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

2

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.

2 Methods

2.1 DNA Barcode Gap Metrics

2.2 A Bayesian Implementation

When p is close to zero, it suggests that the probability of intraspecific distances being larger than interspecific distances is low on averqge, while the probability of interspecific distances being larger than intraspecific distances is high on average; that is, there is evidence for a DNA barcode gap.

When q is close to one, it indicates that the probability of intraspecific distances being larger than interspecific distances is high on average, while the probability of interspecific distances being larger than intraspecific distances is low on average; there is no evidence for a DNA barcode gap.

If the minimum interspecfic distance is relatively large and maximum intraspecfic distance is relatively small, then p_lwr represents the extent to which intraspecific distances tend to be larger than interspecific distances at and beyond the minimum interspecific distance and at and below the maximum intraspecific distance.

If the maximum intraspecfic distance is relatively large and the minimum interspecfic distance is relatively small, p_upr represents the extent to which interspecific distances tend to be larger than intraspecific distances below the maximum intraspecific distance and beyond the minimum interspecific distance.

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/Phillips-et-al.-Seafood-Fraud-Paper.

Acknowledgements

Funding

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.