A hierarchical Bayesian approach to estimate endosymbiont infection rates

Zachary H. Marion[1,2] and Christopher A. Hamm[3]

May 17, 2016

[1] Corresponding author:
email: zmarion@vols.utk.edu
[2] Department of Ecology and Evolutionary Biology
University of Tennessee
Knoxville, TN 37996
[3] Department of Ecology and Evolutionary Biology
University of Kansas

Invited submission to a special edition of Frontiers in Microbiology

Abstract

Endosymbionts may play an important role in the evolution of the Insecta. Bacteria such as *Wolbachia, Cardinium,* and *Rickettsia* are known to manipulate their host' reproduction to facilitate their own. Indeed, there are many well know 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 this is not an easy endeavor and different approaches have yielded different estimates. One aspect of this problem that may be underappreciated arises when multiple yet independent samples are taken within a species. When independent samples within species are treated as levels of a hierarchy the problem is greatly simplified because error propagates through the model in a realistic and intuitive manner. Here, we present a hierarchical Bayesian approach to estimate infection frequency where multiple independent samples were collected across multiple taxonomic levels. We apply this model to estimate the rates of infection for *Wolbachia* in the Lepidoptera, and apply the model with a correction to account for phylogenetic non-independence. In addition, we highlight the present body of knowledge regarding *Wolbachia* and its effects with regards to the Lepidoptera. Our model estimates suggests that the rate of endosymbiont infection in the Lepidoptera is lower than previously estimated. Given our limited knowledge regarding the phenotypes induced by these endosymbionts, we urge caution when interpreting the results of a positive assays.

1 Introduction

Bacterial endosymbionts have been known to inhabit insects for decades. These endosymbionts are maternally transmitted through the cytoplasm of the egg. *Wolbachia* was the first of these endosymbionts to be discovered when **?** examined the adult ovaries and testes of *Culex pipiens* (hence the specific epithet *Wolbachia pipientis*) **?**. Some years later, **?** observed that male *C. pipiens* from one geographic area may not successfully reproduce with females from a different area, and 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, which was later determined to be *Wolbachia* (**?**).

Contemporary researchers detect the presence of *Wolbachia* via the polymerase chain reaction (PCR). Today, any sample can be screened in a quick, easy, and relatively inexpensive manner (**??**), 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 required such effort that they were employed once a researcher had an *a priori* reason 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. It was under these circumstances that a researcher would observe a likely reproductive manipulation phenotype (*male killing, feminization,* or *parthenogenesis*) and then attribute it to *Wolbachia*.

Careful laboratory work is required to determine what phenotype (if any) is induced by an endosymbiont. With the arrival of PCR and Sanger sequencing it became feasible to conduct exploratory investigations for the presence of *Wolbachia*, though few studies conducted the experimental work to determine if any reproductive manipulation was occurring. The effects of *Wolbachia* infection are complex and depend on an 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 (**??**). Additionally, there may be extensive genomic differences between closely related strains of *Wolbachia* (**?**). Though most famous for its status as

a "reproductive parasite," *Wolbachia* infections have been shown to induce no manipulation at all (**???**). Without careful experimentation, it is not scientific to assume that *Wolbachia* will manipulate a host simply because of a positive PCR assay.

The Lepidoptera (Arthropoda: Insecta) represent the best studied order of animals. Because of historic interest in their physical beauty and their contemporary economic importance the literature is replete with detailed knowledge regarding their distribution and life history. The Lepidoptera is a large Order containing approximately 160,000 species in 124 families, which is approximately 13% all species currently known **?**. In addition to research focused on the Lepidoptera for pure biological reasons, the Lepidoptera are also well represented on lists of endangered or threatened species (**?**). Researchers have tended to focus on certain groups of Lepidoptera, 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.

Experiments to determine if a manipulative phenotype exits have been conducted for six species of Lepidoptera and report that *cytoplasmic incompatibility, male killing,* and *feminization* occur (Table 1). We note that the report of *male killing* in *Ephestia kuhniella* is a result of *Wolbachia* transfected from *Ostrinia scapulalis*. Because of the high level of interest in Lepidoptera research, there is a considered enthusiasm for investigating the role that *Wolbachia* has played in its evolution. A vital first step towards this goal is the estimation of *Wolbachia* infection rates in the Lepidoptera.

ZACH, I'LL NEED TO YOU SET THE LAST PARAGRAPH OF THE INTRO. I CAN'T WRITE ABOUT IT WELL BUT MY ABORTED ATTEMPTS ARE BELOW.

Here, we develop and employ a novel approach to the estimation of *Wolbachia* infection frequencies across the Lepidoptera. Our model explicitly accounts for issues that arise with real world data, such as those relating to estimating infection levels at different scales. 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 considered and accounted for with hierarchical Bayesian approaches that explicitly incorporate phlogenetic correlations.

2 Materials & Methods

2.1 Motivating data and previous analyses

Both **? ?** used a likelihood-based approach to describe the distribution of *Wolbachia* infection across arthropods and Lepidoptera, respectively. Both studies used beta-binomial models to estimate the mean proportion of individuals infected within a given species (**?**). They used the same distribution to calculate the incidence of infection as well, where incidence was the proportion of species infected above a threshold frequency c (i.e.,, one infection in 1000 individuals, or 0.001) (**?**).

In the case of *Wolbachia*, insects screened for this bacterium may either be positive or not positive. It is important to state that not positive is the appropriate state here because an infection could have been missed for a number of reasons, including low density infections (**?**). 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 likelihood deals with error at each level, or rather how it does not. 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, let us say that two individuals from a species were assayed for *Wolbachia*, and one tested positive. For this example, the proportion infected in this species is 0.5, however, 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 (each species), rather than at the level of the study.

2.2 Data

We used the data set synthesized by **?**, which contains records from thousands of individual sampling efforts across the Arthropoda. These data were arranged such that each row represented one independent sampling event (though each row may contain multiple individuals) and contained information on the family, genus, species, endosymbiont genus, number of individuals assayed, and number of positive individuals. We filtered these data such that they contained only *Wolbachia* assays of Lepidoptera. The filtered data set contained 1037 sampling events on 10860 individual Lepidoptera, of which 3607 screened positive for *Wolbachia* infection. We imported these data into the program R v3.2 (R Core Development Team) and there conducted all subsequent analyses. All data and code necessary to reproduce the analyses and figures in this paper are freely available on FigShare (DOI TBD: *NB*, the data will be accessioned to FigShare once the manuscript and code are in their final form).

To correct for any influence of the relatedness among families in our analysis, we used the Lepidoptera phylogeny of **?**, which contained 115 of the 124 families in the order. The tree was pruned to remove duplicate families and those not present in the **?** dataset. We then made the tree ultrametric following the penalized likelihood method of **?** using tools in the *ape* package (**?**). 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 neighboring taxa share that trait due to common ancestry (**?**). We also constructed matrices that assumed trait evolution followed an Ornstein-Uhlenbeck (OU) process, which places constraints around which a character evolves (**?**). 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) (**?**). 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).

2.3 Bayesian hierarchical models

In contrast to **?**, we adopted a hierarchical Bayesian approach to estimate the probability of infection prevalence within and among species of Lepidoptera using a subset of the data from **?**. Each observation ($N = 1037$)—the number of *Wolbachia*-infected individuals—was nested within species ($S = 419$) 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 (**?**). 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 lepidopteran species family
162 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 a shrinkage prior (**??**) 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.

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

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

166 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:

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

168 where $\boldsymbol{\eta}$ is a $k \times k$ diagonal matrix with the overall standard deviation on the diagonals and

$\mathbf{\Omega}$ is a $k \times k$ phylogenetic correlation matrix. We then put regularizing priors on both γ and η:

$$\gamma \sim \text{Normal}(0, 5)$$
$$\eta \sim t^{+}_{\nu}(0, 5) \tag{6}$$

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 (**?**) via the RStan interface (**?**). For each model, four MCMC chains were used with 5,000 iterations each. The first 2,500 iterations for each chain were adaptive and thus 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 and calculation of effective sample size (ESS) and the Gelman and Rubin convergence diagnostic (**??**). In particular, model convergence was assessed by inspecting the diagnostics of the log-posterior density.

We used WAIC (the widely applicable or Watanabe-Akaike information criterion; **??**) to compare models with different phylogenetic correlation matrices (e.g., Brownian motion vs. OU processes) using functions in the loo package (**?**).

3 Results

After filtering the **?** data to contain only Lepidoptera that were screened for *Wolbachia* we retained 1037 independent sampling events of 411 unique species from 28 families, representing a total of 10860 individual assays. Of these, 3607 samples from 163 species and members of 19 families were scored PCR positive for *Wolbachia*.

ZACH, please describe the model results (diagnostics, chain mixing, posterior predictive.

1. WAIC scores among models

2. Median estimate for Order, almost identical regardless of model

3. Discuss family level estimate (is there a correlation between sample size and CI?)

4 Discussion

Previous research on *Wolbachia* in the Lepidoptera has estimated the infection frequency at very high levels, 80% **?** and 77% **?**.

Our model predicts that median *Wolbachia* infection frequency in the Lepidoptera is significantly lower than previous studies have reported.

It is interesting to consider that the median infection frequency estimates for the Lepidoptera do not significantly change when the model considers relatedness by incorporating phylogenetic information. Also, the WAIC scores did not significantly var

Whether we incorporate a phylogeny or not, use pure BM process or an OU with varying alpha values, we get the same answer.

Seven species (out of ¿450 assayed) in four families of Lepidoptera have been assayed for a *Wolbachia* induced phenotype.

In many respects, scientific research with regards to *Wolbachia* is still in its ”natural history” phase, wherein we describe the distribution and effects of infection.

Lower than **?** mean is near **?** and **?**, but variance is much higher

Likely to be differences between male and female infection levels if there is an induced phenotype (indirect evidence). The data presented here pool males and females, which may bias the results.

Bias in sampling must be accounted for, make a plot of families we have within the Lepidoptera.

Given the potential use *Wolbachia* as a potential biological control agent

It is a bit simplistic to just assume that there will be a reproductive manipulation (**??**). We strongly encourage that assays for endosymbionts be conducted prior to the translocation or

interbreeding of insects of conservation concern and, if to conduct controlled matings as guided by the results of those assays.

It could still be in the genome, (example from *Drosophila* and *Nasonia* REFS.

Dont forget **?**, which demonstrated that vertically transmitted reproductive *Wolbachia* should evolve to minimize harm to the host.

?

?

?

Make point that Fem & MK likely manipulate the piRNA - REF!

We conclude that the science of microbes in the Lepidoptera, especially with regards to the endosymbiont *Wolbachia*, is still in its natural history phase wherein discovery is still largely in the descriptive phase, and as such we urge caution when interpreting positive *Wolbachia* assays and extrapolating consequences.

The standard Wolbachia paradigm holds that the endosymbiont is a reproductive parasite that manipulates its hosts reproduction to facilitate its own. The phenotypic effects of infection are known for only a handful of Lepidoptera, given the recent advances in *Wolbachia* research in other taxa (particularly *Drosophila*) which demonstrate that infection does necessarily result in reproductive manipulation, we urge caution when interpreting positive *Wolbachia* assays and extrapolating consequences.

We present a new analysis estimating the infection frequency of *Wolbachia* in the Lepidoptera, which contrasts with a recently published estimate. Our model treats error hierarchically, which we feel is appropriate given the structure of the data.

Slow your ever-loving role which will translate into we urge caution. The model cited in **?** make a lot of assumptions as to how *Wolbachia* works, please note that they did not estimate a phenotype. Wolbachia research in the Lepidoptera is still in the Natural History / Discovery phase

Indirect evidence - Look for signature of reproductive manipulation: biased sex ratios, mito nuclear discord in phylogenies.

Model of *Wolbachia* infection frequency - Handles error hierarchically (just like the data), consider binomial confidence interval for 1+ of 2. Its huge! - Our model is awesome, generates credible intervals that are more reasonable - More like 50% with HPD of 30 60. Plot the posteriors.

Some families only had one sample, in which case there was little information for the model and inferences for these groups were driven by the prior and had large CIs (**Figure 3**).

5 Tables & Figures

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	**?**
Acraea encedon	Nymphalidae	MK	**?**
*Ephestia kuehniella**	Pyralidae	MK	**?**
Eurema hecabe	Pieridae	CI	**?**
Hypolimnas bolima	Nymphalidae	MK	**??**
Ostrinia scapulalis	Crambidae	MK & Fem	**?**
Ostrinia furnacalis	Crambidae	Fem	**?**

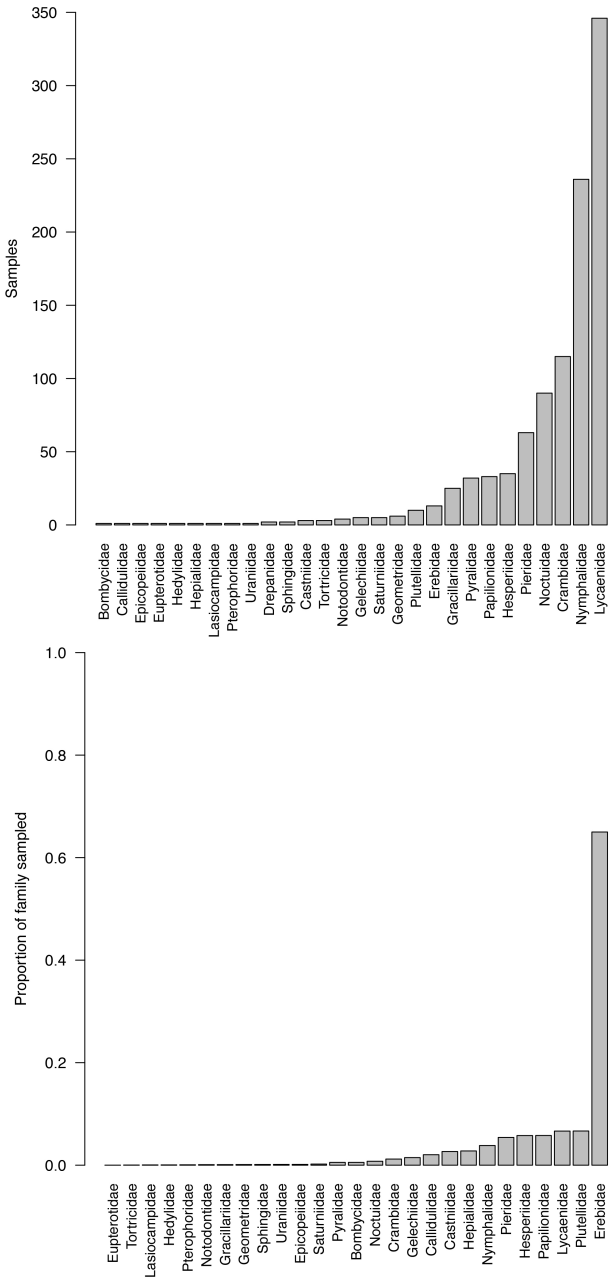

Figure 1: *Wolbachia* sampling by family for total number (Top) and scaled by proportion of species sampled in each family (Bottom)

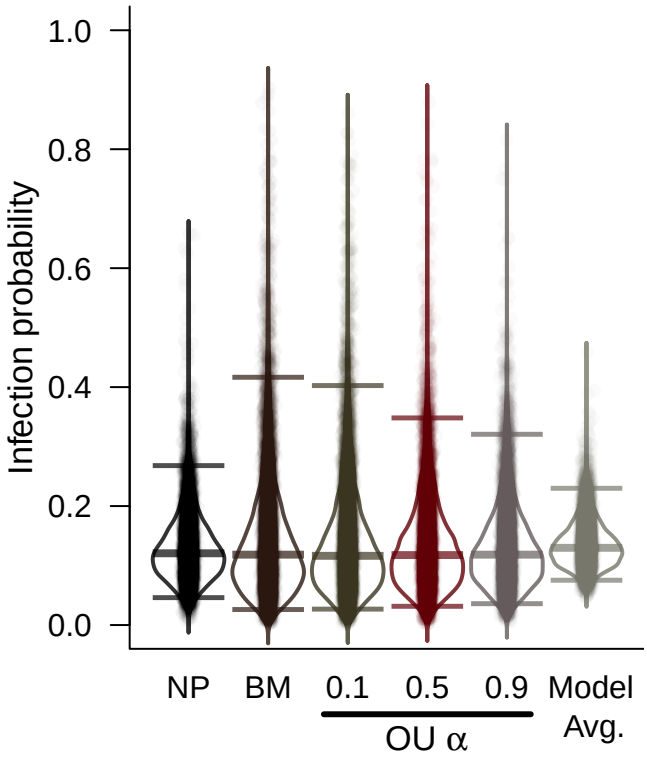

Figure 2: Posterior density plots for the median frequency of *Wolbachia* infection in the Lepidoptera. Models (L to R): NP = No Phylogenetic correction; BM = Brownian Motion; OU = Ornstein-Uhlenbeck with varying levels of α (0.1, 0.5, 0.9); Model Avg = AIC model weighted averaging.

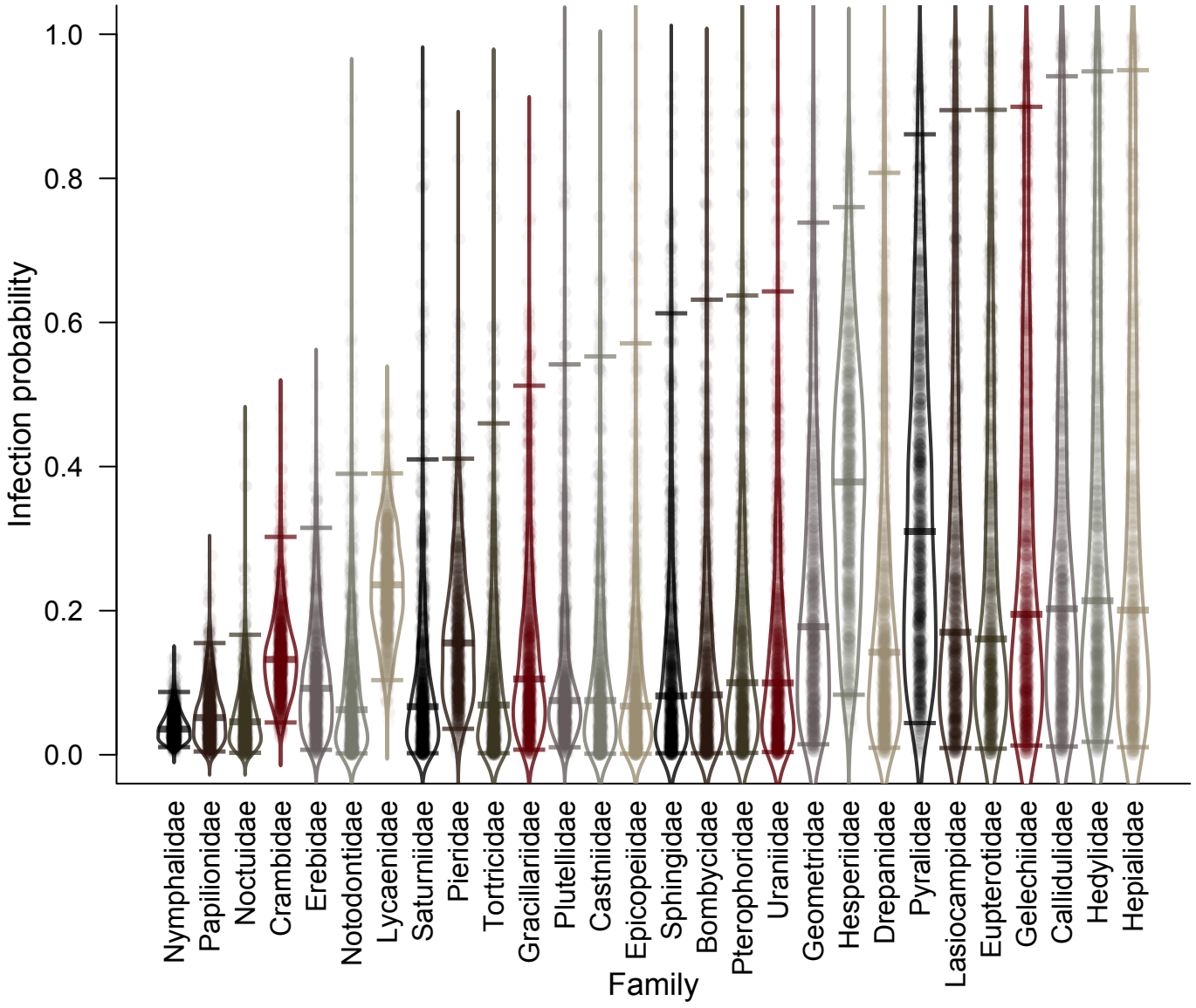

Figure 3: Posterior density plots for the probabilities of *Wolbachia* infection for 28 families of Lepidoptera