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 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 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. 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: DNA barcoding, intraspecific genetic distance, interspecific genetic distance, Stan

1 Introduction

2 Methods

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

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