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 novel 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 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. The proposed maximum likelihood estimators and Bayesian model are demonstrated on CYTB gene sequence data from two *Agabus* diving beetle species exhibiting limits in the extent of representative taxonomic sampling. Obtained results have a strong potential to shed light on interesting 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 use of DNA sequences to support broad microevolutionary and macroevolutionary hypotheses at wide taxonomic levels such as birds, fishes, and insects, is not a new idea (*e.g.* Forensically Important Nucleotide Sequences (FINS); Bartlett and Davidson (1992)). Since its inception over 20 years ago, DNA barcoding (Hebert et al., 2003a,b) 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 haploid 5'-COI for animals. However, the success of the 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 idea that the maximum pairwise 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), thereby compromising rapid matching of unknown samples to expertly-validated references, and 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, and 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 (Rannala and Yang, 2003; Yang and Rannala, 2017) were recently outlined and shown to hold strong promise when applied to predatory *Agabus* (Coleoptera: Dytiscidae) diving beetles (Phillips et al., 2024), a group, which, despite their ease of sampling and well-established taxonomy, possess few morphologically-distinct taxonomic characters that readily facilitate their assignment to the species level (Bergsten et al., 2012). Further, the 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 CYTB) obtained from BOLD and GenBank, results highlight that DNA barcoding has been a one-sided argument. Results point to the need to balance both the sufficient collection of specimens, as well as the extensive sampling of species (Phillips et al., 2024). DNA barcode libraries are biased toward the latter. The coalescent (Kingman, 1982) 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 (Hebert et al., 2003a), which asserts that DNA sequences differing by at least 2% at sequenced genomic regions should be expected to originate from different biological species, and the 10× rule, which suggests that sequences display 10 times more genetic variation among species than within taxa, is evidence for a distinct evolutionary origin (Hebert et al., 2004). In addition, the reliance on visualization approaches, such as frequency histograms, dotplots, and quadrant plots to expose DNA barcoding's limitations, have 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 pairwise genetic distances exists. The proposed statistics quantify the extent of asymmetric directionality of proportional pairwise genetic distance distribution overlap/separation for species within well-sampled taxonomic genera based on a straightforward distance count. The metrics can be employed in a variety of ways, including to assess performance of marker genes for species identification (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 (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, along an unbiased way to compute the statistical accuracy of the recommended 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 pairwise genetic distances for species of interest.

2 Methods

2.1 DNA Barcode Gap Metrics

Recently, Phillips et al. (2024) proposed novel nonparametric maximum likelihood estimators (MLEs) of proportional overlap/separation between intraspecific and interspecific pairwise genetic distance distributions for a given species (x) to aid assessment of the DNA barcode gap 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 pairwise genetic distances within species, d_{XY} are paireise genetic 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, and the maximum intraspecific distance, respectively (**Figure 1**). Hence, Equations (1)-(4) are simply empirical partial means of pairwise genetic 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) for details). Pairwise genetic distances form a continuouss 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). 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 interspecfic genetic 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 pairwise interspecific distance is used. While these schemes more accurately account for species' coalescence processes inferred from contemporaneous samples of DNA sequences, such as interspecific hybridization/introgression events (Phillips et al., 2024). Within equations (3) and (4), the degree of pairwise genetic 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), and thus can be used as a criterion to assess the failure of DNA barcoding in recently radiated taxonomic groups, among other plausible biological explanations. Note, pairwise genetic distances are constrained to the unit interval [0, 1], whereas the metrics are defined only on $[a/a', b]$. Values of the estimators obtained from equations (1)-(4) close to or equal to zero give evidence for separation between intraspecific and interspecific pairwise genetic 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 gap.

3 Mathematical Details

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 empirical cumulative distribution functions (ECDFs) as follows, since the true underlying CDFs, $F(\cdot)$, are unknown *a priori*, and therefore must be estimated using available data:

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

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

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

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

noting that there are no pairwise genetic distances less than the minimum and greater than the maximum, and that all pairwise genetic distances are greater than or equal to the minimum and less than or equal to the maximum. 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) herein clearly demonstrate the asymmetric directionality of the proposed metrics. Furthermore, calculation of the DNA barcode gap estimators is straightforward as they implicitly account for total distribution area (including overlap). Nevertheless, the total areas bounded by intraspecific and interspecific distributions on $[a, b]$, and combined distributions on $[a', b]$ (see **Figure 1**) are given by

$$\begin{aligned} A &= p_x + q_x \\ &= \mathbb{P}(d_{ij} \geq a) + \mathbb{P}(d_{XY} \leq b) \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \end{aligned} \tag{12}$$

$$\begin{aligned} A' &= p'_x + q'_x \\ &= \mathbb{P}(d_{ij} \geq a') + \mathbb{P}(d'_{XY} \leq b) \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \end{aligned} \tag{13}$$

whose values lie in [0, 2].

The above derivations lay a good foundation for the remainder of this paper.

3.1 A Bayesian Implementation

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. This problem is especially pronounced in the case of binomial proportions. The estimated Wald standard error (SE) of the sample proportion, is given by

$$\widehat{SE[\hat{p}]} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{14}$$

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 is problematic for several reasons. First, Equation (14) 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 barcode gap when one sctually 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 proprtions are exactly equal to zero or one, resulting SEs will be exactly zero, rendering Equation (14) 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 is lacking and sample sizes are small, through integration with the likelihood function associated with true generating processes. The posterior distribution is given by Bayes' rule up to a proportionality

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) \tag{15}$$

where θ are unobserved parameters and Y are known data. 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, since, under the Bayesian paradigm, entire posterior distributions, along with their summaries, 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.

3.2 The Model

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 pairwise genetic 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; however, this requires careful selection of the bandwidth parameter, among other considerations. Here, for simplicity, a different route is taken. Counts, y, of overlapping pairwise genetic 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 pairwise genetic distances, respectively, for a target species along with its nearest neighbour species, and $k = M$ is a total count vector for all pairwise 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. Such a scheme is quite convenient since the beta distribution is conjugate to the binomial distribution. Thus, the posterior distribution is also beta distributed. 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. As a result, 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. 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. With sufficient data, the choice of prior distribution becomes less important since the posterior will be dominated by the likelihood. The full univariate Bayesian model for species x is thus given by

$$\begin{aligned} y_{\text{lwr}} &\sim \text{Binomial}(N, p_{\text{lwr}}) \\ y_{\text{upr}} &\sim \text{Binomial}(M, p_{\text{upr}}) \\ y^{'}_{\text{lwr}} &\sim \text{Binomial}(N, p^{'}_{\text{lwr}}) \\ y^{'}_{\text{upr}} &\sim \text{Binomial}(C, p^{'}_{\text{upr}}) \\ p_{\text{lwr}}, p_{\text{upr}}, p^{'}_{\text{lwr}}, p^{'}_{\text{upr}} &\sim \text{Beta}(1, 1). \end{aligned} \tag{16}$$

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 (16) for distinction between MLEs and Bayesian posterior estimates.

The model, which is inherently vectorized to allow processing of multiple species datasets simultaneously, was fitted 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 (Stan Development Team, 2023). 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. A random seed was set to ensure reproducibility of model results.

Since the DNA barcode gap metrics often attain values very close to zero (suggesting no overlap and complete separation of pairwise genetic 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, which 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, was also attempted. However, this resulted in several divergent transitions, among other pathologies, imposed by complex geometry (*i.e.*, curvature) in the posterior space, despite remedies to resolve them, such as lowering the step size of the HMC/NUTS sampler. Note that this prior is Jeffreys' prior, which is proportional to the square root of the Fisher information and has several desirable statistical properties, most notably invariance to model reparameterization.

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 (7), 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 herein has a straightforward interpretation (**Table 1**).

4 Case Study

Herein, the *Agabus* CYTB dataset analyzed by Phillips et al. (2024) is revisited. 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 taxon sampling, whicj arise 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); yet, due to the level of convenience sampling inherent in taxonomic collection efforts for this genus, adequate representation 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 cutoffs of one and thousands of samples, respectiely, indicating chained 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 via Equation (5) given by

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{17}$$

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. 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 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 or shortest probability intervals (SPIn) (Liu et al., 2015) is warranted. As such 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 show evidence of complete overlap in pairwise intraspecific, interspecific, and combined genetic distances for *A. bipustulatus* in both the p/q and p'/q' directions (**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 more 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% CI for $q_x^{'}$ extends to negative values at the left endpoint, due to the correponding SE of 0.070 being too high as a result of the extremely low sample size of $n = 2$ individuals sampled (**Table 2**). Simulated counts of overlapping specimen records from the posterior predictive distribution (**Table 2**) were found to be very close to observed counts for both species, indicating that the proposed model adequately captures underlying variation.

5 Conclusion

Herein,

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 analysed and interpreted all model results.

References

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.

Collins, R. A. and R. H. Cruickshank
2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 13(6):969–975.

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. 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.

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

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.

Pentinsaari, M., H. Salmela, M. Mutanen, and et al.
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.

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 2.21.8.

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.

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.

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

Figures and Tables

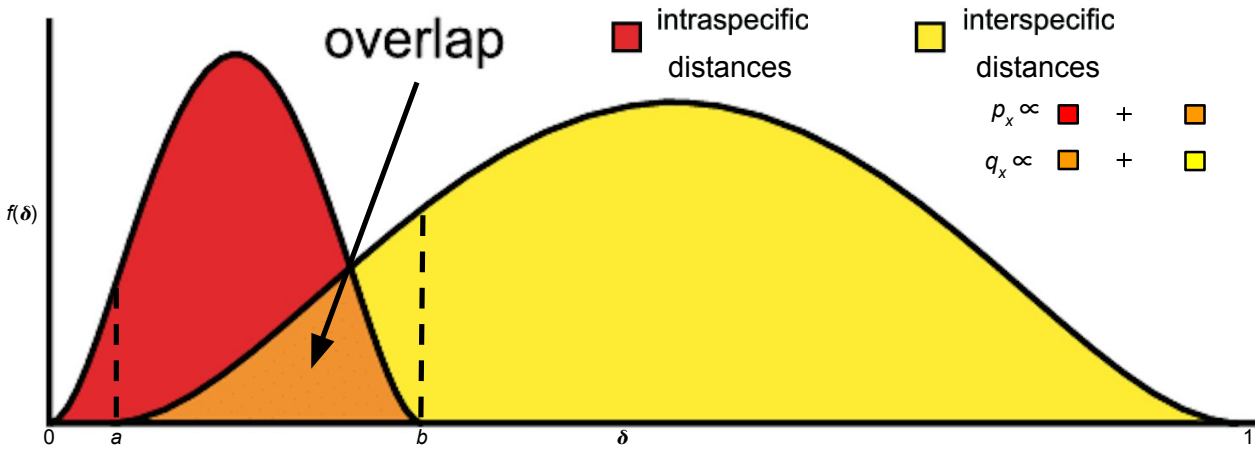

Figure 1: Modified depiction from Meyer and Paulay (2005) and Phillips et al. (2024) of the overlap/separation of pairwise intraspecific and interspecific pairwise genetic distances (δ) for calculation of the DNA barcode gap metrics (p_x and q_x) for a hypothetical species x. The minimum interspecific 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 pairwise genetic 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 (smaaler) 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.

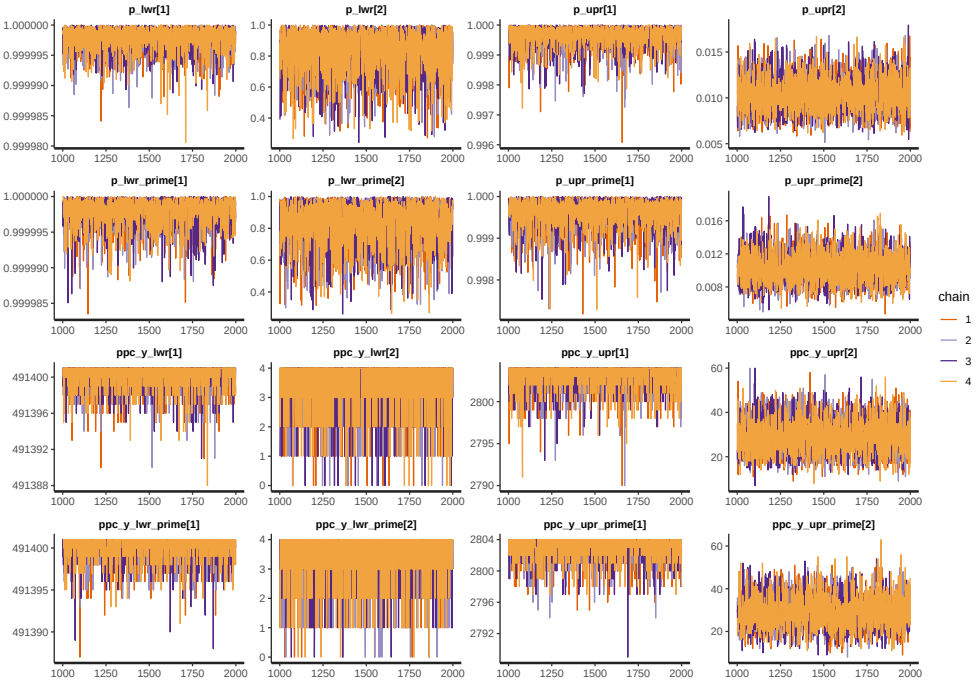

Figure 2: MCMC parameter traceplots applied to *A. bipustulatus* ([1]; $n = 701$) and *A. nevadensis* ([2]; $n = 2$) for CYTB across 2000 iterations.

Table 2: Nonparametric frequentist and Bayesian estimates of pairwise genetic 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/Variable**	**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. bipustulatus	y_{lwr}	491401.000	491400.018 (1.378, 491396.000-491401.000)
A. bipustulatus	y_{upr}	2804.000	2803.019 (1.433, 2799.000-2804.000)
A. bipustulatus	$y_{\text{lwr}}^{'}$	491401.000	491400.008 (1.412, 491396.000-491401.000)
A. bipustulatus	$y_{\text{upr}}^{'}$	2804.000	2802.992 (1.429, 2799.000-2804.000)
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)
A. nevadensis	y_{lwr}	4.000	3.355 (0.888, 1.000-4.000)
A. nevadensis	y_{upr}	28.000	29.151 (7.620, 16.000-45.000)
A. nevadensis	$y_{\text{lwr}}^{'}$	4.000	3.325 (1.000-4.000)
A. nevadensis	$y_{\text{upr}}^{'}$	28.000	28.942 (15.000-45.000)

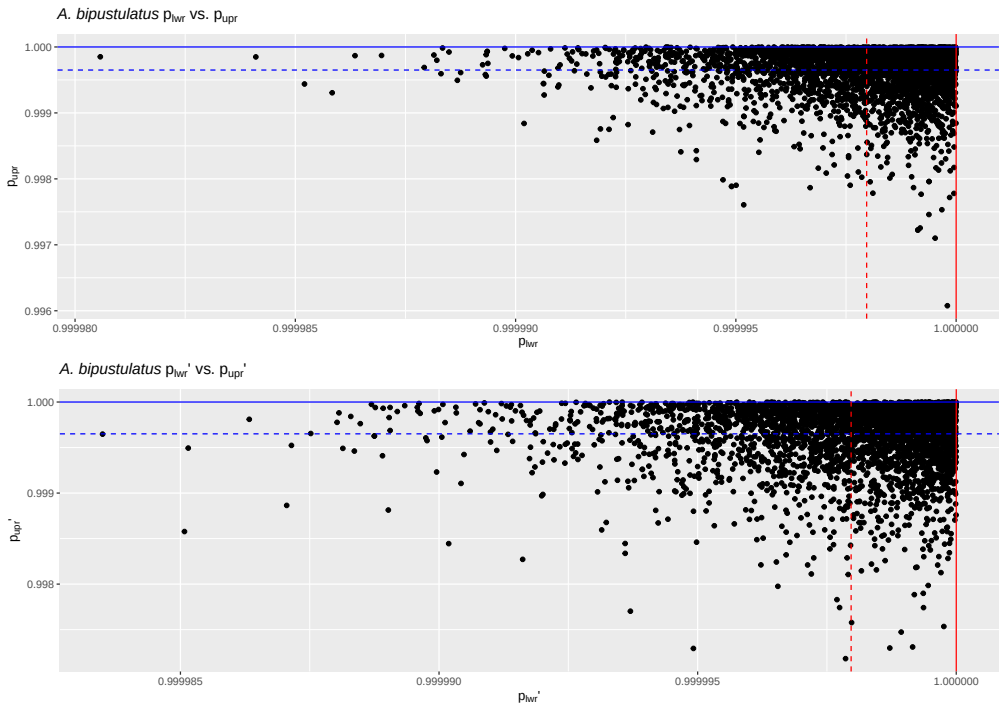

Figure 3: Bayesian posterior draws ($n = 4000$; black solid points) for *A. bipustulatus* ($n = 701$) across CYTB. MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.

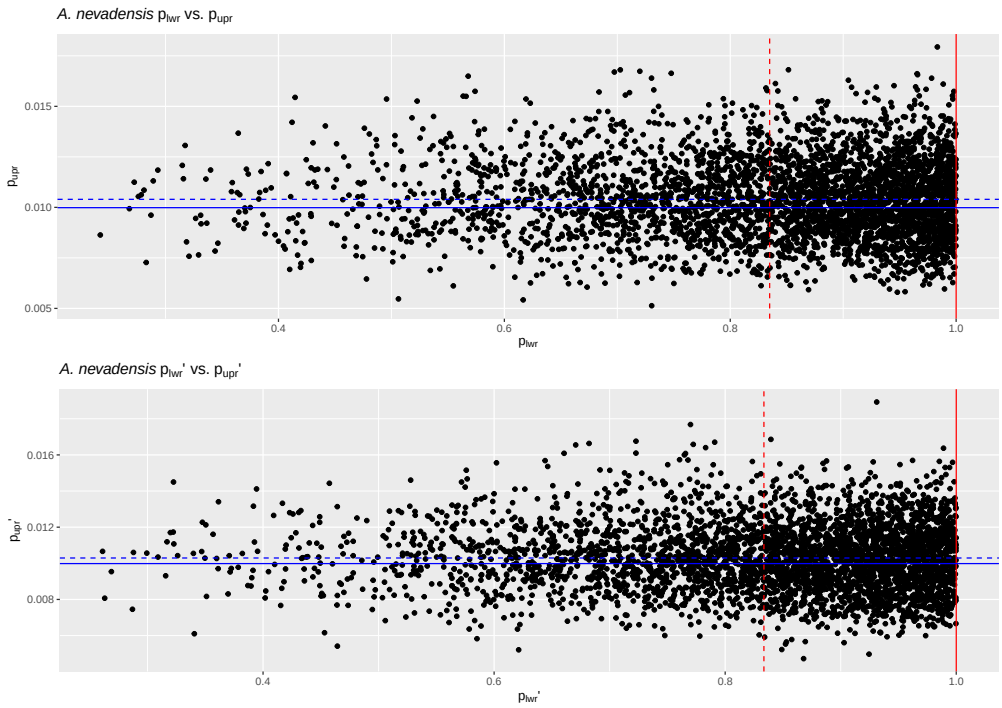

Figure 4: Bayesian posterior draws ($n = 4000$; black solid points) for *A. nevadensis* ($n = 2$) across CYTB. MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.