The incidence of bacterial endosymbionts in terrestrial arthropods

Lucy A. Weinert[1†], Eli V. Araujo-Jnr[2†], Muhammad Z. Ahmed[3], John J. Welch[2*]

March 19, 2015

1. Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, UK.

2. Department of Genetics, University of Cambridge, Downing Street, Cambridge, UK.

3. Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA.

† These authors contributed equally to this work.

* Author for correspondence: John Welch, e-mail: j.j.welch@gen.cam.ac.uk

Running title:

Symbiont incidence in arthropods

Key words:

Rickettsia, *Wolbachia*, *Cardinium*, Maximum Likelihood, infection

Abstract

Intracellular endosymbiotic bacteria are found in many terrestrial arthropods, and have a profound influence on host biology. A basic question about these symbionts is why they infect the hosts that they do, but estimating symbiont incidence (the proportion of potential host species that are actually infected) is complicated by dynamic or low prevalence infections. We develop a maximum likelihood approach to estimating incidence, and testing hypotheses about its variation. We apply our method to a database of screens for bacterial symbionts, containing >3600 distinct arthropod species, and >150,000 individual arthropods. After accounting for sampling bias, we estimate that 52% (CIs: 48-57) of arthropod species are infected with *Wolbachia*, 24% (CIs: 20-42) with *Rickettsia* and 13% (CIs: 13-55) with *Cardinium*. We then show that these differences stem from the significantly reduced incidence of *Rickettsia* and *Cardinium* in most hexapod orders, which might be explained by evolutionary differences in the arthropod immune response. Finally, we test the prediction that symbiont incidence should be higher in speciose host clades. But while some groups do show a trend for more infection in species-rich families, the correlations are generally weak and inconsistent. These results argue against a major role for parasitic symbionts in driving arthropod diversification.

Introduction

Terrestrial arthropods carry an array of intracellular endosymbiotic bacteria. These bacteria have a profound influence on their hosts, and are thought to affect areas of biology ranging from reproductive mode and resistance to viruses, to effective population size and rate of speciation [1-9]. Some of the bacteria are also remarkable for the breadth of their host range [3,9-11], but relatively little is known about why they infect the hosts that they do.

Several authors have suggested that symbiont infection frequency might vary predictably with host biology [e.g., 1,3,8,9,12-19]. For example, two distinct arguments predict that symbionts should be more common in host taxa that are species rich. First, some symbionts might cause reproductive isolation in their hosts, thus increasing the number of species in infected groups, relative to uninfected groups [20-22]. Second, if symbionts occasionally switch hosts, and if these switches take place preferentially between closely-related host species, then symbionts should be more common in host groups with many closely-related species, since these relatives can act as ready sources of infection [23,24].

Hypotheses of this kind are common, but difficult to test in a rigorous comparative framework (though see, e.g., [8,18]). This is partly because symbiont incidence (i.e., the proportion of potential host species that are actually infected) is not easy to measure. While symbionts can be detected with PCR-based screens, infections vary in their prevalence (i.e., the proportion of individuals infected), and so it follows that low prevalence infections will be difficult to detect, that symbiont absence is impossible to prove, and that the number of infected samples might grossly underestimate the number of infected populations [e.g., 11,25]. Furthermore, the infection status of a population can change rapidly [e.g., 3,23,26,27], and this makes any single sample a mere snapshot of the ongoing ecological dynamics.

A way to mitigate these problems is to combine data from several populations, and estimate the distribution of prevalences across a group of potential hosts [11]. This distribution might be relatively stable, even when infection in any single species changes rapidly [23,24,28,29], and the distribution allows us to infer the number of unobserved, low prevalence infections, even when few populations were sampled in depth. Such an approach to estimating incidence was pioneered by Hilgenboecker *et al.* [11], and has since been applied to several bacterial symbionts [11,29,30]. Here, we extend the approach of [11] in a full likelihood-based framework; this allows us to place proper confidence intervals on our incidence estimates, and to formally test hypotheses about whether and why incidence varies.

We apply our method to a newly collated database of published screens for three genera of bacterial secondary symbionts: *Wolbachia*, *Rickettsia* and *Cardinium.* Each genus employs a range of transmission strategies, but is best known as a reproductive parasite, manipulating the sexual biology of hosts to facilitate vertical transmission via the egg cytoplasm [1,3,5,9]. Most importantly, each genus has been extensively studied, and so our database contains screens of over 150,000 individuals from over 3500 distinct arthropod species.

Methods

Data collection

We searched the literature for PCR-based screens of *Wolbachia*, *Rickettsia* and *Cardinium* in wild populations of terrestrial arthropods (full details of our inclusion criteria are given in supplementary information section 1). For each population, we recorded the host species, the number of individuals screened, and the number of individuals found to be infected. The final database included data from 361 distinct source publications, comprising screens of over 10,000 populations. Each screened arthropod was classified according to up-to-date taxonomy, and a taxonomic breakdown of the database is shown in supplementary Figure S1. The full database is included as online supplementary information. To estimate the relative species richness of different arthropod groups, we used estimates of the number of described species. These will, of course, be a crude proxy for true species number, but are acceptable for our purposes, given the many well-known difficulties in extrapolation [31]. All estimates used are listed in supplementary Table S2 and online supplementary information.

Model

We estimate symbiont incidence by first inferring the distribution of infection prevalences across many populations. Following [11], we initially assume that between-population variation in prevalences can be adequately described by a beta distribution, whose parameters are estimated from the screen data. From the best fit distribution, we then calculate the proportion of species infected above a given threshold frequency c, and denote this estimate of incidence as x_c [11]. Most results reported below use $c = 0.001$, and thus we define a population as "infected" if it has a prevalence of greater than one in a thousand individuals. We use this threshold frequency for expedience, but it is clear that the proportion of species in which no single individual is infected will be difficult to estimate with much confidence or precision (and this is borne out by results reported below). Furthermore, the threshold frequency reflects a biologically meaningful distinction between established infections and very low prevalence "dead-end" infections, which are unlikely to persist in the host population [32,33]. All equations associated with the model, and details of the numerical methods are found in the supplementary methods section 2.

Results

Estimating symbiont incidence

We begin by estimating symbiont incidence across the terrestrial arthropods as a whole. Figure 1 shows three such estimates for each bacterial genus. The initial estimates (labelled (a) in Fig. 1), were obtained from fitting a simple beta distribution to our complete database of screens. However, due to the shape of the beta distribution, these estimates entail the assumption that no population is completely free from infection (with a prevalence of exactly zero), and no population is completely infected (with a prevalence of exactly one, as with an obligate or primary symbiont). To relax this questionable assumption, we developed a method of fitting a doubly-inflated beta distribution [34] which does allow for completely uninfected and completely infected host populations, as well as populations with intermediate prevalence. Comparing the two models on simulated data shows that the doubly-inflated distribution is much more accurate when, in reality, a large fraction of populations do not harbour the symbionts (see supplementary methods section 3 and Fig. S2 for full details). However, for our real data, fitting the doubly-inflated distribution had almost no effect on the incidence estimates (see results labelled (b) in Fig. 1). This suggests that model inadequacy is not greatly influencing our results.

More fundamentally, we are most interested in the incidence across arthropod species (i.e., the proportion of species infected), and for this purpose, estimates from our complete database will be biased in at least three ways. First, and most obviously, some species are represented by only a single population sample, and some by a very large number. For example, the vectors of Rickettsial disease (Parasitiformes and Siphonaptera) are hugely overrepresented (Supp. Figure S1). Second, there are clear taxonomic biases in the sampled species. In particular, minor arthropod orders are overrpresented (presumably from studies of symbiont host range; Supp. Figure S1). Third, and more subtly, sampling might be biased by the concentration of research effort on populations and species that were already known to contain infection [11].

To mitigate and test for these biases, we developed a three-stage process, which we call "standardised sampling". First, we subsampled our data, retaining only the single largest screen from each species. Second, we devised a method of estimating incidence for each of the major groups of arthropods, and then combining these estimates in a weighted sum, weighting the estimate from each group by its contribution to total arthropod biodiversity. Third, we tested for differences in symbiont prevalence between multi-individual screens (which are more likely to be carried out on species known to carry infection), and single-individual screens (which are most likely to resemble a quasi-random sample of species). Full details of the "standardised sampling" are given in supplementary methods section 4, and Tables S2 and S3, and results obtained with this approach are labelled (c) in Figure 1.

These improved estimates (Fig. 1c) are substantially lower than the estimates from the complete database, but they remain remarkably high. We estimate that just over half of terrestrial arthropod species are infected with *Wolbachia* at a non-neglible frequency (52%, CIs 48-57), around a quarter infected with *Rickettsia* (24% CIs 20-42) and around an eighth infected with *Cardinium* (13%, CIs 13-55). Furthermore, we cannot reject the possibility that *Rickettsia* and *Cardinium* incidences are much higher (Fig. 1). Their large and skewed confidence intervals reflect the shape of the underlying distributions of prevalences that we inferred from the data. In particular, for all bacteria, we inferred that most single species were subject to either very high, or very low levels of infection at any given time (Table S1; [11,29,30]). However, estimates of mean prevalence levels were much lower for *Cardinium* and *Rickettsia* (<6%) than for *Wolbachia* (24%; Table S1). As such, for *Cardinium* and *Rickettsia*, it was difficult to distinguish between low incidence, and a high incidence of low-prevalence infections; this uncertainty is reflected in the high upper bounds on our estimates (Fig. 1).

Variation in incidence between bacteria and major arthropod host groups

We next tested for differences in symbiont incidence between bacteria and between major host groups. Figure 2 compares estimates of incidence for the best-sampled subphyla of arthropods, namely Hexapoda (insects and relatives), and Chelicerata (represented solely by arachnids in our database), after applying our standardised sampling approach. Results show no significant difference in the incidences of *Wolbachia* in hexapods (51%) versus chelicerates (61%), and no significant difference in the incidences of the three bacteria in chelicerates (*Wolbachia* 61%, *Rickettsia* 51% and *Cardinium* 60%). As such, the clearest pattern in our data is significantly lower incidences, in hexapod hosts, of *Rickettsia* (22%) and especially *Cardinium* (8%) [17]. (We note that results are quite different when standardised sampling is not applied, confirming the benefits of this approach; Figure S4).

The pattern in Fig. 2 might be explained in many ways, but one possibility is differences in arthropod innate immunity [35-37]. Comparative genomics has shown that chelicerates lack key components of the IMD immune pathway [37], which is primarily responsive to gram-negative bacteria [38], and activated by DAP-type peptidoglycans [39]. Crucially, peptidoglycans are not thought to be produced by *Wolbachia* [40,41], but are produced by *Cardinium* [40,42], and also by *Rickettsia*, albeit sometimes at very low levels [43,44]. Therefore, the low incidence in hexapods of *Rickettsia* and especially *Cardinium*, might be due to their eliciting an additional immune response, not found in chelicerates, and not induced by *Wolbachia* in any host group. Suggestive support for this hypothesis comes from estimates of incidence within the hexapod host orders, where the paraneopteran orders Hemiptera (true bugs), and Psocodea (lice) are also known to lack components of the IMD pathway [36]. Figure 3 shows that for *Cardinium*, the six arthropod groups with the highest estimated incidence (the five sampled chelicerate groups, and Hemiptera) all lack IMD components. This pattern is weakly present in *Rickettsia* (where it applies the three groups with highest incidence), and wholly absent in *Wolbachia*.

Regardless of its cause, Figure 3 suggests that closely-related groups of host might have similar levels of symbiont incidence. This is borne out in formal tests, where *Cardinium* and *Rickettsia*, but not *Wolbachia* show weak evidence of phylogenetic signal in their incidence levels (see supplementary methods section 5.1 and Table S4 for full details).

Species richness and symbiont incidence

We next tested the prediction that infection levels in a host group will tend to increase with its species richness [21-24]; this is best tested with many taxonomic groups of similar age, and a rough biological similarity [14,16,21], and so we considered arthropod families or genera within major orders. For the best sampled orders, we asked whether higher symbiont incidence is found in families containing more species. We did this by fitting the linear model $\hat{y}_i = a + b\log_{10}(S_i)$ where S_i is the number of described species in family i, and $y_i \equiv \ln(x_{c,i}/(1 - x_{c,i}))$ is the logit transformed incidence for that family. This model was fit directly to the likelihood surface of the $x_{c,i}$, and so all of the uncertainty in our incidence estimates was taken into account (see supplementary information section 5.2 for further details).

As with Figures 1-2, we first considered a species "infected" if more than 1/1000 individuals harboured the bacteria (i.e., we used $x_{0.001}$ as our response variable), but it is unlikely that, say, a speciation event would be caused by a very low prevalence infection, and so we also repeated all analyses considering only host species infected at prevalences greater than 50% and 90% (i.e., using $x_{0.5}$ and $x_{0.9}$ as response variables). All results are shown in Table S5 and some illustrative cases are plotted in Figure 4.

Considered together, no clear pattern emerges from the results. For example, in Coleoptera (beetles), a significant positive relationship between incidence and species richness is found for both *Wolbachia* and *Rickettsia* (Fig. 4a). However, Araneae (spiders) show the opposite result - a significant tendency for higher *Wolbachia* incidence in species-poor groups (Fig. 4c-d). Furthermore, the explanatory power of the model is very low in almost all cases (only with *Wolbachia* infections at >50% prevalence in Coleoptera does the model yield a pseudo-r^2 above 10%).

Discussion

Terrestrial arthropods cannot be understood without considering their bacterial symbionts, and one key to understanding symbiont biology is to explain why particular symbionts are present or absent in particular groups of potential hosts [e.g., 9,13,17,19]. Following Hilgenboecker *et al.* [11], we have introduced a maximum likelihood estimator of symbiont incidence (the proportion of potential host species that are actually infected), and applied our estimator to a large database of PCR screens for *Wolbachia*, *Rickettsia* and *Cardinium*. We have also introduced methods to account for the most serious sources of sampling bias in our data, including weighting estimates from different arthropod groups by their contribution to total arthropod diversity. Of course, biases will remain (nobody could hope to obtain a truly random sample of all arthropod species) and it remains practically impossible to prove the absence of a symbiont in a given species; for example, infections might have been missed due to primer issues [11,45] or PCR inhibitors, particularly with older methods of DNA extraction [46]. Most seriously, our estimates will be reliable only if prevalences in the sampled populations are representative of the species range as a whole. When only a tiny proportion of the species range has been sampled, it is impossible to know if the sample is representative. For while migration between populations and species-specific susceptibilities will act to homogenise prevalences across populations, geographical isolation, habitat variation, or intraspecific genetic variation, could lead to large, sustained differences in prevalence across a species range. In the worst case scenario, prevalences across populations of a given species would be completely uncorrelated with each other. In such a case, the x_c that we estimate would correspond, roughly, to the incidence across populations, while the incidence across species would tend towards 100% (since, with uncorrelated prevalences, it becomes extremely unlikely for all populations of a given species to be free from infection at any given time). Reality must lie somewhere between these two extremes, and so our incidence estimates are probably downwardly biased (see also Supp. Figure S3).

With these caveats, we have estimated that *Wolbachia*, *Rickettsia* and *Cardinium* infect, respectively, around a half, a quarter and an eighth of terrestrial arthropod species (Fig. 1). These differences mask remarkably similar incidences in chelicerates (Fig. 2), and stem from the significantly lower incidences of *Rickettsia* and especially *Cardinium* in hexapod hosts (Fig. 2; see also [17,28]). This results in the incidence of *Cardinium* and *Rickettsia* - though not *Wolbachia* - showing weak phylogenetic signal across host orders (Table S4). We have speculated that this pattern might reflect evolutionary changes in the arthropod immune system [35-37], which strongly affect *Cardinium*, but have no effect on *Wolbachia*. When transferred to a novel host species, all three bacteria can induce an immune response [38,47], but they probably do so in different ways [40]. We have noted that many arthropod groups lack key components of the IMD pathway, which is activated by DAP-type peptidoglycan, a common components of the gram-negative cell wall [36,37]. We have also noted that the incidence of *Cardinium*, which produces peptidoglycan [40,42], is highest in those host groups, while *Wolbachia*, which does not produce peptidoglycan [40,41], shows no such pattern.

Finally, we tested the prediction that incidence levels would be higher in host groups that are more speciose (Table S5; Fig. 4). This prediction follows from (i) the observation that horizontally-transferred parasites often establish more easily on novel hosts that are closely related to their existing hosts [14,16,48] which implies that speciose host groups contain more sources of potential infection, leading to higher incidence [23,24], and (ii) suggestions that symbionts might cause speciation in their hosts [20-22]. Data from some groups, such as Coleoptera, supported the prediction, but across all host groups, the correlations were inconsistent and generally weak (Table S5; Fig. 4).

There are three possible explanations of these negative results. First, there is the limited power of our tests: inconsistent sampling, screening errors, and reliance on described species numbers, might all have made a true correlation difficult to detect. Furthermore, a general conclusion of this study is that symbiont incidence is difficult to estimate with high precision - even with very large samples (e.g., Figs. 3-4). Nevertheless, several of the data sets did yield significant results - but not consistently in the predicted direction (Table S5).

Second, a confounding factor might have masked a true underlying correlation. Several alternatives are possible. For example, competitive exclusion among symbionts might lead to high incidence of one bacterium being predictably associated with low incidence of another (though see, e.g., [25,49]), or species richness might correlate with clade age, which might also affect symbiont incidence [50]. Alternatively, symbionts might induce speciation without transferring to the new daughter species, or parasitic symbionts might drive their hosts extinct [e.g., 51,52], creating a negative correlation between species richness and infection.

Third, and finally, there might be no causal relationship between species number and symbiont incidence. Regarding host shifting, there is evidence of between-species transfer in all three bacteria, both from phylogenetic incongruence [3,9,53-56] and, in *Wolbachia*, from experimental transfers [45,57]. There is also evidence that transfer success increases with host relatedness - but this evidence is largely indirect, coming from phylogenetic clustering [9,14,16,27,53-55], and strong experimental evidence comes solely from *Spiroplasma*, an ecologically similar, but phylogenetically distant endosymbiont [48]. The evidence for symbiont-mediated speciation is even sparser. Host reproductive isolation might arise as a passive byproduct of host-symbiont coevolution (since any genomic change, whether in host or symbiont, might have negative epistatic fitness effects in a hybrid background [4,22]). But the most plausible route to rapid speciation is through reproductive manipulations which cause reproductive isolation, such as cytoplasmic incompatibility, or host parthenogenesis. Not all manipulations have been observed in all host-parasite combinations, and so this might explain the inconsistency in our results (Table S5). Nevertheless, taken together, our results must count as evidence against the claim that symbionts are a major cause of diversification across the arthropods as a whole.

But while no consistent effect of species-richness has been found, many further hypotheses about the causes of endosymbiont incidence in nature remain to be tested, and we hope that the methods presented here will prove useful for this purpose.

References

1. O'Neill, S. L., Hoffman, A. & Werren, J. H., editors 1997 *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction.* Oxford: Oxford University Press.

2. Moran, N. A. 2006 Symbiosis. *Current Biology* **16**, R866-R871. (doi:10.1016/j.cub.2006.09.019)

3. Werren, J. H., Baldo, L. & Clark, M. E. 2008 *Wolbachia*: Master Manipulators of Invertebrate Biology. *Nature Reviews Microbiology* **6**, 741-751. (doi:10.1038/nrmicro1969)

4. Zilber-Rosenberg, I. & Rosenberg, E. 2008 Role of Microorganisms in the Evolution of Animals

and Plants: The Hologenome Theory of Evolution. *FEMS Microbiology Reviews* **32**, 723-735. (doi:10.1111/j.1574-6976.2008.00123.x)

5. Engelstädter, J. & Hurst, G. D. D. 2009 The Ecology and Evolution of Microbes that Manipulate Host Reproduction. *Annual Review of Ecology, Evolution, and Systematics* **40**, 127-149. (doi:10.1146/annurev.ecolsys.110308.120206)

6. Sharon, G., Segal, D., Ringo, J. M., Hefetz, A., Zilber-Rosenberg, I. & Rosenberg, E. 2010 Commensal Bacteria Play a Role in Mating Preference of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences* **107**, 20051-20056. (doi:10.1073/pnas.1009906107)

7. Zchori-Fein, E. & Bourtzis, K., editors 2012 *Manipulative tenants: bacteria associated with arthropods*. Boca Raton: Taylor & Francis.

8. Henry, L. M., Peccoud, J., Simon, J.-C., Hadfield, J. D., Maiden, M. J. C., Ferrari, J. & Godfray, H. C. J. 2013 Horizontally Transmitted Symbionts and Host Colonization of Ecological Niches. *Current Biology* **23**, 1713-1717. (doi:10.1016/j.cub.2013.07.029)

9. Weinert, L. A. 2014 The Diversity and Phylogeny of *Rickettsia*. In *Parasite Diversity and Diversification: Evolutionary Ecology Meets Phylogenetics* (eds S. Morand B. Krasnov & T. Littlewood), Cambridge University Press.

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

11. Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A. & Werren, J. H. 2008 How Many Species Are Infected with *Wolbachia*? A Statistical Analysis of Current Data. *FEMS Microbiology Letters* **281**, 215-220. (doi:10.1111/j.1574-6968.2008.01110.x)

12. Hurst, L. D. 1991 The Incidences and Evolution of Cytoplasmic Male Killers. *Proc. R. Soc. Lond. B* **244**, 91-99. (doi:10.1098/rspb.1991.0056)

13. Werren, J. H., Windsor, D. & Guo, L. 1995 Distribution of *Wolbachia* among Neotropical Arthropods. *Proc. R. Soc. Lond. B* **262**, 197-204. (doi:10.1098/rspb.1995.0196)

14. Jiggins, F. M., Randerson, J. P., Hurst, G. D. D. & Majerus, M. E. N. 2002 How can sex ratio distorters reach extreme prevalences? Male-killing *Wolbachia* are not suppressed and have near-perfect vertical transmission efficiency in *Acraea encedon*. *Evolution* **56**, 2290-2295.

15. Reuter, M., Pedersen, J. S. & Keller, L. 2004 Loss of *Wolbachia* infection during colonisation in the invasive Argentine ant *Linepithema humile*. *Heredity* **94**, 364-369. (doi:10.1038/sj.hdy.6800601)

16. Baldo, L., Ayoub, N. A., Hayashi, C. Y., Russell, J. A., Stahlhut, J. K. & Werren, J. H. 2008 Insight into the routes of *Wolbachia* invasion: high levels of horizontal transfer in the spider genus *Agelenopsis* revealed by *Wolbachia* strain and mitochondrial DNA diversity. *Molecular Ecology* **17**, 557-569. (doi:10.1111/j.1365-294X.2007.03608.x)

17. Martin, O. Y. & Goodacre, S. L. 2009 Widespread Infections by the Bacterial Endosymbiont *Cardinium* in Arachnids. *Journal of Arachnology* **37**, 106-108. (doi:10.1636/SH08-05.1)

18. Toju, H. & Fukatsu, T. 2011 Diversity and Infection Prevalence of Endosymbionts in Natural Populations of the Chestnut Weevil: Relevance of Local Climate and Host Plants. *Molecular Ecology* **20**, 853-868. (doi:10.1111/j.1365-294X.2010.04980.x)

19. Russell, J. A., Funaro, C. F., Giraldo, Y. M., Goldman-Huertas, B., Suh, D., Kronauer, D. J. C., Moreau, C. S. & Pierce, N. E. 2012 A Veritable Menagerie of Heritable Bacteria from Ants, Butterflies, and Beyond: Broad Molecular Surveys and a Systematic Review. *PLoS ONE* **7**, e51027. (doi:10.1371/journal.pone.0051027)

20. Wallin, I. E. 1927 *Symbionticism and the origin of species*. Baltimore : Williams & Wilkins Company.

21. Bordenstein, S. R. 2003 Symbiosis and the origin of species. In *Insect Symbiosis* (eds K. Bourtzis & T. A. Miller), New York: CRC Press.

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

23. Waxman, D., Weinert, L. A. & Welch, J. J. 2014 Inferring Host Range Dynamics from Comparative Data: The Protozoan Parasites of New World Monkeys. *The American Naturalist* (doi:10.1086/676589)

24. Engelstädter, J. & Hurst, G. D. D. 2006 The Dynamics of Parasite Incidence Across Host Species. *Evol Ecol* **20**, 603-616. (doi:10.1007/s10682-006-9120-1)

25. Weinert, L. A., Tinsley, M. C., Temperley, M. & Jiggins, F. M. 2007 Are We Underestimating the Diversity and Incidence of Insect Bacterial Symbionts? A Case Study in Ladybird Beetles. *Biology Letters* **3**, 678-681. (doi:10.1098/rsbl.2007.0373)

26. Weeks, A. R., Turelli, M., Harcombe, W. R., Reynolds, K. T. & Hoffmann, A. A. 2007 From Parasite to Mutualist: Rapid Evolution of *Wolbachia* in Natural Populations of *Drosophila*. *PLoS Biol* **5**, e114. (doi:10.1371/journal.pbio.0050114)

27. Himler, A. G., Adachi-Hagimori, T., Bergen, J. E., Kozuch, A., Kelly, S. E., Tabashnik, B. E., Chiel, E., Duckworth, V. E., Dennehy, T. J., Zchori-Fein, E. *et al.* 2011 Rapid Spread of a Bacterial Symbiont in an Invasive Whitefly Is Driven by Fitness Benefits and Female Bias. *Science* **332**, 254-256. (doi:10.1126/science.1199410)

28. Werren, J. H. & Windsor, D. M. 2000 *Wolbachia* Infection Frequencies in Insects: Evidence of a Global Equilibrium? *Proceedings of the Royal Society B: Biological Sciences* **267**, 1277-1285. (doi:10.1098/rspb.2000.1139)

29. Ahmed, M. Z., Greyvenstein, O. F. C., Erasmus, C., Welch, J. J. & Greeff, J. M. 2013 Consistently High Incidence of *Wolbachia* in Global Fig Wasp Communities. *Ecol Entomol* **38**, 147-154. (doi:10.1111/een.12002)

30. Zug, R. & Hammerstein, P. 2012 Still a Host of Hosts for *Wolbachia*: Analysis of Recent Data Suggests That 40% of Terrestrial Arthropod Species Are Infected. *PLoS ONE* **7**, e38544. (doi:10.1371/journal.pone

31. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. 2011 How Many Species Are There on Earth and in the Ocean? *PLoS Biology* **9**, e1001127. (doi:10.1371/journal.pbio.1001127)

32. Fine, P. E. M. 1978 On the Dynamics of Symbiote-Dependent Cytoplasmic Incompatibility in Culicine Mosquitoes. *Journal of Invertebrate Pathology* **31**, 10-18. (doi:10.1016/0022-2011(78)90102-7)

33. Woolhouse, M. E. J., Haydon, D. T. & Antia, R. 2005 Emerging Pathogens: The Epidemiology and Evolution of Species Jumps. *Trends in Ecology & Evolution* **20**, 238-244. (doi:10.1016/j.tree.2005.02.009)

34. Ospina, R. & Ferrari, S. L. P. 2010 Inflated Beta Distributions. *Statistical Papers* **51**, 111-126. (doi:10.1007/s00362-008-0125-4)

35. Waterhouse, R. M., Kriventseva, E. V., Meister, S., Xi, Z., Alvarez, K. S., Bartholomay, L. C., Barillas-Mury, C., Bian, G., Blandin, S., Christensen, B. M. *et al.* 2007 Evolutionary Dynamics of Immune-Related Genes and Pathways in Disease-Vector Mosquitoes. *Science* **316**, 1738-1743. (doi:10.1126/science.1139862)

36. Gerardo, N. M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S. M., de Vos, M., Duncan, E. J., Evans, J. D., Gabaldón, T., Ghanim, M. *et al.* 2010 Immunity and Other Defenses in Pea Aphids, *Acyrthosiphon pisum*. *Genome Biology* **11**, R21. (doi:10.1186/gb-2010-11-2-r21)

37. Palmer, W. J. 2014 Personal Communication / Submitted Article.

38. Harris, H. L., Brennan, L. J., Keddie, B. A. & Braig, H. R. 2010 Bacterial Symbionts in Insects: Balancing Life and Death. *Symbiosis* **51**, 37-53. (doi:10.1007/s13199-010-0065-3)

39. Kaneko, T., Goldman, W. E., Mellroth, P., Steiner, H., Fukase, K., Kusumoto, S., Harley, W., Fox, A., Golenbock, D. & Silverman, N. 2004 Monomeric and Polymeric Gram-Negative Peptidoglycan but Not Purified LPS Stimulate the *Drosophila* IMD Pathway. *Immunity* **20**, 637-649. (doi:10.1016/S1074-7613(04)00104-9)

40. Nakamura, Y., Gotoh, T., Imanishi, S., Mita, K., Kurtti, T. J. & Noda, H. 2011 Differentially Expressed Genes in Silkworm Cell Cultures in Response to Infection by *Wolbachia* and *Cardinium* Endosymbionts. *Insect Molecular Biology* **20**, 279-289. (doi:10.1111/j.1365-2583.2010.01056.x)

41. Vollmer, J., Schiefer, A., Schneider, T., Jülicher, K., Johnston, K. L., Taylor, M. J., Sahl, H.-G., Hoerauf, A. & Pfarr, K. 2013 Requirement of Lipid II Biosynthesis for Cell Division in Cell Wall-less *Wolbachia*, Endobacteria of Arthropods and Filarial Nematodes. *International Journal of Medical Microbiology* **303**, 140-149. (doi:10.1016/j.ijmm.2013.01.002)

42. Penz, T., Schmitz-Esser, S., Kelly, S. E., Cass, B. N., Müller, A., Woyke, T., Malfatti, S. A., Hunter, M. S. & Horn, M. 2012 Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in *Cardinium hertigii*. *PLoS Genetics* **8**, e1003012. (doi:10.1371/journal.pgen.1003012)

43. Amano, K., Tamura, A., Ohashi, N., Urakami, H., Kaya, S. & Fukushi, K. 1987 Deficiency of Peptidoglycan and Lipopolysaccharide Components in *Rickettsia tsutsugamushi*. *Infect. Immun.* **55**, 2290-2292.

44. Pang, H. & Winkler, H. H. 1994 Analysis of the Peptidoglycan of *Rickettsia prowazekii*. *J. Bacteriol.* **176**, 923-926.

45. Schneider, D. I., Riegler, M., Arthofer, W., Merçot, H., Stauffer, C. & Miller, W. J. 2013 Uncovering *Wolbachia* Diversity upon Artificial Host Transfer. *PLoS ONE* **8**, e82402. (doi:10.1371/journal.pone.008240

46. Beckmann, J. F. & Fallon, A. M. 2012 Decapitation Improves Detection of Wolbachia pipientis (Rickettsiales: Anaplasmataceae) in Culex pipiens (Diptera: Culicidae) Mosquitoes by the Polymerase Chain Reaction. *Journal of Medical Entomology* **49**, 1103-1108. (doi:10.1603/ME12049)

47. Pelc, R. S., McClure, J. C., Sears, K. T., Chung, A., Rahman, M. S. & Ceraul, S. M. 2014 Defending the Fort: a Role for Defensin-2 in Limiting *Rickettsia montanensis* Infection of *Dermacentor variabilis*. *Insect Molecular Biology* (doi:10.1111/imb.12094)

48. Tinsley, M. C. & Majerus, M. E. 2007 Small Steps or Giant Leaps for Male-Killers? Phylogenetic Constraints to Male-Killer Host Shifts. *BMC Evolutionary Biology* **7**, 238. (doi:10.1186/1471-2148-7-238)

49. Duron, O., Bouchon, D., Boutin, S., Bellamy, L., Zhou, L., Engelstadter, J. & Hurst, G. D. 2008 The Diversity of Reproductive Parasites Among Arthropods: *Wolbachia* Do Not Walk Alone. *BMC Biology* **6**, 27. (doi:10.1186/1741-7007-6-27)

50. Zug, R., Koehncke, A. & Hammerstein, P. 2012. Epidemiology in Evolutionary Time: the Case of *Wolbachia* Horizontral Transmission Between Arthropod Host Species. *Journal of Evolutionary Biology* **25**, 2149-2160. (doi:10.1111/j.1420-9101.2012.02601)

51. Anderson, R. M. 1978 The regulation of host population growth by parasitic species. *Parasitology* **76**, 119-157. (doi:10.1017/S0031182000047739)

52. Nice, C. C., Gompert, Z., Forister, M. L. & Fordyce, J. A. 2009 An Unseen Foe in Arthropod Conservation Efforts: The Case of *Wolbachia* Infections in the Karner Blue Butterfly. *Biological Conservation* **142**, 3137-3146. (doi:10.1016/j.biocon.2009.08.020)

53. Russell, J. A., Goldman-Huertas, B., Moreau, C. S., Baldo, L., Stahlhut, J. K., Werren, J. H. & Pierce, N. E. 2009 Specialization and Geographic Isolation Among Wolbachia Symbionts from Ants and Lycaenid Butterflies. *Evolution* **63**, 624-640. (doi:10.1111/j.1558-5646.2008.00579.x)

54. Weinert, L. A., Werren, J. H., Aebi, A., Stone, G. N. & Jiggins, F. M. 2009 Evolution and diversity of *Rickettsia* bacteria. *BMC Biology* **7**, 6. (doi:10.1186/1741-7007-7-6)

55. Perlman, S. J., Magnus, S. A. & Copley, C. R. 2010 Pervasive Associations Between *Cybaeus* Spiders and the Bacterial Symbiont *Cardinium*. *Journal of Invertebrate Pathology* **103**, 150-155. (doi:10.1016/j.jip.2009.12.009)

56. Ros, V. I. D., Fleming, V. M., Feil, E. J. & Breeuwer, J. A. J. 2012 Diversity and recombination in *Wolbachia* and *Cardinium* from *Bryobia* spider mites. *BMC Microbiology* **12**, S13. (doi:10.1186/1471-2180-12-S1-S13)

57. Carrington, L. B., Hoffmann, A. A. & Weeks, A. R. 2010 Monitoring Long-term Evolutionary Changes Following *Wolbachia* Introduction into a Novel Host: the *Wolbachia* Popcorn Infection in *Drosophila simulans*. *Proceedings of the Royal Society B: Biological Sciences* **277**, 2059-2068. (doi:10.1098/rspb.2010.0166)

Acknowledgements

We are very grateful to J. Beckmann, F. Jiggins, B. Longdon, J. Martinez, W. Palmer, A. Rambaut, J. Thomas, D. Waxman and J. Werren for their help and advice, and to two anonymous reviewers whose comments substantially improved the manuscript.

Figure Captions

Figure 1

Estimates of symbiont incidence, $x_{0.001}$ (i.e., the proportion of species infected at a prevalence of greater than 1/1000) in terrestrial arthropods. Estimates obtained from (a) fitting a beta distribution to the complete database; (b) fitting a doubly-inflated beta distribution to the complete database, and so allowing for completely uninfected or completely infected species; (c) standardised sampling (i.e., a weighted sum of estimates from the largest arthropod taxa, using the single largest population sample from each sampled species within each taxon).

Figure 2

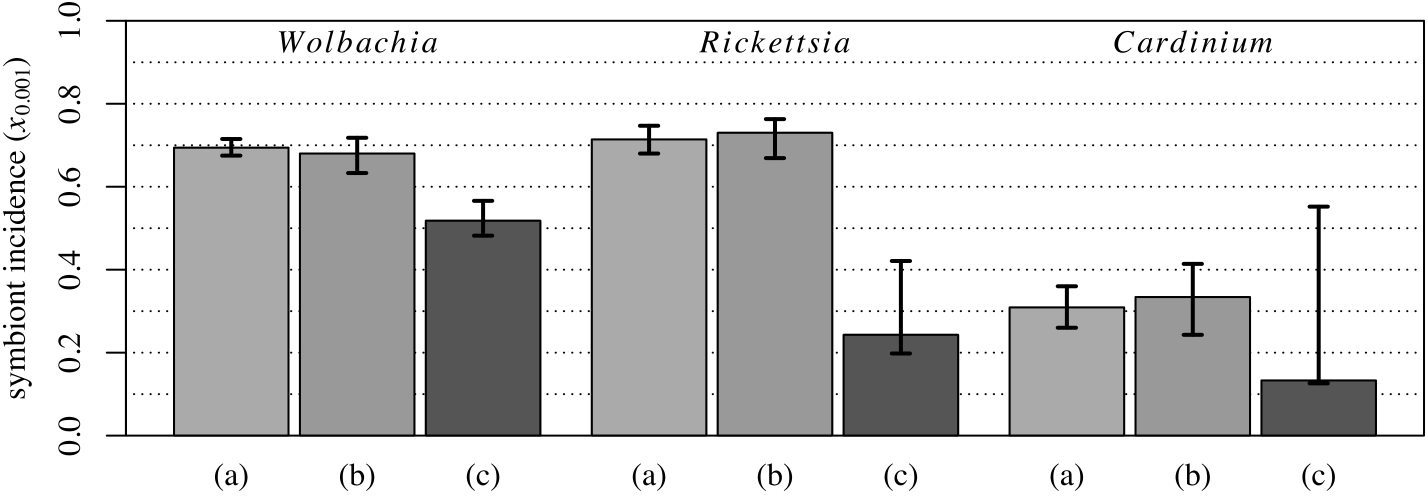
Estimates of symbiont incidence, $x_{0.001}$ (i.e., the proportion of species infected at a prevalence of greater than 1/1000) in the two major subphyla of arthropoda. Each pair of bars shows the incidence of a different bacterial genus, and compares estimates for Hexapoda (left-hand bar) and Chelicerata (right-hand bar). Estimates used “standardised sampling” (see main text). p-values above each set of bars are from a Likelihood Ratio Tests of heterogeneity in the estimates.

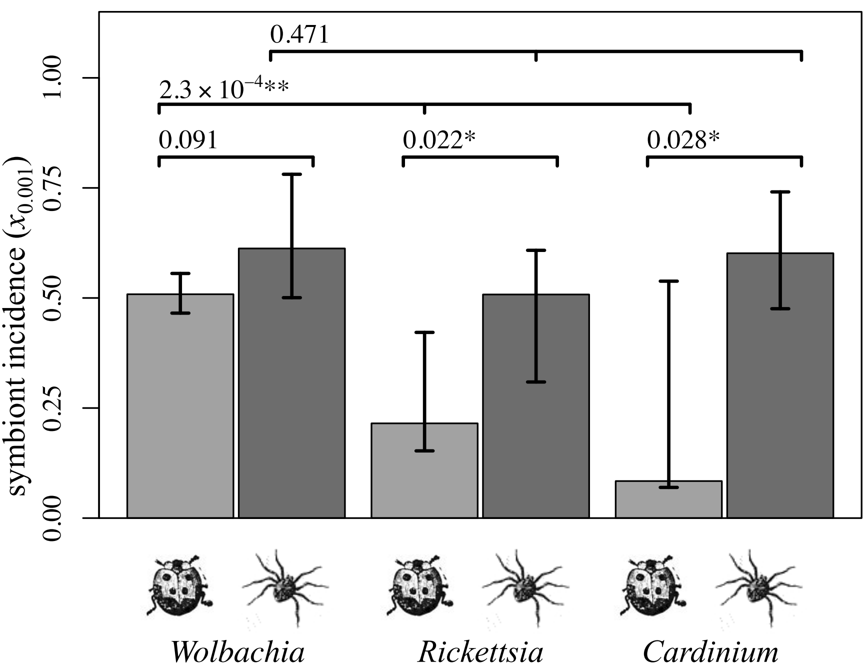
Figure 3

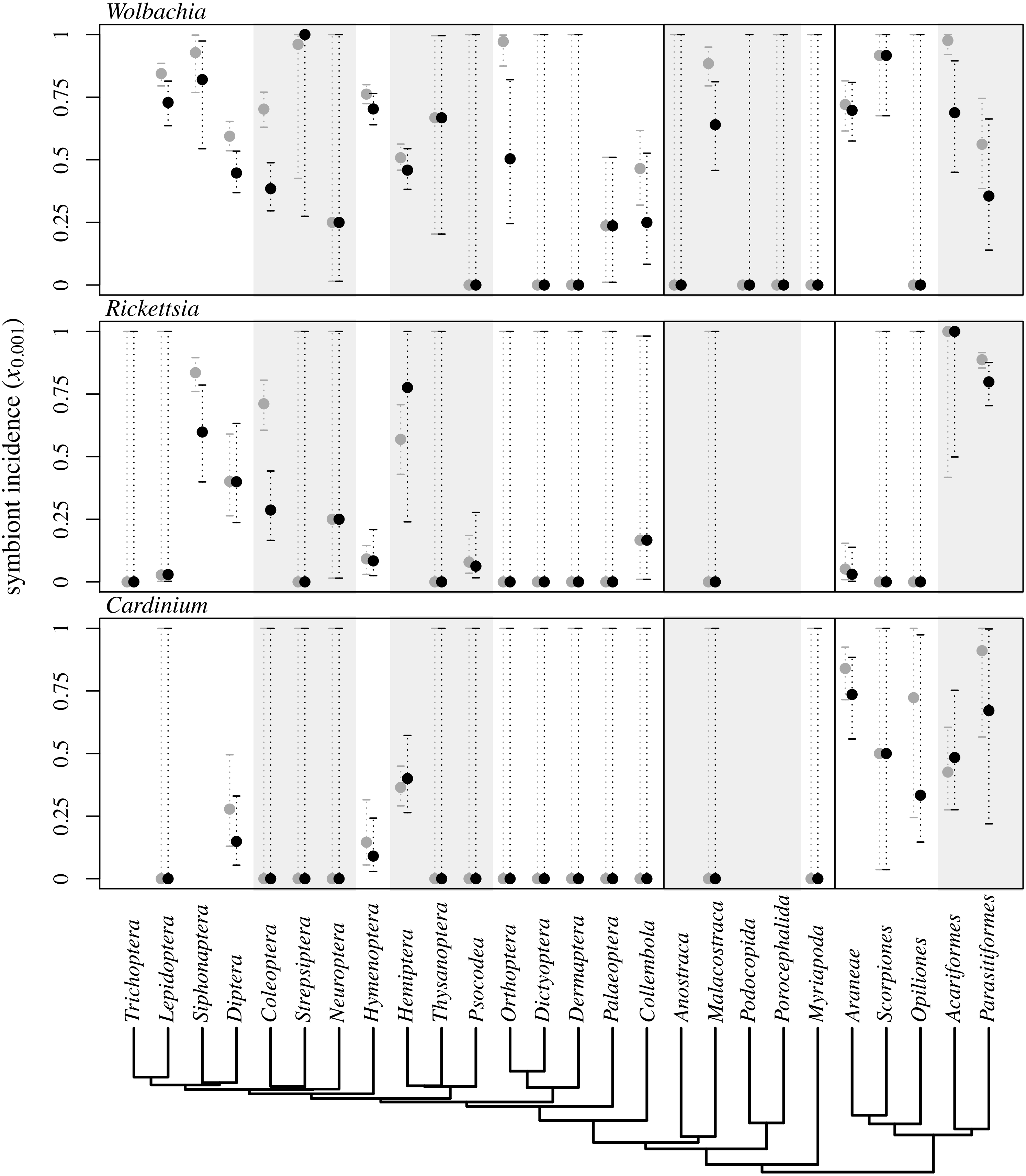
Estimates of symbiont incidence, $x_{0.001}$ (i.e., the proportion of species infected at a prevalence of greater than 1/1000) for three genera of bacterial endosymbionts, across orders (and some superordinal groups) of terrestrial arthropods. Grey points show estimates from our complete database, and black points show estimates with standardised sampling, in which all sampled species in each group were represented by the single largest population sample. Shading and vertical lines demarcate some major host groups, including Hexapoda (left-hand panel) and Chelicerata (right-hand panel).

Figure 4

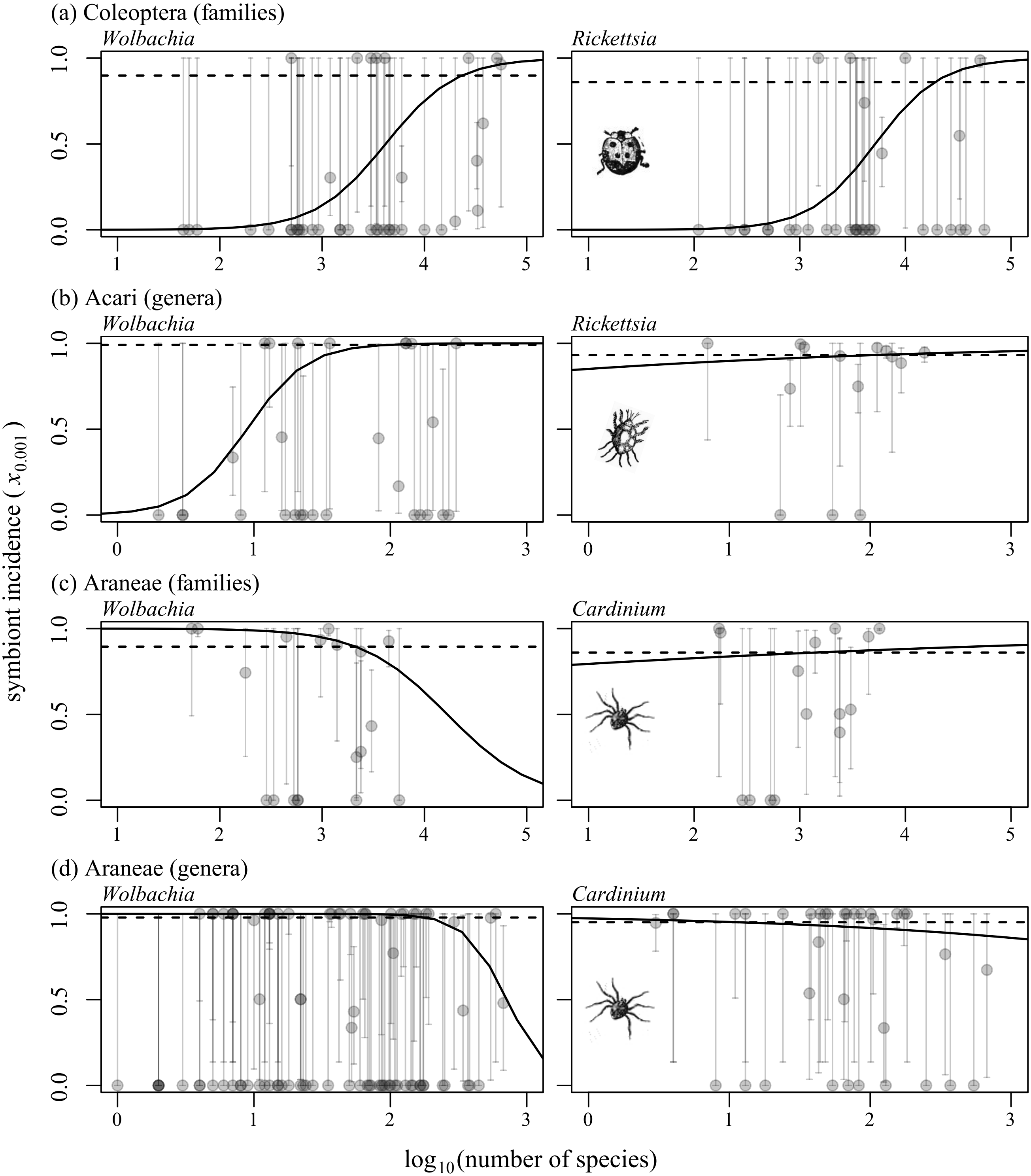
Estimated incidence of bacterial endosymbionts for individual families of terrestrial arthropods, plotted against the number of described species in that family. Each point represents the estimated proportion of populations in a single family infected at a prevalence of greater than 1/1000 ($x_{0.001}$). Solid lines show the best-fit line linking symbiont incidence and host species richness (see main text), while the dashed lines show the best-fitting null model (in which all families have the same expected incidence). Results are shown only for host groups that were well sampled for two bacteria.


symbiont incidence ($x_{0.001}$)
1.0
0.8
0.6
0.4
0.2
0.0
Wolbachia
Rickettsia
Cardinium
(a)
(b)
(c)
(a)
(b)
(c)
(a)
(b)
(c)


symbiont incidence ($x_{0.001}$)
0.00
0.25
0.50
0.75
1.00
0.471
2.3×10^{-4}**
0.091
0.022*
0.028*
Wolbachia
Rickettsia
Cardinium



Wolbachia
Rickettsia
Cardinium
symbiont incidence ($x_{0.001}$)
1
0.75
0.5
0.25
0
Trichoptera
Lepidoptera
Siphonaptera
Diptera
Coleoptera
Strepsiptera
Neuroptera
Hymenoptera
Hemiptera
Thysanoptera
Psocodea
Orthoptera
Dictyoptera
Dermaptera
Palaeoptera
Collembola
Anostraca
Malacostraca
Podocopida
Porocephalida
Myriapoda
Araneae
Scorpiones
Opiliones
Acariformes
Parasitiformes



(a) Coleoptera (families)
Wolbachia
Rickettsia
(b) Acari (genera)
Wolbachia
Rickettsia
(c) Araneae (families)
Wolbachia
Cardinium
(d) Araneae (genera)
Wolbachia
Cardinium
symbiont incidence ($x_{0.001}$)
$\log_{10}$(number of species)
0.0
0.5
1.0
0
1
2
3
4
5

Supplementary Methods for:
The incidence of bacterial endosymbionts in terrestrial arthropods

Lucy A. Weinert, Eli V. Araujo-Jnr, Muhammad Z. Ahmed, John J. Welch

correspondence to: j.j.welch@gen.cam.ac.uk

1 Database collation

Each entry in our database comprised data from a single arthropod population. In general, it was not always possible to use a single consistent definition of a "population", but where possible, we split the data by sampling location, date of collection and host subspecies, while for vertebrate-associated arthropods, we treated samples from different vertebrate host species in the same geographical location as belonging to different populations. For most studies, arthropod individuals were screened individually, but screens of multi-individual pools were also included, as these too can inform estimates [e.g., 1; see below], but we excluded any population where pool size was variable, or unreported. In addition, we excluded any population that had been kept in a laboratory for more than twelve months, or where individuals were screened long after death, unless stored in solution or frozen immediately after death. Source publications used a variety of primers and protocols, but we did exclude studies using long PCR - which is highly sensitive to very low titre infections, but might yield a high rate of false positives [2,3].

During the collation of the database, many authors provided important clarifications or additional data, and we are very grateful to all of the following: A. N. Alekseev, C. S. Apperson, H.-N. Chai, G. A. Dasch, Y.-Z. Du, M. Eremeeva, K. D. Floate, N. Guz, S. Hornok, L. Hun, M.-X. Jiang, T. Kurtti, M. L. Levin, Z. Lijuan, J. H. McQuiston, O. Mediannikov, C. S. Moreau, N. Nakamura, M.-M. Nogueras, J. A. Oteo Revuelta, Y. Peng, A. Portillo, R. Rajagopal, A. Richards, Y. Sakamoto, P. Shimabukuro, P.-Y. Shu, C. Silaghi, M. Škaljac, C. Strube, L. Tomassone, A. Troyo, K.-H. Tsai, J. Walochnik, M. Wijnveld, and K. Wilson.

2 Likelihood function and numerical methods

2.1 The likelihood function

We estimate symbiont incidence by first inferring the distribution of prevalence values across arthropod populations. Let us first assume that the true prevalence of bacterial infection in a single population is q, where $0 \leq q \leq 1$, and that we are estimating this prevalence by screening n pools, each containing m randomly-sampled individuals - and thereby screening nm individuals in total (for data sets where each arthropod was individually screened, we simply set $m = 1$). In this case, the probability that a given pool will be free of infection is $(1-q)^m$, and the probability of observing k infected pools is

$$p(k;n,m,q) = \binom{n}{k}(1-(1-q)^m)^k(1-q)^{m(n-k)} \tag{1}$$

We must now make some assumptions about the distribution of prevalences [3]. Initially, we assume that the across-population distribution of prevalences can be adequately described by a beta distribution

$$P(q;\alpha,\beta) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)} \tag{2}$$

In eq. (2), $\alpha,\beta > 0$ are shape parameters, and $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function.

The likelihood of observing our data can now be derived by combining eqs. (1) and (2).

$$\begin{aligned} L(k,n,m;\alpha,\beta) &= \int\limits_{q=0}^{1} P(q;\alpha,\beta)p(k;n,m,q)dq & (3) \\ &= \frac{\binom{n}{k}}{B(\alpha,\beta)} \sum_{i=0}^{k} (-1)^i \binom{k}{i} B(\alpha,\beta+m(n-k+i)) & (4) \end{aligned}$$

The complete log likelihood function follows from including screens from many different populations:

$$\ln L = \sum_{i}^{pops.} \ln L(k_i,n_i,m_i;\alpha,\beta) \tag{5}$$

The maximum likelihood estimates of the parameters α and β are the values that maximise this function, and likelihood-based confidence intervals (as reported in the main text) are the values that reduce this maximised log likelihood by two units [4]. For our major datasets, we also produced confidence intervals by bootstrapping the data, and these were nearly identical to the likelihood-based intervals (not shown).

2.2 Meaningful parameterisation

The shape parameters in eq. (2) are not readily interpretable in biological terms, but the distribution can be written in terms of two alternative parameters, namely the mean prevalence (denoted $\bar{q}$), and the proportion of the total variance in infection status that is distributed between arthropod populations, as opposed to within populations (denoted F).

$$\bar{q} \equiv E[q] \tag{6}$$

$$F \equiv \frac{\mathrm{Var}[q]}{\bar{q}(1-\bar{q})} \tag{7}$$

For the beta distribution, these more meaningful parameters can be derived via

$$\bar{q} = \alpha/(\alpha+\beta) \tag{8}$$

$$F \equiv 1/(1+\alpha+\beta) \tag{9}$$

We note that the parameter F is defined by analogy with Wright's F_{st} [5]. Its value ranges between $F = 0$, when all populations have the same prevalence, and $F = 1$ when there is no variation in infection status within populations, such that each population is either completely infected, or completely uninfected. As such, it is also be defined as the correlation in infection probability among members of the same population. Given this definition, the value of F is undefined if all populations are free from infection ($\bar{q} = 0$) or if all populations are completely infected ($\bar{q} = 1$). Similarly, the parameter is not identifiable for data sets without multi-individual screens (i.e., when all $n_i = 1$). For this reason, to generate bootstrap confidence intervals on parameters, we sampled single- and multi-individual screens separately. We also set the maximum likelihood value of F at $\hat{F} = 1$ for data sets that contained no infected individuals (i.e., for which $\hat{\bar{q}} = 0$). This is because $\hat{F} = 1$ maximizes the likelihood with when $\bar{q} > 0$ for any data set that contains no partially infected samples.

Finally, following Hilgenboecker *et al.* [3], we define the incidence as the proportion of populations that are infected above a certain threshold prevalence, $0 \leq c \leq 1$. This is found from:

$$\begin{aligned} x_c \equiv \Pr(q > c) &= \int_{q=c}^{1} P(q;\alpha,\beta)dq \\ &\approx 1 - \frac{c^{\alpha}}{\alpha B(\alpha,\beta)}, \quad c \ll 1 \end{aligned} \tag{10}$$

Confidence intervals on these compound parameters can also be generated as described above [4].

2.3 Numerical methods

In general, either eq. (3) or eq. (4) can be used to calculate the likelihood. However, there are also simplifications and transformations that can be used in some regions of parameter space. First, and most importantly, when arthropods were screened individually, solving the integral in eq. (3) with $m = 1$, shows that the summation in eq. (4) simplifies, as in standard beta-binomial modelling [3].

$$\sum_{i=0}^{k}(-1)^i \binom{k}{i} B(\alpha, \beta+n-k+i) = B(\alpha+k, \beta+n-k) \tag{11}$$

There are also further simplications that arise in data sets without partially infected samples, e.g., when F is not defined (see above). Finally, in some parameter combinations (e.g., when $n = k$, $m > 1$ and $\beta < 1$), both eqs. (3) and (4) can become numerically unstable. In such regimes, we used an exponential transformation of eq. (3) [6,7]. This transformation is as follows:

$$\int_0^1 f(q)dq = \int_{-\infty}^{\infty} f(\varphi(t))\varphi'(t)dt \tag{12}$$

where

$$\begin{aligned} \varphi(t) &\equiv \frac{1}{2}[\tanh(\pi \sinh(t)/2)+1] \qquad (13) \\ \varphi'(t) &= \frac{\pi \cosh(t)}{4\cosh^2(\pi \sinh(t)/2)} \end{aligned}$$

Working with the compound parameters was straightforward for $\bar{q}$ and F, because the likelihood function can be easily rewritten as a function of these parameters. For x_c, we first calculated the likelihood for a fine grid of $\bar{q}$ and F values, and then used the observation that the likelihood surface for x_c was smooth and unimodal. This allowed us to generate this surface using linear interpolation. Computer code to calculate and maximise the likelihood was written in *R* [8] and is included as supplementary information.

3 More complex models and simulations

3.1 Inflated beta distributions

A major limitation of the beta distribution is that, in most cases, it does not allow for a substantial fraction of populations to be completely free of infection (with $q = 0$), or completely infected (with $q = 1$). This is why we had to define a non-zero threshold prevalence, c (eq. (10)) because - given the form of the beta distribution - an infinitesimal fraction of populations will contain exactly no infection, except in the special cases of $\bar{q} = 0$ (i.e., when all populations are infection free), or when $F = 1$ (i.e., when all populations are either completely uninfected or completely infected). In all other cases, therefore, $x_0 = 1$.

An alternative that avoids this limitation is the doubly-inflated beta distribution [9], i.e., a beta distribution combined with two spikes of probability at the extreme values.

$$P(q;\phi,\gamma,\alpha,\beta) \;=\; \begin{cases} \phi(1-\gamma), & \text{if}\, q = 0 \\ \phi\gamma, & \text{if}\, q = 1 \\ (1-\phi)\frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}, & \text{if}\, q \in (0,1) \end{cases} \qquad (14)$$

In eq. (14), $\alpha,\beta > 0$ are the shape parameters, and $1 \geq \phi,\gamma \geq 0$ control the weight of the spikes. The meaningful quantities, $\bar{q}$, F and x_c (eqs. (6), (7) and (10)) can then be derived for this new distribution.

$$\begin{aligned} \bar{q} &= \phi\gamma + \frac{(1-\phi)\alpha}{\alpha+\beta} \\ F &= 1 - \frac{(1-\phi)\alpha\beta}{\bar{q}(1-\bar{q})(\alpha+\beta)(1+\alpha+\beta)} \\ x_c &= \phi\gamma + (1-\phi)\int\limits_{q=c}^{1} P(q;\alpha,\beta)dq \end{aligned} \qquad (15)$$

Note that $x_0 < 1$ is now possible, even when some populations do contain intermediate levels of infection. Furthermore, we can define two new useful parameters that define the proportion of species that are completely uninfected or completely infected.

$$\begin{aligned} p_0 &\equiv \Pr(q=0) = \phi(1-\gamma) \\ p_1 &\equiv \Pr(q=1) = \phi\gamma \end{aligned} \tag{16}$$

3.2 Performance of estimators on simulated data sets

To compare the performance of the two models (eqs. (2) and (14)), and to compare the performance of our maximum likelihood approach to the moment-based estimators of Hilgenboecker *et al.* ([3]; see their eqs. 1-5), we generated a large number of simulated data sets with known parameters, and then reestimated these parameters using the three methods. To generate the simulated data, we assumed that the true distribution of prevalences followed the doubly-inflated beta distribution (eq. (14)), with a range of different parameter values. Each simulated data set contained the same number of screens and individuals as our true *Wolbachia* data (the largest of our data sets), but for each screen, the number of individuals infected was generated by (i) drawing a true prevalence at random from the doubly-inflated beta distribution (eq. (14)), and (ii) drawing the sample prevalence, at random, from a binomial distribution parameterised with the randomly generated prevalence level (eq. (1)). Because the moment-based approach cannot be used with pooled samples ($m > 1$), we removed all screens of pooled samples before carrying out the simulation procedure described above. However, this had little effect on the performance of the likelihood-based methods (not shown).

Figure S2 shows the results of the simulations. Each column of panels (a)-(j) contains results for a different set of true parameters, while each row contains estimates for one of the important parameters (x_c, $\bar{q}$, F), under each of the three methods, with the true values - used to simulate the data - indicated in red. The first message of Figure S2 is that the heuristic moment-based estimators for the beta distribution [3] have a similar level of accuracy to our maximum likelihood approach, but the moment-based estimates are generally less precise (i.e., there is much wider spread of estimates for the same true parameter values); this is a benefit of using the full likelihood approach. Second, Figure S2 shows that fitting a standard beta distribution can yield misleading results when, in reality, a substantial fraction of species are either completely infected, or completely uninfected (see particularly panels (a), (c)-(e)) - i.e., when the doubly-inflated distribution is the true model. In such cases, fitting a doubly-inflated beta distribution does provide a substantial improvement in accuracy. Furthermore, the doubly-inflated distribution shows good performance even when the smaller, beta distribution is the true model.

3.3 Performance of estimators on the real data

Table S1 contains parameter estimates for our major data sets from the beta distribution (a) and the doubly-inflated distribution (b). In both cases, the estimates correspond to the incidence estimates shown in Figure 1 labelled (a) and (b). Table S1 also contains Akaike weights for the two models, i.e., the probability that this distribution, and not the alternative, minimises the information loss [10,11]. Table S1 shows that the doubly-inflated distribution is strongly preferred for two of our three data sets (*Rickettsia* and *Cardinium*). However, as with the incidence estimates (Fig. 1), none of the parameter estimates is substantially changed, and the estimates from the larger model are less precise. Furthermore, the additional parameter p_0 is very imprecisely estimated; for example, for *Cardinium*, we cannot reject the proportion of completely uninfected species being as low as 0% or as high as 75%. Furthermore, simulation results suggest that when the two models give similar parameters estimates, both methods will be reasonably accurate (Fig. S2). These results explain why we continue to report incidence assuming a threshold cutoff of 1/1000 infected individuals, and why we used the simpler beta distribution to calculate estimates for individual host orders (e.g., Fig. 3).

4 Standardised sampling

4.1 Unequal representation of species

Some arthropod species are represented in our database by many populations, and others by only one.

To balance the sampling, we chose to subsample our database, retaining only a single population sample from each species. To determine which sample to retain, we preferred samples with larger numbers of pools (larger n_i), and in the case of ties (equal numbers of pools), samples with larger pool sizes (larger m_i). For our database, in almost every case, we were not forced to choose between screens with identical numbers of pools and pool sizes but unequal sample prevalences, and so the subsampling involved no random choice. The sole exception was the Acariformes ("true mites"), in which there were very few screens for *Rickettsia*, and so we retained all of these data, but merged the samples of *Tetranychus urticae* [14], as if they had come from a single population. Since our database contained a large number of samples where taxonomy was incomplete (Figure S1), we treated each unidentified species as if it were unique. This maximised the use of the data, and is probably reasonably accurate, given that the taxonomy was least complete for very large, speciose groups, and that many of the unidentified species came from families or genera that were not otherwise represented.

To test the robustness of our results, we also examined a second approach to equalising the representation of all species, namely, merging all samples with a common pool size from each species, and treating them as a single sample, and then retaining the largest "merged" sample for each species. This approach includes information from across the species range when it is available, and so it could mitigate any downward bias in estimates of incidence. We rejected this approach for our main results, however, as it could upwardly bias incidence estimates when samples were obtained on different dates and prevalence varied over time.

Figure S3 compares parameter estimates for the major terrestrial arthropod orders (see below and Table S2), obtained with these two approaches to sampling. Figure S3 shows two cases where the different approaches to sampling did create substantial differences in the estimated incidence (one each in panels (f) and (i)). In particular, our "single largest sample" approach led to substantially lower incidence estimates for *Rickettsia* in Diptera, and for *Cardinium* in Opiliones. However, for the remaining 28/30 cases, incidence estimates were generally highly congruent between the two approaches, and particularly for our largest, *Wolbachia* data set (panel (c)). Overall, the similarity of the estimates must partly reflect the trivial fact that the single largest sample of each species often comprises a substantial fraction of the total number of individuals sampled for that species, but it also reflects the fact that the largest samples were often taken over larger sections of the species range, and might thus be more representative.

4.2 Sampling bias towards minor orders

Even after subsampling our data, our database contained a highly unrepresentative sample of arthropod species (Fig. S1). To correct for this taxonomic bias, we used weighted sums of the incidence estimates from each of the major terrestrial arthropod orders (Table S2; Fig. S3), weighting each estimate by the (estimated) contribution of that order to total arthropod biodiversity. In particular, for symbiont incidence we used

$$x_c = \sum_i f_i x_{c,i} \tag{17}$$

where $x_{c,i}$ is the estimated incidence for host order i, and f_i is the proportion of all arthropod species that are members of order i (such that $\sum_i f_i = 1$). For the results reported, we used only the largest orders of hexapods and/or chelicerates, and estimated f_i from the number of described species in those groups, as obtained from [12]. The estimates that we used are found in Table S2. So, for example, to obtain an estimate for chelicerates alone (Fig. 2), for Araneae we calculated $f_i = 43678/(43678 + 41939 + 12338 + 6534) = 0.418$, thus assuming that ~42% of chelicerate species are spiders.

To generate confidence intervals on this estimate, and to use its likelihood surface for model fitting, we wrote $x_{c,1} = \left(x_c - \sum_{i>1} f_i x_{c,i}\right)/f_1$, and then found the values of the $x_{c,i}$ that maximised the likelihood, conditional on x_c taking a given value. This was the approach used to produce the estimates labelled (c) in Figure 1, and all estimates in Figure 2.

Table S1 also applies the same approach to the other quantities of interest. These were calculated from:

$$\begin{aligned}\bar{q} &= \sum_i f_i \bar{q}_i \qquad (18)\\ F &= \frac{\sum_i f_i F_i \bar{q}_i (1 - \bar{q}_i)}{\bar{q}(1 - \bar{q})}\end{aligned}$$

although confidence intervals could not be provided for F, which is a ratio of variances, and not a simple sum.

4.3 Sampling bias towards infected populations

Another source of potential sampling bias is the overrepresentation in our database of species or populations already known to contain infection [3]. This bias is clearly evident from noting the species that are represented by a large number of screens (e.g., *Ixodes ricinus,* the castor bean tick, which is a known vector of rickettsial pathogens). This bias will be mitigated by the subsampling of screens, since no species will represented by more than one sample, but it could still remain. To test for this bias, we note that it is least likely to affect screens of single individuals from a large number of haphazardly-sampled arthropod species (e.g., [13]), and most likely to affect large multi-individual screens designed as stand-alone studies [3]. Accordingly, a suitable test is to compare estimates of the mean prevalence, $\bar{q}$, from single-individual screens and multi-individual screens (noting that $\bar{q}$ is the sole parameter than can be estimated from single-individual screens alone). If multi-individual screens have a significantly higher mean prevalence, this indicates that at least some of the screened populations were selected on the basis of prior knowledge of infection. To carry out this test, we fitted a model in which single-individual screens and multi-individual screens were each assigned their own value of $\bar{q}$ (the sole parameter that can be estimated from single-individual screens alone), and compared results to a model in which all screens had the same mean prevalence. Results shown in Table S3 suggest that this source of sampling bias is substantial across our data set as a whole: for all three symbionts, the two-$\bar{q}$ model provides a significantly better fit to the data, and the estimates for multi-individual screens were always substantially higher than those from single-individual screens (Table S3). Furthermore, for all three symbionts, the difference in $\bar{q}$ estimates from single- and multi-individual screens was always greater than the differences between estimates obtained from equivalently-sized but randomised divisions of the data (not shown).

However, we then applied the test to the subsampled data from each of the major arthropod orders (Table S2), which we used to produce our most reliable estimates. For the subsampled *Wolbachia* data, no order showed a significant difference in $\bar{q}$ estimates between single- and multi-individual screens (Table S2). For *Rickettsia* and *Cardinium*, four groups did show a significant difference, but there was no consistent tendency for the multi-individual screens to have a higher prevalence (as predicted if sampling were biased towards known infection). For example, for Araneae infected with *Cardinium*, and Diptera infected with *Rickettsia*, we found significantly higher levels of infection in the single-individual screens (Table S2). Thus, we concluded that the subsampled database showed no evidence of this kind of sampling bias.

The importance of the standardising sampling procedure, described above, is evident from Figure 1. To show how sampling bias can also affect between host-group comparisons, we repeated the analysis shown in Figure 2, but without applying standardised sampling. Results, shown in Figure S4, would lead us to conclude that there was a significantly higher incidence in chelicerates for all three bacteria, and significant differences between the bacteria within both groups; these results differ qualitatively from those shown in Figure 2, and reported in the main text.

5 Tests of predictors of incidence and prevalence

5.1 Tests for phylogenetic signal

To obtain a dated phylogeny of the higher arthropod taxa (Figure 3), we combined phylogenetic trees from published sources [15-17]. Since these trees included no dates for the Thysanoptera/Hemiptera split, we dated that node at 270.6 MA, which is consistent with fossil evidence [18], and with the dates of its parental nodes. In most cases, we divided the data by order, but we included some monophyletic superordinal groups where sampling was sparse.

To test for phylogenetic signal in symbiont incidence, we compared the fit of models in which these parameter values were assumed to have evolved over the true arthropod phylogeny (Fig. 3), to a model in which they evolved over a star phylogeny. Formally, we assumed that the logit transformed mean prevalence, $\ln(x_c/(1-x_c))$, for each order, evolved over the phylogeny by Brownian motion. This meant that the likelihood equation (eq. (5)), was combined with a multivariate normal distribution, with a covariance matrix determined by the phylogeny. The parameters of this distribution, namely its variance ("evolutionary rate"), and mean ("ancestral mean prevalence"), were then estimated along with the other model parameters. To assess the support for the non-nested models, we again used Akaike weights [10,11].

Table S4 shows results, using standardised sampling within each order. Results show that an explicit phylogenetic model gives a superior fit to the data for *Cardinium* and *Rickettsia*, but not for *Wolbachia*. However, the non-phylogenetic model could not be rejected in any case (Table S4), and this is consistent with the wide confidence intervals on estimates for many poorly-sampled orders (Fig. 3).

5.2 Species number and incidence

To test whether species rich families have higher levels of incidence, we used estimates of described species number from 39 published sources (see online supplementary information for full details). We then fit the linear model $\hat{y}_i = a + b\log(S_i)$ where S_i is the number of described species for host family i and $y_i = \ln(x_{c,i}/(1-x_{c,i}))$ is the logit transformed incidence. This model was fit directly to the sample prevalence data using the likelihood surface of eq. (5) expressed as a function of x_c, and compared to the fit of the null model (with $b = 0$) using a Likelihood Ratio Test. As a goodness of fit measure, we used McFadden's [19] pseudo-r^2, which is defined as

$$r^2 \equiv 1 - \ln\hat{L}_{lm}/\ln\hat{L}_{null} \tag{19}$$

where $\hat{L}_{lm}$ and $\hat{L}_{null}$ are the maximised likelihood values under, respectively, the linear model (with b free to vary) and the null model (with $b = 0$). To be a meaningful test, we required host groups that contained sufficient variation in both predictor and response variables. Therefore, Table S5 contains results from only those orders (or superordinal groups), which contained 5 or more families whose maximum likelihood estimate of incidence was intermediate (i.e., $0 < \hat{x}_{c,i} < 1$). We retained all families in these groups, including those represented by only a single screened individual, because the uncertainty in the incidence estimate for poorly sampled families is taken into account during the model fitting. The heterogeneity in the precision of the parameter estimates for individual families can be seen clearly in the large confidence intervals shown in Figure 4.

Supplementary references

1. Chiang, C. L. & Reeves, W. C. 1962 Statistical Estimation of Virus Infection Rates in Mosquito Vector Populations. *Am J Hyg* **75**, 377-391.

2. Jeyaprakash, A. & Hoy, M. A. 2000 Long PCR Improves *Wolbachia* DNA Amplification: wsp Sequences Found in 76% of Sixty-three Arthropod Species. *Insect Molecular Biology* **9**, 393-405. (doi:10.1046/j.1365-2583.2000.00203.x)

3. Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A. & Werren, J. H. 2008 How Many Species Are Infected with *Wolbachia*? A Statistical Analysis of Current Data. *FEMS Microbiology Letters* **281**, 215-220. (doi:10.1111/j.1574-6968.2008.01110.x)

4. Edwards, A. W. F. 1992 *Likelihood*. Expanded ed. Baltimore: Johns Hopkins University Press.

5. Wright, S. 1951 The Genetical Structure of Populations. *Annals of Eugenics* **15**, 323-354.

6. Takahasi, H. & Mori, M. 1974 Exponential Formulas for Numerical Integration. *Publ. RIMS, Kyoto Univ.* **9**: 721-741.

7. Mori, M. 2005. Discovery of the Exponential Transformation and its Developments. *Publ. RIMS, Kyoto Univ.* **41**: 897–935.

8. R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. *Vienna, Austria: the R Foundation for Statistical Computing*. ISBN: 3-900051-07-0. Available online at http://www.R-project.org/.

9. Ospina, R. & Ferrari, S. L. P. 2010 Inflated Beta Distributions. *Statistical Papers* **51**, 111-126. (doi:10.1007/s00362-008-0125-4)

10. Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**, 716-723. (doi:10.1109/TAC.1974.1100705)

11. Burnham, K. P. & Anderson D. R. 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

12. Zhang, Z.-Q. 2013 Phylum Arthropoda. *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (Addenda 2013). Zootaxa* **3703**, 17-26.

13. Russell, J. A., Funaro, C. F., Giraldo, Y. M., Goldman-Huertas, B., Suh, D., Kronauer, D. J. C., Moreau, C. S. & Pierce, N. E. 2012 A Veritable Menagerie of Heritable Bacteria from Ants, Butterflies, and Beyond: Broad Molecular Surveys and a Systematic Review. *PLoS ONE* **7**, e51027. (doi:10.1371/journal.pone.0051027)

14. Hoy MA, Jeyaprakash A 2005. Microbial diversity in the predatory mite *Metaseiulus occidentalis* (Acari: Phytoseiidae) and its prey, *Tetranychus urticae* (Acari: Tetranychidae). Biological Control 32: 427-441.

15. Wiegmann, B. M. 2009 Holometabolous insects (Holometabola). In *The Timetree of Life* (eds S. B. Hedges & S. Kumar), pp. 260-263. Oxford: Oxford University Press.

16. Wheat, C. W. & Wahlberg, N. 2013 Phylogenomic Insights into the Cambrian Explosion, the Colonization of Land and the Evolution of Flight in Arthropoda. *Systematic Biology* **62**, 93-109. (doi:10.1093/sysbio/sys074)

17. Johnson, K. P., Yoshizawa, K. & Smith, V. S. 2004 Multiple Origins of Parasitism in Lice. *Proceedings of the Royal Society B: Biological Sciences* **271**, 1771-1776. (doi:10.1098/rspb.2004.2798)

18. Martynov, A. 1935 A find of Thysanoptera in the Permian Deposits. *Compte Rendus (Doklady) de l'Academie des Sciences de l'URSS* **3**, 333-336.

19. McFadden, D. 1973 Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics* (ed P. Zarembka), pp. 105-142. New York: Academic Press.

Supplementary Tables

Table S1: The estimated distribution of symbiont prevalences in terrestrial arthropods

Symbiont		$\bar{q}$	F	p_0	p_1	w
Wolbachia	(a)	0.313 (0.303, 0.325)	0.718 (0.704, 0.732)	0	0	0.815
	(b)	0.313 (0.302, 0.324)	0.715 (0.699, 0.731)	0.115 (0.000, 0.282)	0.027 (0.000, 0.066)	0.185
	(c)	0.236 (0.219 , 0.255)	0.747	-	-	-
Rickettsia	(a)	0.146 (0.136, 0.158)	0.443 (0.416, 0.472)	0	0	0.016
	(b)	0.144 (0.134, 0.155)	0.443 (0.412, 0.473)	0.000 (0.000, 0.205)	0.011 (0.004, 0.019)	0.984
	(c)	0.051 (0.040, 0.069)	0.577	-	-	-
Cardinium	(a)	0.108 (0.095, 0.123)	0.734 (0.680, 0.780)	0	0	0.091
	(b)	0.105 (0.092, 0.119)	0.712 (0.656, 0.764)	0.407 (0.000, 0.745)	0.038 (0.017, 0.055)	0.909
	(c)	0.059 (0.051, 0.073)	0.596	-	-	-

Parameter values show maximum likelihood estimates, with confidence intervals in parentheses. Parameters estimated are $\bar{q}$: the mean prevalence; F: the proportion of the variance in infection status that is due to between-species variation in prevalence; p_0: the proportion of species free from infection; p_1: the proportion of species that are completely infected (as with a primary symbiont). As with Figure 1, estimates were obtained from (a) fitting a beta distribution to the complete database; (b) fitting a doubly-inflated beta distribution to the complete database; (c) standardised sampling (i.e., a weighted sum of estimates from the largest arthropod taxa, using an equalised number of screens per sampled species within in each taxon); w is the Akaike weight associated with the chosen form of the distribution of prevalences, i.e., the probability that this model, and not the alternative, minimises the information loss [10,11].

Table S2: Numbers of described species, and tests of sampling bias for major arthropod groups

			Wolbachia	*Rickettsia*	*Cardinium*
	Group	No. spp.	all (SIS/MIS)	all (SIS/MIS)	all (SIS/MIS)
Hexapoda	Coleoptera	389,487	0.211 (0.225, 0.188)	0.049 (0.015, 0.110)*	0.000 (0.000, 0.000)
	Lepidoptera	158,423	0.283 (0.277, 0.292)	0.029 (0.038, 0.000)	0.000 (0.000, 0.000)
	Hymenoptera	153,088	0.346 (0.324, 0.369)	0.006 (0.003, 0.008)	0.022 (0.022, 0.024)
	Diptera	156,774	0.182 (0.183, 0.180)	0.143 (0.154, 0.027)*	0.057 (0.000, 0.087)*
	Paraneoptera	118,867	0.206 (0.227, 0.195)	0.027 (0.030, 0.008)	0.114 (0.035, 0.171)*
	Orthoptera	23,830	0.233 (0.179, 0.311)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Chelicerata	Araneae	43,678	0.192 (0.233, 0.170)	0.030 (0.000, 0.037)	0.446 (0.667, 0.323)*
	Acariformes	41,939	0.282 (0.211, 0.350)	0.076 (0.000, 0.078)	0.394 (0.412, 0.371)
	Parasitiformes	12,338	0.157 (0.214, 0.114)	0.191 (0.273, 0.188)	0.092 (0.000, 0.165)
	Opiliones	6,534	0.000 (0.000, 0.000)	0.000	0.333 (0.313, 0.500)

No. spp.: Estimated number of described species [12]; Remaining columns show estimates of the mean prevalence, $\hat{\hat{q}}$, for subsamples of the data, with equalised representation of each species in the data set. Estimates in parentheses show the same estimates for single-individual screens (SIS), and multi-individual screens (MIS) for each subset of the data. * indicates a significant improvement in model fit when SIS and MIS were allowed to have their own mean prevalences (Likelihood Ratio Test, with significance at the 5% level).

Table S3: Evidence of sampling bias in the full data sets

	No. screens		$\hat{\bar{q}}$			
Symbiont	SIS	MIS	SIS	MIS	$\Delta \ln L$	p
Wolbachia	2965	3222	0.249	0.355	47.428	$< 10^{-6}$
Rickettsia	1427	1427	0.107	0.165	13.516	$< 10^{-6}$
Cardinium	1095	672	0.056	0.174	35.301	$< 10^{-6}$

SIS: single-individual screens; MIS: multi-individual screens. No. screens: the number of screens of each type; $\hat{\bar{q}}$: maximum likelihood estimates of the mean prevalence from screens of each type; $\Delta \ln L$: the improvement in log likelihood obtained by allowing SIS and MIS to have different mean prevalences; p: p-value of Likelihood Ratio Test comparing one- and two-$\bar{q}$ models.

Table S4: Phylogenetic signal in symbiont incidence

	$\ln\hat{L}$		
Symbiont	star phylogeny	true phylogeny	p
Wolbachia	-3135.16	-3137.35	0.101
Rickettsia	-859.25	-856.91	0.088
Cardinium	-392.00	-389.54	0.079

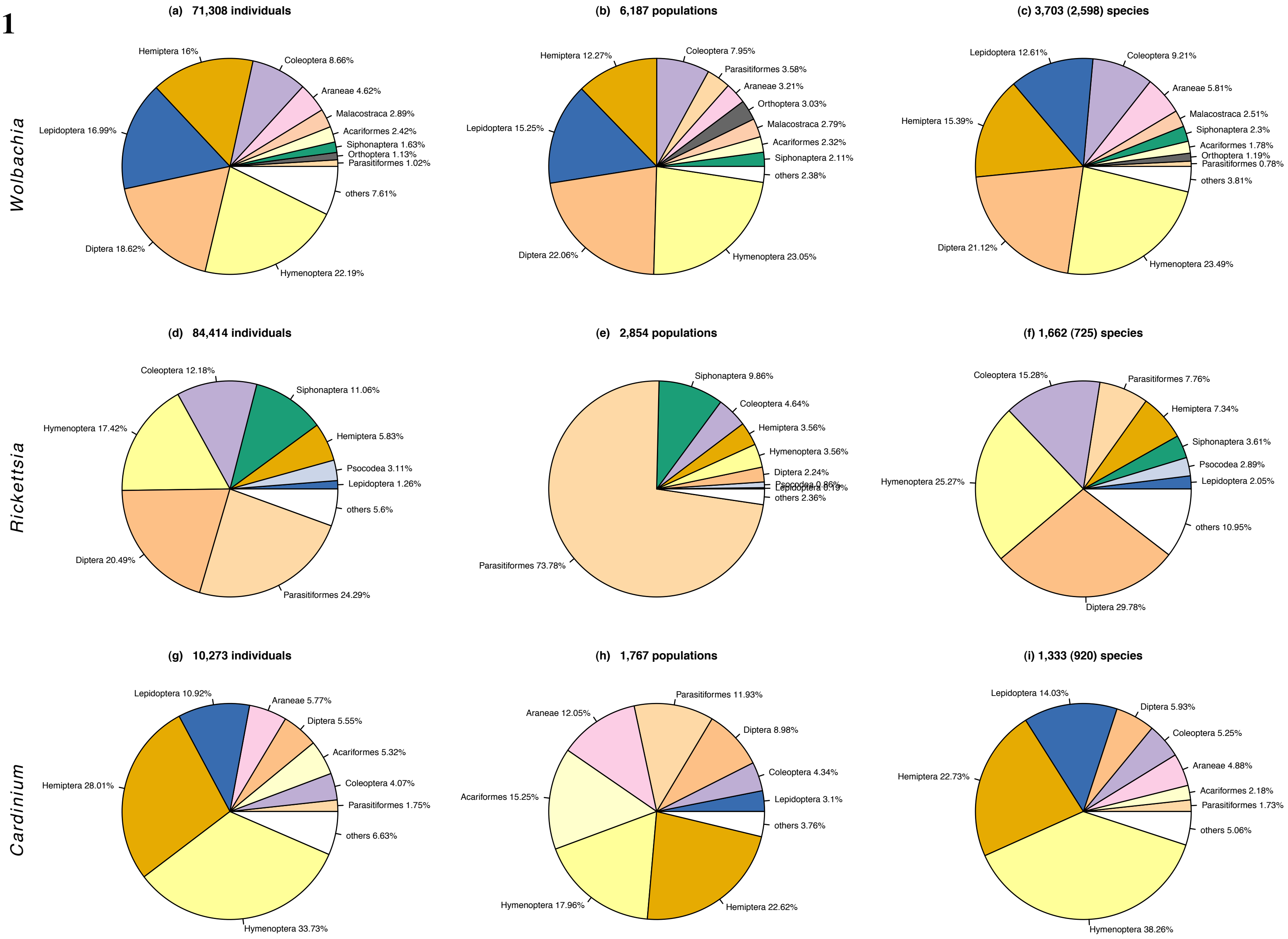
$\ln\hat{L}$: the maximised log likelihood under a model in which the logit transformed incidence ($x_{0.001}$) in each arthropod group was assumed to have evolved over a star phylogeny, or the true phylogeny, by Brownian motion. The higher likelihood for each data set is underlined, and p is the probability that this higher-likelihood model minimises the information loss (calculated using Akaike weights; [10,11]).

Table S5: The relationship between species richness and symbiont incidence

				>0.1% prevalence ($x_{0.001}$)			>50% prevalence ($x_{0.5}$)			>90% prevalence ($x_{0.9}$)		
Symbiont	Host group	Tax. level	n	$\hat{b}$	pseudo-r^2	p	$\hat{b}$	pseudo-r^2	p	$\hat{b}$	pseudo-r^2	p
Wolbachia	Coleoptera	family	40	2.93	0.040	$< 10^{-6}$**	6.25	0.154	$< 10^{-6}$**	5.97	0.078	$< 10^{-6}$**
	Lepidoptera	family	29	0.82	0.002	0.038*	-0.32	0.001	0.11	-0.73	0.002	0.017*
	Hymenoptera	family	40	0.02	0.000	0.91	-0.54	0.008	$< 10^{-6}$**	-0.73	0.010	$< 10^{-6}$**
	Diptera	family	45	0.04	0.000	0.85	-0.19	0.001	0.07	0.02	0.000	0.89
	Hemiptera	family	56	-2.29	0.028	$< 10^{-6}$**	-1.21	0.024	$< 10^{-6}$**	-0.67	0.006	0.001**
	Araneae	family	19	-2.39	0.007	0.05*	-0.78	0.007	0.04*	-0.73	0.001	0.40
		genus	93	-6.47	0.032	0.0002**	-1.67	0.014	0.007**	-0.59	0.001	0.59
	Acari	genus	28	4.57	0.032	0.0003**	1.73	0.029	0.0001**	0.29	0.000	0.84
	Malacostraca	family	32	0.69	0.001	0.41	0.35	0.002	0.26	0.76	0.002	0.36
Rickettsia	Coleoptera	family	33	3.24	0.043	$< 10^{-6}$**	1.15	0.008	0.029*	0.01	0.000	0.96
	Acari	genus	14	0.43	0.000	0.27	-0.98	0.004	0.0002*	-1.84	0.001	0.045*
	Siphonaptera	genus	36	0.17	0.000	0.87	-2.45	0.040	$< 10^{-6}$**	-3.14	0.007	0.017*
Cardinium	Araneae	family	15	0.22	0.000	0.82	-0.66	0.007	0.15	2.87	0.006	0.18
		genus	35	-0.59	0.001	0.68	0.35	0.006	0.19	0.65	0.005	0.24

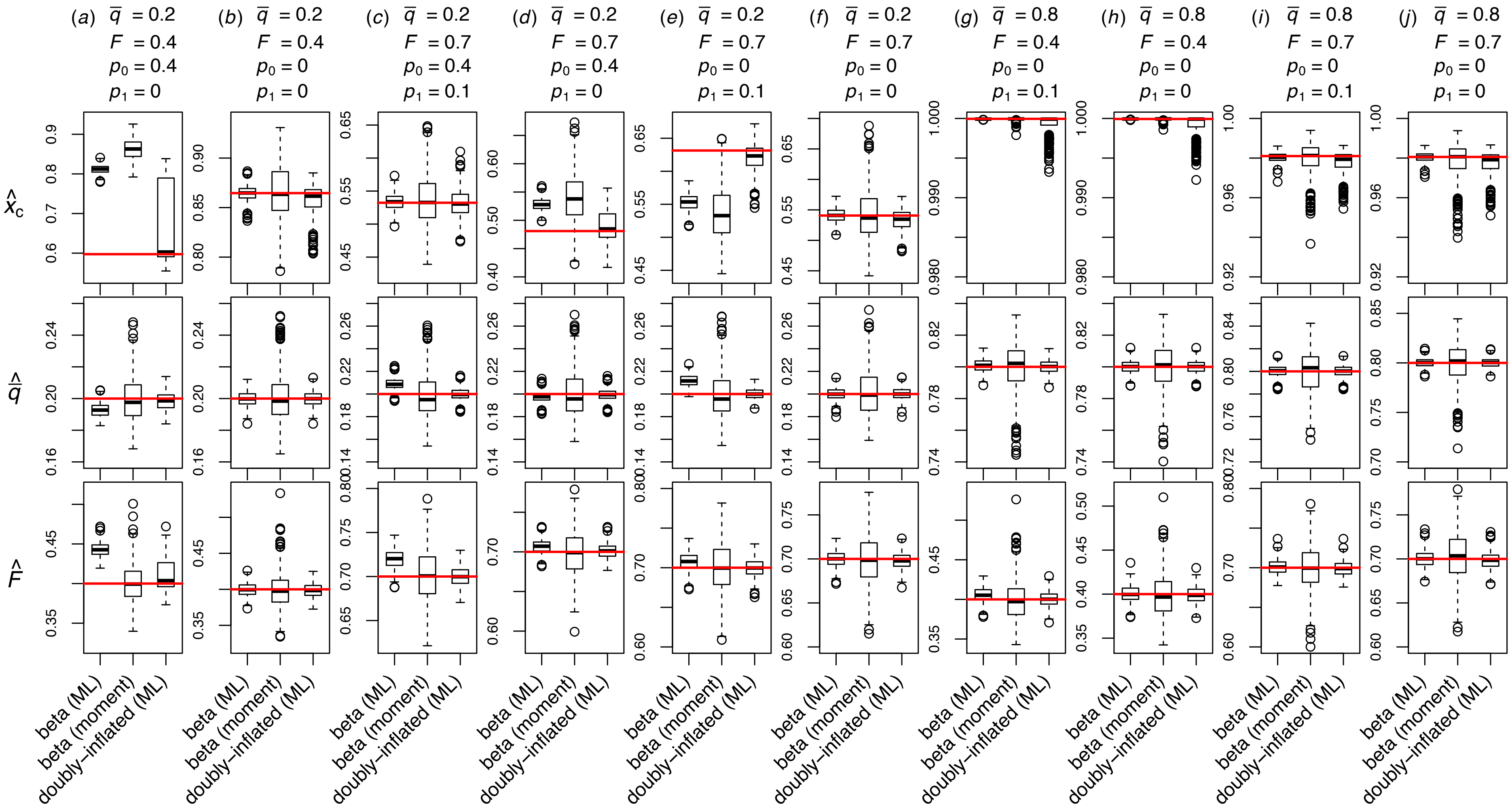
n: number of sampled families or genera within each host group; $\hat{b}$: best-fitting slope in the linear model connecting symbiont incidence (proportion of species infected above a given prevalence) in an arthropod family or genus to the species richness of that family or genus; pseudo-r^2: the goodness-of-fit measure, eq. (19); p: p-value from a Likelihood Ratio Test of the null model $b = 0$; * $p < 0.05$; ** $p < 0.01$.

Figure S1



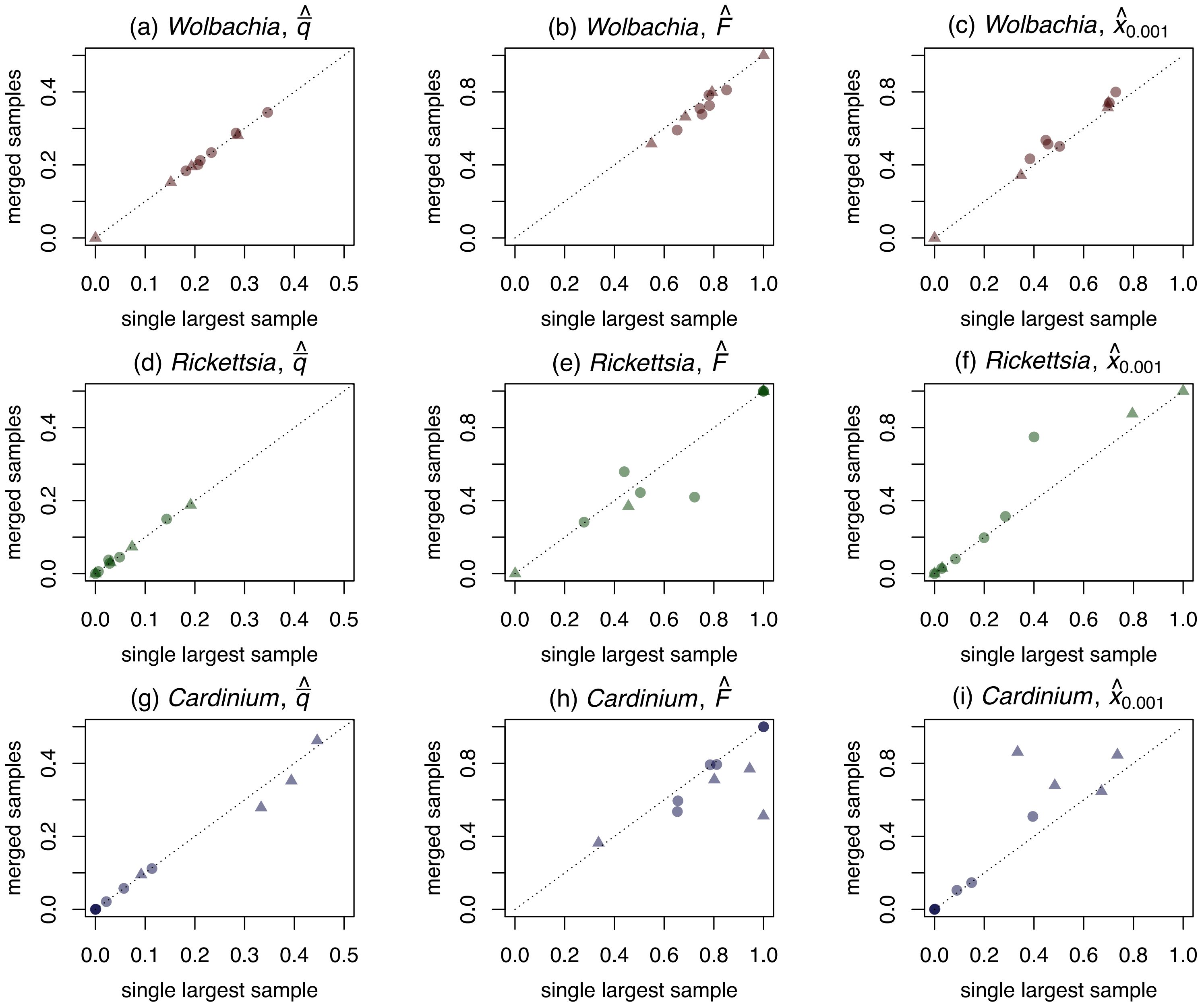
A summary of our database of arthropod screens for three genera of endosymbiotic bacteria, namely *Wolbachia* (a)-(c), *Rickettsia* (d)-(f), and *Cardinium* (g)-(i). The content of the database is summarised in terms of host taxonomy. Left-to-right, columns show plots for arthropod individuals; populations (each of which might be represented by one or more individuals); and species (each of which might be represented by one or more populations). For the number of species, two values are listed. The larger value treats each population with incomplete taxonomy as if it came from a unique species, otherwise absent from the database. The smaller value, in parentheses, counts only those species whose taxonomy was complete. As such, these two numbers represent upper and lower bounds on the true numbers of species sampled. The full database is provided as online supplementary information.

Figure S2



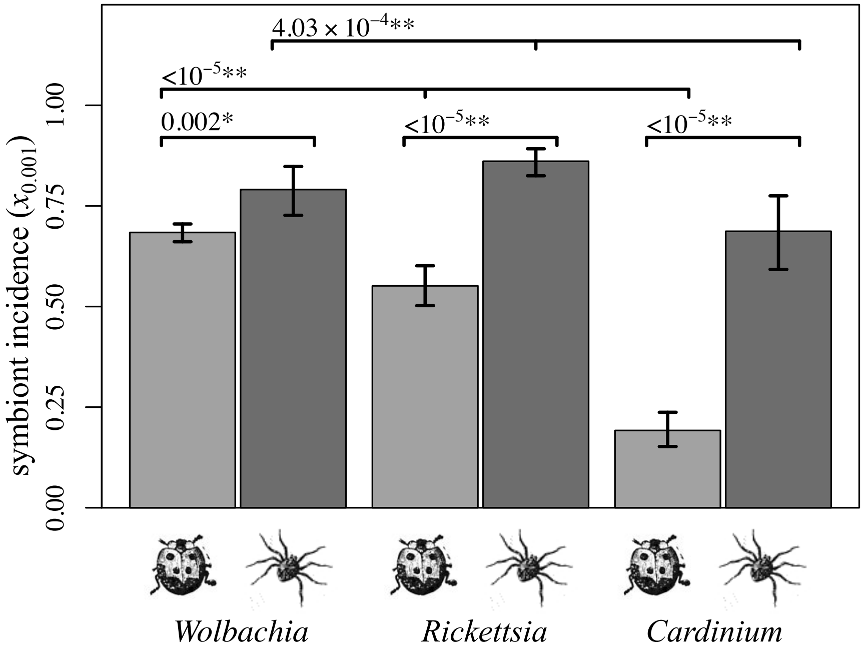
Estimated parameters of the distribution of across-species prevalences for simulated data sets based on the real *Wolbachia* data. Each column of panels (a)-(j) contains results for data simulated under a different set of parameters for the true distribution of prevalences across species, while each row of panels shows estimates for a different parameter, namely, the proportion of species estimated to be infected at a prevalence above 0.001 (x_c), the mean prevalence (q), and the proportion of the variance in infection status due to between-species variation in prevalence (F). The true values of these parameters - used to simulate the data - are shown in red. Each plot compares parameter estimates from a maximum likelihood fitting of a beta distribution (eq. 2), a moment-based approach to estimating these same parameters [3], and maximum likelihood fitting of a doubly-inflated beta distribution (eqs. 14-15). For the moment-based approach, we used estimators of the shape parameters α and β reported by Hilgenboecker *et al*. ([3]; their eqs. 1-4), and then used eqs. (8)-(10). The box-and-whiskers were generated using the *boxplot* function in *R* [8] with default settings.

Figure S3



Comparison of maximum likelihood parameter estimates for the major groups of arthropods under two methods of equalising the representation of each species, to better estimate the distribution of prevalences across arthropod species. The x-axis shows estimates obtained from retaining only the single largest population sample from each species (the approach used in the main text). The y-axis shows equivalent estimates when all of the samples from each species were combined, and treated as if they came from a single population. In each plot, points correspond to the ten arthropod taxa listed in Table S2, with hexapod groups shown as circles, and chelicerates as triangles.

Figure S4



Estimates of symbiont incidence in the two major subphyla of arthropoda. Estimates used our complete database, without applying “standardised sampling”. All other details match Figure 2.