A Bayesian Model of the DNA Barcode Gap

Jarrett D. Phillips[1,2*] (ORCID: 0000-0001-8390-386X)

[1]*School of Computer Science, University of Guelph, Guelph, ON., Canada, N1G2W1*

[2]*Department of Integrative Biology, University of Guelph, Guelph, ON., Canada, N1G2W1*

***Corresponding Author**: Jarrett D. Phillips[1]

Email Address: jphill01@uoguelph.ca

Running Title: Bayesian inference for DNA barcode gap estimation

Abstract

A simple statistical model of the DNA barcode gap is outlined. Specifically, accuracy of recently introduced nonparametric metrics, inspired by coalescent theory, to characterize the extent of proportional overlap/separation in maximum and minimum pairwise genetic distances within and among species, respectively, is explored in both frequentist and Bayesian contexts. The empirical cumulative distribution function (ECDF) is utilized to estimate probabilities associated with positively skewed extreme tail distribution quantiles bounded on the closed unit interval [0, 1] based on a straightforward binomial distance overlap count. Using R and Stan, the proposed maximum likelihood estimators and Bayesian model are demonstrated on cytochrome *b* (CYTB) gene sequences from two *Agabus* diving beetle species exhibiting limits in the extent of representative taxonomic sampling. Large-sample theory and MCMC simulations show much uncertainty in parameter estimates, particularly when specimen sample sizes for target species are small. Findings highlight the promise of the Bayesian approach using a conjugate beta prior for reliable posterior uncertainty estimation when available data are sparse. Obtained results can shed light on foundational and applied research questions concerning DNA-based specimen identification and species delineation for studies in evolutionary biology and ecology, as well as biodiversity conservation and management, of wide-ranging taxa.

Keywords: Bayesian/frequentist inference, DNA barcoding, intraspecific genetic distance, interspecific genetic distance, specimen identification, species discovery

1 Introduction

The routine use of DNA sequences to support broad evolutionary hypotheses and questions concerning demographic processes like gene flow and speciation in diverse and spatially-distributed taxonomic lineages such as birds, fishes, insects, and arachnids took flight in the late 1980s (Avise et al., 1987). Despite this, the application of genomic data

to applied fields like biodiversity forensics, conservation, and management for the molecular identification of unknown specimen samples came later (*e.g.*, Forensically Informative Nucleotide Sequencing (FINS); Bartlett and Davidson (1992)). Since its inception over 20 years ago, DNA barcoding (Hebert et al., 2003a,b) built on earlier work and has emerged as a robust method of specimen identification and species delimitation across myriad Eukaryotic groups which have been sequenced at short, standardized gene regions like the cytochrome *c* oxidase subunit I (5'-COI) mitochondrial locus for animals. However, the success of the single-locus approach, particularly for regulatory and forensic applications, depends crucially on two important factors: (1) the availability of high-quality specimen records found in public reference sequence databases such as the Barcode of Life Data Systems (BOLD; http://www.barcodinglife.org) (Ratnasingham and Hebert, 2007) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and (2) the establishment of a DNA barcode gap — the notion that the maximum genetic distance observed within species is much smaller than the minimum degree of marker variation found among species (Meyer and Paulay, 2005; Meier et al., 2008). Early work has demonstrated that the presence of a DNA barcode gap hinges strongly on extant levels of species haplotype diversity gauged from comprehensive specimen sampling at wide geographic and ecological scales (Bergsten et al., 2012; Čandek and Kuntner, 2015). Despite this, many taxa lack adequate separation in their pairwise intraspecific and interspecific genetic distances due to varying rates of evolution in both genes and taxa (Pentinsaari et al., 2016). This can compromise rapid matching of unknown samples to expertly-validated references, leading to cases of false positives (taxon oversplitting) and false negatives (excessive lumping of taxa) as a result of incomplete lineage sorting, hybridization/introgression, species synonymy, cryptic species diversity, and misidentifications (Hubert and Hanner, 2015; Phillips et al., 2022).

Recent work has argued that DNA barcoding, in its current form, is lacking in statistical rigor, as most studies rely strongly on heuristic distance-based measures to infer taxonomic identity. Of these studies, few report measures of uncertainty, such as standard errors (SEs)

and confidence intervals (CIs), around estimates of intraspecific and interspecific variation, calling into question the existence of a true species' DNA barcode gap (Čandek and Kuntner, 2015; Phillips et al., 2022). To support this notion, novel nonparametric locus- and species-specific metrics based on the multispecies coalescent (MSC) (Yang and Rannala, 2010, 2017) were recently outlined. The statistics have been shown to hold strong promise for reliable DNA barcode gap assessment when applied to predatory *Agabus* (Coleoptera: Dytiscidae) diving beetles (Phillips et al., 2024). Despite their ease of sampling and well-established taxonomy, this group possesses few morphologically-distinct taxonomic characters that readily facilitate their assignment to the species level (Bergsten et al., 2012). Further, the proposed metrics indicate that sister species pairs from this taxon are often difficult to distinguish on the basis of their DNA barcode sequences (Phillips et al., 2024). Using sequence data from three mitochondrial cytochrome markers (5'-COI, 3'-COI, and cytochrome *b* (CYTB)) obtained from BOLD and GenBank, results highlight that DNA barcoding has been a one-sided argument. Phillips et al.'s 2024 findings point to the need to balance both the sufficient collection of specimens, as well as the extensive sampling of species: DNA barcode libraries are biased toward the latter (Phillips et al., 2024). The coalescent (Kingman, 1982a,b) encompasses a backwards continuous-time stochastic Markov process of allelic sampling within natural, neutrally-evolving, species populations towards the most recent common ancestor (MRCA). The estimators from Phillips et al. (2024) represent a clear improvement over simple, yet arbitrary, distance heuristics such as the 2% rule noted by Hebert et al. (2003a) and the 10× rule (Hebert et al., 2004). The former asserts that DNA sequences differing by at least 2% at sequenced genomic regions should be expected to originate from different biological species, whereas the latter suggests that sequences displaying 10 times more genetic variation among species than within taxa is evidence for a distinct evolutionary origin. In addition, the reliance on visualization approaches, such as frequency histograms, dotplots, and quadrant plots to expose DNA barcoding's limitations, has also been criticized (Collins and Cruickshank, 2013; Phillips et al., 2022). Up until the

work of Phillips et al. (2024), the majority of studies (*e.g.*, Young et al. (2021)) have treated the DNA barcode gap as a binary response. However, given poor sampling depth for most taxa, a Yes/No dichotomy is inherently flawed because it can falsely imply a DNA barcode gap is present for a taxon of interest when in fact no such separation in distances exists. The proposed statistics quantify the extent of asymmetric directionality of proportional distance distribution overlap/separation for species within well-sampled taxonomic genera based on a straightforward distance count, in a similar vein to established measures of statistical similarity such as f-divergence. The metrics can be employed in a variety of ways, including to validate performance of marker genes for specimen identification to the species level (as in Phillips et al. (2024)), as well as to assess whether computed values are consistent with population genetic-level parameters like effective population size (N_e), mutation rates (μ) and divergence times (τ) for species under study in a statistical phylogeographic setting (Knowles and Maddison, 2002; Rannala and Yang, 2003; Mather et al., 2019).

While introduction of the metrics is a step in the right direction, what appears to be missing is a rigorous statistical treatment of the DNA barcode gap. This includes an unbiased way to compute the statistical accuracy of Phillips et al.'s (2024) estimators arising through problems inherent in frequentist maximum likelihood estimation for probability distributions having bounded positive support on the closed unit interval [0, 1]. To this end, here, a Bayesian model of the DNA barcode gap coalescent is introduced to rectify such issues. The model allows accurate estimation of posterior means, posterior standard deviations (SDs), posterior quantiles, and credible intervals (CrIs) for the metrics given datasets of intraspecific and interspecific distances for species of interest.

2 Methods

2.1 DNA Barcode Gap Metrics

The novel nonparametric maximum likelihood estimators (MLEs) of proportional overlap/separation between intraspecific and interspecific distance distributions for a given species (x) to aid assessment of the DNA barcode gap are as follows:

$$p_x = \frac{\#\{d_{ij} \geq a\}}{\#\{d_{ij}\}} \tag{1}$$

$$q_x = \frac{\#\{d_{XY} \leq b\}}{\#\{d_{XY}\}} \tag{2}$$

$$p'_x = \frac{\#\{d_{ij} \geq a'\}}{\#\{d_{ij}\}} \tag{3}$$

$$q'_x = \frac{\#\{d'_{XY} \leq b\}}{\#\{d'_{XY}\}} \tag{4}$$

where d_{ij} are distances within species, d_{XY} are distances among species for an entire genus of concern, and d'_{XY} are combined interspecific distances for a target species and its closest neighbouring species. The notation # reflects a count. Quantities a, a', and b correspond to $\min(d_{XY})$, $\min(d'_{XY})$, and $\max(d_{ij})$, the minimum interspecific distance, the minimum combined interspecific distance, and the maximum intraspecific distance, respectively (**Figure 1**). Hence, Equations (1)-(4) are simply empirical partial means of distances falling at and below, or at and exceeding, given distribution thresholds. Notice further that a/a', and b are also the first and nth order statistics, $X_{(1)}$ and $X_{(n)}$, respectively. Equations (1)-(4) can also be expressed in terms of empirical cumulative distribution functions (ECDFs) (see next section). Distances form a continuous distribution and are easily computed from a model of DNA sequence evolution, such as uncorrected or corrected p-distances (Jukes and Cantor, 1969; Kimura, 1980); however, values are *not* independent and identically distributed (IID). The approach of Phillips et al. (2024) differs markedly from the traditional definition

of the DNA barcoding gap laid out by Meyer and Paulay (2005) and Meier et al. (2008) in that the proposed metrics incorporate interspecific distances which include the target species of interest. Furthermore, if a focal species is found to have multiple nearest neighbours, then the species possessing the smallest average distance is used. These schemes more accurately account for species' coalescence processes inferred from contemporaneous samples of DNA sequences leading to instances of barcode sequence sharing, such as interspecific hybridization/introgression events (Phillips et al., 2024). Within equations (3) and (4), the degree of distance distribution overlap between a target taxon and its nearest neighbouring species, gauged from magnitudes of p'_x and q'_x, is directly proportional to the amount of time in which the two lineages diverged from the MRCA (Phillips et al., 2024). Thus, the quantities can be used as a criterion to assess the failure of DNA barcoding in recently radiated taxonomic groups, among other plausible biological explanations. Note, distances are constrained to the unit interval [0, 1], whereas the metrics are defined only on the interval $[a/a', b]$. Values of the estimators obtained from equations (1)-(4) close to or equal to zero give evidence for separation between intraspecific and interspecific distance distributions; that is, values suggest the presence of a DNA barcode gap for a target species. Conversely, values near or equal to one give evidence for distribution overlap; that is, values likely indicate the absence of a DNA barcode gap.

2.2 The Model

Before delving into the derivation of the proposed DNA barcode gap metrics, review of some fundamental statistical theory is necessary.

For a given random variable X, its cumulative distribution function (CDF) is defined by

$$F_X(t) = \mathbb{P}(X \leq t) = 1 - \mathbb{P}(X > t). \tag{5}$$

Rearranging Equation (5) gives

$$\mathbb{P}(X > t) = 1 - F_X(t), \tag{6}$$

from which it follows that

$$\mathbb{P}(X \geq t) = 1 - F_X(t) + \mathbb{P}(X = t). \tag{7}$$

Equations (1)-(4) can thus be expressed in terms of ECDFs as follows, since the true underlying CDFs, $F(\cdot)$, are unknown *a priori*, and therefore must be estimated using available data:

$$\begin{aligned} p_x &= \mathbb{P}(d_{ij} \geq a) \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \end{aligned} \tag{8}$$

$$\begin{aligned} q_x &= \mathbb{P}(d_{XY} \leq b) \\ &= \hat{F}_{d_{XY}}(b) \end{aligned} \tag{9}$$

$$\begin{aligned} p'_x &= \mathbb{P}(d_{ij} \geq a') \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') \end{aligned} \tag{10}$$

$$\begin{aligned} q'_x &= \mathbb{P}(d'_{XY} \leq b) \\ &= \hat{F}_{d'_{XY}}(b) \end{aligned} \tag{11}$$

From this, it can be seen that $\hat{F}_{d_{ij}}(b) = 1$ in Equations (8) and (10). Given n increasing-ordered data points, the (discrete) ECDF, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[x_i \leq t]}$, comprises a step function having jump discontinuities of size $\frac{1}{n}$ at each sample observation (x_i), excluding ties (or steps of weight $\frac{i}{n}$ with duplicate observations), where $\mathbb{1}(x)$ is the indicator function. Note, $\mathbb{P}(X = t) \neq 0$. Equations (8)-(11) clearly demonstrate the asymmetric directionality of the proposed metrics. Furthermore, calculation of the DNA barcode gap estimators is convenient as they implicitly account for total distribution area (including overlap).

A major criticism of large sample (frequentist) theory is that it relies on asymptotic properties of the MLE (whose population parameter is assumed to be a fixed but unknown quantity), such as estimator normality and consistency as the sample size approaches infinity. This problem is especially pronounced in the case of binomial proportions (Newcombe, 1998). The estimated Wald standard error (SE) of the sample proportion, is given by $\widehat{SE[\hat{p}]} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p} = \frac{Y}{n}$ is the MLE, Y is the total number of successes ($Y = \sum_{i=1}^{n} y_i$) and n is the total number of trials (*i.e.*, sample size). However, the above formula for the standard error is problematic for several reasons. First, it is a Normal approximation which makes use of the central limit theorem (CLT); thus, large sample sizes are required for reliable estimation. When few observations are available, SEs will be large and inaccurate, leading to low statistical power to detect a true DNA barcode gap when one actually exists. Further, resulting interval estimates could span values less than zero or greater than one, or have zero width, which is practically meaningless. Second, when proportions are exactly equal to zero or one, resulting SEs will be exactly zero, rendering $\widehat{SE[\hat{p}]}$ given above completely useless. In the context of the proposed DNA barcode gap metrics, values obtained at the boundaries of their support are often encountered. Therefore, reliable calculation of SEs is not feasible. Given the importance of sufficient sampling of species genetic diversity for DNA barcoding initiatives, a different statistical estimation approach is necessary.

Bayesian inference offers a natural path forward in this regard since it allows for straightforward specification of prior beliefs concerning unknown model parameters and

permits the seamless propagation of uncertainty, when data are lacking and sample sizes are small, through integration with the likelihood function associated with true generating processes. The posterior distribution ($\pi(\theta|Y)$) is given by Bayes' theorem up to a proportionality $\pi(\theta|Y) \propto \pi(Y|\theta)\pi(\theta)$, where θ are unobserved parameters, Y are known data, $\pi(Y|\theta)$ is the likelihood, and $\pi(\theta)$ is the prior. As a consequence, because parameters are treated as random variables, Bayesian models are much more flexible and generally more easily interpretable compared to frequentist approaches. Under the Bayesian paradigm, entire posterior distributions, along with their summaries (*e.g.*, CrIs) are outputted, rather than just long run behaviour reflected in sampling distributions, p-values, and CIs as in the frequentist case, thus allowing direct probability statements to be made.

Essentially, from a statistical perspective, the goal herein is to nonparametrically estimate probabilities corresponding to extreme tail quantiles for positive highly skewed distributions on the unit interval (or any closed subinterval thereof). Here, it is sought to numerically approximate the extent of proportional overlap/separation of intraspecific and interspecific distance distributions within the subinterval $[a/a', b]$. This is a challenging computational problem within the current study as detailed in subsequent sections. The usual approach employs kernel density estimation (KDE), along with numerical or Monte Carlo integration and invocation of extreme value theory (EVT); however, this requires careful selection of the bandwidth parameter, among other considerations. This becomes problematic when fitting finite mixture models where nonidentifiability is rampant. For DNA barcode gap estimation, this would correspond to a two-component mixture (one for intraspecific distance comparisons, and the other for interspecific comparisons), with one or more curve intersection points between components, and the presence of zero distance inflation. This makes parameter estimation difficult using methods like the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Here, for simplicity, a different route is taken to avoid these obstacles. Counts, y, of overlapping distances (as expressed in the numerator of Equations (1)-(4)) are treated as binomially distributed with expectation $\mathbb{E}[Y] = k\theta$, where $k = \{N, C\}$

are total count vectors of intraspecific and combined interspecfic distances, respectively, for a target species along with its nearest neighbour species, and $k = M$ is a total count vector for all interspecific species comparisons. This follows from the fact that the ECDF is binomially distributed. The quantity $\theta = \{p_x, q_x, p_x^{'}, q_x^{'}\}$.

The metrics encompassing θ are presumed to follow a Beta(α, β) distribution, with real shape parameters α and β, which is a natural choice of prior on probabilities. The beta distribution has a prior mean of $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$ and a prior variance equal to $\mathbb{V}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. In the case where $\alpha = \beta$, all generated Beta(α, β) distributions will possess the same prior expectation, whereas the prior variance will shrink as both α and β increase. Such a scheme is quite convenient since the beta distribution is conjugate to the binomial distribution. Thus, the posterior distribution is also beta distributed, specifically, Beta($\alpha+Y$, $\beta+n-Y$), having expectation $\mathbb{E}[\theta|Y] = \frac{\alpha+Y}{\alpha+\beta+n}$ and variance $\mathbb{V}[\theta|Y] = \frac{(\alpha+Y)(\beta+n-Y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$. In the context of DNA barcoding, it is important that the DNA barcode gap metrics effectively differentiate between extremes of no overlap/complete separation and complete overlap/no separation, corresponding to values of the metrics equal to 0 and 1 (equivalent to total distance counts of 0 and n), respectively. These extremes yield a posterior expectation of $\mathbb{E}[\theta|Y = 0] = \frac{\alpha}{\alpha+\beta+n}$ and a posterior variance of $\mathbb{V}[\theta|Y] = 0] = \frac{\alpha(\beta+n)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$ and $\mathbb{E}[\theta|Y = n] = \frac{\alpha+n}{\alpha+\beta+n}$ and $\mathbb{V}[\theta|Y = n] = \frac{(\alpha+n)\beta}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$. Note, the posterior variances are equivalent at these thresholds for all $\alpha = \beta$.

Parameters were given an uninformative Beta(1, 1) prior, which is equivalent to a standard uniform (Uniform(0, 1)) prior since it places equal probability on all parameter values within its support. This distribution has an expected value of $\mu = \frac{1}{2}$ and a variance of $\sigma^2 = \frac{1}{12}$. Further, the posterior is Beta($Y + 1$, $n - Y + 1$), from which various moments such as the expected value $\mathbb{E}[Y] = \frac{Y+1}{n+2}$ and variance $\mathbb{V}[Y] = \frac{(Y+1)(n-Y+1)}{(n+2)^2(n+3)}$, and other quantities, can be easily calculated. Clearly, $\mathbb{E}[\theta|Y = 0] = \frac{1}{n+2}$ and $\mathbb{V}[\theta|Y = 0] = \frac{n+1}{(n+2)^2(n+3)}$, and $\mathbb{E}[\theta|Y = n] = \frac{n+1}{n+2}$ and $\mathbb{V}[\theta|Y = n] = \frac{n+1}{(n+2)^2(n+3)}$. In general however, when possible, it is always advisable to incorporate prior information, even if only weak, rather than

simply imposing complete ignorance in the form of a flat prior distribution. In the case of unimodal distributions, the (estimated) posterior mean often possesses the property that it readily decomposes into a convex linear combination, in the form of a weighted sum, of the (estimated) prior mean and the MLE. That is $\hat{\mu}_{\text{posterior}} = w\hat{\mu}_{\text{prior}} + (1-w)\hat{\mu}_{\text{MLE}}$, where for the beta distribution, $w = \frac{\alpha+\beta}{\alpha+\beta+n}$. Therefore, with sufficient data, $w \to 0$ as $n \to \infty$, regardless of the values of α and β, and the choice of prior distribution becomes less important since the posterior will be dominated by the likelihood. For the Beta(1, 1), $w = \frac{2}{2+n}$, with $n = 2$ giving $w = \frac{1}{2}$; that is, the posterior is the arithmetic average of the prior and the likelihood. The full Bayesian model for species x is thus given by

$$\begin{aligned} y_{\text{lwr}} &\sim \text{Binomial}(N, p_{\text{lwr}}) \\ y_{\text{upr}} &\sim \text{Binomial}(M, p_{\text{upr}}) \\ y^{'}_{\text{lwr}} &\sim \text{Binomial}(N, p^{'}_{\text{lwr}}) \\ y^{'}_{\text{upr}} &\sim \text{Binomial}(C, p^{'}_{\text{upr}}) \\ p_{\text{lwr}}, p_{\text{upr}}, p^{'}_{\text{lwr}}, p^{'}_{\text{upr}} &\sim \text{Beta}(1, 1). \end{aligned} \tag{12}$$

Note that p_x, q_x, $p^{'}_x$, and $q^{'}_x$ in Equations (1)-(4) are denoted p_{lwr}, p_{upr}, $p^{'}_{\text{lwr}}$, $q^{'}_{\text{upr}}$ within Equation (12) for distinction between MLEs and Bayesian posterior estimates. The above statistical theory and derivations lay a good foundation for the remainder of this paper.

The proposed model is inherently vectorized to allow processing of multiple species datasets simultaneously. Model fitting was achieved using the Stan probabilistic programming language (Carpenter et al., 2017) framework for Hamiltonian Monte Carlo (HMC) via the No-U-Turn Sampler (NUTS) sampling algorithm (Hoffman and Gelman, 2014) through the `rstan` R package (version 2.32.6) (Stan Development Team, 2023) in R (version 4.4.1) (R Core Team, 2024). Four Markov chains were run for 2000 iterations each in parallel across four cores with random parameter initializations. Within each chain, a total

of 1000 samples was discarded as warmup (*i.e.*, burnin) to reduce dependence on starting conditions and to ensure posterior samples are reflective of the equilibrium distribution. Further, 1000 post-warmup draws were utilized per chain. Because HMC/NUTS results in dependent samples that are minimally autocorrelated, chain thinning is not required. Each of these reflect default Markov Chain Monte Carlo (MCMC) settings in Stan to control both bias and variance in the resulting draws. All analyses in the present work were carried out on a 2023 Apple MacBook Pro with M2 chip and 16 GB RAM running macOS Ventura 13.2. A random seed was set to ensure reproducibility of model results. Outputted estimates were rounded to three decimal places of precision. Posterior distributions were visualized as KDE plots using the `ggplot2` R package (version 3.5.1) (Wickham, 2016) with the default Gaussian kernel and optimal smoothness selection.

Convergence was assessed both visually and quantitatively as follows: (1) through examining parameter traceplots, which depict the trajectory of accepted MCMC draws as a function of the number of iterations, (2) through monitoring the Gelman-Rubin $\hat{R}$ statistic (Gelman and Rubin, 1992; Vehtari et al., 2021), which measures the concordance of within-chain *versus* between-chain variance, and (3) through calculating the effective sample size (ESS) for each parameter, which quantifies the number of independent samples generated Markov chains are equivalent to. Mixing of chains was deemed sufficient when traceplots looked like "fuzzy caterpillars", $\hat{R} < 1.01$, and effective sample sizes were reasonably large (Gelman et al., 2020). After sampling, a number of summary quantities were reported, including posterior means, posterior SDs, and posterior quantiles from which 95% CrIs could be computed to make probabilistic inferences concerning true population parameters.To validate the overall correctness of the proposed statistical model given by Equation (12), as a means of comparison, posterior predictive checks (PPCs) were also employed to generate binomial random variates in the form of counts from the posterior predictive distribution; that is $\gamma = \{Np_x, Mq_x, Np_x^{'}, Cq_x^{'}\}$ to verify that the model adequately captures relevant features of the observed data. The proposed Bayesian model outlined here has a straightforward

interpretation (**Table 1**).

3 Results and Discussion

The *Agabus* CYTB dataset analyzed by Phillips et al. (2024) is revisited herein. Specifically, the proposed Bayesian model is demonstrated on the species *A. bipustulatus* and *A. nevadensis*, since these taxa were the sole representatives for this locus, with the most and the least specimen records, respectively ($N = 701$ and $N = 2$) across all three assessed molecular markers. Note, DNA barcode gap estimation is only possible for species having at least two specimen records. This dataset is a prime illustrative example highlighting the issue of inadequate taxon sampling, which arises frequently in large-scale phylogenetic and phylogeographic studies, in several respects. First, from a statistical viewpoint, sample sizes reflect extremes in reliable parameter estimation. Second, from a DNA barcoding perspective, *Agabus* comprises about 200 extant species according to the Global Biodiversity Information Facility (GBIF) (https://www.gbif.org); yet, due to the level of convenience sampling inherent in taxonomic collection efforts for this genus, adequate representation of species and genetic diversity is far from complete.

MCMC parameter traceplots showed rapid mixing of chains to the stationary distribution (**Figure 2**). Further, all $\hat{R}$ and ESS values (not shown) were close to their recommended cutoffs of one and thousands of samples, respectively, indicating chains are both well-mixed and have converged to the posterior distribution.

Bayesian posterior estimates were reported alongside frequentist MLEs, in addition to SEs, posterior SDs, 95% CIs and 95% CrIs (**Table 2**). CIs were calculated using the usual large sample $(1-\alpha)100\%$-level interval estimate given by $\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $z_{1-\frac{\alpha}{2}} = 1.960$ for 95% confidence and α is the stated significance level (here, 5%). Given a $(1-\alpha)100\%$ CI, with repeated sampling, on average $(1-\alpha)100\%$ of constructed intervals will contain the true parameter of interest; on the other hand, any given CI will either capture or exclude the true

parameter with 100% certainty. This in stark contrast to a CrI, where the true parameter is contained within said interval with $(1-\alpha)100\%$ probability. Note, by default Stan computes equal-tailed (central) CrIs such that there is equal area situated in the left and right tails of the posterior distribution. For a 95% CrI, this corresponds to the 2.5th and 97.5th percent quantiles. However, constructed intervals are usually only valid for symmetric or nearly symmetric distributions. Given the bounded nature of the DNA barcode gap metrics, whose posterior distributions, as expected, show considerable skewness, a different approach to reporting CrIs, such as Highest Posterior Density (HPD) intervals (Chen and Shao, 1999) or shortest probability intervals (SPIn) (Liu et al., 2015) is warranted. As such asymmetric intervals generally attain greater statistical efficiency (in the form of smaller Mean Squared Error (MSE) or variance) and higher coverage probabilities than more standard interval estimates, careful in-depth comparison is left for future work.

Findings based on nonparametric MLEs and Bayesian posterior means were quite comparable with one another and show evidence of complete overlap in intraspecific, interspecific, and combined interspecific distances for *A. bipustulatus* in both the p/q and p'/q' directions since the metrics attain magnitudes very close to one (**Table 2**). As a result, this likely indicates that no DNA barcode gap is present for this species. Such findings are strongly reinforced by the very tight clustering of posterior draws (**Figure 3**) and associated interval estimates owing to the large number of specimens sampled for this species. On the other hand, the situation for *A. nevadensis* is more nuanced, as posterior values are further spread out (**Table 2** and **Figure 4**), suggesting less overall certainty in true parameter values given the low specimen sampling coverage for this taxon. Of note, the 95% CIs and 95% CrIs are quite wide for *A. nevadensis*, consistent with much uncertainty regarding the computed frequentist and Bayesian posterior means of the DNA barcode gap metrics. For instance, the Bayesian analysis for *A. nevadensis* suggests that the data are consistent with both p_{lwr} and p'_{lwr} ranging from approximately 0.250-1.000. Further, regarding the frequentist analysis for the same species, the 95% CI for q'_x extends to negative values at the left endpoint, due

to the corresponding SE of 0.070 being too high as a result of the extremely low sample size of $n = 2$ individuals sampled (**Table 2**). Since the 95% CI truncated at the lower endpoint includes the value of zero, the null hypothesis for the presence of a DNA barcode gap cannot be rejected. Despite this, it is worth noting that truncation is not standard statistical practice and will likely lead to an interval with less than 95% nominal coverage. In such cases, more appropriate confidence interval methods like the Wilson score interval, the exact (Clopper-Pearson) interval, or the Agresti-Coull interval should be employed (Newcombe, 1998; Agresti and Coull, 1998). KDEs for *A. bipustulatus* are strongly left (negatively) skewed (**Figure 5**), whereas those for *A. nevadensis* exhibit more symmetry, especially for p_{upr} and $p^{'}_{\text{upr}}$ (**Figure 6**). These differences are likely due to the stark contrast in sample sizes for the two examined species. Nevertheless, simulated counts of overlapping specimen records from the posterior predictive distribution (**Table 3**) were found to be very close to observed counts for both species, indicating that the proposed model adequately captures underlying variation. Obtained results suggest that use of the Beta(1, 1) prior may not be appropriate given a low number of collected individuals for most taxa in DNA barcoding efforts. This suggests that further consideration of more informative beta priors is worthwhile.

4 Conclusion

Herein, the accuracy of the DNA barcode gap was analyzed from a rigorous statistical lens to expedite both the curation and growth of reference sequence libraries, ensuring they are populated with high quality, statistically defensible specimen records fit for purpose to address standing questions in ecology, evolutionary biology, management, and conservation. To accomplish this, recently proposed, easy to calculate nonparametric MLEs were formally derived using ECDFs and applied to assess the extent of overlap/separation of distance distributions within and among two species of predatory water beetles in the genus *Agabus* sequenced at CYTB using a Bayesian binomial count model with conjugate beta priors.

Findings highlight a high level of parameter uncertainty for *A. nevadensis*, whereas posterior estimates of the DNA barcode gap metrics for *A. bipustulatus* are much more certain. Based on these results, it is imperative that specimen sampling be prioritized to better reflect actual species boundaries.

Since the DNA barcode gap metrics often attain values very close to zero (suggesting no overlap and complete separation of distance distributions) and/or very near one (indicating no separation and complete overlap), in addition to more intermediate values, a noninformative Beta($\frac{1}{2}$, $\frac{1}{2}$) prior may be more appropriate over complete ignorance imposed by a Beta(1, 1) prior. The former distribution is U-shaped symmetric and places greater probability density at the extremes of the distribution due to its heavier tails, while still allowing for variability in parameter estimates within intermediate values along its domain. Note that this prior is Jeffreys' prior density (Jeffreys, 1946), which is proportional to the square root of the Fisher information $\mathcal{I}(\theta)$. That is $\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$. Jeffreys' prior has several desirable statistical properties as a prior: that it is inversely proportional to the standard deviation of the binomial distribution, and most notably, that it is invariant to model reparameterization (Gelman et al., 2014). However, this prior can lead to divergent transitions, among other pathologies, imposed by complex geometry (*i.e.*, curvature) in the posterior space since many iterative stochastic MCMC sampling algorithms experience difficulties when exploring high density distribution regions. Thus, remedies to resolve them, such as lowering the step size of the HMC/NUTS sampler, should be attempted in future work, along with other approaches such as empirical Bayes estimation to approximate beta prior hyperparameters from observed data through the MLE or other methods of parameter estimation, such as the method of moments. Alternatively, hierarchical modelling could be employed to estimate separate distribution model hyperparameters for each species and/or compute distinct estimates for the directionality/comparison level of the DNA barcode gap metrics (*i.e.,* lower *vs.* upper, non-prime *vs.* prime) separately within the genus under study. This would permit greater flexibility through incorporating more fine-grained structure seen in the data; however, low

taxon sample sample sizes may preclude valid inferences to be reasonably ascertained due to the large number additional parameters which would be introduced through the specification of the hyperprior distributions. Methods outlined in Gelman et al. (2014), such as dealing with non-exchangeability of observations and alternate model parameterizations like the logit, may prove useful in this regard. Even though more work remains, it is clear that both frequentist and Bayesian inference hold much promise for the future of molecular biodiversity science.

Supplementary Information

None declared.

Data Availability Statement

Raw data, R, and Stan code can be found on GitHub at:
https://github.com/jphill01/Bayesian-DNA-Barcode-Gap-Coalescent.

Acknowledgements

We wish to recognise the valuable comments and discussions of Daniel (Dan) Gillis, Robert (Bob) Hanner, Robert (Rob) Young, and XXX anonymous reviewers.

We acknowledge that the University of Guelph resides on the ancestral lands of the Attawandaron people and the treaty lands and territory of the Mississaugas of the Credit. We recognize the significance of the Dish with One Spoon Covenant to this land and offer our respect to our Anishinaabe, Haudenosaunee and Métis neighbours as we strive to strengthen our relationships with them.

Funding

None declared.

Conflict of Interest

None declared.

Author Contributions

JDP wrote the manuscript, wrote R and Stan code, as well as analyzed and interpreted all model results.

References

Agresti, A. and B. A. Coull

1998. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Avise, J., J. Arnold, R. Ball, Jr., E. Bermingham, T. Lamb, J. Neigel, C. Reeb, and N. Saunders

1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.*, 18:489–522.

Bartlett, S. and W. Davidson

1992. FINS (forensically informative nucleotide sequencing): A procedure for identifying the animal origin of biological specimens. *BioTechniques*, 12(3):408—411.

Bergsten, J., D. Bilton, T. Fujisawa, M. Elliott, M. Monaghan, M. Balke, L. Hendrich, J. Geijer, J. Herrmann, G. Foster, I. Ribera, A. Nilsson, T. Barraclough, and A. Vogler

2012. The effect of geographical scale of sampling on DNA barcoding. *Systematic biology*, 61(5):851–869.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell

2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1.

Chen, M.-H. and Q.-M. Shao

1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92.

Collins, R. A. and R. H. Cruickshank

2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 13(6):969–975.

Dempster, A. P., N. M. Laird, and D. B. Rubin

1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Gelman, A., J. Carlin, H. Stern, D. Duncan, A. Vehtari, and D. Rubin

2014. *Bayesian Data Analysis*, third edition. Chapman and Hall/CRC.

Gelman, A. and D. Rubin

1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gelman, A., A. Vehtari, D. Simpson, C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák

2020. Bayesian workflow.

Hebert, P., A. Cywinska, S. Ball, and J. deWaard

2003a. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321.

Hebert, P., S. Ratnasingham, and J. de Waard

2003b. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S96–S99.

Hebert, P. D., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis

2004. Identification of birds through DNA barcodes. *PLoS Biol*, 2(10):e312.

Hoffman, M. and A. Gelman

2014. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Hubert, N. and R. Hanner

2015. DNA barcoding, species delineation and taxonomy: A historical perspective. *DNA Barcodes*, 3:44–58.

Jeffreys, H.

1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Jukes, T. and C. Cantor

1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro, ed., Pp. 21–132. New York: Academic Press.

Kimura, M.

1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(1):111–120.

Kingman, J.

1982a. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248.

Kingman, J.

1982b. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Knowles, L. L. and W. P. Maddison

2002. Statistical phylogeography. *Molecular Ecology*, 11(12):2623–2635.

Liu, Y., A. Gelman, and T. Zheng

2015. Simulation-efficient shortest probability intervals. *Statistical Computing*, 25:809–819.

Mather, N., S. Traves, and S. Ho

2019. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 10(1):579–589.

Meier, R., G. Zhang, and F. Ali

2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology*, 57(5):809–813.

Meyer, C. and G. Paulay

2005. DNA barcoding: Error rates based on comprehensive sampling. *PLOS Biology*, 3(12):e422.

Newcombe, R. G.

1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872.

Pentinsaari, M., H. Salmela, M. Mutanen, and T. Roslin

2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports*, 6:35275.

Phillips, J., D. Gillis, and R. Hanner

2022. Lack of statistical rigor in DNA barcoding likely invalidates the presence of a true species' barcode gap. *Frontiers in Ecology and Evolution*, 10:859099.

Phillips, J., C. Griswold, R. Young, N. Hubert, and H. Hanner

2024. *A Measure of the DNA Barcode Gap for Applied and Basic Research*, Pp. 375–390. New York, NY: Springer US.

R Core Team

2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. and Z. Yang

2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656.

Ratnasingham, S. and P. Hebert

2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, 7(3):355–364.

Stan Development Team

2023. RStan: The R interface to Stan. R package version 2.32.6.

Čandek, K. and M. Kuntner

2015. DNA barcoding gap: Reliable species identification over morphological and geographical scales. *Molecular Ecology Resources*, 15(2):268–277.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner

2021. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

Wickham, H.

2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yang, Z. and B. Rannala

2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107:9264–9269.

Yang, Z. and B. Rannala

2017. Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Molecular Ecology*, 26:3028–3036.

Young, R., R. Gill, D. Gillis, and R. Hanner

2021. Molecular Acquisition, Cleaning and Evaluation in R (MACER) - A tool to assemble molecular marker datasets from BOLD and GenBank. *Biodiversity Data Journal*, 9:e71378.

Figures and Tables

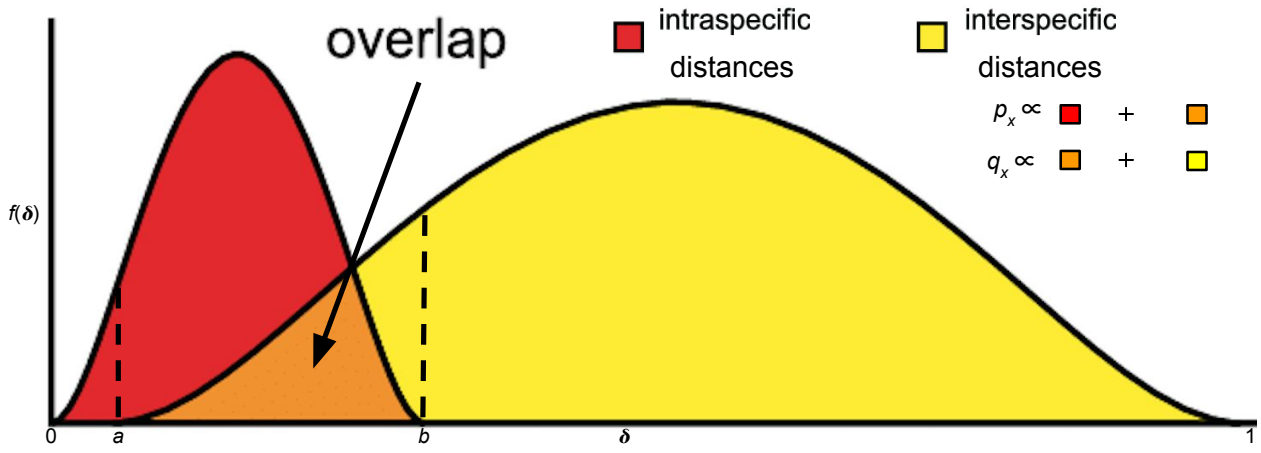

Figure 1: Modified depiction from Meyer and Paulay (2005) and Phillips et al. (2024) of the overlap/separation of intraspecific and interspecific distances (δ) for calculation of the DNA barcode gap metrics (p_x and q_x) for a hypothetical species x. The minimum interspecific distance is denoted by a and the maximum intraspecific distance is indicated by b. The quantity $f(\delta)$ is akin to a kernel density estimate of the probability density function of distances. A similar visualization can be displayed for $p_x^{'}$ and $q_x^{'}$ within the interval $[a', b]$.

Table 1: Interpretation of the DNA barcode gap estimators within $[a/a', b]$

Parameter	Explanation
p_x/p_{lwr}	When p_{lwr} is close to 0 (1), it suggests that the probability of intraspecific (interspecific) distances being larger (smaller) than interspecific (intraspecific) distances is low (high) on average, while the probability of interspecific (intraspecific) distances being larger (smaller) than intraspecific (interspecific) distances is high (low) on average; that is, there is (no) evidence for a DNA barcode gap.
q_x/p_{upr}	When p_{upr} is close to 0 (1), it suggests that the probability of interspecific (intraspecific) distances being larger (smaller) than intraspecific (interspecific) distances is high (low) on average, while the probability of intraspecific (interspecific) distances being larger (smaller) than interspecific (intraspecfic) distances is low (high) on average; that is, there is (no) evidence for a DNA barcode gap.
$p_x^{'}/p_{\text{lwr}}^{'}$	When $p_{\text{lwr}}^{'}$ is close to 0 (1), it suggests that the probability of intraspecific (combined interspecific distances for a target species and its nearest neighbour species) distances being larger than combined interspecific distances for a target species and its nearest neighbour species (intraspecific distances) is low (high) on average, while the probability of combined interspecific distances for a target species and its nearest neighbour species (intraspecfic distances) being larger than intraspecific distances (combined interspecific distances for a target species and its nearest neighbour species) is high (low) on average; that is, there is (no) evidence for a DNA barcode gap.
$q_x^{'}/p_{\text{upr}}^{'}$	When $p_{\text{upr}}^{'}$ is close to 0 (1), it suggests that the probability of combined interspecific distances for a target species and its nearest neighbour species (intraspecific distances) being larger than intraspecific distances (combined interspecific distances for a target species and its nearest neighbour species) is high (low) on average, while the probability of intraspecific distances (combined interspecific distances for a target species and its nearest neighbour species) being larger than combined interspecific distances for a target species and its nearest neighbour species (intraspecific distances) is low (high) on average; that is, there is (no) evidence for a DNA barcode gap.

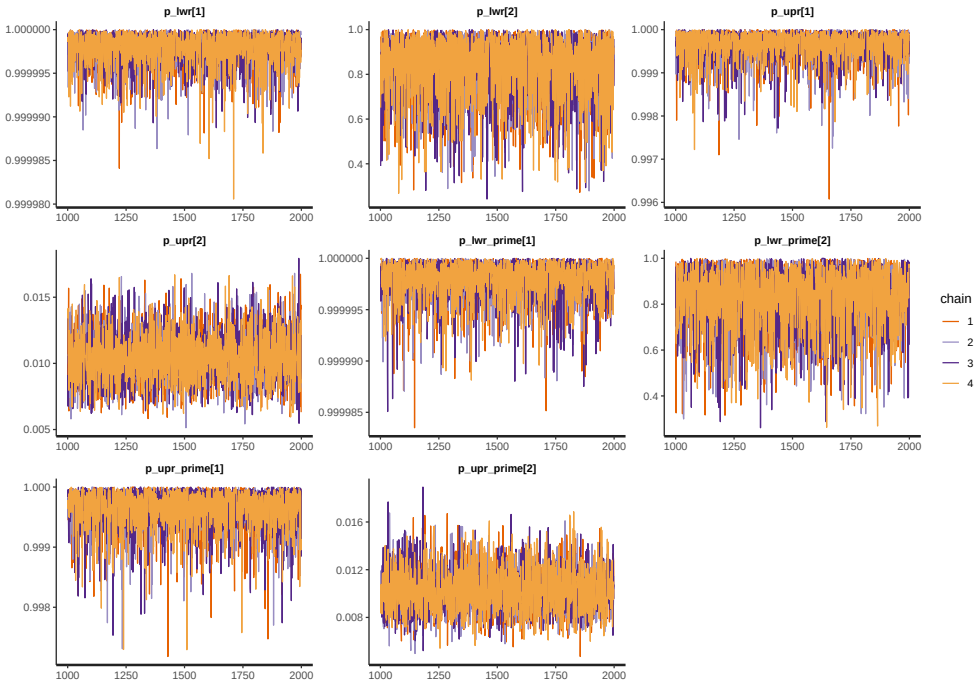

Figure 2: MCMC parameter traceplots applied to *A. bipustulatus* ([1]; $N = 701$) and *A. nevadensis* ([2]; $N = 2$) for CYTB across 1000 post-warmup iterations.

Table 2: Nonparametric frequentist and Bayesian estimates of distance distribution overlap/separation for the DNA barcode gap coalescent model parameters applied to *A. bipustulatus* ($N = 701$) and *A. nevadensis* ($N = 2$) for CYTB, including 95% CIs and CrIs. CrIs are based on 4000 posterior draws. All parameter estimates are reported to three decimal places of precision.

Species	**Parameter**	**MLE (SE, 95% CI)**	**Bayes Est. (SD; 95% CrI)**
A. bipustulatus	p_x/p_{lwr}	1.000 (0.000; 1.000-1.000)	1.000 (0.000; 1.000-1.000)
A. bipustulatus	q_x/p_{upr}	1.000 (0.000; 1.000-1.000)	1.000 (0.000; 0.999-1.000)
A. bipustulatus	$p_x^{'}/p_{\text{lwr}}^{'}$	1.000 (0.000; 1.000-1.000)	1.000 (0.000; 1.000-1.000)
A. bipustulatus	$q_x^{'}/p_{\text{upr}}^{'}$	1.000 (0.000; 1.000-1.000)	1.000 (0.000; 0.999-1.000)
A. nevadensis	p_x/p_{lwr}	1.000 (0.000; 1.000-1.000)	0.835 (0.144; 0.470-0.996)
A. nevadensis	q_x/p_{upr}	0.010 (0.002; 0.006-0.014)	0.010 (0.002; 0.007-0.014)
A. nevadensis	$p_x^{'}/p_{\text{lwr}}^{'}$	1.000 (0.000; 1.000-1.000)	0.834 (0.138; 0.481-0.994)
A. nevadensis	$q_x^{'}/q_{\text{upr}}^{'}$	0.010 (0.070; -0.128-0.148)	0.010 (0.002; 0.007-0.014)

Table 3: Posterior predictive checks of distance distribution overlap/separation for the DNA barcode gap coalescent model parameters applied to *A. bipustulatus* ($N = 701$) and *A. nevadensis* ($N = 2$) for CYTB. CrIs are based on 4000 posterior draws. All parameter estimates are reported to three decimal places of precision.

Species	**Variable**	$\boldsymbol{Y}$	$\boldsymbol{n}$	**Bayes Est. (SD; 95% CrI)**
A. bipustulatus	y_{lwr}	491401.000	491401.000	491400.018 (1.378, 491396.000-491401.000)
A. bipustulatus	y_{upr}	2804.000	2804.000	2803.019 (1.433, 2799.000-2804.000)
A. bipustulatus	y'_{lwr}	491401.000	491401.000	491400.008 (1.412, 491396.000-491401.000)
A. bipustulatus	y'_{upr}	2804.000	2804.000	2802.992 (1.429, 2799.000-2804.000)
A. nevadensis	y_{lwr}	4.000	4.000	3.355 (0.888, 1.000-4.000)
A. nevadensis	y_{upr}	28.000	2804.000	29.151 (7.620, 16.000-45.000)
A. nevadensis	y'_{lwr}	4.000	4.000	3.325 (0.884, 1.000-4.000)
A. nevadensis	y'_{upr}	28.000	2804.000	28.942 (7.409, 15.000-45.000)

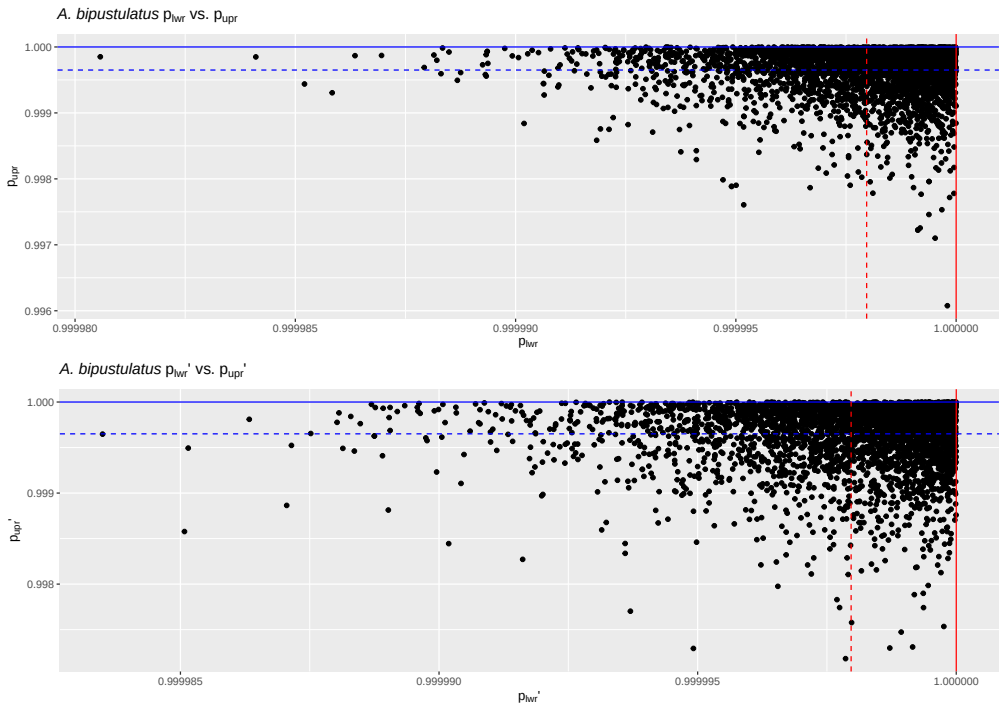

Figure 3: Scatterplot of 4000 Bayesian posterior draws (black solid points) for *A. bipustulatus* ($N = 701$) across CYTB. MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.

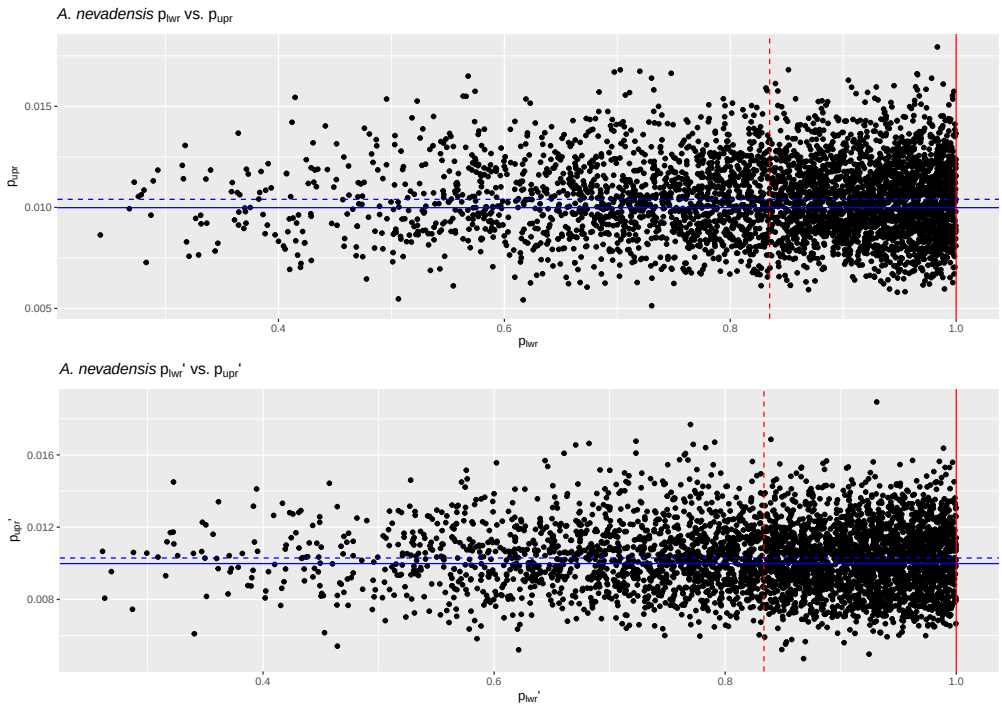

Figure 4: Scatterplot of 4000 Bayesian posterior draws (black solid points) for *A. nevadensis* ($N = 2$) across CYTB. MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.

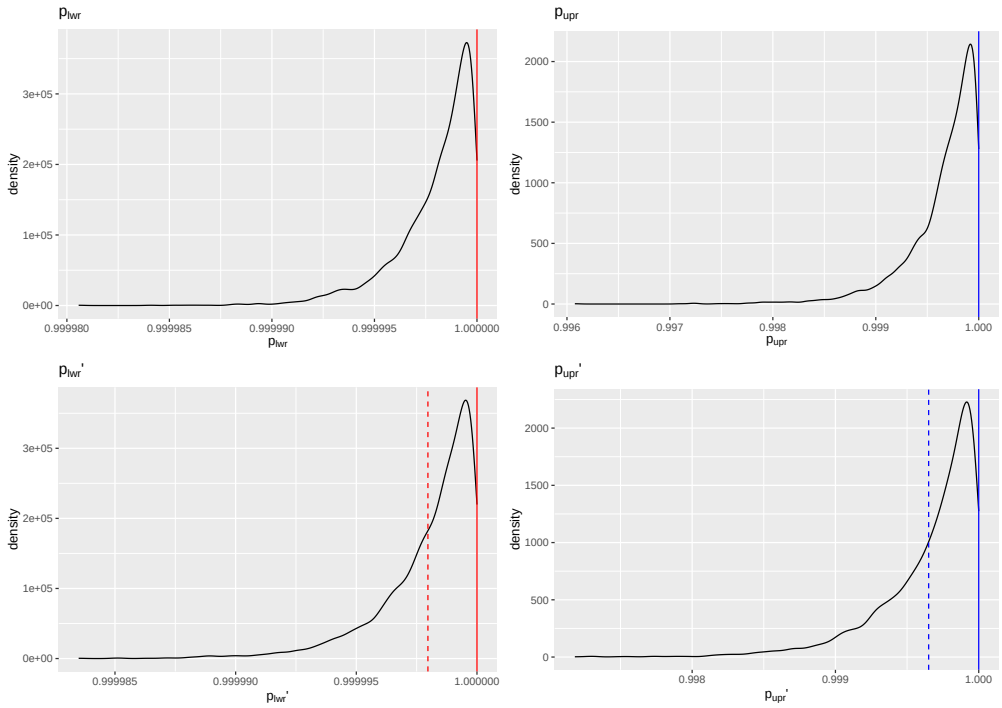

Figure 5: Posterior distributions based on 4000 draws of the DNA barcode gap metrics depicted as density plots for *A. bipustulatus* ($N = 701$). MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.

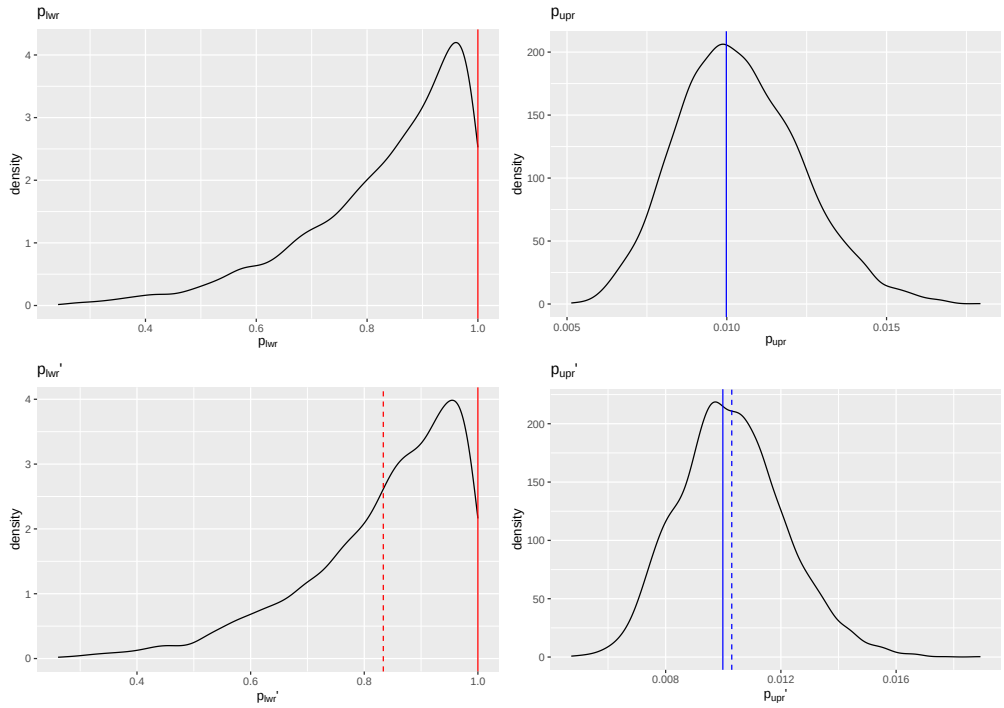

Figure 6: Posterior distributions based on 4000 draws of the DNA barcode gap metrics depicted as density plots for *A. nevadensis* ($N = 2$). MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.