ANSI/ASB Standard 180, First Edition
2024

Standard for the Selection and Evaluation of GenBank® Results for Taxonomic Assignment of Wildlife

Standard for the Selection and Evaluation of GenBank® Results for Taxonomic Assignment of Wildlife

ASB Approved September 2023

ANSI Approved May 2024

410 North 21st Street
Colorado Springs, CO 80904

Foreword

This standard defines the minimum requirements that shall be met when comparing evidentiary sequences to those in GenBank® for taxonomic assignment of non-human samples. The aim is to provide a framework that will result in consistency in the wildlife forensic DNA community. Forensic scientists using these standards are expected to have a working understanding of DNA sequencing, taxonomy, and phylogeny.

This standard is intended to assist those using GenBank® for the taxonomic identification of wildlife in forensic casework.

The American Academy of Forensic Sciences established the Academy Standards Board (ASB) in 2015 with a vision of safeguarding Justice, Integrity and Fairness through Consensus Based American National Standards. To that end, the ASB develops consensus based forensic standards within a framework accredited by the American National Standards Institute (ANSI), and provides training to support those standards. ASB values integrity, scientific rigor, openness, due process, collaboration, excellence, diversity and inclusion. ASB is dedicated to developing and making freely accessible the highest quality documentary forensic science consensus Standards, Guidelines, Best Practices, and Technical Reports in a wide range of forensic science disciplines as a service to forensic practitioners and the legal system.

This document was revised, prepared, and finalized as a standard by the Wildlife Consensus Body of the AAFS Standards Board. The draft of this standard was developed by the Biology/Wildlife Forensic Biology Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science.

Questions, comments, and suggestions for the improvement of this document can be sent to AAFS-ASB Secretariat, asb@aafs.org or 401 N 21st Street, Colorado Springs, CO 80904.

All hyperlinks and web addresses shown in this document are current as of the publication date of this standard.

ASB procedures are publicly available, free of cost, at www.aafs.org/academy-standards-board.

Keywords: *GenBank®, BLAST, DNA, Public sequence databases, Taxonomic identification, Wildlife*

Table of Contents

Standard for the Selection and Evaluation of GenBank® Results for Taxonomic Assignment of Wildlife

1 Scope

This standard provides the minimum requirements for selection and evaluation of DNA sequences retrieved from the National Center for Biotechnology Information's GenBank® and their subsequent use as reference material for taxonomic identification of wildlife[a].

This standard does not cover the use of DNA sequences from other public sequence databases (e.g., BOLD, UNITE), the protocol for downloading sequences from GenBank® for inclusion in in-house databases, or the use of custom BLAST searches against GenBank®. However, the criteria can be conceptually applied to other sequence databases.

2 Normative References

The following reference is indispensable for the application of the standard. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ANSI/ASB Standard 048, *Wildlife Forensic DNA Standard Procedures*, First Edition, 2019[b]

3 Terms and Definitions

For purposes of this document, the following definitions apply.

3.1
alignment
An arrangement of two or more nucleotide or protein sequences that is used to illustrate similarity among those sequences.

3.2
Basic Local Alignment Search Tool
BLAST
The a) BLAST algorithm, and b) a suite of database search programs that implement variations of this algorithm to generate alignments between a nucleotide or protein sequence in a query, and nucleotide or protein sequences within a database.

3.3
expectation value
e-value
The number of distinct alignments expected by chance; the default sorting metric in BLAST search results.

[a] For the purposes of this document, "wildlife" species are defined as non-human multicellular animals and plants, whether wild, captive-bred, or domesticated.

3.4
GenBank®
A public repository of DNA sequences maintained by the National Center for Biotechnology Information, part of the U.S. National Institutes of Health.

3.5
hit(s)
Sequence(s) returned from GenBank® when performing a BLAST search. Also known as a "subject sequence."

3.6
interspecific
Between members of different species.

3.7
intraspecific
Between members of the same species.

3.8
National Center for Biotechnology Information
NCBI
The U.S. National Center for Biotechnology Information is located in Bethesda, Maryland and is part of the United States National Library of Medicine (a branch of the National Institutes of Health). NCBI houses a series of databases relevant to biotechnology and biomedicine and provides several bioinformatics tools for searching and analyzing the housed data.

3.9
phylogram
A branching diagram that illustrates putative relationships amongst organisms.

NOTE Phylograms are typically generated using genetic sequences and/or morphological characters.

3.10
query
(n) The nucleotide or protein sequence that has an unknown source (i.e., evidence sequence), or (v) the action of searching an unknown sequence against a database.

3.11
query coverage
The percent of the query sequence length that is included in the aligned segment with a hit.

3.12
sequence identity
The percentage or number of nucleotides or amino acids that are identical between two sequences.

3.13
taxonomic identification
Analyses to establish the classification of biological evidence to family, genus, species, etc. These analyses are based on class characters (e.g., morphological, genetic) that are diagnostic for the taxonomic level in question.

3.14
topology
The branching structure of a phylogram.

3.15
voucher specimen
Biological specimen that is representative of its species in accordance with the relevant taxonomic authority and is therefore valid for comparative purposes. Voucher specimens are of known identity, and are curated with available associated geographic, field collection, and life history data.

4 Significance and Use

4.1 These are minimum requirements applicable to the taxonomic assignment of evidentiary items to the lowest appropriate taxonomic level (e.g., species).

4.1.1 As these are minimum requirements, they may not be sufficient for accurate taxonomic assignment of unknown taxa in all circumstances.

4.1.2 No single sequence dissimilarity threshold can apply to all taxa because rates of genome evolution and intraspecific divergence vary by species.

4.1.3 Accurate taxonomic assignment of a sequence of interest depends on:

— the use of validated methods (e.g., ANSI/ASB Std 019);

— training and experience acquired through appropriate education, general scientific knowledge, and sound professional judgment (see ANSI/ASB Std 022).

5 Requirements

5.1 Details about the operation of BLAST can be found in Madden (2013), and detailed information on the terms in the BLAST output can be found in the NCBI *Field Guide Glossary*. The minimum requirements and recommendations in Section 5 address criteria for the preparation and submission of evidentiary query sequences (5.2) and evaluation and interpretation of BLAST results from GenBank® (5.3, 5.4), which should take into account whether the returned hit(s) is attributed to the correct species and whether the hit(s) is a close enough match for the taxon in question. Section 5.5 addresses appropriateness of reference data and level of assignment, and 5.6 addresses reporting of results (5.6).

5.2 Prior to performing a BLAST search, an evidentiary query sequence:

a) shall be prepared by removing non-template flanking regions (e.g., primers);

b) shall meet quality criteria as defined by the laboratory (see ANSI/ASB Standard 048)

c) shall, if from a protein coding region, be examined to ensure it does not contain premature stop codons (e.g., by translation).

5.3 To ensure that a hit(s) on which conclusions are based are of high quality, an initial assessment of the BLAST results:

a) shall ensure the hit(s) belongs to the expected broader taxonomic group (e.g., macerated plant tissue returns matches to sequences from the plant kingdom, not the bacterial kingdom);

NOTE In situations involving a complete unknown, it may not be possible to complete this assessment.

b) shall ensure that any hit(s) that is an anomaly among the returned results is not used; this would be indicated by being the only representative of its species interleaved among many in a different taxonomic group; this could be an indication of human error in sequence labeling during sequence preparation prior to GenBank® upload;

c) shall ensure the hit(s) does not originate from an environmental sample (e.g., bulk soil extraction, bacterial swab) or low copy sample;

NOTE The original publication can often be consulted to determine the source of the sequence. In some instances, this determination may not be possible.

d) shall include a review for descriptors or characteristics that indicate the sequence was not reviewed prior to uploading in GenBank®;

NOTE Sequences that have not been reviewed for quality may include descriptors such as “NGS”, “MPS”, “EST”, “shotgun”, “library”, and “WGS”; these may have been batch uploaded directly from the sequencing platform. Unedited sequences may also have a higher number of “Ns” or degenerate bases at the ends, or contain non-template flanking (e.g., primer, adapter) sequences.

e) shall include a review for ambiguous bases;

NOTE Treat ambiguous bases with caution, as they can indicate poor-quality sequence, but they can also indicate heteroplasmic sites within a high-quality sequence.

f) shall ensure the hit(s) from a protein coding region does not contain premature stop codons.

5.4 Any hit(s) on which conclusions are based shall be evaluated to determine if the returned sequence is attributed to the correct species based on the criteria listed below. This section is to determine if returned sequences are appropriate for interpretations as outlined in 5.5. These criteria confer either strong or moderate support to the attribution

5.4.1 Strong criteria (not all of these criteria have to be met, see 5.5 for more information about how to evaluate relevant criteria).

a) Sequence(s) is derived from a voucher specimen that bears a unique identifier and is accessioned in a curated collection.

NOTE Use of the word “voucher” in the GenBank® description is not sufficient to confirm that the sequence is derived from a morphologically confirmed voucher specimen. “Voucher” designation may be confirmed by a review of associated publications or institutional databases (e.g., a museum database).

b) Sequence(s), when downloaded, aligned with sequences from closely related species and used to construct a phylogram, results in a species-level topology concordant with expectations from the peer-reviewed literature.

c) Sequence(s) is from a study published in a peer-reviewed journal and,

1) the study addresses the phylogeny or taxonomy of the taxon of interest and,

2) the publication or accompanying metadata makes it clear that the source specimen(s) was morphologically identified by a taxonomic expert.

d) Sequence(s) is part of a population genetic study for the given species published in a peer-reviewed journal.

NOTE Typically a population genetic study characterizes numerous individuals from the studied species in order to explore intraspecific variation (sample sizes will vary based on genetic variability and rareness of the species in question; published studies will have sample sizes that capture the relevant genetic diversity of the taxon in question). The individuals may either be from the same geographic region, or from distinct populations within the known distributional range.

5.4.2 Moderate criteria (not all of these criteria have to be met, see 5.5 for more information about how to evaluate relevant criteria).

a) Sequence(s) is from a study published in a peer-reviewed journal; the study includes additional data establishing species identity (e.g., morphological evidence, museum specimen), but it is not clear that the source specimen was a voucher (5.4.1a) or was morphologically identified by a taxonomic expert (5.4.1c).

b) Sequence(s) is from a phylogenetic study in a peer-reviewed journal; the study addresses phylogeny or taxonomy of the taxon of interest and:

1) includes most or all members of the genus in question, and

2) the locus shows resolution at the species level (see 5.5.2).

c) Sequence(s) is one of multiple identical or near-identical sequences for the same locus and species from different submitters or geographic sample collection locations .

d) Sequence(s) is not from a peer-reviewed study on the taxon of interest, but is accompanied by additional metadata concerning the source individual (e.g., location life history stage, name of collector, name of taxonomic expert who rendered the source individual's identification).

5.5 Sections 5.5.1 and 5.5.2 should be evaluated to determine the appropriate level for taxonomic assignment. This section is to assess whether locus selection and taxonomic representation is appropriate, and if the taxa in question are well-separated.

5.5.1 Determine whether all likely candidate species in the taxonomic group in question are represented amongst the returned hit(s); if relevant taxa are missing, use alternate or additional loci or additional reference material to provide further support for the comparison.

NOTE 1 Complete taxon sampling is ideal, but often not feasible. Species that are distantly related based on published phylogenies or those that do not occur in the geographic area of interest may be exempted from the comparison if sequences are not available. See section 4.5.2 in ANSI/ASB 019 and section 3.5 in ANSI/ASB Std 029.

NOTE 2 Peer-reviewed literature or internal validation for the species/marker of interest provides the foundation for evaluating whether hits are appropriate and comprehensive enough to provide accurate interpretation for reporting.

5.5.2 Determine whether the interspecific distance for the taxonomic group of interest at the surveyed locus is greater than intraspecific distance.

NOTE If interspecific and intraspecific distances are similar, consider using a different locus or additional loci, or limiting identification to a higher taxonomic level.

5.6 Requirements for reporting from BLAST results are in 5.6.1 through 5.6.2.

5.6.1 The analyst may report to the species level when all of the following criteria are met:

a) the evidentiary sequence(s) has been prepared as outlined in 5.2;

b) the hit(s) on which conclusions are to be based:

   1) meets the quality criteria as defined in 5.3;

   2) meets at least two strong support criteria (as defined in 5.4.1), or at least one strong and one moderate (as defined in 5.4.2) support criteria;

   3) has been evaluated against the criteria defined in 5.5;

   4) and when aligned to the evidentiary query sequence, shows 99% to 100% sequence identity (inclusive).

NOTE 1 99% is a conservative threshold, to be applied in instances where no other information is available for the target taxon. For most species, intraspecific distance will be greater than 1%; in cases where additional information (e.g., other loci, taxonomies based on morphological features) indicates species are well-separated, identities lower than 99% may still warrant a species level identification.

NOTE 2 By default, BLAST results are sorted by E-value, which gives more weight to matches with higher query coverage. This can result in shorter sequences with higher percent sequence identity being displayed after longer sequences with lower percent sequence identity. The list may be sorted by the identity value to reveal the highest-similarity matches. It is critical to consider both the percent identity and the length of the match when evaluating BLAST results.

5.6.2 It is appropriate to report to a higher taxonomic level when all of these criteria are met:

a) the evidentiary sequence(s) has been prepared as outlined in 5.2;

b) the hit(s) meets the quality criteria as defined in 5.3;

c) the hit(s) has been evaluated against the criteria defined in 5.5;

d) the hit(s) does not meet the support criteria given in 5.6.1 b) 2), but is from a peer-reviewed publication and:

   1) the most similar sequences returned by a query are <99% identical and there is little definitive information on interspecific distance;

OR

2) all top hits represent a single taxonomic level (i.e., genus, family, order), but there is a discrepancy at a lower taxonomic level (e.g., hits represent different species, but they all belong to a single genus).

Annex A
(informative)

Empirical Assessment of the Performance of ANSI/ASB Standard 180

A number of published studies (e.g., Nilsson et al. 2006, Ross et al. 2008, Collins and Cruickshank 2012, Sonet et al. 2013, Seah et al. 2017, Kolter and Gemeinholzer 2021) have demonstrated the utility and accuracy of taxonomic assignment of unknown specimens using genetic data, including via GenBank® BLAST. Ross et al. (2008), in particular, compared BLAST, distance- and tree-based methods and explored the scenarios under which each are most likely to succeed or fail. BLAST methods where the identification of an unknown is based on the top BLAST hits were the most successful of the tested methods. Comparisons are most difficult when there is incomplete taxon sampling in the reference dataset (addressed in 5.5.1), and when species are very closely related and/or the DNA region used lacks sufficient resolution [5.4.2 b) 2) and 5.5.2]. Another difficulty is that sequences in the GenBank® dataset can be erroneously labeled [addressed in 5.3 a) through 5.3 c)], or not of sufficient quality [5.3 c) through 5.3 f)].

Though there are many publications exploring the accuracy of BLAST identifications for various taxonomic groups, those studies do not determine how this standard would work in forensic practice. The accuracy of results using a portion of this standard were recently assessed in a study by Patel et al. (2023; NIST award 70NANB21H128). The study showed that 100% of identifications to genus were correct, while 98.3% of assignments to species were correct. Given the choice to report to species or to a higher taxonomic level (e.g., genus; see 5.5 and 5.6), analysts were much more likely to be conservative (e.g., reporting to genus when they could have gone to species) than to overreach (e.g., report to species when they should have gone to genus or higher). However, in a few instances, analysts determined that sequences met the criteria to report to higher taxonomic level (5.6.2) but reported to species anyway. In the few instances where overreach happened, the analysts assigned the species correctly. The high accuracy and correctness in this study suggest that this standard is appropriate for accurate taxonomic assignment of unknown sequences using GenBank® BLAST.

It is worthwhile to scrutinize the circumstances under which the ~2% of incorrect taxonomic assignments at the species level occurred. In the Patel et al. study, participants were given raw Sanger sequence data to identify to species via GenBank® BLAST. Because participants did not have access to the original tissues or DNA extracts, they could only follow part of the proposed standard. In an actual casework scenario, if BLAST returned ambiguous results or revealed very poor taxon sampling, analysts could choose to sequence alternate loci and/or augment their BLAST results with in-house databases (as suggested in 5.5.1 and 5.5.2). These options were not available to study participants, as the locus was chosen for them. The one incorrect species assignment was due to a fish sample, which was sequenced at a locus with both poor taxon sampling and poor resolution, and thus if encountered in casework, should have been re-amplified and sequenced at an appropriate locus (see 5.5.1). Study participants also could not evaluate the sequences in the context of others in the same run, to assess how the controls performed and whether all sequences were of similar quality. Participants also lacked relevant case information, such as likely geographic origin of the samples, which is often an important part of a species' description. Additionally, while best attempts were made to align taxa with laboratory expertise, some of the taxa laboratories were asked to assign would not normally occur in their jurisdictions.

The high accuracy and correctness of identifications in this study suggest that ANSI/ASB Standard 180 is appropriate for accurate taxonomic assignment of unknown sequences using GenBank® BLAST. This standard will perform even better in practice when analysts can see their sequences in the context of the rest of the run, choose informative loci for the taxa in question, and have access to case-relevant information. Future studies conducted using case-type samples that are analyzed from DNA extraction through data interpretation will provide a more complete picture of error in forensic taxonomic assignments.

Annex B
(informative)

Bibliography

This is not meant to be an all-inclusive list as the group recognizes other publications on this subject may exist. At the time this standard was drafted, these were the publications available for reference. Additionally, any mention of a particular software tool or vendor as part of this bibliography is purely incidental, and any inclusion does not imply endorsement.

1] Altschul SF. (2014). "BLAST Algorithm." In: eLS, John Wiley & Sons, Ltd (Ed.). doi: 10.1002/9780470015902.a0005253.pub2.

2] ANSI/ASB Standard 019, *Wildlife Forensics General Standards*, First Edition, 2019[b].

3] ANSI/ASB Standard 022, *Standard for Forensic DNA Analysis Training Programs*, First Edition, 2019[b].

4] ANSI/ASB Standard 029, *Report Writing in Wildlife Forensics: Morphology and Genetics*, First Edition, 2019[b].

5] ANSI/ASB Standard 048, *Wildlife Forensic DNA Standard Procedures*, First Edition, 2019[b].

6] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2013). "GenBank." *Nucleic Acids Research* 41(D1):D36-42[c].

7] BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008[d].

8] Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 16, *Molecular Phylogenetics*[e].

9] International Organization for Standardization. (2017). ISO/IEC 17025:2005 *General Requirements for the Competence of Testing and Calibration Laboratories*. 28 pp.

10] Lee TRC, Anderson SJ, Tran-Nguyen LTT, Sallam N, Le Ru BP, Conlong D, Powell K, Ward A, Mitchell A. 2019. "Towards a global DNA barcode reference library for quarantine identifications of lepidopteran stemborers, with an emphasis on sugarcane pests." *Scientific Reports* 9: 7039[f].

11] Lorenz JG, Jackson WE, Beck JC, Hanner R. (2005). "The problems and promise of DNA barcodes for species diagnosis of primate biomaterials." *Philosophical Transactions of the Royal Society* B 360, 1869–1877.

[b] Available from: https://www.aafs.org/academy-standards-board
[c] Available from: https://www.ncbi.nlm.nih.gov/genbank/
[d] Available from: https://www.ncbi.nlm.nih.gov/books/NBK279690/
[e] Available from: https://www.ncbi.nlm.nih.gov/books/NBK21122/
[f] Available from: Doi: https://doi.org/10.1038/s41598-019-42995-0

12] Madden T. (2013). "The BLAST Sequence Analysis Tool." In: *The NCBI Handbook*, 2nd ed. Bethesda, MD[g].

13] NCBI Field Guide Glossary[h].

References for Annex A:

14] Collins, R. A., and R. H. Cruickshank. 2012. "The Seven Deadly Sins of DNA Barcoding." *Molecular Ecology Resources* 13 (6): 969–75. https://doi.org/10.1111/1755-0998.12046.

15] Crocetta F, P. Mariottini, D. Salvi, & M. Oliverio. "Does GenBank provide a reliable DNA barcode reference to identify small alien oysters invading the Mediterranean Sea?" *J. Mar. Biol. Assoc.* U.K. 2015;95: 111-122.

16] Kolter A, B. Gemeinholzer, "Plant DNA barcoding necessitates marker-specific efforts to establish more comprehensive reference databases." *Genome.* 2021 Mar;64(3):265-298.

17] Nilsson R.H., M. Ryberg, E. Kristiansson, K. Abarenkov, KH. Larsson, U. Koljalg. "Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective." *PLoS One.* 2006;1: e59.

18] Patel, Scheible, Meiklejohn (2023). "Interlaboratory study to assess the practical utility of OSAC proposed standard 2021-S-0006: Standard for the Use of GenBank for Taxonomic Assignment of Wildlife." *Forensic Science International: Animals and Environments*. 2023 Dec; 4:100067. https://doi.org/10.1016/j.fsiae.2023.100067

19] Ross, H.A., S. Murugan, and W.L.S. Li. "Testing the Reliability of Genetic Methods of Species Identification via Simulation." *Systematic Biology.* 2008. 57 (2): 216–30.[i]

20] Seah Y.G., A.F. Ariffin, & T.N.A.M. Jaafar. "Levels of COI divergence in Family Leiognathidae using sequences available in GenBank and BOLD Systems: A review on the accuracy of public databases." *Aquac. Aquar. Conserv. Legis. Int. J. Bioflux Soc*. 2017;10: 391-401.

21] Sonet G, K. Jordaens, Y. Braet, L. Bourguignon, E. Dupont, T. Backeljau, et al. "Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France." *ZooKeys*. 2013;365: 307-328.

[g] Available from: https://www.ncbi.nlm.nih.gov/books/NBK153387/
[h] Available from: https://www.ncbi.nlm.nih.gov/Class/FieldGuide/glossary.html#
[i] Available from: https://doi.org/10.1080/10635150802032990.

Academy Standards Board
410 North 21st Street
Colorado Springs, CO 80904

www.aafs.org/academy-standards-board