1

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

4

Mathé-Hubert Hugo[1,*], Heidi Kaech[1,2], Corinne Hertaeg[1,3], Christoph Vorburger[1,2]

6

[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

13

* author for correspondence:

Hugo Mathé-Hubert, Tel : +33 977216666 , E-mail: hugomh@gmx.fr

16

Short running title: Endosymbiont associations in pea aphids

18

Abstract

19

20   Virtually all higher organisms form holobionts with associated microbiota. To understand the

21 biology of holobionts we need to know how species assemble and interact. Controlled experiments are

22 suited to study interactions between particular symbionts, but they can only inform about a tiny portion

23 of the diversity within each species. Alternatively, interactions can be inferred from associations

24 among symbionts in the field that are more or less frequent than expected under random assortment.

25 However, random assortment may not be a valid null hypothesis for maternally transmitted symbionts

26 in finite populations, where drift alone can result in associations. Here we report results from a

27 European field survey of endosymbionts in the pea aphid (*Acyrthosiphon pisum*), and we develop a

28 model to study the effect of drift on symbiont associations under different population sizes,

29 considering varying rates of horizontal and maternal transmission. The model showed that even though

30 horizontal transmissions and maternal transmission failures tend to randomise symbiont associations,

31 drift can induce significant departures from random assortment, at least in moderate-sized populations.

32 Based on these results, we carefully interpret our field survey and we re-visit the association between

33 *Spiroplasma* and *Wolbachia* in *Drosophila neotestacea* reported by Jaenike *et al.* (2010). For this and

34 for several significant associations between symbionts in European pea aphids we conclude that under

35 reasonable assumptions of effective population size, they are indeed likely to be maintained by biased

36 co-transmission or selection. Our study shows that formulating appropriate null expectations can

37 strengthen the biological inference from co-occurrence patterns in the field.

38   Keywords: Coalescence, Symbiotic community, *Spiroplasma*, *Acyrthosiphon pisum*, drift

Introduction

40   Some of the interactions between organisms are so tight and durable that a new level of

41 organisation has been defined to describe them: the holobiont (Queller & Strassmann 2016). These

42 interactions are rarely bipartite and instead generally involve a host with a microbial community of

43 varying degree of complexity. From the host's perspective, these associations often lead to the

44 acquisition of novel traits, allowing the host to expand its ecological niche (e.g., Oliver *et al.* 2010;

45   Henry *et al.* 2013; Brucker & Bordenstein 2012). Understanding the evolutionary ecology of these

46   interactions requires identifying how species assemble to form holobionts, both at the ontogenetic and

47   evolutionary levels.

48   Large-scale screens for well-known species like *Wolbachia*, *Cardinium* or *Spiroplasma* suggest that

49   the majority of arthropod species are infected with heritable endosymbionts (Zchori-Fein & Perlman

50   2004; Duron *et al.* 2008; Hilgenboecker *et al.* 2008; Regassa 2014). However, there is considerable

51   variability in the effects these symbionts have on their hosts and in their prevalence among species.

52   *Wolbachia* is probably the most widespread of these endosymbionts. It has been estimated to occur in

53   66% of arthropod species, and it typically has either low (<10%) or very high (>90%) prevalence

54   within species (Hilgenboecker *et al.* 2008). *Wolbachia* is mainly known as a reproductive parasite

55   (Werren *et al.* 2008), but it may also protect its host against parasites (e.g., Hedges *et al.* 2008;

56   Teixeira *et al.* 2008; Faria *et al.* 2016) and is sometimes necessary for successful offspring production

57   (Dedeine *et al.* 2001; Kremer *et al.* 2009). Other widespread endosymbionts of arthropods are bacteria

58   of the genus *Spiroplasma*, infecting 4-7% of species, often with a low prevalence (Duron *et al.* 2008;

59   Regassa 2014). Known effects of *Spiroplasma* also include reproductive parasitism (e.g.: Tabata *et al.*

60   2011; Sanada-Morimura *et al.* 2013; Anbutsu *et al.* 2016) as well as defense against at least three

61   different kinds of parasites (Lukasik *et al.* 2013; Xie *et al.* 2014; Ballinger & Perlman 2017; Frago *et*

62   *al.* 2017).

63   The pea aphid, *Acyrthosiphon pisum*, is one of the main biological models of endosymbiosis. It can

64   be host to at least eight facultative heritable endosymbionts (Vorburger 2018), including *Spiroplasma*.

65   Interestingly, Ferrari *et al.* (2012) showed that the communities of facultative symbionts differed

66   strongly among host plant-associated biotypes of the pea aphid (Peccoud *et al.* 2009), although the

67   prevalence of *Spiroplasma* is only weakly affected by biotype, which explains only 9% of the variance

68   (Ferrari *et al.* 2012). A symbiont can spread in a host population as a reproductive parasite, for which

69   there is some limited evidence from pea aphid *Spiroplasma* (Simon et al. 2011), or by providing a

70   benefit to offset the cost it inflicts on the host. For example, *Spiroplasma* may protect pea aphids

71 against entomopathogenic fungi (Łukasik *et al.* 2013) or parasitoid wasps (Frago *et al.* 2017).

72 However, this cost-benefit balance varies depending on the environment, which is thought to be the

73 main reason for the observed polymorphism of facultative symbiont communities. For example,

74 defensive symbioses depend on the presence of some parasites of the host, and some symbioses help

75 the host to cope with warm environments (e.g., Russell & Moran 2006). The cost-benefit balance may

76 also depend on the associations with other symbionts. If two symbionts provide the same service, then

77 one of them might be redundant an thus too costly to the host. This may be the reason why defensive

78 bacterial symbionts are less frequent in aphids protected by ants (Henry *et al.* 2015), or why the two

79 defensive symbionts *Serratia symbiotica* and *Hamiltonella defensa* rarely co-occur in pea aphids

80 (Oliver *et al.* 2006). Also, interactions between symbionts can lead to non-additive effects in a

81 symbiont- and host-specific manner, making the outcome of a given association difficult to predict.

82 For instance, in *A. pisum*, *H. defensa* increases the titer of *S. symbiotica*, but *S. symbiotica* does not

83 affect the titer of *H. defensa* (Oliver *et al.* 2006). In the presence of *Spiroplasma*, *H. defensa* decreases

84 the fecundity of its host *A. pisum* while it increases the fecundity of the aphid *Sitobion avenae*

85 (Lukasik *et al.* 2013).

86 Interactive effects that vary from one symbiont strain to the other limit the utility of controlled

87 laboratory experiments, which usually include only a few particular strains, for making predictions

88 about the overall interactions among symbionts in natural populations. For this reason, results from

89 controlled experiments are often compared to analyses of field surveys (for several examples, see

90 Zytynska & Weisser 2016). These analyses notably aim at identifying pairs of symbionts for which the

91 co-occurrence is more or less frequent than expected under the null hypothesis of random assortment

92 (hereafter, positive and negative associations). Three kinds of mechanisms are generally considered

93 when trying to explain such deviation from random assortment. Firstly, the symbionts could interact in

94 a way that increases or decreases the rate of maternal transmission failures (e.g. Rock *et al.* 2017),

95 which should lead to negative or positive associations, respectively. Secondly, the symbionts could

96 have an interactive effect on host fitness, enhancing or hindering their co-transmission to the next

97   generation (e.g. Oliver *et al.* 2006). Thirdly, Jaenike (2012) and Smith *et al.* (2015) suggested that

98   neutral or even slightly costly maternally transmitted symbionts could spread in the host population by

99   chance association with another symbiont that is beneficial to the host. Faithful maternal transmission

100   would maintain the association between the two symbionts even if it were advantageous for the host to

101   lose one. This symbiont hitchhiking is analogous to the genetic hitchhiking (or draft), where a neutral

102   or slightly deleterious mutation spreads in the population because of its linkage disequilibrium with a

103   beneficial mutation (Felsenstein 1974). This symbiont hitchhiking might be responsible for the

104   evolutionary maintenance of the symbiont called *X-type*, which is costly to its host, has not been found

105   to provide any counterbalancing benefit, but is positively associated with the defensive symbiont *H.*

106   *defensa* (Doremus & Oliver 2017).

107   Jaenike (2012) argued that because most symbionts show some degree of maternal transmission

108   failure, associations due to symbiont hitchhiking should disappear rapidly. Thus, in most cases, the

109   presence of positive (or negative) associations between symbionts suggests an interaction that favors

110   (or hinders) their co-occurrence. Jaenike *et al.* (2010) showed that *Spiroplasma* and *Wolbachia* in

111   *Drosophila neotestacea* are positively associated despite imperfect maternal transmission. By

112   combining these observations with a mathematical model, they suggested that these two symbionts are

113   likely to be interacting positively with each other. As we will show in this paper, positive and negative

114   associations are also expected to appear and persist by drift, implying that without information about

115   the effective female population size, one needs to be cautious in assigning biological meaning to such

116   associations.

117   In the first part of this study, we used a field survey of *A. pisum* symbiotic infections to identify

118   positive and negative associations among symbionts. We report some new associations and confirm

119   some already known associations. We also show that there are three sub-clades of *Spiroplasma* in pea

120   aphids, which tend to be associated with very different symbiont communities. In the second part of

121   this study, we developed a simulation model of maternally and horizontally transmitted symbiont

122   communities in female host populations of finite size. We used this model to assess the effect of drift

123   on the tendency of symbionts to be randomly assorted to each other or not, thus aiding our

124   interpretation of observed associations in the field. We also applied the model to the association

125   between *Spiroplasma* and *Wolbachia* in *D. neotestacea* described by Jaenike *et al.* (2010) as a useful

126   test case, and, under a reasonable assumption for its effective population size, we come to the

127   conclusion that this association is indeed likely to reflect a positive interaction between symbionts.

Materials and Methods
128
129

130   Natural symbiont co-occurrence

Field sampling and symbiont screening
131
132
133   We sampled 498 aphids in France, Switzerland, Germany and Denmark during autumn 2014 and

134   spring and summer 2015. We selected colonies that were at least 2 meters apart from each other to

135   limit the risk of sampling offspring of the same mother. For each sample, we recorded the host plant

136   and the GPS coordinates. We characterised the presence of seven known facultative endosymbionts by

137   diagnostic PCR using symbiont-specific primers to amplify a part of the 16S rRNA gene (Table S1).

138   DNA was extracted from individual aphids using the 'salting out' protocol (Sunnucks & Hales 1996)

139   and the PCR cycling conditions are described by Henry *et al.* (2013). We also ran a diagnostic PCR for

140   the obligate endosymbiont *Buchnera aphidicola*, which is present in all aphids and thus served as an

141   internal positive control for the quality of the DNA preparation. Only samples testing positive for *B.*

142   *aphidicola* were included in the final dataset.

Phylogeny of *Spiroplasma* in pea aphids
143
144
145   Because we had a special interest in *Spiroplasma* infecting pea aphids (Mathé-Hubert et *al.* in

146   prep.), we also analysed the distribution of intraspecific diversity in this symbiont. This analysis

147   included the 26 strains found in the above-mentioned field sampling as well as 11 strains that were

148   kindly provided by Ailsa McLean (Department of Zoology, University of Oxford, UK; Table S2). The

149   diversity of *Spiroplasma* was characterised using a phylogeny based on the dnaA gene and the

150   rpoB gene, including its surrounding regions (445 pb and 2731 pb, respectively, including primers;

151   Table S1). The PCR cycling conditions are as described by Henry *et al.* (2013), except that the

152   elongation time of the primer pair RpoBF1.ixod was doubled (Table S1). We deposited all dnaA and

153   rpoB sequences in Genbank (accession numbers: MG288511 to MG288588).

154   We inferred the phylogenetic tree of the 37 *Spiroplasma* strains using *Spiroplasma sp.* in *Ostrinia*

155   *zaguliaevi* as outgroup, which is the most closely related species to *Spiroplasma* of the pea aphid for

156   which we could obtain sequences for both the rpoB and dnaA genes. The substitution model "GTR +

157   gamma + invariant sites" was identified by AICc ("phangorn" R package v 2.2.0, Schliep 2011) as the

158   best-fitting for the MAFFT-aligned and then concatenated dnaA and rpoB sequences. It was used to

159   build a maximum likelihood tree with the software MEGA6 (Tamura *et al.* 2013). This analysis

160   identified three main clades of *Spiroplasma* from pea aphids that are later referred to as clade 1, 2 and

161   3.

Statistical analysis

164   All analyses were performed using the R software (version 3.4.4; R Core Team 2018). To detect

165   associations of symbionts that are more or less abundant than expected under random assortment while

166   accounting for spatiotemporal non-independence among samples we predicted the presence or absence

167   of each symbiont species with a regression random forest model (RF). In each RF, the following

168   explanatory variables were used: latitude, longitude, season (number of days since the start of the

169   year), host plant on which the aphid has been sampled, aphid colour (pink or green), presence or

170   absence of the six other symbionts (one variable per symbiont) and the total number of other symbiont

171   species infecting the aphid.

172   In order to handle the multicollinearity as accurately as possible, RFs were grown with conditional

173   inference trees and the effect of variables estimated with conditional importance (package *party*;

174   Hothorn *et al.* 2006). Our methodology mostly follows the advice of Jones and Linder (2015). The

175   only difference is that we applied the permutation approach developed by Hapfelmeier & Ulm (2013)

176   to the conditional importance of explanatory variables to estimate their *p*-values that were then

177   adjusted to keep the false discovery rate at 5% (Benjamini & Yekutieli 2001). With this approach,

178   there is one model per symbiont, the presence or absence of the focal symbiont being explained by the

179   presence or absence of the six other symbionts. Thus each pair of symbiont (a and b) is considered by

180   two models, one explaining the symbiont a and the other explaining the symbiont b. This approach

181   thus leads to two *p*-values per couple of symbionts. To facilitate the interpretations of the results, we

182   repeated this analysis by restricting the dataset to aphids sampled on *Medicago sativa* and to aphids

183   sampled on *Trifolium spp*., which represent 30 and 33% of all field samples, respectively. This

184   restriction avoids lumping together aphids from multiple host races and thus simplifies the

185   interpretation. We refer hereafter to these three types of models as RF_{WD} (whole dataset), RF_M (*M.*

186   *sativa*) and RF_T (*Trifolium spp*.). For RF_M and RF_T the host plant was removed from the set of

187   explanatory variables.

188   These analyses revealed that some symbionts are less frequent in aphids already containing other

189   symbiont species, while others are not significantly affected by the presence of other symbionts. To

190   further investigate this, we tested if there was a link between the frequency of each symbiont species

191   and the average number of additional symbiont species with which it co-occurs. Such a link is

192   expected to occur even in the absence of interaction between symbionts, because under occasional

193   horizontal transmission, common symbionts are more likely to become associated with their own kind

194   than rare symbionts. We used a Wald test (function *regTermTest* of the package *survey*; Lumley &

195   Scott 2014) to test if the observed slope of the relationship between symbiont frequency and the mean

196   number of co-infecting symbionts was different from the slope expected under the null assumption of

197   no interaction between symbionts. The formula used to compute the expected slope is provided in

198   Appendix S1.

199   To investigate the intraspecific distribution of *Spiroplasma*, we used a classification RF to predict

200   the phylogenetic clade of *Spiroplasma* for each *Spiroplasma* infected aphid. The *Spiroplasma* strain

201   S362 was excluded from the analysis because it was not assignable to one of the three phylogenetic

202   clades. The variables were the same as in the previous analysis except that the aphid colour was not

203   available for all aphids and was thus not used. For significant variables, we compare clades to each

204   other in a pairwise fashion to characterise inter-clade differences. Since some of these significant

205   variables are continuous and some are categorical, we used Wilcoxon or Fisher's exact tests,

206   respectively to perform these pairwise comparisons.

Simulations of the symbiont co-occurrences

207

208

209   We investigated the effect of drift on deviations from random assortment of symbionts by

210   simulating populations of hosts with two species of maternally transmitted symbionts. In short, we

211   simulated populations of female hosts reproducing with non-overlapping generations, and being

212   infected by zero, one or two symbionts species (or strains). Symbionts are maternally transmitted with

213   a varying degree of efficiency and they are also horizontally transmitted at varying rates. Specifically,

214   we simulated 3000 replicates of all combinations of the following sets of parameters: Effective female

215   population sizes (N_eF: 10^3, 10^4, 10^5, 10^6, and 10^7), Successful maternal transmission rates (M_T: 1,

216   0.999, 0.99, 0.90), Horizontal transmission rates (i.e.: Average number of horizontal transmission

217   events caused by each infected host; H_T: 0, 0.001, 0.01, 0.1).

218   To initiate the populations we first randomly chose the frequencies of the two symbiont species

219   from a uniform distribution and computed the frequency of the four kinds of symbiont communities

220   (no symbiont, symbiont 1, symbiont 2, symbionts 1 and 2) according to the assumption of random

221   assortment and the frequencies of the symbionts. Then, the population sizes of the four kinds of

222   communities were sampled in a multinomial distribution, according to the previously computed

223   probabilities.

224   Population evolution was simulated for 10^5 generations or stopped if the polymorphism of infection

225   was lost. In the absence of selection, for most combinations of M_T and H_T, the symbionts get rapidly

226   fixed or lost which prevents assessing deviations from random assortment. To limit this loss of

227   polymorphism, we have set the direction and the strength of the selection on the presence of each

228   symbiont in a way that the expected equilibrium frequency of each symbiont equals their initial

229　frequency. For more detail, see Appendix S2. This procedure was more efficient at delaying the loss of

230　polymorphism due to the maternal transmission failure, than the loss due to horizontal transmissions.

231　This is because the amount of horizontal transmissions has a positive feedback on itself, while the

232　amount of vertical transmission failure has a negative feedback on itself (Appendix S2).

233　At each generation, 500 individuals were randomly sampled from the population and used to test

234　the significance of the deviation from the assumption of random assortment using a χ^2-test (Yates

235　1934) and to assess the sign of the deviation. The *p*-values are recorded at generations 0, 10, 10^2, 10^3,

236　10^4 and 10^5. We also assessed if, as it is often assumed, associations lasting for several years are

237　unlikely driven by drift. We estimated the number of generations needed for a previously significantly

238　positive association to become significantly negative. This was computed for each replicate as the

239　number of simulated generations divided by the number of inversion of the sign of significant trends.

240　The detailed description of the model and the model itself are in Appendix S2.

241　Jaenike *et al.* (2010) argued that *Wolbachia* and *Spiroplasma* in *D. neotestacea* are probably

242　interacting in a way that enhances the fitness of co-infected hosts because they are positively

243　associated in natural populations, despite having a maternal transmission rate of around 0.96, which

244　should rapidly randomise them. We used our model to assess how robust this conclusion is to drift. We

245　simulated, as described above, populations of female hosts infected by up to two symbionts with initial

246　frequencies of 0.8135 and 0.4289, which are the mean frequencies of infection with *Wolbachia* and

247　*Spiroplasma* in *D. neotestacea* measured by Jaenike *et al.* (2010). In this simulation, we set the

248　maternal transmission rate to 0.96, which is the mean of the estimates obtained by Jaenike *et al.* (2010)

249　for both *Wolbachia* and *Spiroplasma.*

Results

Natural co-occurrence of pea aphid symbionts

252　The RFs analysing the associations of symbionts in all field-sampled aphids revealed three positive

253　associations and six negative associations. Of these associations, all were detected in RF$_{WD}$ (Fig. 1A)

254   six were detected in RF_T (Fig. 1B), and only two were detected in RF_M (Fig. 1C). The lower number of

255   significant associations in *M. sativa* (RF_M) than in *Trifolium spp.* (RF_M) is unlikely due to a lower

256   statistical power since the sample sizes and the average number of symbionts per aphid were similar in

257   these two groups (*M. sativa*: 148 aphids with 0.97 symbionts per aphid on average; *Trifolium*: 161,

258   0.77). Of the 11 significant associations already identified by other studies on pea aphids, six were also

259   found in this study, and all associations reported by several studies (including ours) were always of the

260   same sign (Table 1).

261   Some symbiont prevalences co-varied negatively with the total number of co-infecting symbiont

262   species, whatever their identity (*H. defensa*: FDR-adjusted *p*-values: $P=0.002$ and $P=0.02$ in RF_{WD}

263   and RF_M, respectively; *R. insecticola*: $P<0.001$ in the three models RF_{WD}, RF_T and RF_M; *S. symbiotica*:

264    $P<0.001$ and in both RF_{WD} and RF_M). For pea aphids from *Trifolium spp.*, the relationship between

265   symbiont prevalence and the mean number of co-infecting symbionts was tight ($R^2=0.98$) and its slope

266   was more negative than the expected slope under the assumption of no interactions among symbionts

267   (Wald test: $P<0.001$, Fig. 2B), but this was mostly driven by *R. insecticola*. However, repeating the

268   analysis without aphids infected by *R. insecticola* also yielded significantly different slopes ($P=0.04$;

269   R²=0.88), mainly because *Rickettsiella* and *Rickettsia*, had a low frequency and tended to occur in

270   aphids already infected by more than one other symbiont species (Fig. 2C). For pea aphids from *M.*

271   *sativa*, the slope was also more negative than expected, but the difference was not significant ($P=0.45$,

272   Fig. 2A).

273   All these models included the variables longitude, latitude, season and in the case of RF_{WD}, the host

274   plant, to account for the non-independence between samples. However, these variables are highly

275   correlated, and although we used conditional inference trees and conditional importance, the results

276   should be interpreted with caution. The effects of these four variables on the frequency of each

277   symbiont are described in Figure S1.

278   *Spiroplasma* intraspecific diversity

279

280   The pairwise comparison of sequences from *Spiroplasma* of pea aphids revealed that on

281   average 0.63% of sites were different (max divergence=1.2%) which corresponds to 38% of the

282   mean divergence between *Spiroplasma* of pea aphids and *Spiroplasma* of the moth *Ostrinia*

283   *zaguliaevi* (1.6% of sites being divergent). The phylogenetic tree suggests that in Europe,

284   *Spiroplasma* of the pea aphid is sub-divided into at least three clades, but clade 3 in particular has a

285   low bootstrap support (Fig. 3). The relative frequencies of these three clades did not depend on the

286   host plant (non-FDR-adjusted P=0.98; Fig. 3) but were strongly dependent on the symbiont

287   community. Clade 2 was more frequent in aphids already infected by other endosymbionts (FDR-

288   adjusted P=0.01) than the other two clades. This was marginally non-significant when comparing it to

289   cade 1 and marginally significant when comparing it to clade 3 (P=0.06 and P=0.03, respectively;

290   Wilcoxon-test). The *Spiroplasma* clades were also differently associated with *H. defensa*, *X-type* and

291   *Rickettsia* (FDR-adjusted P=0.02, 0.003 and 0.003, respectively). Specifically, clade 3 co-occurs less

292   frequently with *H. defensa* than clades 1 and 2 (P=0.02 and 0.01; Fisher-exact test) and more

293   frequently with *X-type* than clades 1 and 2 (P=0.003 and 0.006; Fisher-exact test; Fig. 1 and 3). Also,

294   clade 2 is more frequently associated with *Rickettsia* than clades 1 and 3 (P<0.001 and <0.001; Fisher-

295   exact tests; Fig. 1 and 3).

Simulations of the symbiont co-occurrences evolving by drift

297   Symbiont associations that are more or less frequent than expected under the hypothesis of random

298   assortment are generally interpreted as a sign that an interaction between the symbionts promotes or

299   prevents their co-occurrence. Our simulations showed that for symbionts with perfect maternal and no

300   horizontal transmission, drift always leads to strong deviations from random assortment (Fig. 4). Such

301   deviations take a longer time to appear in large populations where drift is weak. Also, the median

302   number of generations between the appearance of a significant deviation from random assortment of a

303   given sign and the inversion of the sign is 54, 117, and 211, respectively, for effective female

304   population sizes of 10^3, 10^4 and 10^5 or more (Fig. S4). As often presumed, less-than-perfect maternal

305   transmission and horizontal transmissions tend to randomize symbiont associations (Fig. 4 and S3).

306   However, our model shows that this effect can be offset by drift. Thus, whether random assortment is a

307   valid expectation for non-interacting symbionts will depend on the effective female population size as

308   well as on the rates of horizontal transmissions and maternal transmission failures (Fig. 4 and S3).

309   Horizontal transmission and maternal transmission failure are also shortening the mean number of

310   generations between inversions of the sign of deviations from random assortment (Fig S3). In case of a

311   maternal transmission rate of 0.9, and a horizontal transmission rate of 0.1, the median number of

312   generations is 38, 50, and 52, respectively, for effective female population sizes of 10^3, 10^4 and 10^5 or

313   more.

314   Our simulations suggest that the observations reported by Jaenike *et al.* (2010) would be

315   compatible with drift if the effective female population size of *D. neotestacea* would be 10^3, but not if

316   it would be 10^4 or more. Indeed, for the symbiont frequency and maternal transmission rate reported

317   by Jaenike *et al.* (2010) and a female population size of 10^3, the type 1 error rate of a test of

318   independence of symbiont occurrence is 14%, while it is around 5% for female population size of 10^4

319   or more (Fig. S4). Our simulations also suggest that on average, the sign of deviations from random

320   assortment that are driven by drift would only happen every 34, 48, and 52 generations, respectively,

321   for female populations size of 10^3, 10^4, and $\geq 10^5$ (Fig. S4).

Discussion

323   Understanding how symbionts associate and interact within a host is important but challenging.

324   Laboratory experiments address this question by controlling all relevant parameters and observing the

325   outcomes, but they can only accommodate a tiny portion of the natural diversity of each interacting

326   species. In addition, such studies have often found that the outcome depends on the genotypes of the

327   interacting partners (e.g.: Russell & Moran 2006; Oliver *et al.* 2009; Vorburger & Gouskov 2011;

328   Hansen *et al.* 2012; Lukasik *et al.* 2013; Weldon *et al.* 2013; Niepoth *et al.* 2018), further complicating

329   predictions about these interactions in natural populations. Comparisons with field observations are

330   therefore essential. When analysing field surveys, interactions between symbionts are tentatively

331   inferred by comparing the observed frequency of co-occurrences to the frequency expected under the

332   hypothesis of random assortment. Departures from random assortment have been reported frequently

333   (e.g., Ferrari *et al.* 2012; Henry *et al.* 2013 in addition to the above-mentioned studies). Of the 21

334   possible pairwise associations among the seven facultative endosymbionts considered here, 11 have

335   already been reported to have significantly higher or lower frequencies than expected under random

336   assortment in earlier studies on pea aphids (Table 1). Six of these associations were also found in our

337   field sampling, and three are reported for the first time. When focusing on *Spiroplasma*, we even

338   found significant associations at the intra-specific level. The three main *Spiroplasma* clades identified

339   in the phylogenetic tree were non-randomly associated with other symbionts, independent of the host

340   plants the aphids were collected from. Such intraspecific variation in a symbiont-symbiont association

341   has also been reported between *X-type* and *H. defensa* in the pea aphid (Doremus & Oliver 2017). But

342   what is the biological meaning of these associations?

Drift induces deviations from random assortment

344   Our simulation model showed that random drift also induces associations among maternally

345   transmitted symbionts, suggesting that random assortment is not an appropriate null model to compare

346   symbiont co-infections against. The reason for that is most easily understood by considering the

347   coalescence framework. Statistical tests used to detect departures from random assortment assume that

348   samples are independent of each other. While this may apply to horizontally transmitted symbionts, it

349   will not apply to maternally transmitted symbionts. Some individuals will have the same symbiont

350   simply because they share a female ancestor that transmitted this particular symbiont community to all

351   of its offspring. In population genetics, this phenomenon is referred to as the coalescence (Balding *et*

352   *al.* 2007; which should not be confounded with the 'community coalescence'; Rillig *et al.* 2015). One

353   of the measures of the strength of drift is the expected coalescent time, the average number of

354   generations between two randomly sampled alleles and their most recent common ancestor. It is equal

355   to $2N_e$ for diploid autosomal genes, but it is only $N_e/2$ for maternally transmitted cytoplasmic genomes

356   (assuming a sex-ratio of 0.5). This is because only females do transmit the cytoplasmic genome, and

357   they have only one copy of it (Moore 1995; Jaenike 2012). Cytoplasmic genomes, including

358   endosymbionts, hence undergo four times more drift than nuclear autosomal genes.

359   Jaenike (2012) investigated how the population genetics framework can be adapted and used to

360   study the evolution of communities of maternally transmitted symbionts by comparing each symbiont

361   to a gene. However, given the generally high fidelity of maternal transmission and the low rate of

362   horizontal transmission of endosymbionts, one could also compare the whole symbiont community to

363   one gene with many alleles. Assuming no mutational bias, mutations increase allelic diversity and

364   maintain the alleles at a similar frequency, while, drift has the opposite effect. This mutation-drift

365   balance is largely analogous to the balance between maternal transmission failures, horizontal

366   transmissions and drift that we studied with our model. The main difference is that while maternal

367   transmission failure effectively acts as a directional mutation pressure, where the number of

368   individuals mutating from one state (infected) to the other (uninfected) is proportional to the number

369   of individual in the original state (infected), this is not true for horizontal transmission. The probability

370   of undergoing a horizontal transmission increases with the frequency of the symbiont, which makes

371   polymorphism less easily maintained in the presence of horizontal transmission.

372   Drift-induced deviations from random assortment can persist for a very long time. In a population

373   of diploid autosomal genes, a neutral mutation that reaches fixation does so, on average, $4N_e$

374   generations after it appeared (Kimura & Ohta 1969), or after N_e generations in a haploid, maternally

375   transmitted gene. Thus, we should expect that drift-induced deviations from random assortment of

376   symbionts should also be somewhat stable in time. In agreement with that, our simulations of two

377   strictly maternally transmitted symbionts show that drift-induced inversions of the sign of significant

378   deviations from random assortment occur every 50 to 200 generations on average, depending on the

379   effective female population size. These numbers should not be used as a general reference because

380   significance depends on the size of the samples used to assess deviations from random assortment (500

381   hosts in our simulations). Departures from random assortment became less stable in presence of

382   horizontal transmissions and maternal transmission failures.

The association between *Wolbachia* and *Spiroplasma* in *Drosophila neotestacea* as a test case

Jaenike *et al.* (2010) studied the maintenance of the positive association between *Wolbachia* and *Spiroplasma* in *D. neotestacea*. They used a mathematical model to show that given the maternal transmission rate estimated at 0.96, the association should disappear very rapidly in the absence of any 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, will push populations towards non-random assortment. Using our model with the same parameter estimates, we show that a significant association between two symbionts could be induced just by drift if the effective female population size of *D. neotestacea* were in the range of 10^3, but not if it were in the range of 10^4 or larger. The effective population size of *D. neotestacea* is not known, but the effective population size of North American *D. melanogaster* has been estimated to be between 3 to 5×10^6 (Garud *et al.* 2015). Unless the population structure of *D. neotestacea* is radically different from *D. melanogaster*, we can reasonably assume that its effective population size is higher than 10^4. Therefore, a stronger mechanism than drift appears to be responsible for maintaining the positive association between *Wolbachia* and *Spiroplasma* in *D. neotestacea*, corroborating the conclusion of Jaenike *et al.* (2010) that positive selection acts on the combination of these two endosymbionts.

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

409   clonal selection during the asexual phase of the life cycle (approx. 7-14 generations in pea aphids;

410   Barker 2016) can be intense (e.g., Vorburger 2006), which will also reduce the effective population

411   size. This clonal selection acts on the three components of genetic variance (additive, epistatic and of

412   dominance), but the optimisation it induces on the non-additive variances is lost at each sexual

413   generation, which maintains the presence of clonal selection from year to year (Lynch & Deng 1994).

414   On the other hand, aphids are good dispersers and exhibit shallow genetic population structure over

415   large geographic scales. For example, Ferrari *et al.* (2012) reported F_{ST}-values ranging from 0.03 to

416   0.11 for pea aphid populations from the same host plants across different European countries, and Via

417   & West (2008) reported a mean F_{ST} of 0.03 for North American populations of the pea aphid. Such

418   high population connectivity should have a positive effect on effective population size. We do not

419   know the effective population size of pea aphids, but DNA sequence-based estimates from other

420   cyclical parthenogens, waterfleas of the genus *Daphnia*, are rather high (300'000 – 600'000; Haag *et*

421   *al.* 2009). If estimates were similarly high for pea aphids, the importance of drift in generating non-

422   random assortment of symbionts would be limited (Fig. 4).

423   Another important aspect to consider is the consistency of the sign of significant associations.

424   While drift will generate associations of random and (slowly) fluctuating sign, selection is expected to

425   consistently favour either positive or negative associations between particular pairs of facultative

426   endosymbionts. For significant associations that were discovered in multiple studies, the sign of the

427   association was always the same (Table 1). Finding particular combinations of symbionts consistently

428   over- or underrepresented across different times and places suggests they are not caused by drift. For

429   example, the European pea aphids population is thought to have colonised North America at least 200

430   years ago, which would represent 1600 to 3000 pea aphid generations, and there is strong genetic

431   differentiation among pea aphids from the two continents today (Brisson *et al.* 2009). Despite this

432   separation, the four associations that have been reported in both continents are of the same sign.

433   However, any of these associations observed one time on each of the two continents are only two

434   independent observations of a kind of event that may also happen by drift.

435   In addition to testing for deviations from random assortment, some studies have also assessed

436   whether symbiont species tend to be differently associated with aphids that are already infected with 0,

437   1, 2 or more other symbiont species (e.g., Ferrari *et al.* 2012; Russell *et al.* 2013; Zchori-Fein *et al.*

438   2014; Rock *et al.* 2017). As these analyses are assuming that all the sampled symbiont communities

439   are independent, they are also affected by drift. In our field survey, we found that *H. defensa*, *S.*

440   *symbiotica* and *R. insecticola* occurred more frequently in aphids containing no or few other symbiont

441   species than expected under the assumption of random assortment, although this was only significant

442   in aphids sampled from *M.sativa*. When further investigating this by characterising the link between

443   the frequency of each symbiont species and the number of co-infecting symbiont species (see Figure

444   2), we found that this link was non-significant absent in aphids sampled on *M. sativa*, and stronger

445   than expected under the assumption of no symbiont interaction in aphids sampled on *Trifolium spp*. At

446   least two non-exclusive mechanisms could have yielded such a pattern. Firstly, rare symbionts might

447   be rare because they need the presence of other symbionts to persist in their host. Secondly, symbionts

448   might have adapted to co-occurrence patterns that are largely a function of their relative frequencies,

449   which entail that frequent symbionts should be less likely to share a host with other symbiont species

450   than rare symbionts. Rare symbionts are thus more strongly selected for persistence in the presence of

451   other symbiont species. This highlights that only abundant symbiont associations are efficiently

452   optimised by natural selection. It is therefore worth considering that associations between symbionts

453   that are currently maintained by a positive interaction may have evolved as a consequence of an

454   association that had originally appeared by drift or draft.

455   Lastly, inference on the biology of particular symbionts or their associations can be strengthened

456   from analyses of seasonal patterns and their comparison with expectations from laboratory

457   experiments. In studies of seasonal dynamics, the effect of drift is ideally ruled out using

458   spatiotemporal replication. For example, Smith *et al.* (2015) reported correlated change in the

459   symbiont frequencies and the parasitoid-induced host mortality which, together with the laboratory

460   evidence for symbiont-conferred resistance against parasitoids, suggested a causal relationship

461   between them. Also, Montllor *et al.* (2002) reported an increase of the frequency of *S. symbiotica*

462   correlated with temperature, which was consistent with this symbiont helping to tolerate heat stress.

463   Our sampling design was not suited for such inference, but the result that *H. defensa* was more

464   abundant in summer than in spring (Fig. S1) was at least consistent with selection by parasitoids as

465   also reported by Smith *et al.* (2015). Field observations are also informative when they do not match

466   the expectations from laboratory work. For example, laboratory experiments suggested that *X-Type*

467   does not provide any detectable benefit to the pea aphid, but it is quite frequent and positively

468   associated to *H. defensa*, suggesting it might have benefited from hitchhiking during the spread of *H.*

469   *defensa* (Doremus & Oliver 2017). Also, Wulff *et al.* (2013) did not find that the symbiont

470   *Arsenophonus* was protecting its *Aphis glycines* host against its main parasites, but it was present at

471   high frequency. This discrepancy between observation and expectation motivated further experiments

472   revealing that *Arsenophonus* provides a general – yet to be described – benefit to the aphid (Wulff &

473   White 2015). Although difficult to interpret, field surveys remain crucial for our understanding of the

474   ecology of symbioses.

Conclusion

476   The fate of holobionts depends on host-symbiont interactions as well as on symbiont-symbiont

477   interactions, but identifying them is not always straightforward. The approach consisting in analysing

478   the frequency of associations in the field is useful. However, the results it yields must be interpreted

479   carefully, in particular in the case of maternally transmitted symbionts, as patterns expected to be

480   produced by interactions between symbionts are also induced by drift. This highlights the need for

481   translating the biological null hypothesis into a valid statistical null hypothesis. Failure to do so may

482   result in biological misinterpretations of survey results.

Acknowledgements

References

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

491   Balding DJ, Bishop M, Cannings C (2007) Population Genetics. In: *Handbook of Statistical Genetics* ,
492   pp. 753–780.

493   Ballinger MJ, Perlman SJ (2017) Generality of toxins in defensive symbiosis: Ribosome-inactivating
494   proteins and defense against parasitic wasps in Drosophila. *PLoS pathogens*, **13**, e1006431.

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

496   Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under
497   dependency. *Annals of statistics*, **29**, 1165–1188.

498   Brisson JA, Nuzhdin S V, Stern DL (2009) Similar patterns of linkage disequilibrium and nucleotide
499   diversity in native and introduced populations of the pea aphid, Acyrthosiphon pisum. *BMC*
500   *Genetics*, **10**.

501   Brucker RM, Bordenstein SR (2012) Speciation by symbiosis. *Trends in Ecology and Evolution*, **27**,
502   443–451.

503   Dedeine F, Vavre F, Fleury F, *et al.* (2001) Removing symbiotic Wolbachia bacteria specifically
504   inhibits oogenesis in a parasitic wasp. *Proceedings of the National Academy of Sciences*, **98**,
505   6247–52.

506   Doremus MR, Oliver KM (2017) Aphid heritable symbiont exploits defensive mutualism. *Applied and*
507   *Environmental Microbiology*, **83**, 1–15.

508   Duron O, Bouchon D, Boutin S, *et al.* (2008) The diversity of reproductive parasites among
509   arthropods: Wolbachia do not walk alone. *BMC Biology*, **6**, 27.

510   Faria VG, Martins NE, Paulo TF, *et al.* (2016) Drosophila adaptation to viral infection through
511   defensive symbiont evolution. *PLoS genetics*, **12**, e1006297.

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

513   Ferrari J, West J a, Via S, Godfray HCJ (2012) Population genetic structure and secondary symbionts
514   in host-associated populations of the pea aphid complex. *Evolution*, **66**, 375–390.

515   Frago E, Mala M, Weldegergis BT, *et al.* (2017) Symbionts protect aphids from parasitic wasps by
516   attenuating herbivore-induced plant volatiles. *Nature Communications*, **8**, 1–9.

517   Garud NR, Messer PW, Buzbas EO, Petrov DA (2015) Recent selective sweeps in north american
518   Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics*, **11**, e1005004.

519   Haag CR, McTaggart SJ, Didier A, Little TJ, Charlesworth D (2009) Nucleotide polymorphism and
520   within-gene recombination in Daphnia magna and D. pulex , two cyclical parthenogens.
521   *Genetics*, **182**, 313–323.

522   Hansen AK, Vorburger C, Moran N a. (2012) Genomic basis of endosymbiont-conferred protection
523   against an insect parasitoid. *Genome Research*, **22**, 106–114.

524 Hapfelmeier A, Ulm K (2013) A new variable selection approach using random forests.
525   *Computational Statistics and Data Analysis*, **60**, 50–69.

526 Hedges LM, Brownlie JC, O'Neill SL, Johnson KN (2008) Wolbachia and virus protection in insects.
527   *Science*, **322**, 702.

528 Henry LM, Maiden MCJ, Ferrari J, Godfray HCJ (2015) Insect life history and the evolution of
529   bacterial mutualism. *Ecology Letters*, **18**, 516–525.

530 Henry LM, Peccoud J, Simon JC, *et al.* (2013) Horizontally transmitted symbionts and host
531   colonization of ecological niches. *Current Biology*, **23**, 1713–1717.

532 Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH (2008) How many species
533   are infected with Wolbachia?-A statistical analysis of current data. *FEMS microbiology letters*,
534   **281**, 215–20.

535 Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: A conditional inference
536   framework. *Journal of Computational and Graphical Statistics*, **15**, 651–674.

537 Jaenike J (2012) Population genetics of beneficial heritable symbionts. *Trends in Ecology and
538   Evolution*, **27**, 226–232.

539 Jaenike J, Stahlhut JK, Boelio LM, Unckless RL (2010) Association between Wolbachia and
540   Spiroplasma within Drosophila neotestacea: an emerging symbiotic mutualism ? *Molecular
541   Ecology*, **19**, 414–425.

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

544 Kremer N, Charif D, Henri H, *et al.* (2009) A new case of Wolbachia dependence in the genus
545   Asobara: evidence for parthenogenesis induction in Asobara japonica. *Heredity*, **103**, 248–56.

546 Łukasik P, van Asch M, Guo H, Ferrari J, Charles J. Godfray H (2013) Unrelated facultative
547   endosymbionts protect aphids against a fungal pathogen. *Ecology Letters*, **16**, 214–218.

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

552 Lumley T, Scott A (2014) Tests for regression models fitted to survey data. *Australian and New
553   Zealand Journal of Statistics*, **56**, 1–14.

554 Lynch M, Deng H-W (1994) Genetic slippage in response to sex. *The American Naturalist*, **144**, 242–
555   261.

556 Mathé-Hubert H, Kaech H, Ganesanandamoorthy P, Vorburger C (in prep.) Evolutionary costs and
557   benefits of *Spiroplasma* infection in pea aphids.

558 Montllor CB, Maxmen A, Purcell AH (2002) Facultative bacterial endosymbionts benefit pea aphids
559   Acyrthosiphon pisum under heat stress. *Ecological Entomology*, **27**, 189–195.

560 Moore WS (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus
561   nuclear-gene trees. *Evolution*, **49**, 718–726.

562 Niepoth N, Ellers J, Henry LM (2018) Symbiont interactions with non-native hosts limit the formation
563   of new symbioses. *BMC Evolutionary Biology*, **18**, 1–12.

564 Oliver KM, Degnan PH, Burke GR, Moran N a (2010) Facultative symbionts in aphids and the

565   horizontal transfer of ecologically important traits. *Annual review of entomology*, **55**, 247–266.

566   Oliver KM, Degnan PH, Hunter MS, Moran NA (2009) Bacteriophages encode factors required for
567   protection in a symbiotic mutualism. *Science*, **325**, 992–994.

568   Oliver KM, Moran NA, Hunter MS (2006) Costs and benefits of a superinfection of facultative
569   symbionts in aphids. *Proceedings of the Royal Society B*, **273**, 1273–1280.

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

573   Queller DC, Strassmann JE (2016) Problems of multi-species organisms: endosymbionts to
574   holobionts. *Biology & Philosophy*, **31**, 855–873.

575   Regassa LB (2014) The family Spiroplasmataceae. In: *The prokaryotes: Firmicutes and Tenericutes*
576   (eds Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F), pp. 551–567. Springer,
577   Berlin, Heidelberg.

578   Rillig MC, Antonovics J, Caruso T, *et al.* (2015) Interchange of entire communities: microbial
579   community coalescence. *Trends in Ecology & Evolution*, **30**, 470–476.

580   Rock DI, Wong N, Joffe J, *et al.* (2017) Context-dependent vertical transmission shapes strong
581   endosymbiont community structure in the pea aphid, Acyrthosiphon pisum. *Molecular Ecology*,
582   1–18.

583   Russell JA, Moran NA (2006) Costs and benefits of symbiont infection in aphids: variation among
584   symbionts and across temperatures. *Proceedings of the Royal Society B*, **273**, 603–610.

585   Russell J a., Weldon S, Smith AH, *et al.* (2013) Uncovering symbiont-driven genetic diversity across
586   North American pea aphids. *Molecular Ecology*, **22**, 2045–2059.

587   Sanada-Morimura S, Matsumura M, Noda H (2013) Male killing caused by a Spiroplasma symbiont in
588   the small brown planthopper, Laodelphax striatellus. *Journal of Heredity*, **104**, 821–829.

589   Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

590   Simon JC, Boutin S, Tsuchida T, *et al.* (2011) Facultative symbiont infections affect aphid
591   reproduction. *PLoS ONE*, **6**.

592   Smith AH, Piotr Ł, Connor MO, *et al.* (2015) Patterns, causes, and consequences of defensive
593   microbiome dynamics across multiple scales. *Molecular Ecology*, **24**, 1135–1149.

594   Sunnucks P, Hales DF (1996) Numerous transposed sequences of mitochondrial cytochrome I-II in
595   aphids of the genus Sitobion (Hemiptera: Aphididae). *Molecular Biology and Evolution*, **13**,
596   510–524.

597   Tabata J, Hattori Y, Sakamoto H, *et al.* (2011) Male killing and incomplete inheritance of a novel
598   Spiroplasma in the moth Ostrinia zaguliaevi. *Microbial Ecology*, **61**, 254–263.

599   Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary
600   genetics analysis version 6.0. , **30**, 2725–2729.

601   Teixeira L, Ferreira A, Ashburner M (2008) The bacterial symbiont Wolbachia induces resistance to
602   RNA viral infections in Drosophila melanogaster. *PLoS Biology*, **6**, e1000002.

603   Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation.
604   *Molecular Ecology*, **17**, 4334–4345.

605   Vorburger C (2006) Temporal dynamics of genotypic diversity reveal strong clonal selection in the

606   aphid Myzus persicae. *European Society for Evolutionary Biology*, **19**, 97–107.

607   Vorburger C (2018) Symbiont-conferred resistance to parasitoids in aphids – Challenges for biological
608   control. *Biological Control*, **116**, 17–26.

609   Vorburger C, Gouskov A (2011) Only helpful when required: A longevity cost of harbouring
610   defensive symbionts. *Journal of Evolutionary Biology*, **24**, 1611–1617.

611   Weldon SR, Strand MR, Oliver KM (2013) Phage loss and the breakdown of a defensive symbiosis in
612   aphids. *Proceedings of the Royal Society B*, **280**, 20122103.

613   Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology.
614   *Nature reviews. Microbiology*, **6**, 741–751.

615   Wulff JA, Buckman KA, Wu K, Heimpel GE, White JA (2013) The endosymbiont Arsenophonus is
616   widespread in soybean aphid, Aphis glycines, but does not provide protection from parasitoids or
617   a fungal pathogen. *PLoS ONE*, **8**, e62145.

618   Wulff JA, White JA (2015) The endosymbiont Arsenophonus provides a general benefit to soybean
619   aphid (Hemiptera: Aphididae) regardless of host plant resistance (Rag). *Environmental*
620   *Entomology*, **44**, 574–581.

621   Xie J, Butler S, Sanchez G, Mateos M (2014) Male killing Spiroplasma protects Drosophila
622   melanogaster against two parasitoid wasps. *Heredity*, **112**, 399–408.

623   Yates F (1934) Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal*
624   *of the Royal Statistical Society*, **1**, 217–235.

625   Zchori-Fein E, Lahav T, Freilich S (2014) Variations in the identity and complexity of endosymbiont
626   combinations in whitefly hosts. *Frontiers in Microbiology*, **5**, 1–8.

627   Zchori-Fein E, Perlman SJ (2004) Distribution of the bacterial symbiont Cardinium in arthropods.
628   *Molecular Ecology*, **13**, 2009–2016.

629   Zytynska SE, Weisser WW (2016) The natural occurrence of secondary bacterial symbionts in aphids.
630   *Ecological Entomology*, **41**, 13–26.

631

Data Accessibility

Author Contributions

638

639   HK, CH, CV and HMH performed the field sampling; HK, CH and HMH carried out the molecular

640   analysis of the field samples; HMH was responsible for the data analysis and developed the model;

641   HMH, HK, CH and CV wrote the paper.

Tables

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
Place	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					−				

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

Figures

## Figure 1: 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.
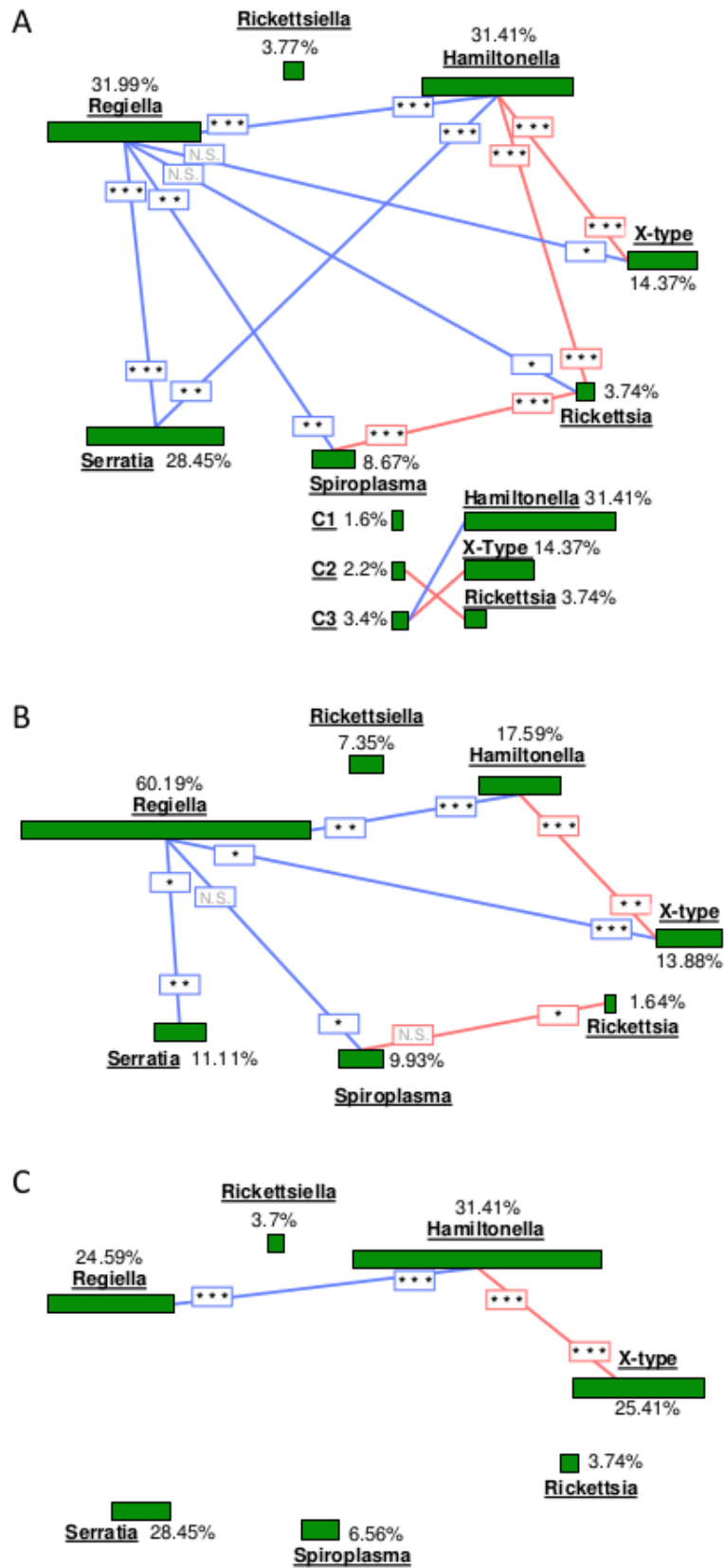
657   **Figure 2: Relationship between symbiont frequency and mean number of other**

658   **symbionts species**

659   Comparison of the actual (black) and expected (grey) relationship between the frequency of

660   endosymbionts species and the mean number of other symbiont species with which they co-occur.

661   Each observed value is connected to its expected value by a dotted line. Stars along these lines indicate

662   the level of significance detected by RF models. Analysis were performed on pea aphids from

663   *Medicago sativa* (A) and *Trifolium spp.* (B). Panel C refers to the analysis performed on aphids from

664   *Trifolium spp.*, but excluding individual infected with *Regiella insecticola* from the analysis. For each

665   of these three cases, we tested if the angle between the two slopes (α) differed significantly from zero.
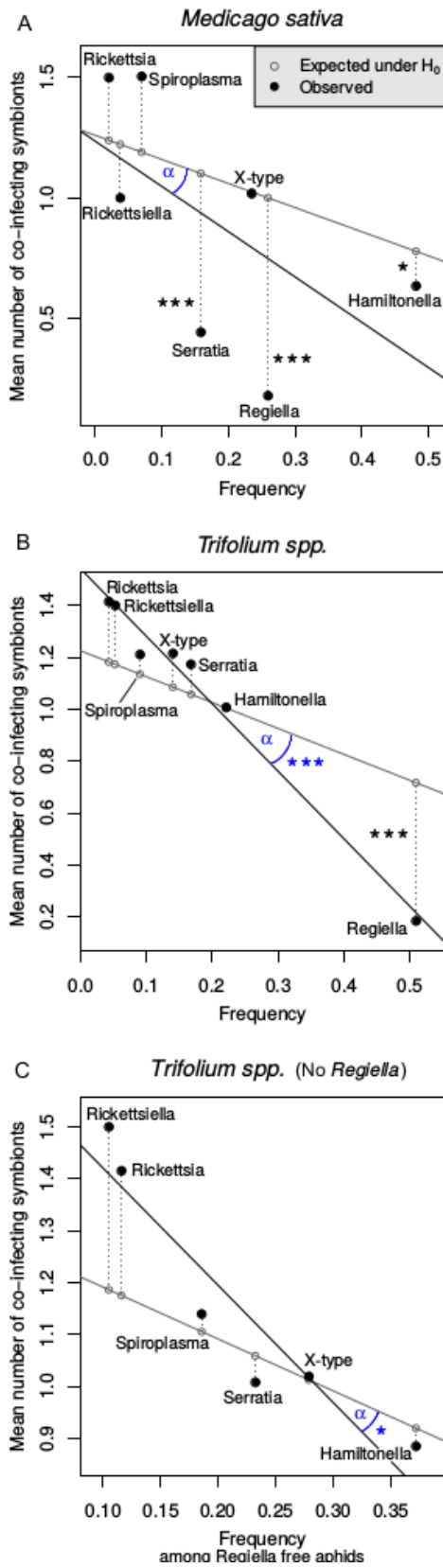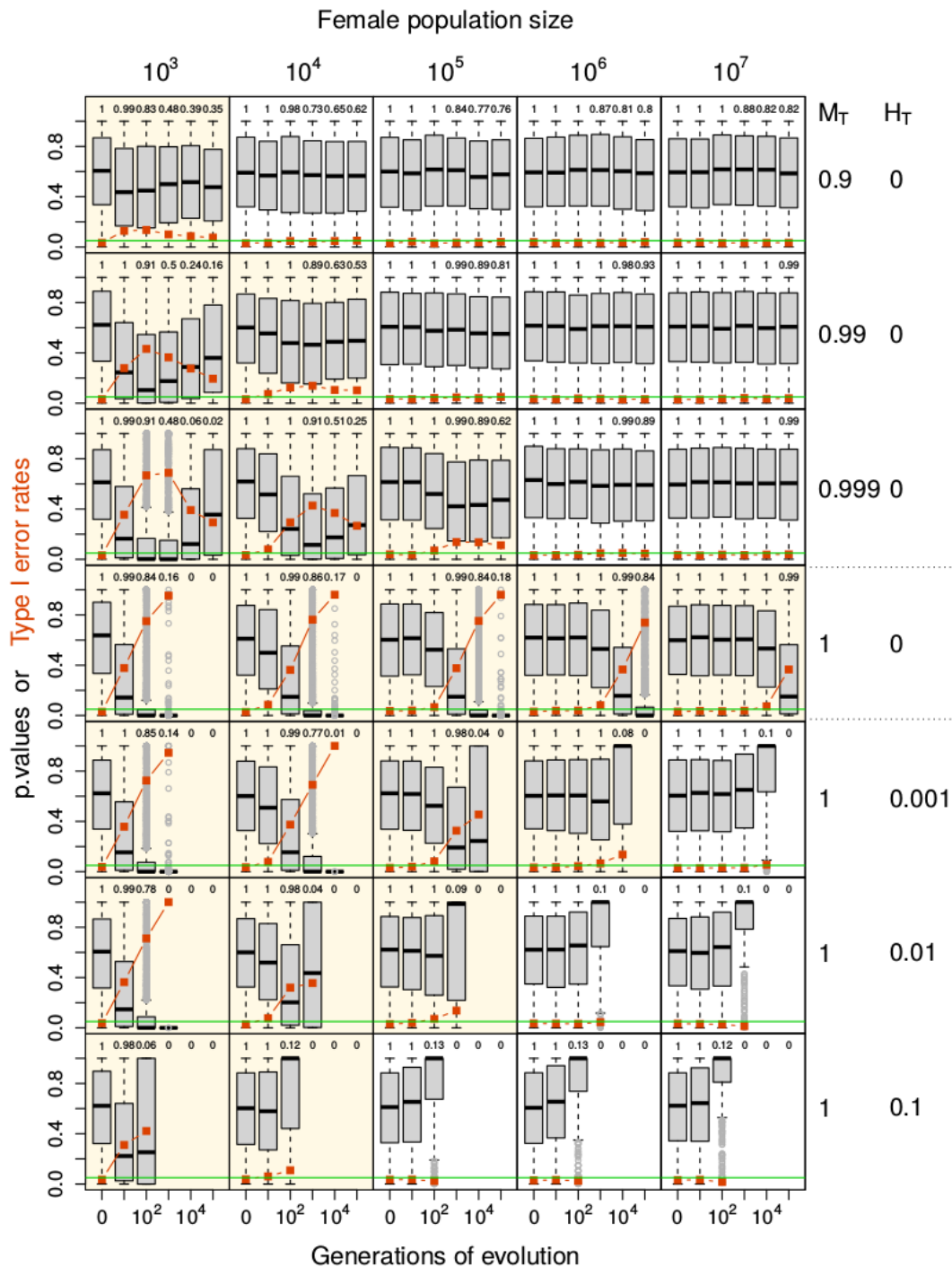
667 ## Figure 3: Phylogenetic tree of *Spiroplasma* from pea aphids

668 . Maximum likelihood phylogenetic tree of *Spiroplasma* strains from pea aphids, reconstructed

669 from concatenated rpoB and dnaA sequqences and using *Spiroplasma* from *Ostrinia zaguliaevi* as

670 outgroup. Values in grey are the bootstrap support for the tree topology. The lists of the co-infecting

671 symbionts and the host plants are indicated on the right side of the strain name. Scale bar indicates the

672 substitution rate. The legend gives the abbreviations of the symbiont names, as well the FDR-adjusted

673 level of significance of these variables in the RF predicting the clade of each *Spiroplasma* strain (1, 2

674 or 3).

675
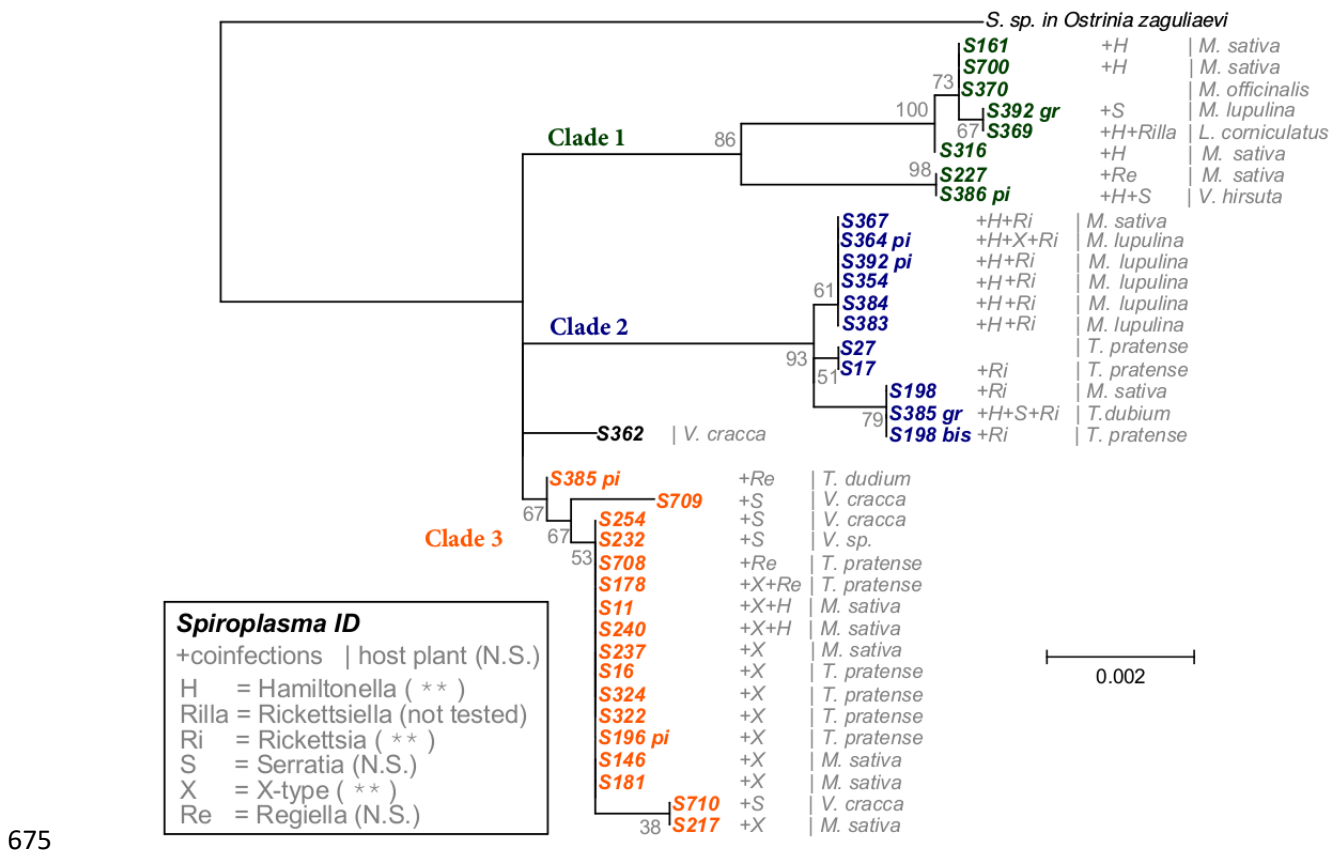
676 ## Figure 4: Deviations from random assortment induced by drift

677 The frequency of two maternally transmitted symbionts evolved for up to 10^5 generations, starting

678 from a population in which symbionts were randomly assorted. Boxplots show the *p*-values of χ^2-tests

679 assessing the deviations random assortment at generations 0, 10, 10^2, 10^3, 10^4 and 10^5. Each set of

680 boxplots corresponds to 3000 populations evolving with the combination of the parameters indicated

681 aside: 'female population size' (columns), 'horizontal transmission rate' (H_T; rows) and 'maternal

682   transmission rate' (M_T; rows). The green horizontal line shows the 0.05 threshold, and the orange

683   squares and lines indicate the type 1 error rate. Analyses of field surveys testing for deviation from

684   random assortment usually assume that the type 1 error rate is 0.05. Combinations of parameters

685   where this is not the case have a yellowish background. The numbers above the boxplots indicate the

686   proportion of populations that still retained some polymorphism of infection by both symbionts.

687