

Supplementary materials

Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional reference library and a public repository

Bourret, A., Nozères, C., Parent, É., Parent, G.J.

A. Creation of Gulf of St. Lawrence regional library (GSL-rl)

Table S1 – List of species retrieved in BOLD under different names

Table S2 – List of taxa BIN number removed

B. Supplementary figures and tables from the manuscript

Fig. S1 – Steps to obtain, filter and select publicly available sequences

Table S3 – Metadata of the GSL-rl, including ranking systems. [csv file]

Table S4 – BINs shared by two or more taxa

Table S5 – Taxa sharing more than one BIN within the GSL-rl

Table S6 – Characteristic of the curated regional library of the Gulf of St. Lawrence (GSL-rl) version 1.0

C. Creation of an eDNA metabarcoding dataset

Fig. S2 – Sampling locations for the metabarcoding dataset in the St. Lawrence

Fig. S3 – Schematic representation of the metabarcoding bioinformatics pipeline

A. Creation of Gulf of St. Lawrence regional library (GSL-rl)

The creation of regional library for the Gulf of St. Lawrence (GSL) metazoans was composed of three major steps: 1) obtaining a preliminary list of species from decision-makers, management personnel and regional taxonomic information, 2) creating an initial regional library using bioinformatics tools and round of reviews (hereafter draft regional library), and 3) enhancing the draft regional library using a metabarcoding dataset (hereafter GSL regional library (GSL-rl); Fig. S1).

Step 1. Reviewed species list

Decision-makers and management personnel provided a list of taxa (species or higher taxonomic ranks), observed with traditional survey methods, or suspected being present. We enhanced this list by adding potential species within this group present in the geographical range based on a checklist of marine animal species of the GSL (Nozères 2017). This list was reviewed by Claude Nozères, a regional marine biodiversity expert, to validate their presence within the GSL and to highlight species with identification problems at the morphological level or in need of further taxonomic studies. The species names list was then matched with the World Register of Marine Species (WoRMS, WoRMS Editorial Board 2020) through an in-house R script. Recent name changes were noted to allow sequence searches with invalid retrieved names (Table S1). Taxonomic classification hierarchy for each focal taxon was also retrieved using WoRMS database.

Step 2. Draft regional library

We then used BOLD to seek reference sequences due to the availability of data for COI, the quality of metadata, and the relative ease to access and obtain the data with bioinformatics tools. In BOLD, large metadata are available with each sequence (e.g., regions, sites, pictures of the specimen, sequencing traces), and a BIN information (i.e., unique identifier of sequences based on genetic distance) is also provided, which are not available while working with NCBI-nt.

We built an R script using the package *bold* (v.1.1.0, Chamberlain 2020) for datamining of COI sequences and metadata for species of interest within BOLD. Searches on BOLD were done with the function *bold_seqspect*. We filtered sequences based on different criteria including markercode as “COI-5P”, less than 1% ambiguous sites over the nucleotides, a preference for sampling site between 35–78 latitude and -78–40 longitude, province_state within Quebec, Nova Scotia, New Brunswick, Newfoundland and Labrador, Prince Edward Island, Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New Jersey, Delaware, Maryland, Virginia, column “region”, “sector” and “exactsite” including St-Laurent, Saint-Laurent, St Lawrence, St. Lawrence, Fundy, Flemish Cap, Maritimes, Scotian Shelf, Georges Bank, Grand Bank, Baffin Bay, Durban Harbour, but excluding Hudson Bay, Hudson Strait, Beaufort, Nord-du-Quebec, Ungava. Entries without bin_uri defined were excluded. Sequences were classified based on their geographic origin (i.e., NWA or not) and a proxy of quality (i.e., proportion of ambiguous nucleotides), the presence of a picture, and the length of the sequences. When Northwest Atlantic sequences were available, up to 50 were kept to largely cover the genetic diversity. When none were within the geographical limits, up to 20 were kept from undefined or other locations. Sequences were then aligned and marker sites (LCO1490, HCO2198) were retrieved and trimmed before dereplication. Sequences shorter than 650 pb were removed. BIN information for each sequences were kept. For each BIN, a reverse search was made to check all associated taxonomic classifications; whenever discrepancy was observed with taxa-BIN identification (e.g., most sequences associated with

another species), further investigations were made to assess the causes (i.e., potential error or genetic similarity to another species; Table S2).

We used the list of species to retrieve the vast majority of sequences. For species with no matching sequences but known recent name change, a search was made with the previous known name (Table S1). Sequences from other species than those from the list were also added if we identified a close relative species present in an area close to the GSL. Finally, for flagged species (i.e., identified a priori as potentially problematic, see Step 1), searches were also made at the genus levels, to retrieve potential sequences within NWA not previously covered by a BIN. We also assessed the number of intraspecific BIN and the number of species within BIN. We computed a genetic distance intra and interspecific between BINs as the Kimura's 2-parameters distance (Kimura 1980) with ape R package (v.5.0, Paradis and Schliep 2019).

Step 3. GSL regional library

To check for missing important species from the draft regional library, ESV observed within the metabarcoding dataset (see section C below on the creation of the eDNA metabarcoding dataset) were compared over NCBI-nt (downloaded 2020-10-23), using the BLAST+ tool *blastn* (v.2.10.1. Camacho et al. 2009). Sequences with at least 95% identity, and with an alignment length \geq 95% of the compared sequence, were kept. We examined sequence names within the Metazoa kingdom to detect potential uncovered taxa (i.e., taxa not in the a priori taxa list but possibly within the NWA). New taxa were used for a second round of data mining on BOLD, as previously described in the draft-local library section above. Resulting aligned, trimmed, and filtered sequences were joined to the draft regional library to create an improved GSL regional library (GSL-rl) based on metabarcoding data results.

Each species within the GSL-rl was annotated based on the completeness and similarity to close relative species. Species sharing a BIN group with other species within the GSL-rl were annotated as "Unreliable due to BIN sharing", species close to another species (i.e., within the same genus) with no reference sequences were classified as "Unreliable due to gaps"; all other species, with unique BIN and reference sequence for all known local species from the same genus, were noted "Reliable" (Table S1). Searches on BOLD and WORMS for the step 2 were performed between 2020-09-30 and 2020-10-05, and between 2021-01-16 and 2021-01-20 for the step 3. Note that *Aulactina stella* was also resolved by hand, because there was an error in the loading of BOLD metadata.

Table S1 List of species retrieved in BOLD under different names following WoRMS current taxonomy (consulted between 2020-09 and 2021-01).

Retrieved name (BOLD)	Current name (WoRMS 2021)	Reason
<i>Myxine glutinosa</i>	<i>Myxine limosa</i>	Other is NE Atlantic only
<i>Flabellina salmonacea</i>	<i>Ziminella salmonacea</i>	Synonym
<i>Melita dentata</i>	<i>Megamoera dentata</i>	Synonym
<i>Dendronotus niveus</i>	<i>Dendronotus elegans</i>	Synonym
<i>Palaemonetes vulgaris</i>	<i>Palaemon vulgaris</i>	Synonym
<i>Palaemonetes pugio</i>	<i>Palaemon pugio</i>	Synonym
<i>Radiella hemisphaerica</i>	<i>Polymastia hemisphaerica</i>	Synonym
<i>Anonyx lilljeborgii</i>	<i>Anonyx lilljeborgi</i>	Synonym
<i>Brada villosa</i>	<i>Bradabyssa villosa</i>	Synonym
<i>Diplocirrus longisetosus</i>	<i>Saphobranchia longisetosa</i>	Synonym
<i>Scoloplos acutus</i>	<i>Leitoscoloplos acutus</i>	Synonym
<i>Halisarca dujardini</i>	<i>Halisarca dujardinii</i>	Synonym

Table S2 List of taxa BIN number removed for a specific species from the regional library during the two rounds of data mining on BOLD, and the reason of their exclusion. Note that these BINs could be present for others species within the GSL-rl.

Species	BIN	Reason of exclusion
<i>Acanthodoris Pilosa</i>	BOLD:ACO2942	Another BIN is more geographically relevant
<i>Acanthodoris pilosa</i>	BOLD:AAX0711	Another BIN is more geographically relevant
<i>Anarhichas denticulatus</i>	BOLD:ACE4267	Highly potential wrong identification (<i>Anarchinas minor</i>)
<i>Arctogadus glacialis</i>	BOLD:AAA9472	Highly potential wrong identification (<i>Boreogadus saida</i>)
<i>Artediellus atlanticus</i>	BOLD:AAB1020	Highly potential wrong identification (<i>Triglops murrayi</i>)
<i>Aspidophoroides monopterygius</i>	BOLD:AAA9929	Highly potential wrong identification
<i>Balanus crenatus</i>	BOLD:AEB7035	Another BIN is more geographically relevant
<i>Balanus crenatus</i>	BOLD:AAB1410	Highly potential wrong identification
<i>Chaetozone setosa</i>	BOLD:AAB5820	Another BIN is more geographically relevant
<i>Cirratulus cirratus</i>	BOLD:AAJ0882	Another BIN is more geographically relevant
<i>Cirratulus cirratus</i>	BOLD:ACJ5831	Another BIN is more geographically relevant
<i>Cirratulus cirratus</i>	BOLD:ACH1141	Another BIN is more geographically relevant
<i>Coryphella verrucosa</i>	BOLD:AAN6579	Highly potential wrong identification
<i>Cyanea capillata</i>	BOLD:AAF9673	Another BIN is more geographically relevant
<i>Cyanea capillata</i>	BOLD:ACM6954	Another BIN is more geographically relevant
<i>Doto coronata</i>	BOLD:AAE8508	Another BIN is more geographically relevant
<i>Doto coronata</i>	BOLD:ACB8287	Another BIN is more geographically relevant
<i>Eualus pusiolus</i>	BOLD:ABW0161	Another BIN is more geographically relevant
<i>Laonice cirrata</i>	BOLD:ACH1038	Another BIN is more geographically relevant
<i>Laonice cirrata</i>	BOLD:ACC2851	Another BIN is more geographically relevant
<i>Laonice cirrata</i>	BOLD:ACI3025	Another BIN is more geographically relevant
<i>Laonice cirrata</i>	BOLD:AAV0397	Another BIN is more geographically relevant
<i>Laonice cirrata</i>	BOLD:ACV6427	Another BIN is more geographically relevant
<i>Limnodrilus hoffmeisteri</i>	BOLD:ADK1554	Another BIN is more geographically relevant
<i>Limnodrilus hoffmeisteri</i>	BOLD:ACW0660	Another BIN is more geographically relevant
<i>Littorina saxatilis</i>	BOLD:AAA7469	Highly potential wrong identification
<i>Micrura varicolor</i>	BOLD:AAK8914	Highly potential wrong identification
<i>Obelia dichotoma</i>	BOLD:AAA7089	Highly potential wrong identification
<i>Obelia geniculata</i>	BOLD:AAA7088	Another BIN is more geographically relevant
<i>Obelia geniculata</i>	BOLD:ADK7351	Another BIN is more geographically relevant
<i>Obelia longissima</i>	BOLD:AAF4295	Highly potential wrong identification
<i>Oithona similis</i>	BOLD:AAD7750	Another BIN is more geographically relevant
<i>Oithona similis</i>	BOLD:ACQ4233	Another BIN is more geographically relevant
<i>Onchidoris muricata</i>	BOLD:AAZ3494	Highly potential wrong identification
<i>Orthopyxis integra</i>	BOLD:AAE5348	Highly potential wrong identification
<i>Orthopyxis integra</i>	BOLD:AAE5349	Highly potential wrong identification
<i>Orthopyxis integra</i>	BOLD:AAE5350	Highly potential wrong identification
<i>Orthopyxis integra</i>	BOLD:AAE5351	Highly potential wrong identification
<i>Pandalus montagui</i>	BOLD:AAB2200	Highly potential wrong identification

<i>Phycis chesteri</i>	BOLD:AAB0057	Highly potential wrong identification (<i>Merluccius bilinearis</i>)
<i>Polymastia boletiformis</i>	BOLD:ACQ0787	Highly potential wrong identification
<i>Praxillella affinis</i>	BOLD:ACH1056	Another BIN is more geographically relevant
<i>Praxillella praetermissa</i>	BOLD:ACH1056	Another BIN is more geographically relevant
<i>Spiophanes bombyx</i>	BOLD:ACH1173	Another BIN is more geographically relevant
<i>Spiophanes bombyx</i>	BOLD:ACH1123	Another BIN is more geographically relevant
<i>Spirorbis spirorbis</i>	BOLD:AAJ3464	Highly potential wrong identification
<i>Stomias boa</i>	BOLD:AAB5302	Another BIN is more geographically relevant
<i>Temora stylifera</i>	BOLD:ADK4239	Another BIN is more geographically relevant
<i>Temora stylifera</i>	BOLD:ADK0428	Another BIN is more geographically relevant
<i>Triglops nybelini</i>	BOLD:AAB1020	Highly potential wrong identification
<i>Triglops nybelini</i>	BOLD:ADQ8959	Highly potential wrong identification
<i>Trochochaeta multisetosa</i>	BOLD:ADH7281	Another BIN is more geographically relevant
<i>Tubifex tubifex</i>	BOLD:AEB6122	Another BIN is more geographically relevant
<i>Tubifex tubifex</i>	BOLD:AAA6805	Another BIN is more geographically relevant
<i>Urasterias lincki</i>	BOLD:AAG0209	Highly potential wrong identification
<i>Weberella bursa</i>	BOLD:ACH5414	Highly potential wrong identification

B. Supplementary figures and tables from the manuscript

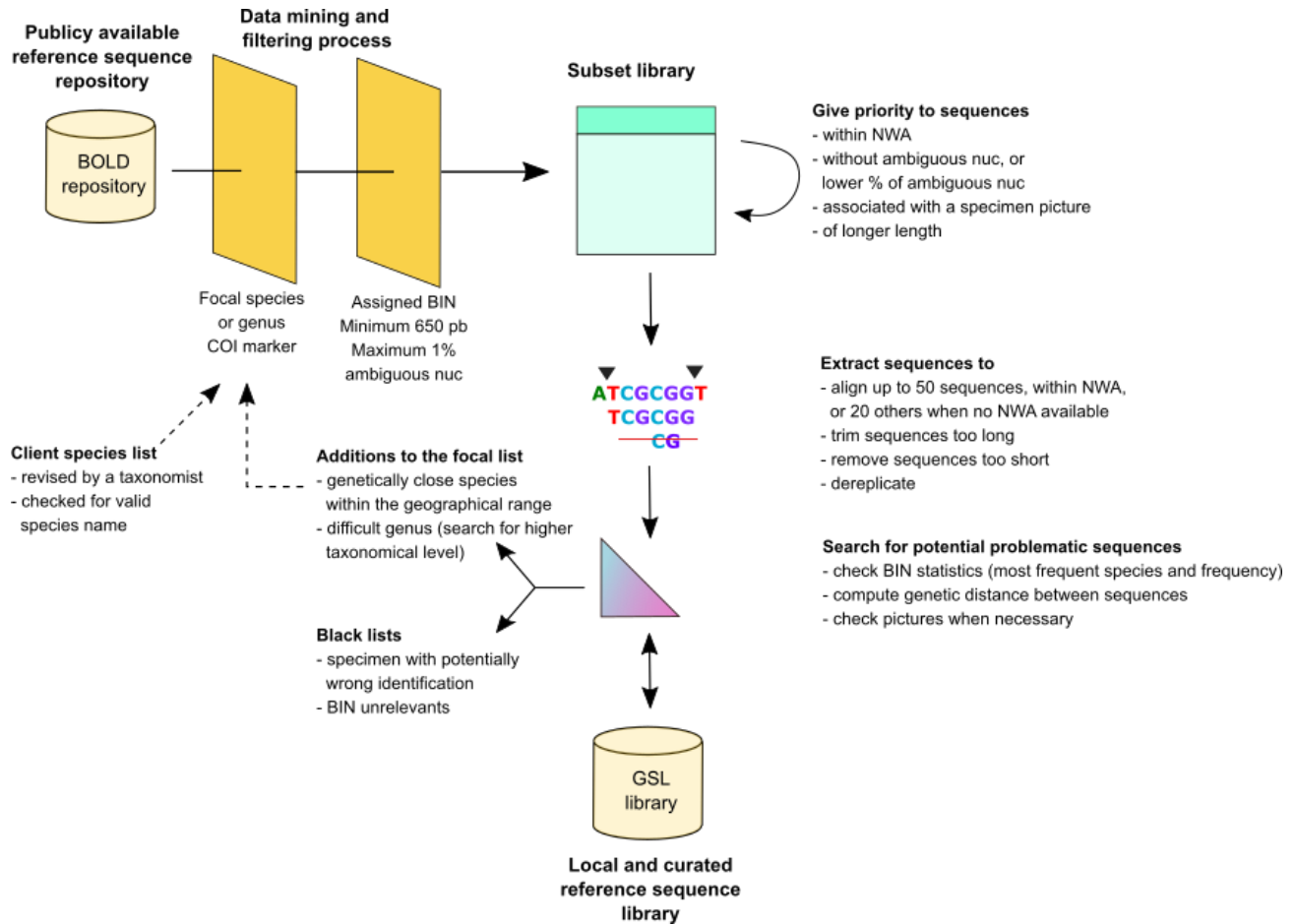


Figure S1 Summary of the steps to obtain, filter and select publicly available sequences to create a regional library. First, a filtration step through BOLD metadata. Second, remaining sequences are ordered based on various criteria. Then sequences are extracted, aligned, trimmed, filtered and dereplicated. Remaining sequences are checked to identify potential identification problems, from BIN information and genetic distances, then added to the curated regional library for the Gulf of St. Lawrence (GSL-rl). Since it is an iterative process, new species or genus can be searched again to further improve the GSL-rl. All these steps can be done through an R code.

Table S4. BINs shared by two or more taxa in the GSL-rl

Category	Phylum	BIN	Species
Vertebrate	Chordata	BOLD:AAB1380	<i>Alosa aestivalis</i> , <i>Alosa pseudoharengus</i>
Vertebrate	Chordata	BOLD:AAB4332	<i>Ammodytes americanus</i> , <i>Ammodytes dubius</i>
Vertebrate	Chordata	BOLD:AAA9928	<i>Aspidophoroides monopterygius</i> , <i>Aspidophoroides olrikii</i>
Vertebrate	Chordata	BOLD:AAB3781	<i>Eumicrotremus spinosus</i> †, <i>Eumicrotremus terraenovae</i>
Vertebrate	Chordata	BOLD:ACA5464	<i>Gymnelus hemifasciatus</i> ‡, <i>Gymnelus viridis</i>
Vertebrate	Chordata	BOLD:AAB4897	<i>Liparis atlanticus</i> , <i>Liparis inquilinus</i> ‡
Vertebrate	Chordata	BOLD:AAC3262	<i>Lycodes lavalaei</i> , <i>Lycodes polaris</i> ‡
Vertebrate	Chordata	BOLD:ACE9999	<i>Sebastes fasciatus</i> , <i>Sebastes norvegicus</i> , <i>Sebastes mentella</i>
Invertebrate	Arthropoda	BOLD:AAA9922	<i>Hyas coarctatus</i> †, <i>Hyas alutaceus</i>
Invertebrate	Cnidaria	BOLD:AAH8240	<i>Clavularia borealis</i> †, <i>Alcyonium siderium</i> , <i>Gersemia rubiformis</i>
Invertebrate	Cnidaria	BOLD:AAF8957	<i>Hormathia nodosa</i> , <i>Hormathia digitata</i>
Invertebrate	Cnidaria	BOLD:AAE6029	<i>Obelia longissima</i> , <i>Obelia dichotoma</i>
Invertebrate	Echinodermata	BOLD:AAB2709	<i>Leptasterias littoralis</i> †, <i>Leptasterias groenlandica</i>
Invertebrate	Mollusca	BOLD:ACR2733	<i>Buccinum hydrophanum</i> ‡, <i>Buccinum scalariforme</i>
Invertebrate	Mollusca	BOLD:AAF2554	<i>Nuculana tenuisulcata</i> , <i>Nuculana pernula</i>
Invertebrate	Porifera	BOLD:ACQ6272	<i>Polymastia thielei</i> , <i>Polymastia hemisphaerica</i>

† : species likely absent from region

‡ : species status uncertain

Table S5 Taxa sharing more than one BIN within the GSL-rl

Category	Phylum	Taxon	N BIN	BIN
Vertebrate	Chordata	<i>Alepocephalus bairdii</i>	2	BOLD:AAD7907, BOLD:ABY4903
Vertebrate	Chordata	<i>Chauliodus sloani</i>	2	BOLD:AAB1178, BOLD:AAB1179
Vertebrate	Chordata	<i>Liparis inquilinus</i>	2	BOLD:AAB4897, BOLD:ACA5502
Vertebrate	Chordata	<i>Lycodes spp</i>	2	BOLD:ABZ0973, BOLD:ACE4278
Vertebrate	Chordata	<i>Mola mola</i>	3	BOLD:ADK1672, BOLD:AAD0035, BOLD:AAD0034
Vertebrate	Chordata	<i>Synaphobranchus kaupii</i>	2	BOLD:AAA8286, BOLD:AAA8287
Invertebrate	Annelida	<i>Aphelochaeta filiformis</i>	2	BOLD:ACH1048, BOLD:ACH1047
Invertebrate	Annelida	<i>Bradabyssa villosa</i>	4	BOLD:ACH1761, BOLD:ABZ8261, BOLD:AAW7084, BOLD:ACM2001
Invertebrate	Annelida	<i>Gattyana cirrhosa</i>	2	BOLD:AAC0429, BOLD:AAG5103
Invertebrate	Annelida	<i>Harmothoe imbricata</i>	3	BOLD:ABZ7769, BOLD:AAB1625, BOLD:AAA9124
Invertebrate	Annelida	<i>Laonice cirrata</i>	3	BOLD:AAD5360, BOLD:AAU3353, BOLD:AAW9888
Invertebrate	Annelida	<i>Limnodrilus hoffmeisteri</i>	4	BOLD:ACB6505, BOLD:AAA6803, BOLD:ADX1968, BOLD:ACM1418
Invertebrate	Annelida	<i>Nicolea zostericola</i>	2	BOLD:AAC0291, BOLD:ADW6705
Invertebrate	Annelida	<i>Pista cristata</i>	2	BOLD:AAH9500, BOLD:ACP4603
Invertebrate	Annelida	<i>Polycirrus medusa</i>	3	BOLD:ACE5004, BOLD:AAF1233, BOLD:ABY4561
Invertebrate	Annelida	<i>Potamothrix moldaviensis</i>	2	BOLD:ACK5400, BOLD:ACK5110
Invertebrate	Annelida	<i>Terebellides spp</i>	2	BOLD:AAD6167, BOLD:AAF2172
Invertebrate	Annelida	<i>Terebellides stroemii</i>	3	BOLD:ACP0580, BOLD:AAE9393, BOLD:AAE9392
Invertebrate	Annelida	<i>Tubifex tubifex</i>	7	BOLD:ADJ9700, BOLD:ADK4845, BOLD:AAA6801, BOLD:ADX0052, BOLD:AAA6802, BOLD:ACV8476, BOLD:ACA9222
Invertebrate	Arthropoda	<i>Acartia clausi</i>	3	BOLD:ACB5005, BOLD:ACB4079, BOLD:AAT9961
Invertebrate	Arthropoda	<i>Acartia hudsonica</i>	2	BOLD:AAJ3150, BOLD:ACL9285
Invertebrate	Arthropoda	<i>Acartia tonsa</i>	2	BOLD:AAA5450, BOLD:AAA5453
Invertebrate	Arthropoda	<i>Anonyx spp</i>	2	BOLD:AAF0743, BOLD:AAA9156
Invertebrate	Arthropoda	<i>Anonyx ochoticus</i>	2	BOLD:ACA9971, BOLD:ADF0772
Invertebrate	Arthropoda	<i>Argis dentata</i>	2	BOLD:ADK3850, BOLD:AAC3186
Invertebrate	Arthropoda	<i>Calanus finmarchicus</i>	2	BOLD:ACP6217, BOLD:ACP5847
Invertebrate	Arthropoda	<i>Cancer irroratus</i>	2	BOLD:AAB7738, BOLD:ACD1651
Invertebrate	Arthropoda	<i>Centropages typicus</i>	2	BOLD:AAW6024, BOLD:ACM7827
Invertebrate	Arthropoda	<i>Chionoecetes opilio</i>	2	BOLD:AAA3656, BOLD:ABZ6341
Invertebrate	Arthropoda	<i>Crangon septemspinosa</i>	2	BOLD:AAB5942, BOLD:ACL7976
Invertebrate	Arthropoda	<i>Eualus gaimardii</i>	2	BOLD:AAE4771, BOLD:AAJ0881

Invertebrate	Arthropoda	<i>Gammarus setosus</i>	4	BOLD:ABZ4044, BOLD:ACE3104, BOLD:AAA3651, BOLD:AAA3653 BOLD:ADP7774, BOLD:ADP7773,
Invertebrate	Arthropoda	<i>Gammarus tigrinus</i>	3	BOLD:AAA4302
Invertebrate	Arthropoda	<i>Ischyrocerus anguipes</i>	2	BOLD:AAC5063, BOLD:AAC0651 BOLD:AAB4526, BOLD:AAG9842, BOLD:AAB4528, BOLD:ACF1644,
Invertebrate	Arthropoda	<i>Lebbeus polaris</i>	6	BOLD:AAB4529, BOLD:AAB4527
Invertebrate	Arthropoda	<i>Nymphon stroemi</i>	2	BOLD:ADF4183, BOLD:AAG4693 BOLD:AAG5172, BOLD:ACK9027,
Invertebrate	Arthropoda	<i>Oithona similis</i>	3	BOLD:ADU0358 BOLD:AAB6143, BOLD:AAB6146,
Invertebrate	Arthropoda	<i>Spirontocaris spinus</i>	4	BOLD:AAB6144, BOLD:ACR1068
Invertebrate	Cnidaria	<i>Cyanea capillata</i>	2	BOLD:AAP1190, BOLD:AAD3480
Invertebrate	Echinodermata	<i>Gorgonocephalus arcticus</i>	2	BOLD:AAC8875, BOLD:ABY6858 BOLD:AAI5449, BOLD:AAE7308,
Invertebrate	Echinodermata	<i>Gorgonocephalus eucnemis</i>	3	BOLD:ACQ0423 BOLD:AAD3463, BOLD:AAD3482,
Invertebrate	Echinodermata	<i>Henricia</i> spp	4	BOLD:AAB3569, BOLD:AAB9183
Invertebrate	Echinodermata	<i>Leptasterias polaris</i>	2	BOLD:ACE6383, BOLD:ADV1115
Invertebrate	Echinodermata	<i>Ophioscolex glacialis</i>	2	BOLD:ACV6677, BOLD:ACV6676
Invertebrate	Echinodermata	<i>Ophiura sarsii</i>	2	BOLD:ACO7183, BOLD:AAD3481
Invertebrate	Echinodermata	<i>Pseudarchaster parelii</i>	2	BOLD:AAD5295, BOLD:AAH8175
Invertebrate	Mollusca	<i>Ancula gibbosa</i>	2	BOLD:ADL9788, BOLD:ACW0116
Invertebrate	Mollusca	<i>Buccinum hydrophanum</i>	2	BOLD:ACR2733, BOLD:ADE0634 BOLD:ADY4343, BOLD:AAO4021, BOLD:ADS1898, BOLD:ADE1312, BOLD:ABA4155, BOLD:ADY4344,
Invertebrate	Mollusca	<i>Cryptonatica affinis</i>	7	BOLD:ADF1526
Invertebrate	Mollusca	<i>Euspira heros</i>	2	BOLD:ABW0896, BOLD:ABW0528
Invertebrate	Mollusca	<i>Hiatella arctica</i>	2	BOLD:ABA8494, BOLD:ADC8266
Invertebrate	Mollusca	<i>Lepeta caeca</i>	2	BOLD:AAX5488, BOLD:ACB8623
Invertebrate	Mollusca	<i>Mytilus edulis</i>	2	BOLD:AAA2184, BOLD:AAA2185 BOLD:ADM2144, BOLD:AEF0228,
Invertebrate	Mollusca	<i>Nuculana minuta</i>	3	BOLD:ACM0814
Invertebrate	Mollusca	<i>Serripes groenlandicus</i>	2	BOLD:AAH9555, BOLD:AAH9554
Invertebrate	Nemertea	<i>Cephalothrix spiralis</i>	2	BOLD:ACQ6349, BOLD:AAG3611 BOLD:ACH4809, BOLD:AAO8118,
Invertebrate	Porifera	<i>Cliona celata</i>	3	BOLD:ACH3517
Invertebrate	Porifera	<i>Halichondria panicea</i>	2	BOLD:ABV8568, BOLD:ADX0707

Table S6. Characteristics of the curated regional library of the Gulf of St. Lawrence (GSL-rl) version 1.0.

Number (N) of targeted species in the improved decision maker list and number (N) of targeted species with COI sequences for each of the three reliability categories, namely “Reliable”, “Unreliable due to BIN sharing” and “Unreliable due to gaps”.

Phylum	Class	N species	N species with COI sequences			
			Reliable	Unreliable due to BIN sharing	Unreliable due to gaps	Total
Cnidaria	Anthozoa	24	4	4	1	9
	Hydrozoa	8	5	2	0	7
	Scyphozoa	1	1	0	0	1
	Staurozoa	1	1	0	0	1
Porifera	Calcarea	7	0	0	0	0
	Demospongiae	39	11	0	8	19
	Hexactinellida	1	0	0	0	0
Bryozoa	Gymnolaemata	8	0	0	0	0
Annelida	Clitellata	8	4	0	2	6
	Polychaeta	99	19	0	24	43
Brachiopoda	Rhynchonellata	2	1	0	0	1
Nemertea	Hoplonemertea	4	3	0	0	3
	Palaeonemertea	1	1	0	0	1
	Pilidiophora	8	1	0	3	4
Sipuncula	Sipunculidea	1	1	0	0	1
Mollusca	Bivalvia	24	9	2	4	15
	Gastropoda	62	30	0	14	44
	Polyplacophora	5	2	0	1	3
Arthropoda	Hexanauplia	22	12	0	7	19
	Malacostraca	81	34	2	29	65
	Pycnogonida	15	1	0	1	2
Echinodermata	Asteroidea	31	12	0	6	18
	Echinoidea	3	3	0	0	3
	Holothuroidea	3	3	0	0	3
	Ophiuroidea	18	8	0	2	10
Chordata	Actinopterygii	150	116	13	12	141
	Appendicularia†	4	0	0	0	0
	Ascidiacea†	4	3	0	0	3
	Elasmobranchii	12	12	0	0	12
	Mammalia	3	3	0	0	3
	Myxini	1	1	0	0	1
	Petromyzonti	1	1	0	0	1

†Invertebrates included within the phylum Chordata

C. Creation of an eDNA metabarcoding dataset

Water samples of 2L were collected from scientific surveys in 2018 in coastal areas from the GSL, both at surface and bottom of the water column (n=61, Fig. S2). Negative field controls (n=2) were also obtained by pouring 2L of clean water at sampling site. Water samples and controls were then frozen at -20°C until their transport to Maurice Lamontagne Institute and afterwards at -40°C until filtration. Water samples and controls were thawed and filtered on glass fiber 47 mm, 1.2 and/or 10 µm pore size (Sigma Aldrich, MO, U.S.) in an ultraclean room (n = 63 filters) and 10 negative filtration controls were added. Qiagen Blood and Tissue Kit (QIAGEN, MD, U.S.) was used for DNA extraction from filters, including 7 extraction controls (elution in 80µl AE buffer). PCR were performed using 5 µl of a 1/10 dilution of the extract at Genome Québec with the primers mICOIntF (Leray et al. 2013) and jgHCO2198 (Geller et al. 2013), targeting a 313 pb section of the COI Folmer region. Three PCR negative controls were added for a total of 83 samples. Each amplification was then indexed and pooled, and then sequenced using Illumina MiSeq PE250 at Genome Québec.

Reads were demultiplexed by the sequencing facility, and quality control of raw reads was assessed using FastQC (Andrews 2010) and MultiQC (Ewels et al. 2016). Adapters were retrieved and removed using cutadapt (Martin 2011). Then, the dada2 R package (Callahan et al. 2016) was used to create exact sequence variants (ESV) from the data. Briefly, the *trimandfilter* function was first used for quality filtering with slightly modified parameters (truncQ = 10, maxEE = 2;), then error rate was computed with the *learnErrors* function. Reads were then dereplicated (*derepFastq* function), samples were inferred (*dada* function) then paired reads were merged (*mergePairs* function, minOverlap = 30, maxMismatch = 0). An ESV table was created (*makeSequenceTable* function), with samples as columns and ESV as rows, and chimeras were removed with the *removeBimeraDenovo* function. The ESV table was further corrected based on negative samples, where we have subtracted from the total number of observations the number observed in related negative controls (from sampling to PCR control). ESVs with only one read detected were removed. More details on the bioinformatics pipeline can be found on Fig. S3.

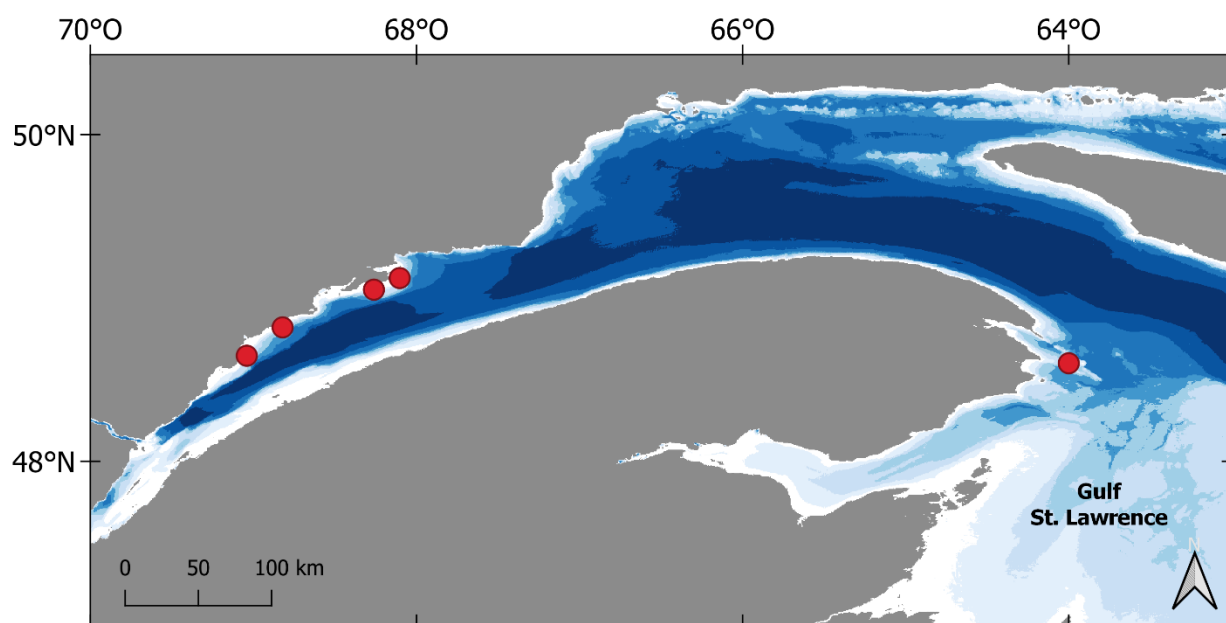


Figure S2 Sampling locations for the metabarcoding dataset in the St. Lawrence.

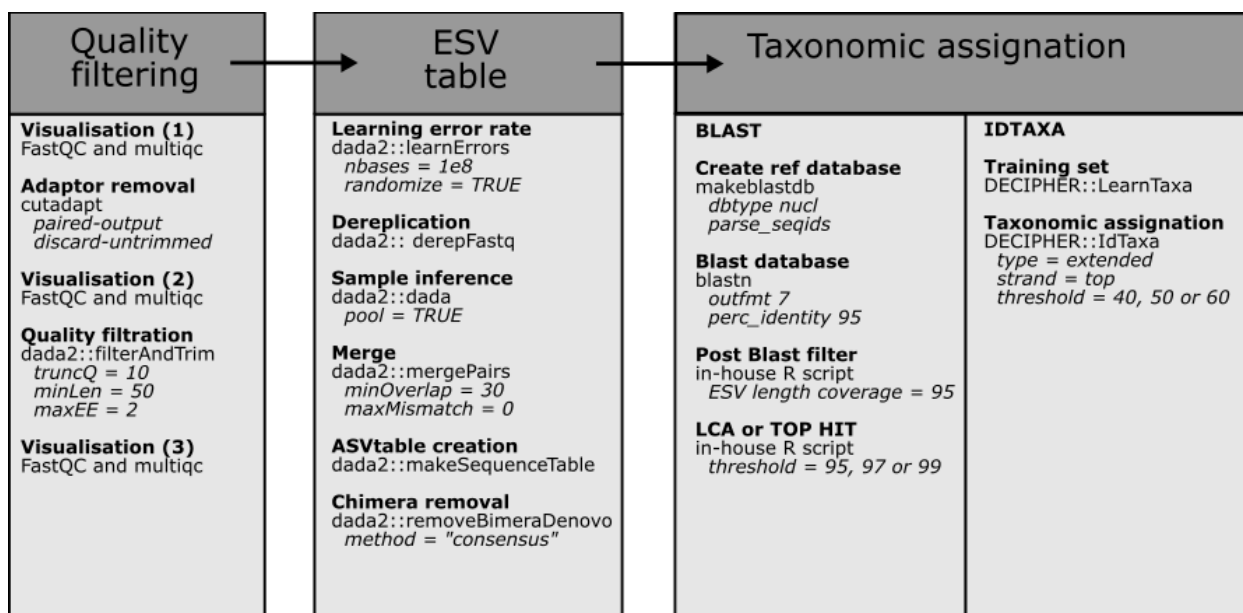


Figure S3 Schematic representation of the metabarcoding bioinformatics pipeline.

References

- Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13: 581–583. <https://doi.org/10.1038/nmeth.3869>
- 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>
- Chamberlain S (2020) bold: Interface to Bold Systems API. R package version 1.1.0. <https://CRAN.R-project.org/package=bold>.
- Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32: 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources* 13: 851–861. <https://doi.org/10.1111/1755-0998.12138>
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111–120. <https://doi.org/10.1007/BF01731581>
- Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology*: 1–14. <https://doi.org/10.1186/1742-9994-10-34>
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10. <https://doi.org/10.14806/ej.17.1.200>
- Nozères C (2017) Preliminary checklist of marine animal species of the Gulf of St. Lawrence, Canada, based on 4 sources. <https://doi.org/10.13140/RG.2.2.10056.62727>
- Paradis E, Schliep K (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- WoRMS Editorial Board (2020) World Register of Marine Species. <https://doi.org/10.14284/170>