

Note

Algae 2020, 35(3): 293-301

<https://doi.org/10.4490/algae.2020.35.9.4>

Open Access



An assessment of the taxonomic reliability of DNA barcode sequences in publicly available databases

Soyeong Jin¹, Kwang Young Kim², Min-Seok Kim³ and Chungoo Park^{1,*}

¹School of Biological Sciences and Technology, Chonnam National University, Gwangju 61186, Korea

²Department of Oceanography, Chonnam National University, Gwangju 61186, Korea

³Dental Science Research Institute, School of Dentistry, Chonnam National University, Gwangju 61186, Korea

The applications of DNA barcoding have a wide range of uses, such as in taxonomic studies to help elucidate cryptic species and phylogenetic relationships and analyzing environmental samples for biodiversity monitoring and conservation assessments of species. After obtaining the DNA barcode sequences, sequence similarity-based homology analysis is commonly used. This means that the obtained barcode sequences are compared to the DNA barcode reference databases. This bioinformatic analysis necessarily implies that the overall quantity and quality of the reference databases must be stringently monitored to not have an adverse impact on the accuracy of species identification. With the development of next-generation sequencing techniques, a noticeably large number of DNA barcode sequences have been produced and are stored in online databases, but their degree of validity, accuracy, and reliability have not been extensively investigated. In this study, we investigated the extent to which the amount and types of erroneous barcode sequences were deposited in publicly accessible databases. Over 4.1 million sequences were investigated in three large-scale DNA barcode databases (NCBI GenBank, Barcode of Life Data System [BOLD], and Protist Ribosomal Reference database [PR2]) for four major DNA barcodes (cytochrome c oxidase subunit 1 [*COI*], internal transcribed spacer [*ITS*], ribulose biphosphate carboxylase large chain [*rbcL*], and 18S ribosomal RNA [*18S rRNA*]); approximately 2% of erroneous barcode sequences were found and their taxonomic distributions were uneven. Consequently, our present findings provide compelling evidence of data quality problems along with insufficient and unreliable annotation of taxonomic data in DNA barcode databases. Therefore, we suggest that if ambiguous taxa are presented during barcoding analysis, further validation with other DNA barcode loci or morphological characters should be mandated.

Key Words: 18S rRNA; COI; DNA barcoding; ITS; rbcL; taxonomic databases

Abbreviations: 16S rRNA, 16S ribosomal RNA; 18S rRNA, 18S ribosomal RNA; BOLD, Barcode of Life Data System; BSTI, barcode sequences with their respective species-level taxonomic identifiers; COI, cytochrome c oxidase subunit 1; EBS, erroneous barcode sequences; iBOL, the International Barcode of Life database; ITS, internal transcribed spacer; NT, NCBI non-redundant nucleotide sequence database; PR2, the Protist Ribosomal Reference database; rbcL, ribulose biphosphate carboxylase large chain



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received July 29, 2020, Accepted September 4, 2020

*Corresponding Author

E-mail: chungoo@jnu.ac.kr

Tel: +82-62-530-1913, Fax: +82-62-530-2199

INTRODUCTION

Accurate and reliable taxonomic identification is a major cornerstone of evolutionary biology and critical for understanding the diversity of biological life. With the lack of taxonomic expertise, several limitations of extant research, such as phenotypic plasticity, genetic variability, and morphologically cryptic taxa, hinder precise morphological taxonomic identification. Recent technological advances in molecular biology have allowed the development of rapid, robust, and sensitive diagnostic methods for species identification that use standardized DNA regions known as DNA barcodes. Since the inception of DNA barcoding in 2003 (Hebert et al. 2003), over 9,800 peer-reviewed scientific articles containing the terms “DNA barcode” or “DNA barcoding” have been hitherto published. These studies range from taxonomic studies that elucidate cryptic species and phylogenetic relationships to analyses of environmental samples (e.g., soil, marine sediments, and seawater) that include biodiversity monitoring and conservation planning.

Since first proposed by Hebert et al. (2003), the mitochondrial gene encoding cytochrome c oxidase subunit 1 (*COI*) has been widely used in the identification of species in many groups of animals including birds (Kerr et al. 2007), amphibians (Smith et al. 2008), spiders (Barrett and Hebert 2005), and butterflies (Burns et al. 2008), and several early papers carried out proof-of-concept studies for the utility of the *COI* barcoding region. Despite the potential power of DNA barcode, several conceptual and methodological limitations still exist regarding the absence of a generally accepted single universal DNA barcode for all organisms (Kress et al. 2015) and DNA amplification bias (Jo et al. 2019).

Once a dataset of DNA barcode sequences is generated from an unidentified specimen, the most common approaches for species discovery and identification are the use of the sequence similarity-based methods, including BLAST search and phylogenetic analysis. To this end, the obtained barcode sequences are first compared to the sequences in the DNA barcode reference databases, such as NCBI GenBank (Sayers et al. 2019), the Barcode of Life Data System (BOLD) (Ratnasingham and Hebert 2007), the Protist Ribosomal Reference database (PR2) (Guillou et al. 2013), and the UNITE database (Koljalg et al. 2005). This implies that the quantity and quality of the barcode data within these databases must be stringently monitored to prevent an adverse impact on species identification accuracy. With the development of high-throughput next-generation sequencing techniques, a noticeably

large number of DNA barcode sequences have been produced and stored in online databases, but their degree of validity, accuracy, and reliability have not been extensively and thoroughly investigated (Kim et al. 2019). For example, Bridge et al. (2003) re-evaluated only 206 published DNA barcode sequences for Fungi and revealed that up to 20% of sequences appeared to be misidentified, dubious, or chimeric. Similar validation studies were carried out restrictedly in each bacterial (4,138 16S ribosomal RNA [*16S rRNA*] sequences) (Ashelford et al. 2005), fungal (51,354 internal transcribed spacer [*ITS*] sequences) (Nilsson et al. 2006), dipteran (85 *COI* sequences) (Sonet et al. 2013), and ponyfish (232 *COI* sequences) (Seah et al. 2017) community.

Thus, in this study, we investigated the amount and types of erroneous barcode sequences (EBS) deposited in publicly accessible databases that are used by molecular taxonomists and geneticists. More than 4.1 million DNA barcode sequences in three large-scale DNA barcode storage databases (NCBI GenBank, BOLD, and PR2) were investigated for four major DNA barcodes (*COI*, *ITS*, ribulose biphosphate carboxylase large chain [*rbcL*], and *18S rRNA*). It was found that approximately 2% of sequences were detected as EBS and, intriguingly, their taxonomic distributions were uneven.

MATERIALS AND METHODS

We used the four most commonly used eukaryotic DNA barcodes including a mitochondrial gene (*COI*), a chloroplast gene (*rbcL*), and nuclear ribosomal regions (*ITS* and *18S rRNA*) (Kress et al. 2015). For generating libraries for each barcode sequence, we collected all the sequences that have any given keywords (listed in Supplementary Table S1) in the annotation section of the sequence database record or gene annotation text field. For all four barcode sequences, we used the NCBI non-redundant nucleotide sequence database (NT) that has the most comprehensive set of sequences (approximately 49 million non-redundant sequences and >185 billion base pairs) collected from myriad organisms from all kingdoms. Specifically, for the *COI* and *rbcL* barcodes, we further generated corresponding barcode libraries collected from the International Barcode of Life (iBOL) database (<http://www.ibol.org>) (Ratnasingham and Hebert 2007), representing the largest biodiversity genomics initiative to date. For the *18S rRNA* barcode, the PR2 database (<https://github.com/vaulot/pr2database>) (Guillou et al. 2013) that currently (version 4.11.1) comprises approxi-

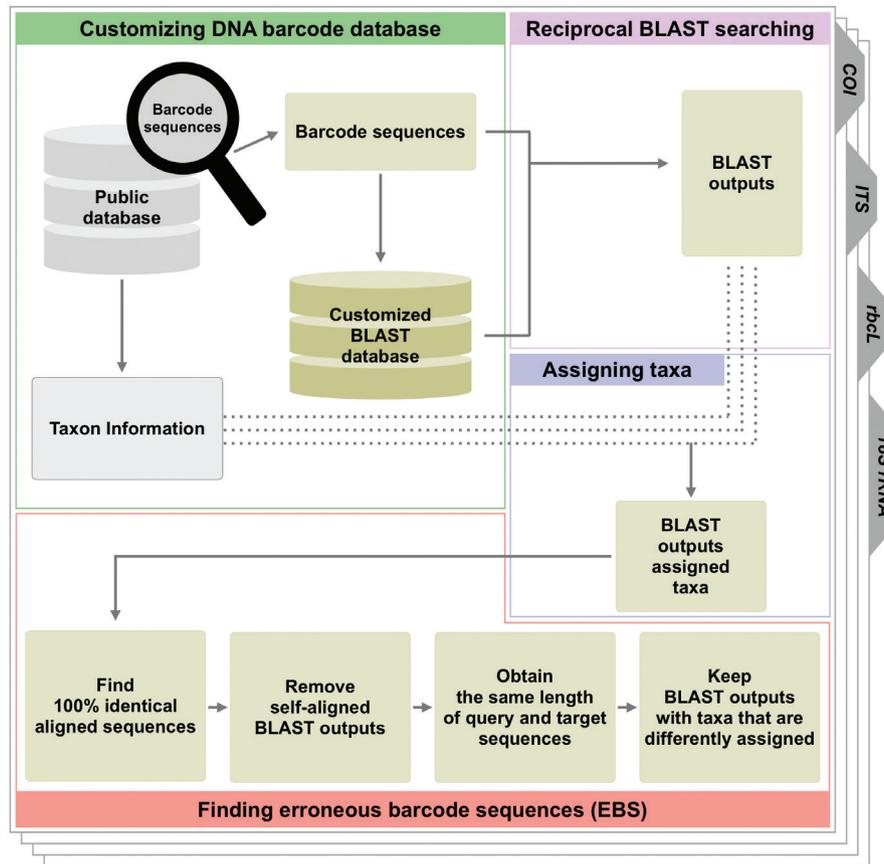


Fig. 1. Bioinformatic workflow for identifying erroneous barcode sequences (EBS) deposited in publicly available databases. Our EBS workflow system consists of four main components: (1) customizing barcode sequence database, (2) reciprocal BLAST searching, (3) assigning taxa, and (4) finding EBS. A solid arrow indicates the next step in the procedure. Three dotted lines represent the linkage between BLAST outputs and their taxonomic information. These analyses are independently repeated for each of the four barcode sequences.

mately 180,000 ribosomal RNA and DNA sequences and represents most eukaryotic phyla was used.

To identify the barcode sequences that are completely identical but with different taxonomic identifiers, hereafter referred to simply as “erroneous barcode sequences” (EBS), we performed the following procedures, illustrated in Fig. 1. Briefly, we first created a BLAST reference database using the *COI* barcode library and the makeblastdb application from NCBI-BLAST+ (v2.3.0) (Camacho et al. 2009). Next, BLAST was used for each query sequence in the *COI* barcode library against the target reference database using BLASTN with default parameters. Because the best BLAST hit would usually correspond to the query sequence itself, we further filtered the BLAST output and identified the EBS if the second best hit corresponding to 100% sequence identify (query sequence coverage 100%) had a taxonomic identifier different from that of the query sequence. We repeated these procedures for the remaining barcode libraries (*rbcL*, *ITS*, and *18S* from

NT, *COI* and *rbcL* from iBOL, and *18S rRNA* from PR2).

RESULTS AND DISCUSSION

From the NCBI NT database, four major DNA barcode sequences including *COI*, *ITS*, *rbcL*, and *18S rRNA* were semi-automatically collected using keyword-based search. A total of 834,252 species were identified representing 66,535 genera, 7,289 families, 1,238 orders, 246 classes, and 62 phyla. Specifically, 585,968 species comprising 59.6% of the total barcode sequences were distinguished by the *COI* barcode and grouped based on phylum, class, and order, such that Arthropoda (78.68%) and Chordata (11.37%) were the major phyla (Supplementary Table S2); Insecta (66.80%) and Arachnida (5.80%) were the major classes (Supplementary Table S3); and Diptera (29.69%), Lepidoptera (11.57%), and Hymenoptera (10.50%) were the major orders (Supplementary Table

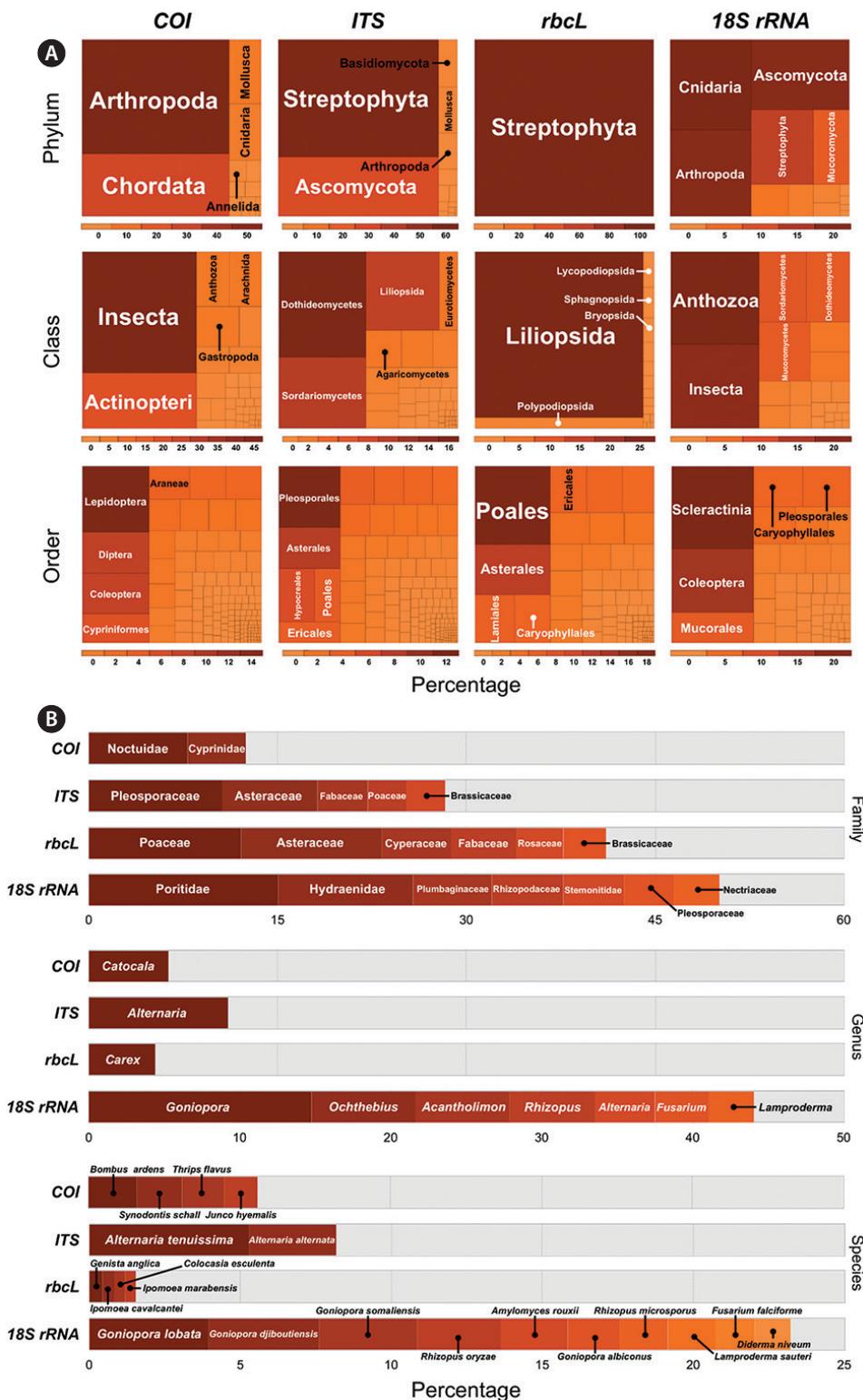


Fig. 2. Percentage of erroneous barcode sequences (EBS) at each taxon level. (A) Treemaps show the percentage of relative EBS abundance in the top 5 taxa at the phylum, class, and order levels. The larger and darker the rectangle, the higher the EBS counts in the corresponding taxon. (B) The EBS abundance data for taxa at the family, genus, and species levels are visualized by bar charts. All taxa with >3% EBS at the family and genus levels are shown, and species with >1% EBS for cytochrome c oxidase subunit 1 (COI), internal transcribed spacer (ITS), and 18S ribosomal RNA (18S rRNA) and >0.3% EBS for rbcL are shown. Taxa with an unknown status are excluded. For these analyses, the EBS search is limited to the NCBI non-redundant nucleotide sequence database (NCBI NT) database.

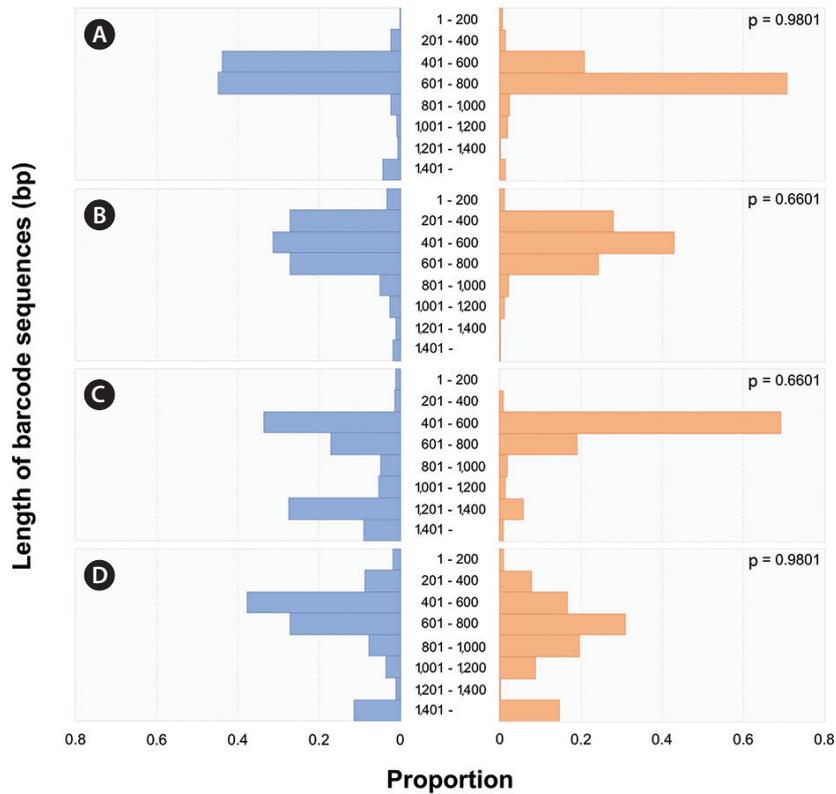


Fig. 3. Length comparison between total barcode sequences and erroneous barcode sequences (EBS). Blue (left) and orange (right) bars represent the proportion of the length of the total barcode sequences and EBS for each bin, respectively. (A) to (D) represent data for the cytochrome c oxidase subunit 1 (*COI*), internal transcribed spacer (*ITS*), ribulose biphosphate carboxylase large chain (*rbcL*), and 18S ribosomal RNA (*18S rRNA*) DNA barcodes, respectively. The lengths of barcode sequences are binned in intervals of 200 bp. The p-value is calculated using the Kolmogorov-Smirnov test. For these analyses, the NCBI non-redundant nucleotide sequence database (NCBI NT) database is used.

S4). For the barcode *ITS* (32.3% of total barcode sequences), 220,527 species from 54 phyla were identified, and Ascomycota (23.90%) and Streptophyta (21.04%) were the major phyla (Supplementary Table S5). As expected, the top 4 classes and top 5 orders belonged to Fungi (Supplementary Tables S6 & S7). For each of the other two barcodes, less than 5% of the total barcode sequences were identified (Supplementary Tables S8-10 for *rbcL* and Supplementary Tables S11-13 for *18S rRNA*).

To demonstrate how many EBS were typically used for barcode-based species identification, all barcode sequences were compared and aligned with each other (see the Materials and Methods section and Fig. 1 for detail). From the approximately 2 million barcode sequences with their respective species-level taxonomic identifiers (hereafter named BSTI), we revealed that approximately 2% were EBS (Table 1). Upon a close examination of the four barcodes (Fig. 2), 6,289 EBS were found in the *COI*

Table 1. Summary of species identification based on barcode sequences for each species in the NCBI non-redundant nucleotide sequence database

Barcode	No. of barcode sequences (%)	BSTI	EBS	Clade	Organelle
<i>COI</i>	2,261,665 (59.6)	1,254,703	6,289	Animals	Mitochondrion
<i>ITS</i>	1,228,044 (32.3)	638,861	12,266	Fungi	Nuclei
<i>rbcL</i>	143,517 (3.8)	120,466	13,184	Plants	Chloroplast
<i>18S rRNA</i>	164,496 (4.3)	56,093	1,262	Eukaryotes	Nuclei
Total	3,797,722	2,070,123	33,001	-	-

BSTI, No. of barcode sequences with species-level taxonomic identifier; EBS, No. of erroneous barcode sequences; *COI*, cytochrome c oxidase subunit 1; *ITS*, internal transcribed spacer; *rbcL*, ribulose biphosphate carboxylase large chain; *18S rRNA*, 18S ribosomal RNA.

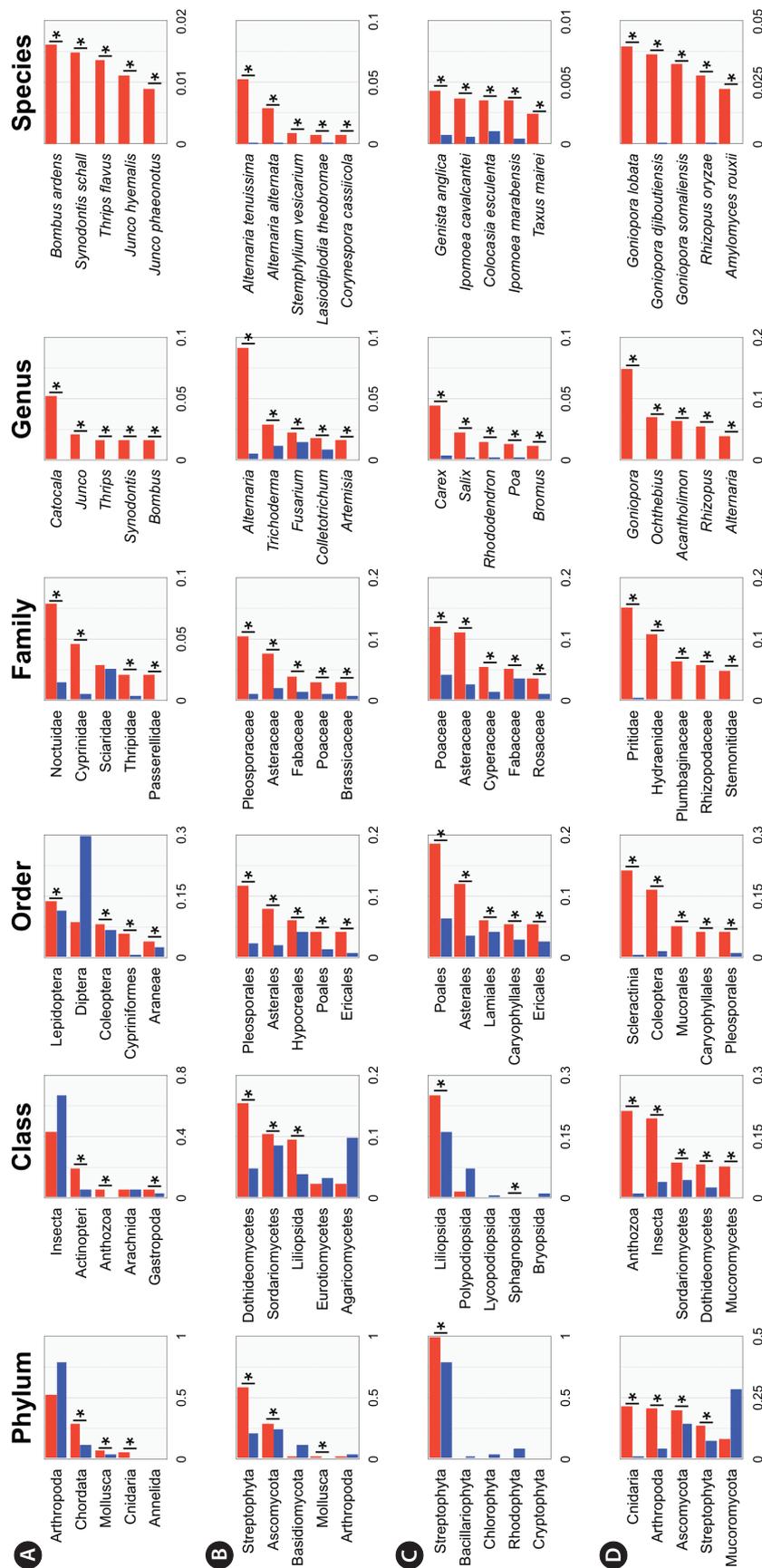


Fig. 4. Significantly enriched erroneous barcode sequences (EBS) regardless of ascertainment bias. Red and blue bars indicate the proportion of EBS and non-EBS in the top 5 taxa from the phylum to the species levels. (A) to (D) represent data for the cytochrome c oxidase subunit 1 (COI), internal transcribed spacer 1 (ITS), ribulose biphosphosphate carboxylase large chain (rbcL), and 18S ribosomal RNA (18S rRNA) DNA barcodes, respectively. Asterisk indicates significant enrichment ($p < 2.2e-16$) of EBS versus non-EBS. The p-value is calculated using the chi-square test. For these analyses, the NCBI non-redundant nucleotide sequence database (NCBI NT) database is used.

barcode database, which represents 19% of the total EBS and 0.5% (6,289 out of 1,254,703) of the corresponding *COI* BSTI. The EBS were most dominant in the phylum Arthropoda (52.35%), class Insecta (43.11%), order Lepidoptera (13.66%), family Noctuidae (7.84%), and genus *Catocala* (5.23%). When classified at the species level, *Bombus ardens* (101 EBS), *Synodontis schall* (94 EBS), *Thrips flavus* (86 EBS), and *Junco hyemalis* (71 EBS) had more than 1% of the *COI* EBS. From the *ITS* barcode database, 12,266 sequences were detected as EBS, and their major taxonomic ranks were mostly from the fungal groups. At the species level, *Alternaria tenuissima* (5.29% of *ITS* EBS) and *Alternaria alternata* (2.88% of *ITS* EBS) were the major species containing EBS. From the *rbcL* barcode database, 13,184 EBS representing 40% of the total EBS and more than 10% of the corresponding *rbcL* BSTI were found. Despite the large numbers of EBS, no dominant species (>1% of *rbcL* EBS) were observed, but the genus *Carex* had more than 3% of *rbcL* EBS. Lastly, we found 1,262 EBS in the *18S rRNA* barcode database. Despite the low numbers of EBS, 10 major species (>1% of *18S rRNA* EBS) and 7 dominant genera (>3% of *18S rRNA* EBS) were observed.

To check whether our findings were not confounded by multiple biases, we performed two sensitivity analyses. First, the length difference between the total barcode sequences and EBS was measured, and no significant differences in length distributions were observed ($p > 0.05$ with the Kolmogorov-Smirnov test) (Fig. 3). Next, to clarify whether our findings were not confounded due to the ascertainment bias, namely uneven taxonomic distribution of EBS, the number of EBS and non-EBS were compared at each taxonomic level from species

to phylum for the four barcode sequences. Except for the Diptera order and Sciaridae family in the *COI* EBS, all top five taxonomic ranks from species to order in all four barcode sequences had significantly enriched EBS ($p < 2.2e-16$ with the chi-square test), suggesting partial ascertainment bias. These biases were evident in some taxonomic groups at the phylum and class levels (Fig. 4). Consequently, our finding of the presence of considerable number of EBS is fairly robust and unequivocal.

To increase the available evidence, we extended our findings to two well-curated DNA barcode databases and observed that there were still considerable numbers of EBS in the iBOL and PR2 databases (Table 2).

CONCLUSION

In this study, we identified the EBS that are completely identical but have different taxonomic identifiers and examined the amount and types of EBS that were deposited in publicly accessible databases. A considerable number of EBS were sparsely unequally dispersed throughout major taxa. Surprisingly, EBS were discovered even in the highly curated iBOL and PR2 databases. Because of the incompleteness and inaccuracy of existing DNA barcode databases, molecular taxonomists must exert caution and careful judgment when identifying species, especially when using only DNA barcode sequence data. If ambiguous species identification occurs during DNA barcoding analysis, we advise performing further evaluation with other DNA barcode loci or morphological characters. Finally, we encourage geneticists and molecular taxonomists to reliably generate authoritative DNA barcode libraries, and report or correct any mistakes or errors detected when working with DNA barcode databases.

Table 2. Summary of species identification based on barcode sequences for each species in the highly curated iBOL and PR2 databases

Database	Barcode sequence	No. of barcode sequences	BSTI	EBS
iBOL ^a	<i>COI</i>	163,325	31,763	61
	<i>rbcL</i>	1,523	845	211
PR2 ^b	<i>18S rRNA</i>	176,813	86,643	4,285

iBOL, the International Barcode of Life database; PR2, the Protist Ribosomal Reference database; BSTI, No. of barcode sequences with species-level taxonomic identifier; EBS, No. of erroneous barcode sequences; *COI*, cytochrome c oxidase subunit 1; *rbcL*, ribulose biphosphate carboxylase large chain; *18S rRNA*, 18S ribosomal RNA.

^aThe version of iBOL database is iBOL_phase_6.50 including a total of 165,237 sequences.

^bThe version of PR2 database is v4.11.1 containing all 176,813 sequences.

ACKNOWLEDGEMENTS

We thank the members of the CSB lab and the anonymous reviewers for their valuable comments. This research was supported by the “Research center for fishery resource management based on the information and communication technology” (ICT to C.P) of the Korea Institute of Marine Science and Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Korea, and the National Research Foundation (NRF) of Korea grant funded by the Korea government (MSIT) (NRF-2020R1A2C3005053 to K.Y.K and NRF-2017R1A2B1007928 to M.S.K).

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Keywords to collect each barcode sequence in the NCBI nucleotide database (<https://e-algae.org/>).

Supplementary Table S2. Number of *COI* sequences of each phylum (<https://e-algae.org/>).

Supplementary Table S3. Number of *COI* sequences of each class (<https://e-algae.org/>).

Supplementary Table S4. Number of *COI* sequences of each order (<https://e-algae.org/>).

Supplementary Table S5. Number of *ITS* sequences of each phylum (<https://e-algae.org/>).

Supplementary Table S6. Number of *ITS* sequences of each class (<https://e-algae.org/>).

Supplementary Table S7. Number of *ITS* sequences of each order (<https://e-algae.org/>).

Supplementary Table S8. Number of *rbcL* sequences of each phylum (<https://e-algae.org/>).

Supplementary Table S9. Number of *rbcL* sequences of each class (<https://e-algae.org/>).

Supplementary Table S10. Number of *rbcL* sequences of each order (<https://e-algae.org/>).

Supplementary Table S11. Number of *18S rRNA* sequences of each phylum (<https://e-algae.org/>).

Supplementary Table S12. Number of *18S rRNA* sequences of each class (<https://e-algae.org/>).

Supplementary Table S13. Number of *18S rRNA* sequences of each order (<https://e-algae.org/>).

REFERENCES

- Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71:7724-7736.
- Barrett, R. D. H. & Hebert, P. D. N. 2005. Identifying spiders through DNA barcodes. *Can. J. Zool.* 83:481-491.
- Bridge, P. D., Roberts, P. J., Spooner, B. M. & Panchal, G. 2003. On the unreliability of published DNA sequences. *New Phytol.* 160:43-48.
- Burns, J. M., Janzen, D. H., Hajibabaei, M., Hallwachs, W. & Hebert, P. D. 2008. DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservación Guanacaste, Costa Rica. *Proc. Natl. Acad. Sci. U. S. A.* 105:6350-6355.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A. -L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P. & Christen, R. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41(Database issue):D597-D604.
- Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270:313-321.
- Jo, J., Lee, H. -G., Kim, K. Y. & Park, C. 2019. SoEM: a novel PCR-free biodiversity assessment method based on small-organelles enriched metagenomics. *Algae* 34:57-70.
- Kerr, K. C. R., Stoeckle, M. Y., Dove, C. J., Weigt, L. A., Francis, C. M. & Hebert, P. D. N. 2007. Comprehensive DNA barcode coverage of North American birds. *Mol. Ecol. Notes* 7:535-543.
- Kim, H. M., Jo, J., Park, C., Choi, B. -J., Lee, H. -G. & Kim, K. Y. 2019. Epibionts associated with floating *Sargassum horneri* in the Korea Strait. *Algae* 34:303-313.
- Kõljalg, U., Larsson, K. -H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjølner, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Vrålstad, T. & Ursing, B. M. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol.* 166:1063-1068.
- Kress, W. J., García-Robledo, C., Uriarte, M. & Erickson, D. L. 2015. DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.* 30:25-35.
- Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. -H. & Kõljalg, U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE* 1:e59.
- Ratnasingham, S. & Hebert, P. D. N. 2007. BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* 7:355-364.
- Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D. & Karsch-Mizrachi, I. 2019. GenBank. *Nucleic Acids Res.* 47:D94-D99.
- Seah, Y. G., Ariffin, A. F. & Jaafar, T. N. A. M. 2017. Levels of COI divergence in Family Leiodnathidae using sequences available in GenBank and BOLD systems: a review on the accuracy of public databases. *AAAC Bioflux* 10:391-

401.

Smith, M. A., Poyarkov, N. A. Jr. & Hebert, P. D. N. 2008. DNA BARCODING: CO1 DNA barcoding amphibians: take the chance, meet the challenge. *Mol. Ecol. Resour.* 8:235-246.

Sonet, G., Jordaens, K., Braet, Y., Bourguignon, L., Dupont, E., Backeljau, T., De Meyer, M. & Desmyter, S. 2013. Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. *Zookeys* 365:307-328.