MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2011) **11**, 941–942 doi: 10.1111/j.1755-0998.2011.03067.x

NEWS AND VIEWS

REPLY

BOLD's role in barcode data management and analysis: a response

S. RATNASINGHAM and P. D. N. HEBERT
Biodiversity Institute of Ontario, University of Guelph, Guelph, ON N1G 2W1, Canada

Abstract

DNA barcoding is a very effective tool for the identification of specimens when a carefully validated and taxonomically comprehensive library of reference DNA barcodes is available. Libraries meeting this criterion are now available for some taxonomic groups in some geographic regions, provoking their use as a tool for the identification of samples that would otherwise remain as unknowns. In this article, we emphasize the need for caution in the interpretation of identifications based on a reference library with entries that have seen limited validation. We also emphasize the need for the deposition of sequence records for 'unknowns' so that presumptive identifications can be tested by other researchers and updated as the barcode reference library gains increased coverage and validation.

Keywords: Barcode of Life Data System, birdstrikes, Database, DNA barcoding

Received 8 July 2011; revision accepted 14 July 2011

In his comment, Federhen (2011) identified two major areas of concern in relation to the deposition and interpretation of DNA barcode data in Waugh *et al.* (2010). As his concerns in relation to data interpretation focused on operational aspects of the Barcode of Life Data System (BOLD), we take this opportunity to clarify its role and the use of its identification engine.

The Barcode of Life Data System is supporting the development of a DNA-based system for species identification (Ratnasingham & Hebert 2007) by providing users with analytical tools, computational resources and access to structured, high-volume data storage. Reflecting its role as a barcode workbench, BOLD allows users to keep certain data elements private and to determine the timing of full data release. However, other data elements are automatically exposed upon submission, reflecting the fact that users are working in a wiki-like environment. Data elements that see immediate public access include taxonomy, geography and images, although it needs emphasis that these attributes are displayed as summaries and are not linked one-on-one to a particular specimen unless the data owner has opted for full data release. Because of its role as a barcode workbench, BOLD unavoidably includes some sequences that reflect analytical error and others that derive from specimens whose taxonomic assignment is uncertain or incorrect. However, it also acts as a repository for sequence and specimen data that have experienced full validation. The differential maturity of barcode records held by BOLD is reflected in the design of its ID Engine.

The Barcode of Life Data System's ID Engine operates in three modes that differ in the inclusivity of their underlying reference sequence set. The mode based on published records has the lowest taxonomic coverage, but the highest reliability. The full data set is, by contrast, parameterized with all public and private sequences, and its use carries risks as many of the private sequence records have seen limited validation and curation. The identification assigned to a particular query sequence by the ID Engine is susceptible to indeterminacy through time as new records are added to the reference library (or subtracted if a record is derived from a misidentified or improperly analysed specimen). The ultimate solution to such indeterminacy involves the construction of a reference database parameterized with carefully validated sequence records for all species from all taxonomic groups, but this goal is distant. Federhen (2011) is correct to emphasize that scientific results must be repeatable. Primary data that have been

Correspondence: S. Ratnasingham, Fax: 1-519-824-5703;
E-mail: sratnasi@uoguelph.ca

used for a publication must be lodged in a persistent repository so that they are available for examination. As well, for those interested in re-examining published species identifications generated from such sequence data, it is critical to be able to test the sequences used for identification against the database originally employed for their taxonomic diagnosis. To meet this need, BOLD has established a set of date-stamped reference sequence databases for its ID engine. These databases are updated annually and each version is archived, providing a basis to test past assignments and to monitor the rising efficacy of the ID Engine as parameterization expands.

The complexities of data analysis and validation are growing as a consequence of the deluge of DNA sequence information. Databases are rising that utilize new models of data curation and release to address this challenge. Projects like Wiki-pathways (Pico *et al.* 2008), the genome re-annotation project (Salzberg 2007) and Transdab wiki (Csosz *et al.* 2009) have been created to empower users, enabling more decentralized control over data. Conventional databases like GenBank retain a critical role as persistent archives for primary data. However, it is clear that informatics platforms using alternate curatorial strategies are essential to channelize the flood of sequence data. BOLD has been developed to meet this need for the DNA barcoding community and is steadily evolving new capacities to better serve its users and to better interface with GenBank and other members of the INSDC.

Acknowledgements

The development of BOLD is supported by the Ontario Ministry of Research and Innovation, by Genome Canada through the Ontario Genomics Institute and by NSERC in aid of the International Barcode of Life Project. We thank Dan Janzen, Justin Schonfeld and Dirk Steinke for revisionary suggestions to this note.

References

Csosz E, Mesko B, Fesus L (2009) Transdab wiki: the interactive transglutaminase substrate database on web 2.0 surface. *Amino Acids*, **36**, 615–617.

Federhen S (2011) Comment on ''Birdstrikes and barcoding: can DNA methods help make the airways safer''. *Molecular Ecology Resources*, doi: 10.1111/j.1755-0998.2011.03054.x, in press.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Elvo C (2008) WikiPathways: pathway editing for the people. *PLoS Biology*, **6**, e184. DOI: 10.1371/journal.pbio.0060184

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364. DOI: 10.1111/j.1471-8286.2006.01678.x

Salzberg LS (2007) Genome re-annotation: a wiki solution?. *Genome Biology*, **8**, 102.

Waugh J, Evans MW, Millar CD, Lambert DM (2010) Birdstrikes and barcoding: can DNA methods help make the airways safer? *Molecular Ecology Resources*, **11**, 38–45.