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. 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, and this might lead to some infections being missed, particularly with older methods of DNA extraction [2-4]. We chose exclude only studies using long PCR - which is highly sensitive to very low titre infections, but might yield a high rate of false positives [2,5].

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 [2]. 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).

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

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 [6]. 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} [7]. 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.* [2], 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 [6].

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 [2].

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

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) [8,9]. This transformation is as follows:

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

where

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

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* [10] 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 [11], 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.

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

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.* ([2]; 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 [2] 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 [12,13]. 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* [16], as if they had come from a single population. Since our database contained a large number of samples where taxonomy was incomplete (Fig. 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 [14]. 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 [2]. 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., [15]), and most likely to affect large multi-individual screens designed as stand-alone studies [2]. 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 (Fig. 3), we combined phylogenetic trees from published sources [17-19]. 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 [20], 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 [12,13].

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). For the best sampled orders, 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 for that family. 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 so all of the uncertainty in our incidence estimates was taken into account. We then compared the fit of the null model (with $b = 0$) using a Likelihood Ratio Test. As a goodness of fit measure, we used McFadden's [21] 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, we calculated 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 S5). We initially 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 of these results are contained in Table S5, and some illustrative cases are plotted in Figure S5.

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. 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)

3. 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.0082402)

4. 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)

5. 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)

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

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

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

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

10. 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/.

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

12. 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)

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

14. 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.

15. 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)

16. 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.

17. 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.

18. 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)

19. 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)

20. 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.

21. 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		$\overline{q}$	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 [12,13].

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 [14]; 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	$\underline{-3135.16}$	-3137.35	0.101
Rickettsia	-859.25	$\underline{-856.91}$	0.088
Cardinium	-392.00	$\underline{-389.54}$	0.079

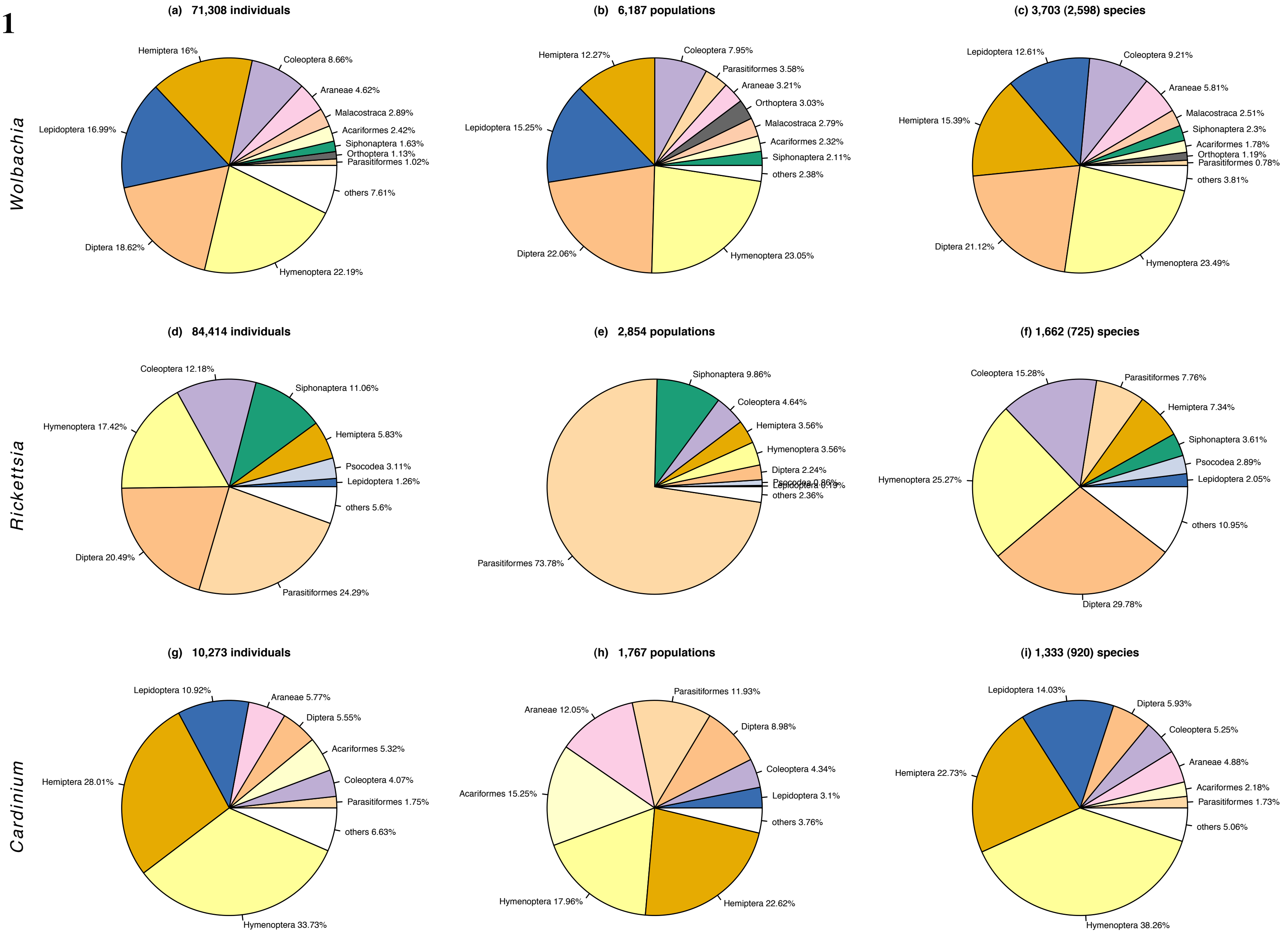
$\ln\hat{L}$: the maximised log likelihood under a model in which the logit transformed incidence ($x_{0.001}$) in each arthropod group was assumed to have evolved over a star phylogeny, or the true phylogeny, by Brownian motion. The higher likelihood for each data set is underlined, and p is the probability that this higher-likelihood model minimises the information loss (calculated using Akaike weights; [12,13]).

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

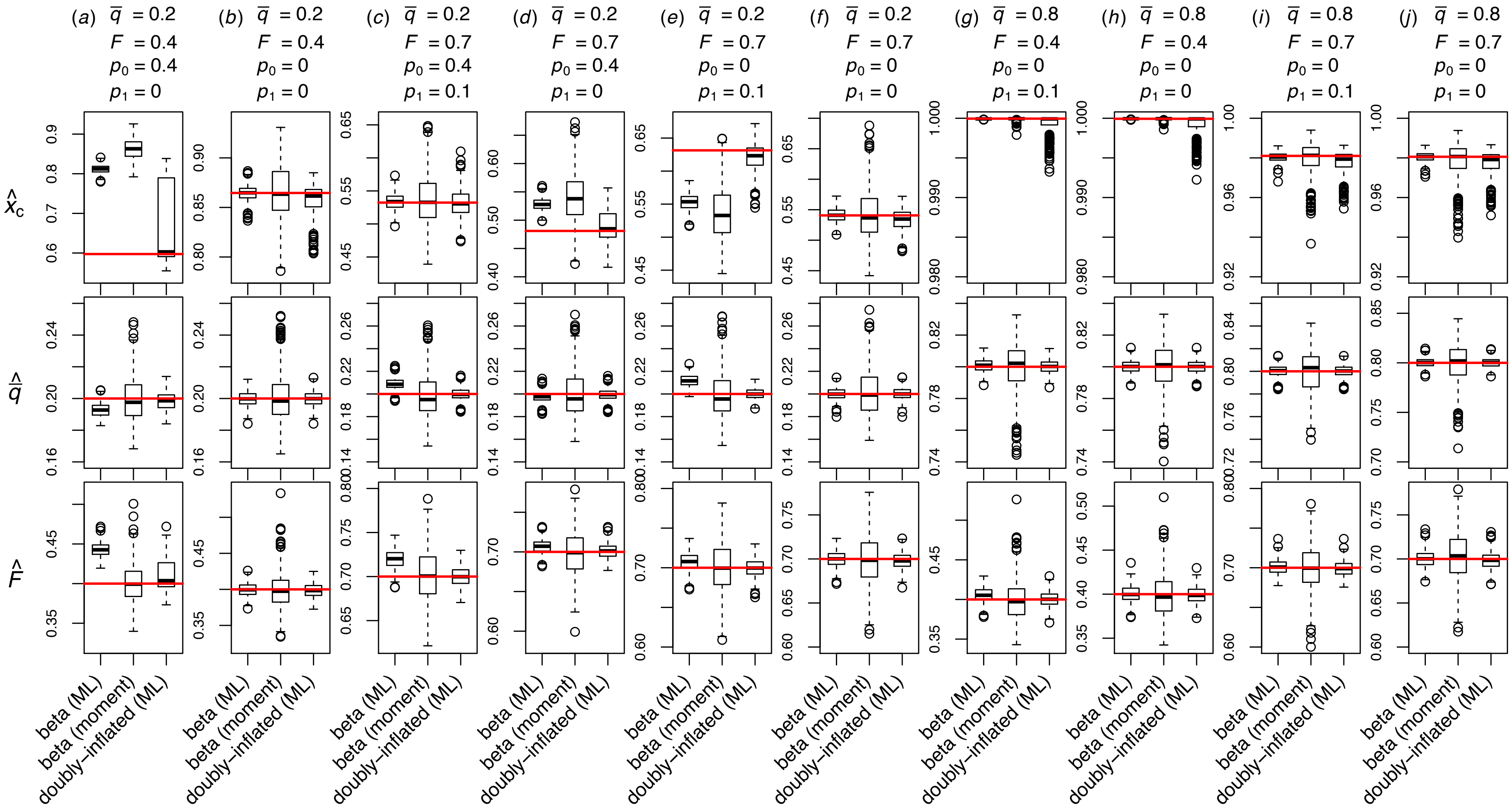
n: number of sampled families or genera within each host group; $\hat{b}$: best-fitting slope in the linear model connecting symbiont incidence (proportion of species infected above a given prevalence) in an arthropod family or genus to the species richness of that family or genus; pseudo-r^2: the goodness-of-fit 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



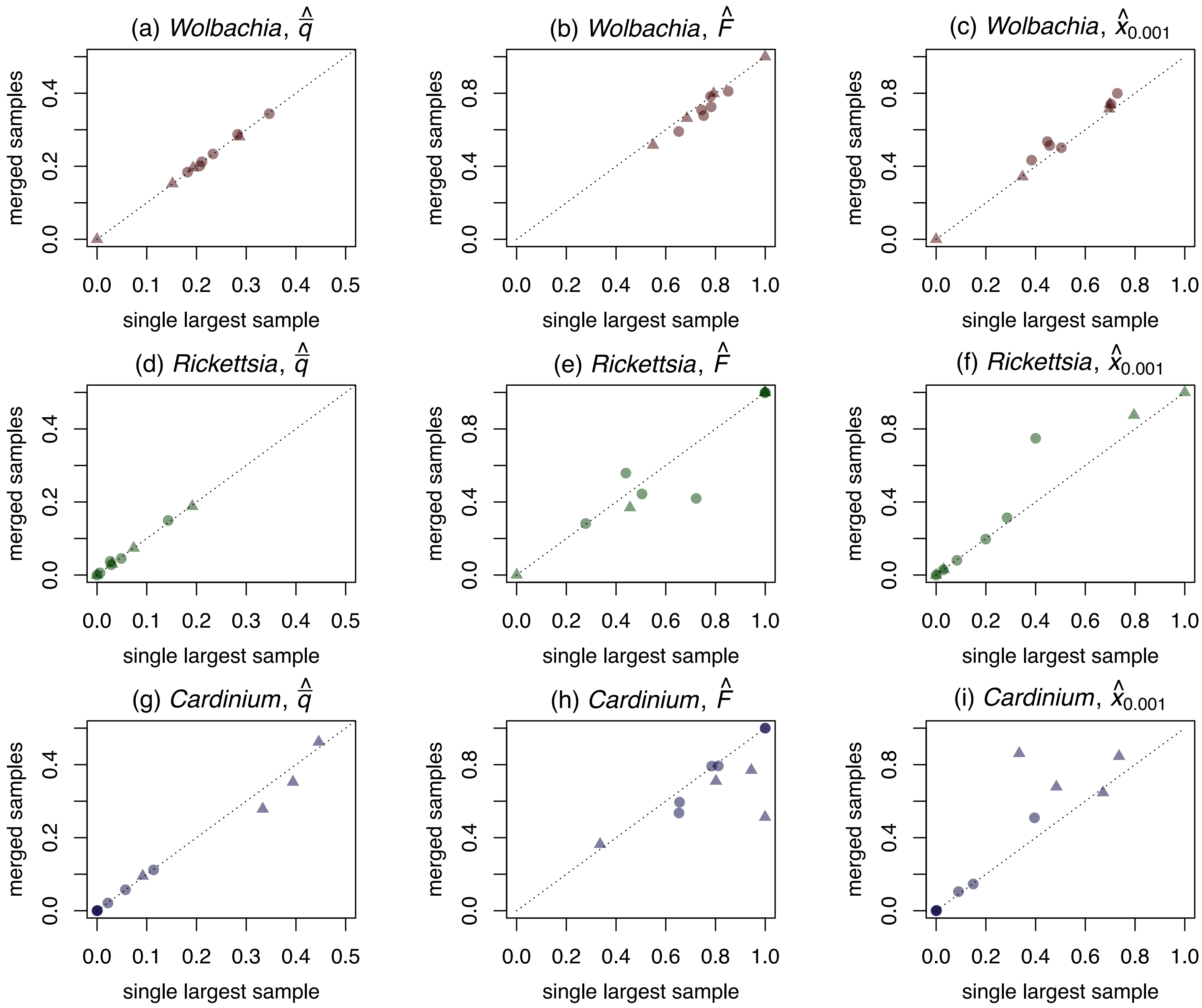
A summary of our database of arthropod screens for three genera of endosymbiotic bacteria, namely *Wolbachia* (a)-(c), *Rickettsia* (d)-(f), and *Cardinium* (g)-(i). The content of the database is summarised in terms of host taxonomy. Left-to-right, columns show plots for arthropod individuals; populations (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



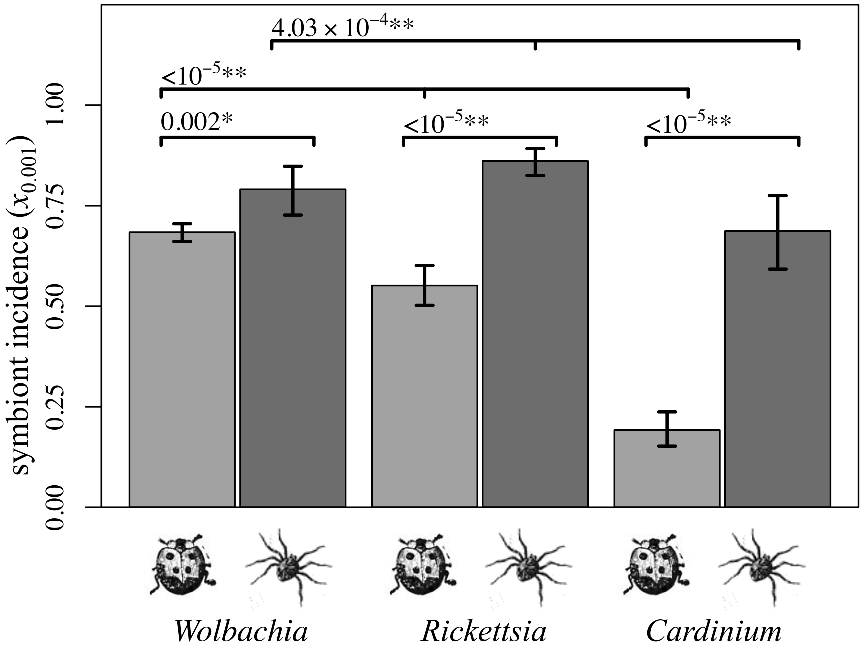
Estimated parameters of the distribution of across-species prevalences for simulated data sets based on the real *Wolbachia* data. Each column of panels (a)-(j) contains results for data simulated under a different set of parameters for the true distribution of prevalences across species, while each 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



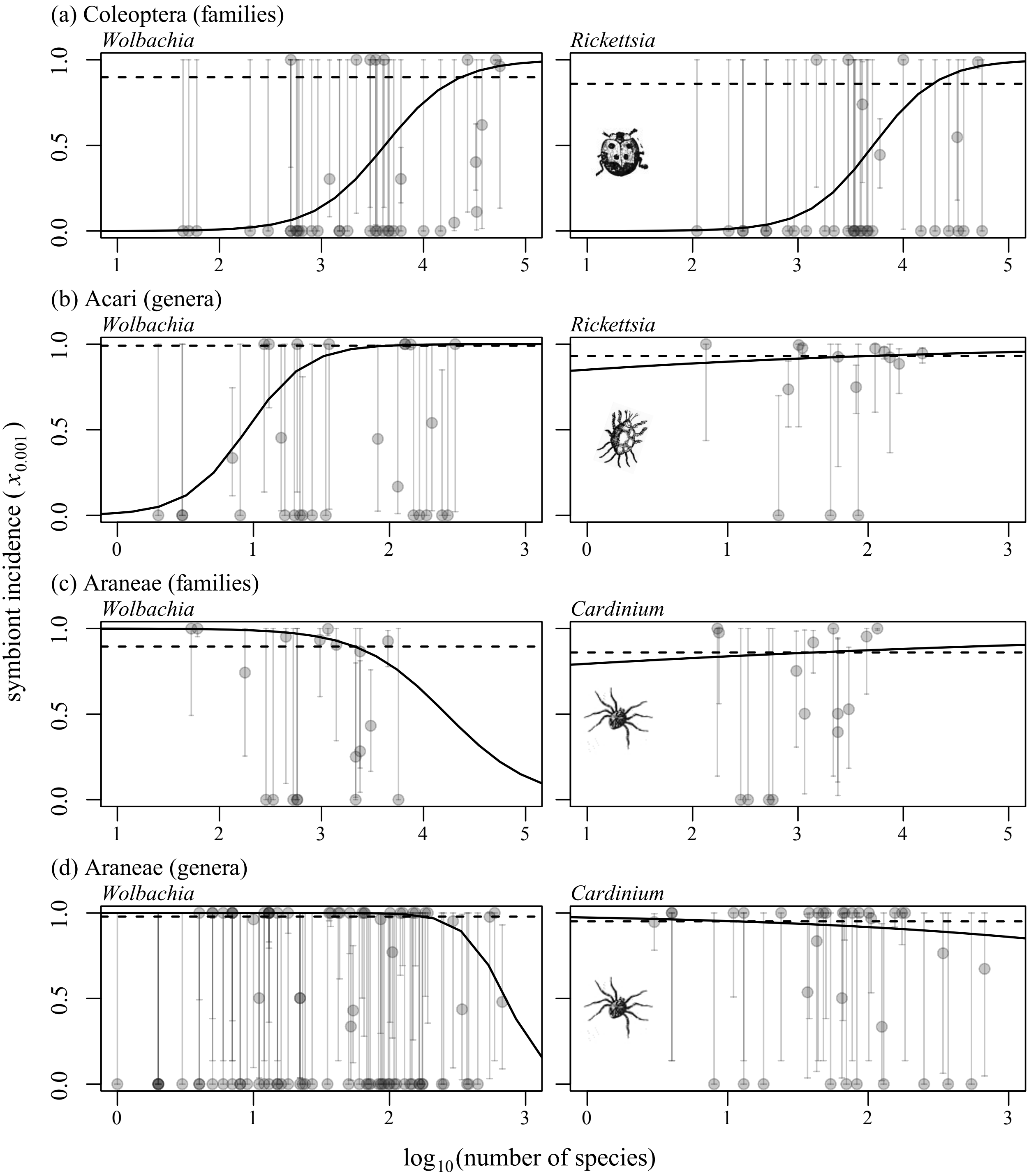
Comparison of maximum likelihood parameter estimates for the major groups of arthropods under two methods of equalising the representation of each species, to better estimate the distribution of prevalences across arthropod species. The x-axis shows estimates obtained from retaining only the single largest 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.

Figure S5



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.