

1 **A Bayesian Model of the DNA Barcode Gap**

2 Jarrett D. Phillips^{1,2*} (ORCID: 0000-0001-8390-386X), ... (others?)

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

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

5 ***Corresponding Author:** Jarrett D. Phillips¹

6 **Email Address:** jphill01@uoguelph.ca

7 **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 (particularly mitochondrial DNA (mtDNA)) to support broad evolutionary hypotheses and questions concerning demographic processes, like gene flow and speciation, that have produced a distinctive and measurable pattern of genetic polymorphism in diverse and spatially-distributed taxonomic lineages such as birds, fishes,

insects, and arachnids, among other extensively studied groups, took flight in the late 1980s
 (Avise et al., 1987). 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 significantly on earlier work and has emerged as a robust method of
 specimen identification and species discovery across myriad multicellular Eukaryotes which
 have been sequenced at easily obtained 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 taxonomic groups 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). Furthermore, it has been well-demonstrated
that the presence of a DNA barcode gap becomes less certain with increasing spatial scale
of sampling since interspecific distances increase, while intraspecific distances shrink as
more closely-related species are sampled (Phillips et al., 2022). This can pose problems
in cases of rare species or monotypic taxa, for instance (Ahrens et al., 2016) and 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, interspecies hybridization, genome 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) were recently outlined by Phillips et al. (2024). Unlike previously proposed MSC algorithmic approaches (of which there are too many to exhaustively list here), which generally assume a strict molecular clock and a simplified model of DNA sequence evolution across closely-related taxa from which an estimated species phylogeny may be constructed (*e.g.*, with or without use of a guide tree) (*e.g.*, Rannala and Yang (2003, 2017); Yang and Rannala (2010, 2014, 2017)), Phillips et al.'s (2024) approach is tree-free and does not require judicious parameter setting. 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

88 extensive sampling of species: DNA barcode libraries are biased toward the latter (Phillips
 89 et al., 2022). The coalescent (Kingman, 1982a,b) encompasses a backwards continuous-time
 90 stochastic Markov process of allelic sampling within natural, neutrally-evolving, species
 91 populations towards the most recent common ancestor (MRCA). The estimators from Phillips
 92 et al. (2024) represent a clear improvement over simple, yet arbitrary, distance heuristics such
 93 as the 2% rule noted by Hebert et al. (2003a) and the 10 \times rule (Hebert et al., 2004) that
 94 form the basis of single-locus species delimitation tools like Automatic Barcode Gap Discovery
 95 (ABGD) (Puillandre et al., 2011), Assemble Species by Automatic Partitioning (ASAP)
 96 (Puillandre et al., 2021), and the Barcode Index Number (BIN) framework (Ratnasingham
 97 and Hebert, 2013). The 2% rule asserts that DNA sequences differing by at least 2% at
 98 sequenced genomic regions should be expected to originate from different biological species,
 99 whereas the 10 \times rule suggests that sequences displaying 10 times more genetic variation
 100 among species than within taxa is evidence for a distinct evolutionary origin. However,
 101 the lack of adoption of an explicit, universally agreed upon, species concept that readily
 102 governs lineage formation and proliferation necessary to establish rigorous taxon definitions
 103 for successful delimitation of hypothesized and heuristic evolutionary units using these
 104 well-known criteria, in conjunction with secondary lines of evidence (*e.g.*, morphology,
 105 ecology, geography, and behaviour) promised by an integrative framework, is
 106 missing (Rannala, 2015; Pante et al., 2015; Wells et al., 2022). In addition, the reliance
 107 on visualization approaches, such as frequency histograms, dotplots, and quadrant plots
 108 to expose DNA barcoding’s limitations, has also been criticized (Collins and Cruickshank,
 109 2013; Phillips et al., 2022). Up until the work of Phillips et al. (2024), the majority of
 110 studies (*e.g.*, Young et al. (2021)) have treated the DNA barcode gap as a binary response.
 111 However, given poor sampling depth for most taxa, a Yes/No dichotomy is inherently flawed
 112 because it can falsely imply a DNA barcode gap is present for a taxon of interest when in
 113 fact no such separation in distances exists. The proposed statistics quantify the extent
 114 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 the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) and other related statistics of 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; Mather et al., 2019). Early on, DNA barcoding was presumed to only work for reciprocally monophyletic groups and thus concerned itself with terminal branches of generated phylogenies rather than more basal lineages occurring deeper in hypothesized species trees (Mutanen et al., 2016). Furthermore, the occurrence of short branches within resolved phylogenies increases the probability of deep coalescence, clouding species delimitations, which often fail or are uncertain in broad parameter space (Carstens et al., 2013; Hickerson et al., 2006; Rannala, 2015). As DNA barcoding is a single-locus approach, it is problematic for evolutionarily young taxa, wherein incomplete lineage sorting within gene genealogies is a common phenomenon due to the ongoing stochastic dynamic of mutation generating population variation, and genetic drift driving variants to fixation (Rannala, 2015). The most promising way forward in this regard seems to be through the use of software such as BPP (Bayesian Phylogenetics and Phylogeography), which permits efficient full Bayesian simulations under various MSC models (*e.g.*, MSC-I (MSC with introgression) or MSC-M (MSC with migration), among others) using MCMC for tree parameter estimation (using the A00 option, for instance) (Flouri et al., 2018), or PHRAPL (Phylogeographic Inference using Approximate Likelihoods) (Jackson et al., 2017), which employs tractable phylogenetic likelihood calculations.

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:

$$p_x = \frac{\#\{d_{ij} \geq a\}}{\#\{d_{ij}\}} \quad (1)$$

$$q_x = \frac{\#\{d_{XY} \leq b\}}{\#\{d_{XY}\}} \quad (2)$$

$$p'_x = \frac{\#\{d_{ij} \geq a'\}}{\#\{d_{ij}\}} \quad (3)$$

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

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).

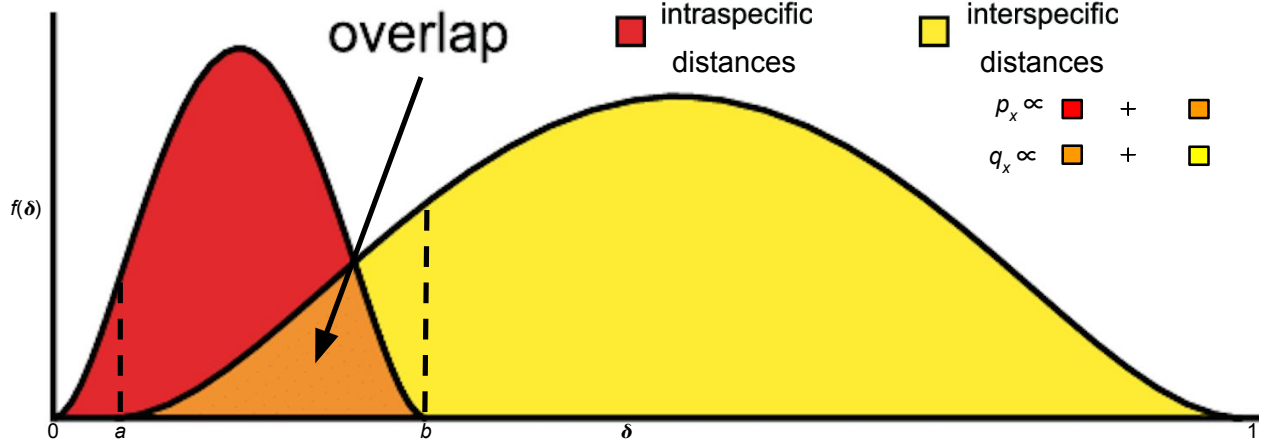


Figure 1: Modified depiction from Meyer and Paulay (2005) and Phillips et al. (2024) of the overlap/separation of intraspecific and interspecific distances (δ) for calculation of the DNA barcode gap metrics (p_x and q_x) for a hypothetical species x . The minimum interspecific distance is denoted by 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 $[a', b]$.

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 n th order statistics, $X_{(1)}$ and $X_{(n)}$, respectively, with $a/a' < b$, which have been pointed out by Phillips et al. (2022) as important for developing a mathematical theory to test the existence of the DNA barcode gap. 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) using, for example, the `dist.dna()` function available in the `ape` R package (Paradis et al., 2004); however, values are *not* independent and identically distributed (IID) because estimated standard errors (SEs) will depend on both the number of species sampled with the genus under study, as well as the number of specimens sampled within a target species. In Phillips et al. (2024), To tease this out, Phillips et al. (2024) suggests plotting estimator values against their estimated SEs, along with a simple random downsampling scheme. In the case of two species comprising a focal genus, one well sampled and the other poorly sampled, values

of the metrics close to zero obtained from Equations (1)-(4) for the sufficiently sampled species will likely possess larger SEs following downsizing to match the number of poorly sampled specimens (Phillips et al., 2024). 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 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

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

199 Rearranging Equation (5) gives

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

200 from which it follows that

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

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

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

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

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

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

204 From this, it can be seen that $\hat{F}_{d_{ij}}(b) = 1$ in Equations (8) and (10). Given n
 205 increasing-ordered data points, the (discrete) ECDF, $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{[x_i \leq t]}$, comprises a step
 206 function having jump discontinuities of size $\frac{1}{n}$ at each sample observation (x_i), excluding ties
 207 (or steps of weight $\frac{i}{n}$ with duplicate observations), where $\mathbf{1}(x)$ is the indicator function. Note,
 208 $\mathbb{P}(X = t) \neq 0$. Equations (8)-(11) clearly demonstrate the asymmetric directionality of the
 209 proposed metrics. Furthermore, calculation of the DNA barcode gap estimators is convenient
 210 as they implicitly account for total distribution area (including overlap).

211 A major criticism of large sample (frequentist) theory is that it relies on asymptotic
 212 properties of the MLE (whose population parameter is assumed to be a fixed but unknown
 213 quantity), such as estimator normality and consistency as the sample size approaches infinity.
 214 This problem is especially pronounced in the case of binomial proportions (Newcombe, 1998).
 215 The estimated Wald SE of the sample proportion, is given by

$$216 \widehat{SE}[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ where } \hat{p} = \frac{Y}{n} \text{ is the MLE, } Y \text{ is the total number of successes } (Y = \sum_{i=1}^n y_i)$$

217 and n is the total number of trials (*i.e.*, sample size). However, the above formula for the
 218 standard error is problematic for several reasons. First, it is a Normal approximation which
 219 makes use of the central limit theorem (CLT); thus, large sample sizes are required for reliable
 220 estimation. When few observations are available, SEs will be large and inaccurate, leading
 221 to low statistical power to detect a true DNA barcode gap when one actually exists. Further,
 222 resulting interval estimates could span values less than zero or greater than one, or have zero
 223 width, which is practically meaningless. Second, when proportions are exactly equal to zero
 224 or one, resulting SEs will be exactly zero, rendering $\widehat{SE}[\hat{p}]$ given above completely useless.
 225 In the context of the proposed DNA barcode gap metrics, values obtained at the boundaries
 226 of their support are often encountered. Therefore, reliable calculation of SEs is not feasible.
 227 Given the importance of sufficient sampling of species genetic diversity for DNA barcoding
 228 initiatives, a different statistical estimation approach is necessary.

229 Bayesian inference offers a natural path forward in this regard since it allows for
 230 straightforward specification of prior beliefs concerning unknown model parameters and
 231 permits the seamless propagation of uncertainty, when data are lacking and sample sizes
 232 are small, through integration with the likelihood function associated with true generating
 233 processes. The posterior distribution ($\pi(\theta|Y)$) is given by Bayes' theorem up to a
 234 proportionality $\pi(\theta|Y) \propto \pi(Y|\theta)\pi(\theta)$, where θ are unobserved parameters, Y are known
 235 data, $\pi(Y|\theta)$ is the likelihood, and $\pi(\theta)$ is the prior. As a consequence, because parameters
 236 are treated as random variables, Bayesian models are much more flexible and generally more
 237 easily interpretable compared to frequentist approaches. Under the Bayesian paradigm, entire
 238 posterior distributions, along with their summaries (*e.g.*, CrIs) are outputted, rather than just
 239 long run behaviour reflected in sampling distributions, p-values, and CIs as in the frequentist
 240 case, thus allowing direct probability statements to be made.

241 Essentially, from a statistical perspective, the goal herein is to nonparametrically estimate
 242 probabilities corresponding to extreme tail quantiles for positive highly skewed distributions
 243 on the unit interval (or any closed subinterval thereof). Here, it is sought to numerically

244 approximate the extent of proportional overlap/separation of intraspecific and interspecific
 245 distance distributions within the subinterval $[a/a', b]$. This is a challenging computational
 246 problem within the current study as detailed in subsequent sections. The usual approach
 247 employs kernel density estimation (KDE), along with numerical or Monte Carlo integration
 248 of complicated probability distribution functions (PDFs), and invocation of extreme value
 249 theory (EVT); however, this requires careful selection of the bandwidth parameter, among
 250 other considerations. This becomes problematic when fitting finite mixture models where
 251 nonidentifiability is rampant. For DNA barcode gap estimation, this would correspond
 252 to a two-component mixture (one for intraspecific distance comparisons, and the other for
 253 interspecific comparisons), with one or more curve intersection points between components,
 254 and the presence of zero distance inflation. This makes parameter estimation difficult using
 255 methods like the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) as
 256 the algorithm may become stuck in suboptimal regions of the parameter search space and
 257 prematurely converge to local optima. Here, for simplicity, a different route is taken to
 258 avoid these obstacles. Counts, y , of overlapping distances (as expressed in the numerator of
 259 Equations (1)-(4)) are treated as binomially distributed with expectation $\mathbb{E}[Y] = k\theta$, where
 260 $k = \{N, C\}$ are total count vectors of intraspecific and combined interspecific distances,
 261 respectively, for a target species along with its nearest neighbour species, and $k = M$ is a
 262 total count vector for all interspecific species comparisons. This follows from the fact that
 263 the ECDF is binomially distributed. The quantity thus being estimated is the parameter
 264 vector $\underline{\theta} = \{p_x, q_x, p'_x, q'_x\}$.

265 The metrics encompassing $\underline{\theta}$ are presumed to follow a Beta(α, β) distribution, with real
 266 shape parameters α and β , which is a natural choice of prior on probabilities. The beta
 267 distribution has a prior mean of $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$ and a prior variance equal to $\mathbb{V}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
 268 In the case where $\alpha = \beta$, all generated Beta(α, β) distributions will possess the same prior
 269 expectation, whereas the prior variance will shrink as both α and β increase. Such a scheme is
 270 quite convenient since the beta distribution is conjugate to the binomial distribution. Thus,

the posterior distribution is also beta distributed, specifically, $\text{Beta}(\alpha + Y, \beta + n - Y)$, 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 $\text{Beta}(1, 1)$ prior, which is equivalent to a standard uniform ($\text{Uniform}(0, 1)$) 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 $\text{Beta}(Y + 1, n - Y + 1)$, 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, $w \rightarrow 0$ as $n \rightarrow \infty$, 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 $\text{Beta}(1, 1)$, $w = \frac{2}{2 + n}$, 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

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

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

300 The proposed model is inherently vectorized to allow processing of multiple species
 301 datasets simultaneously. Model fitting was achieved using the Stan probabilistic
 302 programming language (Carpenter et al., 2017) framework for Hamiltonian Monte Carlo
 303 (HMC) via the No-U-Turn Sampler (NUTS) sampling algorithm (Hoffman and Gelman,
 304 2014) through the **rstan** R package (version 2.32.6) (Stan Development Team, 2023) in R
 305 (version 4.4.1) (R Core Team, 2024). Four Markov chains were run for 2000 iterations each in
 306 parallel across four cores with random parameter initializations. Within each chain, a total
 307 of 1000 samples was discarded as warmup (*i.e.*, burnin) to reduce dependence on starting
 308 conditions and to ensure posterior samples are reflective of the equilibrium distribution.
 309 Further, 1000 post-warmup draws were utilized per chain during the sampling phase. Because
 310 HMC/NUTS results in dependent samples that are minimally autocorrelated, chain thinning
 311 is not required. Each of these tuning parameters reflect default Markov Chain Monte Carlo
 312 (MCMC) settings in Stan to control both bias and variance respectively in the resulting
 313 draws. All analyses in the present work were carried out on a 2023 Apple MacBook Pro
 314 with M2 chip and 16 GB RAM running macOS Ventura 13.2. A random seed was set to
 315 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. To successfully run the Stan program, end users must have installed an appropriate compiler (such as GCC or Clang) which is compatible with their operating system, such as macOS.

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 potential scale reduction factor statistic (\hat{R}) (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 $\underline{\gamma} = \{Np_x, Mq_x, Np'_x, Cq'_x\}$ 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. Briefly, using the R package `MACER` (Young et al., 2021), DNA sequences were downloaded

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 (intraspecific) distances is low (high) on average; that is, there is (no) evidence for a DNA barcode gap. |
| p'_x/p'_{lwr} | When p'_{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 (intraspecific 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'_{upr} | When p'_{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. |

from GenBank and BOLD and processed to obtain a 343 bp FASTA alignment representing 46 unique haplotypes. Genetic distances were calculated using uncorrected p-distances. 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. Further, *A. bipustulatus* comprised 46 total haplotypes, whereas *A. nevadensis* possessed two haplotypes. 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* currently 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 (**Supplementary Figure 1**). 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**).

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) |
|------------------------|------------------------|-----------------------------|----------------------------|
| <i>A. bipustulatus</i> | p_x/p_{lwr} | 1.000 (0.000; 1.000-1.000) | 1.000 (0.000; 1.000-1.000) |
| <i>A. bipustulatus</i> | q_x/p_{upr} | 1.000 (0.000; 1.000-1.000) | 1.000 (0.000; 0.999-1.000) |
| <i>A. bipustulatus</i> | p'_x/p'_{lwr} | 1.000 (0.000; 1.000-1.000) | 1.000 (0.000; 1.000-1.000) |
| <i>A. bipustulatus</i> | q'_x/p'_{upr} | 1.000 (0.000; 1.000-1.000) | 1.000 (0.000; 0.999-1.000) |
| <i>A. nevadensis</i> | p_x/p_{lwr} | 1.000 (0.000; 1.000-1.000) | 0.835 (0.144; 0.470-0.996) |
| <i>A. nevadensis</i> | q_x/p_{upr} | 0.010 (0.002; 0.006-0.014) | 0.010 (0.002; 0.007-0.014) |
| <i>A. nevadensis</i> | p'_x/p'_{lwr} | 1.000 (0.000; 1.000-1.000) | 0.834 (0.138; 0.481-0.994) |
| <i>A. nevadensis</i> | q'_x/p'_{upr} | 0.010 (0.070; -0.128-0.148) | 0.010 (0.002; 0.007-0.014) |

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 (SPIIn) (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 2**) and associated interval estimates owing to the large number of specimens sampled for this species.

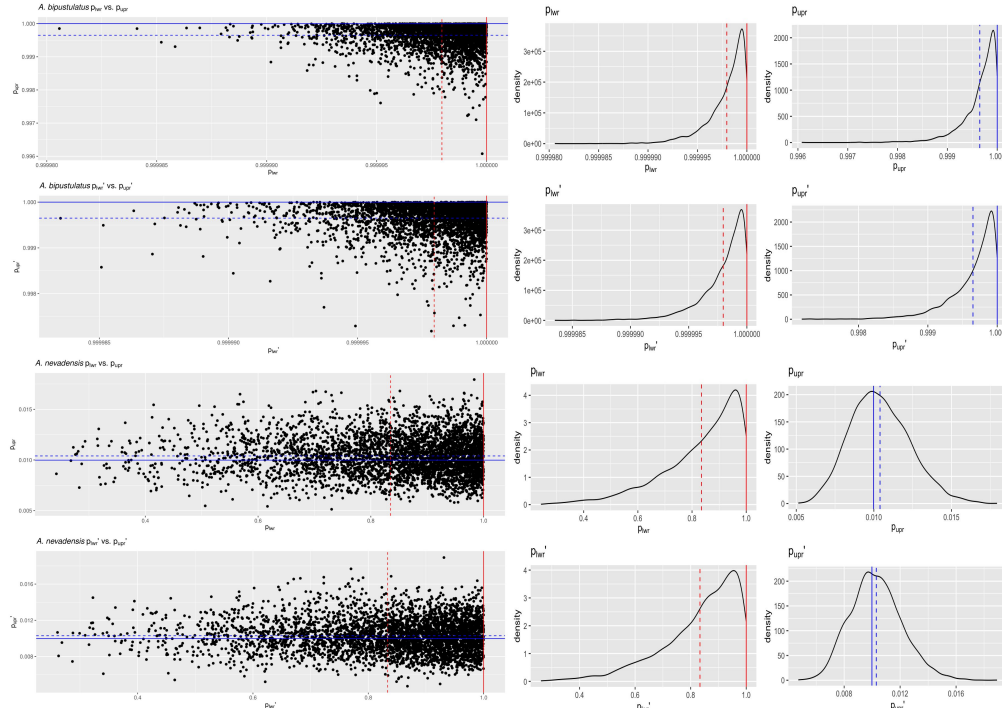


Figure 2: Scatterplots (black solid points) and distributions (black solid lines) depicting the DNA barcode gap metrics for *A. bipustulatus* ($N = 701$) and *A. nevadensis* ($N = 2$) across CYTB based on 4000 Bayesian posterior draws. MLEs and posterior means are displayed as coloured (red/blue) solid and dashed lines for the metrics, respectively.

On the other hand, the situation for *A. nevadensis* is more nuanced, as posterior values are further spread out (**Table 2** and **Figure 2**), 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 2**), whereas those for *A. nevadensis* exhibit more symmetry, especially for p_{upr} and p'_{upr} (**Figure 2**). 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 (**Supplementary Table 1**) 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. More generally, apart from the metrics being employed to
 better highlighting the importance of within-species genetic diversity versus between-species
 divergence, it is expected that the approach developed herein will be of broad utility in applied
 fields, such as DNA-based detection of seafood fraud within global supply chains, and in the
 determination of species occupancy/detection probabilities at ecological sites of interest using
 active and passive environmental DNA (eDNA) methods such as metabarcoding.

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
 $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ prior may be more appropriate over complete ignorance imposed by a $\text{Beta}(1, 1)$
 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 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 accessed via Dryad at:

<http://datadryad.org/stash/share/>

RZIfMixcEODe0RWP7eyXWQewSVbqEIA9UTrH3ZVKyn4.

A GitHub repository can be found 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.

Ahrens, D., F. Fujisawa, H.-J. Krammer, J. Eberle, S. Fabrizi, and A. Vogler

2016. Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*, 65(3):478–494.

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.

499 Čandek, K. and M. Kuntner
 500 2015. DNA barcoding gap: Reliable species identification over morphological and
 501 geographical scales. *Molecular Ecology Resources*, 15(2):268–277.

502 Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker,
 503 J. Guo, P. Li, and A. Riddell
 504 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1.

505 Carstens, B. C., T. A. Pelletier, N. M. Reid, and J. D. Satler
 506 2013. How to fail at species delimitation. *Molecular Ecology*, 22(17):4369–4383.

507 Chen, M.-H. and Q.-M. Shao
 508 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of*
 509 *Computational and Graphical Statistics*, 8(1):69–92.

510 Collins, R. A. and R. H. Cruickshank
 511 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*,
 512 13(6):969–975.

513 Dempster, A. P., N. M. Laird, and D. B. Rubin
 514 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the*
 515 *Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

516 Flouri, T., X. Jiao, B. Rannala, and Z. Yang
 517 2018. Species tree inference with BPP using genomic sequences and the multispecies
 518 coalescent. *Molecular Biology and Evolution*, 35(10):2585–2593.

519 Gelman, A., J. Carlin, H. Stern, D. Duncan, A. Vehtari, and D. Rubin
 520 2014. *Bayesian Data Analysis*, third edition. Chapman and Hall/CRC.

521 Gelman, A. and D. Rubin
 522 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*,
 523 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.
- Hickerson, M. J., C. P. Meyer, and C. Moritz
2006. DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, 55(5):729–739.
- 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.
- Jackson, N. D., A. E. Morales, B. C. Carstens, and B. C. O’Meara
2017. PHRAPL: Phylogeographic inference using approximate likelihoods. *Systematic Biology*, 66(6):1045–1053.

548 Jeffreys, H.
549 1946. An invariant form for the prior probability in estimation problems. *Proceedings*
550 *of the Royal Society of London. Series A, Mathematical and Physical Sciences*,
551 186(1007):453–461.

552 Jukes, T. and C. Cantor
553 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro,
554 ed., Pp. 21–132. New York: Academic Press.

555 Kimura, M.
556 1980. A simple method for estimating evolutionary rates of base substitutions
557 through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*,
558 16(1):111–120.

559 Kingman, J.
560 1982a. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248.

561 Kingman, J.
562 1982b. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

563 Knowles, L. L. and W. P. Maddison
564 2002. Statistical phylogeography. *Molecular Ecology*, 11(12):2623–2635.

565 Kullback, S. and R. Leibler
566 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

567 Liu, Y., A. Gelman, and T. Zheng
568 2015. Simulation-efficient shortest probability intervals. *Statistical Computing*, 25:809–819.

569 Mather, N., S. Traves, and S. Ho
570 2019. A practical introduction to sequentially Markovian coalescent methods for estimating
571 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.

Mutanen, M., S. M. Kivelä, R. A. Vos, C. Doorenweerd, S. Ratnasingham, A. Hausmann,

P. Huemer, V. Dincă, E. J. van Nieukerken, C. Lopez-Vaamonde, R. Vila, L. Aarvik,

T. Decaëns, K. A. Efetov, P. D. N. Hebert, A. Johnsen, O. Karsholt, M. Pentinsaari,

R. Rougerie, A. Segerer, G. Tarmann, R. Zahiri, and H. C. J. Godfray

2016. Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in european lepidoptera. *Systematic Biology*, 65(6):1024–1040.

Newcombe, R. G.

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

Pante, E., N. Puillandre, A. Viricel, S. Arnaud-Haond, D. Aurelle, M. Castelin, A. Chenuil,

C. Destombe, D. Forcioli, M. Valero, F. Viard, and S. Samadi

2015. Species are hypotheses: Avoid connectivity assessments based on pillars of sand.

Molecular Ecology, 24(3):525–544.

Paradis, E., J. Claude, and K. Strimmer

2004. Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*,

20(2):289–290.

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.

Puillandre, N., S. Brouillet, and G. Achaz

2021. Asap: assemble species by automatic partitioning. *Molecular Ecology Resources*, 21(2):609–620.

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz

2011. Abgd, automatic barcode gap discovery for primary species delimitation. *Molecular Ecology*, 21(8):1864–1877.

R Core Team

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

Rannala, B.

2015. The art and science of species delimitation. *Current Zoology*, 61(5):846–853.

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.

Rannala, B. and Z. Yang

2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, 66(5):823–842.

620 Ratnasingham, S. and P. Hebert
621 2007. BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular*
622 *Ecology Notes*, 7(3):355–364.

623 Ratnasingham, S. and P. D. N. Hebert
624 2013. A dna-based registry for all animal species: The barcode index number (bin) system.
625 *PLoS One*, 8(7):e66213.

626 Stan Development Team
627 2023. RStan: The R interface to Stan. R package version 2.32.6.

628 Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner
629 2021. Rank-normalization, folding, and localization: An improved \hat{R} for assessing
630 convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

631 Wells, T., T. Carruthers, P. Muñoz-Rodríguez, A. Sumadijaya, J. R. I. Wood, and R. W.
632 Scotland
633 2022. Species as a heuristic: Reconciling theory and practice. *Systematic Biology*,
634 71(5):1233–1243.

635 Wickham, H.
636 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

637 Yang, Z. and B. Rannala
638 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the*
639 *National Academy of Sciences*, 107:9264–9269.

640 Yang, Z. and B. Rannala
641 2014. Unguided species delimitation using dna sequence data from multiple loci. *Molecular*
642 *Biology and Evolution*, 31(12):3125–3135.

643 Yang, Z. and B. Rannala

644 2017. Bayesian species identification under the multispecies coalescent provides significant
645 improvements to DNA barcoding analyses. *Molecular Ecology*, 26:3028–3036.

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

647 2021. Molecular Acquisition, Cleaning and Evaluation in R (MACER) - A tool to assemble
648 molecular marker datasets from BOLD and GenBank. *Biodiversity Data Journal*, 9:e71378.