GenBank is a reliable resource for 21st century biodiversity research

Matthieu Leraya, Nancy Knowltonb,1, Shian-Lei Ho, Bryan N. Nguyenb,de, and Ryuji J. Machidac,1

aSmithsonian Tropical Research Institute, Smithsonian Institution, Panama City, 0842-03092, Republic of Panama; bNational Museum of Natural History, Smithsonian Institution, Washington, DC 20560; cBiodiversity Research Centre, Academia Sinica, 115-29 Taipei, Taiwan; dDepartment of Biological Sciences, The George Washington University, Washington, DC 20052; eComputational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052

Contributed by Nancy Knowlton, September 15, 2019 (sent for review July 10, 2019; reviewed by Ann Bucklin and Simon Creer)

Traditional methods of characterizing biodiversity are increasingly being supplemented and replaced by approaches based on DNA sequencing alone. These approaches commonly involve extraction and high-throughput sequencing of bulk samples from biologically complex communities or samples of environmental DNA (eDNA). In such cases, vouchers for individual organisms are rarely obtained, often unidentifiable, or unavailable. Thus, identifying these sequences typically relies on comparisons with sequences from genetic databases, particularly GenBank. While concerns have been raised about biases and inaccuracies in laboratory and analytical methods, comparatively little attention has been paid to the taxonomic reliability of GenBank itself. Here we analyze the metazoan mitochondrial sequences of GenBank using a combination of distance-based clustering and phylogenetic analysis. Because of their comparatively rapid evolutionary rates and consequent high taxonomic resolution, mitochondrial sequences represent an invaluable resource for the detection of the many small and often undescribed organisms that represent the bulk of animal diversity. We show that metazoan identifications in GenBank are surprisingly accurate, even at low taxonomic levels (likely <1% error rate at the genus level). This stands in contrast to previously voiced concerns based on limited analyses of particular groups and the fact that individual researchers currently submit annotated sequences to GenBank without significant external taxonomic validation. Our encouraging results suggest that the rapid uptake of DNA-based approaches is supported by a bioinformatic infrastructure capable of assessing both the losses to biodiversity caused by global change and the effectiveness of conservation efforts aimed at slowing or reversing these losses.

environmental DNA | metabarcoding | taxonomic assignments

Human activities have resulted in deteriorating environmental conditions worldwide, prompting urgent calls to increase our ability to assess biodiversity across space and through time (1, 2). Identifying organisms for these assessments typically relies on observations or, for small organisms, reference to voucher specimens housed in collections. Traditional morphological identifications of collected materials are increasingly being replaced by DNA-based identifications. This change has been accelerated by the adoption of DNA barcoding (3) because of its accuracy, speed, and now lower costs. However, for very small and abundant organisms (the vast majority of diversity on the planet), even traditional barcoding can be impractical. Thus, sequencing of bulk DNA samples (from collections of organisms) or environmental DNA (eDNA; DNA shed from organisms into the environment) via metabarcoding (the characterization of multiple DNA sequences from a single sample) is increasingly being used to characterize biodiversity in terrestrial, freshwater, and marine habitats (4, 5). Materials studied include bulk soil (6), sediment, and benthic aquatic samples (7); bulk samples from organismal traps (e.g., Malaise traps for insects) (8); gut contents (9); plankton tow (9); filtered water (10); and ancient DNA (11, 12).

Identifying animal (metazoan) sequences obtained from these environmental samples typically relies on comparisons with GenBank (13), the largest repository of genetic data for biodiversity (14, 15). In many cases, no vouchers are available to independently confirm identification, because the organisms are tiny, very difficult or impossible to identify, or lacking entirely (in the case of eDNA). While concerns have been raised about biases and inaccuracies in laboratory and analytical methods used in metabarcoding (16–18), comparatively little attention has been paid to the taxonomic reliability of GenBank itself, whose >4.7 million mitochondrial gene sequences representing >170,000 metazoan species have never been comprehensively assessed for taxonomic accuracy.

Mitochondrial sequences represent an invaluable resource because of their comparatively rapid evolutionary rates and consequent high taxonomic resolution (19). The majority of metazoan mitochondrial genomes contain 37 genes, including 2 coding for ribosomal RNA, 13 coding for proteins, and 22 tRNAs (20). The tRNAs generally are very short (approximately 60 to 70 bp) and thus not useful as taxonomic markers. However, the other 15 genes are generally much longer than 100 bp, making them potentially useful markers. For the many small and often undescribed organisms

Significance

As loss of biodiversity and ecosystem degradation become major concerns worldwide, scientists increasingly depend on DNA-based characterization of animal communities for monitoring and impact assessments. These analyses ultimately depend on the taxonomic reliability of genetic databases for taxonomic assignments. Concerns have been raised about the reliability of GenBank, the largest and most widely used genetic database. We show that, contrary to expectations, the proportion of mislabeled sequences in GenBank is surprisingly low. Major taxonomic errors are vanishingly small (0.01% at the class level, 0.05% at the order level), and likely <1% even at the genus level. These results show that GenBank is much more reliable for a range of applications, including studies of environmental change, than previously thought.


Reviewers: A.B., University of Connecticut; and S.C., Bangor University.

The authors declare no competing interest.

This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

Data deposition: Raw data files (FASTA files: https://figshare.com/articles/Fasta_files/7642202), data analysis scripts (https://figshare.com/articles/R_Script_for_estimation_of_minimum_and_maximum_error_/7856180), clustering output files (https://figshare.com/articles/MultipleTaxaClusteringPhylumClassOrder/7853567), and a table summarizing the estimated minimum and maximum number of mislabeled sequences per phylum and per gene (https://figshare.com/articles/Summarized_Table/7856156) are available in Figshare.

1To whom correspondence may be addressed. Email: knowlton@si.edu or ryujimachida@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1911714116/-/DCSupplemental.

First published October 21, 2019.

www.pnas.org/cgi/doi/10.1073/pnas.1911714116

PNAS | November 5, 2019 | vol. 116 | no. 45 | 22651–22656
that represent the bulk of animal diversity (21), they represent a cost-effective way of broadly characterizing individuals, populations, species, and communities.

Incorrect taxonomic annotations of DNA sequence data can be caused by inadvertent amplification of laboratory contaminants (22), intimately associated organisms including bacteria (23, 24), or nontarget genes, such as pseudogenes (25); by incorrect identification of the study organisms (26); or by mistakes made during data entry at various stages. These annotation errors could be pervasive and highly problematic, but there have been only limited attempts to quantify their magnitude. We analyzed the scale, patterns, and causes of metazoan mitochondrial sequence mislabeling in the GenBank BLAST NT database at the genus level and above to address previous critiques of GenBank that have focused on major taxonomic errors. We used VSEARCH (27) to group closely related metazoan sequences of 13 protein-coding and 2 ribosomal RNA-coding mitochondrial genes (*SI Appendix, Table S1*) at 97%, 98%, 99%, and 100% thresholds. Clusters containing sequences belonging to different phyla, classes, orders, families, or genera were flagged for further investigation. This approach assumes that clusters of highly similar sequences (97% or greater) for these hypervariable genes should typically contain conspecifics but not sequences of multiple genera, families, orders, classes, and phyla (unless they are mislabeled). We did not attempt to identify errors at the species level because to do so would require case-by-case assessments of far more clusters coupled with detailed taxonomic evaluations, since multispecies clusters could often be the result of insufficient taxonomic resolution of the genes or taxonomic disagreements and revisions.

**Results and Discussion**

A total of 4,714,864 gene sequences downloaded from GenBank yielded 279,899, 304,804, 354,463, and 440,800 multisquence clusters (groups containing more than 1 sequence) at 97%, 98%, 99% and 100% clustering thresholds, respectively (*SI Appendix, Table S1*) provides the breakdown per gene at the 97% threshold). The remaining 332,834, 385,985, 511,172, and 1,547,404 “clusters,” respectively, had just a solitary sequence, which by definition could not be used to test for labeling errors since the test depends on having more than 1 higher taxon in a cluster. As expected, sequences in multisquence clusters represented a larger proportion of all sequences at lower similarity thresholds (Fig. 1). Cytochrome oxidase subunit I (CO1) and Cytochrome b apoenzyme (Cytb), the 2 most common mitochondrial gene sequences in GenBank, with 54.5% and 10.5% of all mitochondrial sequences, respectively, also had somewhat higher proportions of sequences in multisquence clusters (e.g., >90% at a 97% threshold; Fig. 1 and *SI Appendix, Fig. S1*), probably because with a greater depth of sampling, sequences that do not form clusters are less likely. Thus, we highlight analyses based on the 97% threshold for these 2 genes, as these estimates are the most inclusive of GenBank’s data and hence likely to be the most reliable. We also present analyses of the small (12S) and large (16S) ribosomal RNA genes, as they are being increasingly targeted in community DNA studies that heavily rely on GenBank for taxonomic annotations (4). The 97% threshold represents a level of sequence similarity typical of intraspecific variation for mitochondrial genes (28); it also provides a reasonable upper bound on the estimate of mislabeled sequences because the resulting larger clusters are more likely to contain genuinely related higher taxa (e.g., congeners) that nevertheless will be flagged as misidentified by virtue of belonging to a single cluster.

To estimate the number of sequences that were incorrectly annotated, we examined all clusters containing multiple phyla, classes, and orders individually and used phylogenetic analyses to determine where the errors occurred. Whenever mislabeled sequences at the order, class, and phylum levels could not be identified unequivocally with phylogenetic analysis, we calculated the minimum and maximum possible number of misannotated sequences, assuming that at least 1 annotation was correct. For the much more numerous clusters containing multiple families or genera, we used the same methods as were used for higher taxonomic ranks for all the clusters with 100 or more sequences (i.e., 375 clusters accounting for 133,598 sequences); for clusters with fewer than 100 sequences, we only estimated the minimum and maximum numbers of mislabeled sequences. We manually examined clusters larger than 100 sequences because our analyses showed that checking these clusters (which represented only 2.0% of all clusters but contained 46.8% of the total sequences) was an efficient way of improving the precision of our estimate of the possible range of error rates (*SI Appendix, Fig. S2*).

The percentage of sequences with incorrect assignments increased with decreasing taxonomic rank but was very low for most taxonomic levels (Table 1 and Fig. 2). Only 0.01% of sequences were incorrectly annotated at or above the level of the class, 0.05% were incorrectly annotated at or above the level of order, and 0.17% to 0.95% were incorrectly annotated at or above the level of family. Error rates rose only at the level of genus and above, with the total for all genes ranging from 0.73% (minimum estimate) to 3.47% (maximum estimate). For CO1 and Cytb, the minimum and maximum estimates ranged from 0.58% to 3.30% and from 0.72% to 5.03%, respectively. A maximum estimate is likely to be an overestimate because it assumes that the taxon with the fewest sequences in a cluster is the one that is correctly identified (*SI Appendix*), whereas the correct annotation was the most common taxon in 95% of the clusters examined (*SI Appendix, Fig. S3*). Other limitations inherent to distance-based clustering approaches with fixed thresholds may also inflate our estimates. Moreover, some groups, such as cichlid fishes, elephants, dolphins, and bovines, as well as the higher-level taxa Anthozoa and Porifera (*Fig. 3*), are known to have genera that cannot be differentiated using mitochondrial gene sequences alone due to recent divergences (29) or slow rates of molecular
Table 1. Estimated numbers of metazoan mitochondrial encoded sequences in GenBank with incorrect taxonomic labels at the phylum, class, order, family, and genus levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>375</td>
<td>375</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Class</td>
<td>537</td>
<td>566</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Order</td>
<td>1,610</td>
<td>2,377</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Family</td>
<td>7,500</td>
<td>41,677</td>
<td>0.17</td>
<td>0.95</td>
</tr>
<tr>
<td>Family*</td>
<td>5,919</td>
<td>34,071</td>
<td>0.14</td>
<td>0.78</td>
</tr>
<tr>
<td>Genus</td>
<td>32,062</td>
<td>152,157</td>
<td>0.73</td>
<td>3.47</td>
</tr>
<tr>
<td>Genus*</td>
<td>29,198</td>
<td>140,212</td>
<td>0.67</td>
<td>3.22</td>
</tr>
</tbody>
</table>

Proportions are calculated based on the total number of sequences in nonsolitary clusters at the 97% clustering threshold.

*Removing sequences of Porifera and Cnidaria (0.6% and 0.09% of nonsolitary clusters at the 97% clustering threshold).

The difficulty of assessing the accuracy of singleton sequences also highlights the importance of replication within taxa, particularly in barcoding initiatives, to allow cross-validation.

To determine the likely sources of taxonomic errors and hence potential remedies, we individually examined all multilocus clusters that contained multiple phyla, classes, or orders (SI Appendix, Table S2 and Fig. S5). For 70.8% of the cases, there was no obvious explanation. For the remainder, the likely source of error was laboratory contaminants (16.3%), data entry errors (8.4%, potentially due to autofill functions of sequence submission platforms), contamination by associated organisms in the sample (3.2%), and pseudogenes (1.1%) (SI Appendix, Table S2). As has been noted previously (33), many of these errors could be detected before submission by performing a BLAST search. For known pseudogenes, errors occur when the organelle is specified as “mitochondrion” because the sequence is most likely in the nuclear genome.

We also examined sources of error at lower taxonomic levels (families and genera), focusing on clusters containing 100 or more sequences. Interestingly, 55.0% of the sequences had the incorrect label because of taxonomic revisions, meaning that they were correct at the time of entry. Other assignment errors at these lower levels were caused by laboratory contamination (4.5%) and data entry errors (2.8%).

Concerns about errors in GenBank have prompted cautionary notes (35–37) and the creation of curated sequence databases for particular taxonomic groups and genes, including BOLD (32), PR2 (38), SILVA (39), PHYMYCO-DB (40), and MIDORI (41). However, our results stand in contrast to most previous estimates of GenBank annotation error rates. For example, Bridge et al. (42) reported that 12 of 51 species of the renowned, highly poisonous fungus Amanita had incorrect genus names and that 16 of 100 fungi of the order Helotiales were misidentified at the genus level; in most of these cases, the annotations were to distantly related fungi or nonfungi. A recent study estimated that 6.9% of all sequences listed as Gasterotricha (43) did not belong to this phylum (our estimate is 2.5%). Similarly, some researchers have warned that the number of misidentified parasite sequences may be increasing (26), although there are no quantitative estimates. There are several possible reasons for the discrepancy between earlier estimates and our results. Specifically, fungi, parasites, and other small metazoans are more taxonomically challenging and more difficult to sample without contamination. Error rates may also have decreased in recent years, with increasing attention to quality control (e.g., GenBank; https://www.ncbi.nlm.nih.gov/books/NBK44940/).

DNA-based biosurveys are rapidly increasing in popularity (4, 44), with the annual number of papers based on these methods having risen >21-fold between 2000 and 2018 (Fig. 4). Our results suggest that while quality control efforts remain essential, metazoan sequences in GenBank are highly reliable for a range of applications, ranging from underpinning ecological inferences to assessing environmental policies.

Materials and Methods

The sequence datasets for 13 protein-coding mitochondrial genes—ATP synthase subunit 6 (A6) and 8 (A8); Cytochrome oxidase subunits I (CO1), II (CO2), and III (CO3); Cytochrome b apoenzyme (Cytb); NADH dehydrogenase subunits 1 to 4 (ND1 to ND4), 4 L (ND4L), 5 (ND5), and 6 (ND6)—and 2 ribosomal RNA-coding mitochondrial genes—large (16S) and small (12S) rRNA—were prepared using BLAST with reference datasets.
The category “Other phyla” includes sequences of Acanthocephala, Brachiopoda, Bryozoa, Chaetognatha, Ctenophora, Cycliophora, Entoprocta, Gastrotricha, Hemichordata, Kinorhyncha, Nematomorpha, Nemertea, Onychophora, Placozoa, Priapulida, Rhombozoa, Rotifera, Tardigrada, and Xenacoelomorpha.

The following BLAST parameters were used for ribosomal RNA-coding genes: blastn -num_alignments 100 -word_size 11 -outfmt 7 -dust blastn -db RefSeq_prepared_db -query metazoan_mitochondrial_nucleotide.fasta

The following BLAST parameters were used for protein-coding genes: blastp -num_alignments 100 -word_size 3 -outfmt 7 -seg blastp -db RefSeq_prepared_db -query metazoan_mitochondrial_amino_acid.fasta

Based on the BLAST result, we separated the dataset by genes (total of 15). In addition, length restriction was performed for the datasets: 12S, 200 to 2,000; 16S, 100 to 2,500; A6, 100 to 1,000; A8, 100 to 500; CO1, 100 to 2,000; CO2, 100 to 1,500; CO3, 100 to 1,300; Cytb, 100 to 1,500; ND1, 50 to 1,200; ND2, 150 to 1,500; ND3, 100 to 600; ND4, 150 to 2,000; ND4L, 100 to 700; NDS, 150 to 2,000; ND6, 150 to 1,500. This length restriction procedure was required to avoid random clustering of short sequences in the subsequent procedure.

To identify taxonomically mislabeled sequences, clustering analyses were performed at 97%, 98%, 99%, and 100% similarity thresholds with VSEARCH (-sortbylength,-cluster_fast). The vast majority of sequences were located in multisequence clusters at 97% for all genes (from 86.2% for 12S to 95.6% for CO1; SI Appendix, Fig. S5). After clustering, we checked incongruences of taxonomic labels within each cluster at all taxonomic levels except species (i.e., phylum, class, order, family, and genus) using custom Perl scripts. Clusters with sequences labeled as different phyla, classes, orders, families, or genera were extracted for downstream analysis.

We attempted to identify which were the mislabeled sequences in all clusters containing multiple phyla, classes, and orders, but we examined only a portion of the much more numerous clusters with multiple families and genera. We focused on clusters containing more than 100 sequences, because our estimates based on order-level misidentifications showed that manually checking larger clusters was a relatively efficient way to narrow the possible range of error rates. Clusters with more than 100 sequences represented only 2.0% of all clusters but contained 46.8% of the total sequences. By simulating random sampling of the order-level data at different depths, we found that randomly selecting sequences and then checking the corresponding clusters was an appropriate, because this method tended to preferentially sample low-error large clusters, resulting in a poor approximation of the true error rate (SI Appendix, Fig. S5). We attempted to identify which were the mislabeled sequences in all clusters containing multiple phyla, classes, and orders, but we examined only a portion of the much more numerous clusters with multiple families and genera. We focused on clusters containing more than 100 sequences, because our estimates based on order-level misidentifications showed that manually checking larger clusters was a relatively efficient way to narrow the possible range of error rates. Clusters with more than 100 sequences represented only 2.0% of all clusters but contained 46.8% of the total sequences. By simulating random sampling of the order-level data at different depths, we found that randomly selecting sequences and then checking the corresponding clusters was an appropriate, because this method tended to preferentially sample low-error large clusters, resulting in a poor approximation of the true error rate (SI Appendix, Fig. S5).
A similarity search using the BLAST server (33) (blastn with “low-complexity region filter” and “mask for lookup table only” functions disabled) was performed using the putative mislabeled sequence as a query, and the distance tree function on the BLAST server was used to examine phylogenetic relationships with 100 close matches to the query. If the taxonomy of the outgroup sequences disagreed with the query taxonomy, we concluded that the query had the wrong taxonomy. Mislabeled sequences at the phylum, class, and order levels detected in this study will be reported to GenBank for re-identification (identification flowchart in SI Appendix, Fig. S5). First, we classified sequences as mislabeled because of “data entry error” if the accession number and genus name had consecutive or nearly consecutive numbers or if it had a very similar genus name. Next, we identified whether they were common laboratory contaminants, such as DNA from humans, common rodents, laboratory model organisms, common human food, mosquitoes, or pets. We then checked whether the sequence was a potential nuclear mitochondrial pseudogene using 2 criteria: a mention of pseudogenes in the definition or project title of the GenBank flat file and a drastic change in sequence similarity between different regions of the sequence. Then the sequence was classified as a contamination from a host or dietary item if the project title of the GenBank flat file mentioned gut content analysis or parasite study. Next, we checked whether the BLAST result indicated high similarity to bacterial sequences. Finally, we searched online databases (e.g., FishBase, WORMS) to check whether a taxonomic revision took place in the group. If none of these categories applied, we classified the sequence as another laboratory contamination or taxonomic misidentification. We did not attempt to identify the cause of sequence misidentification in phyla Porifera and Cnidaria clusters because of the lower taxonomic resolution in these groups for mitochondrial genes (30, 31).

Raw data files (FASTA files: https://figshare.com/articles/Fasta_files/7642202), data analysis scripts (https://figshare.com/articles/R_Script_for_estimation_of_minimum_and_maximum_error/7856180), clustering output files (https://figshare.com/articles/MultipleTaxaClustersFamilyGenus/7853561; https://figshare.com/articles/MultipleTaxaClusteringPhylumClassOrder/7853567), and a table summarizing the estimated minimum and maximum numbers of mislabeled sequences per phylum and per gene (https://figshare.com/articles/Summarized_Table/7856156) are available at FigShare (https://figshare.com).

ACKNOWLEDGMENTS. We acknowledge help received from members of the Machida Laboratory at the Biodiversity Research Centre, Academia Sinica. This work is supported by the Ministry of Science and Technology of Taiwan (Grant 105-2621-B-001-003, to R.J.M.), Academia Sinica (R.J.M.), the Smithsonian Training and Research Institute (M.L.) and the Sant Chair for Marine Science of the Smithsonian National Museum of Natural History (N.K.).