Nonrandom associations of maternally transmitted symbionts in insects: The roles of drift versus biased cotransmission and selection

Hugo Mathé-Hubert[1] | Heidi Kaech[1,2] | Corinne Hertaeg[1,3] | John Jaenike[4] | Christoph Vorburger[1,2]

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

[2]Institute of Integrative Biology, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[3]Institute of Agricultural Sciences, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[4]Department of Biology, University of Rochester, Rochester, NY, USA

Correspondence
Hugo Mathé-Hubert, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
Email: hugomh@gmx.fr

Present address
Hugo Mathé-Hubert, CNRS UMR 5525, Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques, Applications, Grenoble (TIMC-IMAG), Université Grenoble Alpes, Grenoble, France

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: CRSII3_154396; US National Science Foundation, Grant/Award Number: 1144581

Abstract

Virtually all higher organisms form holobionts with associated microbiota. To understand the biology of holobionts we need to know how species assemble and interact. Controlled experiments are suited to study interactions between particular symbionts, but they only accommodate a tiny portion of the diversity within each species. Alternatively, interactions can be inferred by testing if associations among symbionts in the field are more or less frequent than expected under random assortment. However, random assortment may not be a valid null hypothesis for maternally transmitted symbionts since drift alone can result in associations. Here, we analyse a European field survey of endosymbionts in pea aphids (*Acyrthosiphon pisum*), confirming that symbiont associations are pervasive. To interpret them, we develop a model simulating the effect of drift on symbiont associations. We show that drift induces apparently nonrandom assortment, even though horizontal transmissions and maternal transmission failures tend to randomise symbiont associations. We also use this model in the approximate Bayesian computation framework to revisit the association between *Spiroplasma* and *Wolbachia* in *Drosophila neotestacea*. New field data reported here reveal that this association has disappeared in the investigated location, yet a significant interaction between *Spiroplasma* and *Wolbachia* can still be inferred. Our study confirms that negative and positive associations are pervasive and often induced by symbiont-symbiont interactions. Nevertheless, some associations are also likely to be driven by drift. This possibility needs to be considered when performing such analyses, and our model is helpful for this purpose.

KEYWORDS
coalescence, drift, symbiont-symbiont interactions, symbiotic community

1 | INTRODUCTION

Some of the interactions between organisms are so tight and durable that a new level of organisation has been defined to describe them: the holobiont (Margulis & Fester, 1991; Queller & Strassmann, 2016). These interactions are rarely bipartite and instead generally involve a host with a microbial community of varying degrees of complexity. From the host's perspective, these associations often lead to the acquisition of novel traits, allowing the host to expand its ecological niche (e.g., Brucker & Bordenstein, 2012; Henry et al., 2013; Oliver, Degnan, Burke, & Moran, 2010). Understanding the evolutionary ecology of these interactions requires identifying how species assemble to form holobionts, both at the ontogenetic and evolutionary levels.

Large-scale screens for well-known species like *Wolbachia*, *Cardinium* or *Spiroplasma* suggest that the majority of arthropod species are infected with heritable endosymbionts (Duron et al., 2008; Hilgenboecker, Hammerstein, Schlattmann, Telschow, & Werren, 2008; Regassa, 2014; Zchori-Fein & Perlman, 2004). However, there is considerable variability in the effects these symbionts have on their hosts and in their prevalence among species. *Wolbachia* is probably the most widespread of these endosymbionts. It has been estimated to occur in 66% of arthropod species, and it typically has either low (<10%) or very high (>90%) prevalence within species (Hilgenboecker et al., 2008). *Wolbachia* is mainly known as a reproductive parasite (Werren, Baldo, & Clark, 2008), but it may also protect its host against parasites (e.g., Faria et al., 2016; Hedges, Brownlie, O'Neill, & Johnson, 2008; Teixeira, Ferreira, & Ashburner, 2008) and is sometimes necessary for successful offspring production (Dedeine et al., 2001; Kremer et al., 2009). Other widespread endosymbionts of arthropods are bacteria of the genus *Spiroplasma*, infecting 4%–7% of species, often with a low prevalence (Duron et al., 2008; Regassa, 2014), although prevalence can be high in some cases, such as in *Myrmica* ants (Ballinger, Moore, & Perlman, 2018) and in *Harmonia axyridis* (Goryacheva, Blekhman, Andrianov, Romanov, & Zakharov, 2018). Known effects of *Spiroplasma* also include reproductive parasitism (e.g., Anbutsu, Lemaitre, Harumoto, & Fukatsu, 2016; Sanada-Morimura, Matsumura, & Noda, 2013; Tabata et al., 2011) as well as defence against at least three different kinds of parasites (Ballinger & Perlman, 2017; Frago et al., 2017; Lukasik, Guo, Van Asch, Ferrari, & Godfray, 2013; Xie, Butler, Sanchez, & Mateos, 2014).

The pea aphid, *Acyrthosiphon pisum*, is one of the main biological models of endosymbiosis. It can be host to at least eight facultative heritable endosymbionts (Vorburger, 2018), including *Spiroplasma* (ixodetis clade; Fukatsu, Tsuchida, Nikoh, & Koga, 2001; Simon et al., 2011). Interestingly, Ferrari, West, Via, and Godfray (2012) showed that the communities of facultative symbionts differed strongly among host plant-associated biotypes of the pea aphid (Peccoud, Ollivier, Plantegenest, & Simon, 2009), although the prevalence of *Spiroplasma* is only weakly affected by biotype, which explains only 9% of the variance (Ferrari et al., 2012). A symbiont that experiences solely vertical transmission can persist in a host population as a reproductive parasite, or by providing a benefit to offset the cost it inflicts on the host. For example, *Spiroplasma* may protect pea aphids against entomopathogenic fungi (Lukasik, van Asch, Guo, Ferrari, & Godfray, 2013) or parasitoid wasps (Frago et al., 2017). However, this cost-benefit balance varies depending on the environment, which is thought to be the main reason for the observed polymorphism of facultative symbiont communities. For example, defensive symbioses depend on the presence of some parasites of the host, and some symbioses help the host to cope with warm environments (e.g., Russell & Moran, 2006). The cost-benefit balance may also depend on the associations with other symbionts. If two symbionts provide the same service, then one of them might be redundant and thus too costly to the host. This may be the reason why defensive bacterial symbionts are less frequent in aphids protected by ants (Henry, Maiden, Ferrari, & Godfray, 2015), or why the two defensive symbionts *Serratia symbiotica* and *Hamiltonella defensa* rarely co-occur in pea aphids (Oliver, Moran, & Hunter, 2006). Also, interactive effects between symbionts make the outcome of a given association difficult to predict. For instance, in *A. pisum*, *H. defensa* increases the titre of *S. symbiotica*, but *S. symbiotica* does not affect the titre of *H. defensa* (Oliver et al., 2006). In the presence of *Spiroplasma*, *H. defensa* decreases the fecundity of its host *A. pisum* while it increases the fecundity of the aphid *Sitobion avenae* (Lukasik, Guo, et al., 2013).

Interactive effects that vary from one symbiont strain to another limit the utility of controlled laboratory experiments, which usually include only a few particular strains, for making predictions about the overall interactions among symbionts in natural populations. For this reason, results from controlled experiments are often compared to analyses of field surveys (for several examples, see Zytynska & Weisser, 2016). These analyses notably aim at identifying pairs of symbionts for which the co-occurrence is more or less frequent than expected under the null hypothesis of random assortment (hereafter, positive and negative associations). Three kinds of mechanisms are generally considered when trying to explain such deviation from random assortment. First, the symbionts could interact in a way that increases or decreases the rate of maternal transmission failures (e.g., Rock et al., 2017), which should lead to negative or positive associations, respectively. Second, the symbionts could have an interactive effect on host fitness, enhancing or hindering their co-transmission to the next generation (e.g., Oliver et al., 2006). Thirdly, Jaenike (2012) and Smith et al. (2015) suggested a mechanism by which neutral or even slightly costly maternally transmitted symbionts could spread in the host population. These symbionts could by chance hitchhike alongside a successful symbiont whose fitness benefits outweigh the costs of the hitchhiker. Rapid spread has been reported for *Rickettsia* and *Spiroplasma* in the whitefly *Bemisia tabaci* and in *Drosophila neotestacea*, respectively (Cockburn et al., 2013; Himler et al., 2011; Jaenike, Unckless, & Cockburn, 2010). If the spreading matriline was initially associated with another symbiont as well, then faithful maternal transmission would drag it along even if it were advantageous for the host to lose the hitchhiking symbiont. This symbiont hitchhiking is analogous to genetic hitchhiking (or draft), where a neutral or slightly deleterious mutation spreads in the population because of its linkage disequilibrium with a beneficial mutation (Felsenstein, 1974). Symbiont hitchhiking might be responsible for the evolutionary maintenance of the dominant strain of the symbiont called X-type in North America. This strain is costly to its host, has not been found to provide any counterbalancing benefit, but is positively associated with the defensive symbiont *H. defensa* (Doremus & Oliver, 2017).

However, most symbionts are not strictly maternally transmitted. For example, *Rickettsia* can be transmitted via plants in whiteflies (Caspi-Fluger et al., 2012), *Spiroplasma* can be transmitted via parasitic mites in flies (Jaenike, Polak, Fiskin, Helou, & Minhas, 2007) and *Hamiltonella* can be transmitted via parasitoids in aphids (Gehrer & Vorburger, 2012). Both *H. defensa* and *Regiella insecticola* show occasional paternal transmission (Moran & Dunbar, 2006). Jaenike (2012) argued that because of these nonmaternal transmission routes and because most symbionts show some degree of maternal transmission failure, associations due to symbiont hitchhiking should disappear rapidly. Thus, in most cases, the presence of positive (or negative) associations between symbionts should suggest an interaction that favours (or hinders) their co-occurrence. Jaenike, Stahlhut, Boelio, and Unckless (2010) showed that *Spiroplasma* and *Wolbachia* in *D. neotestacea* are positively associated despite imperfect maternal transmission. By combining these observations with a mathematical model, they suggested that these two symbionts are likely to be interacting positively with each other. As we will show in this paper, positive and negative associations are also expected to appear and persist by drift, implying that without information about the effective female population size, one needs to be cautious in assigning biological meaning to such associations.

In the first part of this study, we used a field survey of *A. pisum* symbiotic infections to identify positive and negative associations among symbionts. This analysis confirmed several previous findings that associations of symbionts often deviate from random assortment (Figure 1a). In the second part of this study, in order to understand the evolutionary meaning of these associations, we developed a model simulating the evolution of the frequency of symbiont communities in the presence of maternal transmission failures, horizontal transmissions, selection and drift. The model shows that associations of symbionts are expected to be produced by drift provided that the rates of maternal transmission failure, of horizontal transmission and the effective female population size are not too high (Figure 1b). In the third part of this study, we used the same model in the approximate Bayesian computation (ABC) framework to reanalyse the observed positive association between *Spiroplasma* and *Wolbachia* in *D. neotestacea* (Jaenike, Stahlhut, et al., 2010), combining old data (2001–2009) with new data (2010–2016). This analysis suggests that the observed dynamics of infection involve a positive interactive effect of the two symbionts on host fitness (Figure 1d).

2 | MATERIALS AND METHODS

2.1 | Natural symbiont co-occurrence

2.1.1 | Field sampling and symbiont screening

We sampled 498 aphids in France, Switzerland, Germany and Denmark during autumn 2014 and spring and summer 2015. We selected colonies that were at least 2 m apart from each other to lower the proportion of clones sampled more than once. For each sample, we recorded the host plant and the GPS coordinates. We characterised the presence of seven facultative endosymbionts by diagnostic PCR using symbiont-specific primers to amplify a part of the 16S rRNA gene (Table S1). We excluded *Wolbachia* from this analysis because of its low frequency. DNA was extracted from individual aphids using the "salting out" protocol (Sunnucks & Hales, 1996) and the PCR cycling conditions are described by Henry et al. (2013). We also ran a diagnostic PCR for the obligate endosymbiont *Buchnera aphidicola*, which is present in all aphids and thus served as an internal positive control for the quality of the DNA preparation. The nine samples that tested negative for *B. aphidicola* were excluded from the final data set. Because we had a particular interest in *Spiroplasma* infecting pea aphids (Mathé-Hubert, Kaech, Ganesanandamoorthy, & Vorburger, 2019), we also analysed the distribution of intraspecific diversity in this symbiont. This phylogenetic analysis is further described in the Supplementary material S1 and uses the strains of *Spiroplasma* described in Table S2. This analysis identified three main clades of *Spiroplasma* from pea aphids that are later referred to as clades 1, 2 and 3.

A natural population of *D. neotestacea* was sampled monthly from May through September from 2010 through 2016. During this time of year, the generation time of *D. neotestacea* is probably on the order of one month or less. Adult flies were collected by sweep netting over mushroom (*Agaricus bisporus*) baits that had been placed in a forested area in the city of Rochester, New York. Flies were screened for *Wolbachia* and *Spiroplasma* infection using the PCR methods described in Jaenike, Stahlhut, et al. (2010).

2.2 | Statistical analysis

All analyses were performed using the R software (version 3.4.4; R Core Team, 2018). Generally, associations of symbionts that are more or less abundant than expected under random assortment would be analysed using statistical tests that assume independence of observations. Our data do not fulfil this assumption as aphid samples were obtained from many different locations and dates. We thus accounted for potential spatiotemporal autocorrelation by predicting the presence or absence of symbiont species with a regression random forest model (RF). This approach is of similar efficiency as usual spatial models (Fouedjio & Klump, 2019; Hengl, Nussbaum, Wright, Heuvelink, & Gräler, 2018). In each RF explaining the presence or absence of one symbiont species in pea aphid individuals, the following explanatory variables were used: latitude, longitude, season (number of days since the start of the year), host plant on which the aphid has been sampled, aphid colour (pink or green), presence or absence of the six other symbionts (one variable per symbiont) and the total number of other symbiont species infecting the aphid. The significance of these explanatory variables was estimated using FDR adjusted *p*-values (hereafter, FDR *p*-values). The details of this analysis are described in the Supplementary material S2.

To avoid lumping together aphids of different biotypes and thus simplify the interpretation, we re-fitted these random forest models separately to aphids sampled on *Medicago sativa* and on *Trifolium* spp., which represent 30% and 33% of all field samples, respectively. We refer hereafter to these three types of models as RF_{WD} (whole data set), RF_{M} (*Medicago*) and RF_{T} (*Trifolium*). For RF_{M} and RF_{T} the host plant was removed from the set of explanatory variables. This analysis was also run to investigate the intraspecific distribution of *Spiroplasma*, by predicting, for each *Spiroplasma* infected aphid, the phylogenetic clade of *Spiroplasma* (clades 1, 2 or 3).

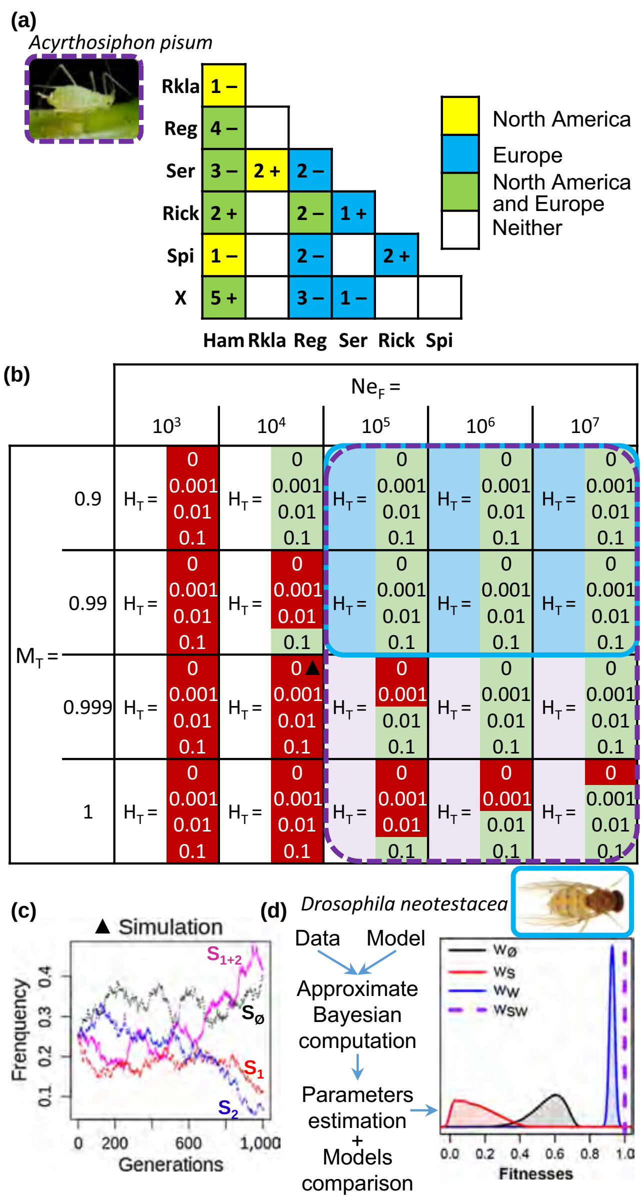


FIGURE 1 Summary of the findings. (a) For 21 pairwise combinations of pea aphid symbionts, the number of studies (including this one) that found positive (+) or negative (−) associations, and geographical locations where the associations were found. Ham, *Hamiltonella*; Rkla, *Rickettsiella*; Reg, *Regiella*; Ser, *Serratia*; Rick, *Rickettsia*; Spiro, *Spiroplasma*; X, X-type. (b) Results of the simulation analysis: combination of parameters where drift induces nonrandom assortment. Red and green values refer to combinations of parameters for which the type 1 error rate is higher and lower than 5%, respectively (based on simulated field samples of $N = 500$). Ne_F, effective female population size; M_T, maternal transmission rate; H_T, horizontal transmission rate. The blue square highlights the range of parameters values that is likely to include the symbionts of *D. neotestacea* while the dotted purple one is likely to include those of the pea aphid (Chen & Purcell, 1997; Moran & Dunbar, 2006; Peccoud et al., 2014; Rock et al., 2017). (c) Example of simulation where drift created a strong positive association. The parameters used for this simulation are pointed (▲) on the panel b: M_T, 0.999; H_T, 0; Ne_F, 10^4. (d) Analysis of the evolution of the *Spiroplasma-Wolbachia* association in *D. neotestacea* (Figure 6) in the approximate Bayesians likelihood framework. The density plot shows the approximate posteriors of the fitnesses of each type of fly infection relative to the fitness of flies infected by both symbionts. $w_ø$, aposymbiotic; w_s, *Spiroplasma* only; w_w, *Wolbachia* only; w_{sw}, coinfected flies. Sources for images: PLoS Biology Issue Image (2010); Werner & Jaenike (2017)

These analyses revealed that some symbionts are less frequent in aphids already containing other symbiont species, while others were not affected. To further investigate this, we characterised the link between the frequency of each symbiont species and the average number of additional symbiont species with which it co-occurs. We also investigated the effect of drift on this link. This analysis is further described in the Supplementary material S3.

2.3 | A model of evolution of symbiont co-occurrences

We developed a model of evolution of maternally transmitted symbiont co-occurrence for two purposes. Firstly, we wanted to assess the effect of drift on deviations from random assortment in the presence of various rates of maternal transmission failure and horizontal transmissions. Secondly, we used this model to analyse the well-documented case of a symbiont association between *Wolbachia* and *Spiroplasma* in *D. neotestacea*, for which estimates of the relevant parameters are available.

In short, considering only the two symbionts case, this model simulates populations of female hosts reproducing with nonoverlapping generations and being infected by zero, one or two different symbionts (species or strains). Symbionts are maternally and horizontally transmitted at varying rates. The strength of this model is that it simulates different events (reproduction and horizontal and maternal transmissions) by performing random samplings in the relevant probability distributions to update the headcount of the different types of infections, which avoids simulating every individual. The different steps for which we generate these randomly sampled values are represented by questions a–f in Figure 2. This allows the model to be fast without assuming an infinite population size. This rapidity is needed to simulate a large number of generations and replicates (simulation study), and to simulate a large number of parameter combination (ABC study). Because this model studies maternally transmitted symbionts, it only simulates females. Fitness in this model thus scales with the capacity of females to produce daughters.

With only two types of symbionts (S_1 and S_2), the population is described by the number of females being aposymbiotic ($S_ø$), having only one of the symbionts (S_1 and S_2), or having both ($S_{1,2}$). At each generation, we simulated horizontal transmissions, reproduction events and maternal transmission failures (Figure 2). The total number of horizontal transmissions is randomly chosen from a Poisson distribution whose mean depends on the horizontal transmission rate (H_T) and the frequency of the transmitted

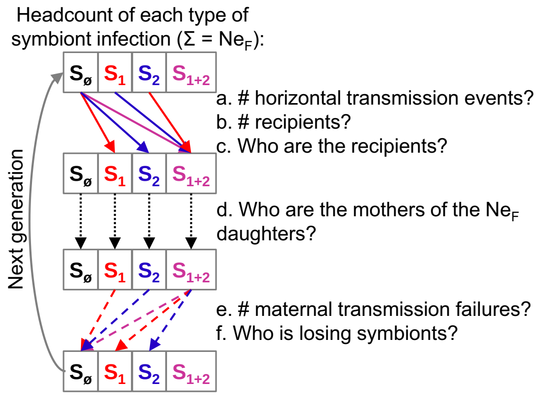


FIGURE 2 Model of evolution of symbiont associations. The population is represented as four cells corresponding to the four types of infection. Each host generation is simulated in three steps. Firstly, horizontal transmissions change some individuals from one category to the other by gaining one (or two) symbionts (solid arrows coloured as a function of the gained symbionts). This is simulated answering the questions a, b and c by randomly sampling the appropriate probability distribution. Then the reproduction is simulated by randomly choosing mothers according to the fitnesses induced by each type of infection (stippled arrows). Finally, maternal transmission failures change some individuals from one category to the other by losing one (or two) symbionts (dotted arrows coloured as a function of the lost symbionts)

symbiont (Figure 2: question a). The number of recipients can be lower than the number of horizontal transmission events when individuals receive the same symbiont more than once. The number of recipients is thus randomly chosen in a binomial distribution in which the mean depends on the number of horizontal transmissions previously drawn randomly and the total number of individuals in the population (Figure 2: question b). Finally, the repartition of the recipients among the four host classes ($S_ø$, S_1, S_2, and $S_{1,2}$) is chosen randomly from a hypergeometric distribution to simulate samplings without replacement (Rice, 2006; Figure 2: question c). Reproduction is simulated by sampling mothers from a multinomial distribution described by the headcount of females with the four kinds of symbiont communities and scaled by their relative fitness, which is determined by the fitness effects of the symbionts, which can be assumed to be multiplicative or interactive in the case of double infections. This simulates samplings with replacement (Rice, 2006; Figure 2: question d).

Maternal transmissions are simulated using the same general logic as horizontal transmissions (Figure 2: questions e and f). For more details, see the Supplementary material S4. The model is available as an R function in Appendix S1.

2.4 | Can deviations from random assortment appear by drift?

The first aim of this model was to investigate the effect of drift on symbiont associations. Hence we did not simulate any interactive effect of the symbionts on host fitness or maternal transmission, but we had to assume some noninteractive effects of the symbionts on host fitness to stabilise the polymorphism of infection, which would otherwise have disappeared rapidly under many parameter combinations (e.g., frequent maternal transmission failures or horizontal transmissions).

Specifically, we simulated 3,000 replicates of all combinations of the following sets of parameters: Effective female population sizes (Ne_F: 10^3, 10^4, 10^5, 10^6, and 10^7), successful maternal transmission rates (M_T: 1, 0.999, 0.99, 0.90), horizontal transmission rates (i.e., Average number of horizontal transmission events caused by each infected host; H_T: 0, 0.001, 0.01, 0.1). The parameter values M_T = 1 and H_T = 0 are unrealistic but were investigated to ensure that even biological systems with extremely low or high values such as associations between strains of obligatory symbionts are in the explored parameter space.

In the absence of selection, for most combinations of M_T and H_T, the symbionts get either rapidly fixed or lost. This absence of polymorphism prevents assessing deviations from random assortment. To slow down the loss of polymorphism, we set the selection on the presence of each symbiont such that it counteracts the effect of maternal transmission failures and of horizontal transmissions. The fitness of aposymbiotic hosts was set to one. Then the fitness of those infected by only one symbiont species was set to the value that, in an infinite population, would keep the frequency of the symbiont constant. For individuals infected by both symbionts, the fitness is the product of the fitnesses induced by each of its symbionts. This multiplicative fitness is similar to the model used by Jaenike, Stahlhut, et al. (2010), and corresponds to an absence of interaction between the symbionts. For more detail, see Supplementary material S5.

Populations were initiated by randomly picking the frequency of each symbiont in a uniform distribution to then set the headcount of the four kinds of symbiont communities ($S_ø$, S_1, S_2, and $S_{1,2}$) according to these frequencies and to the assumption of random assortment. The evolution of these populations of randomly assorted symbionts was then simulated for 10^5 generations or stopped if the polymorphism of infection was lost. This large number of generations was needed because the initial state of the populations, where symbionts are randomly assorted, might have actually never existed in natural populations. Therefore, the time needed for drift to induce apparent nonrandom assortment should be interpreted as an estimation of the strength of the effect of drift. This also allowed to assess the stability of deviations from random assortment once they appeared, which can take a long time in large populations.

At each generation, 500 individuals were randomly sampled from the population and used to test the significance of the deviation from the assumption of random assortment using a Chi-square test and to assess the sign of the deviation. The *p*-values were computed at every generation and recorded at generations 0, 10, 10^2, 10^3, 10^4 and 10^5. The *p*-values computed at every generation were used to assess if, as it is often assumed, associations lasting for multiple generations are unlikely to be caused by drift. We estimated the number of generations needed for a previously significantly positive association to become significantly negative and vice-versa. This was computed for each replicate as the number of generations between the first significant deviation from random assortment and the end of the simulation divided by the number of such inversions.

2.5 | Analysing a real data set while accounting for drift

Jaenike, Stahlhut, et al. (2010) argued that *Wolbachia* and *Spiroplasma* in *D. neotestacea* are probably interacting in a way that enhances the fitness of coinfected hosts. Indeed, these two symbionts are positively associated in natural populations, despite having a maternal transmission rate of approximately 0.96, which should rapidly randomise them. We used our model in the ABC framework (Approximate Bayesian Computation) to assess how robust this conclusion is to drift. We combined the data gathered from 2001 to 2009 and analysed by Jaenike, Stahlhut, et al. (2010) with additional data gathered from 2010 to 2016 (Table S3). We fitted to these data an interaction model and a no interaction model, the latter being similar to the model of Jaenike, Stahlhut, et al. (2010). We tested for the interaction on the host fitness twice. Firstly, we looked at the interaction model and tested whether the distribution of the approximate posteriors of the interactions included zero. Secondly, we compared the quality of fit of the two models.

We used the function "ABC_rejection" of the R package "EasyABC" (Jabot, Faure, Dumoulin, & Albert, 2015) to estimate the relative fitness induced by the different kinds of infections. This approach compares observed data to the data simulated with varying values for the parameters to be estimated. To compare observed and simulated data, we assigned each field sample to one generation assuming that there were five generations per year (as Jaenike, Stahlhut, et al., 2010; details in Table S3). According to this assumption, the data set spans 77 *Drosophila* generations.

In these simulations, the parameters that are not estimated need to be fixed. These parameters are Ne_F, M_T and H_T. For M_T, we used the estimates of Jaenike, Stahlhut, et al. (2010) that range from 0.945 to 0.981. To be conservative in inferring potential interactions, we used values for Ne_F and H_T that should overestimate the effect of drift. We assumed no horizontal transmissions ($H_T = 0$) because they decrease the effect of drift (result of the simulation study). This assumption is reasonable given that a high association between the type of symbiotic infection and the mitochondrial haplotype has been observed (Jaenike, Stahlhut, et al., 2010). The effective population size (Ne) was approximately estimated using the formula $Ne = \pi/4\mu$ where μ is the mutation rate of *Drosophila melanogaster* ($\mu = 2.8 \times 10^{-9}$; 95% CI = $[10^{-9}; 6.1 \times 10^{-9}]$; Keightley, Ness, Halligan, & Haddrill, 2014) and π the nucleotide diversity ($\pi = 0.0237$; 95% CI = [0.0135; 0.0337] estimated by bootstrapping autosomal loci; Pieper & Dyer, 2016). This gives an estimate of $Ne = 2 \times 10^6$, but to be conservative, we used an underestimation of the Ne using the lower CI of π and the upper CI of μ, which gives $Ne_{F\_min} = 2.8 \times 10^5$, assuming a sex-ratio of 0.5.

The other parameters were estimated by randomly sampling their values from uniform priors. These parameters are the initial frequencies of *Spiroplasma* and *Wolbachia*, their initial association (measured with the phi coefficient; Everitt & Skrondal, 2010) and the fitnesses induced by the different combinations of symbionts ($w_ø$, w_s, w_w and w_{sw}). For the initial association and frequencies, the uniform prior ranged from −1 to 1 and 0 to 1 respectively. For the fitnesses, we used the same approach as Jaenike, Stahlhut, et al. (2010) which modelled the cost of not having a symbiont and took the fitness of coinfected individuals as reference by setting $w_{sw} = 1$. For the no interaction model, we estimated the fitness effect of the two other types of infected individuals (w_s and w_w) using the uniform prior ranging from 0 to 2 and we constrained $w_ø$ to be equal to $w_w \times w_s$, which assumes a multiplicative fitness effect as in Jaenike, Stahlhut, et al. (2010). For the interaction model, we also estimated $w_ø$ using the same priors as for w_s and w_w. This allowed the absence of *Spiroplasma* and *Wolbachia* to have an interactive effect on the host fitness.

For both models, the priors were randomly sampled 10^8 times. For each simulation, the randomly drawn initial symbiont frequencies and coefficient of association were used to initiate the population whose evolution was simulated by the model for 77 generations according to randomly drawn fitness effects of symbionts and the fixed parameters (Ne_F, H_T, M_T). The two sets of 10^8 simulated data sets were summarised and compared to the summary of the observed data set. These summaries contain the mean frequencies of the four types of infections at start, midpoint and end (details in Table S3). Simulations were "accepted" and used to estimate parameters when the Euclidean distance between their summary and the summary of the observed data was below the tolerance threshold of 0.153. This tolerance was chosen to accept at least 1,000 simulations per model, which is 0.001% of the simulations.

We estimated the cost of not having *Wolbachia*, *Spiroplasma* or their synergetic effect by applying a similar formalism as Jaenike, Stahlhut, et al. (2010) to the distribution of the approximate posteriors of the fitnesses. Jaenike, Stahlhut, et al. (2010) modelled the cost of not having a symbiont by setting $w_{sw} = 1$; $w_s = 1 - s_w$; $w_w = 1 - s_s$; $w_ø = (1 - s_w) \times (1 - s_s)$, where s_s and s_w are the cost of not having *Spiroplasma* or *Wolbachia*, respectively. This corresponds to the situation modelled by the no interaction model, while for the interaction model, we extended this formalism by setting $w_ø = (1 - s_w) \times (1 - s_s) \times (1 - s_{sw})$, where s_{sw} is the cost of not having the synergetic effect between *Spiroplasma* and *Wolbachia*.

These fitnesses and costs, as well as the initial state of the population, were estimated using the mode of posterior distributions of the interaction and no interaction models, and their 95% confidence interval using the 0.025 and 0.975 percentiles. We performed a pairwise comparison of the estimated fitness effect of the four types of infections. For each pair of infection type we tested the significance of the differences by subtracting their approximate posterior distributions and assessing the extent to which the resulting distribution overlaps with zero. Specifically, we tested the null hypothesis $w_1 = w_2$ using the two-sided Bayesian p-value: $\begin{cases} 2\times \text{Freq.}(w_1 < w_2) & \text{if } \overline{w_1} > \overline{w_2} \\ 2\times \text{Freq.}(w_1 > w_2) & \text{if } \overline{w_1} < \overline{w_2} \end{cases}$ as where the frequencies (Freq.) are estimated over the posteriors (i.e., the "accepted" simulations). We further tested the interaction by comparing the quality of fit of the two models through the delta of the Bayesian predictive information criterion (BPIC; Ando, 2007; Turner, Sederberg, & McClelland, 2014).

3 | RESULTS

3.1 | Natural co-occurrence of pea aphid symbionts

The random forests (RFs) model analysing the associations of symbionts in field-sampled aphids revealed three positive associations and six negative associations. Of these associations, all were detected in the whole data set (RF_{WD}, Figure 3a) six were detected in aphids from *Trifolium* spp. (RF_T, Figure 3b), and only two were detected in aphids from *M. sativa* (RF_M, Figure 3c). The sample size and the average number of symbionts per aphid were similar in the *M. sativa* (RF_M) and in the *Trifolium* spp. (RF_T) group (*M. sativa*: 148 aphids with 0.97 symbionts per aphid on average; *Trifolium*: 161, 0.77). Therefore, it is unlikely that the lower number of significant associations in *M. sativa* (RF_M) is caused by lower statistical power. Of the 11 significant associations already identified by other studies on pea aphids, six were also found in this study, and all associations reported by several studies (including ours) were always of the same sign (Table 1; Figure 1a). Particularly noteworthy are the consistently negative associations between the common symbionts *H. defensa* and *R. insecticola*, and the consistently positive associations between *H. defensa* and X-type.

To account for the nonindependence between samples, these models included the variables longitude, latitude, season and host plant (only RF_{WD}). However, these variables are highly correlated, and although we used conditional inference trees and conditional importance, the results should be interpreted with caution. The effects of these four variables on the frequency of each symbiont are described in Figure S1.

Some symbiont prevalences covaried negatively with the total number of coinfecting symbiont species, whatever their identity (*H. defensa*: FDR p-values = .002 and .02 in RF_{WD} and RF_M, respectively; *R. insecticola*: FDR p-value < .001 in the three models RF_{WD}, RF_T and RF_M; *S. symbiotica*: FDR p-values < .001 and in both RF_{WD} and RF_M; Figure 4). For pea aphids from *Trifolium* spp., the relationship between symbiont prevalence and the mean number of coinfecting symbionts was tight (R^2 = 0.98). The slope was more negative than −1 which is the slope expected under random assortment (slope = −2.16; p-value < .001, Figure 4b). This observation was mostly driven by *R. insecticola*. However, repeating the analysis without aphids infected by *R. insecticola* did not change the result much (R^2 = 0.88; slope = −2.26; p-value = .003). For pea aphids from *M. sativa*, there was no detectable relationship between the frequency of symbionts and the number of coinfecting symbiont species (R^2 = 0.14; slope = −1.61; p-value = .35, Figure 4a). The simulations described in Supplementary material S3 revealed that this relationship is also affected by drift, which increases variation in the slopes around the expected value of −1 and moderately decreases the proportion of variance explained.

3.2 | *Spiroplasma* intraspecific diversity

The phylogenetic tree indicates that in Europe, pea aphid infecting *Spiroplasma* are subdivided into at least three clades, although clade 3 has low bootstrap support (Figure S2). The relative frequencies of these three clades did not depend on the host plant (p-value = .98; Figure S2) but were strongly dependent on the symbiont community. Clade 2 was more frequent in aphids already infected by other endosymbionts (FDR p-value = .01) than the other two clades. The difference of clade 2 to clade 1 was marginally nonsignificant, while the difference to clade 3 was marginally significant (p-values = .06 and .03, respectively; Wilcoxon-test). The *Spiroplasma* clades were also differently associated with *H. defensa*, X-type and *Rickettsia* (FDR p-values = .02, .003 and .003, respectively). Specifically, clade 3 co-occurs less frequently with *H. defensa* than clades 1 and 2 (p-values = .02 and .01; Fisher-exact test) and more frequently with X-type than clades 1 and 2 (p-values = .003 and .006; Fisher-exact test; Figure 3 and Figure S2). Also, clade 2 is more frequently associated with *Rickettsia* than clades 1 and 3 (p-values < .001 in both cases; Fisher-exact tests; Figure 3 and Figure S2).

3.3 | Simulations of the symbiont co-occurrences evolving by drift

Symbiont associations that are more or less frequent than expected under random assortment are generally thought to be the signature of an interaction between the symbionts that promotes or prevents their co-occurrence. Our simulations showed that when $M_T = 1$ and $H_T = 0$, drift always leads to strong deviations from random assortment, although associations take longer to establish in large populations where drift is weak (Figure 5). As expected, less-than-perfect maternal transmission or horizontal transmission tend to randomize symbiont associations (Figure 5 and Figure S3). However, our model shows that this effect can be offset by drift, in particular under effective population sizes lower than 10^6 (Figure 5). For effective female population sizes (Ne_F) of 10^3, 10^4 and 10^5 or more, it takes a median number of 54, 117, and 211 generations to inverse the sign of a significant deviation from random assortment (Figure S4). Symbiont associations due to drift alone can thus be quite persistent in time.

3.4 | *Spiroplasma-Wolbachia* association in *D. neotestacea*: Drift or selection?

The positive association between *Spiroplasma* and *Wolbachia* in *D. neotestacea* reported in Jaenike, Stahlhut, et al. (2010) has declined slowly from 2001 to 2016 and now seems absent. The frequency

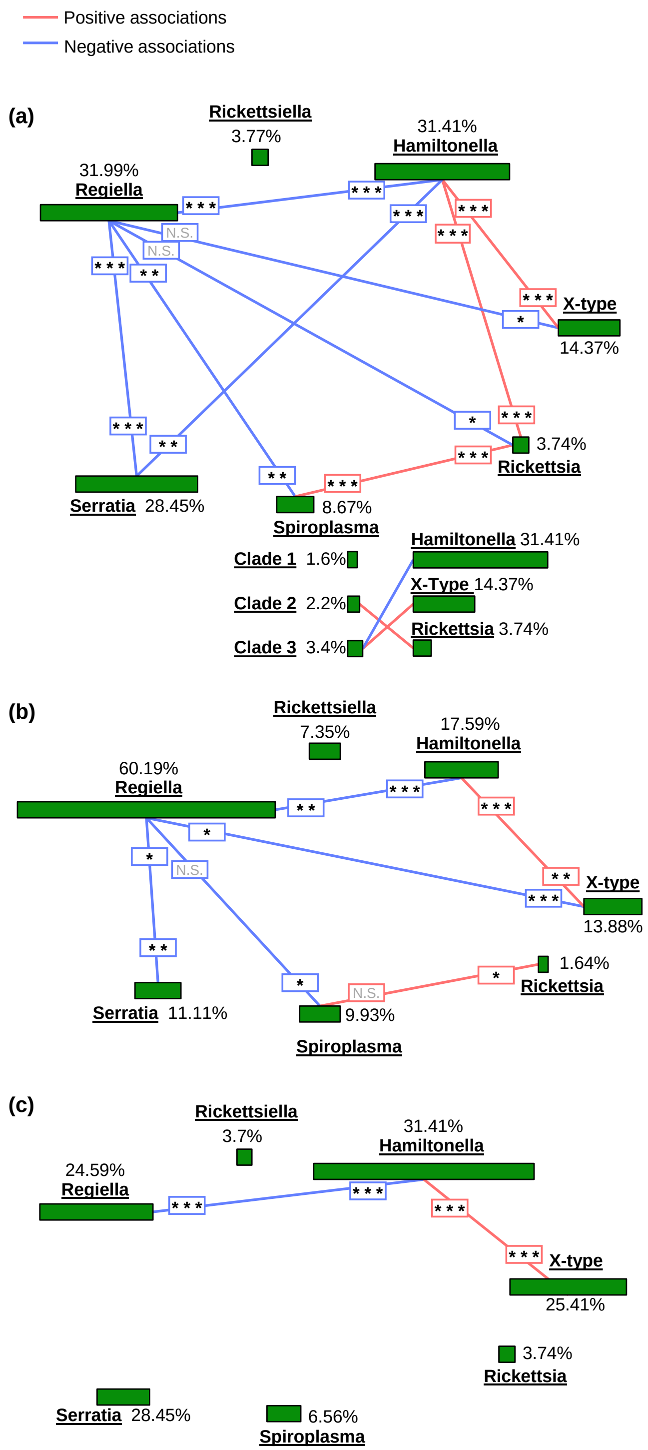


FIGURE 3 Patterns of symbiont co-occurrence. The seven symbiont species are represented by green boxes whose size is proportional to the overall prevalence of the symbiont in the whole data set (a; N = 498), in aphids from *Trifolium* spp. (b; N = 161) and in aphids from *Medicago sativa* (c; N = 148). Red and blue lines connect symbionts that co-occur more or less often than expected under random assortment, respectively. Stars indicate the FDR-adjusted level of significance of these associations and are placed close to the symbiont that was the dependent variable in the random forest models

of coinfected flies has shifted from approximately 0.75 to 0.4 while the frequency of flies infected by *Wolbachia* only has shifted from approximately 0.2 to 0.5 (Figure 6). The disappearance of the positive association calls into question the previous conclusion of an interactive fitness effect of these two symbionts. However, when we compare a no interaction and an interaction model fitted to these data, we do indeed find support for a positive interaction on host fitness. The relationship between the values of the parameters and the distance between the summaries of the simulated and observed data sets are shown in Figure S5. The estimated parameters revealed a clear interaction since the fitnesses of the four host categories had the following rank order: $w_s < w_ø < w_w < w_{ws} = 1$ (p-values = .006, <.001, <.001 respectively; Figure 1d; estimated values are in Table 2). This sorting resulted in a cost of not having the synergetic effect of *Spiroplasma* and *Wolbachia* that is twice higher than the cost of not having the beneficial effect of *Wolbachia* (s_{ws} = 0.78 vs. s_w = 0.38; interaction model in Table 2). This interaction is also demonstrated by a delta of Bayesian predictive information criterion of 7.75. The interaction model also revealed that infection by *Spiroplasma* would actually be costly to the host, since the cost of not having *Spiroplasma* was negative (s_s = −3.13, 95% CI = [−90.4; −0.36]; Table 2). Consistent with this result, strains of *D. neotestacea* only infected by *Spiroplasma* can be difficult to maintain in the laboratory (John Jaenike, personal observation).

4 | DISCUSSION

Understanding how symbionts associate and interact within a host is important but challenging. Laboratory experiments address this question by controlling all relevant parameters and observing the outcomes, but they can only accommodate a tiny portion of the natural diversity of each interacting species. In addition, such studies have often found that the outcome depends on the genotypes of the interacting partners (e.g., Hansen, Vorburger, & Moran, 2012; Lukasik, Asch, et al., 2013; Niepoth, Ellers, & Henry, 2018; Oliver, Degnan, Hunter, & Moran, 2009; Russell & Moran, 2006; Vorburger & Gouskov, 2011; Weldon, Strand, & Oliver, 2013), further complicating general predictions about these interactions in natural populations. Comparisons with field observations are therefore essential. When analysing field surveys, interactions between symbionts are tentatively inferred by comparing the observed frequency of co-occurrences to the frequency expected under the hypothesis of random assortment. Departures from random assortment have been reported frequently in pea aphids. Indeed, of the 21 possible pairwise associations among the seven facultative endosymbionts considered here, 11 have already been reported to have significantly higher or lower frequencies than expected under random assortment in earlier studies on pea aphids (Figure 1a and Table 1). Six of these associations were also found in our field sampling, and three are reported for the first time. When focusing on *Spiroplasma*, we even found significant associations at the intraspecific level. The three main *Spiroplasma* clades identified in the phylogenetic tree

TABLE 1 Patterns of symbiont co-occurrence in this study and in other studies on pea aphids

	RF_{WD}	RF_T	RF_M	Oliver et al. (2006)	Rock et al. (2017)	Ferrari et al. (2012)		Russell et al. (2013)	Henry et al. (2013)
Host plant	Many	T	M	M	M	T	V	T V M	Many
Geographic location	E.	E.	E.	E.	N.A.	E.	E.	N.A.	N.A.& E. (14 countries)
Regiella/Serratia	–	–							
Regiella/Spiroplasma	–	–							
Regiella/Rickettsia	–				–				
Regiella/X-type	–	–				–			
Regiella/Hamiltonella	–	–	–			–		–	
Serratia/Rickettsia						+			
Serratia/X-type							–		
Serratia/Hamiltonella	–			–	–				
Serratia/Rickettsiella					+			+	
Spiroplasma/Rickettsia	+	+							
Spiroplasma/Hamiltonella					–				
Rickettsia/Hamiltonella	+				+				
X-type/*Hamiltonella*	+	+	+		+				+
Hamiltonella/Rickettsiella					–				

Note: E, Europe; M, *Medicago sativa*; N.A., North America; T, *Trifolium* spp.; V, *Vicia*.

were nonrandomly associated with other symbionts, independent of the host plants the aphids were collected from. Such intraspecific variation in a symbiont-symbiont association has also been reported between X-type and *H. defensa* in the pea aphid (Doremus & Oliver, 2017). But what is the biological meaning of these pervasive associations?

4.1 | Drift induces deviations from random assortment

Our simulation model showed that, albeit a random phenomenon, drift alone can induce associations among maternally transmitted symbionts, suggesting that random assortment is not an appropriate null model to compare symbiont coinfections against. The reason is most easily understood by considering the coalescence framework. Statistical tests used to detect departures from random assortment assume that samples are independent of each other. While this may apply to horizontally transmitted symbionts, it will not apply to maternally transmitted symbionts. Some individuals will have the same symbiont association simply because they share a female ancestor that transmitted this particular symbiont community to all of its offspring. In population genetics, this phenomenon is referred to as coalescence (Balding, Bishop, & Cannings, 2007), which should not be confounded with the "community coalescence" (Rillig et al., 2015). One of the measures of the strength of drift is the expected coalescent time, the average number of generations between two randomly sampled alleles and their most recent common ancestor. It is equal to 2Ne for diploid autosomal genes, but it is only Ne/2 for maternally transmitted cytoplasmic genomes (assuming a sex-ratio of 0.5). This is because only females transmit the cytoplasmic genome, and they have only one copy of it (Jaenike, 2012; Moore, 1995). Cytoplasmic genomes, including endosymbionts, hence undergo four times more drift than nuclear autosomal genes.

Jaenike (2012) investigated how the population genetics framework can be adapted and used to study the evolution of communities of maternally transmitted symbionts by comparing each symbiont to a gene. However, given the generally high fidelity of maternal transmission and the low rate of horizontal transmission of endosymbionts, one could also compare the whole symbiont community to one gene with many alleles. Mutations increase allelic diversity, while drift has the opposite effect. This mutation-drift equilibrium is largely analogous to the balance between maternal transmission failures, horizontal transmissions and drift that we studied with our model. The main difference is that maternal transmission failure effectively acts as a directional mutation pressure, where the number of individuals mutating from one state (infected) to the other (uninfected) is proportional to the number of individuals in the original state (infected), which is not true for horizontal transmission. The probability of undergoing horizontal transmissions increases with the frequency of the symbiont, which makes polymorphism less easily maintained in the presence of horizontal transmission.

Drift-induced deviations from random assortment can persist for a very long time. In a population of diploid autosomal genes, a neutral mutation that reaches fixation does so, on average, 4Ne generations

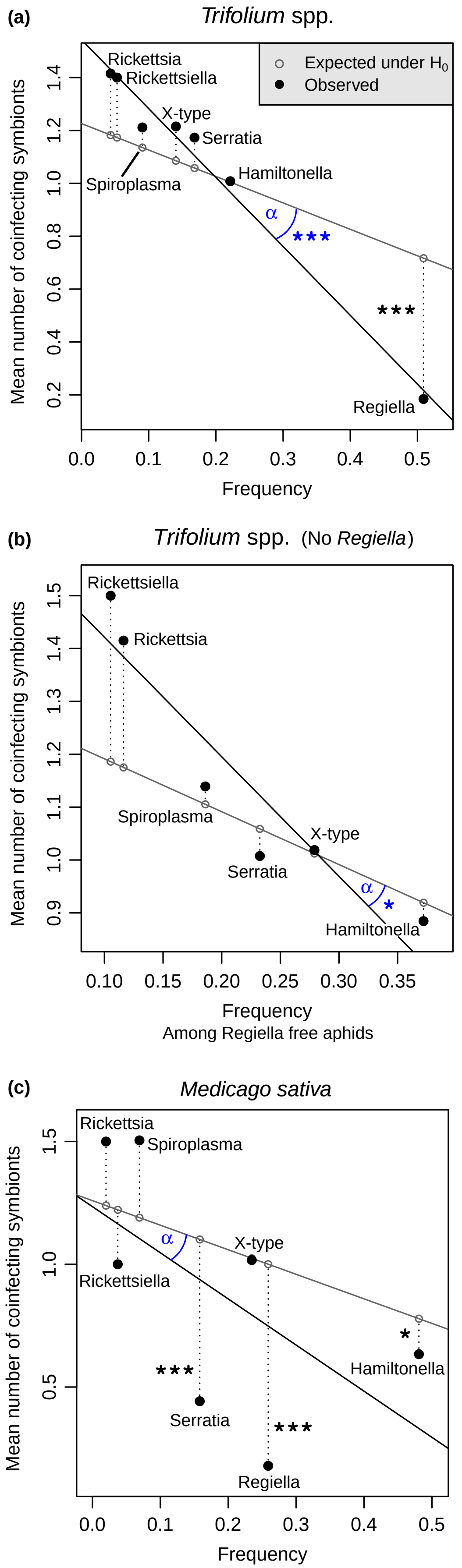


FIGURE 4 Relationship between symbiont frequency and mean number of other symbionts species. Comparison of the actual (black) and expected (grey) relationship between the frequency of endosymbiont species and the mean number of other symbiont species with which they co-occur. Each observed value is connected to its expected value by a dotted line. Stars along these lines indicate the FDR adjusted level of significance detected by random forest models. Analysis was performed on pea aphids from *Trifolium* spp. (a and b) and *Medicago sativa* (c). Panel b refers to the analysis performed on aphids from *Trifolium* spp., but excluding individual infected with *Regiella insecticola* from the analysis. For each of these three cases, we tested if the angle between the two slopes (α) differed significantly from zero

after it appeared (Kimura & Ohta, 1969), or after Ne generations in a haploid, maternally transmitted gene. Thus, we should expect that drift-induced deviations from random assortment of symbionts should also be somewhat stable in time. In agreement with that, our simulations of two strictly maternally transmitted symbionts show that drift-induced inversions of the sign of significant deviations from random assortment occur every 50–200 generations on average, depending on the effective female population size. These numbers should not be used as a general reference, however, because significance depends on the size of the samples used to assess deviations from random assortment (500 hosts in our simulations). Departures from random assortment became less stable in the presence of horizontal transmissions and maternal transmission failures.

4.2 | *Spiroplasma-Wolbachia* association in *D. neotestacea*: Drift or selection?

Jaenike, Stahlhut, et al. (2010) studied the maintenance of the positive association between *Wolbachia* and *Spiroplasma* in *D. neotestacea*. They used a deterministic mathematical model to show that given the maternal transmission rate estimated at 0.96, the association should disappear very rapidly in the absence of positive interactions between the two symbionts. While it is true that this relatively imperfect maternal transmission will push a population towards random assortment, their model only considered the frequency of the symbionts. Thus, it implicitly assumed an infinite population size and omitted drift which, as we have shown, pushes populations towards nonrandom assortment. The additional data collected since this study revealed that, at least in Rochester NY, the association has disappeared. Specifically, the frequency of coinfected flies decreased while the frequency of flies only infected by *Wolbachia* increased.

At a first glance, the disappearance of the association seems to reinforce the view that it could have been driven by drift. However, considering that the effective female population size is probably above 2.8×10^5, and the maternal transmission rate below 0.99, our simulation study revealed that drift alone is unlikely to induce such significant deviations from random assortment as they have been observed between 2001 and 2009. As discussed by Jaenike (2012), such associations could be driven by symbiont hitchhiking, if one of

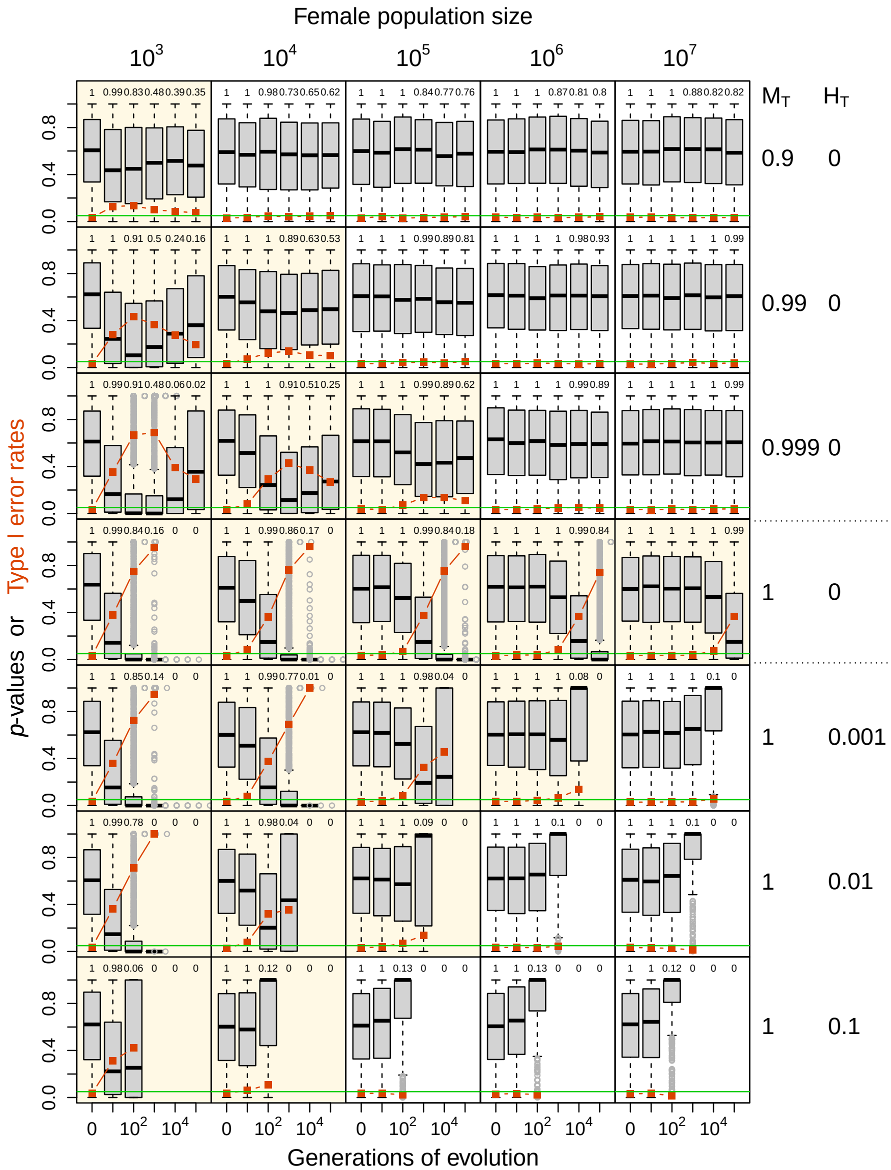


FIGURE 5 Deviations from random assortment induced by drift. The frequency of two maternally transmitted symbionts evolved for up to 10^5 generations, starting from a population in which symbionts were randomly assorted. Boxplots show the *p*-values of Chi-square tests assessing the deviation from random assortment at generations 0, 10, 10^2, 10^3, 10^4 and 10^5. Each set of boxplots corresponds to 3,000 populations evolving with the combination of the parameters indicated on the side: female population size (columns), horizontal transmission rate (H_T, rows) and maternal transmission rate (M_T, rows). The green horizontal line shows the 0.05 threshold, and the orange squares and lines indicate the type 1 error rate. Analyses of field surveys testing for deviation from random assortment usually assume that the type 1 error rate is 0.05. Combinations of parameters where this is not the case have a yellowish background. The numbers above the boxplots indicate the proportion of populations that still retained a polymorphism of infection by both symbionts

the two symbionts is beneficial and spreads in the population from a matriline also containing another symbiont. *Spiroplasma* has actually undergone such a spread (Cockburn et al., 2013; Jaenike, Unckless, et al., 2010), probably because of the protection it provides against the parasitic nematode *Howardula aoronymphium* (Jaenike, Unckless, et al., 2010). This spread could strongly decrease the female effective population size, which was only partially accounted for in our ABC analysis since we assumed that males and females have the same effective population sizes. On the other hand, we used a conservatively low estimate of the effective female population size and the analysis still supported a strong interactive effect of *Spiroplasma* and *Wolbachia* on host fitness. Indeed, in the presence of *Wolbachia*, *Spiroplasma* infected flies had the highest estimated fitness while in the absence of *Wolbachia* they had the lowest estimated fitness. Importantly, we did not estimate parameters explaining the initial association but parameters explaining the evolution of the association. Thus this interactive fitness effect is not deduced from the presence of the association, which could have been due to symbiont hitchhiking, but from the dynamic of its disappearance, which was slower than expected given the relatively high rates of maternal transmission failures (Jaenike, Stahlhut, et al., 2010). This analysis also revealed that whatever the presence of *Spiroplasma*, *Wolbachia* always increases the fitness of its host. A more unexpected result of this analysis is that in the absence of *Wolbachia*, infection with *Spiroplasma* is inferred to be costly to the host. This estimated cost contrasts with the result of Jaenike, Unckless, et al. 2010), that *Spiroplasma* is beneficial by protecting its host from the sterilising effect of the parasitic nematode *H. aoronymphium*, while having no detectable effect on the egg count per ovary.

This surprising result of the ABC analysis results from the fact that the frequency of flies infected only by *Wolbachia* increased while the frequency of flies infected only by *Spiroplasma* remained

TABLE 2 Parameters estimated by the ABC analysis fitting the model. Three kinds of parameters were estimated; the initial population state, the fitnesses corresponding to the different types of symbiont infections, and the corresponding costs of not having a symbiont. The cells in grey correspond to parameters that were estimated by solving the equations shown in the second column of the table (see Materials and methods). The 95% confidence intervals of the parameters are given in brackets

	No interaction BPIC = 25.31	Interaction BPIC = 17.56
Initial population		
f_S	0.64 (0.61; 0.68)	0.66 (0.61; 0.72)
f_W	0.64 (0.60; 0.69)	0.66 (0.59; 0.73)
Phi	0.70 (0.62; 0.78)	0.67 (0.54; 0.78)
Fitnesses (w_{sw} = 1)		
$w_{ø} = (1 - s_w) \times (1 - s_s) \times (1 - s_{ws})$	0.29 (0.10; 0.41)	0.57 (0.34; 0.69)
$w_s = (1 - s_w) \times (1 - s_{ws})$	0.31 (0.11; 0.45)	0.13 (0; 0.37)
$w_w = (1 - s_s) \times (1 - s_{ws})$	0.93 (0.92; 0.93)	0.93 (0.92; 0.93)
Costs of not having symbionts		
s_s	0.07 (0.07; 0.08)	−3.13 (−90.4; −0.36)
s_w	0.68 (0.55; 0.89)	0.38 (0.25; 0.63)
s_{sw}	Set to 0	0.78 (0.32; 0.99)

constant. We assumed that the imperfect maternal transmissions estimated by Jaenike, Stahlhut, et al. (2010) are exact and representative of the considered time series. According to this assumption and in the absence of selection, maternal transmission failures would convert coinfected flies into flies only infected by *Wolbachia* or by *Spiroplasma* (at rates of 3% and 4%, respectively) and these flies would be converted into aposymbiotic flies (at a rate of 5% and 2%, respectively). With these conversion rates and an initial coinfection frequency of 60%, about 2.4% of flies should become infected by *Spiroplasma* only every generation, yet such flies remained at a constantly low frequency, revealing the cost of *Spiroplasma* in the absence of *Wolbachia*.

This cost of *Spiroplasma* contrasts with its known protective effect, which is conditional on the presence of the parasitic nematode. Possibly, *Spiroplasma* has some fitness costs that are not detected through the egg count per ovary used as a fitness proxy by Jaenike, Unckless, et al. (2010). Nevertheless, we should also consider that this result could arise from some of the necessary approximations in our analysis. For example, we considered that the rates of successful maternal transmissions estimated by Jaenike, Stahlhut, et al. (2010) were constant over time. However, the maternal transmission rate is under strong selection and it can vary with temperature, which could have influenced our inferences. This highlights that the ABC approach applied here can be useful to test hypotheses on field data, but the resulting parameter estimates must be interpreted cautiously.

Another assumption we made is the absence of horizontal transmissions. This assumption is reasonable given the high association

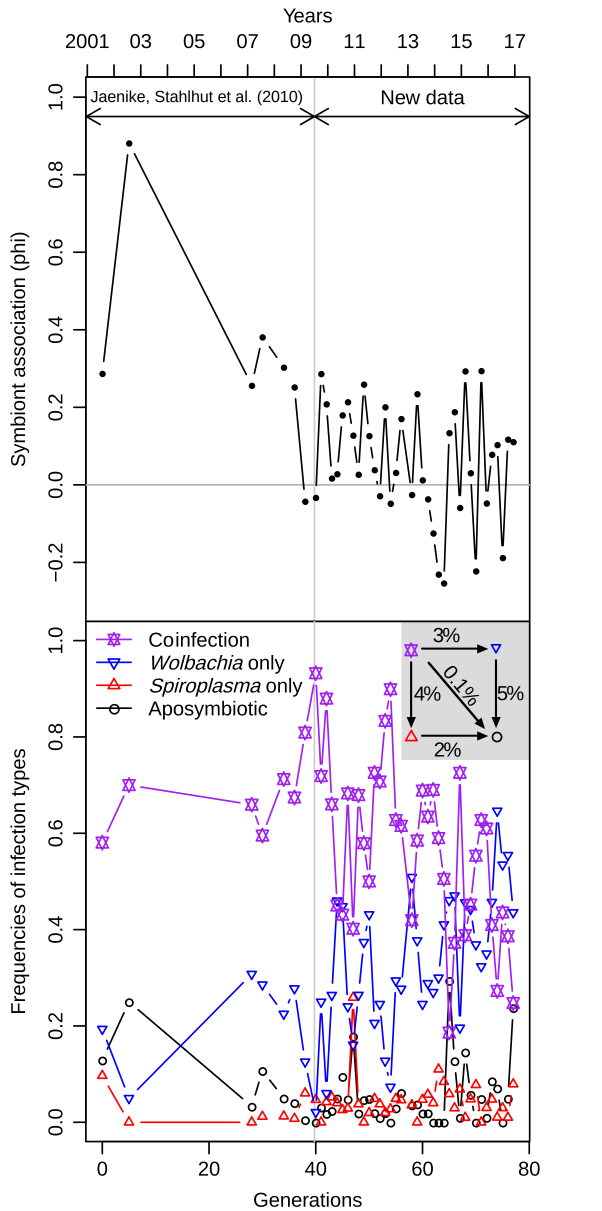


FIGURE 6 *Spiroplasma-Wolbachia* association in *D. neotestacea* in Rochester NY. The evolution of the symbiont association is shown on the upper panel while the frequencies of the four kinds of associations are shown on the lower panel. The time in years is shown at the top and the corresponding number of generations at the bottom. The diagram in the top right corner of the lower panel describes the effect of maternal transmission failures on the rates of conversion between the four types of infections. These rates were estimated by Jaenike, Stahlhut, et al. (2010) and these data were combined in the ABC framework to estimate the fitness effect of the four types of infection

observed between the infection status of the flies and their mitochondrial haplotype (Jaenike, Stahlhut, et al., 2010). This knowledge has strongly increased the statistical power of our analysis. For this reason, for any study that would plan to use such ABC approach to infer symbiont-symbiont interaction from field surveys, we would recommend to also sequence the COI gene. Then haplotypes could be included in the analysis by considering them as a symbiont species with a known perfect maternal transmission and a null rate of horizontal transmission. With such a setting, our model could also estimate the rates of horizontal transmissions of symbionts or paternal inheritance.

Fromont, Adair, and Douglas (2019) have found in a Rochester, NY population of *D. neotestacea* that the density of *Wolbachia* did not differ significantly between *Spiroplasma*-infected and uninfected flies, whereas the density of *Spiroplasma* was positively and significantly correlated with that of *Wolbachia* among coinfected flies. Together, these findings suggest that *Wolbachia* has a positive effect on *Spiroplasma* density, and thus perhaps on maternal transmission fidelity, but that *Spiroplasma* does not have such an effect on *Wolbachia*. *Wolbachia* benefits from the presence of *Spiroplasma* because of the latter's restoration of fertility in nematode-parasitized females.

4.3 | Symbiont associations in pea aphids: Selection or drift?

After emphasizing the importance of considering drift as a source of nonrandom assortment among symbionts, we return to the interpretation of positive and negative associations among facultative endosymbionts observed in pea aphids. Are they maintained by interactions among symbionts or just a consequence of drift? Good estimates of effective female population size would obviously help. Unfortunately, this is a tricky problem in aphids and other cyclical parthenogens. Although aphids can reach enormous population sizes, they undergo a bottleneck each winter, and clonal selection during the asexual phase of the life cycle (approximately 7–14 generations in pea aphids; Barker, 2016) can be intense (e.g., Vorburger, 2006), which will also reduce the effective population size. This clonal selection acts on the three components of genetic variance (additive, epistatic and of dominance), but the optimisation it induces on the nonadditive variances is lost at each sexual generation, which maintains the presence of clonal selection from year to year (Lynch & Deng, 1994). On the other hand, aphids are good dispersers and exhibit shallow genetic population structure over large geographic scales. For example, Ferrari et al. (2012) reported F_{ST}-values ranging from 0.03 to 0.11 for pea aphid populations from the same host plants across different European countries, and Via and West (2008) reported a mean F_{ST} of 0.03 for North American populations of the pea aphid. Such high population connectivity should have a positive effect on effective population size. We do not know the effective population size of pea aphids, but DNA sequence-based estimates from other cyclical parthenogens, waterfleas of the genus *Daphnia*, are rather high (300,000–600,000; Haag, McTaggart, Didier, Little, & Charlesworth, 2009). If estimates were similarly high for pea aphids, the importance of drift in generating nonrandom assortment of symbionts would be limited (Figure 5).

Another important aspect to consider is the consistency of the sign of significant associations. While drift will generate associations of random and (slowly) fluctuating sign, selection is expected to consistently favour either positive or negative associations between particular pairs of facultative endosymbionts. For significant associations that were discovered in multiple studies, the sign of the association was always the same (Table 1). Finding particular combinations of symbionts consistently over- or underrepresented across different times and places suggests they are not caused by drift. For example, the European pea aphids population is thought to have colonised North America at least 200 years ago, which would represent 1,400–2,800 pea aphid generations, and there is strong genetic differentiation among pea aphids from the two continents today (Brisson, Nuzhdin, & Stern, 2009). Despite this separation, the four associations that have been reported in both continents are of the same sign. This strongly suggest that at least some of them are driven by an interaction between the symbionts. Indeed, if these associations were inherited from the pea aphids that invaded North America, then it has been stable for more than 1,000 generations, which is unlikely for associations driven by drift (Figure S4).

In addition to testing for deviations from random assortment, some studies have also assessed whether symbiont species tend to be differently associated with aphids that are already infected with 0, 1, 2 or more other symbiont species (e.g., Ferrari et al., 2012; Rock et al., 2017; Russell et al., 2013; Zchori-Fein, Lahav, & Freilich, 2014). In our field survey, we found that *H. defensa*, *S. symbiotica* and *R. insecticola* occurred more frequently in aphids containing no or few other symbiont species than expected under the assumption of random assortment, although this was only significant in aphids sampled from *M. sativa*.

We further investigated this by characterising the link between the frequency of symbionts and the number of coinfecting symbiont species. This link is expected to be strong because frequent symbionts are less likely to share a host with other symbiont species than rare symbionts, leading to an expected slope of −1 (Supplementary material S3). This reveals that rare symbionts are more strongly selected to cope with other symbiont species than abundant symbionts (this is also true for horizontally transmitted symbionts). We found that this slope was nonsignificant in aphids sampled on *M. sativa*, and significantly lower than the expected value (−1) in aphids sampled on *Trifolium* spp. (Figure 4). These results might be the consequence of drift, constraints, or adaptations. For example, rare symbionts might be rare because they need the presence of other symbionts to persist in the host population. Such constraint would reinforce the expected relationship. Alternatively, since rare symbionts are expected to co-occur on average with more symbiont species than abundant ones, these rare symbionts might have become better adapted to the presence of other symbiont species, thus reinforcing the expected pattern. This highlights that only abundant symbiont associations are efficiently optimised by natural selection. It is therefore worth considering that associations between symbionts that are currently maintained by a positive interaction may have evolved as a consequence of an association that had initially appeared by drift or hitchhiking.

Lastly, inference on the biology of particular symbionts or their associations can be strengthened from analyses of seasonal patterns and their comparison with expectations from laboratory experiments. In studies of seasonal dynamics, the effect of drift is ideally ruled out using spatiotemporal replication. For example, Smith et al. (2015) reported a correlated change in the symbiont frequencies and the parasitoid-induced host mortality which, together with the laboratory evidence for symbiont-conferred resistance against parasitoids, suggested a causal relationship between them. Also, Montllor, Maxmen, and Purcell (2002) reported an increase in the frequency of *S. symbiotica* correlated with temperature, which was consistent with this symbiont helping to tolerate heat stress. Our sampling design was not suited for such inference, but the result that *H. defensa* was more abundant in summer than in spring (Figure S1) was at least consistent with selection by parasitoids as also reported by Smith et al. (2015). Field observations are also informative when they do not match the expectations from laboratory work. For example, laboratory experiments suggested that X-Type does not provide any detectable benefit to the pea aphid, but it is quite frequent and positively associated to *H. defensa*, suggesting it might have benefited from hitchhiking during the spread of *H. defensa* (Doremus & Oliver, 2017). Also, Wulff, Buckman, Wu, Heimpel, and White (2013) did not find that the symbiont *Arsenophonus* was protecting its *Aphis glycines* host against its main parasites, but it was present at high frequency. This discrepancy between observation and expectation motivated further experiments revealing that *Arsenophonus* provides a general – yet to be described – benefit to the aphid (Wulff & White, 2015). Although difficult to interpret, field surveys remain crucial for our understanding of the ecology of symbioses.

In conclusion, the fate of holobionts depends on host-symbiont interactions as well as on symbiont-symbiont interactions, but identifying them is not always straightforward. The approach consisting in analysing the frequency of associations in the field is useful. However, the results it yields must be interpreted carefully, particularly in the case of maternally transmitted symbionts, as patterns expected to be produced by interactions between symbionts are also induced by drift. The model we developed can help this task. The study of this model highlights that holobionts are not only a source of additional units of selection, but also a source of additional units of drift.

ACKNOWLEDGEMENTS

We are very grateful to Paula Rodriguez for her help with insect rearing and to Marco Thali and Joelle Schmid for assistance with the molecular analyses of the pea aphid field samples. We are grateful to Tom Brekke, Meghan Jacobs, Amanda Shaver, Danielle Doi, Evan Tandy, and Katie Mead, who carried out the 2010–2016 screens for *Wolbachia* and *Spiroplasma* infection in *D. neotestacea*. We also thank Ailsa McLean for sharing several *Spiroplasma*-infected aphid lines. This work was supported by a Sinergia grant from the Swiss National Science Foundation (grant No. CRSII3_154396 to CV) and a grant from the US National Science Foundation (grant No. 1144581 to JJ).

AUTHOR CONTRIBUTIONS

H.K., C.H., C.V., J.J., and H.M.H. performed the field sampling; H.K., C.H., and H.M.H. carried out the molecular analysis of the field samples; H.M.H. was responsible for the data analysis and developed the model; H.M.H., H.K., C.H., J.J., and C.V. wrote the paper.

DATA AVAILABILITY STATEMENT

The DNA sequences used in this study are available in Genbank (accession numbers: MG288511–MG288588). The main data set for *D. neotestacea* and *A. pisum* as well as the R function implementing the model are available on Dryad (10.5061/dryad.ch4dp8n).

ORCID

Hugo Mathé-Hubert https://orcid.org/0000-0003-4785-433X
Heidi Kaech https://orcid.org/0000-0002-2149-8050
Corinne Hertaeg https://orcid.org/0000-0002-2730-4042
Christoph Vorburger https://orcid.org/0000-0002-3627-0841

REFERENCES

Anbutsu, H., Lemaitre, B., Harumoto, T., & Fukatsu, T. (2016). Male-killing symbiont damages host's dosage-compensated sex chromosome to induce embryonic apoptosis. *Nature Communications*, *7*, 1–12.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, *94*, 443–458. https://doi.org/10.1093/biomet/asm017

Balding, D. J., Bishop, M., & Cannings, C. (2007). Population genetics. In D. J. Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of statistical genetics* (pp. 753–780). Chichester, UK: John Wiley & Sons, Ltd.

Ballinger, M. J., Moore, L. D., & Perlman, J. (2018). Evolution and diversity of inherited *Spiroplasma* symbionts in *Myrmica* ants. *Applied and Environmental Microbiology*, *84*, e02299–17.

Ballinger, M. J., & Perlman, S. J. (2017). Generality of toxins in defensive symbiosis: Ribosome-inactivating proteins and defense against parasitic wasps in *Drosophila*. *PLoS Path*, *13*, e1006431. https://doi.org/10.1371/journal.ppat.1006431

Barker, B. (2016). *Aphids in pulse crops*. Saskatchewan Pulse Growers; 8–10.

Brisson, J. A., Nuzhdin, S. V., & Stern, D. L. (2009). Similar patterns of linkage disequilibrium and nucleotide diversity in native and introduced populations of the pea aphid, *Acyrthosiphon pisum*. *BMC Genetics*, *10*, 22. https://doi.org/10.1186/1471-2156-10-22

Brucker, R. M., & Bordenstein, S. R. (2012). Speciation by symbiosis. *Trends in Ecology and Evolution*, *27*, 443–451. https://doi.org/10.1016/j.tree.2012.03.011

Caspi-Fluger, A., Inbar, M., Mozes-Daube, N., Katzir, N., Portnoy, V., Belausov, E., ... Zchori-Fein, E. (2012). Horizontal transmission of the insect symbiont *Rickettsia* is plant-mediated. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 1791–1796.

Chen, D., & Purcell, A. H. (1997). Occurrence and transmission of facultative endosymbionts in aphids. *Current Microbiology*, *34*, 220–225. https://doi.org/10.1007/s002849900172

Cockburn, S. N., Haselkorn, T. S., Hamilton, P. T., Landzberg, E., Jaenike, J., & Perlman, S. J. (2013). Dynamics of the continent-wide spread of a *Drosophila* defensive symbiont. *Ecology Letters*, *16*, 609–616.

Dedeine, F., Vavre, F., Fleury, F., Loppin, B., Hochberg, M. E., & Bouletreau, M. (2001). Removing symbiotic *Wolbachia* bacteria specifically inhibits oogenesis in a parasitic wasp. *Proceedings of the National Academy of Sciences of the USA*, *98*, 6247–6252.

Doremus, M. R., & Oliver, K. M. (2017). Aphid heritable symbiont exploits defensive mutualism. *Applied and Environmental Microbiology*, *83*, 1–15. https://doi.org/10.1128/AEM.03276-16

Duron, O., Bouchon, D., Boutin, S., Bellamy, L., Zhou, L., Engelstadter, J., & Hurst, G. D. (2008). The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. *BMC Biology*, *6*, 27. https://doi.org/10.1186/1741-7007-6-27

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*, 4th ed. Edinburgh, UK: Cambridge University Press.

Faria, V. G., Martins, N. E., Magalhães, S., Paulo, T. F., Nolte, V., Schlötterer, C., ... Teixeira, L. (2016). Drosophila adaptation to viral infection through defensive symbiont evolution. *PLOS Genetics*, *12*, e1006297. https://doi.org/10.1371/journal.pgen.1006297

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*, 737–756.

Ferrari, J., West, J. A., Via, S., & Godfray, H. C. J. (2012). Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. *Evolution*, *66*, 375–390. https://doi.org/10.1111/j.1558-5646.2011.01436.x

Fouedjio, F., & Klump, J. (2019). Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*, *78*, 1–24. https://doi.org/10.1007/s12665-018-8032-z

Frago, E., Mala, M., Weldegergis, B. T., Yang, C., McLean, A., Godfray, H. C. J., ... Dicke, M. (2017). Symbionts protect aphids from parasitic wasps by attenuating herbivore-induced plant volatiles. *Nature Communications*, *8*, 1–9. https://doi.org/10.1038/s41467-017-01935-0

Fromont, C., Adair, K. L., & Douglas, A. E. (2019). Correlation and causation between the microbiome, *Wolbachia* and host functional traits in natural populations of drosophilid flies. *Molecular Ecology*, *28*, 1826–1841.

Fukatsu, T., Tsuchida, T., Nikoh, N., & Koga, R. (2001). *Spiroplasma* symbiont of the pea aphid, *Acyrthosiphon pisum* (Insecta: Homoptera). *Applied and Environmental Microbiology*, *67*, 1284–1291. https://doi.org/10.1128/AEM.67.3.1284-1291.2001

Gehrer, L., & Vorburger, C. (2012). Parasitoids as vectors of facultative bacterial endosymbionts in aphids. *Biology Letters*, *8*, 613–615. https://doi.org/10.1098/rsbl.2012.0144

Goryacheva, I., Blekhman, A., Andrianov, B., Romanov, D., & Zakharov, I. (2018). *Spiroplasma* infection in *Harmonia axyridis* – Diversity and multiple infection. *PLoS ONE*, *13*, e0198190. https://doi.org/10.1371/journal.pone.0198190

Haag, C. R., McTaggart, S. J., Didier, A., Little, T. J., & Charlesworth, D. (2009). Nucleotide polymorphism and within-gene recombination in *Daphnia magna* and *D. pulex*, two cyclical parthenogens. *Genetics*, *182*, 313–323.

Hansen, A. K., Vorburger, C., & Moran, N. A. (2012). Genomic basis of endosymbiont-conferred protection against an insect parasitoid. *Genome Research*, *22*, 106–114. https://doi.org/10.1101/gr.125351.111

Hedges, L. M., Brownlie, J. C., O'Neill, S. L., & Johnson, K. N. (2008). *Wolbachia* and virus protection in insects. *Science*, *322*, 702. https://doi.org/10.1126/science.1162418

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518. https://doi.org/10.7717/peerj.5518

Henry, L. M., Maiden, M. C. J., Ferrari, J., & Godfray, H. C. J. (2015). Insect life history and the evolution of bacterial mutualism. *Ecology Letters*, *18*, 516–525. https://doi.org/10.1111/ele.12425

Henry, L. M., Peccoud, J., Simon, J. C., Hadfield, J. D., Maiden, M. J. C., Ferrari, J., & Godfray, H. C. J. (2013). Horizontally transmitted symbionts and host colonization of ecological niches. *Current Biology*, *23*, 1713–1717. https://doi.org/10.1016/j.cub.2013.07.029

Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A., & Werren, J. H. (2008). How many species are infected with *Wolbachia*? – A statistical analysis of current data. *FEMS Microbiology Letters*, *281*, 215–220.

Himler, A. G., Adachi-Hagimori, T., Bergen, J. E., Kozuch, A., Kelly, S. E., Tabashnik, B. E., ... Hunter, M. S. (2011). Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. *Science*, *332*, 254–256. https://doi.org/10.1126/science.1199410

Jabot, F., Faure, T., Dumoulin, N., & Albert, C. (2015). *EasyABC: Efficient approximate bayesian computation sampling schemes*. R package version 1.5.

Jaenike, J. (2012). Population genetics of beneficial heritable symbionts. *Trends in Ecology and Evolution*, *27*, 226–232. https://doi.org/10.1016/j.tree.2011.10.005

Jaenike, J., Polak, M., Fiskin, A., Helou, M., & Minhas, M. (2007). Interspecific transmission of endosymbiotic *Spiroplasma* by mites. *Biology Letters*, *3*, 23–25.

Jaenike, J., Stahlhut, J. K., Boelio, L. M., & Unckless, R. L. (2010). Association between *Wolbachia* and *Spiroplasma* within *Drosophila neotestacea*: An emerging symbiotic mutualism? *Molecular Ecology*, *19*, 414–425. https://doi.org/10.1111/j.1365-294X.2009.04448.x

Jaenike, J., Unckless, R., Cockburn, S. N., Boelio, L. M., & Perlman, S. J. (2010). Adaptation via symbiosis: Recent spread of a *Drosophila* defensive symbiont. *Science*, *329*, 212–215. https://doi.org/10.1126/science.1188235

Keightley, P. D., Ness, R. W., Halligan, D. L., & Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*, *196*, 313–320.

Kimura, M., & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, *61*, 763–771.

Kremer, N., Charif, D., Henri, H., Bataille, M., Prévost, G., Kraaijeveld, K., & Vavre, F. (2009). A new case of *Wolbachia* dependence in the genus *Asobara*: Evidence for parthenogenesis induction in *Asobara japonica*. *Heredity (Edinb)*, *103*, 248–256. https://doi.org/10.1038/hdy.2009.63

Lukasik, P., Guo, H., Van Asch, M., Ferrari, J., & Godfray, H. C. J. (2013). Protection against a fungal pathogen conferred by the aphid facultative endosymbionts *Rickettsia* and *Spiroplasma* is expressed in multiple host genotypes and species and is not influenced by co-infection with another symbiont. *Journal of Evolutionary Biology*, *26*, 2654–2661.

Lukasik, P., van Asch, M., Guo, H., Ferrari, J., & Godfray, H. C. J. (2013). Unrelated facultative endosymbionts protect aphids against a fungal pathogen. *Ecology Letters*, *16*, 214–218. https://doi.org/10.1111/ele.12031

Lynch, M., & Deng, H.-W. (1994). Genetic slippage in response to sex. *The American Naturalist*, *144*, 242–261. https://doi.org/10.1086/285673

Margulis, L., & Fester, R. (1991). *Symbiosis as a source of evolutionary innovation: Speciation and morphogenesis*, 1st ed. Amherst, MA: University of Massachusetts.

Mathé-Hubert, H., Kaech, H., Ganesanandamoorthy, P., & Vorburger, C. (2019). Evolutionary costs and benefits of infection with diverse strains of *Spiroplasma* in pea aphids. *Evolution*, *73*, 1–16.

Montllor, C. B., Maxmen, A., & Purcell, A. H. (2002). Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. *Ecological Entomology*, *27*, 189–195. https://doi.org/10.1046/j.1365-2311.2002.00393.x

Moore, W. S. (1995). Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, *49*, 718–726. https://doi.org/10.2307/2410325

Moran, N. A., & Dunbar, H. E. (2006). Sexual acquisition of beneficial symbionts in aphids. *Proceedings of the National Academy of Sciences of the USA*, *103*, 12803–12806.

Niepoth, N., Ellers, J., & Henry, L. M. (2018). Symbiont interactions with non-native hosts limit the formation of new symbioses. *BMC Evolutionary Biology*, *18*, 1–12. https://doi.org/10.1186/s12862-018-1143-z

Oliver, K. M., Degnan, P. H., Burke, G. R., & Moran, N. A. (2010). Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annual Review of Entomology*, *55*, 247–266. https://doi.org/10.1146/annurev-ento-112408-085305

Oliver, K. M., Degnan, P. H., Hunter, M. S., & Moran, N. A. (2009). Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science*, *325*, 992–994. https://doi.org/10.1126/science.1174463

Oliver, K. M., Moran, N. A., & Hunter, M. S. (2006). Costs and benefits of a superinfection of facultative symbionts in aphids. *Proceedings of the Royal Society B: Biological Sciences*, *273*, 1273–1280. https://doi.org/10.1098/rspb.2005.3436

Peccoud, J., Bonhomme, J., Mahéo, F., de la Huerta, M., Cosson, O., & Simon, J. C. (2014). Inheritance patterns of secondary symbionts during sexual reproduction of pea aphid biotypes. *Journal of Insect Science*, *21*, 291–300. https://doi.org/10.1111/1744-7917.12083

Peccoud, J., Ollivier, A., Plantegenest, M., & Simon, J.-C. (2009). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of Sciences of the USA*, *106*, 7495–7500.

Pieper, K. E., & Dyer, K. A. (2016). Occasional recombination of a selfish X-chromosome may permit its persistence at high frequencies in the wild. *Journal of Evolutionary Biology*, *29*, 2229–2241. https://doi.org/10.1111/jeb.12948

PLOS Biology Issue Image (2010). The pea aphid *Acyrthosiphon pisum*, an emerging genomic model. *PLOS Biology*, *8*(2), ev08.i02.

Queller, D. C., & Strassmann, J. E. (2016). Problems of multi-species organisms: Endosymbionts to holobionts. *Biology and Philosophy*, *31*, 855–873. https://doi.org/10.1007/s10539-016-9547-x

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Regassa, L. B. (2014). The family *Spiroplasmataceae*. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The prokaryotes: Firmicutes and tenericutes* (pp. 551–567). Berlin, Heidelberg, Germany: Springer.

Rice, J. A. (2006). *Mathematical statistics and data analysis*, 3rd ed. Berkeley, CA: University of California.

Rillig, M. C., Antonovics, J., Caruso, T., Lehmann, A., Powell, J. R., Veresoglou, S. D., & Verbruggen, E. (2015). Interchange of entire communities: Microbial community coalescence. *Trends in Ecology and Evolution*, *30*, 470–476. https://doi.org/10.1016/j.tree.2015.06.004

Rock, D. I., Smith, A. H., Joffe, J., Albertus, A., Wong, N., O'Connor, M., … Russell, J. A. (2017). Context-dependent vertical transmission shapes strong endosymbiont community structure in the pea aphid, *Acyrthosiphon pisum*. *Molecular Ecology*, *27*, 2039–2056.

Russell, J. A., & Moran, N. A. (2006). Costs and benefits of symbiont infection in aphids: Variation among symbionts and across temperatures. *Proceedings of the Royal Society B: Biological Sciences*, *273*, 603–610. https://doi.org/10.1098/rspb.2005.3348

Russell, J. A., Weldon, S., Smith, A. H., Kim, K. L., Hu, Y. I., Łukasik, P., … Oliver, K. M. (2013). Uncovering symbiont-driven genetic diversity across North American pea aphids. *Molecular Ecology*, *22*, 2045–2059. https://doi.org/10.1111/mec.12211

Sanada-Morimura, S., Matsumura, M., & Noda, H. (2013). Male killing caused by a *Spiroplasma* symbiont in the small brown planthopper, *Laodelphax striatellus*. *Journal of Heredity*, *104*, 821–829. https://doi.org/10.1093/jhered/est052

Simon, J.-C., Boutin, S., Tsuchida, T., Koga, R., Le Gallic, J.-F., Frantz, A., … Fukatsu, T. (2011). Facultative symbiont infections affect aphid reproduction. *PLoS ONE*, *6*(7), e21831. https://doi.org/10.1371/journal.pone.0021831

Smith, A. H., Łukasik, P., O'Connor, M. P., Lee, A., Mayo, G., Drott, M. T., … Russell, J. A. (2015). Patterns, causes, and consequences of defensive microbiome dynamics across multiple scales. *Molecular Ecology*, *24*, 1135–1149. https://doi.org/10.1111/mec.13095

Sunnucks, P., & Hales, D. F. (1996). Numerous transposed sequences of mitochondrial cytochrome I-II in aphids of the genus *Sitobion* (*Hemiptera*: *Aphididae*). *Molecular Biology and Evolution*, *13*, 510–524.

Tabata, J., Hattori, Y., Sakamoto, H., Yukuhiro, F., Fujii, T., Kugimiya, S., … Kageyama, D. (2011). Male killing and incomplete inheritance of a novel *Spiroplasma* in the moth *Ostrinia zaguliaevi*. *Microbial Ecology*, *61*, 254–263. https://doi.org/10.1007/s00248-010-9799-y

Teixeira, L., Ferreira, A., & Ashburner, M. (2008). The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. *PLOS Biology*, *6*, e1000002. https://doi.org/10.1371/journal.pbio.1000002

Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2014). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, *72*, 191–199. https://doi.org/10.1016/j.jmp.2014.10.001

Via, S., & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, *17*, 4334–4345. https://doi.org/10.1111/j.1365-294X.2008.03921.x

Vorburger, C. (2006). Temporal dynamics of genotypic diversity reveal strong clonal selection in the aphid *Myzus persicae*. *European Society for Evolutionary Biology*, *19*, 97–107. https://doi.org/10.1111/j.1420-9101.2005.00985.x

Vorburger, C. (2018). Symbiont-conferred resistance to parasitoids in aphids – Challenges for biological control. *Biological Control*, *116*, 17–26. https://doi.org/10.1016/j.biocontrol.2017.02.004

Vorburger, C., & Gouskov, A. (2011). Only helpful when required: A longevity cost of harbouring defensive symbionts. *Journal of Evolutionary Biology*, *24*, 1611–1617. https://doi.org/10.1111/j.1420-9101.2011.02292.x

Weldon, S. R., Strand, M. R., & Oliver, K. M. (2013). Phage loss and the breakdown of a defensive symbiosis in aphids. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 2012–2103. https://doi.org/10.1098/rspb.2012.2103

Werner, T., & Jaenike, J. (2017). *Drosophilids of the Midwest and Northeast*, 1st ed. Rochester, NY: River Campus Libraries, University of Rochester.

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia*: Master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*, 741–751. https://doi.org/10.1038/nrmicro1969

Wulff, J. A., Buckman, K. A., Wu, K., Heimpel, G. E., & White, J. A. (2013). The endosymbiont *Arsenophonus* is widespread in soybean aphid, *Aphis glycines*, but does not provide protection from parasitoids or a fungal pathogen. *PLoS ONE*, *8*, e62145. https://doi.org/10.1371/journal.pone.0062145

Wulff, J. A., & White, J. A. (2015). The endosymbiont *Arsenophonus* provides a general benefit to soybean aphid (*Hemiptera*: *Aphididae*) regardless of host plant resistance (Rag). *Environmental Entomology*, *44*, 574–581. https://doi.org/10.1093/ee/nvv031

Xie, J., Butler, S., Sanchez, G., & Mateos, M. (2014). Male killing *Spiroplasma* protects *Drosophila melanogaster* against two parasitoid wasps. *Heredity (Edinb)*, *112*, 399–408. https://doi.org/10.1038/hdy.2013.118

Zchori-Fein, E., Lahav, T., & Freilich, S. (2014). Variations in the identity and complexity of endosymbiont combinations in whitefly hosts. *Frontiers in Microbiology*, *5*, 1–8. https://doi.org/10.3389/fmicb.2014.00310

Zchori-Fein, E., & Perlman, S. J. (2004). Distribution of the bacterial symbiont *Cardinium* in arthropods. *Molecular Ecology*, *13*, 2009–2016. https://doi.org/10.1111/j.1365-294X.2004.02203.x

Zytynska, S. E., & Weisser, W. W. (2016). The natural occurrence of secondary bacterial symbionts in aphids. *Ecological Entomology*, *41*, 13–26. https://doi.org/10.1111/een.12281

SUPPORTING INFORMATION

Additional supporting information may be found on next pages

Supplemental Information for:

Non-random associations of maternally transmitted symbionts in insects: The roles of drift versus biased co-transmission and selection

Mathé-Hubert Hugo, Heidi Kaech, Corinne Hertaeg, John Jaenike, Christoph Vorburger

Table of Contents:

Supplementary material S1: Phylogeny of Spiroplasma in pea aphids

The analysis of the distribution of the intraspecific diversity of *Spiroplasma* included the 26 strains found in our previous field sampling mentioned in the main text as well as 11 strains that were kindly provided by Ailsa McLean (Department of Zoology, University of Oxford, UK; Table S2). The diversity of Spiroplasma was characterised using a phylogeny based on the dnaA gene and the rpoB gene, including its surrounding regions (445 bp and 2731 bp, respectively, including primers; Table S1). The PCR cycling conditions are as described by Henry *et al.* (2013), except that the elongation time of the primer pair RpoBF1.ixod was doubled (Table S1). We deposited all dnaA and rpoB sequences in Genbank (accession numbers: MG288511 to MG288588).

We inferred the phylogenetic tree of the 37 *Spiroplasma* strains (Fig. S2) using *Spiroplasma sp.* in *Ostrinia zaguliaevi* as outgroup, which is the most closely related species to *Spiroplasma* of the pea aphid for which we could obtain sequences for both the rpoB and dnaA genes. The substitution model "GTR + gamma + invariant sites" was identified by AICc ("phangorn" R package v 2.2.0, Schliep 2011) as the best-fitting for the MAFFT-aligned and then concatenated dnaA and rpoB sequences. It was used to build a maximum likelihood tree with the software MEGA6 (Tamura et al. 2013).

Supplementary material S2: Detecting positive and negative associations among pea aphid symbionts with a random forest approach

To detect pair of symbionts that were co-occuring more or less frequently than expected under random assortment, we fitted the presence of each symbiont species with a random forest (RF) model. In order to handle the multicollinearity as accurately as possible, RFs were grown with conditional inference trees and the effect of variables was estimated with conditional importance (package *party*; Hothorn *et al*. 2006). Our methodology mostly follows the advice of Jones and Linder (2015). The only difference is that we applied the permutation approach developed by Hapfelmeier & Ulm (2013) to the conditional importance of explanatory variables to estimate their *p*-values. These were then adjusted to keep the false discovery rate at 5% (Benjamini and Yekutieli 2001). With this approach, there is one model per symbiont, the presence or absence of the focal symbiont being explained by the presence or absence of the six other symbionts. Thus, each pair of symbiont (*a* and *b*) is considered by two models, one explaining the presence of the symbiont *a* and the other explaining the presence of the symbiont *b*. This approach thus leads to two *p*-values per couple of symbionts.

To investigate the intraspecific distribution of *Spiroplasma*, we used a classification RF to predict the phylogenetic clade of *Spiroplasma* for each *Spiroplasma* infected aphid. The *Spiroplasma* strain S362 was excluded from the analysis because it was not assignable to one of the three phylogenetic clades. The explanatory variables were the same as in the previous analysis except that the aphid colour was not available for all aphids and was thus not used. For significant variables, we further compared clades to each other in a pairwise fashion to characterise among-clade variance. Since some of these significant variables are continuous and some are categorical, we used Wilcoxon or Fisher's exact tests, respectively to perform these pairwise comparisons.

Supplementary material S3: Analysis of the link between the frequency of symbiont species and the average number of additional symbiont species with which they co-occur.

The random forest analysis detected that some symbionts are less frequent in aphids already containing other symbiont species, while others are not significantly affected by the presence of other symbionts. To further investigate this, we tested if there was a link between the frequency of each symbiont species and the average number of additional symbiont species with which it co-occurs. We characterised this link with a linear model where the average number of additional symbiont species is explained by the frequency of other symbionts. Under random assortment between symbionts, the slope of this model is not expected to be flat (0) because the rarest symbiont species can only co-occur with species that are more abundant than itself, while on the opposite, the most abundant species can only co-occur with species that are less abundant than itself. Thus, under random assortment rare species should on average co-occur with more other species than common symbionts would. In consequence, the common ones should be more often found alone or with only a few other species.

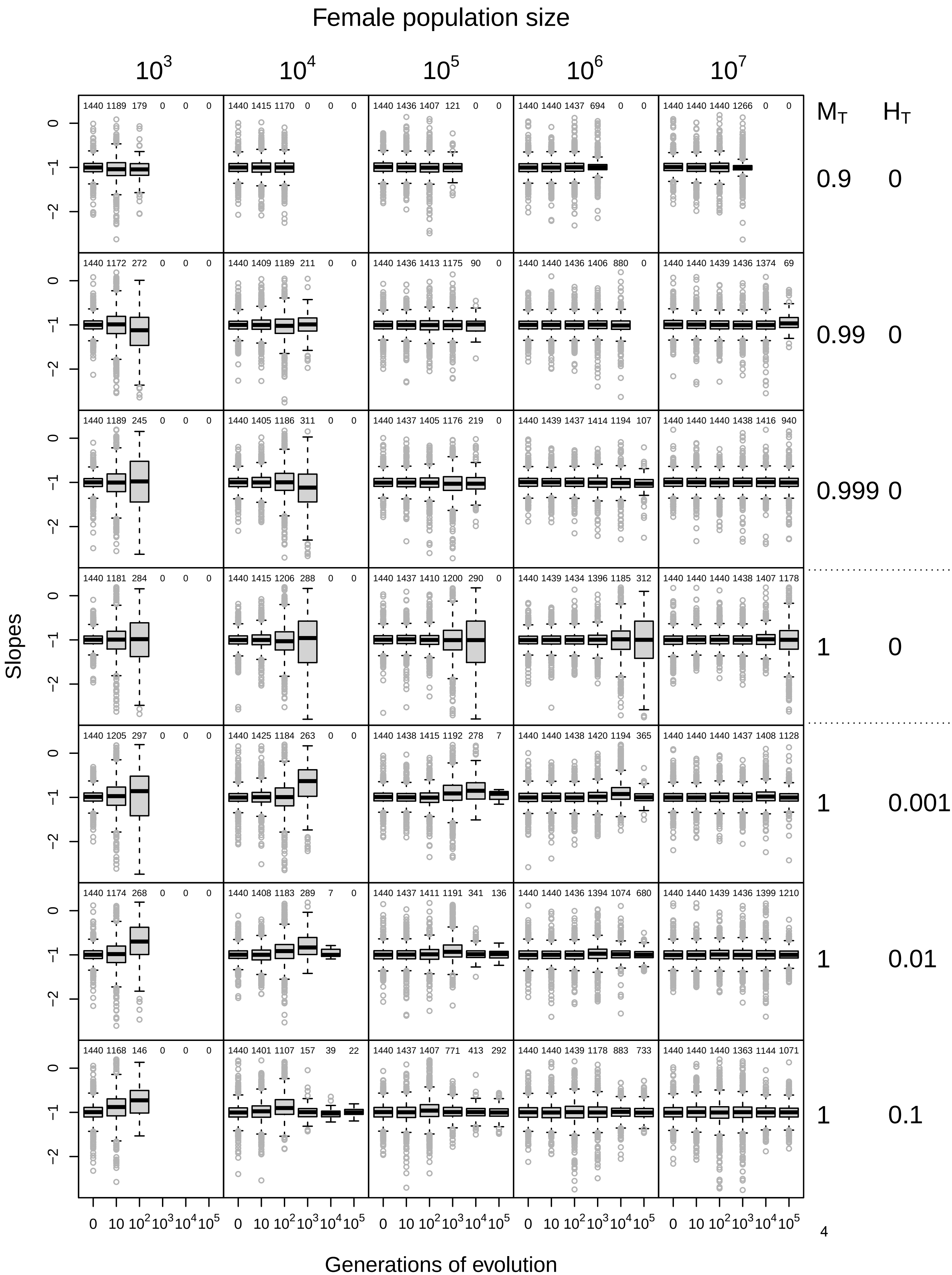
Formally, for any symbiont species x of a set of K symbionts, the average expected number of other species with which x co-occurs is the average number of symbionts the hosts contain other than x (noted hereafter $\overline{N_{coinf.\backslash x}}$). It is simply the sum of the frequencies of symbionts other than x: $\overline{N_{coinf.\backslash x}} = \sum_{y=1\,;\,y\neq x}^{K} F_y$, where F_y is the frequency of the symbiont y.

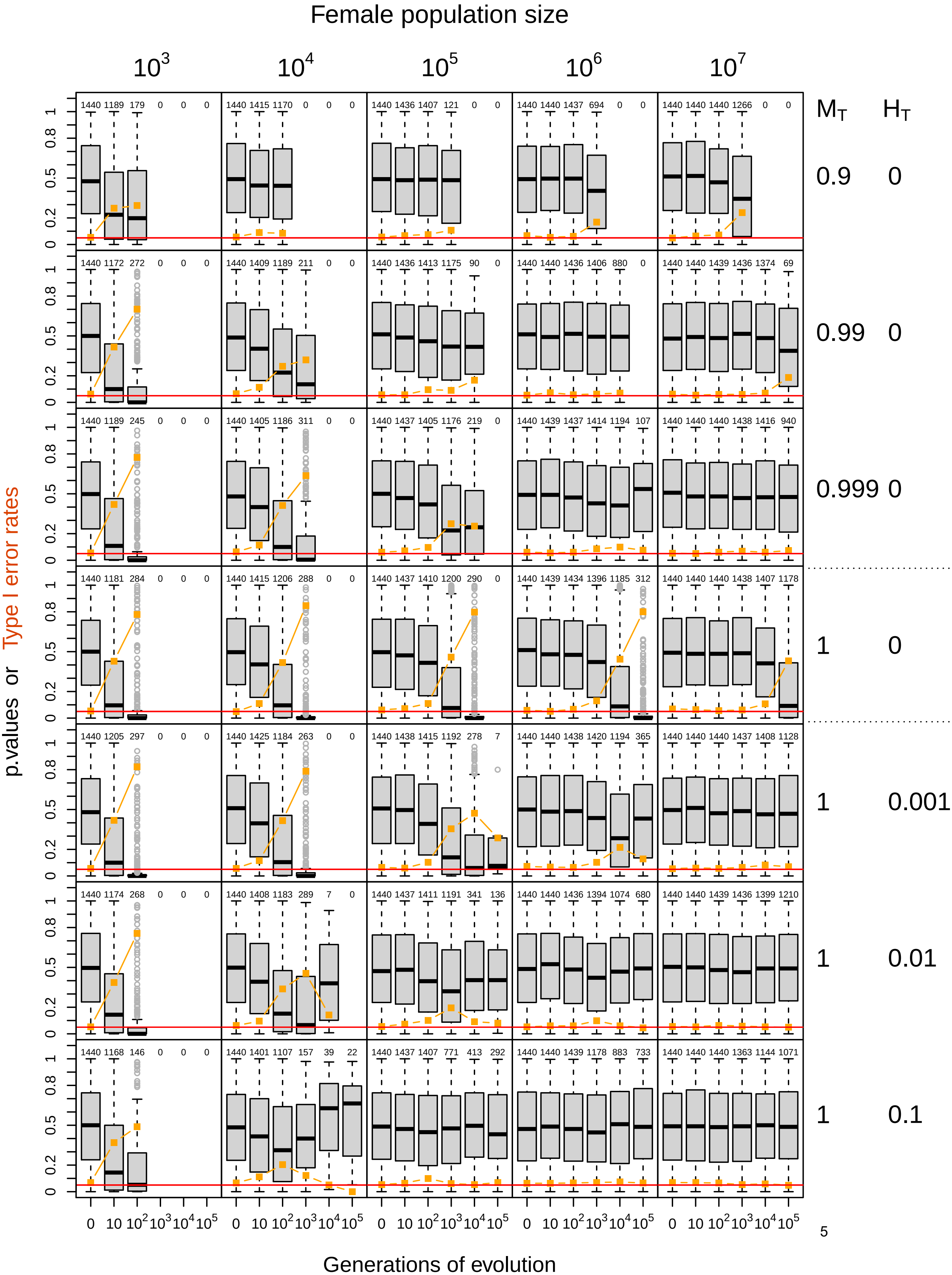
The linear model thus have the following form: $\overline{N_{coinf.\backslash x}} = a \times F_x + b + e;\ e{\sim}N(0;\sigma_e)$ which can also be written as $\sum_{y=1\,;\,y\neq x}^{K} F_y = a \times F_x + b + e;\ e{\sim}N(0;\sigma_e)$. The coefficients can be estimated, for example, from the two extreme values, $F_x = 0$ and $F_x = 1$ for which $\sum_{y=1\,;\,y\neq x}^{K} F_y$ is equal to $\sum_{y=1}^{K} F_y$ and to $\left(\sum_{y=1}^{K} F_y\right) - 1$, respectively. This gives $b = \sum_{y=1}^{K} F_y$ and $a = \left(\sum_{y=1}^{K} F_y\right) - 1 - b = -1$.

However, F_y and F_x are partly estimated from the same individuals when they have both symbionts x and y. Thus even the noise in the explained and explanatory variables are not independent, meaning that classical linear model tests cannot be used. Instead we estimated the *p*-value by bootstrapping the observed sample of the population and refitting the linear model on each bootstrap sample. This was used to test if the observed slop is equal to –1. This analysis was performed separately for aphids sampled from *Medicago sativa* and *Trifolium spp*. Because for *Trifolium spp*. the trend was mostly driven by *Regiella insecticola*, we also repeated the analysis without *R. insecticola* infected aphids.

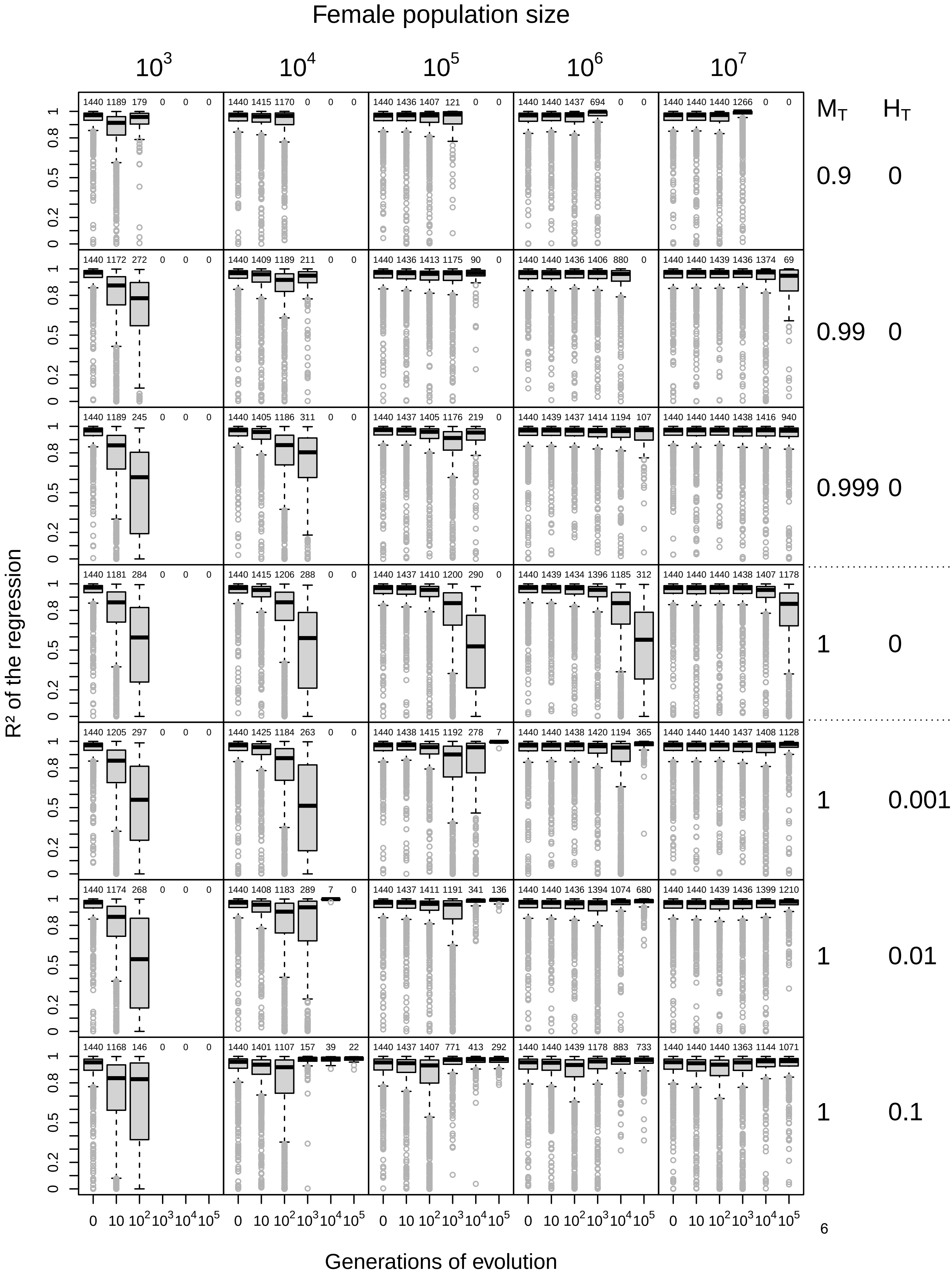
We additionally investigated the effect of drift on this test by simulating the evolution of 3000 replicate populations for 10^5 generations. Individuals of these populations where infected by up to 7 symbiont species, meaning that the population is characterised by the frequencies of 128 (2^7) kinds of symbiont communities instead of four when only two symbionts are considered. These simulations were otherwise identical to the simulations described in the main text, and we used the same combinations of parameters. We computed the same test as described above at generations 0, 10, 10^2, 10^3, 10^4, and 10^5.

The result of this simulation is shown on the four following pages. The first page shows the effect of drift, horizontal transmissions and maternal transmission failures on the slope of the linear model explaining the mean number of other symbionts ($N_{coinf.\backslash x}$) by the frequency of the symbiont (F_x). The *p*-values of the tests described above and the corresponding type 1 error rate are shown on the second page. The third page gives the proportion of variance explained by the models (R^2), and the fourth the number of different symbiont communities present in the population. The numbers above the boxplots indicate the proportion of populations that still retained some polymorphism of infection by the seven symbionts.

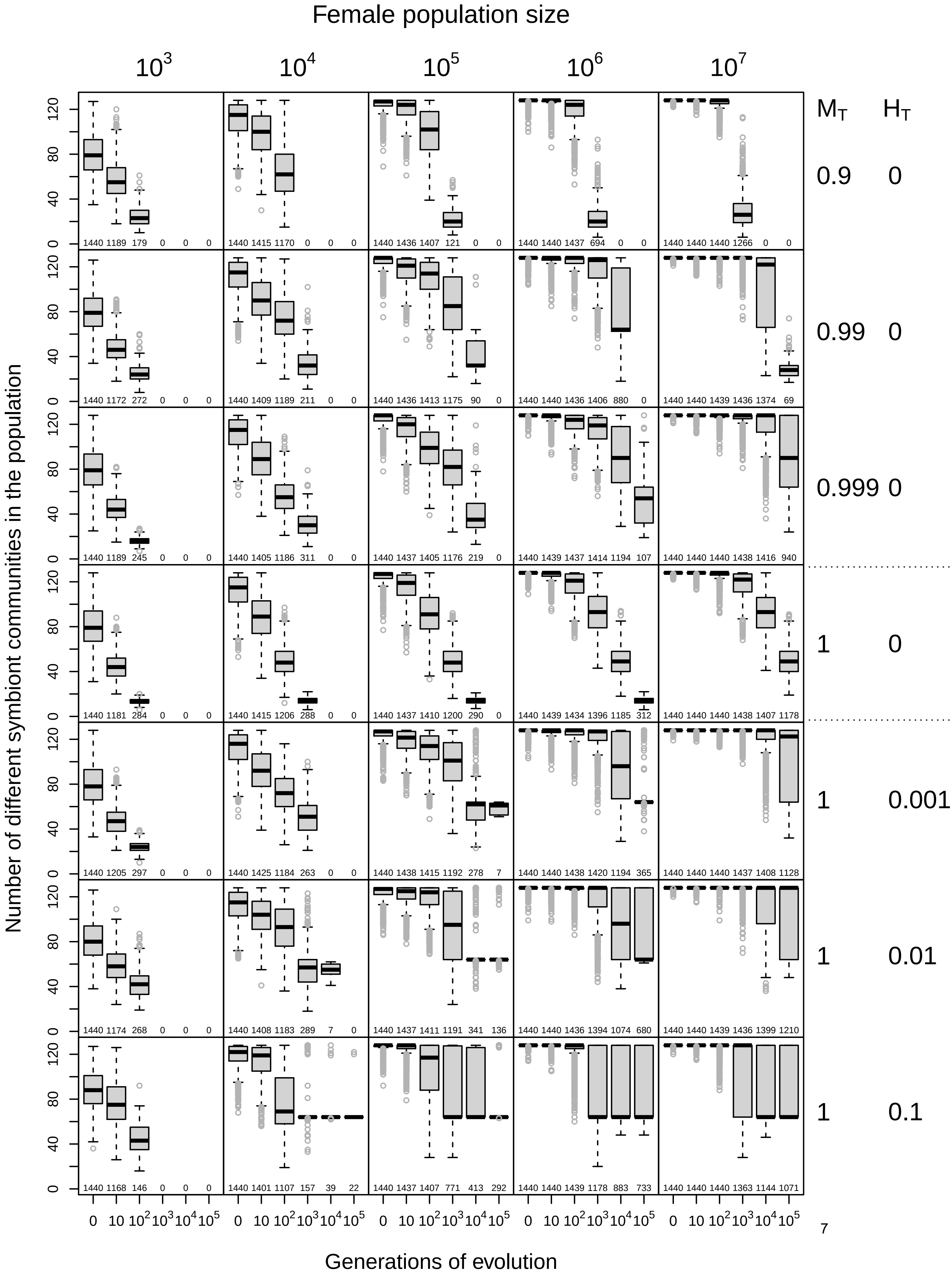

Female population size
10^3
10^4
10^5
10^6
10^7
M_T
H_T
1440 1189 179 0 0 0
1440 1415 1170 0 0 0
1440 1436 1407 121 0 0
1440 1440 1437 694 0 0
1440 1440 1440 1266 0 0
0.9
0
1440 1172 272 0 0 0
1440 1409 1189 211 0 0
1440 1436 1413 1175 90 0
1440 1440 1436 1406 880 0
1440 1440 1439 1436 1374 69
0.99
0
1440 1189 245 0 0 0
1440 1405 1186 311 0 0
1440 1437 1405 1176 219 0
1440 1439 1437 1414 1194 107
1440 1440 1440 1438 1416 940
0.999
0
1440 1181 284 0 0 0
1440 1415 1206 288 0 0
1440 1437 1410 1200 290 0
1440 1439 1434 1396 1185 312
1440 1440 1440 1438 1407 1178
1
0
1440 1205 297 0 0 0
1440 1425 1184 263 0 0
1440 1438 1415 1192 278 7
1440 1440 1438 1420 1194 365
1440 1440 1440 1437 1408 1128
1
0.001
1440 1174 268 0 0 0
1440 1408 1183 289 7 0
1440 1437 1411 1191 341 136
1440 1440 1436 1394 1074 680
1440 1440 1439 1436 1399 1210
1
0.01
1440 1168 146 0 0 0
1440 1401 1107 157 39 22
1440 1437 1407 771 413 292
1440 1440 1439 1178 883 733
1440 1440 1440 1363 1144 1071
1
0.1
Slopes
0
−1
−2
0 10 10^2 10^3 10^4 10^5
Generations of evolution



Female population size
10^3 10^4 10^5 10^6 10^7
M_T H_T
0.9 0
0.99 0
0.999 0
1 0
1 0.001
1 0.01
1 0.1
p.values or Type I error rates
0 0.2 0.5 0.8 1
0 10 10^2 10^3 10^4 10^5
Generations of evolution
1440 1189 179 0 0 0 | 1440 1415 1170 0 0 0 | 1440 1436 1407 121 0 0 | 1440 1440 1437 694 0 0 | 1440 1440 1440 1266 0 0
1440 1172 272 0 0 0 | 1440 1409 1189 211 0 0 | 1440 1436 1413 1175 90 0 | 1440 1440 1436 1406 880 0 | 1440 1440 1439 1436 1374 69
1440 1189 245 0 0 0 | 1440 1405 1186 311 0 0 | 1440 1437 1405 1176 219 0 | 1440 1439 1437 1414 1194 107 | 1440 1440 1440 1438 1416 940
1440 1181 284 0 0 0 | 1440 1415 1206 288 0 0 | 1440 1437 1410 1200 290 0 | 1440 1439 1434 1396 1185 312 | 1440 1440 1440 1438 1407 1178
1440 1205 297 0 0 0 | 1440 1425 1184 263 0 0 | 1440 1438 1415 1192 278 7 | 1440 1440 1438 1420 1194 365 | 1440 1440 1440 1437 1408 1128
1440 1174 268 0 0 0 | 1440 1408 1183 289 7 0 | 1440 1437 1411 1191 341 136 | 1440 1440 1436 1394 1074 680 | 1440 1440 1439 1436 1399 1210
1440 1168 146 0 0 0 | 1440 1401 1107 157 39 22 | 1440 1437 1407 771 413 292 | 1440 1440 1439 1178 883 733 | 1440 1440 1440 1363 1144 1071



Female population size
10^3 10^4 10^5 10^6 10^7
M_T H_T
0.9 0
0.99 0
0.999 0
1 0
1 0.001
1 0.01
1 0.1
1440 1189 179 0 0 0 | 1440 1415 1170 0 0 0 | 1440 1436 1407 121 0 0 | 1440 1440 1437 694 0 0 | 1440 1440 1440 1266 0 0
1440 1172 272 0 0 0 | 1440 1409 1189 211 0 0 | 1440 1436 1413 1175 90 0 | 1440 1440 1436 1406 880 0 | 1440 1440 1439 1436 1374 69
1440 1189 245 0 0 0 | 1440 1405 1186 311 0 0 | 1440 1437 1405 1176 219 0 | 1440 1439 1437 1414 1194 107 | 1440 1440 1440 1438 1416 940
1440 1181 284 0 0 0 | 1440 1415 1206 288 0 0 | 1440 1437 1410 1200 290 0 | 1440 1439 1434 1396 1185 312 | 1440 1440 1440 1438 1407 1178
1440 1205 297 0 0 0 | 1440 1425 1184 263 0 0 | 1440 1438 1415 1192 278 7 | 1440 1440 1438 1420 1194 365 | 1440 1440 1440 1437 1408 1128
1440 1174 268 0 0 0 | 1440 1408 1183 289 7 0 | 1440 1437 1411 1191 341 136 | 1440 1440 1436 1394 1074 680 | 1440 1440 1439 1436 1399 1210
1440 1168 146 0 0 0 | 1440 1401 1107 157 39 22 | 1440 1437 1407 771 413 292 | 1440 1440 1439 1178 883 733 | 1440 1440 1440 1363 1144 1071
0 0.2 0.5 0.8 1
R^2 of the regression
0 10 10^2 10^3 10^4 10^5
Generations of evolution
6



Female population size
10^3 10^4 10^5 10^6 10^7
M_T H_T
0.9 0
0.99 0
0.999 0
1 0
1 0.001
1 0.01
1 0.1
Number of different symbiont communities in the population
0 40 80 120
1440 1189 179 0 0 0 1440 1415 1170 0 0 0 1440 1436 1407 121 0 0 1440 1440 1437 694 0 0 1440 1440 1440 1266 0 0
1440 1172 272 0 0 0 1440 1409 1189 211 0 0 1440 1436 1413 1175 90 0 1440 1440 1436 1406 880 0 1440 1440 1439 1436 1374 69
1440 1189 245 0 0 0 1440 1405 1186 311 0 0 1440 1437 1405 1176 219 0 1440 1439 1437 1414 1194 107 1440 1440 1440 1438 1416 940
1440 1181 284 0 0 0 1440 1415 1206 288 0 0 1440 1437 1410 1200 290 0 1440 1439 1434 1396 1185 312 1440 1440 1440 1438 1407 1178
1440 1205 297 0 0 0 1440 1425 1184 263 0 0 1440 1438 1415 1192 278 7 1440 1440 1438 1420 1194 365 1440 1440 1440 1437 1408 1128
1440 1174 268 0 0 0 1440 1408 1183 289 7 0 1440 1437 1411 1191 341 136 1440 1440 1436 1394 1074 680 1440 1440 1439 1436 1399 1210
1440 1168 146 0 0 0 1440 1401 1107 157 39 22 1440 1437 1407 771 413 292 1440 1440 1439 1178 883 733 1440 1440 1440 1363 1144 1071
0 10 10^2 10^3 10^4 10^5
Generations of evolution

These simulations showed that the test for the slope describing the relationship between the frequencies of each symbiont species with the average number of other symbiont species with which they co-occur is also highly affected by drift. Drift strongly increases the variation the slopes around the expected value of -1 which increases type 1 error rate. This effect is associated with a decrease of R^2. However, this effect is moderate in presence of horizontal transmission or maternal transmission failures. Finally, it appeared that horizontal transmissions tended to erase the relationship expected under random assortment by increasing the slope. This unexpected effect might be due to that the number of horizontal transmissions of a symbiont increases with its frequency which counter-balance the fact that while it is more frequent, it tends to co-occurs more often with itself than with another species.

Supplementary material S4: Modelling the evolution of the frequency of symbionts communities

Model parameters

Ne_F: Size of the population of females being simulated

$Nsymbionts$: Number of symbiont species simulated. Notice that the computational cost of the model is approximately proportional to $2^{Nsymbionts}$.

M_T: Rates of successful maternal transmission (one rate per symbiont species)

H_T: Rates of horizontal transmission (one rate per symbiont species). For each infected host and symbiont species it contains, the symbiont is transmitted on average H_T times. Horizontal transmission happens indifferently toward infected and uninfected individuals. When an individual receives a symbiont it already contains, this does not have any consequences.

W: Fitnesses induced by each kind of symbionts community; for each offspring produced, for any female, the probability to be the mother of that offspring will be proportional to this number (W).

$\overline{w}$: Average fitness accounting for the frequency of each community

f: Frequency of the symbionts.

Description of the model

We are modelling a population of females of constant size. Each female can contain 0 to $Nsymbionts$ symbiont species. The dynamic of the symbiont communities depends on M_T, H_T and W. In the current implementation of the model, M_T and H_T have the same value for all symbionts and there are no interactions between symbionts. This is because the purpose of the model was simply to show that deviation from the random assortment hypothesis happens by drift, without any particular interaction between symbionts. However, the model can easily be modified to implement symbiont interactions or symbiont specific values for M_T and H_T.

Structure of the population in R and package dependency

In R, the population (object **pop**) is a named vector containing the number of females infected by each kind of symbiont community. The names of this vector indicate the kind of symbiont community by a series of zeros and ones. The 1st character indicates the presence (1) or the absence (0) of symbiont 1, the 2nd character indicates the presence (1) or the absence (0) of symbiont 2, and so on.

The model relies on the function **permutations** from the package **gtools**, **rmultinom** from the package **stats** and **rmvhyper** from the package **extraDistr**. The function **permutations** is just used once to create or check the **pop** object. The function **rmultinom** is used to simulate sampling in the populations with replacement (multinomial distribution) and the function **rmvhyper** for sampling without replacement (multivariate hypergeometric distribution).

Sampling with replacement happens at each replicate, to instantiate populations and to choose the mothers of the offspring population. Sampling without replacement happens when the individuals receiving the symbiont from horizontal transmission are chosen and when the offspring that don't receive a symbiont from the mother (maternal transmission failure) are randomly chosen. This function can also be used to simulate random samples from the population.

Horizontal transmission: choosing the number of recipients

For the horizontal transmission, we need to account for the fact that for any given generation, each individual can receive a symbiont several times, which has the same effect as receiving it only one time.

For a given symbiont species of frequency $\boldsymbol{f}$, in a population of size Ne_F, there will be on average $f \times Ne_F \times H_T$ horizontal transmission events in each generation. In each of these transmission events, every individual has a probability of $\frac{1}{Ne_F}$ to be the recipient. Thus, the number of horizontal transmission events undergone by each individual follows a binomial distribution of parameters $n = f \times Ne_F \times H_T$ and $p = \frac{1}{Ne_F}$. The probability of receiving the symbiont once or several times is thus

$$1 - P\{Never\ receiving\ the\ symbiont\}$$

$$= 1 - P\{Binomial(n, p) = 0\}$$

$$= 1 - \binom{n}{0} \times p^0 \times (1-p)^{n-0}$$

$$= 1 - \frac{n\,!}{0\,! \times (n-0)\,!} \times 1 \times (1-p)^n$$

$$= 1 - \frac{n\,!}{1 \times (n)\,!} \times (1-p)^n$$

$$= 1 - 1 \times (1-p)^n$$

$$= 1 - (1-p)^n$$

Thus, given that $n = f \times Ne_F \times H_T$ and $p = \frac{1}{Ne_F}$, in each generation, the average proportion of individuals receiving a symbiont at least once is $1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}$. This means that for each symbiont species, the number of individuals receiving the symbiont follows a new binomial distribution of parameters $n = N_e F$ and $p = 1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}$. Some of the individuals will already have the symbiont from their mother and the horizontal transmission event will not change their symbiont community, while, for the previously uninfected individuals, the transmission event will change the category of the symbiont community to which they belong.

Detailed model workflow

For each generation:

1) Horizontal transmission: For each symbiont (x)
   a. Choose the number of horizontal transmission event (Nht) in a Poisson distribution of mean $H_T \times$ number of individuals infected by x.
   b. Choose the number of recipient individuals ($Nrecipients$) in a binomial distribution of parameters $n = Ne_F$ and $p = 1 - \left(1 - \frac{1}{Ne_F}\right)^{Nht}$. Choose the $Nrecipients$ recipient individuals in a hypergeometric distribution.
   c. Modify the number of individuals of each kind of community of symbionts according to b.
2) Reproduction: create the offspring population according to the number of individuals of each kind of community and their relative fitness.
3) Maternal transmission failure:

a. For each symbiont (x), choose the number of maternal transmission failure events in a Poisson distribution of mean $(1 - M_T) \times$ number of offspring infected by x.

b. Modify the number of individuals of each kind of community of symbionts according to a.

4) Assessment of the deviation from random assortment

R code used for the simulations

A R function corresponding to this model is provided along with this supplementary material as well as a R script illustrating the use of this function. These are in the appendix S1.

Supplementary material S5: Setting the selection parameter to maintain polymorphism

When using this model (Supplementary material S4) to perform simulations without any selection for most of the combinations of parameters of M_T and H_T, the polymorphism is rapidly lost. This prevents assessing whether the frequency of co-occurrences deviates from random assortment.

We therefore use the selection (parameter W) to maintain polymorphism. For each replicate of every combination of parameters M_T and H_T, and for each symbiont separately, we randomly pick the desired expected equilibrium frequency (f^*) in a uniform distribution [0.01; 0.99] (this f^* is also the initial frequency of the symbionts). This is done separately for every symbiont. Then we use the deterministic model described hereafter to calculate the value of the W parameter that, given the values of M_T and H_T, is expected to lead to the equilibrium frequency f^*. In this model, we only consider one symbiont. The fitness of aposymbiotic individuals ($w_ø$) is used a reference and is set to 1.

Description of the deterministic model

f' is the frequency of a symbiont at the next generation, and f its frequency in the current generation.

$f' =$ **frequency induced by selection** + **horizontal transmissions** – **failed vertical transmission**

Recall that the probability of receiving a given symbiont species by horizontal transmission at least once is $1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}$ (see § “Horizontal transmission: getting the number of recipients”). This gives us:

$$f' = \frac{f \times w_S}{\overline{w}} + \left(1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}\right) \times (1 - f) - f \times (1 - M_T)$$

$$f' = f \times \left(\frac{w_S}{\overline{w}} + \left(1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}\right) \times \frac{1 - f}{f} - (1 - M_T)\right)$$

$$f' = f \times \left(\frac{w_S}{(1 - f) \times w_ø + f \times w_S} + \left(1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}\right) \times \frac{1 - f}{f} - 1 + M_T\right)$$

$$f' = f \times \left(\frac{w_S}{(1 - f) \times w_ø + f \times w_S} + \left(1 - \left(1 - \frac{1}{Ne_F}\right)^{f \times Ne_F \times H_T}\right) \times \frac{1 - f}{f} - 1 + M_T\right)$$

$$f' = f \times \left(\frac{1}{(1-f) \times \frac{w_{ø}}{w_{S}} + f \times 1} + \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} - 1 + M_T \right)$$

$$f' = f \times \left(\frac{1}{(1-f) \times \frac{1}{w_{S}} + f \times 1} + \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} - 1 + M_T \right)$$

$$f' = f \times \left(\frac{1}{\frac{1-f}{w_{S}} + f} + \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} - 1 + M_T \right)$$

We want to find combinations of parameters H_T, V_T and s for which we will have $f' = f = f^*$.

Thus, we will have $f' = f$ if $f = 0$ or if

$$\frac{1}{\frac{1-f}{w_{S}} + f} + \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} - 1 + V_T = 1$$

$$\frac{1}{\frac{1-f}{w_{S}} + f} + \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} - 1 + V_T = 1$$

$$\Leftrightarrow \frac{1}{\frac{1-f}{w_{S}} + f} = 1 - \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} + 1 - V_T$$

$$\Leftrightarrow \frac{1-f}{w_{S}} + f = \frac{1}{1 - \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} + 1 - V_T}$$

$$\Leftrightarrow \frac{1-f}{w_{S}} = \frac{1}{1 - \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} + 1 - V_T} - f$$

$$w_{S} = \frac{1-f}{\frac{1}{1 - \left(1 - \left(1 - \frac{1}{Ne_F} \right)^{f \times Ne_F \times H_T} \right) \times \frac{1-f}{f} + 1 - V_T} - f}$$

This formula is used to choose the fitness of individuals infected by only one symbiont species. The fitness of individuals infected by several symbiont species, is the product of the fitness induced by each symbiont species it has. This multiplicative fitness effect is similar to the model of Jaenike et al. (2010).

We see that the value of w_{S}, which is expected to maintain a constant symbiont frequency, depends on the current symbiont frequency f. This means that as drift induces random variation of f and W will not be at its optimal value anymore. This will be particularly problematic in cases of high rates of horizontal transmission (H_T) and low rates of maternal transmission failure (high M_T). Indeed, horizontal transmissions and maternal transmission failures regulate each other because both the probability of receiving a symbiont and the number of maternal transmission failures are proportional to the symbiont frequency. Additionally, in the absence of horizontal transmissions ($H_T = 0$), maternal transmission failures push the symbiont frequency toward zero. However, the number of these events also decreases with symbiont frequency, meaning that this is a slowing down process that can be counteracted by selection. On the opposite, in the absence of maternal transmission failures ($M_T = 1$) horizontal transmission events increase symbiont frequency which leads to fixation, but the number of these events also increase with symbiont frequency, meaning that this is an accelerating process that can hardly be counteracted by selection. Indeed, at a given symbiont frequency, horizontal transmissions represent a given evolutionary strength that can be offset by a given strength of selection. However, drift will induce variation of the symbiont frequencies which makes the selection either too strong or too weak to maintain stable symbiont frequencies. Indeed, horizontal transmissions tend to amplify the effect of drift, because increasing or decreasing f increases and decreases the number of horizontal transmission events, respectively. This is the opposite for maternal transmission failures.

Numeric application (checking for calculus error)

```
update_f = function(f,Ws,H,V,N){  return(
  f*(1/((1-f)/Ws+f)+(1-(1-1/(N))^(f*N*H ) )*(1-f)/f-1+V )
)}

N_ = S = MT_ = HT_ = obs = exp = NULL
for(l in 1:1000){
  N = runif(1,100,10^7)
  MT=V=runif(1)
  HT=H=runif(1)
  f = f_ = runif(1)
  Ws = (1-f)/(1/(1-(1-(1-1/(N))^(f*N*H))*(1-f)/f+1-V )-f)

  F = f ; for(g in 1:100){F[g+1] = f = update_f(f,Ws=Ws,H=HT,V=MT,N=N)}
  # plot(F)

  obs[l]=f
  exp[l]=f_
  MT_[l] = MT
  HT_[l] = HT
  S[l]=s
  N_[l] = N
}

plot(obs,exp ,xlim = c(0,1))
```

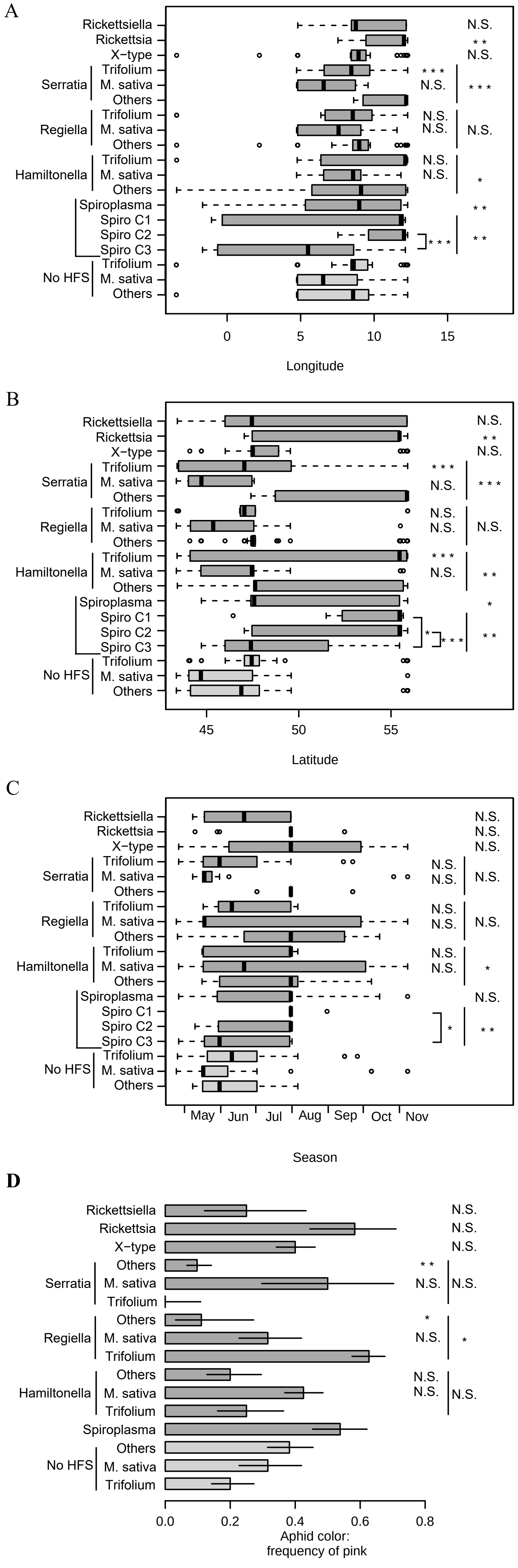


Figure S1: Occurrence in host plants, space and time

A-B: Four symbionts were spatially structured: *Rickettsia* was more frequent in the north-east (FDR-adjusted p=9.10^{-3}), *S. symbiotica* from *Trifolium spp.* in the south-west (FDR-adjusted p<1.10^{-3}), *H. defensa* from *Trifolium spp.* as well as *Spiroplasma* in the north (FDR-adjusted p=0.05 and p<1.10^{-3}).

C: The frequency of *H. defensa* was also increasing during the season (FDR-adjusted p=0.01).

D: Aphids sampled on *Trifolium spp.* contained more often *S. symbiotica* or *R. insecticola*, when they were green and pink, respectively (FDR-adjusted p<1.10^{-3}; p=0.02).

There was no effect of the year of sampling. The frequency of *H. defensa*, *R. insecticola*, and *S. symbiotica* also depended on the host plant. They were respectively most abundant on *M. sativa*, *Trifolium spp.*, and others plants (respective FDR-adjusted p-values: p<1.10^{-3}; p=9.10^{-3}; p<1.10^{-3}).

A-B: The frequencies of the three *Spiroplasma* clades depend on the longitude (FDR-adjusted p=3.10^{-3}) - clade 3 being more abundant in the west, and clade 2 in the east (p<1.10^{-3} ; Wilcoxon-test comparing clade 2 and 3), and on the latitude (FDR-adjusted p=7.10^{-3}) - the clade 3 being more abundant in the south (p=0.01 and 6.10^{-3} respectively when comparing clade 3 to clades 1 and 2; Wilcoxon-test; Fig. S1).

C: The frequencies of the clades were also affected by the season (FDR-adjusted p=3.10^{-3}) - clade 3 being most abundant early in the season (May-June), and clade 1 being most abundant late in the season (July-August; p=0.01; Wilcoxon-test comparing clades 1 and 3).

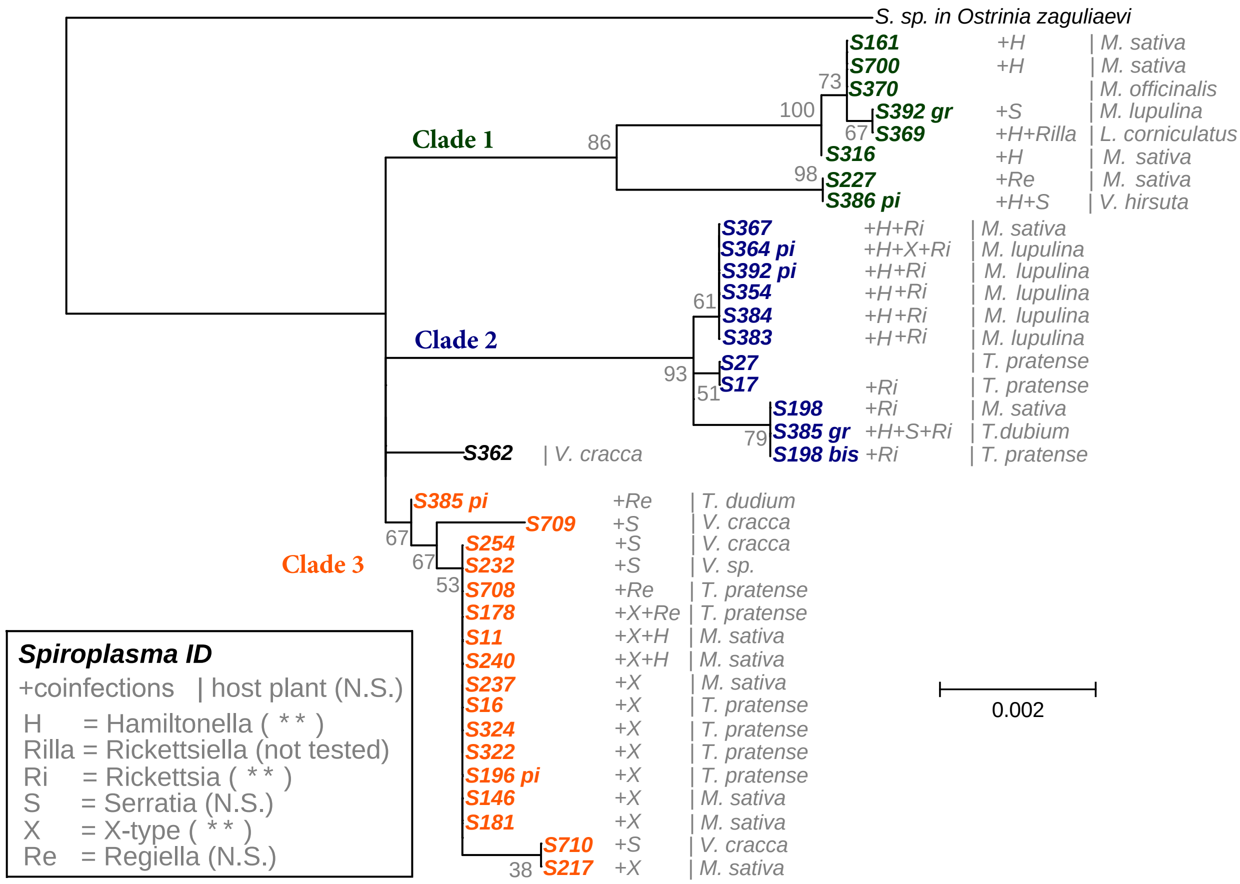


Figure S2: Phylogenetic tree of Spiroplasma from pea aphids

Maximum likelihood phylogenetic tree of Spiroplasma strains from pea aphids, reconstructed from concatenated rpoB and dnaA sequqences and using Spiroplasma from Ostrinia zaguliaevi as outgroup. Values in grey are the bootstrap support for the tree topology. The lists of the co-infecting symbionts and the host plants are indicated on the right side of the strain name. Scale bar indicates the substitution rate. The legend gives the abbreviations of the symbiont names, as well the FDR-adjusted level of significance of these variables in the RF predicting the clade of each Spiroplasma strain (1, 2 or 3).

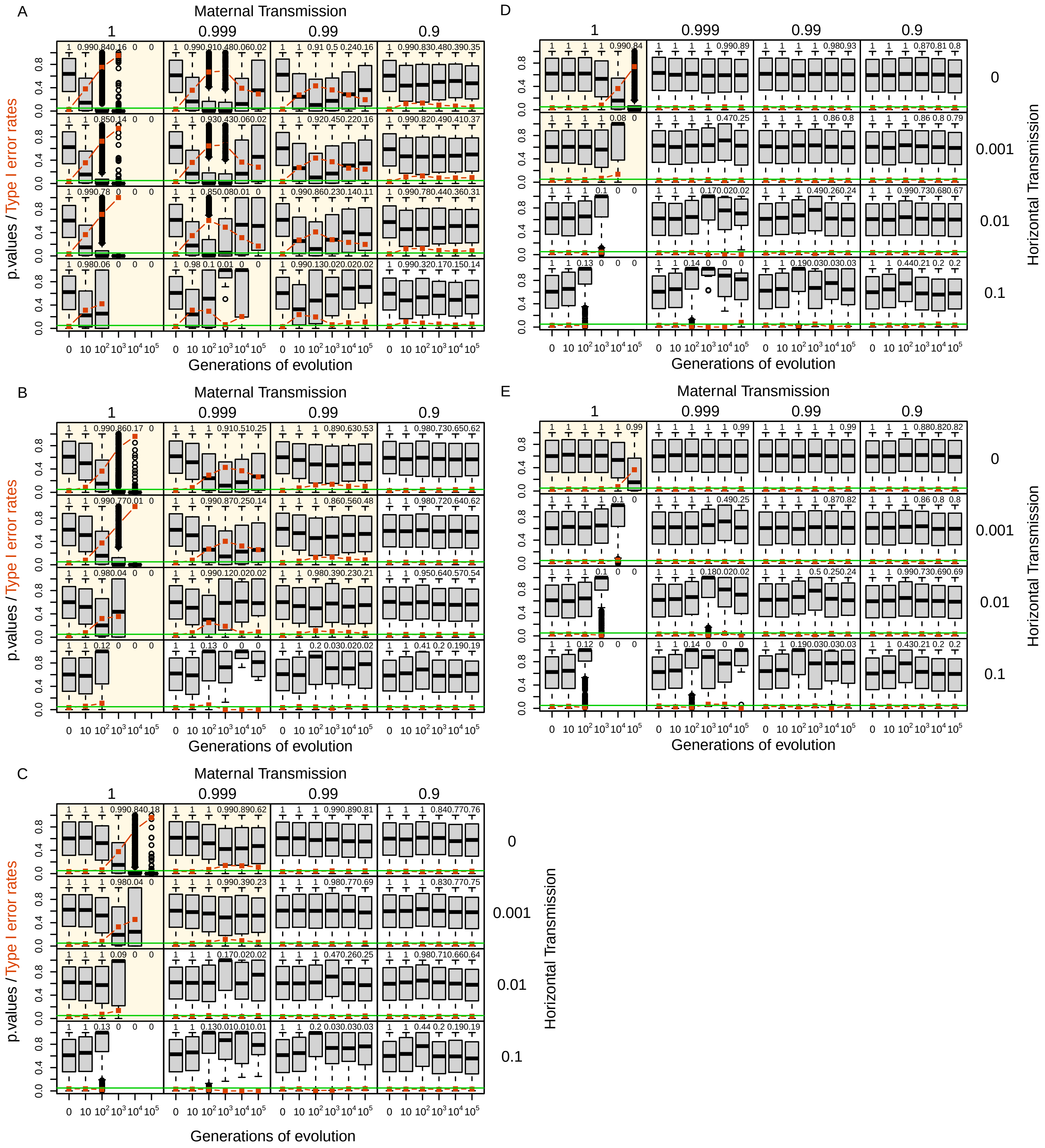


Figure S3: Deviations from random assortment induced by drift – full combination of parameters.

The frequency of two maternally transmitted symbionts evolved for up to 10^5 generations, starting from a population where symbionts were randomly assorted. Boxplots show the p-values of χ^2-tests assessing the deviations random assortment at generations 0, 10, 10^2, 10^3, 10^4 and 10^5. Panels A, B, C, D, and E correspond to female population sizes of 10^3, 10^4, 10^5, 10^6 and 10^7, respectively. In each panel, each set of boxplots corresponds 3000 populations evolving with the combination of the parameters 'horizontal transmission rate' (HT; columns) and 'maternal transmission rate' (MT; rows). The green horizontal line shows the 0.05 threshold, and the orange squares and lines indicate the type 1 error rate. Analyses of field surveys testing for deviation from random assortment usually assume that the type 1 error rate is of 0.05. Combinations of parameters where this is not the case have a yellowish background. The numbers above the boxplots indicate the proportion of populations that still have some polymorphism of infection by both symbionts.

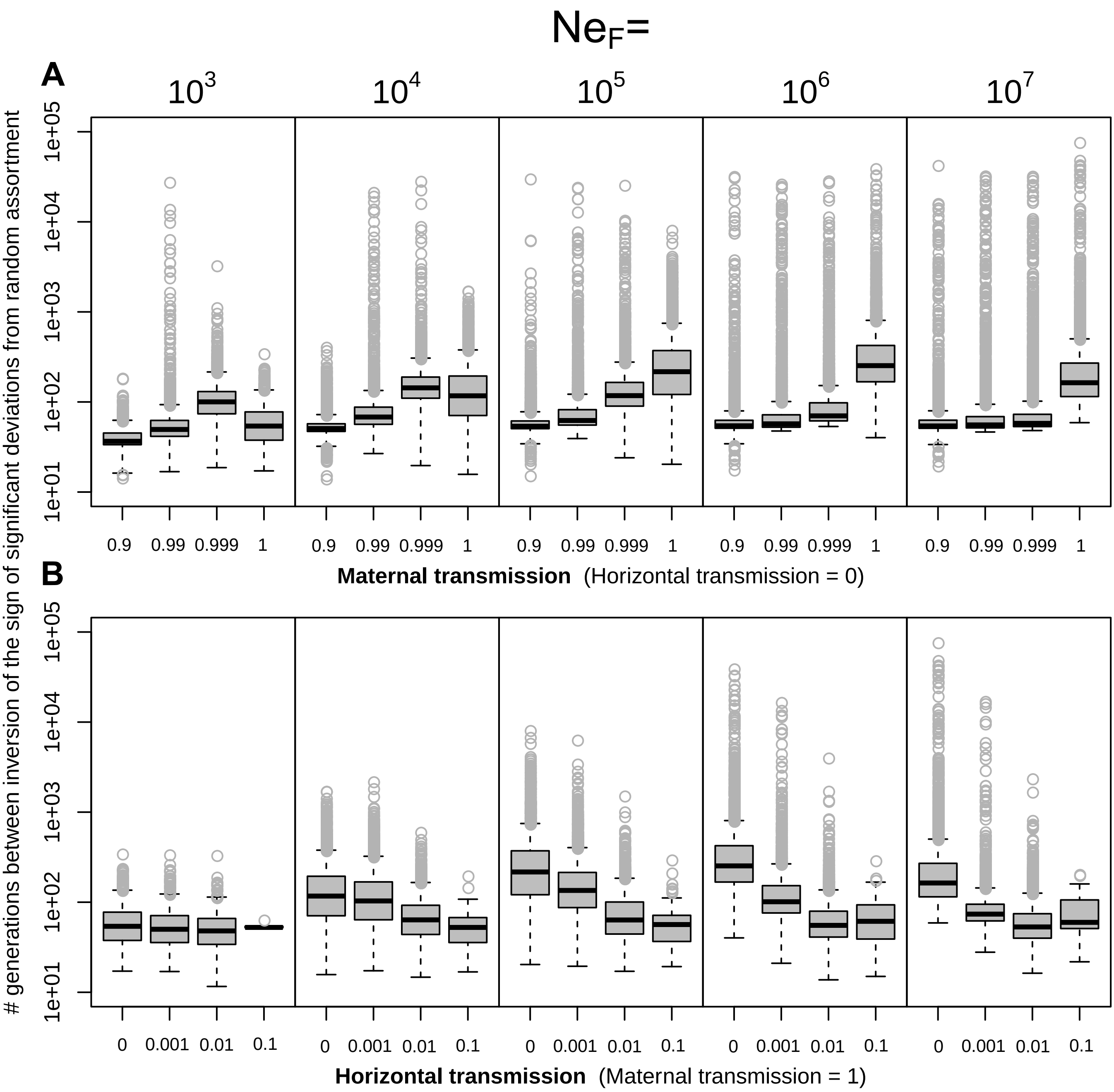


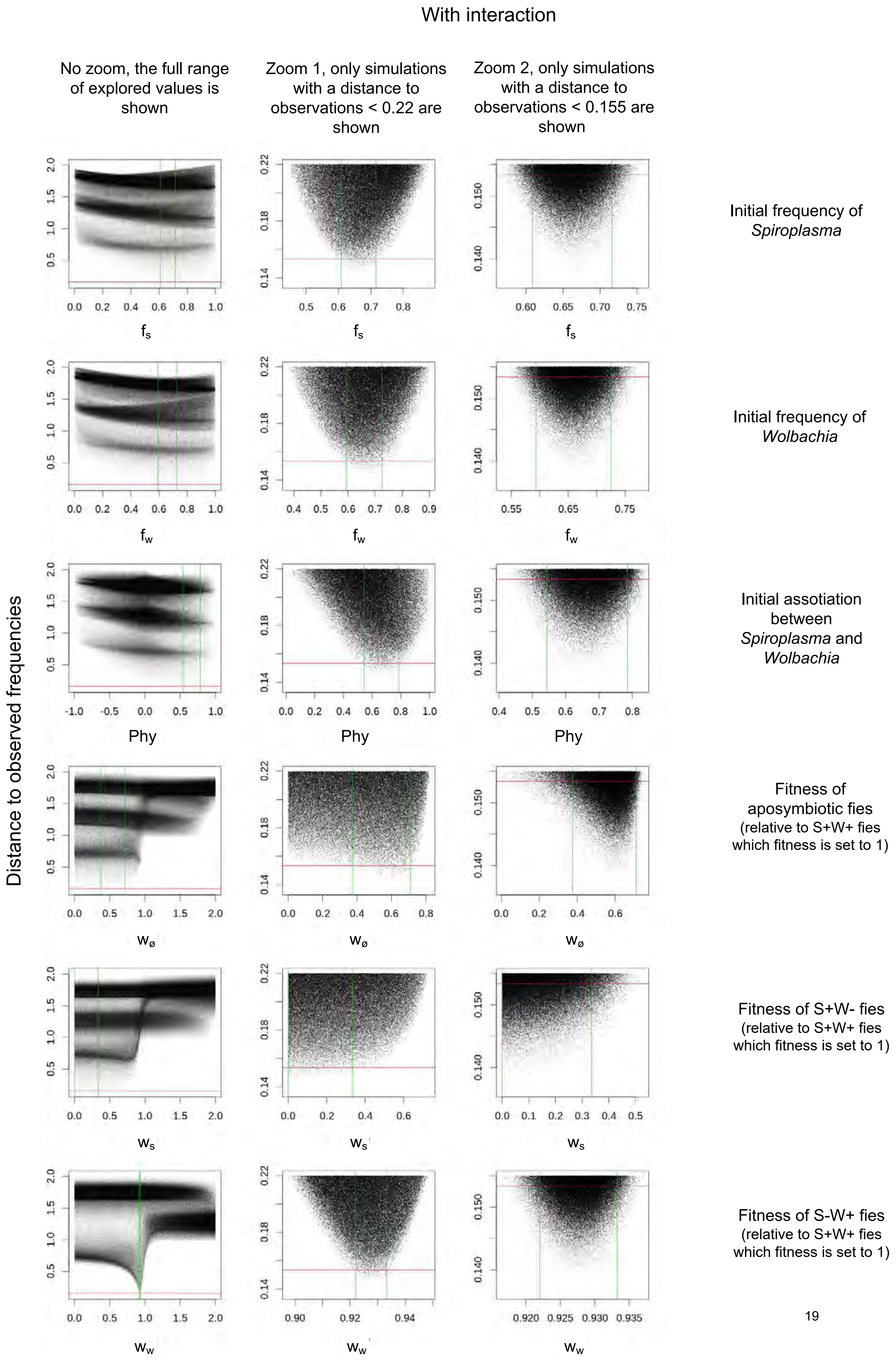
Figure S4: Frequency of inversion of significant deviations from random assortment.

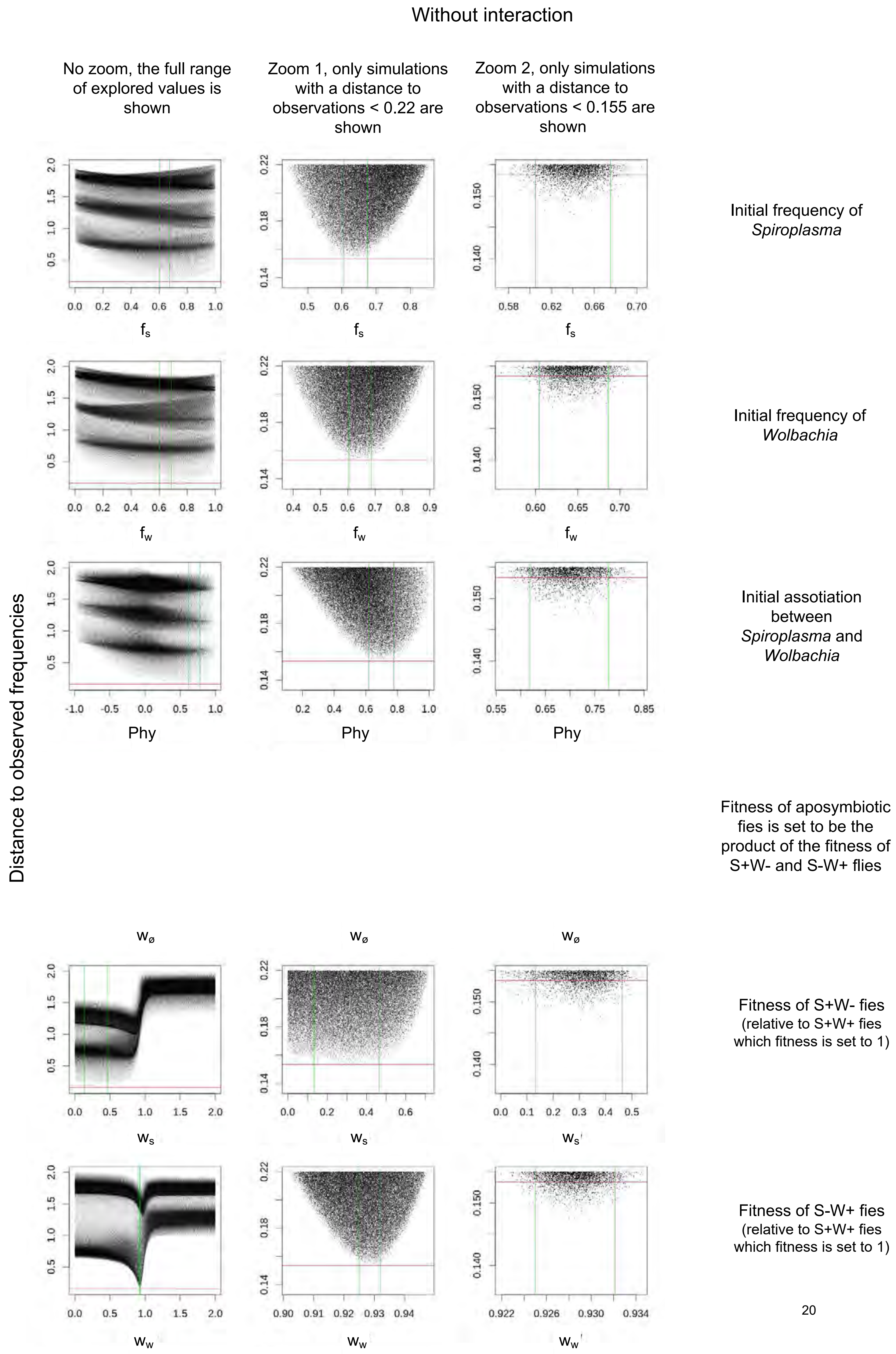
Association of pair of symbionts can be positive or negative, i.e; enrichment or depletion of co-occurence of symbionts, respectively. Drift is though to be unlikely to maintain these associations for many generations. We assesed this common wisdom by measuring the number of generation needed for a significant association to become unsignificant and then significant but of the opposit sign (Y-axis). This analysis was done for each combination of parameters shown in Fig. 5. Combinaisions where the rate of horizontal trasmission (H_T) is of 0 are swhown in panel A while those where the rate of successful maternal transmission (M_T) is of 1 are shown on panel B.

Figure S5:

For the 'interaction' and 'no interaction' models, the relationships between the values of the parameters (x-axes) and the Euclidean distance between the summaries of the simulated and observed datasets (y-axes) are shown on the two following pages. For each parameter (row), the left column shows, the full range of values sampled in the priors while the two others shows a zoom on the range of parameters leading to some small distances. The horizontal red line shows the tolerance threshold and parameters leading smaller distance are the approximate posteriors. The vertical green lines give their 95% CI.

Figures are on the two following pages


With interaction
No zoom, the full range of explored values is shown
Zoom 1, only simulations with a distance to observations < 0.22 are shown
Zoom 2, only simulations with a distance to observations < 0.155 are shown
Distance to observed frequencies
Initial frequency of *Spiroplasma*
Initial frequency of *Wolbachia*
Initial assotiation between *Spiroplasma* and *Wolbachia*
Fitness of aposymbiotic fies (relative to S+W+ fies which fitness is set to 1)
Fitness of S+W- fies (relative to S+W+ fies which fitness is set to 1)
Fitness of S-W+ fies (relative to S+W+ fies which fitness is set to 1)
f_s
f_w
Phy
$w_ø$
w_s
w_w


Without interaction
No zoom, the full range of explored values is shown
Zoom 1, only simulations with a distance to observations < 0.22 are shown
Zoom 2, only simulations with a distance to observations < 0.155 are shown
Distance to observed frequencies
Initial frequency of *Spiroplasma*
f_s
Initial frequency of *Wolbachia*
f_w
Initial assotiation between *Spiroplasma* and *Wolbachia*
Phy
Fitness of aposymbiotic fies is set to be the product of the fitness of S+W- and S-W+ flies
w_ø
Fitness of S+W- fies (relative to S+W+ fies which fitness is set to 1)
w_s
Fitness of S-W+ fies (relative to S+W+ fies which fitness is set to 1)
w_w