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

Mathé-Hubert Hugo$^{1,*}$, 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, Überlandstrasse 133, 8600 Dübendorf, Switzerland

$^2$Institute of Integrative Biology, Department of Environmental Systems Science, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland

$^3$Institute of Agricultural Sciences, Department of Environmental Systems Science, ETH Zürich, Schmelzbergstrasse 9, 8092 Zürich, Switzerland

$^4$Department of Biology, University of Rochester, Rochester, New York, United States of America

* author for correspondence:

Hugo Mathé-Hubert, E-mail: hugomh@gmx.fr

$^5$Present address: LIEC UMR 7360, Université de Lorraine, CNRS, Metz, France

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.15206

This article is protected by copyright. All rights reserved.
Short running title: Endosymbiont associations, drift or selection?

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 non-random 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 re-visit the association between Spiroplasma and Wolbachia in Drosophila neotestacea reported by Jaenike et al. (2010a). 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, Symbiotic community, Symbiont-symbiont interactions, Drift
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 and Fester 1991; Queller and 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., Oliver et al. 2010; Henry et al. 2013; Brucker & Bordenstein 2012). 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 (Zchori-Fein and Perlman 2004; Duron et al. 2008; Hilgenboecker et al. 2008; Regassa 2014). 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 et al. 2008), but it may also protect its host against parasites (e.g., Hedges et al. 2008; Teixeira et al. 2008; Faria et al. 2016) 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 et al. 2018) and in Harmonia axyridis (Goryacheva et al. 2018). Known effects of Spiroplasma also include reproductive parasitism (e.g.: Tabata et al. 2011; Sanada-Morimura et al. 2013; Anbutsu et al. 2016) as well as defense against at least three different kinds of parasites (Łukasik et al. 2013a; Xie et al. 2014; Ballinger and Perlman 2017; Frago et al. 2017).
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 et al. 2001; Simon et al. 2011). Interestingly, Ferrari et al. (2012) showed that the communities of facultative symbionts differed strongly among host plant-associated biotypes of the pea aphid (Peccoud et al. 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 (Łukasik et al. 2013b) 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 et al. 2015), or why the two defensive symbionts *Serratia symbiotica* and *Hamiltonella defensa* rarely co-occur in pea aphids (Oliver et al. 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 titer of *S. symbiotica*, but *S. symbiotica* does not affect the titer 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* (Łukasik et al. 2013a).

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. Firstly, 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. Secondly, 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 neostetacea*, respectively (Jaenike et al. 2010b; Himler et al. 2011; Cockburn et al. 2013). 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 and 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 et al. 2007) and *Hamiltonella* can be transmitted via parasitoids in aphids (Gehrer and Vorburger 2012). Both *H. defensa* and *Regiella insecticola* show occasional paternal transmission (Moran and Dunbar 2006). Jaenike (2012) argued that because of these non-maternal 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 et al. (2010a) 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 (Fig. 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 (Fig. 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, et al. 2010a), 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 (Fig. 1D).

Materials and Methods

Natural symbiont co-occurrence

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 meters 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 and 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 dataset. Because we had a particular interest in Spiroplasma infecting pea aphids (Mathé-Hubert et al. 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 Drosophila 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 et al. (2010a).

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 fulfill 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.

This article is protected by copyright. All rights reserved.
as usual spatial models (Hengl et al. 2018; Fouedjio and Klump 2019). 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, FDRp-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<sub>WD</sub> (whole dataset), RF<sub>M</sub> (*Medicago*) and RF<sub>T</sub> (*Trifolium*). For RF<sub>M</sub> and RF<sub>T</sub> 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* (clade 1, 2 or 3).

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.

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 non-overlapping 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) to 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\textsubscript{1} and S\textsubscript{2}), the population is described by the number of females being aposymbiotic (S\textsubscript{ø}), having only one of the symbionts (S\textsubscript{1} and S\textsubscript{2}), or having both (S\textsubscript{1,2}). At each generation, we simulated horizontal transmissions, reproduction events and maternal transmission failures (Fig. 2). The total number of horizontal transmissions is randomly chosen from a Poisson distribution whose mean depends on the horizontal transmission rate (H\textsubscript{T}) and the frequency of the transmitted symbiont (Fig. 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 (Fig. 2: question b.). Finally, the repartition of the recipients among the four host classes (S\textsubscript{ø}, S\textsubscript{1}, S\textsubscript{2}, and S\textsubscript{1,2}) is chosen randomly from a hypergeometric distribution to simulate sampling without replacement (Rice 2006;Fig. 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 sampling with replacement (Rice 2006: Fig. 2: question d.).

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

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 non-interactive 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 3000 replicates of all combinations of the following sets of parameters: Effective female population sizes (Ne: \(10^3, 10^4, 10^5, 10^6, \) 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 1. 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 et al.
(2010a), 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 (S₀, S₁, S₂, and S₁,₂) according to these frequencies and to the assumption of random assortment. The evolution of these populations of randomly assorted symbionts was then simulated for 10⁵ 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 non-random 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 χ²-test and to assess the sign of the deviation. The p-values were computed at every generation and recorded at generations 0, 10, 10², 10³, 10⁴ and 10⁵. 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.

Analysing a real dataset while accounting for drift.

Jaenike et al. (2010a) argued that Wolbachia and Spiroplasma in D. neotestacea are probably interacting in a way that enhances the fitness of co-infected 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 et al. (2010a) 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 et al. (2010a). 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 et al. 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 et al. 2010a; details in Table S3). According to this assumption, the dataset spans 77 Drosophila generations.

In these simulations, the parameters that are not estimated need to be fixed. These parameters are NeF, Mt and Ht. For Mt, we used the estimates of Jaenike et al. (2010a) that range from 0.945 to 0.981. To be conservative in inferring potential interactions, we used values for NeF and Ht that should overestimate the effect of drift. We assumed no horizontal transmissions (Ht = 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 et al. 2010a). The effective population size (Ne) was roughly estimated using the formula \(Ne = \pi/4\mu\) where \(\mu\) is the mutation rate of D. melanogaster (\(\mu = 2.8 \times 10^{-9}\); 95% CI = \([10^{-9}; 6.1 \times 10^{-9}]\); Keightley et al. 2014) and \(\pi\) 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 \(\pi\) and the upper CI of \(\mu\), 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 \(w_o, w_s, w_w\) 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 et al. (2010a) which modelled the cost of not having a symbiont and took the fitness of co-infected 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 \(w_i\) and \(w_o\) using the uniform prior ranging from 0 to 2 and we constrained \(w_o\) to be equal to \(w_i \times w_s\), which assumes a multiplicative fitness effect as in Jaenike et al. (2010a). For the ‘interaction’ model, we also estimated \(w_o\) using the same priors as for \(w_s\) and \(w_w\). This allows 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 \((N_E, H_T, M_T)\). The two sets of \(10^8\) simulated datasets were summarised and compared to the summary of the observed dataset. These summaries contain the mean frequencies of the four types of infections at start, mid-point 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 1000 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 et al. (2010a) to the distribution of the approximate posteriors of the fitnesses. Jaenike et al. (2010a) modelled the cost of not having a symbiont by setting \(w_{sw} = 1; w_s = 1 - s_w; w_w = 1 - s_w; w_o = (1 - s_w) \times (1 - s_s)\), 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 $w_0 = (1 - s_w) \times (1 - s_s) \times (1 - s_{sw})$, 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 estimated the two sided Bayesian $p$-value as $p.\text{value}(H_0; w_1 = w_2) = \begin{cases} 2 \times \text{Freq.}(w_1 < w_2) & \text{if } \overline{w}_1 > \overline{w}_2 \\ 2 \times \text{Freq.}(w_1 > w_2) & \text{if } \overline{w}_1 < \overline{w}_2 \end{cases}$ where the frequencies 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 et al. 2014).

Results

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 dataset (RFWD, Fig. 3A) six were detected in aphids from Trifolium spp. (RF_T, Fig. 3B), and only two were detected in aphids from M. sativa (RF_M, Fig. 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

This article is protected by copyright. All rights reserved.
(including ours) were always of the same sign (Table 1; Fig. 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 non-independence 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 co-varied negatively with the total number of co-infecting symbiont species, whatever their identity (*H. defensa*: FDRp-values=0.002 and 0.02 in RF WD and RF M, respectively; *R. insecticola*: FDRp-value<0.001 in the three models RF WD, RF T and RF M; *S. symbiotica*: FDRp-values<0.001 and in both RF WD and RF M; Fig. 4). For pea aphids from *Trifolium spp.*, the relationship between symbiont prevalence and the mean number of co-infecting 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<0.001, Fig. 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=0.003). For pea aphids from *M. sativa*, there was no detectable relationship between the frequency of symbionts and the number of co-infecting symbiont species ($R^2$=0.14; slope=−1.61; p-value=0.35, Fig. 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.

**Spiroplasma** intraspecific diversity

The phylogenetic tree indicates that in Europe, pea aphid-infecting *Spiroplasma* are sub-divided into at least three clades, although clade 3 has low bootstrap support (Fig. S2). The relative frequencies of these three clades did not depend on the host plant (p-value=0.98; Fig. S2) but were strongly dependent on the symbiont community. Clade 2 was more frequent in aphids already infected...
by other endosymbionts (FDRp-value=0.01) than the other two clades. The difference of clade 2 to clade 1 was marginally non-significant, while the difference to clade 3 was marginally significant (p-values=0.06 and 0.03, respectively; Wilcoxon-test). The Spiroplasma clades were also differently associated with H. defensa, X-type and Rickettsia (FDRp-values=0.02, 0.003 and 0.003, respectively). Specifically, clade 3 co-occurs less frequently with H. defensa than clades 1 and 2 (p-values=0.02 and 0.01; Fisher-exact test) and more frequently with X-type than clades 1 and 2 (p-values=0.003 and 0.006; Fisher-exact test; Fig. 3 and S2). Also, clade 2 is more frequently associated with Rickettsia than clades 1 and 3 (p-values<0.001 in both cases; Fisher-exact tests; Fig. 3 and S2).

**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 (Fig. 5). As expected, less-than-perfect maternal transmission or horizontal transmission tend to randomize symbiont associations (Fig. 5 and S3). However, our model shows that this effect can be offset by drift, in particular under effective population sizes lower than $10^6$ (Fig. 5). For effective female population sizes (N_{eF}) 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 (Fig. S4). Symbiont associations due to drift alone can thus be quite persistent in time.

**Spiroplasma - Wolbachia association in D. neotestacea, drift or selection?**

The positive association between Spiroplasma and Wolbachia in D. neotestacea reported in Jaenike et al. (2010a) has declined slowly from 2001 to 2016 and now seems absent. The frequency of co-infected 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 (Fig. 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

This article is protected by copyright. All rights reserved.
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 datasets are shown in figures S5. The estimated parameters revealed a clear interaction since the fitnesses of the four host categories had the following rank order: \( w_\alpha < w_\beta < w_w < w_{ws} = 1 \) (\( p \)-values = 0.006, <0.001, <0.001 respectively; Fig. 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 = [-0.4; -0.36]; Table 2). Consistent with this result, strains of \( D. \) neotestacea only infected by Spiroplasma can be difficult to maintain in the lab (John Jaenike, personal observation).

**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.: Russell & Moran 2006; Oliver et al. 2009; Vorburger & Gouskov 2011; Hansen et al. 2012; Lukasik et al. 2013; Weldon et al. 2013; Niepoth et al. 2018), 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

This article is protected by copyright. All rights reserved.
frequencies than expected under random assortment in earlier studies on pea aphids (Fig. 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 intra-specific level. The three main Spiroplasma clades identified in the phylogenetic tree were non-randomly 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 and Oliver 2017). But what is the biological meaning of these pervasive associations?

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 co-infections 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 et al. 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 endsymbionts, 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, $4N_e$ generations after it appeared (Kimura and Ohta 1969), or after $N_e$ 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 to 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.

**Spiroplasma - Wolbachia** association in *D. neotestacea*, drift or selection?

Jaenike et al. (2010a) 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 non-random assortment. The additional data collected since this study revealed that, at least in Rochester NY, the association has disappeared. Specifically, the frequency of co-infected 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 likely above $2.8 \times 10^5$, 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 the two symbionts is beneficial and spreads in the population from a matriline also containing another symbiont. Spiroplasma has actually undergone such a spread (Jaenike et al. 2010b; Cockburn et al. 2013), likely because of the protection it provides against the parasitic nematode Howardula aoronymphium (Jaenike et al. 2010b). 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, et al. 2010a). 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

This article is protected by copyright. All rights reserved.
contrasts with the result of (Jaenike et al. 2010b), that Spiroplasma is beneficial by protecting its host from the sterilising effect of the parasitic nematode Howardula 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 constant. We assumed that the imperfect maternal transmissions estimated by Jaenike et al. (2010a) are exact and representative of the considered time series. According to this assumption and in the absence of selection, maternal transmission failures would convert co-infected 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 et al. (2010b). 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, et al. 2010a) 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 observed between the infection status of the flies and their mitochondrial haplotype (Jaenike et al. 2010a). 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

This article is protected by copyright. All rights reserved.
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 et al. (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.

**Symbiont associations in pea aphids – selection or drift?**

After emphasizing the importance of considering drift as a source of non-random 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 (approx. 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 non-additive variances is lost at each sexual generation, which maintains the presence of clonal selection from year to year (Lynch and 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 & 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 et al. 2009). If estimates were similarly high for pea aphids, the importance of drift in generating non-random assortment of symbionts would be limited (Fig. 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 1400 to 2800 pea aphid generations, and there is strong genetic differentiation among pea aphids from the two continents today (Brisson et al. 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 1000 generations, which is unlikely for associations driven by drift (Fig 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; Russell et al. 2013; Zchori-Fein et al. 2014; Rock et al. 2017). 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*.

This article is protected by copyright. All rights reserved.
We further investigated this by characterising the link between the frequency of symbionts and the number of co-infecting 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 non-significant in aphids sampled on *M. sativa*, and significantly lower than the expected value (-1) in aphids sampled on *Trifolium spp.* (Fig. 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 et al. (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 (Fig. 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 and Oliver 2017). Also, Wulff et al. (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 and White 2015). Although difficult to interpret, field surveys remain crucial for our understanding of the ecology of symbioses.

**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 nr. CRSII3_154396 to CV) and a grant from the US National Science Foundation (grant nr. 1144581 to JJ).
References


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


Data Accessibility

The DNA sequences used in this study are available in Genbank (accession numbers: MG288511 to MG288588). The main dataset for D. neotestacea and A. pisum as well as the R function implementing the model are available on Dryad (DOI: https://doi.org/10.5061/dryad.ch4dp8n).
Author Contributions

HK, CH, CV, JJ and HMH performed the field sampling; HK, CH and HMH carried out the molecular analysis of the field samples; HMH was responsible for the data analysis and developed the model; HMH, HK, CH, JJ and CV wrote the paper.

Tables

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

<table>
<thead>
<tr>
<th>Geographic location</th>
<th>RF_{WD}</th>
<th>RF_{T}</th>
<th>RF_{M}</th>
<th>Oliver et al. 2006</th>
<th>Rock et al. 2017</th>
<th>Ferrari et al. 2012</th>
<th>Russell et al. 2013</th>
<th>Henry et al. 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host plant</td>
<td>Many</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>V</td>
<td>T V M</td>
<td>Many</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>N.A.</td>
<td>E</td>
<td>E</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

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 –

T: Trifolium spp.; M: Medicago sativa; V: Vicia; E: Europe; N.A.: North America
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 in the second column of the table (see material and methods). The 95% confidence intervals of the parameters are given in brackets.

<table>
<thead>
<tr>
<th></th>
<th>‘no interaction’</th>
<th>‘interaction’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPIC=25.31</td>
<td>BPIC=17.56</td>
</tr>
<tr>
<td>Initial population state</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f_S</td>
<td>0.64 (0.61; 0.68)</td>
<td>0.66 (0.61; 0.72)</td>
</tr>
<tr>
<td>f_W</td>
<td>0.64 (0.60; 0.69)</td>
<td>0.66 (0.59; 0.73)</td>
</tr>
<tr>
<td>Phi</td>
<td>0.70 (0.62; 0.78)</td>
<td>0.67 (0.54; 0.78)</td>
</tr>
<tr>
<td>Fitnesses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w_o=(1-s_o)×(1-s_s)×(1-s_w)</td>
<td>0.29 (0.10; 0.41)</td>
<td>0.57 (0.34; 0.69)</td>
</tr>
<tr>
<td>w_s=(1-s_o)×(1-s_w)</td>
<td>0.31 (0.11; 0.45)</td>
<td>0.13 (0; 0.37)</td>
</tr>
<tr>
<td>w_w=(1-s_w)×(1-s_w)</td>
<td>0.93 (0.92; 0.93)</td>
<td>0.93 (0.92; 0.93)</td>
</tr>
<tr>
<td>Costs of not having symbionts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s_s</td>
<td>0.07 (0.07; 0.08)</td>
<td>-3.13 (-90.4; -0.36)</td>
</tr>
<tr>
<td>s_w</td>
<td>0.68 (0.55; 0.89)</td>
<td>0.38 (0.25; 0.63)</td>
</tr>
<tr>
<td>s_sw</td>
<td>Set to 0</td>
<td>0.78 (0.32; 0.99)</td>
</tr>
</tbody>
</table>

Figures

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 and X: X-Type. B. Results of the simulation analysis: combination of parameters where drift induces non-random assortment. Red and green values refer to combinations of parameters for which the type I 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 and Purcell 1997; Moran and 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 (Fig. 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. \(w_o\): aposymbiotic; \(w_s\): Spiroplasma only; \(w_w\): Wolbachia only; \(w_{sw}\): co-infected flies.

Sources for pictures:
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).

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 dataset (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.
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-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.
Trifolium spp.

A

B

Trifolium spp. (No Regiella)

C

Medicago sativa

This article is protected by copyright. All rights reserved.
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 $\chi^2$-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 3000 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.
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 et al. (2010a) and these data were combined in the ABC framework to estimate the fitness effect of the four types of infection.