B144 Public Sequence Databases: An Assessment of Their Reliability for Identifying Non-Human Biological Material

Kelly A. Meiklejohn, PhD, ORISE/FBI Laboratory, 2501 Investigation Parkway, Quantico, VA 22204; Natalie Damaso, PhD, ORISE/FBI Laboratory, 2334 Brookmoor Lane, Woodbridge, VA 22191; and James M. Robertson, PhD, CFSRU, FBI Laboratory, 2501 Investigation Parkway, Quantico, VA 22135*

After attending this presentation, attendees will understand: (1) the differences between the two main public databases of DNA barcode data (GenBank and Barcode Of Life Data Systems (BOLD)) with regard to the number of sequences, curation, and quality checks; (2) the success of both databases for obtaining the correct taxonomic assignment for insects, plants; and fungal taxa; and, (3) the precautions necessary when using public sequence databases in a forensic setting.

This presentation will impact the forensic science community by identifying varied challenges of using public sequence databases for the identification of unknown biological material encountered in casework.

Crime laboratories routinely receive evidence which contains non-human biological material. If identified, such material could help reduce the search area in provenance cases or identify the best method for isolating a plant toxin. DNA barcoding permits species-level identification of biological materials through the comparison of the unknown barcode sequence to a reference database.

There are two main public sequence databases containing barcode data, BOLD and GenBank, the latter for which the data is not curated. This study performed an initial assessment of both the quality and reliability of the DNA barcode data contained in these databases, a prerequisite for their use in a forensic setting. To achieve this, curated reference material was sourced from national collections, with taxa chosen based on their inclusion in BOLD but also to represent the main lineages of plants, macro-fungi, and insects (total *n*, ~150). The relevant barcode sequences from these reference samples (rbcL, matK, trnH-psbA, ITS, and COI) were generated and used for searching against both databases. The ability of each database was assessed to obtain the correct taxonomic assignment (genus and species), when using the default search parameters; GenBank outperformed BOLD for insect taxa (86% and 50%, respectively) whereas for plant and fungal taxa, both databases performed comparably (~78% and ~64%, respectively). Considering that the correct match was often not discernible among the top matches, modified searches against each database were performed to assess whether resolution improvements were possible. Given that the underlying algorithm and associated parameters for searching BOLD are fixed, modified searches were limited to changing the subset of barcode sequences against which an unknown is compared (i.e., all records, only records with species level identifications, only full-length sequences). For a blast search against GenBank, the impact of altering parameters, including word-size and the penalties/rewards for mismatches and gaps, was systematically assessed.

This presentation will outline the optimal search parameters needed to consistently obtain the correct identification of an unknown when using either the BOLD or GenBank database. Additionally, some precautions needed when using public sequence databases in a forensic setting will be identified.

Public Sequence Databases, GenBank, BOLD

*Presenting Author