Supporting Information

Oaks, J. R., C. W. Linkem, and J. Sukumaran. Implications of uniformly distributed, empirically informed priors for phylogeographical model selection: A reply to Hickerson et al.

1 An error in Hickerson et al.'s re-analysis of the Philippines data

Hickerson et al. (2014) re-analyzed the dataset of Oaks et al. (2013) using a model-averaging approach, where they placed a discrete uniform prior over eight different prior models (see Table 1 of Hickerson et al. (2014)). However, there was an error in their methodology; their model mixes different units of time.

Each of the eight prior models used in the re-analysis by Hickerson et al. (2014) has one of two priors on the mean size of the descendant populations of each taxon pair: $\theta_D \sim U(0.0001, 0.1)$ or $\theta_D \sim U(0.0005, 0.04)$. As described in Oaks et al. (2013), the divergence-time parameters in the model implemented in `msBayes` are in generations scaled relative to a constant reference-population size, θ_C. This reference-population size is defined in terms of the upper limit of the uniform prior on the mean size of the descendant populations, θ_D, such that for the prior $\theta_D \sim U(a_{\theta_D}, b_{\theta_D})$, the size of the constant reference population is $\theta_C = b_{\theta_D}/2$. Thus, the model used by Hickerson et al. (2014) mixes two different units of time. In other words, some of their prior and posterior samples are in units of $0.05/\mu$ generations, whereas others are in units of $0.02/\mu$ generations.

A fundamental assumption of the `msBayes` model and post hoc regression adjustment is that all possible values of the parameter of interest (divergence times) are in the same units. Thus, the results in sections "Using ABC Model Comparison to Weight Alternative Priors for the Philippine Vertebrate Data" and "Improved Sampling Efficiency by Prior Weighting Supports Asynchronous and Recent Divergence for the Philippines Vertebrate Data" and presented in Figure 2 of Hickerson et al. (2014) are invalid and should be disregarded. The error is easily illustrated by re-plotting their results with the different time units indicated (Figure S2).

2 Theoretical implications of empirical priors for Bayesian model choice—A simple example

The distinctions between Bayesian parameter estimation and model choice discussed in the main text can be illustrated with a simple example. Let us say we are interested in the fairness of a particular coin, and we denote the unknown probability of it landing heads as θ. More specifically, we are interested in the probability of two models, M_1 and M_2. In both models the outcomes of flipping the coin are assumed to be binomially distributed, but under M_1 the coin is weighted toward landing heads (i.e., $\theta > 0.5$)), whereas under M_2, the coin is weighted toward landing tails (i.e., $\theta < 0.5$). We already have data from flipping a different coin 20 times that landed both heads and tails 10 times each, and so we decide to

use these data in specifying a beta prior on fairness of the new coin of $beta(a = 10, b = 10)$ (Figure S1). We collect data by flipping the coin of interest $N = 10$ times, $y = 3$ of which land heads. Given the beta distribution is a conjugate prior for a binomial likelihood, the posterior distribution has the nice analytical form $\theta \mid y, N \sim beta(a + y, b + N - y)$, which for the new dataset is simply $beta(13, 17)$ (Figure S1). The maximum a posteriori (MAP) estimate of the probability of heads is 0.429, and following Equation 2 in the main text the marginal likelihoods of our models of interest are

$$p(y = 3, N = 10 \mid M_1) = \int_{0.5}^{1} p(y = 3, N = 10 \mid \theta, M_1) p(\theta \mid M_1) d\theta \approx 0.029, \tag{4}$$

and

$$p(y = 3, N = 10 \mid M_2) = \int_{0}^{0.5} p(y = 3, N = 10 \mid \theta, M_2) p(\theta \mid M_2) d\theta \approx 0.097. \tag{5}$$

Given the models have equal probability under our prior, we can calculate the posterior probability of Model 1 as

$$p(M_1 \mid y = 3, N = 10) = \frac{p(y = 3, N = 10 \mid M_1)}{p(y = 3, N = 10 \mid M_1) + p(y = 3, N = 10 \mid M_2)} \approx 0.23. \tag{6}$$

This is the correct posterior probability of Model 1 given our prior and data.

To give the data more weight relative to the prior, we could use it twice, and calculate an empirical Bayes estimate using a prior of $beta(13, 17)$. This results in a "posterior" distribution of $beta(16, 24)$ (Figure S1), with a MAP estimate of 0.395, and $p(M_1 \mid y = 3, N = 10) = 0.10$. The estimated posterior distribution of the parameter, and resulting MAP estimate, is similar whether or not an empirically informed prior is used. However, the posterior probability of Model 1 is very sensitive to the empirical prior, decreasing by 56%. By using the empirically informed prior, we ignored prior uncertainty, leading to an underestimate of our posterior uncertainty (Figure S1). While this did not greatly affect our estimate of θ, it misled us to be overconfident in Model 2.

3 Validation analyses

Following Oaks et al. (2013), we characterize the model-choice behavior of the model-averaging approach of Hickerson et al. (2014) under the ideal conditions where the prior is correct (i.e., the data are generated from parameters drawn from the same prior distributions used in the analysis). We used the same prior models as above (M_1–M_5; Table 1), and simulated 50,000 datasets under this prior (10,000 from each model). We used a simulated data structure of eight population pairs, with a single 1000 base-pair locus sampled from 10 individuals from each population. We then analyzed each of these replicate datasets using the same prior with 2.5 million samples (500,000 from each of the five prior models), retaining 1000 posterior samples. Our results are very similar to Oaks et al. (2013), but we note that they are not directly comparable as our simulations contained eight population pairs rather than 10 (Figure 8). We find that the approach of Hickerson et al. (2014) estimates

the posterior probability of divergence models reasonably well when all assumptions of the method are met (i.e., the prior is correct) and the unadjusted posterior estimates are used. Similar to Oaks et al. (2013), we find that the regression-adjusted estimates of the model probabilities are biased.

4 A difficult inference problem

In the main text, we discuss how the prior assumption of uniformly distributed divergence times in `msBayes` leads to posteriors that are difficult to interpret. However, it is also important to consider the difficult inference problem with which `msBayes` is faced. When applying `msBayes` to the dataset of Oaks et al. (2013) with 22 taxon pairs, there are 581–602 free parameters that model highly stochastic coalescent and mutational processes. Under this rich stochastic model, the method is estimating the probability of 1002 divergence models (i.e., the number of integer partitions of $Y = 22$; Oaks et al., 2013). Furthermore, all the information in the sequence alignment of each taxon pair is distilled into four summary statistics. This gives us a total of 88 summary statistics (four from each of the 22 taxon pairs) that contain minimal information about many of the ≈ 600 parameters in the model. More summary statistics can be used in `msBayes`, but most are highly correlated with the four default statistics, and thus contribute little additional information about the parameters from the sequence data. The large number of parameters and divergence models relative to the amount of information in the data is undoubtedly another reason the method lacks robustness to prior conditions.

5 Additional clarifications from Hickerson et al. (2014)

5.1 Saturation of summary statistics

Hickerson et al. (2014) claim the priors used by Oaks et al. (2013) "cause much of the explored parameter space to be beyond the threshold of saturation in most mtDNA genes." To explore this possibility, we simulated datasets under prior settings that match two of the three priors used by Oaks et al. (2013): $\theta_D \sim U(0.0005, 0.04)$ and $\theta_A \sim U(0.0005, 0.02)$. Under this prior, we randomly sample divergence-time parameters from a uniform distribution of $U(0, 20)$ coalescent units, simulate datasets, and plot the τ values against the summary statistics calculated from the resulting datasets (Figure 9). Clearly, the priors used by Oaks et al. (2013) with upper limits on τ of five and 10 coalescent units suffered little to no effect from saturation. Even at divergence times of 20 coalescent units, there is still signal in the summary statistics used by `msBayes` (Figure 9). Thus, the assertion of Hickerson et al. (2014) that the priors used by Oaks et al. (2013) sample parameter space in which the mtDNA alignments are saturated by substitutions is incorrect and, as a result, does not explain the bias they found.

5.2 Graphical prior comparisons

Hickerson et al. (2014) advocate the use graphical checks of prior models. This prior-predictive approach entails generating a small number (1000) of random samples from the prior and plotting the resulting summary statistics in comparison to the observed statistics to see if they coincide (see Figure 1 of Hickerson et al. (2014)). Given the richness of the `msBayes` model ($\approx$ 600 parameters for the Philippine dataset analyzed by Hickerson et al. (2014)), we do not expect that 1000 *random* draws from the vast prior parameter space will yield data and summary statistics consistent with the observed data. In fact, when such random draws are tightly clustered around the observed statistics, this can be an indication that the prior is over-fit, as we show in the main text (Table 1 and Figure S3). Thus, using such plots to select priors should be avoided, and the use of posterior-predictive analyses would be much more informative about the overall fit of models.

5.3 Differing utilities of Ψ and Ω in `msBayes`

The primary component of the `msBayes` model is the vector of divergence times for each of the taxon pairs, $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_Y\}$ (Oaks et al., 2013). Hickerson et al. (2014) argue that the dispersion index of this vector, Ω, is a better model-choice estimator than the number of divergence-time parameters within the vector, Ψ. They present a plot of Ψ against Ω (Fig. S1 of Hickerson et al. (2014)), which is essentially a plot of sample size versus variance. This plot shows that Ω has very little information about the number of divergences among taxa. Nonetheless, Hickerson et al. (2014) conclude Ω is more informative and biogeographically relevant than Ψ. However, the number of divergence-time parameters within the vector and their values contains all of the information about the temporal distribution of divergences, and is much more informative than the variance (i.e., the dispersion index is not a sufficient statistic for $\boldsymbol{\tau}$). Hickerson et al. (2014) also argue that `msBayes` can estimate Ω much better than Ψ. However, Oaks et al. (2013) demonstrate that even when all assumptions of the model are met, Ω is a poor model-choice estimator (see plots B, D & F of Figure 4 in Oaks et al. (2013)), whereas Ψ performs better.

Importantly, Ω is limited to estimating the probability of only a single model (the one-divergence model), and thus its utility for model-choice is very limited. I.e., it can only be informative about the probability of whether there is one divergence shared among the taxa ($\Omega = 0.0$) or there is greater than one divergence ($\Omega > 0.0$). As a result, not only is its model-choice utility limited, but it is also very difficult to estimate. Ω can range from zero to infinity, and the point density that it is at its lower limit of zero will always be zero. Thus, an arbitrary threshold (0.01 is used throughout the `msBayes` literature) must be chosen to make the probability of "simultaneous" divergence estimable. Even with this arbitrary threshold, it is still not surprising to see that it is numerically difficult to obtain reliable estimates of the probability that Ω is "near" its lower limit of zero. It is easier, less subjective, and more interpretable to estimate the probability of the model with one divergence-time parameter (i.e., $\Psi = 1$). Thus, it is not surprising that Oaks et al. (2013) find that Ψ is a better estimator of model probability than Ω.

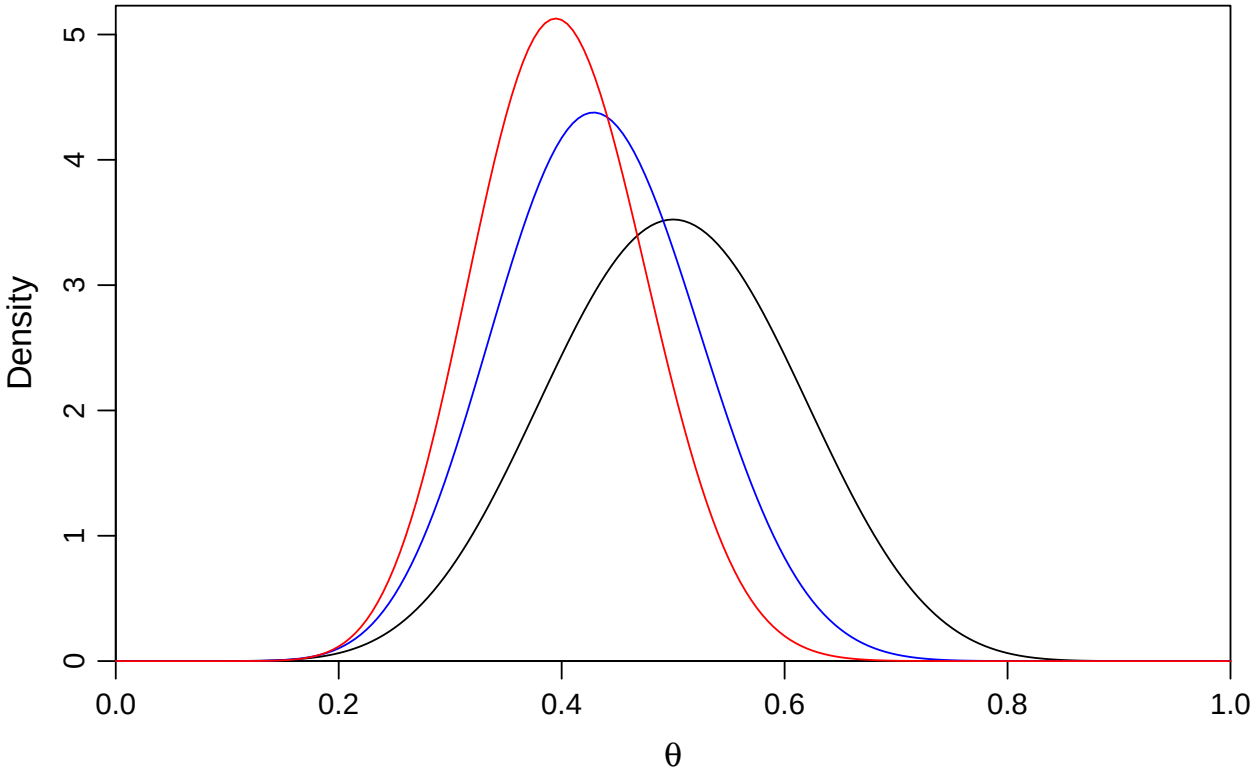

Figure S1: A plot of three beta probability density functions that represent a prior (black; $beta(10, 10)$), posterior (blue; $beta(13, 17)$), and empirical Bayes density (red; $beta(16, 24)$) for a dataset of 10 coin flips, three of which are successes.

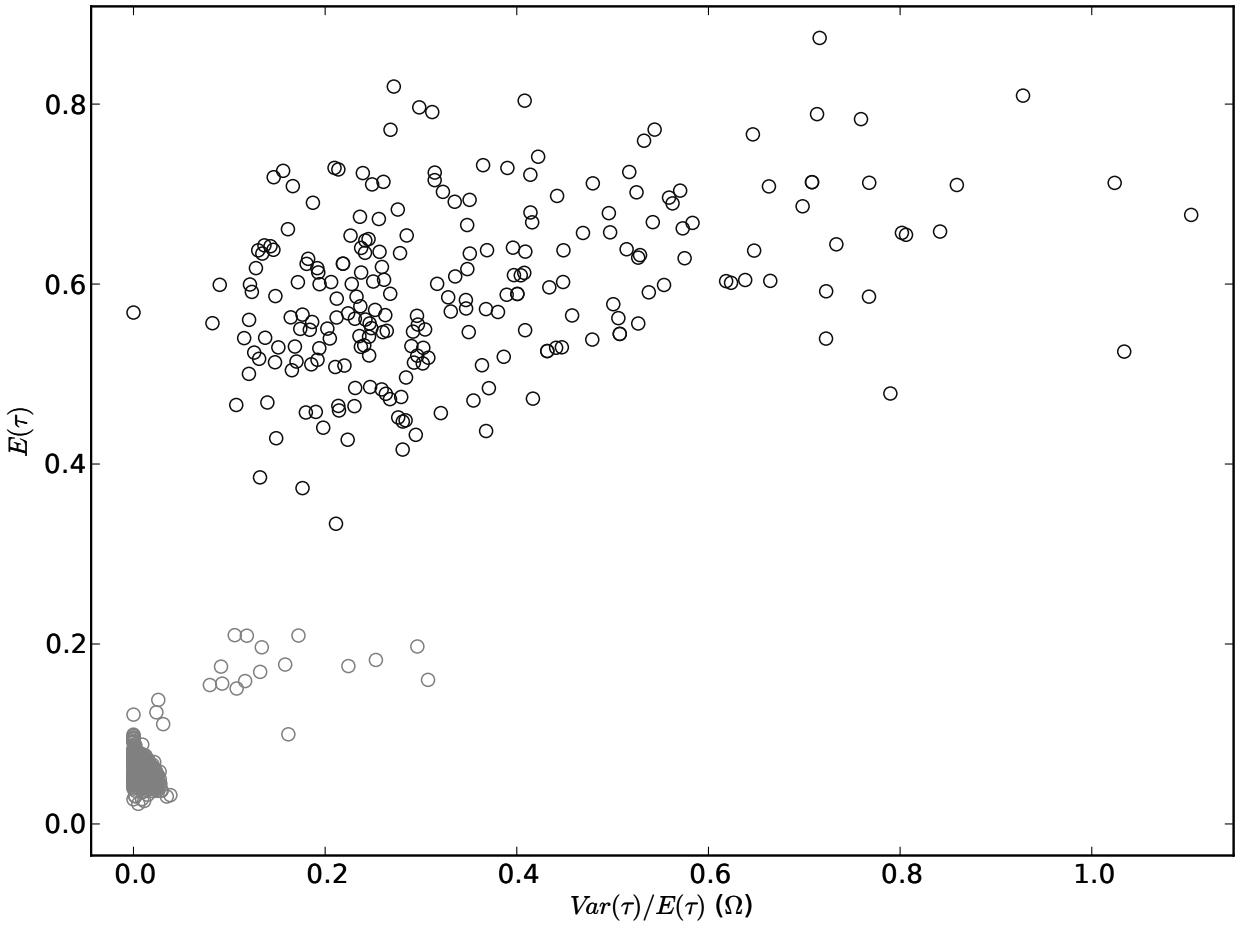

Figure S2: The joint posterior of the mean ($E(\tau)$) and dispersion index ($\Omega = Var(\tau)/E(\tau)$) of divergence times for 22 vertebrate taxon pairs as estimated by Hickerson et al. (2014) (see Figure 2B of Hickerson et al. (2014)). The posterior samples are color-coded to indicate the erroneous mixture of timescales in the analysis of Hickerson et al. (2014); grey = $0.05/\mu$ generations and black = $0.02/\mu$ generations.

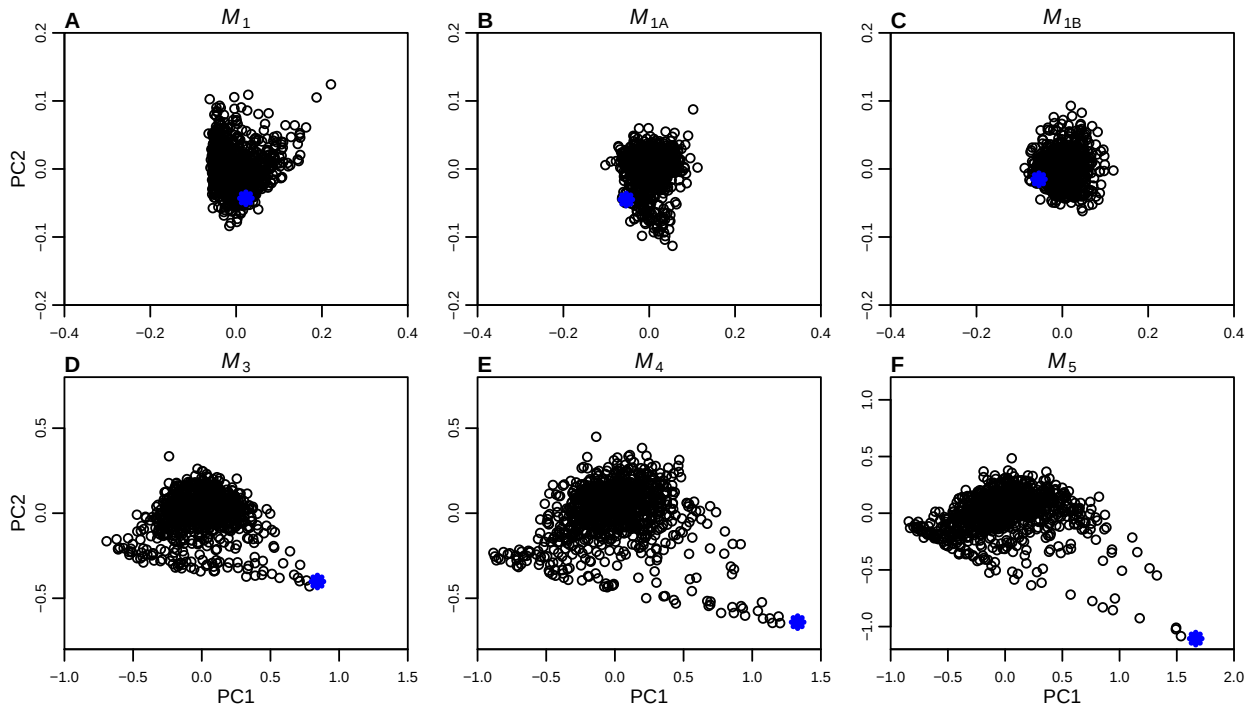

Figure S3: The prior predictive graphical checks recommended by Hickerson et al. (2014) for six prior models: (A) M_1 ($\tau \sim U(0, 0.1)$), (B) M_{1A} ($\tau \sim U(0, 0.01)$), (C) M_{1B} ($\tau \sim U(0, 0.001)$), (D) M_3 ($\tau \sim U(0, 5)$), (E) M_4 ($\tau \sim U(0, 10)$), and (F) M_5 ($\tau \sim U(0, 20)$). The three models that likely exclude true values of some divergence times of the 22 pairs of Philippine taxa (A–C) appear to have a "better fit" than the valid priors that likely cover the true divergence times (D–F). The plots project the summary statistics from 1000 random samples from each model onto the first two orthogonal axes of a principle component analysis, with the blue dot representing the observed summary statistics from the 22 population pairs of Philippine vertebrates.

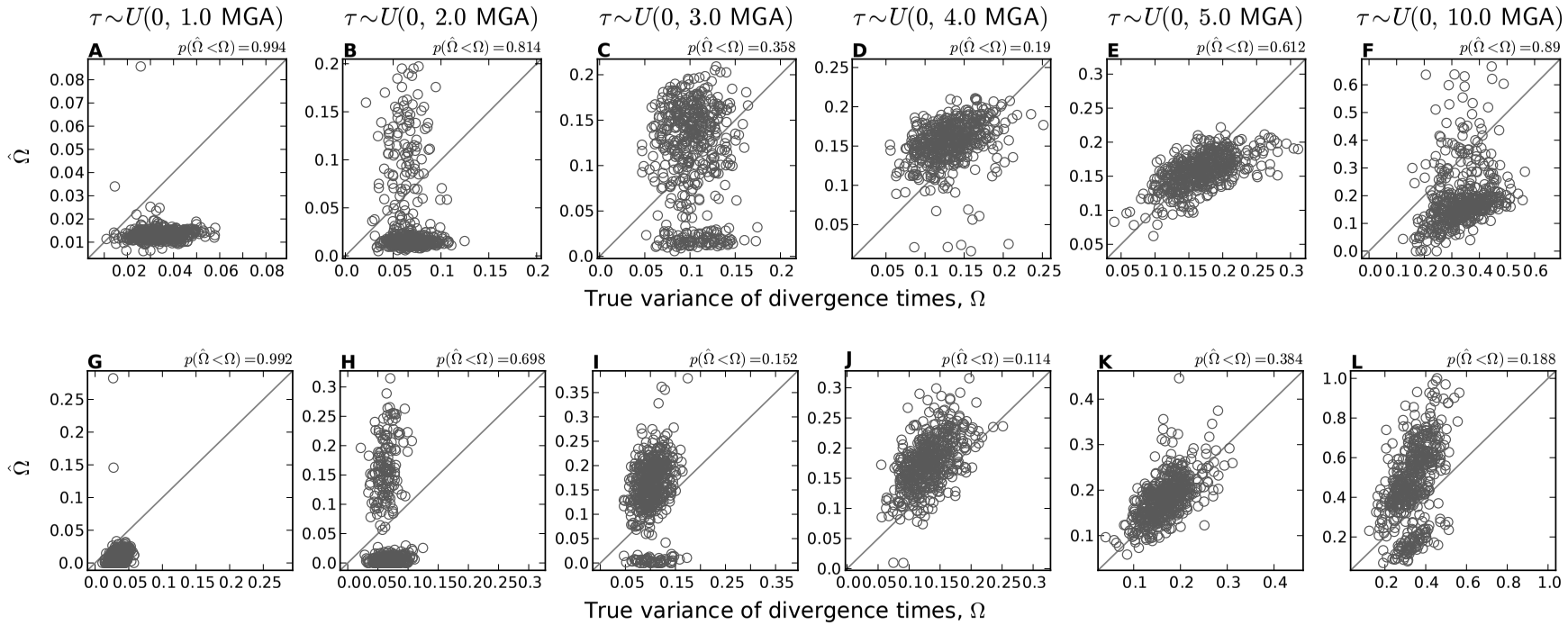

Figure S4: The accuracy of (A–F) unadjusted and (G–L) GLM-adjusted estimates of the dispersion index of divergence times (Ω) when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simlated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions $\tau \sim U(0, \tau_{max})$ indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of $1{\times}10^{-8}$ mutations per generation).

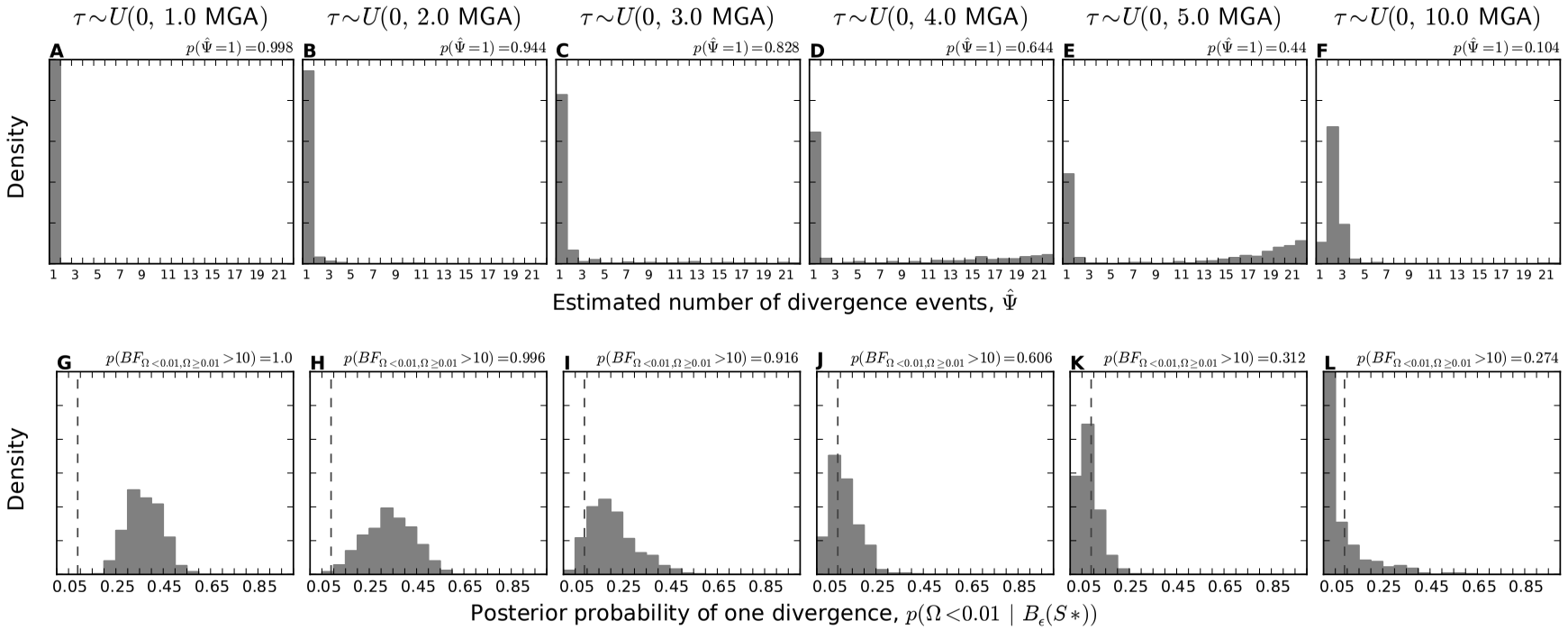

Figure S5: The tendency of the empirically informed model-averaging approach of Hickerson et al. (2014) to (A–F) infer clustered divergences and (G–L) support the extreme model of one divergence when applied to simulated datasets in which the divergence times of 22 pairs of populations are randomly drawn from the uniform distributions $\tau \sim U(0, \tau_{max})$ indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of $1{\times}10^{-8}$ mutations per generation).

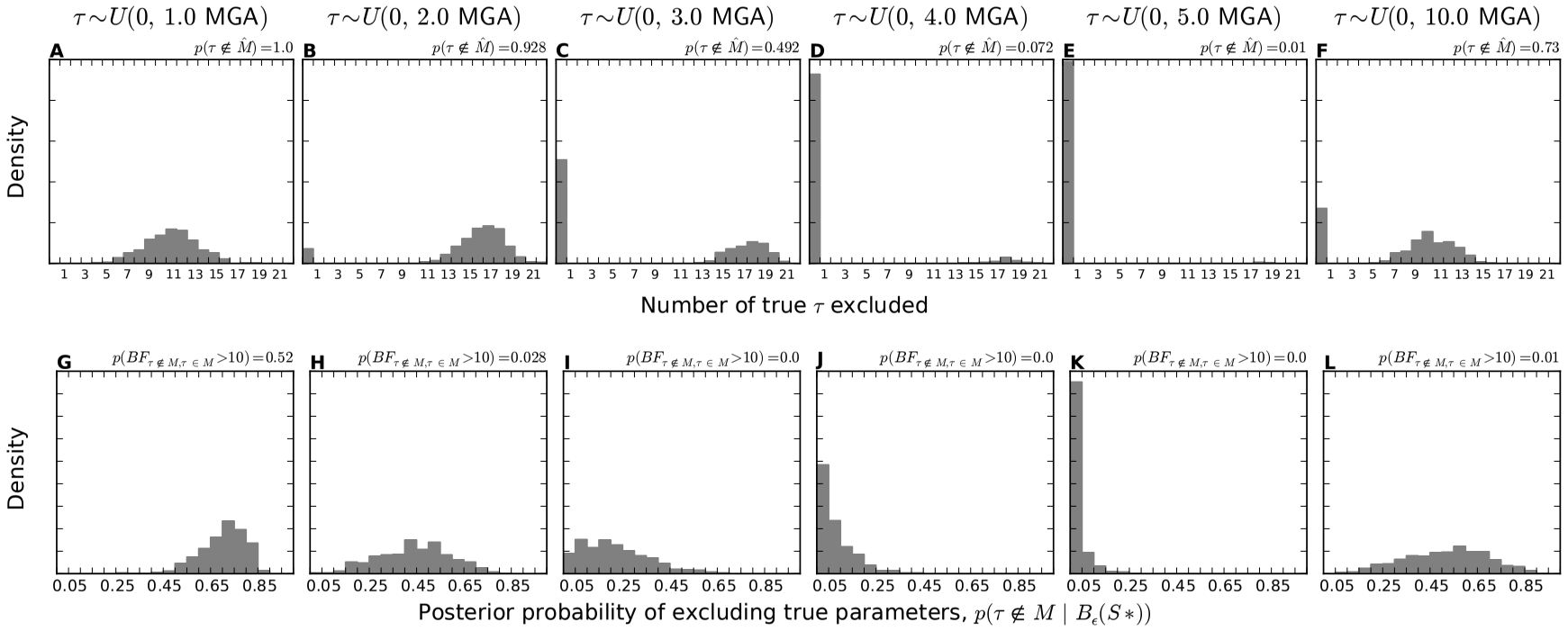

Figure S6: Histograms of the (A–F) number of true divergence times excluded from the preferred model and the (G–L) posterior probability of excluding at least one true divergence time when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simlated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions $\tau \sim U(0, \tau_{max})$ indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of 1×10^{-8} mutations per generation).

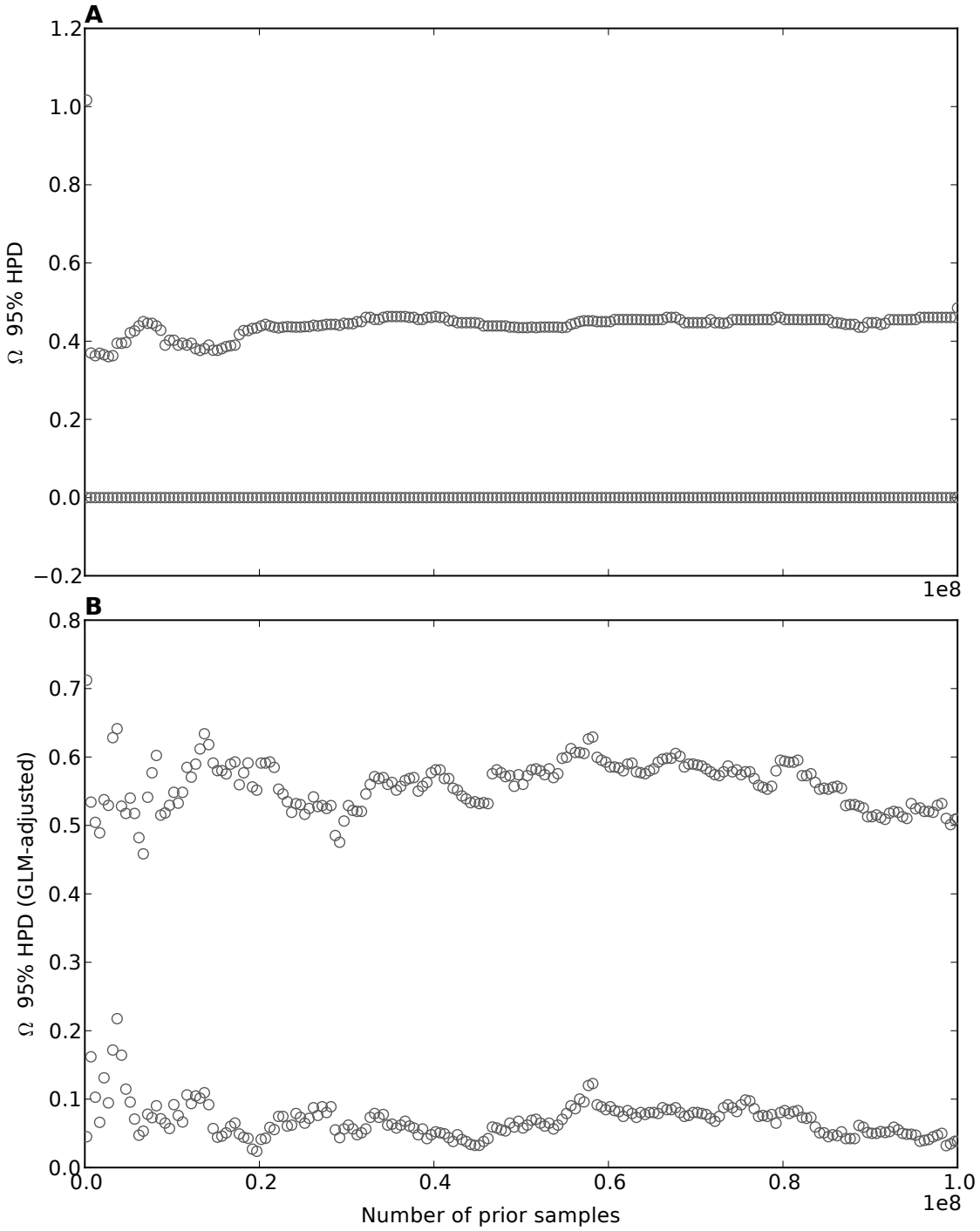

Figure S7: Traces of the estimated lower and upper limits of the 95% highest posterior density (HPD) interval of Ω (the dispersion index of divergence times) as 100 million prior samples are accumulated. Each pair of points is based on 1000 posterior samples retained from the prior. Both (A) unadjusted and (B) GLM-regression-adjusted estimates are shown. The data analyzed were the 22 pairs of Philippine taxa from Oaks et al. (2013). Prior settings were $\tau \sim U(0, 10)$, $\theta_D \sim U(0.0005, 0.04)$, and $\theta_A \sim U(0.0005, 0.02)$.

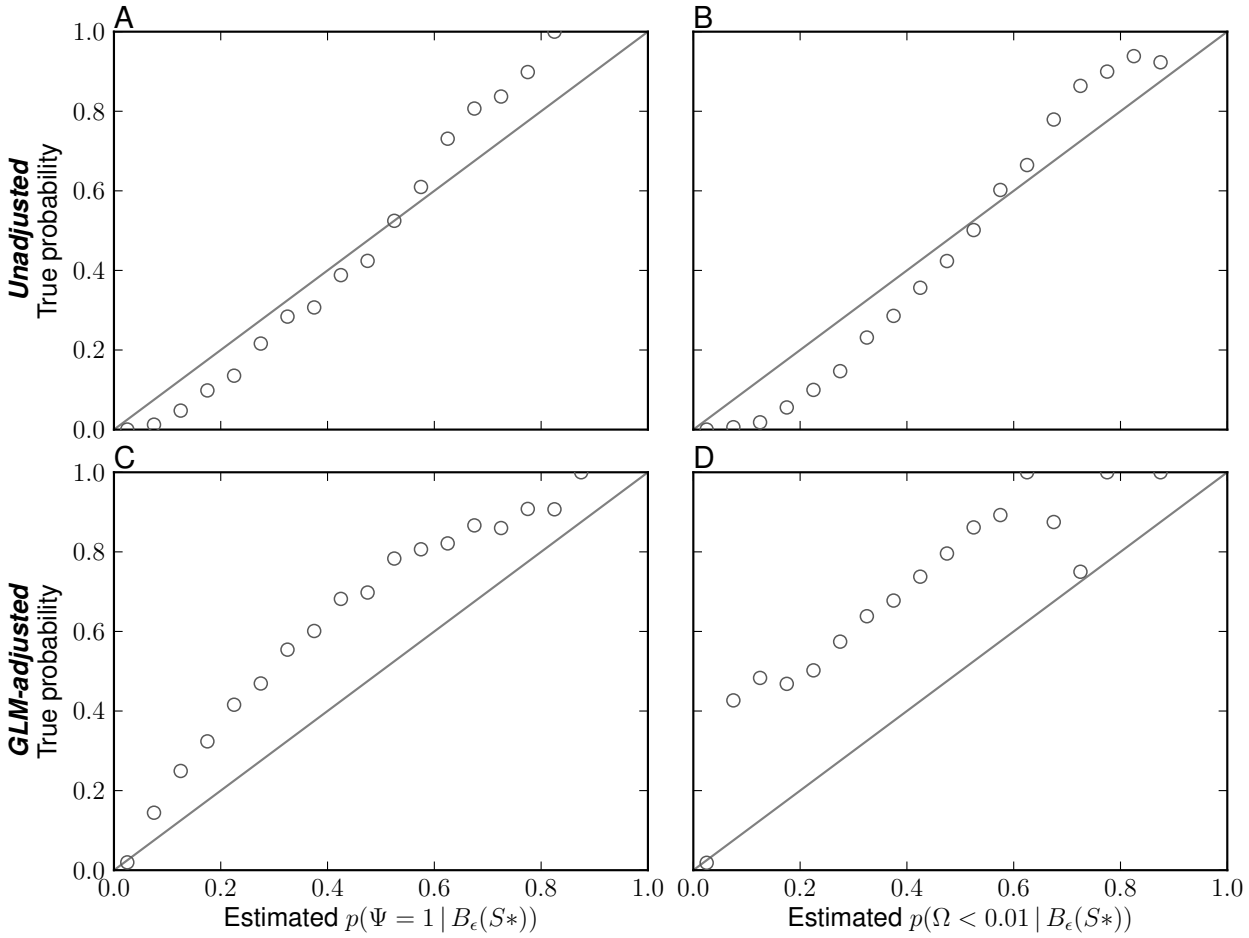

Figure S8: An assessment of the approximate Bayesian model-averageing approach of Hickerson et al. (2014) under the ideal conditions when the prior model is correct (i.e., the datasets are simulated from parameters drawn from the same prior distributions used in the analysis). The plots show the relationship between the estimated posterior and true probability of (A & C) $\Psi = 1$ and (B & D) $\Omega < 0.01$, based on 50,000 simulations. The results summarize the (A & B) unadjusted and (C & D) GLM-adjusted posterior estimate from each simulation replicate. The prior settings for all replicates included five prior models with $\theta_D \sim U(0.0001, 0.1)$ and $\theta_A \sim U(0.0001, 0.05)$ for all five models, and $M_1 : \tau \sim U(0, 0.1)$, $M_2 : \tau \sim U(0, 1)$, $M_3 : \tau \sim U(0, 5)$, $M_4 : \tau \sim U(0, 10)$, and $M_5 : \tau \sim U(0, 20)$. The number of samples from the prior was 2.5×10^6. The simulated data structure was 8 population pairs, with a single 1000 bp locus sampled from 10 individuals from each population. The 50,000 estimates of the posterior probability of one divergence event were assigned to 20 bins of width 0.05. The estimated posterior probability of each bin is plotted against the proportion of replicates in that bin with a true value consistent with one divergence event (i.e., $\Psi = 1$ or $\Omega < 0.01$).

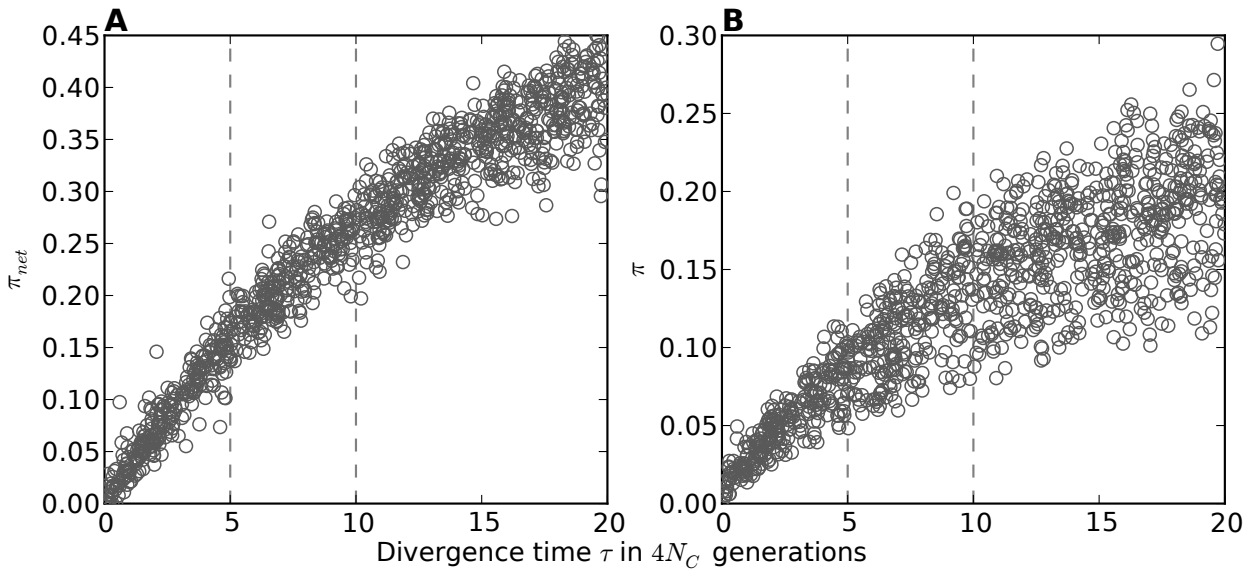

Figure S9: The summary statistics π (Tajima, 1983) and π_{net} (Takahata and Nei, 1985) as a function of divergence time between populations. Each plot represents 1100 pairs of parameter draws and summary statistics calculated from the simulated data. Prior settings for the simulations were $\tau \sim U(0, 20)$, $\theta_D \sim U(0.0005, 0.04)$, and $\theta_A \sim U(0.0005, 0.02)$.

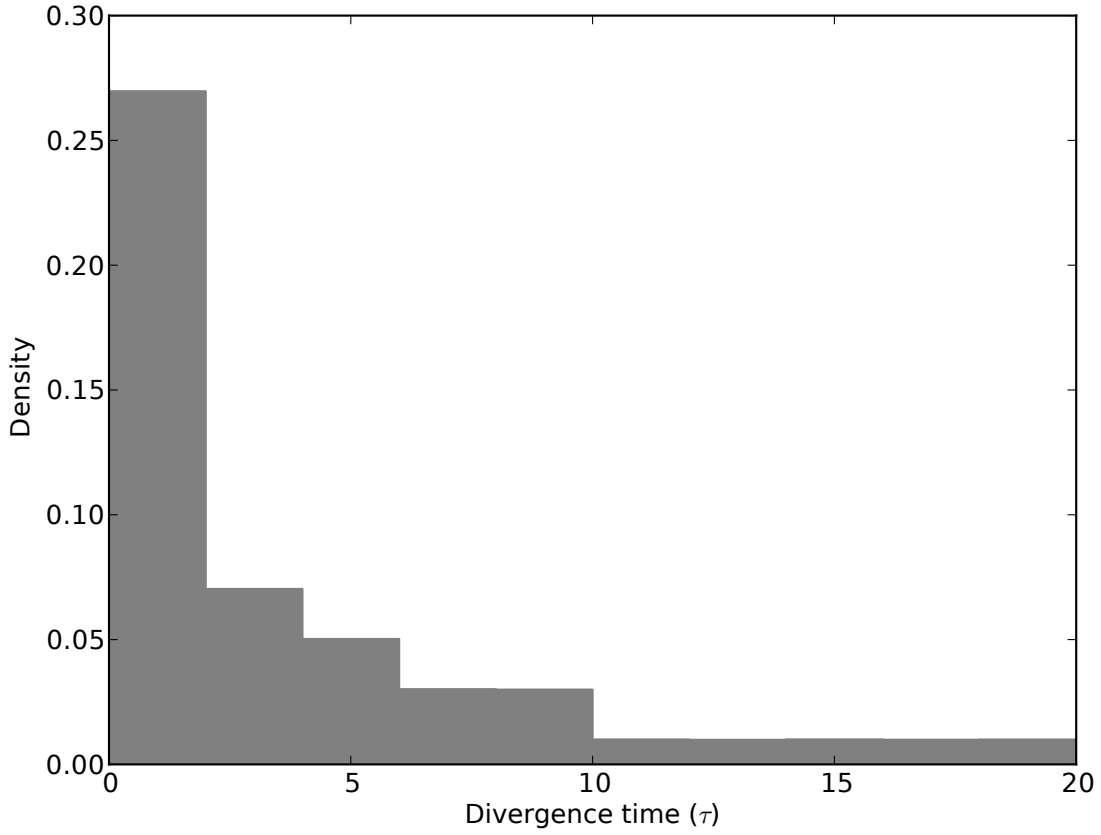

Figure S10: The prior distribution on divergence times imposed by the model-averaging prior comprised of five models with different uniform priors on τ: M_1 ($\tau \sim U(0, 0.1)$), M_2 ($\tau \sim U(0, 1)$), M_3 ($\tau \sim U(0, 5)$), M_4 ($\tau \sim U(0, 10)$), M_5 ($\tau \sim U(0, 20)$).