OTHERS

REVISTA DE
Biología Tropical

https://doi.org/10.15517/rev.biol.trop..v72i1.59992

A response to: Evaluating the reliability of DNA Barcoding for Central American Pacific shallow water echinoderms identification

Spencer Kelvin Monckton[1]; https://orcid.org/0000-0002-9879-9118
Dirk Steinke[1,2]; https://orcid.org/0000-0002-8992-575X
Kevin Charles Robert Kerr[1,2*]; https://orcid.org/0000-0002-6784-3884

1. Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada; dsteinke@uoguelph.ca, smonckto@uoguelph.ca, kkerr@uoguelph.ca (*Correspondence)
2. Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada

Received 16-V-2024. Corrected 27-VII-2024. Accepted 06-VIII-2024.

ABSTRACT

Introduction: Chacón-Monge et al. (2024) sought to test the accuracy of DNA barcoding for species identification in Pacific Central American shallow water echinoderms. They used cytochrome *c* oxidase I (COI) sequences derived from new material collected as part of the BioMar-ACG project in Costa Rica. Using their set of 348 echinoderm sequences, they compared species identification results from two online platforms: the National Center for Biotechnology Information (NCBI) GenBank using the nucleotide Basic Local Alignment Search Tool (BLASTn), and the Barcode of Life Data Systems (BOLD) Identification Engine.
Objective: The present article is a response to their results and conclusions.
Methods: We reinterpreted the results from the authors' Appendix 2 to enable an objective comparison between the BOLD Identification Engine and BLASTn in GenBank.
Results: While the authors found that both platforms were limited by the number of reference sequences available in their respective databases, they concluded that GenBank outperformed BOLD for identification; however, we identify several methodological flaws in their analysis. These include pseudoreplication amongst query sequences, contaminated sequences stemming from sampling errors, and a lack of standardization when interpreting results from the two platforms. Their assessment of the BOLD Identification Engine was also limited by improper selection of a reference database.
Conclusion: Addressing these errors, we reinterpret their results and demonstrate that there is no difference in performance between the two platforms.

Key words: BOLD; GenBank; BLASTn; mitochondrial DNA; sequencing errors; Costa Rica.

RESUMEN

Una respuesta a: Evaluando la fiabilidad del Código de Barras de ADN para la identificación de equinodermos en aguas poco profundas del Pacífico de América Central

Introducción: Chacón-Monge et al. (2024) intentaron probar la precisión de los códigos de barras de ADN para identificar especies de equinodermos de aguas poco profundas del Pacífico Centroamericano. Para ello, utilizaron secuencias de citocromo *c* oxidasa I (COI) provenientes de material recolectado recientemente como parte del proyecto BioMar-ACG en Costa Rica. Utilizando 348 secuencias de equinodermos, compararon los resultados de identificación de especies de dos plataformas en línea: GenBank del Centro Nacional de Información Biotecnológica (NCBI) empleando la herramienta de búsqueda de alineación local básica de nucleótidos (BLASTn) y la herramienta de identificación del Sistema de Datos del Código de Barras de la Vida (BOLD).

Objetivo: El presente artículo es una respuesta a sus resultados y conclusiones.
Métodos: Reinterpretamos los resultados presentados por los autores en Apéndice 2 para comparar objetivamente el sistema de identificación de BOLD y el de BLASTn en GenBank.
Resultados: Si bien los autores encontraron que ambas plataformas estaban limitadas por la cantidad de secuencias de referencia disponibles en sus bases de datos, concluyeron que GenBank superó a BOLD en la identificación de especies; sin embargo, notamos varias fallas metodológicas en su análisis. Estas incluyeron la pseudorreplicación entre las secuencias consultadas, el uso de secuencias contaminadas derivadas de errores de muestreo y falta de estandarización al interpretar los resultados de las dos plataformas. Su evaluación del sistema de identificación de BOLD se vio limitada por la selección inadecuada de una base de datos de referencia.
Conclusión: Teniendo en cuenta estos errores, reinterpretamos sus resultados y demostramos que no existe una diferencia significativa en el rendimiento de ambas plataformas.

Palabras clave: BOLD; GenBank; BLASTn; ADN mitocondrial; errores de secuenciación; Costa Rica.

INTRODUCTION

Chacón-Monge et al. (2024) sought to test the accuracy of DNA barcoding for species identification in Pacific Central American shallow water echinoderms using sequences derived from newly collected material as part of the BioMar-ACG project in the Área de Conservación Guanacaste, Costa Rica (Cortés & Joyce, 2020). They obtained cytochrome *c* oxidase I (COI) sequences from 348 out of 475 echinoderm specimens collected during their survey (approximately 72 %). They used morphological characters to identify 325 specimens to species, five to genus, 14 to family, and four just to class, totalling 51 unique taxonomic assignments (according to information in their Appendix 2). The authors then compared the performance of two online platforms for species identification: the National Center for Biotechnology Information (NCBI) GenBank using the nucleotide Basic Local Alignment Search Tool (BLASTn; Altschul et al., 1990), and the Barcode of Life Data Systems (BOLD) Identification Engine (Ratnasingham & Hebert, 2007). The authors concluded that 53.5 % of the specimens they had morphologically identified to species were "correctly" identified using GenBank, whereas only 33.9 % of them were when using BOLD. They attributed the generally low performance of these platforms to misidentifications in their respective databases and insufficient regional representation.

While we agree with the authors' observation that DNA barcodes provide complementary information to identify Pacific Central American shallow water echinoderms and that increased sampling would improve the available species identification tools, we also identified a number of flaws in their methodology that impede their assessment, particularly with respect to BOLD. We outline these errors below and provide a reinterpretation of their results using data reported in their Appendix, which demonstrates that the performance of the two platforms is actually very similar.

Statistical errors: The authors did not sample evenly across taxonomic entities, which invalidates the numbers they have reported for species identification success rates for the two platforms. Considering both morphospecies and provisional species together, the modal value of replicates per species was only 2, whereas the mean was 6.8 on account of representation being heavily right-skewed (skewness = 2.41; kurtosis = 7.37). *Holothuria impatiens* was notably represented by 40 specimens. This sampling distribution is not necessarily problematic and is in fact expected when samples are derived from a biological survey, where more common species are likely to be sampled more frequently by chance alone. However, if intraspecific variation is low, simply treating each identification result of individuals within a species as an independent observation leads to pseudoreplication (Hurlbert, 1984).

This issue is apparent in the case of *H. impatiens*, where 22 of the samples appear to share an identical sequence and had a match via BLASTn but not BOLD. By the authors' analysis, this was counted as 22 independent cases of GenBank outperforming BOLD; rather, this should reflect only one instance of differing performance. Consequently, their approach has resulted in inflated values where they report identification error rates (e.g., see Table 1 in Chacón-Monge et al., 2024), obfuscating the real difference in performance between the two platforms. Alternatively, the authors could have taken advantage of these replicates to examine intraspecific variation, as has been done in prior studies (e.g., Layton et al., 2016). Contrasting intraspecific and interspecific variation would have yielded further insights into the performance of COI barcodes for species delimitation in Central American echinoderms.

Sampling Errors: The authors' sweeping interpretation of mismatched species identifications as failures of the identification platforms is flawed, as there are alternative, more parsimonious explanations in a number of cases. For example, sequence BMAR368-19 was supposedly derived from an easily recognizable sea star, *Nidorellia armata* (a member of the order Valvatida), but had > 97 % sequence similarity to records from *Toxopneustes* spp. (sea urchins from the order Camarodonta) in *both* GenBank and BOLD, which the authors scored as an identification error. Conversely, sequence BMAR369-19 was meant to represent *T. roseus*, but was identified by the BOLD Identification Engine as *N. armata*, which was also interpreted as a misidentification. Neither of these outcomes is very likely to be correct; rather, the obvious interpretation is that the two samples were swapped during sampling or subsampling, with this field error later being mistakenly attributed to the two platforms. We found 13 such instances, in which a sequence shared a high degree of similarity to an unrelated species (i.e., different genus, order, or class) that was included in the sampling effort, suggesting sample mix-ups or contamination (see SMT1). In another 17 instances, we noted probable contamination or misidentification either due to identifications being mismatched at the rank of class or higher, or due to the query sequence failing to match an available congener sequence in one or the other database. We additionally noted eight instances of possible mix-ups or contamination between samples of related species of *Holothuria*, making it difficult to determine whether the resulting species identifications represented true errors or false negatives. This reinforces the importance of interpreting results carefully, such as considering sequencing results for each species holistically.

Lack of Standardization Between Platforms: There are fundamental differences in the operation of the two molecular identification platforms used by the authors, which were not addressed in their study. The BLASTn tool is not intended to provide a species-level identification, but rather to align to the *most* similar sequence(s) in the database. Thus, when highly similar sequences are missing from the database, matches can still be returned from distantly related taxa. In contrast, BOLD employs divergence thresholds to avoid returning distantly related taxa as a "species match" and will abort the identification algorithm if a sequence match exceeding 97 % is not found. Thus, quantifying the cases of "no match" does not provide a meaningful comparison of the two platforms because only one of them is likely to yield this result.

This distinction is especially important when considering genus-level identifications because the BOLD Identification Engine is not designed for this purpose. To illustrate, consider a case where the best available matches for a query sequence are at most 96.9 % identical, and these sequences are present in both databases: these would appear in BLASTn results and could readily be interpreted as genus-level matches, whereas BOLD would simply return "no match" (i.e., no species match). We note that BLASTn failed to match a sequence in GenBank with greater than 97 % similarity in 166 cases, which is not significantly different than the number of cases with BOLD (n = 178). Considering that BOLD will return a list of hits in order of similarity, akin to output from BLASTn, comparing hit tables from each platform for the remainder of cases would have provided a more appropriate comparison of their performance.

Reference Database: Lastly, while access to sequences through BOLD is limited to the public database, private unpublished sequences are nevertheless available for comparison via the BOLD Identification Engine. By default, the Identification Engine searches against all "Species Level Barcode Records", nearly five million sequences at least 500 bp in length that have species-level identifications. The "Public Record Barcode Database" used by Chacón-Monge et al. (2024) includes only published records, is less than half the size (2.39 M records) and tends to deliver fewer species-level identifications. This distinction is not paralleled in GenBank, which is entirely public, so although the Public Record Barcode Database might arguably provide a more direct comparison of the two platforms in terms of data availability, the default Species Level Barcode database is the more appropriate choice for comparing species identification capabilities between the two platforms. For example, the sequence from BMAR484-19 was identified by the authors as *Labidodemas maccullochi* based on morphological analysis but came up with no match in the Identification Engine when the authors queried the sequence; however, there were four representative sequences from the species in BOLD at the time of their analysis (a fact easily verified using various search tools in BOLD), and thus they likely would have received a correct species ID had they used the "Species Level Barcode Records" database (a genus-level match was not noted by the authors because congeners for this species in BOLD had less than 97 % similarity). This is an exceptional case because a second sequence (i.e., BMAR889-20) was also supposed to represent this species but had a low-level match via BLASTn to *Ophionereis reticulata* instead, as did their sequences from *O. annulate* specimens, suggesting that the sequence from BMAR889-20 was the result of contamination from specimens of the latter species. Because there was a sequence from *O. annulata* in BOLD, this contamination would have been more apparent had the "Species Level Barcode Records" database been used.

MATERIALS AND METHODS

We reinterpreted the results from the authors' Appendix 2 (see our SMT 1) to enable an objective comparison between the BOLD Identification Engine and BLASTn in GenBank. We interpret a match as any result with concordant taxonomy that exceeds 97 % pairwise identity. Non-matching identifications above 97 % are counted as errors, while non-matches and hits below 97 % are counted as gaps. We also counted instances of improved identifications, wherein the sequence delivered an identification with more specific taxonomy than indicated by the morphological identification.

To tally the results, we excluded instances of contamination or sample mix-ups and counted each distinct outcome per taxon only once (see our SMT 2). For example, all five samples of *Nidoriella armata* resulted in a match from BOLD and a gap from GenBank, so are counted only once; meanwhile, five samples of *Pharia pyramidata* returned gaps from both BOLD and BLASTn, while one sample returned a gap from BOLD and an error from BLASTn – these are counted as two distinct outcomes. To compare identification success between the two platforms, we used the 'mcnemar.test' function from the stats package (v.4.3.1) in R (R Core Team, 2023).

RESULTS

Our counts of the identification outcomes of BLASTn and BOLD are provided in Table 1. To compare genus and species level identifications between the two platforms, we only counted matches above 97 %. For both platforms, identification success was below 51 % at the genus level and below 44 % at the species level. Contrary to the results of Chacón-Monge et al. (2024), we found no significant difference in success rate between the two platforms for either genus or species level identifications (McNemar's chi-squared = 0.5, df = 1, p = 0.48; McNemar's chi-squared = 0.94, df = 1, p = 0.33).

Table 1

Summary of re-interpreted results from Chacón-Monge et al. (2024) comparing sequence-based identification of Central American Pacific echinoderms with GenBank and BOLD. Counts are of distinct outcomes for each taxon represented among the query sequences.

ID	GenBank BLASTn	BOLD ID Engine
Match	24	19
Improved	2	4
Error	7	6
Gap	31	35

DISCUSSION

We do not find that the data presented by Chacón-Monge et al. (2024) in any way support their claim that "GenBank outperforms BOLD" in terms of identification accuracy. If anything, we would point to the fact that the BOLD Identification Engine produced one fewer error and provided improved identifications for two additional records compared to BLASTn, suggesting that BOLD delivers better accuracy overall. Perhaps the most important difference between the two platforms is in the number of gaps in BOLD that are now notably filled by Chacón-Monge et al.'s own data; however, their data remains private and thus only benefits users that perform a BOLD Identification Engine search using the full species level barcode database. Unfortunately, the fact that the authors have not yet made their data publicly available on either BOLD or GenBank prevents them from addressing the very issue they identified (i.e., the gap in reference sequences available for Central American echinoderms) and limits the reproducibility of their analysis.

Ethical statement: the authors declare that they all agree with this publication and made significant contributions; that there is no conflict of interest of any kind; and that we followed all pertinent ethical and legal procedures and requirements. All financial sources are fully and clearly stated in the acknowledgments section. A signed document has been filed in the journal archives.

See supplementary material a39v72n1-MS1

ACKNOWLEDGMENTS

We thank the editors for the opportunity to publish this response article. This work was funded, in part, by awards in support of BIOSCAN-Canada and the Centre for Biodiversity Genomics from the Government of Canada through the Canada Foundation for Innovation (2023 Major Science Infrastructure Program; 42450) and Genome Canada and Ontario Genomics (Large-Scale Applied Research Project; OGI-208).

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Chacón-Monge, J. L., Abarca-Odio, J. I., & González-Sánchez, K. (2024). Evaluating the reliability of DNA Barcoding for Central American Pacific shallow water echinoderms identification: a molecular taxonomy and database accuracy analysis. *Revista de Biología Tropical*, *72*(S1), e58997. https://doi.org/10.15517/REV.BIOL.TROP..V72IS1.58997

Cortés, J., & Joyce, F. (2020). BioMar-ACG: A successful partnership to inventory and promulgate marine biodiversity. *Biotropica*, *52*(6), 1103–1106. https://doi.org/10.1111/BTP.12841

Hurlbert, S. H. (1984). Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, *54*(2), 187–211. https://doi.org/10.2307/1942661

Layton, K. K. S., Corstorphine, E. A., & Hebert, P. D. N. (2016). Exploring Canadian Echinoderm Diversity through DNA Barcodes. *PLOS ONE*, *11*(11), e0166118. https://doi.org/10.1371/JOURNAL.PONE.0166118

R Core Team (2023). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/J.1471-8286.2007.01678.X