A Bayesian Model of the DNA Barcode Gap

Jarrett D. Phillips[1,3]* (ORCID: 0000-0001-8390-386X)

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

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

Email Address: jphill01@uoguelph.ca

Running Title:

Abstract

Since its inception over 20 years ago, DNA barcoding has emerged as a robust method of specimen identification and species delimitation across myriad taxonomic groups which have been sequenced at short, standardized gene regions like 5'-COI for animals. However, the success of the approach depends crucially on two important factors: (1) the availability of high-quality specimen records found in public reference sequence databases such as BOLD, and (2) the establishment of a DNA barcode gap — the idea that the maximum genetic distance observed within species is much smaller than the minimum degree of marker variation found among species. 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. Despite this, many taxa lack adequate separation in their pairwise intraspecific and interspecific genetic distances, thereby compromising rapid matching of unknown samples to expertly-validated references.

Recent work has argued that DNA barcoding, in its current form, is lacking in statistical rigor, calling into question the existence of a true species' DNA barcode gap. To support this notion, novel nonparametric locus-specific metrics based on the multispecies coalescent were recently outlined and shown to hold strong promise when applied to *Agabus* diving beetles. The metrics quantify the extent of asymmetric directionality of proportional genetic distance distribution overlap/separation for species within well-sampled genera based on a straightforward distance count. Values of the metrics close to zero suggest the existence of DNA barcode gaps, whereas values near one lend credence for the absence of gaps. However, what appears to be missing is an unbiased way to compute the statistical accuracy of the recommended estimators arising through problems inherent in frequentist maximum likelihood estimation for discrete probability distributions having bounded support. Here, a Bayesian model of the DNA barcode gap coalescent, written using the Stan software, is introduced to rectify such issues. The model allows accurate estimation of posterior means, posterior standard deviations, posterior quantiles, and credible intervals for the metrics given datasets of intraspecific and interspecific genetic distances for species of interest.

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

1 Introduction

DNA barcoding (Hebert et al., 2003a,b) was conceived more than two decades ago as an immediate and automatic solution to the taxonomic impediment during a time of ongoing biodiversity crisis. The technique promised the rapid and accurate identification of unknown specimens to known Linnaean binomens, as well as the unambiguous resolution of species boundaries across the eukaryotic Tree of Life through the leveraging of easily obtained genetic variation found in short, universal segments of DNA. In animals, a *c.* 658 bp fragment from the 5' end of the mitochondrial cytochrome *c* oxidase subunit I gene is employed as a DNA barcode due to its ease of isolation, amplification, and sequencing. As a result, dedicated community-curated genomic sequence databases like the Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert, 2007) have accumulated appreciable levels of intraspecific genetic variation across a wide range of taxa and geographic regions to accelerate specimen identification and species discovery tasks essential to addressing novel research questions in ecology and evolutionary biology. Despite this,

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:

$$p_x = \frac{\#\{d_{ij} \geq \min(d_{XY})\}}{\#\{d_{ij}\}} \tag{1}$$

$$q_x = \frac{\#\{d_{XY} \leq \max(d_{ij})\}}{\#\{d_{XY}\}} \tag{2}$$

where d_{ij} and d_{XY} are distances within and among species, respectively, and the notation # reflects a count. Distances are easily computed from a model of DNA sequence evolution, such as p distance. Similar expressions (denoted $p_x^{'}$ and $q_x^{'}$) for nearest neighbour species were also given (see Phillips et al. (2024)), in which d_{XY} included only interspecific distances between the species of interest and its closest neighbouring species. If a focal species is found to have multiple nearest neighbours, then the species possessing the smallest average pairwise interspecfic distance is used. While these schemes differ considerably from the usual definition of the DNA barcode gap laid out by Meyer and Paulay (2005) and Meier et al. (2008), they more accurately account for species' coalescence histories inferred from contemporaneous DNA sequences. such as hybridization/introgression events (Phillips et al., 2024). Note, distances (and hence the metrics) are constrained to the closed interval [0, 1]. Values of the estimators obtained from equations (1) and (2) close to or equal to zero give evidence for separation between intraspecIfic and interspecific genetic distance distributions; that is, values suggest the presence of a DNA barcode gap. Conversely, values near or equal to one give evidence for distribution overlap; that is, values likely indicate the absence of a gap.

2.2 A Bayesian Implementation

A major criticism of large sample (frequentist) theory is that it relies on asymptotic properties of the MLE (which is assumed to be fixed but unknown), such as normality. This problem is especially pronounced in the case of binomial proportions. The estimated Wald SE of the sample proportion, is given by

$$\widehat{SE[\hat{p}]} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{3}$$

where $\hat{p} = \frac{Y}{n}$ is the MLE, Y is the number of successes ($Y = \sum_{i=1}^{n} y$) and n is the number of trials. However, the above formula is problematic for several reasons. First, Equation (3) 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. 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 (3) completely uselesss. 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 direct specification of prior beliefs concerning unknown model parameters and permits the seamless propagation of uncertainty when data is lacking through integration with the likelihood function.

2.3 The Model

Counts, y, of overlapping genetic distances (as expressed in the numerator of Equations (1) and (2)) are treated as binomially distributed with expectation $\mathbb{E}[Y] = k\theta$, where $k = \{N, M, C\}$ are total counts of intraspecific, interspecific, and combined genetic distances for a target species, and $\theta = \{p_x, q_x, p_x^{'}, q_x^{'}\}$. Although the metrics encompassing θ are presumed to follow a beta distribution, for simplicity, they are given uninformative standard uniform ($U(0, 1)$) priors, which is mathematically equivalent to a Beta(1, 1) distribution. Such a scheme is quite convenient since the beta distribution is conjugate to the binomial distribution. Thus, the posterior distribution is also beta distributed. Specifically, the posterior is Beta$(Y + 1,\, n - Y + 1)$ which has expected value $\mathbb{E}[Y] = \frac{Y+1}{n+2}$. Thus, the full univariate Bayesian model is given by

3 Results

4 Discussion

5 Conclusion

Supplementary Information

Information accompanying this article can be found in Supplemental Information.pdf.

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, 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, approved all developed code as well as analysed and interpreted all experimental results.

References

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.

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.

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.

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