Title: Maximizing the reliability and the number of species assignments in metabarcoding studies

Running title: Maximizing reliability in species assignments

Authors :

Audrey Bourret[a,1*], Claude Nozères[b], Eric Parent[a], Geneviève J. Parent[a]

a Laboratory of genomics, Maurice Lamontagne Institute, Fisheries and Oceans Canada, Mont-Joli, QC, G5H 3Z4, Canada

b Maurice Lamontagne Institute, Fisheries and Oceans Canada, Mont-Joli, QC, G5H 3Z4, Canada

Email addresses:

1. audrey.bourret@dfo-mpo.gc.ca*
2. claude.nozeres@dfo-mpo.gc.ca
3. eric.parent@dfo-mpo.gc.ca
4. genevieve.parent@dfo-mpo.gc.ca

*Correspondence author

Abstract

The use of environmental DNA (eDNA) for biodiversity assessments has increased rapidly over the last decade. However, the reliability of taxonomic assignments in metabarcoding studies is variable, and affected by the reference databases and the assignment methods used. Species level assignments are usually considered as reliable using regional libraries but unreliable using public repositories. In this study, we aimed to test this assumption for metazoan species detected in the Gulf of St. Lawrence, in the Northwest Atlantic. We first created a regional library with COI barcode sequences including a reliability ranking system for species assignments. We then estimated the accuracy of the public repository NCBI-nt for species assignments using sequences from the regional library, and contrasted assigned species and their reliability using NCBI-nt or the regional library with a metabarcoding dataset and popular assignment methods. With NCBI-nt and sequences from the regional library, Blast-LCA was the most accurate method for species assignments but the proportions of accurate species assignments were higher with Blast-TopHit (>80 % overall taxa, between 70 and 90 % amongst taxonomic groups). With the metabarcoding dataset, the reliability of species assignments was greater using the GSL-rl compared to NCBI-nt. However, we also observed that the total number of reliable species assignments could be maximized using both GSL-rl and NCBI-nt, and their optimal assignment methods, which differed. The use of a two-step approach in species assignments, using a regional library and a public repository, could improve the reliability and the number of detected species in metabarcoding studies.

Keywords: Genbank, reference library, COI, marine species

Introduction

The use of environmental DNA (eDNA) for biodiversity assessments and monitoring has increased rapidly over the last decade given the high potential of this non-intrusive approach to uncover biodiversity with limited effort (Taberlet et al. 2012, Makiola et al. 2020). eDNA metabarcoding surveys collect and detect traces of a diversity of organisms in various types of environmental samples using high-throughput sequencing or PCR-based approaches (Taberlet et al. 2012, Yu et al. 2012). Surveys of eDNA generally involve a series of steps such as sample collection, extraction, targeted amplification, high-throughput sequencing, and bioinformatic processing, which includes taxonomic assignments to reference sequences from a public repository or a regional library (Deiner et al. 2017). Only a small fraction of detected eDNA sequences in environmental samples can currently be assigned to a species-level identity owing to a lack of data and taxonomic resolution in publicly available resources (Deiner et al. 2017, Leite et al. 2021, Zafeiropoulos et al. 2021). The reliability and precision of taxonomic assignments is affected by the quality and availability of sequences in repositories and the assignment methods, thereby limiting confidence in the use of eDNA for biodiversity monitoring and targeted species detections (Coissac et al. 2012, McGee et al. 2019, Meiklejohn et al. 2019, Gold et al. 2021, Hleap et al. 2021).

Several public repositories exist and can be used as reference databases to provide taxonomic assignments in metabarcoding studies. The public National Center for Biotechnology Information Nucleotide database (NCBI-nt, including the well-known GenBank database) is the largest sequence repository and is widely used in eDNA metabarcoding studies (Porter and Hajibabaei 2018b, 2020). However, the presence of mislabeled specimens, the large variation in quality of sequences available, and gaps in species coverage (i.e., unrepresented species) result in erroneous species identification when directly comparing unknown sequences to NBCI-nt (Bidartondo 2008, Mioduchowska et al. 2018, Leray et

al. 2019). The Barcode of Life Data Systems (BOLD) is another sequence repository specific to the most common barcode regions, including the cytochrome c oxidase I (COI) gene which is the widely used gene region for animal DNA barcoding (Ratnasingham and Hebert 2007, Porter and Hajibabaei 2018b). BOLD displays mandatory (e.g., institution storing voucher specimen, sampling country) and optional (e.g., sampling location, specimen photos) metadata, performs groupings of similar sequences into Barcode Index Number (BIN), and permits editing or updating of records, all of which assists with data quality control. However, like GenBank, it is also vulnerable to submissions of misidentified specimens (McCusker et al. 2013, Oliveira et al. 2016, Fontes et al. 2021, Radulovici et al. 2021). As reliable taxonomic assignments at the species-level are expected under many regulatory contexts (e.g., environmental status assessment, monitoring of invasive species or species at risks; Aylagas et al. 2014, Hering et al. 2018, Bush et al. 2019, Piper et al. 2019), some metabarcoding studies have questioned the value of using public repositories (e.g., von Ammon et al. 2018, Locatelli et al. 2020, Gold et al. 2021). Characterizing the proportion of accurate species assignments using NCBI-nt would be highly valuable to understand the extent of uncertainty in species eDNA detection and consequently, enable an accurate interpretation of a metabarcoding study's results.

Alternatively, curated regional libraries have been shown to reduce errors in species assignments (Gold et al. 2021). Regional libraries are limited to species expected in predefined areas, and can be created by data mining and curating existing sequences from public repositories and/or from generating sequences from specimens. They have the advantage to limit spurious assignments to related but non-local species, and to reveal gaps (i.e., missing sequences) in taxonomic groups (Weigand et al. 2019, Ramirez et al. 2020, Jazdzewska et al. 2021). Examples of regional libraries are widely available in the northern hemisphere for multiple taxonomic groups (e.g., Knebelsberger et al. 2014, Hänfling et al. 2016, Stoeckle

et al. 2017, Fraija-Fernández et al. 2020, Gold et al. 2021, Van Den Bulcke et al. 2021). Some of these reference libraries present ranking systems to ensure high taxonomic reliability (e.g., Costa et al. 2012, Knebelsberger et al. 2014). Ranking systems are often provided to target future barcoding efforts and improvements in reference sequences. No explicit ranking system about the uncertainty of species assignment has yet been presented within metabarcoding studies. Such a system would be highly valuable to provide clear indications on the reliability of species assignments for eDNA end-users.

Another source of variability in species assignments are the bioinformatics software and pipelines used in metabarcoding studies. Recently, studies have started to evaluate the accuracy of taxonomic assignments using various bioinformatic methods (O'Rourke et al. 2020, Hleap et al. 2021, Mathon et al. 2021). These studies compared taxonomic assignment methods that are based on strategies such as alignment, composition, or modelling (Richardson et al. 2017, see also four strategies in Hleap et al. 2021). The Basic Local Alignment Search Tool (BLAST) is an alignment-based approach extensively used in metabarcoding studies that relies on a nearest-neighbor approach to return best hits between unknown sequences and records from a reference database (Camacho et al. 2009). The taxonomic identity of the unknown sequence may be inferred in conjunction with a least common ancestor (LCA) or a Top Hit approach with identity threshold, usually between 95 and 99%. These thresholds should reflect expected inter-species divergence, but high variation among taxonomic groups may cause pitfalls in assignments (Wang et al. 2007, Alberdi et al. 2018). Classifiers using machine-learning algorithms are from composition-based approaches that have shown good performance in some contexts of species assignments (Richardson et al. 2017, Murali et al. 2018, Porter and Hajibabaei 2018a). They can take advantage of phylogeny between reference sequences and thus are less affected by shared divergence between groups. Classifiers are trained on a reference library, and pre-trained classifiers are increasingly

106 available (e.g., Porter and Hajibabaei 2018a). However, recent benchmarking studies have shown lower
107 performance of classifiers compared to BLAST (O'Rourke et al. 2020, Hleap et al. 2021, Mathon et al.
108 2021).

109 This study aimed to estimate the accuracy of species assignments using NCBI-nt and to contrast the
110 reliability of using NCBI-nt or a regional library on a metabarcoding dataset with popular assignment
111 methods (Fig. 1). To achieve these objectives, we first created a curated regional library (GSL-rl) using
112 publicly available sequences from BOLD for COI barcode locus of metazoans from the Gulf of St.
113 Lawrence. The regional library presents a reliability ranking system for species assignments based on
114 sequence availability and similarity that can be understood by any eDNA end-users, scientists or not. We
115 then used sequences from GSL-rl to estimate the accuracy of NCBI-nt. We also compared the assigned
116 species in a metabarcoding dataset using NCBI-nt or GSL-rl and their reliability. We reached the
117 conclusion that using a two-step approach, i.e., species assignments with a regional library and a public
118 repository, is desirable to maximize the reliability and the number of species assignments in
119 metabarcoding studies.

120

Methods

Creation of a regional library for the Gulf of St. Lawrence (GSL-rl) with a reliability ranking system

The creation of a curated regional library for the Gulf of St. Lawrence was composed of three major steps: 1) obtaining a list of marine faunal species from both decision-makers and available regional taxonomical information (Nozères 2017), 2) creating an initial regional library using bioinformatics tools with BOLD and rounds of revisions based on quality and similarity of sequences, and 3) enhancing the draft regional library using a metabarcoding dataset (hereafter GSL regional library: GSL-rl; Fig. S1). All sequences retained in the GSL-rl had names at the genus or species level and were already published on BOLD. More details about the creation of the GSL-rl are provided in the supplementary material.

We created the GSL-rl to identify molecular operational taxonomic units (MOTUs) at the species level. Each species in the GSL-rl was ranked based on sequence availability and similarity (Table S1). Species with reference sequences for itself and closely related species (i.e., from the same genus) acknowledged to be present in the Gulf of St. Lawrence were ranked as “Reliable” if they did not share BOLD’s barcode index number (BIN; i.e., a unique identifier of sequences based on genetic distance, Ratnasingham and Hebert 2013). Species with reference sequences for itself, but not for all congeners acknowledged to be present in the Gulf of St. Lawrence (i.e., other species of the same genus) were ranked as “Unreliable due to gaps”. Species with reference sequences sharing BIN with other species were ranked as “Unreliable due to BIN sharing”. Common causes of BIN sharing are genetic similarities between species or specimen misidentification. For the GSL-rl, curation and validation process done during its creation should limit the BIN sharing due to specimen misidentification. Taxonomic assignments belonging to one of the two “Unreliable” categories should be interpreted with caution, and preferably not at the species-level.

Evaluating the accuracy of species assignments using the public NCBI-nt repository

We used the curated sequences from the GSL-rl to estimate three performance parameters using NCBI-nt: 1) the proportion of assignments, 2) the proportion of accurate assignments, and 3) the accuracy.

Taxonomic assignments were performed using NCBI-nt (downloaded 2020-10-23) and the BlastX tool *blastn* (v2.10.1, Camacho et al. 2009) combined with the least common ancestor (LCA; hereafter Blast-LCA) or the Top Hit methods (hereafter, Blast-TopHit) at three identity thresholds (95, 97 and 99%) from an in-house R script. The LCA method assigns the higher taxonomic rank shared by all hits above the identity threshold while the Top Hit method only used the hits with the highest probability. We excluded hits containing “environmental sample”, “uncultured” or “predicted” in their description.

We assessed the proportion of assignments at the species level, and we considered that an assignment was accurate if it matched the species identity associated in the GSL-rl. We measured accuracy as the proportion of accurate assignments over all assignments at the species level.

Contrasting species assignments using the regional library or the public NBCI-nt repository, and popular assignment methods

We compared the detection results from an eDNA metabarcoding dataset using GSL-rl and NCBI-nt and three assignment methods (Fig. 1). The eDNA metabarcoding dataset was obtained from the analysis of water samples collected from scientific surveys in 2018 in coastal areas of the GSL, both at the surface and bottom of the water column (see supplementary material for details on the field, laboratory and bioinformatics works underlying the eDNA metabarcoding dataset). The three assignment methods were Blast-LCA, Blast-TopHit and the IDtaxa (Murali et al. 2018). Blast assignment methods were used as

described in the previous section with both GSL-rl and NCBI-nt. NCBI-nt Blast results were filtered to retain only metazoan detections and remove non-marine taxa (i.e., *Homo sapiens*, Arachnida, Insecta). IDtaxa is a classifier implemented within the DECIPHER R package (Wright 2016), and was trained only with the GSL-rl. IDtaxa classifier was selected since it would be less prone to "over classification", i.e., classification to an erroneous group when the real group is absent from the training set, compared to the popular Ribosomal Database Project (RDP) classifier (Murali et al. 2018). Taxonomic assignments with IDtaxa were obtained at three confidence thresholds (i.e., weighted fraction of bootstrap replicates assigned to a given taxa) representing moderate confidence (40%), high confidence (50%), and very high confidence (60%) in species assignments (Murali et al. 2018).

We contrasted results obtained using GSL-rl and NCBI-nt with distinct ranking systems. Species detected with the GSL-rl were classified according to the three categories of the reliability ranking system previously created: "Reliable", "Unreliable due to gaps", "Unreliable due to BIN sharing" (Fig. 1B, Table S2). For species assignments with NCBI-nt, we used geographic and habitat filters to classify them as "Likely" if they were part of the Gulf of St. Lawrence checklist (Nozères 2017) or present in the areas based on the World Register of Marine Species (WoRMS, WoRMS Editorial Board 2020), and "Unlikely" if not (Fig. 1A). Such filters are often applied in metabarcoding studies but the source of information for the likeliness of a species to be present is often obscure.

Data and R scripts availability

All data and R scripts used for the creation of the GSL-rl are provided on Github:

https://github.com/GenomicsMLI-DFO/GSL_COI_ref_library

Results

A COI regional library with a reliability ranking system for metazoans from the Gulf of St. Lawrence (GSL-rl)

The first version of the GSL-rl comprised 1304 sequences covering 439 species (158 species of vertebrates from phylum Chordata from 68 families; 281 species of invertebrates from 129 families and 9 phyla) and 11 other taxa at the genus level only (Vertebrates: 3 genera from 2 families and phylum Chordata; Invertebrates: 8 genera from 8 families and 4 phyla; Fig 2). It represented 67.4% of the taxa on the target list (651 species) used and improved during the GSL-rl creation (Vertebrates: 94.6%; Invertebrates: 58.1%; Table S1). The sequences were retrieved mostly from the Northwest Atlantic Ocean (67.8%). A total of 525 BINs were represented (Vertebrates: 159; Invertebrates: 366), with 16 BINs that were shared by at least two taxa (Vertebrates: 8; Invertebrates: 8; Table S2), and 58 taxa occupied more than one BIN (6 vertebrates with up to 3 BINs; 52 invertebrates with up to 7 BINs; Table S3). Median sequence length was 658 pb (range: 640 – 664 pb) while the mean (± sd) of missing values (N's) was 0.002 ± 0.034 % (max < 1 %). Genetic distances were on average 0.005 (range: 0.000–0.023) within BIN and 0.122 (range: 0.012–0.347) between intraspecific BINs.

We then provided a reliability ranking to each species within the GSL-rl based on the completeness of sequences available (Fig 2, Table S4; see methods for more details). Species within the "Reliable" category accounted for the largest proportion of the species with sequences from the regional library (302 species or 68.8%; 133 vertebrates, 169 invertebrates). Species classified to the "Unreliable due to BIN sharing" and the "Unreliable due to gaps" categories represented 5.2% (23 species; 13 Chordata, 10 invertebrates) and 26.0% (114 species; 12 vertebrates, 102 invertebrates) of the GSL-rl species,

204 respectively. The GSL-rl (version 1.0 and future versions) is available on GitHub
205 (https://github.com/GenomicsMLI-DFO/MLI_GSL-rl).

206 ## Accuracy of species assignments using NCBI-nt and two assignment methods

207 The proportions of species assignments overall taxa were higher with the Blast-TopHit method (range:
208 85.5% – 87.9%) than the Blast-LCA method (range: 47.6 – 71.0%) with any identity thresholds (Fig 3A).
209 Overall taxa, the proportions of species assignments increased for the Blast-LCA method while they
210 decreased with the Blast-TopHit method with increasing identity thresholds (Fig. 3A). Across taxonomic
211 groups, the proportions of species assignments were also consistently greater at all thresholds for the
212 Blast-TopHit method compared to the Blast-LCA method. The proportions of species assignments at the
213 97% similarity threshold varied with the Blast-TopHit from 74.4% for Annelida, Brachiopoda, Nemertea
214 to 93.1% for Arthropoda method and with the Blast-LCA method from 35.8% for Cnidaria, Porifera to
215 75.2% for Arthropoda (Fig. 3B).

216 The proportions of accurate species assignments were higher with the Blast-TopHit method compared to
217 the Blast-LCA method, overall taxa and in each taxonomic group at all identify thresholds (Fig. 3A, B).
218 Overall taxa, proportions of accurate species assignments varied between 80.1 and 82.5% for Blast-
219 TopHit method and between 42.7 and 68.0% for the Blast-LCA method for the three identity thresholds
220 tested (Fig. 3A). For each taxonomic group, proportions of accurate assignments at the species level were
221 consistently higher at all thresholds with the Blast-TopHit method compared to the Blast-LCA method.
222 The proportions of accurate species assignments at the 97% threshold varied with the Blast-TopHit
223 method from 69.6% for Annelida, Brachiopoda, Nemertea to 89.6% for Arthropoda and with the Blast-
224 LCA method from 34.0% for Cnidaria and Porifera to 73.6% for Arthropoda (Fig. 3B).

The accuracy was greater for the Blast-LCA method compared to those of the Blast-TopHit method overall taxa at all thresholds (Blast-LCA range: 95.7 – 96.9 %, Blast-TopHit range: 93.8 – 94.4%; Fig. 3A), and in most taxonomic groups at the 97% threshold (Blast-LCA range: 92.3 – 99.2%, Blast-TopHit range: 89.6 – 96.3%; Fig. 3B).

Species assignments using GSL-rl and NCBI-nt, three assignment methods, and a metabarcoding dataset

We used an eDNA metabarcoding dataset to compare the number and the reliability of species assigned using GSL-rl and NCBI-nt, and three assignment methods. The five possible combinations of repository/library and assignment methods were GSL-rl and NCBI-nt with Blast-LCA (1, 2), GSL-rl and NCBI-nt with Blast-TopHit (3,4), and GSL-rl with IDtaxa (5; Fig 1). A total of 80 species were assigned with the five combinations of repository/library and assignment methods (Fig. 4A). Detected species differed using NCBI-nt and GSL-rl and the three assignment methods (Fig. 4A).

Across all combinations, the highest and lowest numbers of species assigned were observed with NCBI-nt and Blast-TopHit95 (66 species) and Blast-LCA95 (44 species), respectively (Fig. 4B). The number of assigned species decreased with increasing thresholds for most combinations, except for Blast-LCA with the NCBI-nt (Fig. 4B). For GSL-rl, proportions of assigned species ranked as “Unreliable due to BIN sharing” or “Unreliable due to gaps” did not increase or decrease linearly with changing thresholds of assignment methods (Fig. 4B). For NCBI-nt, decreasing proportions of “Unlikely” species were assigned with increasing identity thresholds of Blast-LCA or Blast-TopHit (Fig. 4B).

The assignment method with the maximum number of assigned species differed between GSL-rl and NCBI-nt. The maximum number of assigned species was 62 species with the GSL-rl and IDtaxa40 and 66 species with NCBI-nt and TopHit95 (Fig. 4B). Out of the 62 species assigned using the GSL-rl/IDtaxa40 combination, 46 species (74.2%) were ranked as “Reliable”. The remaining assigned species were ranked as “Unreliable due to BIN sharing” (4 species, 6.5%) or “Unreliable due to gaps” (12 species, 19.4%; Fig. 4B). With the NCBI-nt/TopHit95 combination, 58 (87.9%) and 8 (12.1%) assigned species were ranked as “Likely” and “Unlikely” present, respectively (Fig. 4B).

Large proportions of detected species were exclusively assigned using GSL-rl or NCBI-bt. A total of 30 species (37.5% of all species detected) were assigned only using GSL-rl (12 species) or NCBI-nt (18 species; Fig. 4AC). For the species only assigned with GSL-rl, 7 species were ranked as “Reliable” whereas 1 and 4 species were ranked as “Unreliable due to BIN sharing” and “Unreliable due to gaps”, respectively (Fig. 4AC). For the species only assigned with NCBI-nt, 10 species were considered likely to be present in the GSL whereas 8 species were considered unlikely to be present. For the other 50 species assigned with both GSL-rl and NCBI-nt, 39 species were ranked as “Reliable” with the GSL-rl (78.0%, Fig. 2C). The remaining species assigned belonged to the “Unreliable due to BIN sharing” (3 species, 6.0%) or the “Unreliable due to gaps” categories (8 species, 16.0%; Fig. 4AC).

Discussion

Species assignments using DNA barcoding and metabarcoding are affected by the quality and taxonomic coverage of reference sequences and assignment methods used. In this study, we first created a regional library (GSL-rl) for a widely used barcoding gene, COI, that includes a reliability ranking system for species level assignments. We then estimated the accuracy of species assignments with NCBI-nt and two assignment methods using the curated sequences from GSL-rl. While Blast-LCA was the most accurate method when using NCBI-nt, the proportion of accurate species assignments was highest with Blast-TopHit (>80 % overall, between 70 and 90% amongst taxonomic groups). We also compared the number and reliability of species assignments using GSL-rl or NCBI-nt and popular assignment methods with a metabarcoding dataset. The reliability of species assignments was the greatest using the regional library but the regional library and the public repository both provided exclusive plausible species detections, highlighting the importance to use both resources as reference databases. Consequently, we also discuss a two-step approach, using first a regional library followed by a public library, to maximize the reliability of species assignments and the number of species detected in metabarcoding studies.

A COI regional library with a reliability ranking system for metazoans from the Gulf of St. Lawrence (GSL-rl)

The GSL-rl, a curated regional library, provides explicit reliability ranking for 651 species observed within the Gulf of St. Lawrence. We used two simple broad categories, namely “Reliable” and “Unreliable”, to characterize the robustness of species assignments in eDNA metabarcoding studies. This simple ranking system should limit inaccuracies in the interpretation of species assignments for anyone, even with limited scientific background. Past studies have shown the importance of a ranking system to limit

282 erroneous species assignments (e.g., Costa et al. 2012, Knebelsberger et al. 2014). However, the ranking
283 systems used in these studies are targeting an audience of barcoding specialists. With the ranking system
284 of species assignments in GSL-rl, we aimed to keep this classification simple to reach the large audience
285 of eDNA users. Still, we used two "Unreliable" subcategories to highlight 1) the taxa necessitating future
286 barcoding efforts, and 2) the relevancy of the COI barcode to discriminate species. This allows any eDNA
287 scientist to discard the COI loci if species of interest are not discriminated.

288 The reliability ranking of species in GSL-rl may change over time, particularly for understudied species.
289 The reliability ranking is based on recent information of regional species distributions and intra- and
290 inter-genetic diversity. In the future, species may be upgraded to the "Reliable" category when further
291 sequencing results to fill in data gaps. Some species may also be downgraded to the "Unreliable"
292 category, particularly for complex taxonomic groups in the region that should be targeted for review
293 (e.g., polychaete worms). For these reasons, continuous review, curation, and versioning of any regional
294 libraries are important to track changes in reliability ranking through times. A biennial review is planned
295 for the GSL-rl given the large amount of effort required for such work.

296 The GSL-rl covers vertebrates and invertebrate species of interest for conservation in the Gulf of St.
297 Lawrence. The GSL-rl contains reference sequences for 439 species of the 651 targeted species in this
298 study (i.e., 67.4%), with reference sequences available for a relatively large proportion of invertebrates
299 (i.e., 59.1%). In Europe, marine invertebrates represented the taxonomic group with the lowest barcode
300 coverage, where only 22.1% had one or more sequences (Weigand et al. 2019). The larger proportion of
301 invertebrates with reference sequences in the GSL-rl is likely due to the species selection to initiate this
302 regional library but also to the smaller study area and the barcoding campaigns for invertebrates in the
303 Northwest Atlantic (e.g., Radulovici et al. 2009, Layton et al. 2016).

304 The “Reliable” category represented the vast majority of species with reference sequences (68.8%, 302
305 species) in GSL-rl. Similar results were obtained for marine fish species from Portugal at the COI locus
306 (73.5%, grade A, Costa et al. 2012). About a quarter of the GSL-rl species was ranked as “Unreliable due
307 to gaps” due to missing reference sequences from a relative species. For example, the common shelf
308 species of purple sunstar *Solaster endeca* was only ranked as “Unreliable due to gaps” because no
309 reference sequence was available for the local but rare, deep water congener species, *Solaster earlli.*
310 Generally, taxonomic groups for which fewer sequences were recovered in the GSL-rl were also those
311 with more recognized gaps (e.g., Mollusca, Arthropoda). Furthermore, the GSL-rl is in its early
312 development (v.1.0) and presently covers only a quarter of the estimated 2200 marine faunal species
313 that may occur in the Gulf of St. Lawrence (Nozères 2017). Some groups were more underrepresented
314 than others in GSL-rl, e.g., only 10% of the 250 described amphipods. Our study summarizes and lists
315 where future barcoding efforts should be done to fill gaps in the Gulf of St. Lawrence and the Northwest
316 Atlantic.

317 The GSL-rl could also improve species assignments in eDNA metabarcoding studies of the Northwest
318 Atlantic and the Arctic Oceans compared to large public databases. The Gulf of St. Lawrence is a
319 transitional marine region where temperate southern species may occur alongside boreal and arctic
320 species (Bourdages et al. 2022). There are no regional libraries covering marine metazoan species at the
321 COI locus in nearby regions, and the GSL-rl could seed the creation of these regional libraries. These can
322 be created by data mining and curating existing sequences from public repositories (e.g., the approach
323 used in this study), completely de novo from barcoding local specimens (e.g., Delrieu-Trottin et al. 2019),
324 or from a combination of both approaches (e.g., Stoeckle et al. 2020, Gold et al. 2021). New tools are
325 now emerging to facilitate the creation of regional reference libraries (e.g., Meta-Fish-lib, Collins et al.

2021; Barcode, Audit & Grade System (BAGS), Fontes et al. 2021). Some of the tools, such as BAGS, even allow for the annotation of species based on concordance between morphological species-based identification and sequence clusters in BOLD (Fontes et al. 2021). Combined with tools to find gaps in reference sequence libraries (e.g., GAPeDNA, Marques et al. 2021), more comprehensive species-level assignments are now possible.

Accuracy of species assignments using NCBI-nt and two assignment methods

We estimated three performance parameters for metazoan species assignment using NCBI-nt. We observed large variation in those parameters with the two assignment methods tested. While the Blast-LCA method provided overall higher accuracy in species assignments, the proportion of accurate species assignments was greater with Blast-TopHit due to the large difference in the proportion of species assigned, accurate or not. This might be explained by the sensitivity of both methods to the prevalence of BIN sharing, gaps, and mislabeling within public repositories. For instance, Blast-LCA method is expected to be less precise and cause under-classification (i.e., assignments at a higher taxonomic level) in the presence of closely related species and BIN sharing, lowering the proportion of assignments at the species level. In contrast, Blast-TopHit is expected to favor assignments at the species level and is more impacted by gaps and mislabeled sequences (e.g., Schenekar et al. 2020), lowering the accuracy of species assignments. The relatively good performance of Blast-TopHit observed here suggests that gaps and misidentified specimens within NCBI-nt are limited for the targeted marine species of the Gulf of St. Lawrence.

Previous studies have shown that assignment methods can affect taxa detected in metabarcoding studies (O'Rourke et al. 2020, Hleap et al. 2021). Our results confirmed those from Hleap and colleagues

(2021) that the Blast-TopHit method outperformed the Blast-LCA method with NCBI-nt to provide higher proportion of assignments. Our results also show that the proportion of accurate species assignments varied largely between taxonomic groups. Relatively well-described marine taxonomic groups such as Arthropoda (i.e., crustaceans) and Chordata (i.e., fishes and mammals) have reached proportions of accurate species assignments ≥85% with the Blast-TopHit method. The proportion of accurate species assignments is much lower for the Blast-LCA approach with the Chordata (43%), probably because this approach is more sensitive to the presence of close relative species (i.e., BIN sharing) reducing the potential of species level identification. Other groups such as Annelida, Brachiopoda, Nemertea, and Mollusca achieved lower proportions of accurate species assignments using both Blast-TopHit and Blast-LCA methods (maximum 70%).

The proportion of accurate species assignments using the sequences from the GSL-rl in our study will be different at the time of reading this article due to the continuous growth of the public repository NCBI-nt. The publication of new sequences of low quality or with incorrect species identification can create unexpected ambiguities in species assignments as public repositories grow (Locatelli et al. 2020, Radulovici et al. 2021). Without a comprehensive versioning system, changes in the NCBI-nt database also limit the reproducibility of species assignments as it is difficult to identify and to access a specific release. Note that starting with Blast v.2.13 launched in March 2022, it is now possible to generate a metadata file describing the database used (Camacho and Madden 2022), which in an important step toward higher traceability.

366 Comparing reliability of species assignments using GSL-rl and NCBI-nt, three assignment methods,
367 and a metabarcoding dataset

368 The method with the maximal number of species assigned in the metabarcoding dataset differed
369 between GSL-rl and NCBI-nt. The IDtaxa40 assignment method provided the highest number of species
370 assigned using the GSL-rl. Sequence composition strategies for species assignments such as IDtaxa,
371 QIIME, and RDP, had contrasting performance results in previous benchmarking studies (O'Rourke et al.
372 2020, Hleap et al. 2021, Mathon et al. 2021). Our results contrast with those from a previous study
373 showing that IDtaxa did not perform as well as Blast with mock communities composed of various
374 freshwater taxonomic groups (Hleap et al. 2021). The contrasting results between the latter and our
375 studies could be explained by the difference in the confidence threshold used (Hleap et al. 2021).
376 Parameter tuning while using any approach may be key to choosing an optimal method for a dataset
377 while more benchmarking studies are undertaken. The relatively better performance of IDtaxa in our
378 study might also be due to the quality of the regional library used to train the classifier. We know little
379 about the impact of using training sets of different qualities on taxonomic assignment using classifiers,
380 and the gains of using regional libraries might be important in this context. With NCBI-nt, the number of
381 detected species was greater with Blast-TopHit compared to Blast-LCA with the metabarcoding dataset.
382 Those results are similar as those obtained with the GSL-rl COI sequences and have been discussed in the
383 previous section.

384 More than a third of the species assigned (n = 33 out of 80) were exclusive to the use of GSL-rl or NCBI-nt
385 with the metabarcoding dataset. For the GSL-rl, the exclusion of non-indigenous species or mislabeled
386 sequences increased the number of species assigned, confirming previous studies results improving
387 species assignments with regional libraries (von Ammon et al. 2018, Gold et al. 2021). The exclusion of

388 non-indigenous species increased the taxonomic resolution with the GSL-rl of the Atlantic bluefin tuna
389 *Thunnus thynnus*. Underclassification is usually observed when using NCBI-nt as the Atlantic bluefin tuna
390 presently shares a BIN (BOLD: AAA7352) with other *Thunnus* species that are not expected to be present
391 in the Gulf of St. Lawrence (Nozères 2017). With NCBI-nt, detections in the "Likely" category comprised
392 species for which sequences were not included in the GSL-rl because of the stringency of quality filtering
393 performed (e.g., Iceland scallop *Chlamys islandica*). Other detections in the "Likely" category were
394 included in the GSL-rl (e.g., the polychaete worm *Terebellides stroemii*) but the inability to detect them
395 suggests that their intra-specific diversity is not fully covered by the GSL-rl. Finally, a few species assigned
396 with NCBI-nt were not listed as present in the GSL but after reconsideration are likely to be found in the
397 target area (e.g., *Pseudocalanus newmani*). With NCBI-nt, we also observed underclassification of the sea
398 star genus *Leptasterias* due to sequence mislabeling, which is expected to occur (Bidartondo 2008,
399 Mioduchowska et al. 2018). The underclassification is due to two misidentified sequences, one is for
400 *Leptasterias littoralis* identified as the sea star *Asterias forbesi* and the other is for *Leptasterias polaris*
401 identified as the butterfly *Polyommatus fulgens*.

402 Contrasting the ranking category of NCBI-nt and GSL-rl revealed an important gain in reliability with our
403 annotated regional library (Fig 1). With NCBI-nt, we provided the likeliness of a species to be present in
404 the Gulf of St. Lawrence due to the availability of a public species list (Nozères 2017). Such information
405 layer is often difficult to obtain without expert knowledge (Pappalardo et al. 2021). Of all the species
406 ranked in the "Likely" category using NCBI-nt, around 78% were classified as "Reliable" in the GSL-rl. Our
407 results showed that the remaining 22% should be interpreted with caution given gaps (16%) or BIN
408 sharing with close relative species (6%; Fig 4C). Our results suggest that species level assignments of a
409 metabarcoding dataset using NCBI-nt and a filter based on geographic plausibility can be misleading. This

410 important hidden and overlooked uncertainty could be acceptable for empirical studies but not under
411 regulatory context where specific species identification can be crucial, such as the identification of
412 species at risk (Gilbey et al. 2021). Evaluation of false-positives in the detections of endangered or
413 invasive species should include potential bias caused by gaps in reference libraries (Cristescu and Hebert
414 2018).

415 ## Maximizing the reliability and the number of species assignments in eDNA metabarcoding studies
416 ## using a regional library and a public repository

417 Our results showed that the use of a regional library increase both the reliability and the number of
418 species detected with an eDNA metabarcoding dataset. Yet, some species likely present in the Gulf of St.
419 Lawrence were only detected with NCBI-nt, as discussed in the previous section. This will be a recurrent
420 problem until reference sequences are available for more marine metazoan species in the Gulf of St.
421 Lawrence. The growth of GSL-rl will increase the number of species that can be detected using the
422 regional library but unexpected species, such as new invasive species or species that have recently
423 expanded their distribution, could remain undetected (Bohmann et al. 2014, Klymus et al. 2017, Piper et
424 al. 2019, Stoeckle et al. 2020, Gold et al. 2021). Restricting species assignments to GSL-rl and avoiding the
425 use of NCBI-nt would limit the maximum number of species detected.

426 Combining the strengths of a regional library and public repositories as a two-step approach is
427 consequently the optimal solution to maximize reliability and the number of species assigned in
428 metabarcoding studies. Taxonomic assignments should be first performed with a regional library, ideally
429 including a reliability ranking system as in the GSL-rl, to maximize the confidence in species assignment.
430 We then strongly advise contrasting species assignment results from a regional library with those using a

public repository to increase the number of species detections (see also Piper et al. 2019, and Xiong et al. 2022 for similar recommendations). This would allow the reader to have a qualitative estimation of the proportion of accurate species assignments. Species assignments relying uniquely on NCBI-nt should also clearly indicate that their reliability is limited.

We also encourage further benchmarking studies for the selection of optimal methods based on a broader comparison of assignment methods and the development of training sets for machine-learning methods. We limited the size of this study by selecting assignment methods often used in eDNA metabarcoding studies that are also performing relatively well in benchmarking assignment studies (O'Rourke et al. 2020, Hleap et al. 2021). Our results emphasize that future benchmarking studies should be done independently for regional libraries and public repositories, given the different properties of these resources, and to maximize the reliability and the number of species assignments.

Acknowledgements

We thank Grégoire Cortial and Jade Larivière for their inputs at the earlier stages of this study. We also thank Nick Jeffery for helpful comments on a previous version of the manuscript. We thank Yanick Gendreau and Sandra Velasquez from the Coastal environmental baseline program and Geneviève Faille and Geneviève Côté from the Banc-des-Américains Marine Protected Area for eDNA sampling and the initial list of marine faunal species of interest.

References

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution 9: 134–147. https://doi.org/10.1111/2041-210X.12849

von Ammon U, Wood SA, Laroche O, Zaiko A, Tait L, Lavery S, Inglis GJ, Pochon X (2018) Combining morpho-taxonomy and metabarcoding enhances the detection of non-indigenous marine pests in biofouling communities. Scientific Reports 8: 1–11. https://doi.org/10.1038/s41598-018-34541-1

Aylagas E, Borja Á, Rodríguez-Ezpeleta N (2014) Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). PLoS ONE 9. https://doi.org/10.1371/journal.pone.0090529

Bidartondo MI (2008) Preserving Accuracy in GenBank. Science 319: 1616. https://doi.org/10.1126/science.319.5870.1616a

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29: 358–367. https://doi.org/10.1016/J.TREE.2014.04.003

Bourdages H, Brassard C, Chamberland J-M, Desgagnés M, Galbraith P, Isabel L, Senay C (2022) DFO Can. Sci. Advis. Sec. Res. Doc. Preliminary results from the ecosystemic survey in August 2021 in the Estuary and northern Gulf of St. Lawrence. DFO.

Van Den Bulcke L, De Backer A, Ampe B, Maes S, Wittoeck J, Waegeman W, Hostens K, Derycke S (2021) Towards harmonization of DNA metabarcoding for monitoring marine macrobenthos: The effect of technical replicates and pooled DNA extractions on species detection. Metabarcoding and Metagenomics 5: 233–247. https://doi.org/10.3897/MBMG.5.71107

Bush A, Compson ZG, Monk WA, Porter TM, Steeves R, Emilson E, Gagne N, Hajibabaei M, Roy M, Baird DJ (2019) Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. Frontiers in Ecology and Evolution 7: 1–12. https://doi.org/10.3389/fevo.2019.00434

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10: 1–9. https://doi.org/10.1186/1471-2105-10-421

Camacho C, Madden T. BLAST+ Release Notes. 2013 Mar 12 [Updated 2022 Mar 11]. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK131777/

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology 21: 1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Collins RA, Trauzzi G, Maltby KM, Gibson TI, Ratcliffe FC, Hallam J, Rainbird S, Maclaine J, Henderson PA,

Sims DW, Mariani S, Genner MJ (2021) Meta-Fish-Lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. Journal of Fish Biology 99: 1446–1454. https://doi.org/10.1111/jfb.14852

Costa FO, Landi M, Martins R, Costa MH, Costa ME, Carneiro M, Alves MJ, Steinke D, Carvalho GR (2012) A ranking system for reference libraries of DNA barcodes: application to marine fish species from Portugal. PloS one 7: 1–9. https://doi.org/10.1371/journal.pone.0035858

Cristescu ME, Hebert PDN (2018) Uses and misuses of environmental DNA in biodiversity science and conservation. Rev. Ecol. Evol. Syst 49: 209–239. https://doi.org/10.1146/annurev-ecolsys-110617-062306

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology 26: 5872–5895. https://doi.org/10.1111/mec.14350

Delrieu-Trottin E, Williams JT, Pitassy D, Driskell A, Hubert N, Viviani J, Cribb TH, Espiau B, Galzin R, Kulbicki M, Lison de Loma T, Meyer C, Mourier J, Mou-Tham G, Parravicini V, Plantard P, Sasal P, Siu G, Tolou N, Veuille M, Weigt L, Planes S (2019) A DNA barcode reference library of French Polynesian shore fishes. Scientific Data 6: 1–8. https://doi.org/10.1038/s41597-019-0123-5

Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21: 573–583. https://doi.org/10.1111/1755-0998.13262

Fraija-Fernández N, Bouquieaux MC, Rey A, Mendibil I, Cotano U, Irigoien X, Santos M, Rodríguez-Ezpeleta N (2020) Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. Ecology and Evolution 10: 7560–7584. https://doi.org/10.1002/ece3.6482

Gilbey J, Carvalho G, Castilho R, Coscia I, Coulson MW, Dahle G, Derycke S, Francisco SM, Helyar SJ, Johansen T, Junge C, Layton KKS, Martinsohn J, Matejusova I, Robalo JI, Rodríguez-Ezpeleta N, Silva G, Strammer I, Vasemägi A, Volckaert FAM (2021) Life in a drop: Sampling environmental DNA for marine fishery management and ecosystem monitoring. Marine Policy 124. https://doi.org/10.1016/j.marpol.2020.104331

Gold Z, Curd EE, Goodwin KD, Choi ES, Frable BW, Thompson AR, Walker HJ, Burton RS, Kacev D, Martz LD, Barber PH (2021) Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. Molecular Ecology Resources 21: 2546–2564. https://doi.org/10.1111/1755-0998.13450

Hänfling B, Lawson Handley L, Read DS, Hahn C, Li J, Nichols P, Blackman RC, Oliver A, Winfield IJ (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Molecular Ecology 25: 3101–3119. https://doi.org/10.1111/mec.13660

Hering D, Borja A, Jones JI, Pont D, Boets P, Bouchez A, Bruce K, Drakare S, Hänfling B, Kahlert M, Leese F, Meissner K, Mergen P, Reyjol Y, Segurado P, Vogler A, Kelly M (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. Water Research 138: 192–205. https://doi.org/10.1016/j.watres.2018.03.003

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21: 2190–2203. https://doi.org/10.1111/1755-0998.13407

Jazdzewska AM, Tandberg AHS, Horton T, Brix S (2021) Global gap-analysis of amphipod barcode library. PeerJ 9: 1–28. https://doi.org/10.7717/peerj.12352

Klymus KE, Marshall NT, Stepien CA (2017) Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS ONE 12: 1–24. https://doi.org/10.1371/journal.pone.0177643

Knebelsberger T, Landi M, Neumann H, Kloppmann M, Sell AF, Campbell PD, Laakmann S, Raupach MJ, Carvalho GR, Costa FO (2014) A reliable DNA barcode reference library for the identification of the North European shelf fish fauna. Molecular Ecology Resources 14: 1060–1071. https://doi.org/10.1111/1755-0998.12238

Layton KKS, Corstorphine EA, Hebert PDN (2016) Exploring canadian echinoderm diversity through DNA barcodes. PLoS ONE 11: 1–16. https://doi.org/10.1371/journal.pone.0166118

Leite BR, Vieira PE, Troncoso JS, Costa FO (2021) Comparing species detection success between molecular markers in DNA metabarcoding of coastal macroinvertebrates. Metabarcoding and Metagenomics 5: 249–260. https://doi.org/10.3897/MBMG.5.70063

Leray M, Knowlton N, Ho SL, Nguyen BN, Machida RJ (2019) GenBank is a reliable resource for 21st century biodiversity research. Proceedings of the National Academy of Sciences of the United States of America 116: 22651–22656. https://doi.org/10.1073/pnas.1911714116

Locatelli NS, McIntyre PB, Therkildsen NO, Baetscher DS (2020) GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. Proceedings of the National Academy of Sciences of the United States of America 117: 32211–32212. https://doi.org/10.1073/pnas.2007421117

Makiola A, Compson ZG, Baird DJ, Barnes MA, Boerlijst SP, Bouchez A, Brennan G, Bush A, Canard E, Cordier T, Creer S, Curry RA, David P, Dumbrell AJ, Gravel D, Hajibabaei M, Hayden B, van der Hoorn B, Jarne P, Jones JI, Karimi B, Keck F, Kelly M, Knot IE, Krol L, Massol F, Monk WA, Murphy J, Pawlowski J, Poisot T, Porter TM, Randall KC, Ransome E, Ravigné V, Raybould A, Robin S, Schrama M, Schatz B, Tamaddoni-Nezhad A, Trimbos KB, Vacher C, Vasselon V, Wood S, Woodward G, Bohan DA (2020) Key questions for next-generation biomonitoring. Frontiers in Environmental Science 7: 1–14. https://doi.org/10.3389/fenvs.2019.00197

Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel J-B (2021) GAPeDNA: Assessing and mapping global species gaps in genetic databases for metabarcoding studies. Diversity and

559 distribution 27: 1880–1892. https://doi.org/10.3897/aca.4.e64884

560 Mathon L, Valentini A, Guérin PE, Normandeau E, Noel C, Lionnet C, Boulanger E, Thuiller W, Bernatchez
561 L, Mouillot D, Dejean T, Manel S (2021) Benchmarking bioinformatic tools for fast and accurate
562 eDNA metabarcoding species identification. Molecular Ecology Resources 21: 2565–2579.
563 https://doi.org/10.1111/1755-0998.13430

564 McCusker MR, Denti D, Guelpen L, Kenchington E, Bentzen P (2013) Barcoding Atlantic Canada's
565 commonly encountered marine fishes. Molecular Ecology Resources 13: 177–188.
566 https://doi.org/10.1111/1755-0998.12043

567 McGee KM, Robinson C V., Hajibabaei M (2019) Gaps in DNA-Based Biomonitoring Across the Globe.
568 Frontiers in Ecology and Evolution 7: 1–7. https://doi.org/10.3389/fevo.2019.00337

569 Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank – Their accuracy and
570 reliability for the identification of biological materials. PLoS ONE 14: 1–14.
571 https://doi.org/10.1371/journal.pone.0217084

572 Mioduchowska M, Czyż MJ, Gołdyn B, Kur J, Sell J (2018) Instances of erroneous DNA barcoding of
573 metazoan invertebrates: Are universal cox1 gene primers too "universal"? Hajibabaei M (Ed.). PLOS
574 ONE 13: e0199609. https://doi.org/10.1371/journal.pone.0199609

575 Murali A, Bhargava A, Wright ES (2018) IDTAXA: a novel approach for accurate taxonomic classification of
576 microbiome sequences. Microbiome 6: 140. https://doi.org/10.1186/s40168-018-0521-5

577 Nozères C (2017) Preliminary checklist of marine animal species of the Gulf of St. Lawrence, Canada,
578 based on 4 sources. https://doi.org/10.13140/RG.2.2.10056.62727

579 O'Rourke DR, Bokulich NA, Jusino MA, MacManes MD, Foster JT (2020) A total crapshoot? Evaluating
580 bioinformatic decisions in animal diet metabarcoding analyses. Ecology and Evolution 10: 9721–
581 9739. https://doi.org/10.1002/ece3.6594

582 Oliveira LM, Knebelsberger T, Landi M, Soares P, Raupach MJ, Costa FO (2016) Assembling and auditing a
583 comprehensive DNA barcode reference library for European marine fishes. Journal of Fish Biology
584 89: 2741–2754. https://doi.org/10.1111/jfb.13169

585 Pappalardo P, Collins AG, Pagenkopp Lohan KM, Hanson KM, Truskey SB, Jaeckle W, Ames CL, Goodheart
586 JA, Bush SL, Biancani LM, Strong EE, Vecchione M, Harasewych MG, Reed K, Lin C, Hartil EC,
587 Whelpley J, Blumberg J, Matterson K, Redmond NE, Becker A, Boyle MJ, Osborn KJ (2021) The role
588 of taxonomic expertise in interpretation of metabarcoding studies. ICES Journal of Marine Science
589 78: 3397–3410. https://doi.org/10.1093/icesjms/fsab082

590 Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ (2019) Prospects and
591 challenges of implementing DNA metabarcoding for high-throughput insect surveillance.
592 GigaScience 8: 1–22. https://doi.org/10.1093/gigascience/giz092

593 Porter TM, Hajibabaei M (2018a) Automated high throughput animal CO1 metabarcode classification.

Scientific Reports 8: 1–10. https://doi.org/10.1038/s41598-018-22505-4

Porter TM, Hajibabaei M (2018b) Over 2.5 million COI sequences in GenBank and growing. Arthofer W (Ed.). PLOS ONE 13: e0200177. https://doi.org/10.1371/journal.pone.0200177

Porter TM, Hajibabaei M (2020) Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. Frontiers in Ecology and Evolution 8: 1–15. https://doi.org/10.3389/fevo.2020.00248

Radulovici AE, Vieira PE, Duarte S, Teixeira MAL, Borges LMS, Deagle BE, Majaneva S, Redmond N, Schultz JA, Costa FO (2021) Revision and annotation of DNA barcode records for marine invertebrates: report of the 8th iBOL conference hackathon. Metabarcoding and Metagenomics 5: 207–217. https://doi.org/10.3897/mbmg.5.67862

Radulovici AE, Sainte-Marie B, Dufresne F (2009) DNA barcoding of marine crustaceans from the Estuary and Gulf of St Lawrence: A regional-scale approach. Molecular Ecology Resources 9: 181–187. https://doi.org/10.1111/j.1755-0998.2009.02643.x

Ramirez JL, Rosas-Puchuri U, Cañedo RM, Alfaro-Shigueto J, Ayon P, Zelada-Mázmela E, Siccha-Ramirez R, Velez-Zuazo X (2020) DNA barcoding in the Southeast Pacific marine realm: Low coverage and geographic representation despite high diversity. PLoS ONE 15: 1–13. https://doi.org/10.1371/journal.pone.0244323

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System: Barcoding. Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0066213

Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources 17: 760–769. https://doi.org/10.1111/1755-0998.12628

Schenekar T, Schletterer M, Lecaudey LA, Weiss SJ (2020) Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. River Research and Applications 36: 1004–1013. https://doi.org/10.1002/rra.3610

Stoeckle MY, Soboleva L, Charlop-Powers Z (2017) Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. PLoS ONE 12: 1–15. https://doi.org/10.1371/journal.pone.0175186

Stoeckle MY, Das Mishu M, Charlop-Powers Z (2020) Improved environmental DNA reference library detects overlooked marine fishes in New Jersey, United States. Frontiers in Marine Science 7. https://doi.org/10.3389/fmars.2020.00226

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation

biodiversity assessment using DNA metabarcoding. Molecular Ecology 21: 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73: 5261–5267. https://doi.org/10.1128/AEM.00062-07

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of The Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

WoRMS Editorial Board (2020) World Register of Marine Species. https://doi.org/10.14284/170

Wright ES (2016) Using DECIPHER v2.0 to analyze big biological sequence data in R. R Journal 8: 352–359. https://doi.org/10.32614/rj-2016-025

Xiong F, Shu L, Zeng H, Gan X, He S, Peng Z (2022) Methodology for fish biodiversity monitoring with environmental DNA metabarcoding: The primers, databases and bioinformatic pipelines. Water Biology and Security 1: 100007. https://doi.org/10.1016/j.watbs.2022.100007

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3: 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zafeiropoulos H, Gargan L, Hintikka S, Pavloudi C, Carlsson J (2021) The Dark mAtteR iNvestigator (DARN) tool: Getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics 5: 163–174. https://doi.org/10.3897/MBMG.5.69657

Data Accessibility

The data and scripts used in this manuscript are stored in the github repository

https://github.com/GenomicsMLI-DFO/GSL_COI_ref_library

The GSL-rl (sequences, reliability ranking and trained dataset) can be found in the github repository

https://github.com/GenomicsMLI-DFO/MLI_GSL-rl

Benefit-Sharing

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author Contributions

A.B. co-conceived and co-developed the ideas underlying the manuscript, wrote scripts, compiled GSL-rl, analyzed data, and co-wrote the manuscript. C.N. initiated the project of a reference library for the GSL, co-conceived and co-developed the ideas, revised GSL-rl, analyzed data and edited all drafts. E.P. initiated the project of a reference library for the GSL and revised the manuscript. G.J.P. co-conceived and co-developed the ideas, co-wrote the manuscript and secured funding.

673 # Figure legends

674

675 **Figure 1**. Schematic representation of the two sources of reference sequences used and compared in this
676 study and their associated methods, A) the public repository NCBI nucleotide database (NCBI-nt), and B)
677 the newly created regional library for metazoans from the Gulf of St. Lawrence (GSL-rl). Using NCBI-nt,
678 exact sequence variants (ESVs) were assigned using the BlastX tool *blastn* (hereafter Blast; Camacho et
679 al. 2009). Assignment results were filtered based on taxonomic identity, then a least common ancestor
680 (LCA) or a TopHit method was used to assign a unique taxon identity to each ESV. The creation of GSL-rl
681 involved data mining of the public repository BOLD and included multiple filtering and auditing steps and
682 a feedback loop to improve it. Taxonomic assignments of ESV were performed with Blast or with the
683 classifier IDtaxa (Murali et al. 2018). For GSL-rl, the species ranking was based on sequence availability
684 and sequence similarity to closely related species in the Gulf of St. Lawrence. For NCBI-nt, the species
685 ranking involved a plausibility filter based on the location (see methods for more details).

686 **Figure 2.** Classification of 651 marine faunal species previously observed in the Gulf of St. Lawrence and
687 included in the GSL-rl, by phylum. Species reliability ranking is based on the availability from local species
688 and sequence similarity to closely related species in the Gulf of St. Lawrence.

689 **Figure 3.** Results of taxonomic assignment of sequences from the GSL-rl using NBCI-nt and Blast-LCA or
690 Blast-TopHit methods. Panels A and B present the proportions of accurate and inaccurate species
691 assignments. Results are presented for all taxonomic groups at the three identity thresholds (95%, 97%,
692 99%; panel A) and by taxonomic group at the 97% threshold (panel B).

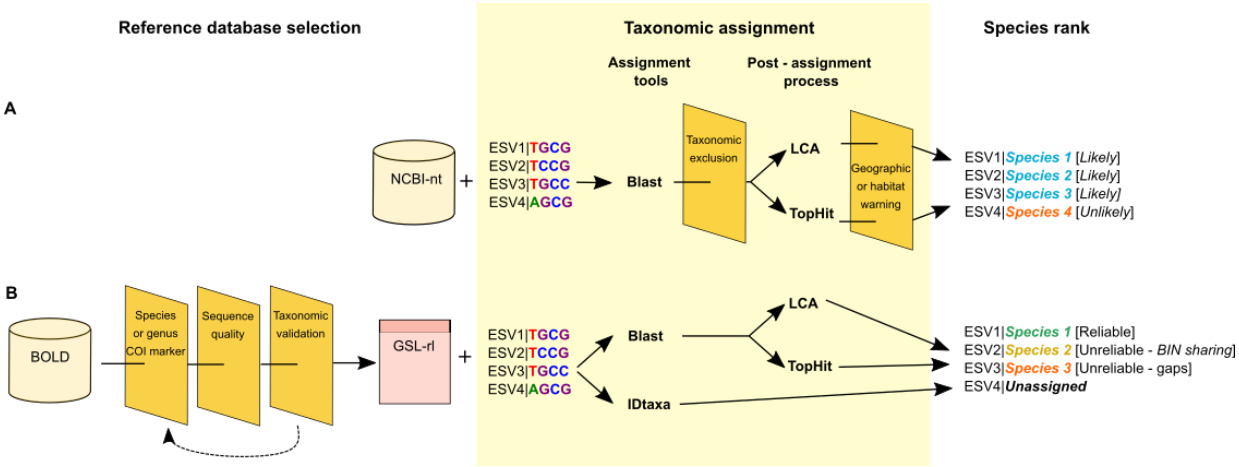
693 **Figure 4.** Assignment results at the species level using a regional library (GSL-rl) or a public repository
694 (NBCI-nt), and popular assignment methods. We used three assignment methods, namely IDtaxa
695 (confidence levels: 40%, 50% and 60%), Blast-LCA, and Blast-TopHit (identity thresholds: 95%, 97% and
696 99%). Panel A detailed the detections for each species, and panel B synthesize the number of species
697 assignments for each source of reference sequences and method. Panel C compared the species rank for
698 all the species assigned with the two sources. Species rank categories are based on sequence availability
699 and sequence similarity to closely related species in the Gulf of St. Lawrence for GSL-rl and on the
700 geographic plausibility for NCBI-nt.

701

702

703 Figures

704

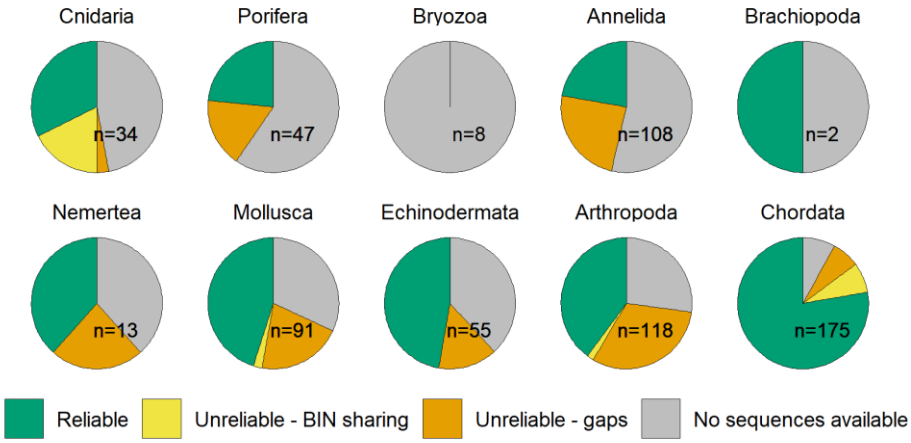


705

706 Fig. 1

707

708

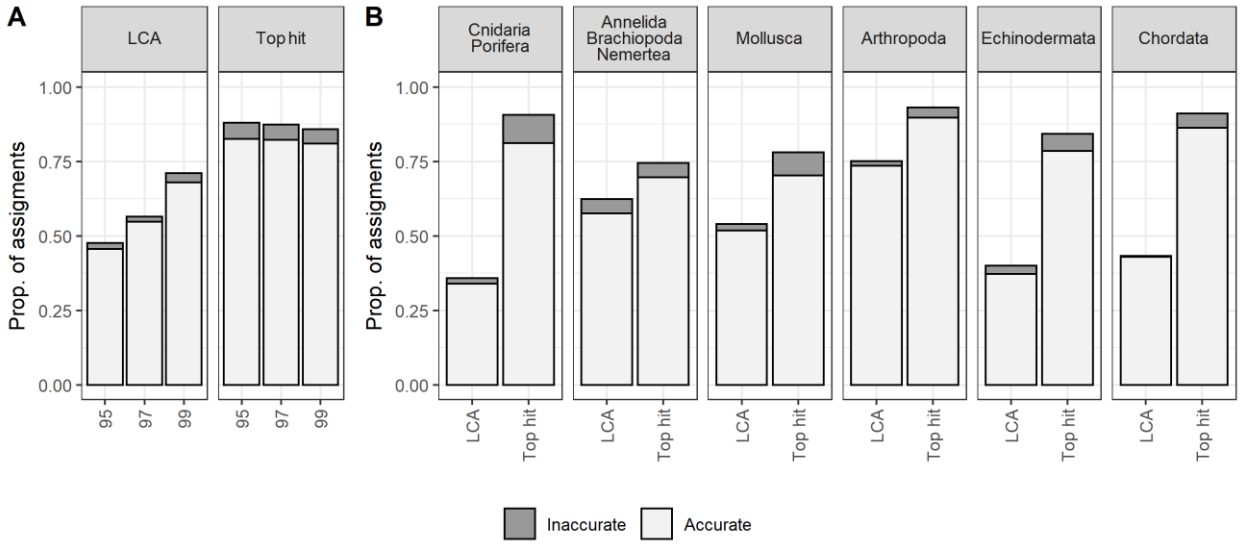


709

710 Fig. 2

711

712



713

714

715 Fig. 3

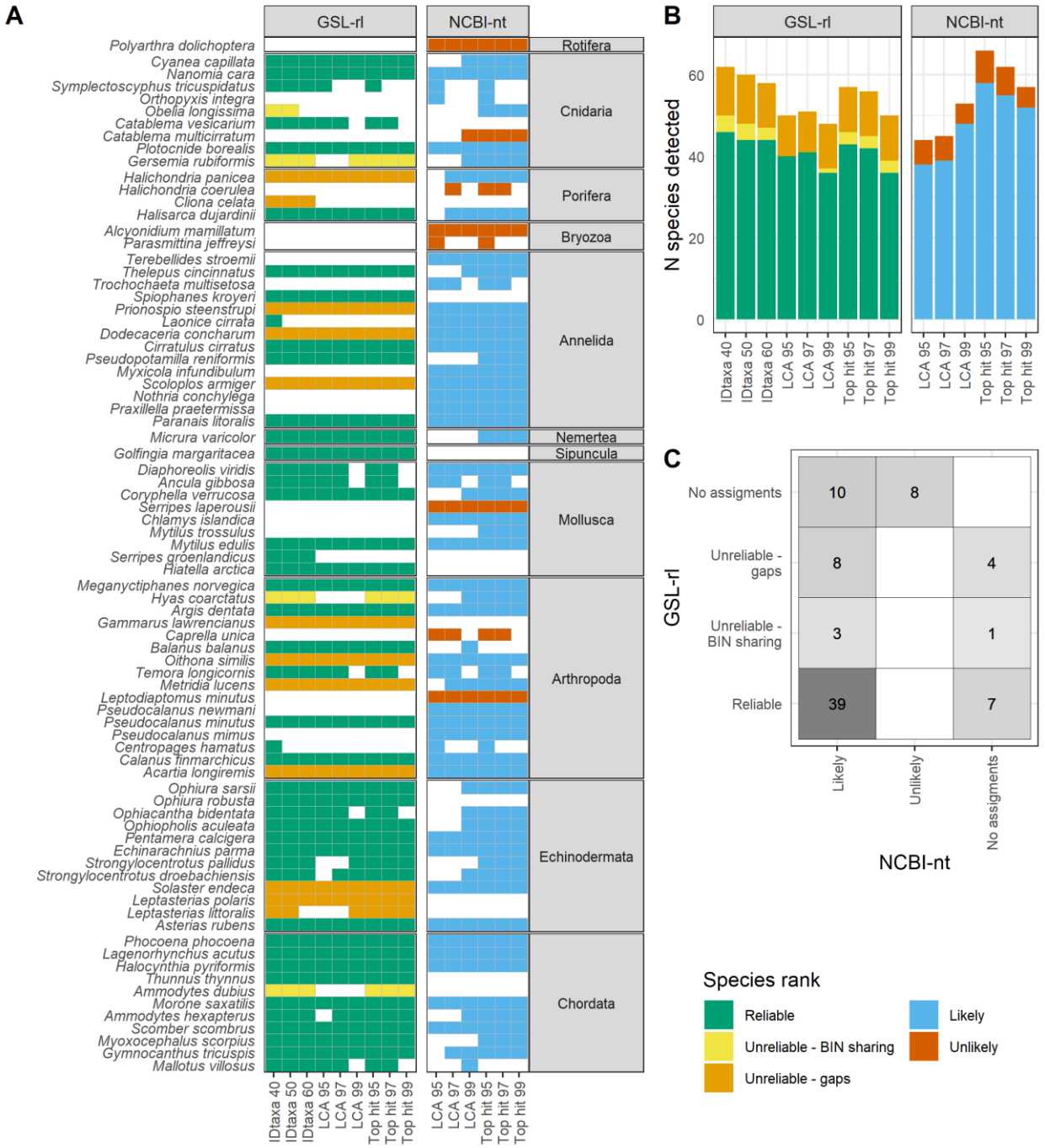


Fig. 4