction of simulations in which the parasites invaded the host tree and survived until the end of the simulations. Solid lines show the median fraction of infected host species at the end of the simulations for those simulations in which the parasites survived, with shadings indicating the interquartile range. Equilibrium host tree size was modified by varying the carrying capacity parameter *K* over a range of values from 60 to 400. All other parameters take standard values.

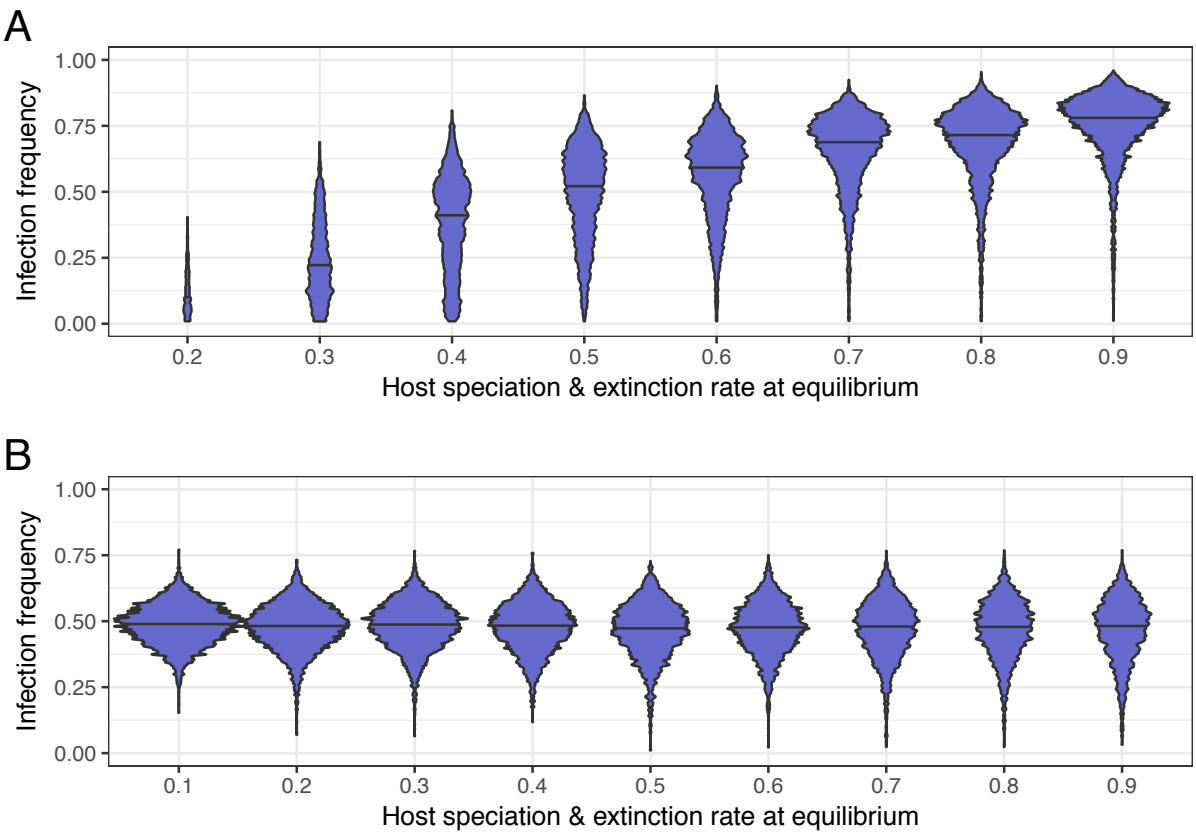

Figure 6. The impact of host speciation and extinction rate at equilibrium on the fraction of infected host species with (A) and without (B) the phylogenetic distance effect. Violins show the distribution of infection frequencies, with the total area of each violin being proportional to the number of simulations where the parasites survived. Equilibrium speciation and extinction rates where varied by using host extinction rates μ ranging from 0.1 to 0.9. At the same time, we varied the host speciation rate λ from 0.6 to 1.4 in order to maintain a constant net diversification rate of $\lambda-\mu=0.5$ during the early stages of host evolution. Parasite parameters take standard PDE and no-PDE values.