frontiers

A hierarchical Bayesian approach to estimate endosymbiont infection rates

Zachary H. Marion [1] **and Christopher A. Hamm** [2,*]

[1] *Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN, USA*

[2] *Center for Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, CA, USA*

Correspondence*:
Christopher A. Hamm
cahamm@ucdavis.edu

Center for Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, One Shields Avenue, Davis, CA, 95616 USA

ABSTRACT

Endosymbionts may play an important role in the evolution of the Insecta. Bacteria such as *Wolbachia, Cardinium,* and *Rickettsia* are known to manipulate host reproduction to facilitate their own. Indeed, there are many documented cases where *Wolbachia* (Alphaproteobacteria: Rickettsiaceae) induces one of four manipulative phenotypes (cytoplasmic incompatibility, male killing, feminization, and parthenogenesis). The scale of infection among species has been a major subject of investigation, but quantification has been difficult because various approaches have yielded different estimates. One under-appreciated aspect of this problem arises when multiple—yet independent—samples are taken within a taxon. When independent samples within a taxon are treated as levels of a hierarchy, the problem is greatly simplified because data are partially pooled according to taxon. Here, we present a hierarchical Bayesian approach to estimate infection frequency where multiple independent samples were collected across several taxonomic levels. We apply this model to estimate the rates of infection for *Wolbachia* in the Lepidoptera, and then extend the model to account for phylogenetic non-independence. In addition, we highlight the current knowledge regarding *Wolbachia* and its effects within Lepidoptera. Our model estimates that the rate of endosymbiont infection for the Lepidoptera is approximately 12%. Given our limited knowledge regarding the phenotypes induced by these endosymbionts and the low infection rate, we urge caution when extrapolating the results of positive assays.

Keywords: *Wolbachia*, Lepidoptera, modeling, butterfly, moth

1 INTRODUCTION

Scientists have long known that bacterial endosymbionts inhabit insects. Many of these endosymbionts are maternally transmitted to offspring through the cytoplasm of the egg. *Wolbachia* (Alphaproteobacteria: Rickettsiaceae) was the first of these endosymbionts to be discovered when Hertig and Wolbach (1924) examined the adult ovaries and testes of *Culex pipiens* (hence the specific epithet of *Wolbachia pipientis*; Hertig, 1936). Some years later, Yen and Barr (1971) observed that male *C. pipiens* from one geographic

area may not successfully reproduce with females from a different area and that reciprocal crosses could produce similar results; this phenomenon was given the name *cytoplasmic incompatibility*. A *Rickettsia*-like organism was determined to be the causative agent—*Wolbachia* (Yen and Barr, 1973). Recently, Hilgenboecker et al. (2008) estimated that about 20% of the Insecta are infected with *Wolbachia*, while Zug and Hammerstein (2012) placed that estimate at roughly 40%.

Today, researchers detect the presence of *Wolbachia* via the polymerase chain reaction (PCR), samples can be screened quickly and relatively inexpensively (Baldo et al., 2006; Simões et al., 2011). However, this development is relatively recent. Prior to the advent of PCR, *Wolbachia* infection was only confirmed through painstaking work that included electron microscopy and other microbiological techniques. Indeed, these methods were so laborious that they were only employed once a researcher had a prior reason (e.g., *male killing, feminization, parthenogenesis, cytoplasmic incompatibility*) to suspect the presence of the bacterium. We are aware of no cases in which exploratory assays for *Wolbachia* were conducted prior to the appearance of PCR.

Exploratory investigations for the presence of *Wolbachia* became feasible with the advent of PCR and Sanger sequencing. Yet few studies conducted the experimental work to determine if any reproductive manipulation was occurring. Careful experimental work is required to determine what (if any) phenotype is induced by an endosymbiont. The effects of *Wolbachia* infection are complex and depend on the interaction between the genomes of the endosymbiont and the host. For example, the phenotypic effects of one strain of *Wolbachia* may be very different if moved into another host (Rigaud et al., 2001; Hoffmann et al., 2011). Additionally, there may be extensive genomic differences between closely related strains (Ishmael et al., 2009). Although *Wolbachia* is most famous for being a "reproductive parasite," *Wolbachia* infections can often result in no manipulation at all (Hamm et al., 2014a; Zhang et al., 2010, 2013). Thus, infections do not necessarily cause reproductive manipulations.

The Lepidoptera (Arthropoda: Insecta) are among the best studied animal orders, containing $\sim 160,000$ species in 124 families, representing approximately 13% of all described life (Regier et al., 2013; van Nieukerken et al., 2011). Because of historic interest in their physical beauty and their contemporary economic importance, the literature is replete with detailed information regarding their distribution and life history. In addition to basic scientific research, the Lepidoptera are also well represented on lists of endangered or threatened species (Hamm et al., 2014b). Yet certain groups of Lepidoptera have garnered the majority of attention, such as the butterflies (e.g. Nymphalidae, Lycaenidae and Pieridae) or groups of economically important pest species such as the Crambidae (which contains the Asiatic rice borer *Chilo suppressalis*) and Noctuidae (which contains the armyworms of the genus *Spodoptera*). This results in a bias towards certain groups and leaves most of the remaining families understudied.

Six species of Lepidoptera have been tested for the existence of a naturally occurring manipulative phenotype with evidence for *cytoplasmic incompatibility, male killing,* and *feminization* (Table 1). We note that the report of *male killing* in *Ephestia kuhniella* is a result of *Wolbachia* transfected from *Ostrinia scapulalis* (Fujii et al., 2001). Given the high level of interest in the Lepidoptera, understanding the role *Wolbachia* plays in the evolution of lepidopterans has received considerable attention. A vital first step towards this understanding is the estimation of *Wolbachia* infection rates across the order. Previous work on the estimation of *Wolbachia* infection frequency have employed maximum likelihood estimation. Ahmed et al. (2015) and Weinert et al. (2015) estimated *Wolbachia* infection frequencies for the Lepidoptera. These studies represent important steps in the estimation of *Wolbachia* infection frequency in the Lepidoptera, and our work here builds upon them. Our primary aim here was to develop a hierarchical partial-pooling strategy and account for phylogenetic non-independence at the family level.

Here, we develop a novel hierarchical Bayesian approach to estimate *Wolbachia* infection frequencies across the Lepidoptera. Our model explicitly accounts for issues that arise with real world data such as quantifying infection levels at different taxonomic levels. In a hierarchical Bayesian approach, a compromise via partial pooling occurs because lower levels of the hierarchy inform higher levels of the hierarchy, and vice versa. Therefore, When there is little information within a grouping (e.g., species with few observations), those estimates are pulled strongly (shrunk) towards the among-group mean. Conversely, parameter estimates for groups with high levels of information experience little shrinkage and instead inform the estimates for groups with less information. For example, there may be multiple observations of infection frequency collected from different populations within a species, often with disparate sample sizes. We do not consider it appropriate for these samples to be completely pooled, as that ignores population differences in infection frequency. Nor should observations within species be considered independent, because of shared ancestry. Similarly, there may be single samples collected from many different species within a family. In this case, individual sampling error should be accounted for when estimating family level infection rates. Finally, we consider that there has been a bias towards studying only a few families of the Lepidoptera. This uneven sampling can cause a few well-studied families to drive estimates of overall infection frequency. Each of these concerns can be specifically addressed using a hierarchical Bayesian model that incorporates phylogenetic relatedness among lepidopteran families.

2 MATERIALS & METHODS

2.1 Motivating data and previous analyses

Early studies on *Wolbachia* prevalence reported the frequency of infection for small groups of insects or arthropods (Jiggins et al., 2001; Werren and Windsor, 2000). More recent and sophisticated models of *Wolbachia* infection in the Lepidoptera used a likelihood-based approach to describe the distribution of *Wolbachia* infection across arthropods (Weinert et al., 2015) and the Lepidoptera specifically (Ahmed et al., 2015). Following Hilgenboecker et al. (2008), both studies used beta-binomial models to estimate the mean proportion of individuals infected within a given species. Both used the same distribution to calculate the incidence of infection as well, where incidence was the proportion of species infected above a threshold frequency (i.e., one infection in 1000 individuals, or 0.001; Weinert et al., 2015).

In the case of *Wolbachia*, tested insects may be either positive or not positive (a band of appropriate length when an electrophoresis gel is run, or no band, respectively). It is important to note that "not positive" is more appropriate than negative here because infections may have been missed for a number of reasons, including low density infections (Schneider et al., 2014). However, for the sake of simplicity, we will treat *Wolbachia* infection status as two mutually exclusive outcomes, (0 or 1; positive or not positive). This makes the question of infection a binomial sampling problem. The issue is the way that some models have accounted for uncertainty at each level. We will demonstrate this problem with two examples. First, let us assume that 200 individuals of a species are assayed for *Wolbachia*, and 100 of those tests are positive for infection. The mean estimate of infection is 0.5 and the 95% exact binomial confidence interval is 0.43–0.57. Next, consider two assayed individuals from a species where one tested positive. Here the proportion infected in this species is also 0.5, yet the 95% confidence interval is 0.01–0.99. It is clear that there is uncertainty around each estimate and that uncertainty varies with sample size. For this error to be properly incorporated into any estimate it must be treated at each level of the analysis (species, family, etc.), rather than pooled at the level of the entire study.

2.2 Data

We used the data compiled by Weinert et al. (2015), which contains records from thousands of individual sampling efforts across arthropods. Each row denotes one independent sampling event (though each row may contain data from multiple individual assays) and contained information on the arthropod family, genus, species, number of individuals assayed, number of individuals positive for infection, and endosymbiont genus. For this study, we only considered *Wolbachia* assays of Lepidoptera. All analyses were conducted in R (R Core Team, 2016, *v3.3*), and all data and code necessary to reproduce our results are freely available on Zenodo (http://doi.org/10.5281/zenodo.166803).

We used the Lepidoptera phylogeny of Regier et al. (2013) as a covariate to account for any influence of the relatedness among families in our analysis. This tree contained representatives from 115 of the 124 families in the order. The tree was pruned to remove duplicate entries at the family level and those not present in the Weinert et al. (2015) dataset. We made the tree ultrametric using the penalized likelihood method of Sanderson (2002) with tools from the *ape* package (Paradis et al., 2004). To incorporate phylogenetic history into the Bayesian model, we used the pruned ultrametric tree to create a series of phylogenetic correlation matrices. We constructed one matrix in which we assumed that *Wolbachia* infection status was distributed according to Brownian Motion (BM), a model of trait evolution that assumes closely related taxa share that trait due to common ancestry (Paradis, 2012). We also constructed matrices which assumed trait evolution followed an Ornstein-Uhlenbeck (OU) process, which places constraints around which a character evolves (Paradis, 2012). Relative to the BM, the OU model has two additional parameters: θ (the "optimal" value for a character), and α (the rate at which θ moves towards α) (Paradis, 2012). The α value can range from 0 - 1; when α is 0 the model is effectively pure BM and becomes less so as α increases. We rescaled the phylogeny using three alpha values to examine their impact: α = 0.1 (similar to BM), α = 0.5, and α = 0.9 (very different than BM). Finally, we used an identity matrix (ones on the diagonal and zeros for the off-diagonal correlations) that assumed no phylogenetic correlation at all. We should note that all these correlation matrices (and the identity matrix) are accounting for relatedness at the family level only because of taxonomic incompleteness at the generic or species level. Because of this taxonomic incompleteness, we assume a star phylogeny for species within families.

2.3 Bayesian hierarchical models

For our hierarchical Bayesian model to estimate the probability of infection prevalence within and among members of Lepidoptera , each observation ($N = 1037$)—the number of *Wolbachia*-infected individuals—was nested within species ($S = 411$) and modeled as:

$$infected_{i,j} \sim \text{Binomial}(n_i, \theta_j). \tag{1}$$

where $i = 1, 2, \ldots, 1037$ and $j = 1, 2, \ldots, 419$. Here $infected_{i,j}$ indicates the number of infected individuals from the ith observation of the jth species, n_i is the total number of screened insects in observation i, and θ_j is the probability of infection for species j.

We then assumed the species-level probabilities of infection were normally-distributed with family-level means (μ_k) and standard deviations (σ_k) where $k = 1, 2, \ldots, 28$ families. For computational efficiency, we used a non-centered parameterization of the normal (Papaspiliopoulos et al., 2007). The normal distribution is unconstrained, but θ is bounded between zero and one. Therefore the species-level θs were logit transformed such that:

$$\text{logit}(\theta_j) \sim \text{Normal}(\mu_k, \sigma_k). \tag{2}$$

The mean (μ_k) describes the average probability of infection within a family on the log-odds scale and can be back-transformed using the inverse-logit function.

The standard deviation (σ) measures how much variation in the probability of infection there is across species. If σ is small, then infection probabilities will be similar among species. Conversely, if σ is large, species-specific probabilities of infection will be more idiosyncratic. Data sparsity can be a problem in hierarchical models, especially for the estimation of scale parameters like variances. Because there were several species with few observations, we used shrinkage priors (Carvalho et al., 2009, 2010) for the species-specific σs:

$$\begin{aligned} \sigma_k &= t_\nu^+(0, \tau) \\ \tau &\sim t_\nu^+(0, 1) \end{aligned} \tag{3}$$

where t_3^+ is half-Student-t distribution with $\nu = 3$ degrees of freedom.

We modeled μ, the vector of log-odds infection probabilities for families using a multivariate normal distribution:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \text{MVNormal}(\gamma, \Sigma). \tag{4}$$

with the mean log-odds probability of infection across Lepidoptera (γ) and covariance matrix Σ. To account for phylogenetic non-independence among families, we constructed sigma as:

$$\Sigma = \boldsymbol{\eta}\, \boldsymbol{\Omega}\, \boldsymbol{\eta} \tag{5}$$

where $\boldsymbol{\eta}$ is a $k \times k$ diagonal matrix with the overall standard deviation on the diagonals and $\boldsymbol{\Omega}$ is a $k \times k$ phylogenetic correlation matrix. We then put regularizing priors (to prevent overfitting) on both γ and η:

$$\begin{aligned} \gamma &\sim \text{Normal}(0, 5) \\ \eta &\sim t_\nu^+(0, 5) \end{aligned} \tag{6}$$

where, again, t_3^+ is half-Student-t distribution with $\nu = 3$ degrees of freedom.

Posterior probabilities for model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling in the Stan programming language (Carpenter et al., 2016) via the *RStan* interface (Stan Development Team, 2016). For each model, four MCMC chains were used with 5,000 iterations each. The first 2,500 iterations for each chain were adaptive and discarded as warm-up. We used several diagnostic tests to confirm that each model had reached a stationary distribution, including visual examination of

MCMC chain history, calculation of effective sample size (ESS), and the Gelman-Rubin convergence diagnostic ($\hat{R}$; Gelman and Rubin, 1992; Brooks and Gelman, 1998). In particular, model convergence was assessed by inspecting the diagnostics of the log-posterior density. Model fit was also assessed by posterior predictive checks by simulating "new" data from the posterior distribution and plotting it against the original data (Gelman et al., 2013).

We used WAIC (the widely applicable or Watanabe-Akaike information criterion; Watanabe, 2010; Gelman et al., 2014) to compare models with different phylogenetic correlation matrices (e.g., Brownian motion vs. OU processes) using the *loo* package (Vehtari et al., 2016).

3 RESULTS

After filtering the Weinert et al. (2015) data to contain only Lepidoptera assayed for *Wolbachia* we retained 1037 independent samples from 411 unique species across 28 families, representing a total of 10860 individual assays (Figure 1a). Of these, 3607 samples from 163 species in 19 families were scored PCR positive for *Wolbachia*.

The $\hat{R}$ diagnostic for all parameters (including the log-posterior density) was 1.0, indicating that each model had reached a stationary posterior distribution. Visual assessment of the MCMC chain history confirmed this. Additionally, the effective sample size for the log-posterior density was > 2000 for all models. Predictive plots of the posterior means of the simulated "new" observations regressed against the original observations (Figure 2) resulted in tight concordance, suggesting the models were doing an excellent job at describing the data (all models: $R^2 = 0.91$). If the model provides a perfect fit, the intercept of the slope should be zero and the slope should equal one. For all models, including the averaged consensus model (see below), the regression intercept was 0.41 (0.12 SE). Given that the data ranged from 0–255, the intercept is effectively zero. Additionally, the slope was 0.88 (0.008 SE), quite close to the theoretical optimum of one.

All models, including those which contained phylogenetic correction, had similar WAIC scores with standard errors that completely overlapped (Table 2). We interpret this to mean that each model was in the same "family" of best models. Additionally, the parameter estimates were almost identical across models, and all models predicted a median infection frequency of $\sim 12\%$. Rather than consider each model separately, we created a consensus model using model weighted averaging based on the Δ_{WAIC} scores (Table 2) to describe *Wolbachia* infection frequency in the Lepidoptera.

Our estimate for the median *Wolbachia* infection frequency in the Lepidoptera was 12.1% (95% Highest Density Interval (HDI) = 0.045–0.33; Figure 3). Estimates of median family-level infection frequencies varied considerably with a positive association between sample size and HDI (Figure 4, Supplemental Figure 1). For example, the Nymphalidae (4060 specimens from 236 species) and Lycaenidae (878 specimens from 346 species) had relatively tight posterior distributions: median estimates of infection were 0.037 (95% HDI: 0.015, 0.082) and 0.24 (95% HDI: 0.13, 0.36) respectively. In, contrast families that had small sample sizes (e.g. Bombycidae, Hedylidae, and Lasiocampidae) generated larger intervals reflecting uncertainty in the estimates. For example, the HDI for Hedylidae was 0.06–0.8.

4 DISCUSSION

Our model predicts an average *Wolbachia* infection rate across lepidoptera of approximately 12%. As with Weinert et al. (2015), we consider that there are three main sources of bias in the dataset: 1) some

species are represented by a single sample (Figure 1b), 2) there is a taxonomic bias in the data (Figure 1c), and 3) research may be focused on groups with known *Wolbachia* infections (e.g., Nice et al., 2009). Additionally, We consider a fourth source of bias: some families were extensively sampled among a small number of different species. This may bias the results towards a few members of an otherwise large family of Lepidoptera. We suggest that our estimates of infection rate are lower than previous research because our hierarchical Bayesian model accounts for these biases at each level and therefore may produce more reliable estimates.

We find it interesting that our median infection frequency estimates for the Lepidoptera do not significantly change when the model considers relatedness by incorporating phylogenetic information (Figure 3). Additionally, the model WAIC scores were within 8 units of one another and their standard errors completely overlapped. This implies that the models were all well within the same "family of best models" (Table 2). We interpret these results to indicate that our model is robust to the differential sampling present in the current data set. We consider our estimate for median *Wolbachia* infection rate for the Lepidoptera, and many of the family-level estimates, to be reliable given the limitations of the data. Estimates of infection frequency for those families with large sample sizes are likely accurate, but we must advise caution when interpreting some of the estimates for families with small sample sizes. In these cases where one sample has been assayed for an entire family (Bombycidae, Callidulidae, Eupterotidae, Hedylidae, Lasiocampidae, Pterophoridae, Uraniidae), the estimates presented in Figure 4 are strongly influenced by families with more complete sampling, and are shrunk towards the hyperprior—i.e., the overall probability of infection across Lepidoptera. Therefore, we have little faith in point estimates (i.e., mean, median) for these families and this is reflected in the large HDIs.

In addition to the lower estimate of *Wolbachia* infection for the Lepidoptera, our family level estimates were also lower than those reported by Ahmed et al. (2015) (we restrict our comparison here to those families with larger sample sizes). In these cases, as with our estimates of order level infection, the highest density intervals for our estimates were also much larger and we attribute this to the hierarchical manner in which error is handled. Using linear regression, we observed a strong association between the number of samples assayed per family and the range of the 95% HPD, where small sample sizes generated larger HPD ranges ($F_{1,26} = 12.56, P = 0.0015$).

Our model—essentially a hierarchical Bayesian extension of phylogenetic generalized linear models (Paradis, 2012)—does make a few assumptions with respect to the phylogenetic information used. First, we incorporate the phylogeny at the family level for Lepidoptera. This implicitly assumes equal branch lengths for species within families, which is clearly not the case. However, we were unable to find a fully resolved species-level phylogeny for the Lepidoptera with adequate coverage for many of the families in our dataset, and we did not want to exclude species that were not represented in the phylogeny. Second, by treating the phylogenetic correlation matrix as a fixed covariate, we are unable to fully account for uncertainty in the phylogenetic relationships among taxa. This is an important assumption that could affect our estimates and conclusions, but one that is often made in these kinds of studies (O'Meara, 2012; Paradis, 2012).

Despite these assumptions and limitations, we are confident in our results for two reasons. First, we used correlation matrices from phylogenies assuming both Brownian motion and OU processes, as well as a matrix assuming no phylogenetic correlation at all. Model selection via WAIC indicated that all of these models were in the same family of "best models" (Table 2), and their parameter estimates (and the estimates from model averaging) were all quite similar, suggesting phylogeny is not playing a large role in the prevalence of *Wolbachia* infection among lineages. Second—and more gratifying—is that the posterior predictive checks for all models were quite similar (Figure 2). More important, data simulated from our

model parameters closely matched the empirical data, indicating our approach is doing an excellent job at modeling the underlying biological process.

There are a number of interesting implications of our results. If the infection frequency for species is on the order of 12%, then perhaps *Wolbachia* is not presently a major player in the reproductive manipulation of this order. It follows that—if *Wolbachia* infection frequency is relatively low in the Lepidoptera—its role in the evolution of the order may not be as significant as with other groups (Miller et al., 2010). Accumulating evidence is demonstrating that *Wolbachia* is not an obligate manipulator of a host's reproductive biology (Hamm et al., 2014a; Zhang et al., 2010, 2013). Indeed, Prout (1994) demonstrated that reproductive manipulator microbes should evolve to minimize harm to its host. Perhaps the paradigm needs to be reevaluated.

The assumption that *Wolbachia* always acts as a reproductive manipulator is clearly incorrect (Nice et al., 2009; Hamm et al., 2014b), and one should take care when extrapolating from the results of a positive *Wolbachia* assay to the real world. Luckily, this assumption is fading away. A *Wolbachia* infection can impart benefits to its host—for example the *wSuz* infection of *Drosophila suzukii* confers resistance to certain viruses (Cattel et al., 2016), can provide nutrition to its host (Hosokawa et al., 2010), and does not induce a manipulative phenotype (Hamm et al., 2014a). Thus, *Wolbachia* detection does not and should not imply detrimental effects to the host. Furthermore, our knowledge of *Wolbachia* as a reproductive manipulator in the Lepidoptera is based on scant evidence. To the best of our knowledge, of the 163 species of Lepidoptera considered positive for *Wolbachia*, only seven species from four families have been assayed for an induced phenotype (Table 1). Reciprocal cross experiments are required to determine what—if any—effects *Wolbachia* have on host reproduction. Until these experiments are conducted for a particular system we urge extreme caution when interpreting a positive PCR assay, and we hope that researchers will conduct the necessary experiments to determine if a manipulative phenotype exists within a particular system.

In many respects, the science of microbes in the insects, especially with regards to the endosymbiont *Wolbachia* and the Lepidoptera, is still in its natural history phase wherein research is largely in the descriptive phase, and as such we urge caution when interpreting positive *Wolbachia* assays and extrapolating consequences. Our model provides a framework to further our understanding of *Wolbachia* infection frequency in the Lepidoptera and, as we acquire more data, can generate new estimates. Ultimately, we hope to extend this model to the Insecta once sufficient data are acquired. We also consider that this model can be applied to other problems where one seeks to estimate infection frequency using hierarchical data.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

ZM and CH conceived of the experiment; ZM and CH conducted the analyses; ZM and CH wrote the manuscript.

FUNDING

The authors received no external funding to conduct this research.

ACKNOWLEDGMENTS

We wish to thank James Fordyce, Ben Fitzpatrick, Christopher Peterson, Brian O'Meara and Jeremy Beaulieu for their input and advice during the preparation of this manuscript. We also wish to thank Francis Jiggins and Jack Welch for fruitful discussions while this project was nascent.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder

REFERENCES

Ahmed, M. Z., Araujo-Jnr, E. V., Welch, J. J., and Kawahara, A. Y. (2015). *Wolbachia* in butterflies and moths: geographic structure in infection frequency. *Frontiers in Zoology* 12, 1–9. doi:{10.1186/s12983-015-0107-z}

Baldo, L., Hotopp, J. C. D., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., et al. (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Applied and Environmental Microbiology* 72, 7098–7110. doi:10.1128/AEM.00731-06

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7, 434–455

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2016). Stan: a probabilistic programming language. *Journal of Statistical Software in press*

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, vol. 5. 73–80

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–4–80. doi:{10.1093/biomet/asq017}

Cattel, J., Martinez, J., Jiggins, F., Mouton, L., and Gibert, P. (2016). *Wolbachia*-mediated protection against viruses in the invasive pest *Drosophila suzukii*. *Insect Molecular Biology* , n/a–n/adoi:10.1111/imb.12245

Dyson, E. A., Kamath, M. K., and Hurst, G. D. D. (2002). *Wolbachia* infection associated with all-female broods in *Hypolimnas bolina* (Lepidoptera: Nymphalidae): evidence for horizontal transmission of a butterfly male killer. *Heredity* 88, 166–171. doi:10.1038/sj.hdy.6800021

Fujii, Y., Kageyama, D., Hoshizaki, S., Ishikawa, H., and Sasaki, T. (2001). Transfection of *Wolbachia* in Lepidoptera: the feminizer of the adzuki bean borer *Ostrinia scapulalis* causes male killing in the Mediterranean flour moth *Ephestia kuehniella*. *Proceedings Of The Royal Society Of London Series B-Biological Sciences* 268, 855–859

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science (CRC Press)

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24, 997–1016

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457–472

Hamm, C. A., Begun, D. J., Vo, A., Smith, C. C. R., Saelao, P., Shaver, A. O., et al. (2014a). *Wolbachia* do not live by reproductive manipulation alone: infection polymorphism in *Drosophila suzukii* and *D. subpulchrella*. *Molecular Ecology* 23, 4871–4885

Hamm, C. A., Handley, C. A., and Pike, A. (2014b). *Wolbachia* infection and Lepidoptera of conservation concern. *Journal of Insect Science* 14, 1–8

Hertig, M. (1936). The *Rickettsia*, *Wolbachia pipientis* (gen. et sp. n.) and associated inclusions of the mosquito, *Culex pipiens*. *Parasitology* 28, 453–486

Hertig, M. and Wolbach, S. B. (1924). Studies on *Rickettsia*-like micro-organisms in insects. *The Journal of Medical Research* 44, 329–374

Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A., and 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

Hoffmann, A. A., Montgomery, B. L., Popovici, J., Iturbe-Ormaetxe, I., Johnson, P. H., Muzzi, F., et al. (2011). Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature* 476, 454–457. doi:10.1038/nature10356

Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X.-Y., and Fukatsu, T. (2010). *Wolbachia* as a bacteriocyte-associated nutritional mutualist. *Proceedings of the National Academy of Sciences* 107, 769–774

Ishmael, N., Hotopp, J. C. D., Ioannidis, P., Biber, S., Sakamoto, J., Siozios, S., et al. (2009). Extensive genomic diversity of closely related *Wolbachia* strains. *Microbiology* 155, 2211–2222. doi:10.1099/mic.0.027581-0

Jiggins, F. M., Bentley, J. K., Majerus, M. E., and Hurst, G. D. (2001). How many species are infected with *Wolbachia*? Cryptic sex ratio distorters revealed to be common by intensive sampling. *Proceedings of the Royal Society B-Biological Sciences* 268, 1123–1126

Jiggins, F. M., Hurst, G. D., and Majerus, M. E. (2000). Sex-ratio-distorting *Wolbachia* causes sex-role reversal in its butterfly host. *Proceedings of the Royal Society B-Biological Sciences* 267, 69–73

Jiggins, F. M., Hurst, G. D. D., and Majerus, M. E. N. (1998). Sex ratio distortion in *Acraea encedon* (Lepidoptera: Nymphalidae) is caused by a male-killing bacterium. *Heredity* 81, 87–91

Kageyama, D., Nishimura, G., Hoshizaki, S., and Ishikawa, Y. (2002). Feminizing *Wolbachia* in an insect, *Ostrinia furnacalis* (Lepidoptera : Crambidae). *Heredity* 88, 444–449. doi:10.1038/sj/hdy/6800077

Miller, W. J., Ehrman, L., and Schneider, D. (2010). Infectious speciation revisited: impact of symbiont-depletion on female fitness and mating behavior of *Drosophila paulistorum*. *PLoS Pathogens* 6, e1001214. doi:10.1371/journal.ppat.1001214.s005

Mitsuhashi, W., Fukuda, H., Nicho, K., and Murakami, R. (2004). Male-killing *Wolbachia* in the butterfly *Hypolimnas bolina*. *Entomologia Experimentalis et Applicata* 112, 57–64

Narita, S., Kageyama, D., Nomura, M., and Fukatsu, T. (2007). Unexpected mechanism of symbiont-induced reversal of insect sex: feminizing *Wolbachia* continuously acts on the butterfly *Eurema hecabe* during larval development. *Applied and Environmental Microbiology* 73, 4332–4341. doi:10.1128/AEM.00145-07

Nice, C. C., Gompert, Z., Forister, M. L., and 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

O'Meara, B. C. (2012). Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics* 43, 267–285

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science* 22, 59–73. doi:{10.1214/088342307000000014}

Paradis, E. (2012). *Analysis of Phylogenetics and Evolution with R*, vol. 386 (Springer), 2 edn.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)* 20, 289–290. doi:10.1093/bioinformatics/btg412

Prout, T. (1994). Some evolutionary possibilities for a microbe that causes incompatibility in its host. *Evolution* 48, 909–911

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Regier, J. C., Mitter, C., Zwick, A., Bazinet, A. L., Cummings, M. P., Kawahara, A. Y., et al. (2013). A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS ONE* 8, e58568. doi:10.1371/journal.pone.0058568.s012

Rigaud, T., Pennings, P. S., and Juchault, P. (2001). *Wolbachia* bacteria effects after experimental interspecific transfers in terrestrial isopods. *Journal of Invertebrate Pathology* 77, 251–257. doi:10.1006/jipa.2001.5026

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19, 101–109

Schneider, D. I., Klasson, L., Lind, A. E., and Miller, W. J. (2014). More than fishing in the dark: PCR of a dispersed sequence produces simple but ultrasensitive *Wolbachia* detection. *BMC Microbiology* 14, 121. doi:10.1186/1471-2180-14-121

Simões, P., Mialdea, G., Reiss, D., Sagot, M. F., and Charlat, S. (2011). *Wolbachia* detection: an assessment of standard PCR protocols. *Molecular Ecology Resources* 11, 567–572

Stan Development Team (2016). *RStan: the R interface to Stan, version 2.9.0*

Sugimoto, T. N. and Ishikawa, Y. (2012). A male-killing *Wolbachia* carries a feminizing factor and is associated with degradation of the sex-determining system of its host. *Biology Letters* 8, 412–415. doi:10.1007/s00114-002-0303-5

van Nieukerken, E., Kaila, l., Kitching, I., Kristensen, N., Lees, D., Minet, J., et al. (2011). Order Lepidoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.), Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa* 3148, 212–221

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv (preprint)*

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research* 11, 3571–3594

Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z., and Welch, J. J. (2015). The incidence of bacterial endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society of London B: Biological Sciences* 282, 1–6

Werren, J. H. and 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

Yen, J. H. and Barr, A. R. (1971). New hypothesis of the cause of cytoplasmic incompatibility in *Culex pipiens* L. *Nature* 232, 657–658

Yen, J. H. and Barr, A. R. (1973). Etiological Agent of Cytoplasmic Incompatibility in *Culex pipiens*. *Journal of Invertebrate Pathology* 22, 242–250

Zhang, H., Zhang, K.-J., and Hong, X.-Y. (2010). Population dynamics of noncytoplasmic incompatibility-inducing *Wolbachia* in *Nilaparvata lugens* and its effects on host adult life span and female fitness. *Environmental Entomology* 39, 1801–1809. doi:10.1603/EN10051

Zhang, Y.-K., Zhang, K.-J., Sun, J.-T., Yang, X.-M., Ge, C., and Hong, X.-Y. (2013). Diversity of *Wolbachia* in natural populations of spider mites (genus *Tetranychus*): evidence for complex infection history and disequilibrium distribution. *Microbial Ecology* doi:10.1007/s00248-013-0198-z

Zug, R. and 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.0038544.t001

FIGURES

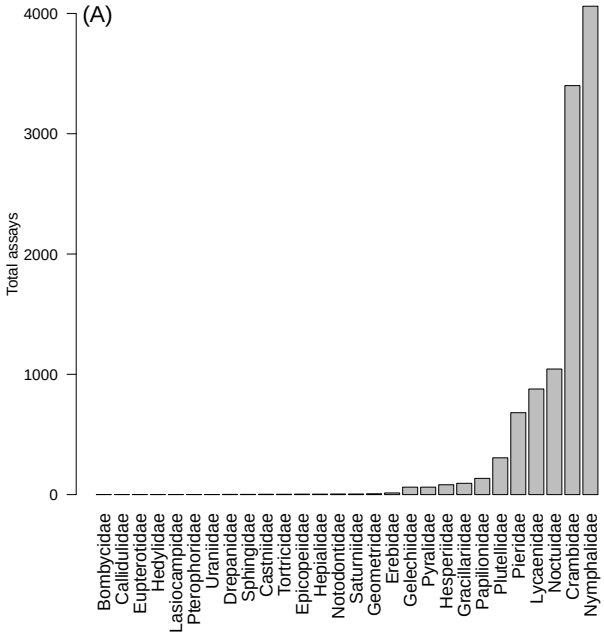

Figure 1a. Total assays by family.

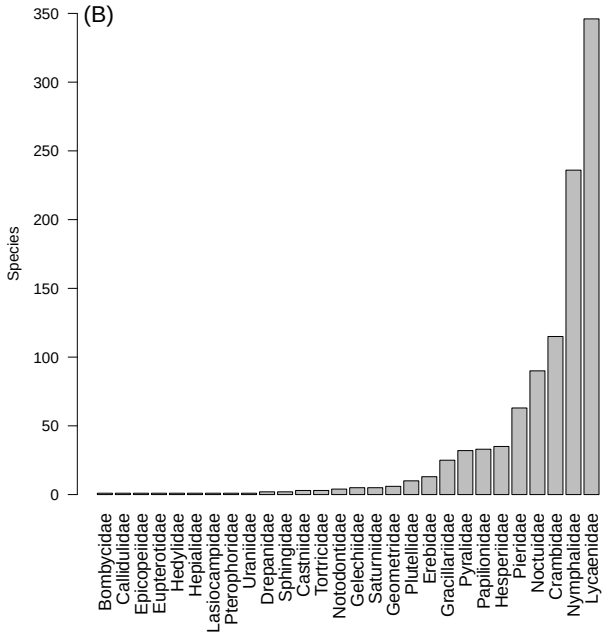

Figure 1b. Number of species sampled by family

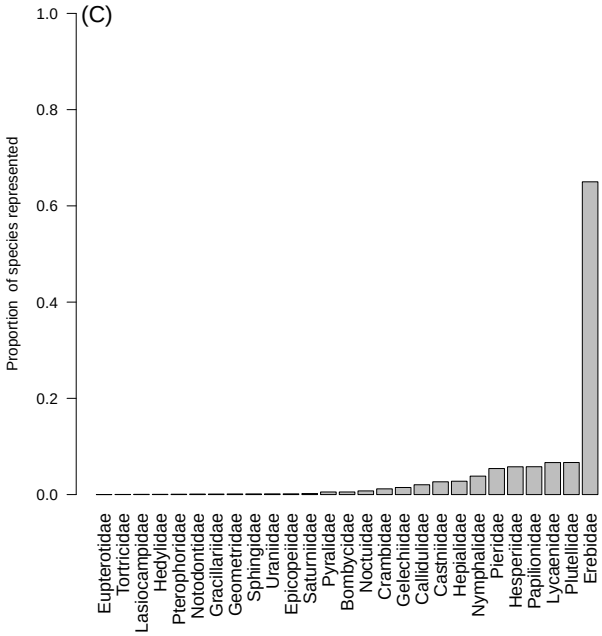

Figure 1c. Proportion of sampling events per family

TABLES

Table 1. Published phenotypic effects of *Wolbachia* on Lepidoptera. Phenotype: MK = male killing, Fem = feminization, CI = cytoplasmic incompatibility. * = induced by transfection with *Wolbachia* strain from *O. scapulalis*.

Species	Family	Phenotype	Reference
Acrea encedana	Nymphalidae	MK	Jiggins et al. (2000)
Acraea encedon	Nymphalidae	MK	Jiggins et al. (1998)
*Ephestia kuehniella**	Pyralidae	MK	Fujii et al. (2001)
Eurema hecabe	Pieridae	CI	Narita et al. (2007)
Hypolimnas bolima	Nymphalidae	MK	Dyson et al. (2002); Mitsuhashi et al. (2004)
Ostrinia scapulalis	Crambidae	MK & Fem	Sugimoto and Ishikawa (2012)
Ostrinia furnacalis	Crambidae	Fem	Kageyama et al. (2002)

Table 2. Models with different phylogenetic correlation structures ranked according to WAIC and their respective model weights. The no-phylogeny model had an identity matrix (ones on the diagonal and zeros on the off-diagonals) in place of a correlation matrix. Smaller WAIC values indicate better estimates. Δ_{waic} is the difference between each WAIC and the lowest WAIC value. SE_{waic} and SE_{Δ} are the standard errors for WAIC and Δ_{waic} respectively.

Model	WAIC	SE_{waic}	p_{waic}	Δ_{waic}	SE_{Δ}	weight
OU: $\alpha = 0.1$	$3,458.7$	318.1	245.2	0.0		0.79
OU: $\alpha = 0.5$	$3,462.8$	319.7	245.7	4.1	2.55	0.10
Brownian Motion	$3,464.1$	320.3	246.6	5.4	3.44	0.05
No Phylogeny	$3,464.9$	320.2	247.9	6.3	3.07	0.03
OU: $\alpha = 0.9$	$3,465.9$	320.1	247.1	7.2	3.13	0.02

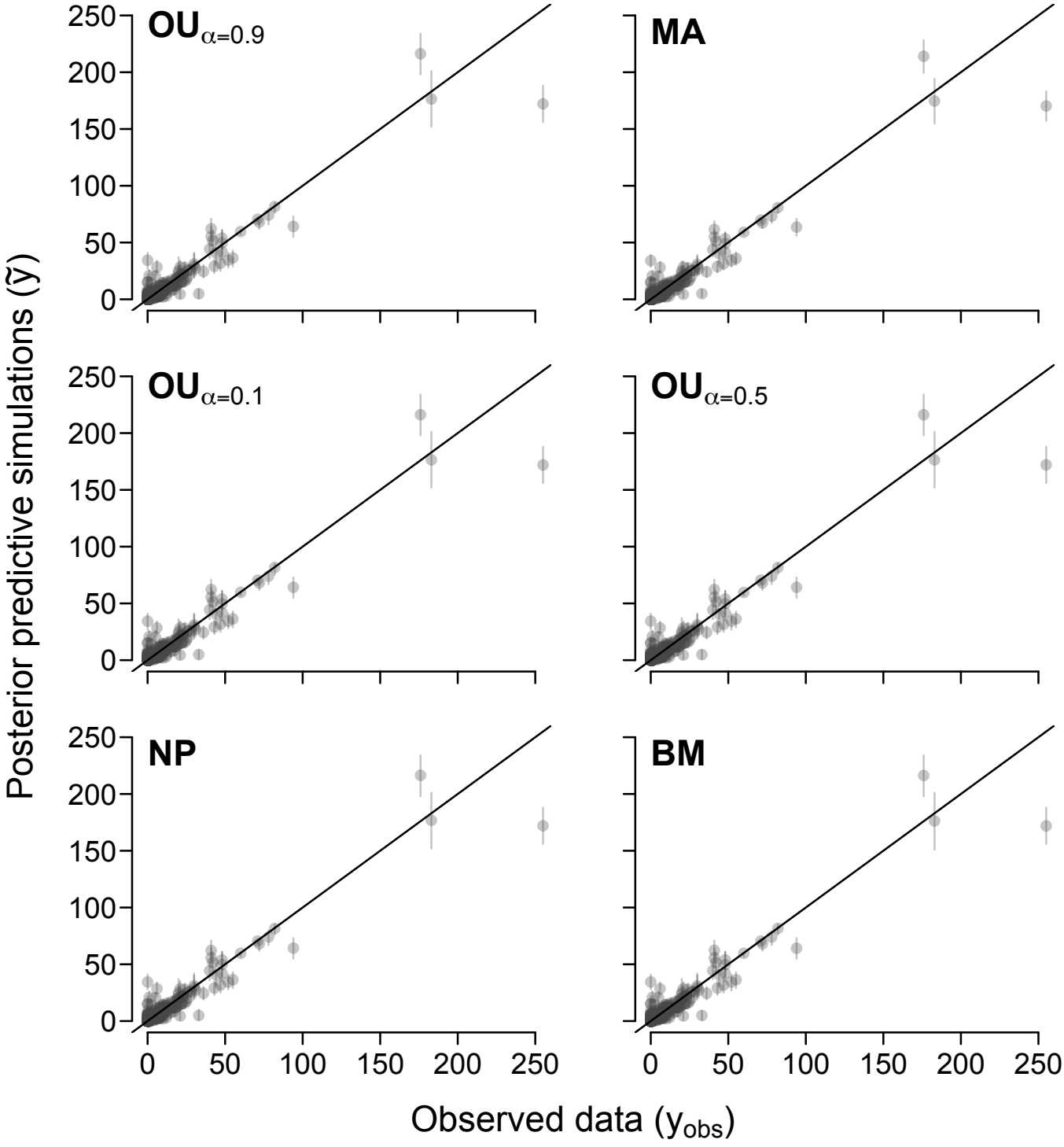

Figure 2. Plots of posterior predictive simulations ($\tilde{y}$) regressed against the observed data (y_{obs}). Points are means of the posterior predictive simulations for each data point, while error bars around each point are the 95% Highest Density Intervals (HDI). Points are partially transparent to show where the majority of the data lie. Ideally, the observed data and the simulated data should have one-to-one correspondence and fall perfectly along a regression line with intercept of zero and slope of one (shown). Different panels represent models with different phylogenetic correlation matrices. **NP** = No Phylogenetic correction; **BM** = Brownian Motion; **OU** = Ornstein-Uhlenbeck with varying levels of α (0.1, 0.5, 0.9); **MA** = WAIC model weighted average.

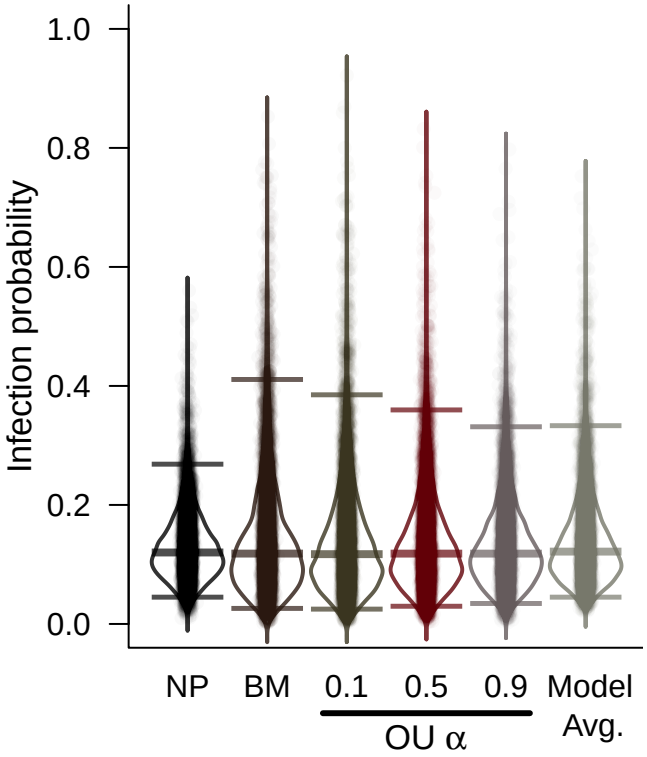

Figure 3. Posterior density plots for the average frequency of *Wolbachia* infection across Lepidoptera. Each posterior estimate is jittered and superimposed on the violins with transparency. Fatter regions of the violins indicate regions of higher posterior density, as do darker regions of jittered points. Horizontal bars indicate the median and upper and lower 95% Highest Density Interval (HDI). Models (L to R): NP = No Phylogenetic correction; BM = Brownian Motion; OU = Ornstein-Uhlenbeck with varying levels of α (0.1, 0.5, 0.9); Model Avg. = WAIC model weighted averaging.

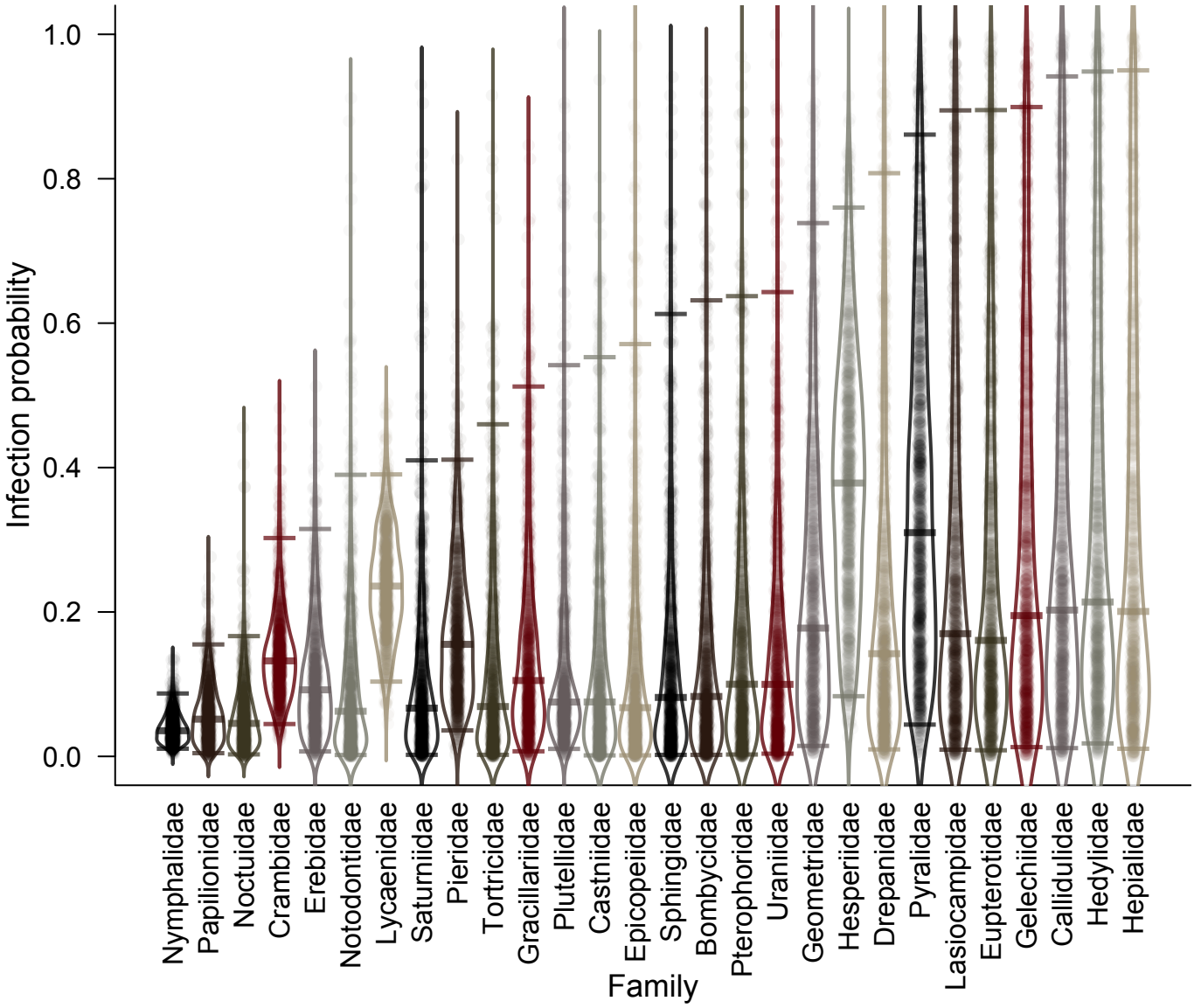

Figure 4. Posterior density plots for the average frequency of *Wolbachia* infection among 28 families of Lepidoptera. Each posterior estimate is jittered and superimposed on the violins with transparency. Fatter regions of the violins indicate regions of higher posterior density, as do darker regions of jittered points. Horizontal bars indicate the median and upper and lower 95% Highest Density Interval (HDI).