MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 4: 35–46
DOI 10.3897/mbmg.4.55162

Research Article

Assessment of species gaps in DNA barcode libraries of non-indigenous species (NIS) occurring in European coastal regions

Sofia Duarte[1,2], **Pedro E. Vieira**[1,2], **Filipe O. Costa**[1,2]

1 *Centre of Molecular and Environmental Biology (CBMA), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*
2 *Institute of Science and Innovation for Bio-sustainability (IB-S), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

Corresponding author: Sofia Duarte (sduarte@bio.uminho.pt)

Academic editor: Baruch Rinkevich | **Received** 8 June 2020 | **Accepted** 23 July 2020 | **Published** 12 August 2020

Abstract

DNA metabarcoding has the capacity to bolster current biodiversity assessment techniques, including the early detection and monitoring of non-indigenous species (NIS). However, the success of this approach is greatly dependent on the availability, taxonomic coverage and reliability of reference sequences in genetic databases, whose deficiencies can potentially compromise species identifications at the taxonomic assignment step. In this study we assessed lacunae in availability of DNA sequence data from four barcodes (COI, 18S, rbcL and matK) for NIS occurring in European marine and coastal environments. NIS checklists were based on EASIN and AquaNIS databases. The highest coverage was found for COI for Animalia and rbcL for Plantae (up to 63%, for both) and 18S for Chromista (up to 51%), that greatly increased when only high impact species were taken into account (up to 82 to 89%). Results show that different markers have unbalanced representations in genetic databases, implying that the parallel use of more than one marker can act complimentarily and may greatly increase NIS identification rates through DNA-based tools. Furthermore, based on the COI marker, data for approximately 30% of the species had maximum intra-specific distances higher than 3%, suggesting that many NIS may have undescribed or cryptic diversity. Although completing the gaps in reference libraries is essential to make the most of the potential of the DNA-based tools, a careful compilation, verification and annotation of available sequences is fundamental to assemble large curated and reliable reference libraries that provide support for rigorous species identifications.

Key Words

BOLD, DNA barcode markers, gap-analysis, GenBank, marine and coastal ecosystems, non-indigenous species

Introduction

Marine and coastal habitats are among the most important, but also the most threatened ecosystems in the world, providing invaluable services, such as provisioning, supporting, regulation and cultural/aesthetic for human well-being (Solan et al. 2004; Rilov and Crooks 2009). Along with climate change, habitat destruction, overexploitation and pollution, the spread of non-indigenous species (NIS), for areas outside their natural occurrence range, is among the five most important direct drivers of biodiversity loss in European coastal regions (MEA 2005). Due to its position as a centre for international trade over centuries, Europe has a large number and diversity of well-established NIS in marine and coastal habitats (Keller et al. 2011; Katsanevakis et al. 2013a,b, 2014; Tsiamis et al. 2019). Many of these species are, or can become, invasive and displace and out-compete native species, leading to severe ecological changes threatening ecosystem integrity (Molnar et al. 2008; Rilov and Crooks 2009; Simberloff et al. 2013). Impacts include community structure alterations, biodiversity loss, habitat modification, harm to human health and economic losses (Keller et al. 2011).

While morphology-based identification of taxa has largely ensured the ascertainment of the current status of NIS occurring in coastal environments in Europe (Keller et al. 2011; Katsanevakis et al. 2013a, 2014; Tsiamis et al. 2019), this process is expertise-demanding, laborious and time consuming. It is also hardly applicable to some poorly studied communities, such as interstitial fauna,

Copyright *Sofia Duarte et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

PENSOFT

which may have moved large distances in ships through ballast waters or sediments (Carlton 1999; Ojaveer et al. 2014; Shang et al. 2019; Shaw et al. 2019). Particularly in aquatic systems, an accurate identification and detection of NIS may be prevented by the presence of life stages not amenable to morphological identification (i.e. eggs, propagules, planktonic larvae, juveniles), or because organisms are not large and distinctive (e.g. meiofauna, microalgae, zooplankton, protists) (Pochon et al. 2015; Zaiko et al. 2015a, 2015b, 2015c, 2016; Pagenkopp Lohan et al. 2016, 2017) or occur in low abundances (Darling and Blum 2007). In the case of NIS, the accuracy of species identifications is paramount since incorrect identifications can lead to biased outcomes and action against harmless species or inaction against problematic ones (Briski et al. 2016; Viard et al. 2019).

DNA-based methods, such as DNA barcoding (Hebert et al. 2003) and DNA metabarcoding (Hajibabaei 2012; Cristescu 2014), offer great promise for reliable species identifications in invasive ecology, having considerable potential to overcome some of the above-mentioned challenges and to improve the monitoring of NIS in marine and coastal ecosystems (Briski et al. 2011, 2016; Zaiko et al. 2015a, 2015b, 2015c; Abad et al. 2016; Brown et al. 2016; Miralles et al. 2016, 2018; Holman et al. 2019; Wood et al. 2019; Rey et al. 2020). In particular, DNA metabarcoding, which allies amplicon barcoding with high throughput sequencing may have a number of potential benefits over traditional methods, including the simultaneous processing of a large number of samples and the simultaneous identification of multiple taxa from various types of environmental samples (Hajibabaei 2012; Taberlet et al. 2012; Shokralla et al. 2012; Cristescu 2014), increased sensitivity and specificity, often revealing hidden diversity (Lindeque et al. 2013; Viard et al. 2019), as well as greater time and cost effectiveness (Briski et al. 2011, 2016; Pochon et al. 2015; Brown et al. 2016; von Ammon et al. 2018; Holman et al. 2019; Rey et al. 2020). In addition, a species may be detected at early developmental stages and before its dispersal and impact become readily apparent and irreversible (Pochon et al. 2015; Holman et al. 2019).

Efficient and accurate species identifications through DNA barcoding or DNA metabarcoding are dependent on reliable reference sequences libraries of known taxa (Briski et al. 2016; Miralles et al. 2016, 2018; Viard et al. 2019; Weigand et al. 2019). The unavailability or under-representation of some taxonomic groups in genetic databases may lead to biased results in biodiversity assessments through DNA-based tools and restrict the resolution and detection capacity of NIS at the taxonomic assignment step (Pochon et al. 2015; Briski et al. 2016; Chain et al. 2016; Zaiko et al. 2016; Lacoursière-Roussel et al. 2018; von Ammon et al. 2018; Rey et al. 2020). In Europe, comprehensive and reliable barcode reference libraries would be mandatory if DNA-based tools are to be implemented in biomonitoring in the context of the European Water Framework Directive (WFD, Directive 2000/60/EC) and the Marine Strategy Framework Directive (MSFD, Directive 2008/56/EC) (Leese et al. 2016, 2018; Hering et al. 2018; Pawlowski et al. 2018; Weigand et al. 2019). However, no recent attempt has been made for assessing the gaps in NIS sequences in publicly accessible databases (i.e. the number of species missing barcode sequences). To our best knowledge the most recent complete gap-analysis was conducted in 2016 (Briski et al. 2016) and although a recent one was performed in 2019 (Ardura 2019), it targeted only high-impact Arthropoda and Mollusca species.

In the current study, the gaps, for the most commonly used barcode markers in DNA-based studies for Animalia (COI and 18S), Chromista (COI, 18S and rbcL) and Plantae (COI, rbcL and matK), were analysed in the genetic databases GenBank and the Barcode of Life Data System (BOLD), with a focus on NIS occurring in European coastal regions by using updated lists retrieved from the European Alien Species Information Network (EASIN) (Katsanevakis et al. 2012) and the Information System on Aquatic Non-indigenous and Cryptogenic species (AquaNIS) (Olenin et al. 2014). This will allow a current appraisal of the status of NIS occurring in European marine and coastal regions that are missing DNA barcodes, and will permit researchers to develop actions to fulfil these gaps, in order to take the most of the potential of NIS identification through DNA-based tools. Actions developing innovative tools for biodiversity monitoring are mandatory for an effective management of biological invasions and the development of mitigation strategies to deal with increasing globalization and environmental change.

Methods

Species checklists

The lists of non-indigenous species (NIS) occurring in European marine coastal regions were assessed using the two most important databases that compile crucial information on non-indigenous species occurring in Europe, on 23th October 2019: the European Alien Species Information Network (EASIN) (https://easin.jrc.ec.europa.eu/easin) (Katsanevakis et al. 2012) and the Information System on Aquatic Non-indigenous and Cryptogenic species (AquaNIS) (http://www.corpi.ku.lt/databases/index.php/aquanis/) (Olenin et al. 2014). The EASIN catalogue, built by the European Commission's Joint Research Center (JRC), is based on an inventory of reported alien species in Europe that was produced by reviewing and standardizing existing information from 43 online databases and selected offline sources, which include the terrestrial and aquatic environments, with 34 of the databases reporting NIS in the marine environment (Katsanevakis et al. 2012). The AquaNIS is an advanced information system that deals in particular with aquatic NIS introduced to marine, brackish and coastal

freshwater environments of Europe and adjacent regions (Olenin et al. 2014). From the AquaNIS list we retrieved 1,172 species using as search criteria the "Recipient region" and the following sub-criteria; "Ocean": Atlantic + Arctic; "Ocean Region": NE Atlantic + Arctic; "LME": 20. Barents Sea + 21. Norwegian Sea + 22. North Sea + 23. Baltic Sea + 24. Celtic-Biscay Shelf + 25. Iberian Coastal + 26. Mediterranean Sea + 59. Iceland Shelf + 60. Faroe Plateau + 62. Black Sea + A1. Macaronesia. From the EASIN list we retrieved 1,566 species using the following criteria; "Environment": Marine + Oligohaline; "Impact": All (High + Low/Unknown); "Species status": Alien + Cryptogenic + Unknown; "Taxonomy": Animalia + Chromista + Plantae and "Pathways": Contaminant + Corridor + Escape + Release + Stowaway + Not assessed + Other + Unknown.

The taxonomic classification and name validation of the NIS compiled in the lists was made through the World Register of Marine Species (WoRMS) database (www.marinespecies.org) and the Algaebase (https://www.algaebase.org/). Both databases adopted the Cavalier-Smith's taxonomic classification system (Cavalier-Smith 1981). In this classification system, Chromista were established to include all chromophyte algae (those with chlorophyll c, not b) considered to have evolved by symbiogenetic enslavement of another eukaryote (a red alga), as well as heterotrophic protists that descended from them by loss of photosynthesis or entire plastids (Cavalier-Smith 1981, Ruggiero et al. 2015), which in our lists include the phyla: Bigyira, Cercozoa, Ciliophora, Cryptophyta, Foraminifera, Haptophyta, Myzozoa and Ochrophyta. All records without species level identifications were removed from the lists. Initially, to conduct the gap-analysis, alternate representations of the species names were maintained in the lists. Later, to simplify the display of the results, all replicated records were removed and the number of sequence hits were merged to accepted names. The final list included species belonging to three Kingdoms: Animalia, Chromista and Plantae. Bacteria and fungi were excluded from the AquaNIS list, because these taxa typically have uncertain status as non-indigenous or native. Birds and mammals were also excluded from the EASIN list.

Data-mining, processing and analyses

For each species in the lists, and within each taxonomic group (i.e. Animalia, Chromista and Plantae), the number of sequences available in the Barcode of Life Data System (BOLD) (www.barcodinglife.org) (Ratnasingham and Hebert 2007) was assessed using the package "bold" implemented in the R 3.6.0 software (R core Team 2019; www.r-project.org) (Chamberlain 2019). For retrieving the number of sequences available on GenBank (www.ncbi.nlm.nih.gov/Genbank) the package "rentrez" was used (Winter 2017). Only public records were retrieved because this information is available to all the users and details on private data cannot be easily accessed (e.g. genetic marker, sequence quality). The following markers were searched for each group: Animalia – COI and 18S; Chromista – COI, 18S and rbcL; Plantae – COI, rbcL and matK. The terms used to filter the sequences from BOLD were: for COI – "COI-5P"; for 18S – "18S" or "18Sa", for rbcL – "rbcL" or "rbcLa" and for matK – "matK". In GenBank, the terms used for the search were (suggested by GenBank for the studied loci): for COI - "COI[Gene] OR CO1[Gene] OR COXI[Gene] OR COX1[Gene] OR complete genome[All Fields] OR mitochondrial genome[All Fields])"; for 18S: "18S ribosomal rna[Title] OR 18S rRNA[Title] OR 18S small subunit ribosomal RNA[Title] OR 18S ribosomal rna[Gene] OR 18S rRNA[Gene] OR 18S small subunit ribosomal RNA[Gene]"; for rbcL: "rbcl[Gene] OR rubisco[Gene]"; and for matK: "maturase k[Gene] OR matk[Gene]". Only sequences with more than 500 base pairs were considered, since this is the minimum length required for a sequence to meet Barcode Compliance standards (Ratnasingham and Hebert 2007) and which has also been used in earlier gap-analysis studies of European aquatic invertebrates (Weigand et al. 2019).

The number of Barcode Index Numbers (BINs; proxy of Molecular operational taxonomic units – MOTUs) (Ratnasingham and Hebert 2013), for each species within each group, were retrieved from BOLD, based on the COI marker. Records co-occurring in both databases were detected through the presence of the tag "Mined from GenBank, NCBI" in BOLD's records and/or availability of a GenBank's accession number, which indicates that those BOLD records were mined/deposited from/to GenBank. A species was considered to be successfully barcoded for each marker if it had at least one compliant sequence in one of the searched databases. All the details of the bioinformatic pipeline, such as the scripts used for each taxonomic group and the markers searched, can be consulted at https://github.com/pedroemanuelvieira/NIS_Europe_GapAnalysis.

Results

Taxonomic composition of the lists

After removal of the records with taxonomic ranks higher than species level and replicated records, the final AquaNIS list had 1,120 species and the final EASIN checklist had 1,554 species (Fig. 1). The taxa in both lists belonged to three kingdoms (data in parentheses correspond to % in AquaNIS and EASIN lists, respectively): i) Animalia (66 and 76%), ii) Chromista (17 and 14%) and iii) Plantae (17 and 10%), comprising 28 phyla (Fig. 1) and 74 classes, 237 orders and 743 families (Suppl. material 1: Tables S1–S3). Both lists shared 667 species (435 Animalia, 125 Chromista and 107 Plantae), while 453 species were exclusive from the AquaNIS list (304 Animalia, 61 Chromista, 88 Plantae) and 887 species were exclusive from the EASIN list (745 Animalia, 99

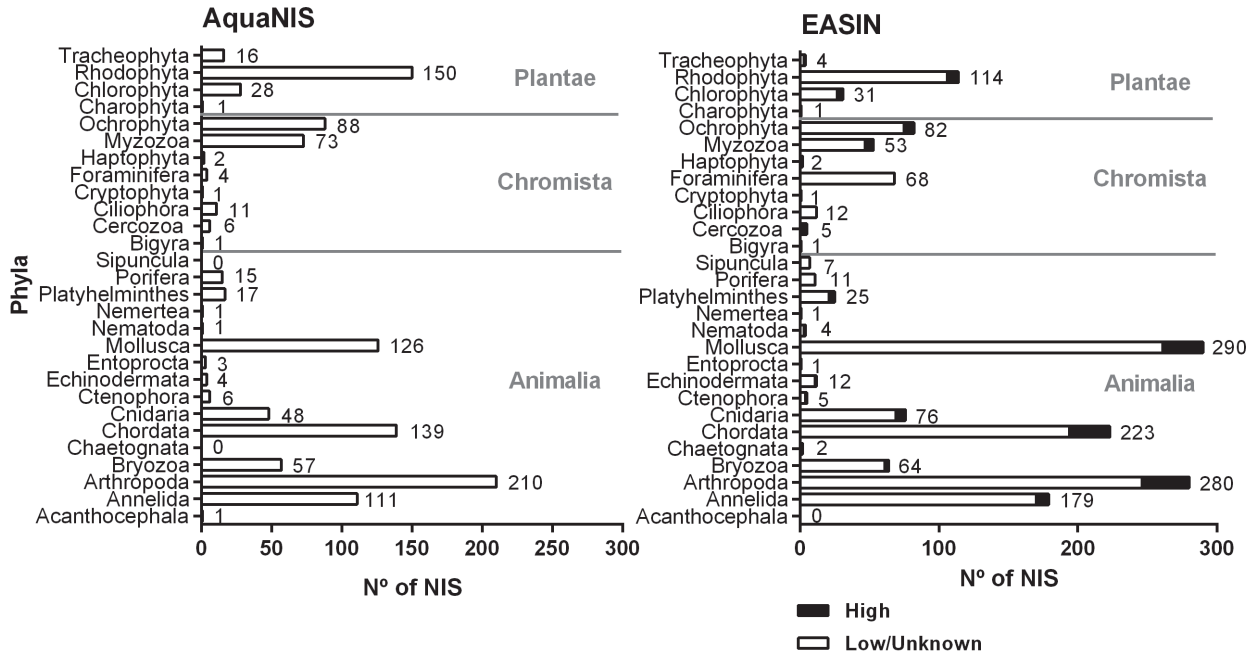

Figure 1. Taxonomic classification. Taxonomic distribution of the species from the AquaNIS and EASIN lists. Numbers on the right of each bar represent the total number of species per phyla. For the EASIN list the species were separated in high and low/unknown impact.

Chromista and 43 Plantae) (Suppl. material 1: Table S1 and Fig. S1). In the EASIN list, 1,294 species have the status of "alien", 174 species of "cryptogenic" and 86 species have the status of "questionable" (Suppl. material 1: Table S1) and 148 out of the 1,554 species (approximately 10% of the total number of species in the list) are classified as high impact species, with 118 species belonging to Animalia, 17 to Chromista and 13 to Plantae (Fig. 1; Suppl. material 1: Tables S1, S3).

Within Animalia, the most well represented phyla were Arthropoda (19 and 18%), Chordata (12 and 14%) and Mollusca (11 and 19%) (data in parentheses correspond to % in AquaNIS and EASIN lists, respectively). Within Chromista, these phyla were Ochrophyta (8%) and Myzozoa (7%), in the AquaNIS list, and Ochrophyta (5%) and Foraminifera (4%), in the EASIN list. Within Plantae, the most well represented phylum was Rhodophyta (13 and 7%, in AquaNIS and EASIN, respectively) in both lists, while the other phyla accounted for less than 5% of the total species (Fig. 1; Suppl. material 1: Table S3). The most well represented classes in AquaNIS and EASIN lists can be consulted in Table S4 (Suppl. material 1).

Gap analysis

For all analysed taxonomic groups (Animalia, Chromista and Plantae), a higher number of records was found on GenBank than on Public BOLD (Table 1). When considering at least the presence of one barcode sequence of at least one marker in at least one genetic database, a total barcode coverage between 58 and 68% and between 50 and 63% was found for the AquaNIS and EASIN list, respectively (Table 1). But the coverage varied considerably among the different taxonomic groups and barcode markers (Table 1). The highest coverage was found in both lists for Animalia and for the COI marker (63 and 51%, for AquaNIS and EASIN, respectively), for Chromista for the 18S marker in the AquaNIS list (51%) and for Plantae for the rbcL marker, in both lists (62 and 63%, for the AquaNIS and EASIN list, respectively) (Table 1). In addition, in particular for Animalia and for the 18S marker, the % of sequences represented by single barcode records in the databases (singletons) was relatively high (38 and 40% for the AquaNIS and EASIN lists, respectively).

For Animalia, in both lists, the phyla with the highest number of total records, taken into account all searched markers in both genetic databases, were Arthropoda (10,863 and 10,148), Chordata (12,478 and 11,808) and Mollusca (7,146 and 6,045, for the AquaNIS and EASIN lists, respectively) (Suppl. material 1: Tables S5, S6). In general, a higher coverage was found for the COI marker than for the 18S marker in both lists (Fig. 2A, B; Table 1), with the exception of Annelida (only for AquaNIS), Ctenophora, Platyhelminthes and Porifera, where a higher coverage was found for 18S (Fig. 2A, B). In the AquaNIS list, and within Animalia, most phyla had a barcode coverage higher than 50% for the COI marker, with the exception of Annelida (41%), Bryozoa (35%), Platyhelminthes (18%) and Porifera (33%), while no barcodes at all were found for Entoprocta (Fig. 2A). For 18S, a barcode coverage near to or higher than 50% was found for Annelida (46%), Arthropoda (49%), Cnidaria (58%), Ctenophora (83%) and Porifera (47%) (Fig. 2A). On the other hand, for the EASIN list, most of the phyla had a barcode coverage lower than 50% with the exception of Arthropoda (52%), Chaetognatha (50%), Chordata (89%), Echinoder-

Table 1. Overall barcode coverage. Overall barcode coverage for selected markers and % of singletons (i.e. species with only one representative sequence) on GenBank and Public BOLD for the major taxonomic NIS groups of the AquaNIS and EASIN lists.

Taxonomic group	Database	No. of species	Marker	No. of records		No. of barcoded species	Singletons (%)
				GenBank	Public BOLD	GenBank + Public BOLD (% barcode coverage)	
Animalia	AquaNIS	739	COI	25,242	21,013	465 (62.9)	9.0
			18S	1,821	7	331 (44.8)	37.8
			COI or 18S			500 (67.7)	
			COI+18S			296 (40.1)	
	EASIN	1,180	COI	23,889	19,154	604 (51.2)	11.9
			18S	1,750	6	352 (29.8)	40.3
			COI or 18S			650 (55.1)	
			COI+18S			306 (25.9)	
Chromista	AquaNIS	186	COI	833	431	60 (32.3)	18.3
			18S	1,190	0	95 (51.1)	18.9
			rbcL	623	224	56 (30.1)	28.6
			COI or 18S or rbcL			108 (58.1)	
			COI+18S+rbcL			30 (16.1)	
	EASIN	224	COI	801	308	51 (22.8)	17.6
			18S	1,123	0	79 (35.3)	19.0
			rbcL	549	209	53 (23.7)	24.5
			COI or 18S or rbcL			113 (50.4)	
			COI+18S+rbcL			18 (8.0)	
Plantae	AquaNIS	195	COI	1,002	494	75 (38.5)	18.7
			rbcL	1,358	718	121 (62.1)	21.5
			matK	67	17	13 (6.7)	23.1
			COI or rbcL or matK			125 (64.1)	
			COI+rbcL+matK			3 (1.5)	
	EASIN	150	COI	802	394	55 (36.7)	12.7
			rbcL	1,216	653	94 (62.7)	16.0
			matK	30	20	5 (3.3)	0
			COI or rbcL or matK			94 (62.7)	
			COI+rbcL+matK			0	

mata (67%) and Nematoda (75%), for COI, and Ctenophora (80%) and Nematoda (75%), for 18S (Fig. 2B).

For Chromista, in both lists, Ochrophyta was the phyla which included the highest number of total records, taking into account all searched markers in both genetic databases (2,188 and 1,983, for the AquaNIS and EASIN respectively) (Suppl. material 1: Tables S5, S6). The barcode coverage among the different markers differed depending on the target phyla (Fig. 2C, D), except for Haptophyta, for which a barcode coverage of 50% and 100% was found for the 3 searched markers (COI, 18S, rbcL), in the AquaNIS and EASIN lists, respectively. For COI, the barcode coverage was always lower than 50% for all remaining analysed phyla, while no COI sequences were found for Bigyra, Cercozoa, Cryptophyta and Foraminifera, in both lists (Fig. 2C, D). Cryptophyta was also not represented by any 18S or rbcL sequence in BOLD and GenBank, for both lists, but it is represented in both lists by only one NIS. The 18S was the most well represented marker in both lists, in particular for Cercozoa (50 and 60%), Ciliophora (46 and 42%), Myzozoa (45 and 42%) and Ochrophyta (58 and 48%, for AquaNIS and EASIN, respectively), while Ochrophyta were better represented by rbcL sequences (59 and 58%, for AquaNIS and EASIN, respectively), but not the other phyla (Fig. 2C, D).

For Plantae, in both lists, Rhodophyta was the phyla which included the highest number of total records in both genetic databases, taking into account all markers (2,362 and 1,931, for the AquaNIS and EASIN lists, respectively) (Suppl. material 1: Tables S5, S6) and similarly to Chromista, the barcode coverage differed among the different markers and the target phyla (Fig. 2E, F). A better barcode coverage was generally found for the rbcL marker and for the four analysed phyla (equal or higher than 60%), in both lists (Fig. 2E, F), while COI sequences were found for Chlorophyta, Rhodophyta and Tracheophyta in the AquaNIS list (25 to 43%), but only for Chlorophyta and Rhodophyta in the EASIN list (23 and 42%, respectively). MatK sequences were exclusively found for Charophyta and Tracheophyta (100 and 75%, respectively, for the AquaNIS, and 100%, for both phyla in the EASIN list) (Fig. 2E, F).

Gap-analysis for high impact species

Considering only the high impact species from the EASIN list, the gap was much lower for all analysed groups and barcode markers, than for the full lists (Fig. 3; Table 2). When considering at least the presence of one barcode sequence of at least one marker in at least one genetic database, a total barcode coverage between 82 and 93% was found for the high impact species (Table 2). In general, coverage was higher than 50% for all analysed groups and barcode markers, with the exception of rbcL for Chromista (35%) and matK for Plantae (8%) (Table 2).

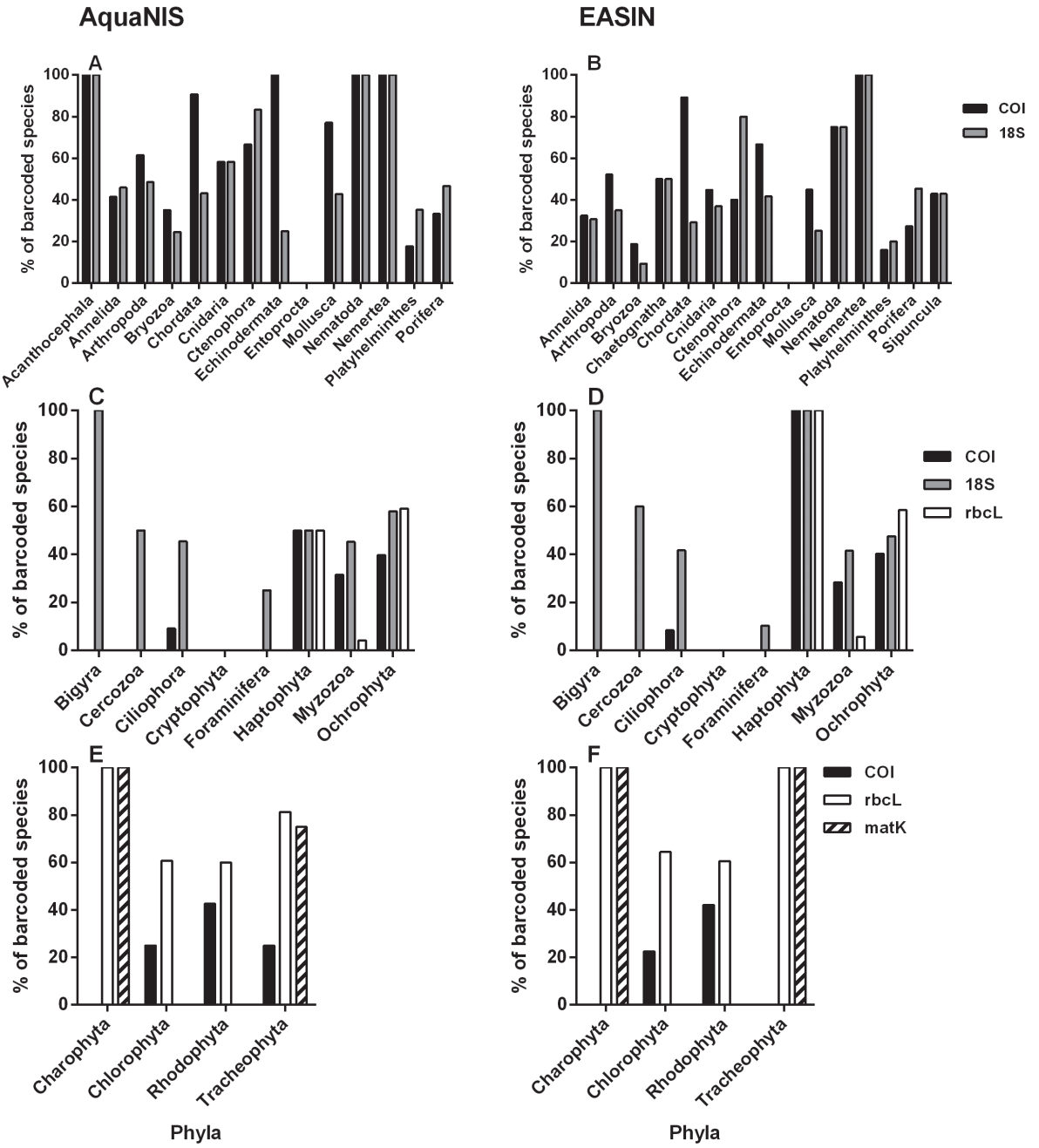

Figure 2. Gap-analysis. Barcode coverage (%) of each searched marker in Public BOLD and GenBank for AquaNIS (left panel) and EASIN (right panel) lists, for each taxonomic group (phyla) within Animalia (**A, B**), Chromista (**C, D**) and Plantae (**E, F**).

For Animalia, the highest number of total records, considering all searched markers in both genetic databases, was found for Arthropoda, Mollusca and Chordata (2,797 to 4,595) (Suppl. material 1: Table S7). At the phyla level a barcode coverage of 100% was found for Ctenophora, Echinodermata and Nematoda, for both markers, and also for Chordata and Cnidaria, for COI (Fig. 3A). For Chromista and Plantae, the highest number of total records were found for Myzozoa (209) and Rhodophyta (210), respectively (Suppl. material 1: Table S7). Within Chromista, a barcode coverage of 100% was found for Haptophyta, for all analysed markers, and for Myzozoa for 18S (Fig. 3B), while for Plantae, for Tracheophyta, for both rbcL and matK (Fig. 3C).

Most remaining phyla containing high impact species, within Animalia, still had a barcode coverage higher than 50%, with the exception of Platyhelminthes for COI (25%), but that was well represented with 18S sequences (75%), and Bryozoa and Chordata for 18S (33 and 45%, respectively) (Fig. 3A). Within Chromista, no COI or rbcL sequences were found for Cercozoa, but 67% of the species were represented with 18S sequences (Fig. 3B). Myzozoa was poorly represented with rbcL sequences (33%), but well represented

Table 2. Overall barcode coverage for high impact species. Overall barcode coverage for selected markers and % of singletons (i.e. species with only one representative sequence) on GenBank and Public BOLD for high impact species (EASIN).

Taxonomic group	No. of species	Marker	No. of records		No. of barcoded species	Singletons (%)
			GenBank	Public BOLD	GenBank + Public BOLD (% barcode coverage)	
Animalia	118	COI	9,968	8,033	105 (89.0)	3.8
		18S	648	3	77 (65.2)	22.1
		COI or 18S			110 (93.2)	
		COI+18S			72 (61.0)	
Chromista	17	COI	75	35	10 (58.8)	10.0
		18S	198	0	14 (82.3)	7.1
		rbcL	62	25	6 (35.3)	16.7
		COI or 18S or rbcL			14 (82.3)	
		COI+18S+rbcL			5 (29.4)	
Plantae	13	COI	84	30	7 (53.8)	14.3
		rbcL	155	94	11 (84.6)	18.2
		matK	1	5	1 (7.7)	0.0
		COI or rbcL or matK			11 (84.6)	
		COI+rbcL+matK			0	

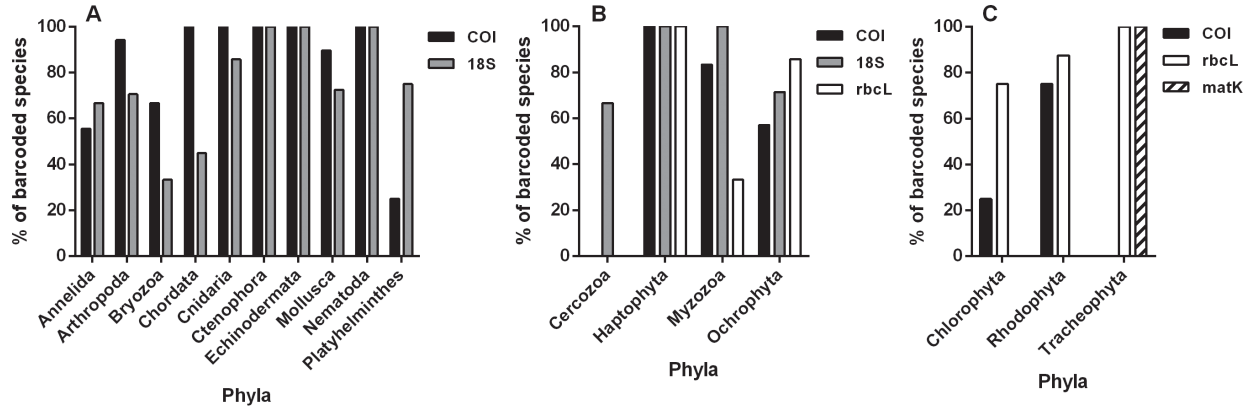

Figure 3. Gap-analysis for high impact species. Barcode coverage (%) of each searched marker in Public BOLD and GenBank for high impact species of the EASIN list for each taxonomic group (phyla) within Animalia (**A**), Chromista (**B**) and Plantae (**C**).

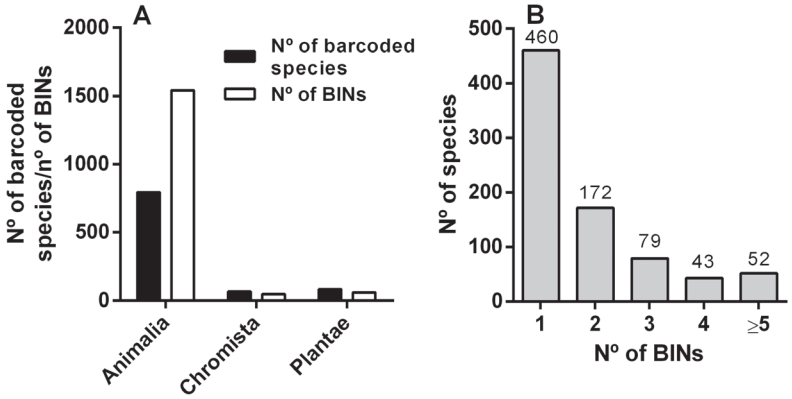

Figure 4. Barcode Index Numbers. Number of barcoded species and number of BINs, based on the COI marker, for each taxonomic group (**A**) and number of species with 1 to ≥5 BINs for the total number of barcoded species found in both lists (**B**). On (**B**) the numbers above bars indicate the number of species.

either with 18S (100%) or COI barcodes (83%) (Fig. 3B). Concerning Plantae, matK sequences were found only for Tracheophyta (100%), and COI for Chlorophyta (25%) and Rhodophyta (75%) (Fig. 3C), while rbcL was the most well represented marker among all Plantae phyla containing high impact species (75 to 100%). More details of the gap-analyses can be found in Suppl. material 1: Tables S5–S10.

Barcode Index Number (BIN) analysis and intra-specific distances

Based on the COI marker, a total number of 1,649 Barcode Index Numbers (BINs) were found for the two lists: 1,541 for Animalia, 48 for Chromista and 60 for Plantae (Fig. 4A). Most species were represented by one BIN, but a high proportion of species have 2 or more BINs (346 species, 37%) (Fig. 4B), including 52 species that were assigned to 5 or more BINs. Species with maximum intra-specific distances higher than 3% constituted 30% of the dataset (16% if we consider the mean intraspecific distances) (Suppl. material 1: Table S11) and included some high impact species such as *Marenzelleria viridis* (Annelida), *Acartia* (*Acanthacartia*) *tonsa* and *Penaeus japonicus* (Arthropoda), *Bugula neritina* (Bryozoa), *Herdmania momus* and *Microcosmus squamiger* (Chordata), *Gonionemus vertens* (Cnidaria), *Acanthaster planci* (Echinodermata), *Xenostrobus securis* and *Arcuatula senhousia* (Mollusca) and *Gyrodactylus salaris* (Platyhelminthes) (Suppl. material 1: Table S11).

Discussion

Our study brings to the forefront two main considerations: first, reference libraries still lack representative sequences for many NIS with extreme cases in some groups, and second, some NIS can be categorised as possible cryptic species. Both these cases may critically impair the detection of NIS and therefore, the current capability for NIS detection and monitoring using molecular tools.

Although the gaps (i.e., NIS still missing barcode sequences) were similar in both lists, the values of missing barcodes clearly differed among taxonomic groups and the barcode markers searched. In both lists the gap was highest for Chromista. In these lists, Chromista include Foraminifera, Myzozoa and Ochrophyta as dominant phyla, that can harbour very small sized species, such as small protists and diatoms and for which obtaining voucher specimens to generate sequences to deposit in genetic databases may be challenging. It has been reported that smaller organisms may have greater invasion opportunities in coastal ecosystems (Ruiz et al. 2000; Pagenkopp Lohan et al. 2016, 2017), but that can be hard to detect by using traditional morphological approaches (Pagenkopp Lohan et al. 2016, 2017). Thus, DNA-based tools are essential for its early detection and accurate identification in recipient ecosystems and fulfilling the gaps in barcode reference libraries is extremely essential for Chromista.

On the other hand, we found a lower gap for Animalia and Plantae. The gaps in BOLD and GenBank were recently analysed for the taxa frequently used in the WFD and the MSFD, under the scope of the COST Action DNAqua-Net (Weigand et al. 2019), and the authors also found that barcode coverage varied strongly among taxonomic groups. In general, groups that were actively targeted in barcode projects were well represented in the barcode libraries, while others have fewer records. Our results support this trend. Under the scope of the public project "WG1.8 Marine Bio-Surveillance" deposited in BOLD, 12 and 17 projects were dedicated to Animalia and Plantae, respectively, with a total of 1,516 sequences, while only 4 projects were dedicated to Chromista, comprising only 105 sequences. In both lists, the phyla with the highest number of records in the two searched genetic databases include a high number of species having a high impact in the environment or species with high economic value (i.e. Chordata, Arthropoda, Mollusca, Ochrophyta, Rhodophyta). These species are generally the focus of a greater number of studies and thus, may display a higher trend of sequence deposition in genetic databases (Briski et al. 2011, 2016; Pyšek et al. 2008; Trebitz et al. 2015; Ardura 2019). In fact, we found among the top ten species with the highest number of sequence records either high impact species, such as *Callinectes sapidus* and *Anguillicoloides crassus*, or species with high economic value such as *Mytilus trossulus*, *Prionace glauca*, *Cyprinus carpio* and *Oncorhynchus mykiss*.

Our results were somewhat discrepant from those obtained in a previous report where the gaps in BOLD and GenBank were analysed for aquatic NIS compiled from literature (Briski et al. 2016). By 2016, 76% of the species in the list, compiled by Briski and colleagues for aquatic NIS (n=1,383), had at least one sequence of 6 searched markers in BOLD or GenBank. In addition, the authors predicted that if the rate of sequence deposition in both genetic databases followed a linear trend, they would expect that all aquatic NIS in their list would be sequenced by 2030. In our study, completion seems to be still a bit far off with only 65% of the species in the AquaNIS and 55% in the EASIN list having at least one of the searched barcode markers in BOLD or GenBank. These disparities probably originated from different compliance criteria and mismatching of the species lists used in the analyses, which in the case of Briski and colleagues (2016) consisted on a list of NIS occurring at a worldwide scale. In addition, only barcode sequences higher than 500 bp were considered in the current gap analysis, while Briski et al. (2016) did not mention if any length filter has been applied to their sequences search. New NIS and new introductions into different recipient regions are reported every year and NIS status can also change (from unknown status to cryptogenic or alien), suggesting that this is a work that needs to be performed from time to time. Fortunately, currently, there are specific databases dedicated to this, and that are constantly updated, such as EASIN and AquaNIS (Katsanevakis et al. 2012; Olenin et al. 2014), which greatly facilitates this task. In addition, the R-based bioinformatic pipeline, developed in our study to retrieve the information relative to each marker from the two genetic databases, will enable to conduct this task effectively and in an automated way when needed (i.e. every time that significant updates are made in the lists).

As above-mentioned, for each taxonomic group, the gap clearly differed among the barcode markers searched.

For Animalia, most phyla were well represented with COI sequences in GenBank and BOLD, but Annelida, Ctenophora, Platyhelminthes and Porifera were better represented with 18S sequences. Within Chromista most phyla were better represented with 18S, but for instance Ochrophyta, which includes brown algae and diatoms, was an exception to this pattern, with the barcode coverage being greatest for rbcL. For Plantae, most phyla were better represented with rbcL sequences. Thus, the simultaneous use of more than one marker can act complimentarily and may greatly increase NIS identification rates through DNA-based tools. Recent studies have highlighted the advantage of using both 18S and COI markers for invasive species detection; the 18S for detecting a much broader range of taxa and the COI for discriminating between many metazoan species (Borrell et al. 2017; von Ammon et al. 2018; Stefanni et al. 2018; Holman et al. 2019; Wood et al. 2019; Rey et al. 2020). In addition, the concomitant use of the rbcL and COI allowed the detection of diatoms and green and yellow algae, in ballast water of a vessel crossing the Atlantic Ocean, which otherwise would remain highly underestimated if communities have been only targeted with COI (Zaiko et al. 2015b).

Approximately 37% of the species displayed more than one BIN, and many of these species displayed mean- and maximum-intraspecific distances higher than 3%, suggesting that many NIS may display hidden diversity or cryptic diversity, which may further complicate taxonomic assignment using DNA-based tools (Viard et al. 2019). In addition, many species were represented by singletons in the genetic databases, thereby preventing detection of possible intraspecific variability or cryptic diversity. At the moment, at least to our knowledge, no dedicated reference sequences database exists for NIS. Ideally, and also suggested by the great proportion of species displaying multiple BINs and high intraspecific distances in the current study, this reference database should cover the full sweep of species in the target ecosystem, with a balanced representation of specimens across each species distribution range in both native and recipient locations, to account for the possible regional variability in targeted barcode genes. In addition, database incompleteness can be somewhat overcome by the addition of DNA sequences for local species. Abad et al. (2016) was able to increase 2 times more the success of the taxonomic assignment of plankton species in the estuary of Bilbao (Spain), by generating DNA barcodes for local species before conducting a metabarcoding-based study.

A closer look at the list of barcoded species with attributed BINs, in particular for COI and Animalia, indicated that many of them displayed discordant BINs (i.e. different species sharing the same BIN), possibly due to incorrect taxonomic assignments of numerous species, that have been repeatedly used in databases without a proper validation. A careful inspection in these BINs would be needed in order to check for potential artefacts such as misidentifications, incomplete taxonomy or sequences that were deposited under different synonyms. Incorrect species identifications could either artificially inflate or depress the number of NIS in an ecosystem, and lead to misdirecting limited resources against harmless species or inaction against problematic ones (Bax et al. 2001; Simberloff 2009). Lacoursière-Roussel et al. (2018) identified *Acartia tonsa* through eDNA metabarcoding, in water samples collected at two Canadian ports, a potential invader that has been previously recorded in the ecoregions of ports connected to Churchill. However, the current available COI sequences for *A. tonsa* form several distinct clades, some of which cluster with *A. hudsonica*, which rose the possibility that the eDNA sequences assigned to *A. tonsa* may belong to the native *A. hudsonica*. Very recently, by examining public databases Viard et al. (2019) also found sequences of *Botrylloides diegensis* erroneously assigned to *B. leachii*. This observation has major implications as the introduced *B. diegensis* can be misidentified as a putatively native species. Unfortunately, these database errors can be frequent, as also suggested by the high proportion of discordant BINs found in the current study, and can delay the implementation of DNA metabarcoding in NIS surveillance in coastal ecosystems.

Final remarks

Although completing the gaps in reference libraries is essential to make the most of the potential of DNA-based tools in NIS surveillance in coastal ecosystems, correct species attribution (by morphology-based methods) and proper management of sequence deposition and voucher storage is vital to preserve correct connections between morphological and molecular data (Briski et al. 2016). This can be particularly challenging for small-sized species that lack unambiguous morphological traits to use in taxonomic diagnosis, such as some particular groups within Chromista (e.g. Myzozoa), for which a higher gap was found in genetic databases. In addition, a careful compilation, verification and annotation of each database record is fundamental to assemble large, curated and reliable reference libraries that provide support for rigorous species identifications through DNA-based tools (Viard et al. 2019; Weigand et al. 2019; Fontes et al. 2020; Leite et al. 2020). This need is particularly acute for the phylogenetically diverse NIS, for which there is highly dispersed data that needs to be compiled and verified. Once this need is fulfilled, the adoption of DNA-based tools for accurate NIS detection and monitoring in marine and coastal ecosystems will very likely accelerate.

Acknowledgements

This work was supported by national funds through the Portuguese Foundation for Science and Technology (FCT, I.P.) in the scope of the project “NIS-DNA: Early detection and monitoring of non-indigenous species

(NIS) in coastal ecosystems based on high-throughput sequencing tools" (PTDC/BIA-BMA/29754/2017). We are also grateful to two reviewers for comments and suggestions that improved the manuscript.

References

Abad D, Albaina A, Aguirre M, Laza-Martínez A, Uriarte I, Uriarte A, Villate F, Estonba A (2016) Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. Marine Biology 163:149. https://doi.org/10.1007/s00227-016-2920-0

Ardura A (2019) Species-specific markers for early detection of marine invertebrate invaders through eDNA methods: gaps and priorities in GenBank as database example. Journal for Nature Conservation 47: 51–57. https://doi.org/10.1016/j.jnc.2018.11.005

Bax N, Carlton JT, Mathews-Amos A, Haedrich RL, Howarth FG, Purcell JE, Rieser A, Gray A (2001) The control of biological invasions in the world's oceans. Conservation Biology 15: 1234–1246. https://doi.org/10.1111/j.1523-1739.2001.99487.x

Borrell YJ, Miralles L, Do Huu H, Mohammed-Geba K, Garcia-Vasquez E (2017) DNA in a bottle-Rapid metabarcoding survey for early alerts of invasive species in ports. PLoS ONE 12: e0183347. https://doi.org/10.1371/journal.pone.0183347

Briski E, Cristescu ME, Bailey SA, MacIsaac HJ (2011) Use of DNA barcoding to detect invertebrate invasive species from diapausing eggs. Biological Invasions 13: 1325–1340. https://doi.org/10.1007/s10530-010-9892-7

Briski E, Ghabooli S, Bailey SA, MacIsaac HJ (2016) Are genetic databases sufficiently populated to detect non-indigenous species? Biological Invasions 18: 1911–1922. https://doi.org/10.1007/s10530-016-1134-1

Brown EA, Chain FJJ, Zhan A, MacIsaac HJ, Cristescu ME (2016) Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports (2016) Diversity and Distributions 22: 1045–1059. https://doi.org/10.1111/ddi.12465

Carlton JT (1999) The scale and ecological consequences of biological invasions in the world's oceans. In: Sandlund OT, Schei PJ, Viken A (Eds) Invasive species and biodiversity management. Kluwer Academic Publishers, Dordrecht, 195–212. https://doi.org/10.1007/978-94-011-4523-7_13

Cavalier-Smith T (1981) Eukaryote kingdoms: seven or nine? Biosystems. 14: 461–481. https://doi.org/10.1016/0303-2647(81)90050-2

Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME (2016) Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. Diversity and Distributions 22: 493–504. https://doi.org/10.1111/ddi.12427

Chamberlain S (2019) bold: Interface to Bold Systems API. R package version 0.9.0. https://CRAN.R-project.org/package=bold

Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology and Evolution 29: 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Darling JA, Blum MJ (2007) DNA-based methods for monitoring invasive species: a review and prospectus. Biological Invasions 9: 751–765. https://doi.org/10.1007/s10530-006-9079-4

European Commission (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. https://eur-lex.europa.eu/eli/dir/2000/60/oj

European Commission (2008) Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). https://eur-lex.europa.eu/eli/dir/2008/56/oj

Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2020) BAGS: an automated Barcode, Audit & Grade System for DNA barcode reference libraries. Authorea. https://doi.org/10.22541/au.159135669.99561145

Hajibabaei M (2012) The golden age of DNA metasystematics. Trends in Genetics 28:535–537. https://doi.org/10.1016/j.tig.2012.08.001

Hebert PDN, Ratnasingham S, DeWaard JR (2003) Barcoding animal life, cytochrome c oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London, series B: Biological Sciences 270: S596–S599. https://doi.org/10.1098/rsbl.2003.0025

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

Holman LE, de Bruyn M, Creer S, Carvalho G, Robidart J, Rius M (2019) Detection of introduced and resident marine species using environmental DNA metabarcoding of sediment and water. Scientific Reports 9: 11559. https://doi.org/10.1038/s41598-019-47899-7

Katsanevakis S, Bogucarskis K, Gatto F, Vandekerkhove J, Deriu I, Cardoso AC (2012) Building the European Alien Species Information Network (EASIN): a novel approach for the exploration of distributed alien species data. BioInvasions Records 4: 235–245. https://doi.org/10.3391/bir.2012.1.4.01

Katsanevakis S, Gatto F, Zenetos A, Cardoso AC (2013a) How many marine aliens in Europe? Management of Biological Invasions 4: 37–42. https://doi.org/10.3391/mbi.2013.4.1.05

Katsanevakis S, Wallentinus I, Zenetos A, Leppäkoski E, Çinar ME, Oztürk B, Grabowski M, Golani D, Cardoso AC (2014) Impacts of invasive alien marine species on ecosystem services and biodiversity: a pan-European review. Aquatic Invasions 9: 391–423. https://doi.org/10.3391/ai.2014.9.4.01

Katsanevakis S, Zenetos A, Belchior C, Cardoso AC (2013b) Invading European Seas: assessing pathways of introduction of marine aliens. Ocean and Coastal Management 76: 64–74. https://doi.org/10.1016/j.ocecoaman.2013.02.024

Keller RP, Geist J, Jeschke JM, Kühn I (2011) Invasive species in Europe: ecology, status, and policy. Environmental Sciences Europe 23: 23. https://doi.org/10.1186/2190-4715-23-23

Lacoursière-Roussel A, Howland K, Normandeau E, Grey EK, Archambault P, Deiner K, Lodge DM, Hernandez C, Leduc N, Bernatchez L (2018) eDNA metabarcoding as a new surveillance approach for coastal Arctic biodiversity. Ecology and Evolution 8: 7763–7777. https://doi.org/10.1002/ece3.4213

Leese F, Altermatt F, Bouchez A, Ekrem T, Hering D, Meissner K, Mergen P, Pawlowski J, Piggott J, Rimet F, Steinke D, Taberlet P, Weigand A, Abarenkov K, Beja P, Bervoets L, Björnsdóttir S, Boets P, Boggero A, Bones A, Borja Á, Bruce K, Bursić V, Carlsson J, Čiampor F,

Čiamporová-Zatovičová Z, Coissac E, Costa F, Costache M, Creer S, Csabai Z, Deiner K, DelValls Á, Drakare S, Duarte S, Eleršek T, Fazi S, Fišer C, Flot J, Fonseca V, Fontaneto D, Grabowski M, Graf W, Guðbrandsson J, Hellström M, Hershkovitz Y, Hollingsworth P, Japoshvili B, Jones J, Kahlert M, Kalamujic Stroil B, Kasapidis P, Kelly M, Kelly-Quinn M, Keskin E, Kõljalg U, Ljubešić Z, Maček I, Mächler E, Mahon A, Marečková M, Mejdandzic M, Mircheva G, Montagna M, Moritz C, Mulk V, Naumoski A, Navodaru I, Padisák J, Pálsson S, Panksep K, Penev L, Petrusek A, Pfannkuchen M, Primmer C, Rinkevich B, Rotter A, Schmidt-Kloiber A, Segurado P, Speksnijder A, Stoev P, Strand M, Šulčius S, Sundberg P, Traugott M, Tsigenopoulos C, Turon X, Valentini A, van der Hoorn B, Várbíró G, Vasquez Hadjilyra M, Viguri J, Vitonytė I, Vogler A, Vrålstad T, Wägele W, Wenne R, Winding A, Woodward G, Zegura B, Zimmermann J (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Research Ideas and Outcomes 2: e11321. https://doi.org/10.3897/rio.2.e11321

Leese F, Bouchez A, Abarenkov K, Altermatt F, Borja Á, Bruce K, Ekrem T, Čiampor F, Čiamporová-Zaťovičová Z, Costa FO, Duarte S, Elbrecht V, Fontaneto D, Franc A, Geiger MF, Hering D, Kahlert M, Kalamujić Stroil B, Kelly M, Keskin E, Liska I, Mergen P, Meissner K, Pawlowski J, Penev L, Reyjol Y, Rotter A, Steinke D, van der Wal B, Vitecek S, Zimmermann J, Weigand AM (2018) Chapter Two – Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-Net COST action. In: Bohan DA, Dumbrell AJ, Woodward G, Jackson M (Eds) Advances in Ecological Research, Next Generation Biomonitoring: Part 1. Academic Press, 63–99. https://doi.org/10.1016/bs.aecr.2018.01.001

Leite B, Vieira PE, Teixeira MAL, Lobo-Arteaga J, Hollatz C, Borges LMS, Duarte S, Troncoso JS, Costa FO (2020) Gap-analysis and annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. Regional Studies in Marine Science 36: 101307. https://doi.org/10.1016/j.rsma.2020.101307

Lindeque PK, Parry HE, Harmer RA, Somerfield PJ, Atkinson A (2013) Next generation sequencing reveals the hidden diversity of zooplankton assemblages. PLoS ONE 8: e81327. https://doi.org/10.1371/journal.pone.0081327

MEA [Millennium Ecosystem Assessment] (2005) Ecosystems and Human Wellbeing: Biodiversity Synthesis. World Resources Institute, Washington, DC. https://www.millenniumassessment.org/documents/document.356.aspx.pdf

Miralles L, Ardura A, Arias A, Borrell YJ, Clusa L, Dopico E, Hernandez de Rojas A, Lopez B, Muñoz-Colmenero M, Roca A, Valiente AG, Zaiko A, Garcia-Vazquez E (2016) Barcodes of marine invertebrates from North Iberian ports: native diversity and resistance to biological invasions. Marine Pollution Bulletin 112: 183–188. https://doi.org/10.1016/j.marpolbul.2016.08.022

Miralles L, Ardura A, Clusa L, Garcia-Vasquez E (2018) DNA barcodes of Antipode marine invertebrates in Bay of Biscay and Gulf of Lion ports suggest new biofouling challenges. Scientific Reports 8: 16214. https://doi.org/10.1038/s41598-018-34447-y

Molnar JL, Gamboa RL, Revenga C, Spalding MD (2008) Assessing the global threat of invasive species to marine biodiversity. Frontiers in Ecology and the Environment 6: 485–492. https://doi.org/10.1890/070064

Ojaveer H, Galil BS, Minchin D, Olenin S, Amorim A, Canning-Clode J, Chainho P, Copp GH, Gollasch S, Jelmert A, Lehtiniemi M, McKenzie C, Mikuš J, Miossec L, Occhipinti-Ambrogi A, Pecarevic M, Pederson J, Quilez-Badia G, Wijsman JWM, Zenetos A (2014) Ten recommendations for advancing the assessment and management of non-indigenous species in marine ecosystems. Marine Policy 44: 160–165. https://doi.org/10.1016/j.marpol.2013.08.019

Olenin S, Narščius A, Minchin D, David M, Galil B, Gollasch S, Marchini A, Occhipinti-Ambrogi A, Ojaveer H, Zaiko A (2014) Making non-indigenous species information systems practical for management and useful for research: an aquatic perspective. Biological Conservation 173: 98–107. https://doi.org/10.1016/j.biocon.2013.07.040

Pagenkopp Lohan KM, Fleischer RC, Carney KJ, Holzer KK, Ruiz GM (2016) Amplicon-based pyrosequencing reveals high diversity of protistan parasites in ships' ballast water: implications for biogeography and infectious diseases. Microbial Ecology 71: 530–542. https://doi.org/10.1007/s00248-015-0684-6

Pagenkopp Lohan KM, Fleischer RC, Carney KJ, Holzer KK, Ruiz GM (2017) Molecular characterisation of protistan species and communities in ships' ballast water across three U.S. coasts. Diversity and Distributions 23: 680–691. https://doi.org/10.1111/ddi.12550

Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L, Beja P, Boggero A, Borja A, Bouchez A, Cordier T, Domaizon I, Feio MJ, Filipe AF, Fornaroli R, Graf W, Herder J, van der Hoorn B, Jones JI, Sagova-Mareckova M, Moritz C, Barquín J, Piggott JJ, Pinna M, Rimet F, Rinkevich B, Sousa-Santos C, Specchia V, Trobajo R, Vasselon V, Vitecek S, Zimmerman J, Weigand A, Leese F, Kahlert M (2018) The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Science of the Total Environment 637–638: 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Pyšek P, Richardson DM, Pergl J, Jarošík V, Sixtová Z, Weber E (2008) Geographical and taxonomic biases in invasion ecology. Trends in Ecology and Evolution 23: 237–244. https://doi.org/10.1016/j.tree.2008.02.002

Pochon X, Zaiko A, Hopkins GA, Banks JC, Wood SA (2015) Early detection of eukaryotic communities from marine biofilm using high-throughput sequencing: an assessment of different sampling devices. Biofouling 31: 241–251. https://doi.org/10.1080/08927014.2015.1028923

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). 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: e66213. https://doi.org/10.1371/journal.pone.0066213

Rey A, Basurko OC, Rodriguez-Ezpeleta N (2020) Considerations for metabarcoding-based port biological baseline surveys aimed at marine nonindigenous species monitoring and risk assessments. Ecology and Evolution 10: 2452–2465. https://doi.org/10.1002/ece3.6071

Rilov G, Crooks J (2009) Biological Invasions in marine ecosystems: Ecological, management and geographic perspectives. Springer, Berlin-Heidelberg, 109–116. https://doi.org/10.1007/978-3-540-79236-9

Ruggiero MA, Gordon DP, Orrell TM, Bailly N, Bourgoin T, Brusca RC, Cavalier-Smith T, Guiry MD, Kirk PM (2015) A higher level classification of all living organisms. PLoS ONE 10: e0119248. https://doi.org/10.1371/journal.pone.0119248

Ruiz GM, Rawlings TK, Dobbs FC, Drake LA, Mullady T, Huq A, Colwell RR (2000) Global spread of microorganisms by ships. Nature 408: 49–50. https://doi.org/10.1038/35040695

Shang L, Hu Z, Deng Y, Liu Y, Zhai X, Chai Z, Liu X, Zhan Z, Dobbs FC, Tang YZ (2019) Metagenomic sequencing identifies highly diverse assemblages of dinoflagellate cysts in sediments from ships' ballast tanks. Microorganisms 7: 250. https://doi.org/10.3390/microorganisms7080250

Shaw JLA, Weyrich LS, Hallegraeff G, Cooper A (2019) Retrospective eDNA assessment of potentially harmful algae in historical ship ballast tank and marine port sediments. Molecular Ecology 28: 2476–2485. https://doi.org/10.1111/mec.15055

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21: 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Simberloff D (2009) We can eliminate invasions or live with them. Successful management projects. Biological Invasions 11: 149–157. https://doi.org/10.1007/s10530-008-9317-z

Simberloff D, Martin J-L, Genovesi P, Maris V, Wardle D, Aronson J, Courchamp F, Galil B, Garcia-Berthou E, Pascal M, Pyšek P, Sousa R, Tabacchi E, Vilà M (2013) Impacts of biological invasions: what's what and the way forward. Trends in Ecology and Evolution 28: 58–66. https://doi.org/10.1016/j.tree.2012.07.013

Solan M, Cardinale BJ, Dowing AL, Engelhardt KAM, Ruesink JL, Srivastava DS (2004) Extinction and ecosystem function in the marine benthos. Science 306: 1177–1189. https://doi.org/10.1126/science.1103960

Stefanni S, David S, Borne D, de Olazabal A, Juretić T, Pallavicini A, Tirelli V (2018) Multi-marker metabarcoding approach to study mesozooplankton at basin scale. Scientific Reports 8: 12085. https://doi.org/10.1038/s41598-018-30157-7

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Trebitz AS, Hoffman JC, Grant GW, Billehus TM, Pilgrim EM (2015) Potential for DNA-based identification of Great Lakes fauna: match and mismatch between taxa inventories and DNA barcode libraries. Scientific Reports 5: 12162. https://doi.org/10.1038/srep12162

Tsiamis K, Palialexis A, Stefanova K, Gladan ŽN, Skejić S, Despalatović M, Cvitković I, Dragičević B, Dulčić J, Vidjak O, Bojanić N, Žuljević A, Aplikioti M, Argyrou M, Josephides M, Michailidis N, Jakobsen HH, Staehr PA, Ojaveer H, Lehtiniemi M, Massé C, Zenetos A, Castriota L, Livi S, Mazziotti C, Schembri PJ, Evans J, Bartolo AG, Kabuta SH, Smolders S, Knegtering E, Gittenberger A, Gruszka P, Kraśniewski W, Bartilotti C, Tuaty-Guerra M, Canning-Clode J, Costa AC, Parente MI, Botelho AZ, Micael J, Miodonski JV, Carreira GP, Lopes V, Chainho P, Barberá C, Naddafi R, Florin AB, Barry P, Stebbing PD, Cardoso AC (2019) Non-indigenous species refined national baseline inventories: A synthesis in the context of the European Union's Marine Strategy Framework Directive. Marine Pollution Bulletin 145: 429–435. https://doi.org/10.1016/j.marpolbul.2019.06.012

Viard F, Roby C, Turon X, Bouchemousse S, Bishop J (2019) Cryptic diversity and database errors challenge non-indigenous species surveys: an illustration with Botrylloides spp. in the English Channel and Mediterranean Sea. Frontiers in Marine Science 6: 615. https://doi.org/10.3389/fmars.2019.00615

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: 16290. https://doi.org/10.1038/s41598-018-34541-1

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 F, Dijkstra KD, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher J, 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

Winter DJ (2017) rentrez: an R package for the NCBI eUtils API. The R Journal 9: 520–526. https://doi.org/10.32614/RJ-2017-058

Wood SA, Pochon X, Laroche O, von Ammon U, Adamson J, Zaiko A (2019) A comparison of droplet digital polymerase chain reaction (PCR), quantitative PCR and metabarcoding for species-specific detection in environmental DNA. Molecular Ecology Resources 19: 1407–1419. https://doi.org/10.1111/1755-0998.13055

Zaiko A, Martinez JL, Ardura A, Clusa L, Borrell YJ, Samuiloviene A, Roca A, Garcia-Vasquez E (2015a) Detecting nuisance species using NGST: Methodology shortcomings and possible application in ballast water monitoring. Marine Environmental Research 112: 64–72. https://doi.org/10.1016/j.marenvres.2015.07.002

Zaiko A, Martinez JL, Schmidt-Petersen J, Ribicic D, Samuiloviene A, Garcia-Vasquez E (2015b) Metabarcoding approach for the ballast water surveillance – an advantageous solution or an awkward challenge? Marine Pollution Bulletin 92: 25–34. https://doi.org/10.1016/j.marpolbul.2015.01.008

Zaiko A, Samuiloviene A, Ardura A, Garcia-Vazquez E (2015c) Metabarcoding approach for nonindigenous species surveillance in marine coastal waters. Marine Pollution Bulletin 100: 53–59. https://doi.org/10.1016/j.marpolbul.2015.09.030

Zaiko A, Schimanski K, Pochon X, Hopkins GA, Goldstien S, Floerl O, Wood SA (2016) Metabarcoding improves detection of eukaryotes from early biofouling communities: implications for pest monitoring and pathway management. Biofouling 32: 671–684. https://doi.org/10.1080/08927014.2016.1186165

Supplementary material 1

Supplementary figures and tables used to analyse the data

Authors: Sofia Duarte, Pedro E. Vieira, Filipe O. Costa

Data type: Lists of species, taxonomic classification, gap-analyses

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.4.55162.suppl1