Exhaustive reanalysis of barcode sequences from public repositories highlights ongoing misidentifications and impacts taxa diversity and distribution.

Antoine Fort, Marcus Mchale, Kevin Cascella, Philippe Potin, Marie-Mathilde Perrineau, Philip D Kerrison, Elisabete da Costa, Ricardo Calado, Maria Do Rosário Domingues, Isabel Costa Azevedo, et al.

DR. ANTOINE FORT (Orcid ID : 0000-0002-2210-7234)

DR. RONAN SULPICE (Orcid ID : 0000-0002-6113-9570)

Article type : Resource Article

Title

Exhaustive reanalysis of barcode sequences from public repositories highlights ongoing misidentifications and impacts taxa diversity and distribution.

Short running title:

Improved taxonomy through barcode reanalysis.

Antoine Fort[1,*], Marcus McHale[1], Kevin Cascella[2], Philippe Potin[2], Marie-Mathilde Perrineau[3], Philip D. Kerrison[3], Elisabete da Costa[4], Ricardo Calado[5], Maria do Rosário Domingues[4], Isabel Costa Azevedo[6], Isabel Sousa-Pinto[6,7], Claire Gachon[3,8], Adrie van der Werf[9], Willem de Visser[9], Johanna E Beniers[9], Henrice Jansen[9], Michael D Guiry[10] and Ronan Sulpice[1].

[1]National University of Ireland - Galway, Plant Systems Biology Lab, Ryan Institute & MaREI Centre for Marine, Climate and Energy, School of Natural Sciences, Galway, H91 TK33, Ireland.

[2]CNRS, Sorbonne Université Sciences, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff, France.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13453

[3]Scottish Association for Marine Science (SAMS), Scottish Marine Institute, PA37 1QA Oban, United Kingdom.

[4]CESAM & LAQV-REQUIMTE, Department of Chemistry, University of Aveiro, Campus Universitário de Santiago, 3810-193, Aveiro, Portugal.

[5]ECOMARE & CESAM, Departamento de Biologia & Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal.

[6]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, Novo Edifício do Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, s/n, 4450-208 Matosinhos, Portugal.

[7]Department of Biology, Faculty of Sciences, University of Porto. R. do Campo Alegre s/n, Porto Portugal.

[8]UMR 7245 - Molécules de Communication et Adaptation des Micro-organismes, Muséum National d'Histoire Naturelle, CNRS, CP 54, 57 rue Cuvier, 75005 Paris, France.

[9]Wageningen University & Research, Wageningen The Netherlands.

[10]AlgaeBase, Ryan Institute, National University of Ireland, Galway, H91 TK33, Ireland.

*Corresponding author:
Dr Antoine Fort, National University of Ireland Galway, Plant Systems Biology Lab, Ryan Institute, Plant and AgriBiosciences Research Centre, School of Natural Sciences, Galway, H91 TK33, Ireland. Email: antoine.fort@nuigalway.ie, tel: +35391493964.

Abstract

Accurate species identification often relies on public repositories to compare the barcode sequences of the investigated individual(s) with taxonomically assigned sequences. However, the accuracy of identifications in public repositories is often questionable, and the names originally given are rarely updated. For instance, species of the Sea Lettuce (*Ulva* spp.; Ulvophyceae, Ulvales, Ulvaceae) are

frequently misidentified in public repositories, including herbaria and gene banks, making species identification based on traditional barcoding unreliable. We DNA barcoded 295 individual distromatic foliose strains of *Ulva* from the North-East Atlantic for three loci (*rbc*L, *tuf*A, ITS1). Seven distinct species were found, and we compared our results with all worldwide *Ulva* spp sequences present in the NCBI database for the three barcodes *rbc*L, *tuf*A and the ITS1. Our results demonstrate a large degree of species misidentification, where we estimate that 24 to 32% of the entries pertaining to foliose species are misannotated and provide an exhaustive list of NCBI sequences reannotations. An analysis of the global distribution of registered samples from foliose species also indicates possible geographical isolation for some species, and the absence of *U. lactuca* from Northern Europe. We extended our analytical framework to three other genera, *Fucus*, *Porphyra* and *Pyropia* and also identified erroneously labelled accessions and possibly new synonymies, albeit less than for *Ulva* spp. Altogether, exhaustive taxonomic clarification by aggregation of a library of barcode sequences highlights misannotations and delivers an improved representation of species diversity and distribution.

Keywords: Sea lettuce, *Ulva*, DNA barcoding, Aquaculture, Phylogeny.

1. Introduction

Species identification of biological specimens is paramount for assessing the diversity of ecosystems (Johannesson & Andre, 2006), identify invasion events (Dunbar et al., 2021; Estoup & Guillemaud, 2010), and qualify the distribution of species of interest (Mendez, Rosenbaum, Subramaniam, Yackulic, & Bordino, 2010). While morphological characteristics can be used for species identification (Dugon, Black, & Arthur, 2012), precise species identification often relies on the analysis of "barcode" sequences, which are small standardized genetic loci used for taxonomic identification of the samples (Valentini, Pompanon, & Taberlet, 2009). Indeed, morphological characters can be a poor indicator of the underlying complexity of the genetic diversity within a genus (Packer, Gibbs, Sheffield, & Hanner, 2009).

For example, due to the phenotypic plasticity of the genus *Ulva* —the type genus of the *Ulvophyceae*, *Ulvales* and *Ulvaceae*— in response to environmental factors, and relatively subtle morphological differences between species (Hofmann, Nettleton, Neefus, & Mathieson, 2010; Malta, Draisma, & Kamermans, 1999), DNA barcoding is necessary to attribute species names to specimens, even for the most common species. DNA barcoding for the purpose of identifying specimens relies on the amplification and sequencing of specific loci in the genome. In plants and algae, it is often through chloroplast markers such as *rbc*L and *tuf*A, but also nuclear markers such as parts of the 45S rRNA repeats [most commonly the Internal Transcribed Spacer 1 (ITS1)] (Coat et al., 1998; Fort, Guiry, & Sulpice, 2018; Fort et al., 2019; Fort, Mannion, Fariñas-Franco, & Sulpice, 2020; Miladi et al., 2018; O'Kelly, Kurihara, Shipley, & Sherwood, 2010). The sequences obtained from those barcodes are then compared with sequences associated to species names which are publicly available in repositories, such as the National Center for Biotechnology Information (NCBI).

Typically, NCBI sequences with high percentage identity compared with the query sequence are considered as belonging to the same species and used as reference for phylogenetic trees when no statistical inference of species delimitation is used (Heesch et al., 2009; Saunders & Kucera, 2010; Steinhagen, Karez, & Weinberger, 2019). The risk in such case is that the species attributed to the matching sequences present in the NCBI can be erroneous, leading to the misidentification of the investigated individual. Indeed, the taxonomic information in the NCBI is not always accurate, and often contains "putative" species names (Garg, Leipe, & Uetz, 2019), erroneous classifications (Chowdhary, Singh, Singh, Khurana, & Meis, 2019; Nasehi et al., 2019), or non-updated species names following nomenclature adjustments (Hughey, Gabrielson, Maggs, & Mineur, 2021a). Therefore, improving the nomenclature and taxonomic classification of sequences of any genus of interest requires a careful exhaustive reanalysis of barcodes sequences, to ensure accurate classification of new specimens, and to provide an updated list of reannotations.

Here, we deployed such an analytical framework to revisit the phylogeny of *Ulva spp.*, a genetically diverse group of green macroalgal species ubiquitous in the world's ocean, brackish and even in freshwater environments. Over 400 *Ulva* names have been coined of which about 90 are

currently recognised as taxonomically valid (Guiry & Guiry 2021), many of which are uncommon or rare and only about 25 are frequently reported (Guiry & Guiry 2021). The morphology of *Ulva* species can be grouped into two general types, one containing foliose "sheet-like" species (distromatic foliose blades commonly known as "Sea Lettuce"), and another with tubular or partially tubular thalli (monostromatic tubes formerly recognized as the genus *Enteromorpha*). However, the phenotypic plasticity between tubular and foliose morphotypes is not solely genetic, but can be based on both abiotic and biotic factors (Wichard et al., 2015). We generated DNA barcodes (*rbc*L, *tuf*A, ITS1) on 185 strains of distromatic foliose *Ulva* from the North East Atlantic, and used data and species delimitation from our previous study containing another 110 strains (Fort, McHale, et al., 2021), as a primer for large-scale phylogenetic analysis of all *Ulva* sequences for the three common barcodes present in the NCBI database. The main goal of this study is to develop an analytical framework allowing to highlight the extent of misannotations in the sequences of any taxa of interest, taking as proof of concept the case of distromatic foliose *Ulva* species. We provide a detailed view of the phylogenetic relationships and possible misannotations between all sequences in the NCBI database, and propose readjustment for misannotated NCBI accessions, a list of appropriate reference vouchers for large foliose species, and a nomenclature adjustment between certain *Ulva* species. Finally, we employed the same analytical framework for three other seaweed genera, *Fucus*, *Porphyra* and *Pyropia* and identified clades containing misannotations and potential new synonymies.

2. Materials and methods

2.1 Foliose *Ulva* sample collection and DNA extraction

We collected individual thalli from foliose *Ulva* individuals with a thalli area > 1000 mm^2 in 34 sites in Ireland, Brittany (France), Spain, Portugal, the United Kingdom and the Netherlands between June 2017 and September 2019. The list of strains and associated metadata are available in **Table S1**. A total of 185 strains were collected for this study. On collection, samples were placed in clip-seal bags filled with local seawater and sent to Ireland in cold insulated boxes. On arrival, thalli were thoroughly washed with artificial seawater and a ~50 mm^2 piece of biomass collected and placed in screw caps tubes (Micronic). The tubes were immediately flash-frozen in liquid nitrogen and stored at

–80 °C. Then, samples were freeze dried, ground to a fine powder using a ball mill (QIAGEN TissueLyser II), and ~5 mg of powder used for DNA extraction, using the magnetic-beads protocol described in Fort et al. (2018).

2.2 DNA amplification and Sanger sequencing

The extracted DNA was amplified using three different primers combinations to obtain partial sequences for the nuclear 45S rRNA repeats (ITS1), as well as the chloroplast *rbc*L and *tuf*A barcodes. The primers used in this study are available in **Table S2**, and originate from (Heesch et al., 2009) and (Saunders & Kucera, 2010) for *rbc*L and *tuf*A, respectively. The ITS1 primers were designed from the dataset obtained in Fort, McHale, et al. (2021), and used in Fort, Linderhof, et al. (2021). PCR amplification was performed in 25 µL reaction volume containing 1 µL of undiluted DNA, 0.65 µL of 20 pmol forward and reverse primers, 9.25 µL of miliQ water and 12.5 µL of MyTaq Red mix (Bioline). The PCR protocol used 35 cycles of denaturation at 95°C for 30 s, annealing at 60 °C for 30 s and extension at 72 °C for 30 s. PCR products were precipitated using 2.5 volumes of 100% EtOH and 0.1 volume of 7.5M ammonium acetate and incubated on ice for 30 min. Pellets were centrifuged at 4,000 *g* for 30 min at 4°C, and washed twice with 75% EtOH. Finally, PCR amplicons were sent to LGC Genomics GmbH (Germany) for Sanger sequencing using the forward primer for each barcode.

2.3 Dataset compilation for phylogenetic analyses

Our phylogenetic analysis aimed to consider all sequences attributed to *Ulva* species (foliose and tubular) in the NCBI database, including tubular and partially tubular species, and detect any evidence of species misannotation therein. We designed an analysis pipeline that could be used in any other taxa of interest, summarised in **Fig. 1**. Command line codes and links to download the software used are available in **File S1**. We downloaded all available sequences in the NCBI for ITS, *rbc*L and *tuf*A (as of 13th of July 2020), in addition to the sequences from our previous study(Fort, McHale, et al., 2021). The search keywords were as follows: "*Ulva* [organism] AND internal transcribed" for ITS

sequences, "*Ulva* [organism] AND *rbc*L [gene] AND plastid [filter]" for *rbc*L sequences, and "*Ulva* [organism] AND tufa [gene] AND plastid [filter]" for *tuf*A sequences. This search strategy yielded 1,679 ITS sequences (1,975 in total including this study and Fort, McHale, et al. (2021), 1,432 *rbc*L sequences (1,732 in total) and 1,114 *tuf*A sequences (1,393 sequences in total).

NCBI entries that did not contain species information (containing "*Ulva sp*" as organism) were then removed from the dataset, by selecting all sequences not containing "*Ulva sp*" in their title, and using Samtools faidx (Li et al., 2009) to extract their corresponding sequences. This filtering yielded 1,726, 1,312 and 1,321 sequences for ITS1, *rbc*L and *tuf*A, respectively. Sequences were then aligned using MAFFT (Katoh, Rozewicki, & Yamada, 2019) using the default settings for *rbc*L and *tuf*A, and the iterative FFT-NS-i method for the ITS1 alignment, due to the numerous gaps present. Because each study might amplify a slightly different portion of the barcodes due to the use of different primers, we then removed nucleotide positions that were absent in i) more than 60% of the sequences using Trimal (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) -gt 0.4 for *rbc*L and *tuf*A, and ii) in more than 91% of the sequences for ITS1 (Trimal -gt 0.09). This step effectively trimmed the 5' and 3' ends of the alignment as to retain informative nucleotides, thereby avoiding large missing positions due to the use of different primers in different studies. Sequences containing more than 50% unknown bases in the trimmed alignments were then removed using Trimal (trimal -Seqoverlap 50) (for *rbc*L and *tuf*A), and more than 70% unknown bases for the ITS1 alignment (trimal -seqoverlap 70). The use of two different filtering methods between the organellar barcodes (*rbc*L and *tuf*A) and ITS1 was because the ITS1 alignment contains gaps that are biologically relevant (the ITS1 length varies between species), while *rbc*L and *tuf*A coding sequences generally do not vary in length, but only in sequence. The filtering steps yielded final alignments containing 1,245 sequences (270 bp), 1,062 sequences (1,231 bp) and 1,320 sequences (801 bp) for ITS1, *rbc*L and *tuf*A, respectively. The 5' and 3' gaps introduced by the presence of missing positions in some of the sequences due to missing data were modified into "n" (i.e., unknown) bases. The missing nucleotides at the beginning and end of the sequences were due to the use of different primers (or sequencing length), and not to genetically relevant differences.

The *Fucus* and *Poyphyra+Pyropia* datasets were generated as above, using the search terms "Fucus [organism] AND (COI [gene] OR COX1[gene])", "Fucus [organism] AND internal transcribed", and

"(porphyra [organism] OR pyropia [organism] OR neoporphyra [organism] OR neopyropia [organism]) AND (COX1[gene] OR COI[gene])". The final alignment datasets contained 174 sequences for *Fucus* COI, 452 sequences for Fucus nrRNAITS and 1,296 sequences for *Porphyra+Pyropia* COI/COX. We kept entries with no taxonomically accepted names to encompass all genetic information available for those clades.

2.4 Phylogenetic analyses

We used both maximum likelihood and Bayesian MCMC phylogenetic analyses for the ITS1, *rbc*L and *tuf*A datasets to create maximum likelihood and Bayesian trees for each barcode. First, the best evolutionary model for each of the three alignments was determined based on their AIC (Akaike Information Criterion) score using jModeltest 2 (Darriba, Taboada, Doallo, & Posada, 2012; Posada & Buckley, 2004). For all three alignments, General Time Reversible + Gamma distribution + Proportion of invariants sites (GTR + G + I) was deemed the most appropriate. Maximum likelihood trees were obtained using RAxML-NG (Kozlov, Darriba, Flouri, Morel, & Stamatakis, 2019) using the "--all" option (20 maximum likelihood inferences, then bootstrap trees). Bootstrapping was stopped automatically using a MRE-based Bootstopping Test (Pattengale, Alipour, Bininda-Emonds, Moret, & Stamatakis, 2010) once reaching convergence values below 0.03. Bootstrap values were computed using the "--bs-metric tbe" option, representing Transfer Bootstrap Expectation (TBE) values, expected to produce higher support for large trees with hundreds of sequences (Lemoine et al., 2018), compared with classical Felsenstein Bootstrap Proportions (FBP). Bayesian MCMC analyses were performed using MrBayes with MPI support (Ronquist et al., 2012), with a varying number of generations between the three datasets, until the average standard deviation of split frequencies reached a maximum of 0.05, and estimated sample sizes (ESSs) were higher than 200 for all parameters.

For species delimitation, we used the same method as per Fort et al. (2019) and Fort, McHale, et al. (2021), with a General Mixed Yule Coalescent model (Fujisawa & Barraclough, 2013; Pons et al., 2006) in BEAST, and 50 million Markov Chain Monte Carlo (MCMC), using the BEAGLE library for decreasing computational time (Suchard & Rambaut, 2009). Convergence was confirmed in

Tracer (Rambaut, Drummond, Xie, Baele, & Suchard, 2018), with an ESS score > 200 for all relevant parameters. Species delimitation was performed using the Rncl and Splits packages in R (Fujisawa & Barraclough, 2013). All trees were visualised using Figtree (http://tree.bio.ed.ac.uk/software/figtree/), and annotated in Inkscape (https://inkscape.org/).

For detecting putative species disagreement within clades, all species names of the accessions present within GMYC clusters were compared and a percentage agreement metric per cluster was generated. For each cluster, the maximum number of accessions with the same species names was divided by the total number of accessions within the clade. This ratio indicates how divergent species names are within the GMYC clade, and all clades below 100% agreement can indicate a possible misannotation or new synonymies. The R script to generate the species delimitation and this ratio is available in **File S2.**

2.5 Taxonomic assignment of sequence names

Regarding foliose *Ulva* species, since several species names have been found to be synonymous, we used the species names listed in **Table 1** as our reference. Where holotype or lectotype reference sequences are available, we attributed the species names of the reference to all sequences within the same GMYC clade. Where such type sequences are not available, we based our species attribution with comparisons from sequences from the literature and the GMYC clustering, with the caveat that indeed the nomenclature of the GMYC clade could change once holotype sequences become available. The rationale behind the selection of reference sequences is detailed in **File S1**.

2.6 Species distribution of distromatic foliose *Ulva* species.

The country of origin, GPS coordinates, specimen name and publication name of all of the NCBI entries in the three datasets were recovered using custom python scripts (**Files S3** and **S4**), restricted to vouchers assigned in our analysis as belonging to the eleven main distromatic foliose *Ulva* species [namely, *U. australis* Areschoug, *U. fenestrata* Postels & Ruprecht, *U. lactuca* Linnaeus, *U. gigantea* (Kützing) Bliding, *U. lacinulata* (Kützing) Wittrock, *U.ohnoi* M.Hiraoka & S.Shimada, *U. rigida*

C.Agardh, *U. expansa* (Setchell) Setchell & N.L.Gardner, *U. arasakii* Chihara and *U. ohiohilulu* H.L.Spalding & A.R.Sherwood], and *Ulva sp. A*. Publications associated with NCBI entries missing GPS coordinates and/or location of origin were manually searched to retrieve GPS coordinates where available. Accessions whose area of origin were uncertain were removed from the analysis. Duplicated specimens (i.e., specimens with more than one barcode sequenced in the NCBI) were removed and only one entry was kept. The complete list of vouchers, specimen, name, publication, GPS coordinates and proposed species attribution is available in **Table S3**. The world map and pie-chart distribution of *Ulva* species was created in R using the package Rworldmap (South, 2011).

3. Results

Using the analysis pipeline we created, we recovered and analysed all *Ulva* sequences in the NCBI, as well as 185 additional strains from the North-East Atlantic sequenced in this study, for the three most common barcodes used in *Ulva* phylogeny, namely *rbc*L, *tuf*A and ITS1.

3.1 Analysis of all *Ulva spp. rbc*L sequences from public repositories

We used the *rbc*L dataset generated in this study, that from Fort, McHale, et al. (2021), as well as all available *rbc*L sequences from *Ulva* entries in the NCBI (see Materials and Methods). From the *rbc*L alignment, we generated a Maximum Likelihood phylogenetic tree containing 1,062 sequences. GMYC analysis revealed the presence of 24 clades containing more than two sequences (confidence interval 19-28) (**Fig. 2).** Of these, ten belong to obligatory distromatic foliose species, namely *U. arasakii*, *U. sp. A*, *U. expansa*, *U. fenestrata*, *U. australis*, *U. gigantea*, *U. ohnoi*, *U. lactuca*, *U. rigida* and *U. lacinulata* (**Table 2**). The GMYC species delimitation, however, failed to discriminate between five species. *U. lacinulata* and *U. sp A* were found to be conspecific, despite previous evidence to the contrary (Fort, McHale, et al., 2021; Heesch et al., 2009), as well as a single clade containing both *U. lactuca* and *U. ohnoi,* and another clade containing *U. rigida* and *U. adhaerens*. The full maximum likelihood tree (including bootstrap support), the Bayesian MCMC analysis tree (including probabilities), and entries species names for *rbc*L can be found in **Fig. S1**, and **Table S3**).

The 177 *rbc*L sequences from this study originating from the North East Atlantic belong to seven distinct clades, with 19 samples identified as *U. rigida*, 21 samples as *U. fenestrata*, 47 as *U. australis*, 13 as *U. gigantea*, 2 as *U. ohnoi*, 12 as *U. sp A* and 63 as *U. lacinulata*.

The clades containing *U. australis* and *U. gigantea* are the most consistent, with minimal discrepancies between species names within the clades. The other clades appear more problematic, with significant species names discrepancies in the *U. fenestrata*, *U. ohnoi*, *U. lacinulata* and *U. rigida* clades (**Fig. S2**).

We found 69 strains belonging to the *U. ohnoi* clade, 2 in this study, 57 *U. ohnoi* vouchers from the NCBI database [described in Hiraoka, Shimada, Uenosono, & Masuda, (2004); Krupnik et al., (2018); Melton, Collado-Vides, & Lopez-Bautista, (2016)], including the type), as well as several likely misannotated entries, including one *U. rigida*, three *U. lactuca*, three *U. fasciata*, one *U. beytensis* Thivy & Sharma, one *U. reticulata* Forsskål and one *U. taeniata* (Setchell) Setchell & N.L.Gardner. Most entries originate from the same unpublished population set (number 452119310). Next, the *U. sp A* clade contains 12 strains from this study, as well as 49 *U. rigida* entries from the NCBI, described in Heesch et al. (2009); Rautenberger et al. (2015) and Loughnane et al. (2008). Finally, the *U. lacinulata* clade containing 138 strains appears to contain several cases of likely species misidentification. This clade contains 63 individuals from this study, 38 individuals from Fort, McHale, et al. (2021) [which are now renamed *U. lacinulata* following nomenclatural reassignment (Hughey et al., 2021a)]and four *U. laetevirens* entries [two from (Kraft, Kraft, & Waller, 2010), and two from China (Du et al., 2014)]. However, 21 entries in the *U. lacinulata* clade were assigned as *U. rigida*. The presence of a large number of *U. laetevirens* individuals in this clade stems from the recent sequencing of the *U. laetevirens* holotype (Hughey, Gabrielson, Maggs, Mineur, & Miller, 2021b), which was found to belong to *U. australis*. Thus, the sequences formerly known as *U. laetevirens* should now be reclassified as *U. lacinulata*, whose sequenced holotype belong to the same clade (Hughey et al., 2021a). Interestingly, all five *U. scandinavica* entries also cluster within the *U. lacinulata* clade, with two out of five *U. scandinavica* entries being indistinguishable from *U. lacinulata* ones, and the other three possessing a single polymorphic site. Altogether, *U. scandinavica* are likely to be synonymous with *U. lacinulata*. Finally, following nomenclatural adjustment via

sequencing of the lectotype (Hughey et al., 2021a), the *U. rigida* clade contains *U. pseudorotundata* sequences.

Of the large foliose species not represented in our dataset, *U. arasakii* is represented by a single individual, and the *U. expansa* clade contains six NCBI entries, four *U. expansa* and two *U. lobata,* which have been shown to be synonymous (Hughey et al., 2019), **Table 1**. Concerning other species, *U. compressa* Linnaeus and *U. intestinalis* Linnaeus are well defined, with no misidentification of *U. intestinalis*, and only three likely misannotated sequences in the *U. compressa* clade: one *U. intestinalis* and two *U. pseudocurvata* entries. The other species are more problematic, with several poorly defined clades containing a mixture of *U. prolifera*, *U. linza*, *U. flexuosa*, *U. californica* and *U. tanneri*.

Altogether, we found a relatively low agreement between the species names assigned to the NCBI vouchers and the GMYC clusters for *rbc*L, with only seven out of 24 GMYC clusters containing 100% of sequences with the same species name annotation (**Fig. S2**). Disagreements between GMYC clades and species names within them do not necessarily indicate misannotations, due to poor detection of species boundaries by the GMYC analysis using this barcode. Nonetheless, the results show that *rbc*L sequences are likely poor at defining *Ulva* species, and that each clade should be investigated in detail, as significant naming discrepancies are present.

3.2 Analysis of all *tuf*A sequences from public repositories

We performed the same analysis using the *tuf*A barcode (**Fig. 3**, **Fig. S3** and **Table S3**). We found significantly more species clusters than for the *rbc*L barcode (40 species clusters, confidence interval 37-46).

For foliose species (**Table 2**), as expected, the *U. fenestrata* clade shows the same name misapplication with *U. lactuca*, with 225 individuals, 21 in this study, 119 *U. fenestrata* entries and 86 *U. lactuca* entries. *U. australis, U. gigantea tuf*A clades appear well defined, with no name misapplication, similar to the *rbc*L results. *U. ohnoi* is also generally well circumscribed. The *U. lacinulata* and *U. sp. A* clades are separated by the GMYC analysis using *tuf*A, and 19 *U. rigida*

sequences are clustering within the *U. lacinulata* clade. Less common foliose species, such as *U. expansa*, *U. arasakii* and *U. ohiohilulu* are represented with more than two entries, each with their separate clades.

For other species, *tuf*A appears more appropriate than *rbc*L for species delimitation, with a clear separation between *U. linza* and *U. prolifera* clades, as well as between *U. californica* and *U. flexuosa*, without apparent misidentifications apart from one *U. mediterranea* Alongi, Cormaci & G.Furnari and one *U. prolifera* vouchers, both displaying 100% identity with *U. flexuosa. Ulva compressa* and *U. intestinalis* are similarly well defined in the *tuf*A dataset.

Consequently, the percentage agreement of species names within GMYC clusters in the *tuf*A dataset is significantly higher than with *rbc*L, with 23/40 GMYC clusters showing complete agreement (**Fig. S2**).

3.3 Analysis of all ITS1 sequences from public repositories

Finally, the analysis was repeated on the ITS1 barcode dataset (**Fig. 4**, **Fig. S4** and **Table S3**). Once again, the results are in general agreement with the previous barcodes, particularly with *tuf*A. Indeed, species delimitation predicts 42 species clusters (compared with 40 with *tuf*A), with a confidence interval of 34 to 59.

The *U. australis*, *U. gigantea* and *U. ohnoi* clades are well conserved, with only minor discrepancies (**Table 2**). The *U. fenestrata* clade however contains 33 *U. fenestrata* accessions and 19 erroneous *U. lactuca* accessions. The *U. lacinulata* clade contains 134 sequences with 62 from this study, the holotype of *U. armoricana* [NCBI accession MT078962, Coat et al., (1998)], and 44 *U. laetevirens.* As for the *rbc*L results, we found *U. scandinavica* within the *U. lacinulata* clade, all of which show 100% identity with most other *U. lacinulata* sequences.

With regard to narrow-tubular species, the "Linza-Procera-Prolifera" (LPP) complex is poorly delimited, with NCBI entries of all three species intertwined within five clades. Outside of the LPP complex, other narrow-tubular *Ulva* species appear well delimited, with two exceptions. The *U. meridionalis* R.Horimoto & S.Shimada (Horimoto, Masakiyo, & Ichihara, 2011) clade contains

twelve likely misannotated *U. prolifera* vouchers. Similarly, the clade containing *U. tepida* Y.Masakiyo & S.Shimada contains several entries annotated as *U. intestinalis, U. shanxiensis* L.Chen, J.Feng & S.-L.Xie *and U. paschisma* F.Bast.

Out of 42 GMYC clusters, only 14 show complete agreement in species names (**Fig. S2**). This shows that a significant number of misannotations are likely present in the ITS sequences of the *Ulva* genera.

3.4 Impact of NCBI accession reanalysis on species distribution

After reassigning species name for each NCBI entry, we generated a world map of the distribution of the eleven large foliose *Ulva* species from which there is genetic evidence (**Fig. S5**). Strikingly, no *U. lactuca* individuals are present in the North Atlantic and the Baltic Sea, outside of a specimen recovered from an aquarium and misannotated as *U. laetevirens* (Vranken et al., 2018), and a single specimen in Massachusetts, USA. As shown above, the reports of *U. lactuca* in many regions are all referable to *U. fenestrata*. Importantly, while the number of misannotations in the NCBI is significant, the problem is even higher in other databases that do not rely on DNA sequencing for reporting species records. For instance, the Ocean Biodiversity Information System (OBIS) contains > 4,700 records for *U. lactuca*, most of which located in the North Atlantic, in seeming contradiction with our results (**Fig. 5**). Hence, reanalysis of barcode sequences can drastically change species distribution.

3.5 Extension of the analytical framework with *Fucus*, *Porphyra* and *Pyropia spp*.

We used the same analytical pipeline to detect possible misannotations or new synonymies in three other genera of economically and ecologically important macroalgae: *Fucus spp.* (*Phaeophyceae*, *Fucaceae*), and two Bangiales genera, *Porphyra* and *Pyropia spp*.

For *Fucus spp*, we used all publicly available sequences for the *COX1* and nrRNA-ITS barcodes, and generated a maximum likelihood tree and species delimitation as for the *Ulva* datasets. The GMYC analysis predicts 8 and 9 species for *COX1* and ITS sequences, respectively (**Fig. 6**), with the *Fucus distichus* clade being split into 6 different predicted species by the GMYC analysis of *COX1*

sequences. For the ITS dataset, the clades containing *Fucus serratus* and *Fucus vesiculosus* species names are separated into two and four predicted clades, respectively. Overall, the species names within the GMYC clusters are well conserved, with 5/8 and 7/9 clusters displaying 100% agreement (**Fig. S2**). However, one clade in both barcode datasets appears problematic. *Fucus vesiculosus* and *Fucus spiralis* sequences are intertwined in both datasets. This indicates that the two species are frequently misannotated. Indeed, sequences with both names are in some cases indistinguishable, with 100% identity. The full maximum likelihood trees are available in **Fig. S6**.

The *Porphyra* and *Pyropia* dataset contains 1,296 *COX1* sequences, separated into 62 GMYC clusters (**Fig. 7,** full tree available in **Fig. S7**). Unlike *Ulva*, the species names within GMYC clusters appear remarkably consistent in this dataset, with only twelve out of 62 GMYC clusters containing sequences with different species names (**Fig. S2**). Furthermore, most of those relate to clusters containing vouchers with undetermined species names, hence do not represent misannotations *per se*. Only one clade is potentially problematic, with sequences named either *Porphyra linearis* or *Porphyra umbilicalis,* despite being identical in sequence.

Altogether, the three additional datasets show a lower extent of potential misannotations than the *Ulva* datasets, even when using a species-rich family such as the Bangiaceae. We generated a histogram of the percentage of agreement in the species names of all GMYC clusters between the three groups of species investigated here (**Fig. 8**), which shows a significant number of GMYC clusters below 100% agreement in *Ulva,* compared to *Fucus*, *Porphyra*, and *Pyropia* datasets.

4) Discussion

4.1) Limitations of species delimitation using single barcodes

In this study, we endeavoured exhaustively to assess the genetic information available for our taxa of interest. We used all publicly available sequences from the NCBI for three common barcodes. Notably, species delimitation using such a large number of sequences yields relatively large species clusters confidence intervals. For instance, using *rbc*L alone did not allow to separate certain taxa that were previously shown to be separate species (Fort, McHale, et al., 2021; Hiraoka et al., 2004;

Hughey et al., 2019), such as *U. sp. A* and *U. lacinulata* or *U. ohnoi* and *U. lactuca*. This could be due to the use of a smaller length of alignment for *rbc*L in this study, as opposed to concatenated *rbc*L + *tuf*A sequences in Fort, McHale, et al. (2021) for the *U. sp A*/*U. armoricana* separation. In addition, such a discrepancy is inherent to large-scale species delimitation analyses when using limited genetic information (Leliaert et al., 2014; Tang, Humphreys, Fontaneto, & Barraclough, 2014). Indeed, the presence of possibly spurious sequences in the entire dataset can skew the speciation threshold of the GMYC analysis, especially when a single barcode containing a limited number of SNPs between species is used. This likely explains the relatively large confidence intervals we observed for *rbc*L. In contrast, using *tuf*A alone we were able to separate *U. lacinulata* and *U. sp A*, which is in agreement with previous studies (Fort, McHale, et al., 2021; Hayden & Waaland, 2002; Heesch et al., 2009; Tan et al., 1999). *tuf*A displays more SNPs than *rbc*L when comparing those two species (nine versus two, respectively), allowing for a species delimitation between the two clades. The ITS1 barcode similarly allowed for the separation of those two species. However, while we are able to separate *U. lactuca* and *U. ohnoi* using *tuf*A, *U. ohnoi* is separated into two different clades. Similarly, *U. linza*, *U. compressa*, *U. intestinalis* and *U. prolifera* clades are separated into several clades. Finally, while seven *U. reticulata* vouchers originating from (Monotilla et al., 2018), are included in the *U. ohnoi* clade using the ITS1 barcode, these likely do not represent erroneous annotation, since in their study, Monotilla et al. (2018) showed that *U. ohnoi* and *U. reticulata* are sexually isolated, despite having little to no sequence divergence in this barcode sequence.

Thus, appropriate species delimitation analysis should ideally be performed on a larger amount of genetic information, such as full organellar genomes, or concatenated sequences from the same specimens. Additionally, other species delimitation algorithms are available, such as Poisson Tree Processes (PTP) or the Automatic Barcode Gap Discovery for primary species delimitation (ABGD) (Puillandre, Lambert, Brouillet, & Achaz, 2012; Zhang, Kapli, Pavlidis, & Stamatakis, 2013). It is likely that using different methodologies for species delimitation will yield a different number of species clades in the same dataset, and a combination of approaches could be used to precisely delimitate all *Ulva* species. Regardless of precise species delimitation however, the methodology described here allows to quickly test putative clades and their associated sequence names for possible misannotations or new synonymies. Notably, the use of "agreement of species names within clade"

(**Fig. S2, Fig. 8**) from the GMYC output helps to identify potentially problematic clades and species names. It provides a visual representation of the diversity within the dataset and serves as a steppingstone for in-depth reassessment of the taxonomy and diversity of genera of interest.

Regarding our findings with *Ulva*, the number of "species names" in the entries from the NCBI dataset is 56, nine of which are classified as synonyms. Of the 47 unique species names remaining, this analysis, despite its limitations, found ~40 species clusters containing more than two sequences, thus broadly agreeing with the present number of species described in NCBI. These numbers are significantly lower than that of the number of currently accepted species taxonomically (84 according to (Guiry & Guiry, 2021)). This apparent discrepancy could be explained by the presence of numerous species entities described morphologically in past studies from which there is no genetic evidence. These specimens should be sequenced if they are available, or their type locality resampled, as the NCBI database likely only contains a subset of all *Ulva* species.

4.2) Nomenclature, taxonomy and species misidentifications in public repositories.

The main issue with the use of public repositories to assign species name to sequences is the underlying quality of the species annotation within the repository. Two issues can be present, a nomenclatural issue, where the naming of the taxa is erroneous, or taxonomical issues, where the relationships between taxa is at fault (de Queiroz, 2006). The analytical framework described here allows us to identify clades that contain sequences with different species names, which could represent new synonymies for nomenclatural adjustments, and/or detect problematic taxonomic relationships when sequences of the same species name are present in different clades. Importantly, both of those points do not require prior knowledge of the nomenclature or taxonomy of the genus. For example, the presence of a significant amount of *U. lactuca* sequences intertwined with *U. fenestrata* accessions in one clade highlights misannotation of many specimens of *U. lactuca*, while multiple clades containing only one species name could represent undescribed new taxa.

However, to resolve the nomenclatural issues highlighted requires the systematic sequencing of all available types or the designation of epitypes. This work in *Ulva* is currently underway by Hughey

and colleagues, leading to nomenclatural adjustments of several species names (Hughey et al., 2021a; Hughey et al., 2021b; Hughey et al., 2019). For example, the clade described here as *U. lacinulata* was previously referred to as *U. laetevirens* and *U. armoricana* (Fort, McHale, et al., 2021; Kirkendale, Saunders, & Winberg, 2013; Miladi et al., 2018). Following sequencing of the *U. laetevirens* lectotype (Hughey et al., 2021b), the name *U. laetevirens* was found to be synonymous with *U. australis*. Recently, the sequencing of *U. lacinulata* lectotype revealed that it was the oldest valid and available name for this clade (Hughey et al., 2021a). We therefore renamed our accessions as *U. lacinulata*. Furthermore, sequencing of the *U. rigida* lectotype revealed that it belongs to the clade previously known as *U. pseudorotundata,* for which the oldest available name is *U. rigida* (Hughey et al., 2021a). Finally, given that the sequences previously assigned as *U. rigida* by us do not currently have a an available name with a sequenced type, these sequences are provisionally referred to as *Ulva sp. A*. This highlights that nomenclature adjustments are likely to continue until all available types sequences become available, a huge task made more difficult by missing types and prohibitions on sampling of types by herbaria. Nonetheless, taxonomically, such adjustments do not impact the clustering of sequences into species clades and the analytical framework described here, which aims to provide an exhaustive view of sequences names, agreements, and species clusters for a genus of interest.

For instance, it was recently reported by Hughey et al. (2019) that several misidentifications were found within the *U. fenestrata* clade. Here, using all sequences available, we found that this misidentification is indeed significant. Some 40% of sequences belonging to *U. fenestrata* are misannotated (127 / 334). Hence, caution should be exercised when comparing *U. fenestrata* sequences using BLAST since some of the best matches will erroneously be referred to "*U. lactuca*." We naturally support the use of *U. fenestrata* type as described by Hughey et al. (2019) as the baseline for this species (**Table 3**). This significant amount of species misannotation lead to a drastic change in the species distribution of *U. lactuca* (**Fig. 5**) and should not be overlooked. Only *Ulva* products labelled as containing "*Ulva lactuca*" are officially authorized for food consumption in Europe outside of France (Barbier et al., 2019). Furthermore, accurate description of the species used in the literature is essential for natural products biodiscovery, nutritional profile and traceability (Leal, Hilário, Munro, Blunt, & Calado, 2016). This highlights the need to both improve the identification of

Ulva species and to change the European food regulation by inclusion of the *Ulva* species which are effectively consumed at present under the name of "*Ulva lactuca*" or to treat "Ulva lactuca" as a commercial name encompassing all foliose *Ulva* species.

Finally, our study shows that *U. "rigida"* (now *U. sp A*) and *U. lacinulata* are also commonly misannotated in public repositories, which was hinted by Miladi et al. (2018). It perhaps is not surprising since both species sequences are relatively close, with only a handful of discriminating SNPs contained within those three barcodes, and the viability of interspecific hybrids (Fort, Linderhof, et al., 2021; Fort, McHale, et al., 2021). However, previous species delimitation analysis on *rbc*L + *tuf*A using different methodologies (GMYC and bPTP), and the sequence identity differences between the organellar genomes of the two clades indicates that they are likely two separate species (Fort, McHale, et al., 2021), and not the single taxon as postulated by Hughey et al. (2021a). While we consider that the *U. lacinulata* clade is fully resolved due to the presence of *U. lacinulata* type within the clade (Hughey et al., 2021a), the sequence of the *U. sp. A* type specimen is not currently available in public repositories. Hence, sequences of the *U. sp. A* clade will need to be renamed when a suitable type is found.

Overall, the analysis of large foliose *Ulva* species showed ~26% of misannotated entries in the NCBI database, a percentage likely much higher when tubular or partially tubular species are considered. A significant amount of the misannotations originates from recent nomenclature changes, which renders the work presented in this study particularly important, as we provide in **Table S3** all of the NCBI accession numbers of the foliose species highlighted here, as well as the updated species attribution. We encourage the *Ulva* scientific community to use the trees described here as potential "accession quality check" for species annotation based on BLAST results. We provide in **Fig. S1, Fig. S3** and **Fig. S4** the trees of all three barcodes in to allow researchers to use the search function of PDF viewers for searching specific NCBI accessions and identifying to which clade they belong. Generally, however, we encourage the use of exhaustive trees for phylogenetic analyses (i.e., including all available NCBI sequences), instead of trees containing a subset of "selected" NCBI entries. For example, a BLAST result of NCBI accession HQ610342.1 shows 11 matches with 100% identity, 10 of which are classified as *U. lactuca*. Therefore, if a tree were generated using the first five NCBI hits as reference, the sequence will likely be classified as *U. lactuca*. Conversely, using the

entire NCBI dataset highlights that all of those *U. lactuca* sequences are misannotated *U. fenestrata*. Including all sequences leads to a significant increase in computational time, but with the use of multithreading by raxml-NG and MrBayes, and the BEAGLE library for BEAST, we found that generating trees and GMYC analyses with > 1,000 sequences takes ~ 48 hours on eight CPU cores, decreasing further to ~10 hours with 64 CPU cores.

Nevertheless, we propose in **Table 3** a list of reference NCBI accessions for all three barcodes of the eleven large foliose *Ulva* species. The rationale for this list is available in **File S1**. As it is simple to update the information associated to NCBI sequences (see https://www.ncbi.nlm.nih.gov/genbank/update/), we encourage authors that have deposited sequences on the NCBI to update, if incorrect, the "organism" information of their accession numbers, thus avoid the amplification and recurrence of misannotated *Ulva* species, such as *U. lactuca,* and to update taxonomic assignments due to nomenclatural adjustments.

Concerning tubular and or partially tubular species, the major hurdle found here lies within the separation of *U. linza, U. procera* and *U. prolifera* individuals. This appears to be an ongoing issue with the delimitation of the species within the Linza-Procera-Prolifera (LPP) complex (Cui et al., 2018; Kang, Kim, Kim, Choi, & Kim, 2014; Leliaert et al., 2009), and will require further re-analysis of the NCBI entries after organelle sequencing of holotype specimens. The precise species delimitation of those clusters is outside the scope of this study but indicates that caution should also be taken when analyzing the sequences of those species, as misidentifications are likely to be present.

The taxonomic groups described here could also be used to study possible introduction event(s) of non-native species. Notably, (Sauriau et al., 2021) recently questioned the introduction of *U. australis* in Europe by using all available NCBI sequences of *U. australis* to infer introduction events. Indeed, the separation of sequences from a given species into haplotypes allows for a more granular analysis of species diversity and the detection of the introduction of new genotypes into the environment (Zhao et al., 2021). The use of haplotype network tools such as POPART (Leigh & Bryant, 2015), together with the output of the analytical framework presented here, could allow to quickly revisit introduction events of any taxa of interest.

4.3) *Ulva spp*, a particularly problematic genus compared to *Fucus* and *Porphyra*/*Pyropia* genera

Altogether, the potential for misidentifications in public repositories should not be overlooked, and in case of *Ulva* is significant. Comparing the results obtained from *Ulva* with those from *Fucus* and *Porphyra*/*Pyropia* demonstrated that *Ulva* is a particularly problematic genus (**Fig. 8**). In the case of *Fucus* spp., we only found a single clade that seems particularly problematic, with apparent misannotations between *Fucus spiralis* and *Fucus vesiculosus.* With the *Porphyra*/*Pyropia* dataset, which contains some 62 GMYC clades, one clade contained a mixture of *Porphyra linearis* Greville and *Porphyra umbilicalis* Kützing. Given that this clade is the only one containing either species' names, it is likely that those two species are synonymous. One species, *Neoporphyra haitanensis* (T.J.Chang & B.F.Zheng) J.Brodie & L.-E.Yang, whose genome has been released (Cao et al., 2020), appears to be frequently misannotated, given that sequences with this species name are present in multiple clades containing other species names.

The striking consistency in the Bangiales dataset over the *Ulva* one (**Fig. 8**) is likely due to the efforts of the Bangiales scientific community, that have collaboratively reassessed the Bangiales taxonomy and nomenclature over the last 20 years (Sutherland et al., 2011; Yang et al., 2020). Perhaps the ubiquitous distribution of *Ulva*, its phenotypical plasticity, and the slow release of holotype/lectotype specimen sequences, contribute to the considerable discrepancies in *Ulva* taxonomy. We believe that a similar approach to that of the Bangiales order is needed to appropriately revisit *Ulva* nomenclature and taxonomy, and the analytical framework described here could be used as the first step towards that goal.

Conclusions

Due to the increasingly large number of sequences being deposited in public repositories, it is becoming important regularly to reassess the genetic information of taxa of interest, to highlight ongoing species identification issues, update NCBI accessions with new nomenclatures, and potentially reassign names to previously uncharacterised synonymous species. Here, we investigated all *Ulva, Fucus* and *Porphyra/Pyropia* sequences in the NCBI public repository for common

barcodes, as a contribution to clarify the species composition and annotation of these three genera. This dataset can be used for future species identification, accession validation and classification purposes, to ensure accurate representation of the species names and taxa within the databases. The analytical framework described here in detail could be transferred to any other taxa of interest, particularly those that show subtle morphological differences between taxa and contain large amount of sequences and suspected misannotations.

Figure legends

Fig. 1: Analysis framework used in this study. The list of scripts and software is available in **File S1**.

Fig. 2: Maximum Likelihood phylogenetic tree of 1,062 *Ulva spp*. *rbc*L sequences, and description of the entries belonging to the main distromatic foliose *Ulva* species. Maximum likelihood tree of the *rbc*L alignment, rooted on *Umbraulva* sequences. Colored clades represent distromatic foliose species found in this study. Shaded clades represent tubular or partially tubular species and/or species with no representative in this study. Numbers, shaded and/or colored clades represent species clusters determined using GMYC. Full trees including bootstrap values and bayesian posterior probabilities are available in **Fig. S1**.

Fig. 3: Maximum Likelihood phylogenetic tree of 1,320 *Ulva spp*. *tuf*A sequences, and description of the entries belonging to the main distromatic foliose *Ulva* species. Maximum likelihood tree of the *tuf*A alignment, rooted on *Umbraulva* sequences. Colored clades represent distromatic foliose species found in this study. Shaded clades represent tubular or partially tubular species and/or species with no representative in this study. Numbers, shaded and/or colored clades represent species clusters determined using GMYC. Full trees including bootstrap values and bayesian posterior probabilities are available in **Fig. S3**.

Fig. 4: Maximum Likelihood phylogenetic tree of 1,245 *Ulva spp*. ITS1 sequences, and description of the entries belonging to the main distromatic foliose *Ulva* species. Maximum likelihood tree of the ITS1 alignment, rooted on *Umbraulva* sequences. Colored clades represent

distromatic foliose species found in this study. Shaded clades represent tubular or partially tubular species and/or species with no representative in this study. Numbers, shaded and/or colored clades represent species clusters determined using GMYC. Full trees including bootstrap values and Bayesian posterior probabilities are available in **Fig. S4**.

Fig. 5: Comparison of *Ulva lactuca* species distribution based on different databases. Each dot represents a single record.

Fig. 6: Maximum Likelihood phylogenetic tree of *Fucus spp* COX1 and nrRNA-ITS sequences. Numbers and shaded clades represent species clusters determined using GMYC. Full ML trees are available in **Fig. S6**.

Fig. 7: Maximum Likelihood phylogenetic tree of *Porphyra+Pyropia* COX1 sequences. Shaded clades represent species clusters determined using GMYC. Full ML tree is available in **Fig. S7**.

Fig. 8: Distribution of species names agreement per GMYC cluster between *Ulva*, *Fucus* and *Porphyra+Pyropia* datasets.

Table 1: Names and synonyms used in this study.

Species	**Synonymous name**	**Reference**
Ulva lactuca Linnaeus	*Ulva fasciata* Delile	Hughey et al, 2019
Ulva australis Areschoug	*Ulva pertusa Kjellman, Ulva*	Kraft et al, 2010, Huhgey et al,

	laetevirens Areschoug	2021
Ulva compressa Linnaeus	*Ulva mutabilis* Föyn	Steinhagen et al. 2019
Ulva expansa (Setchell) Setchell & N.L.Gardner	*Ulva lobata* (Kützing) Harvey	Hughey et al, 2019
	Ulva scandinavica Bliding, *Ulva armoricana* (Dion, Reviers	
Ulva lacinulata (Kützing)	& Coat)	Hughey et al, 2021, this study

Supplementary Data

Table S1: List of samples and metadata of *Ulva* strains collected in this study.

Table S2: List of primers used in this study.

Table S3: List of NCBI vouchers belonging to the eleven main foliose *Ulva* species, proposed name attribution and GPS coordinates.

Fig. S1: Complete ML and Bayesian trees of *rbc*L alignment.

Fig. S2: Agreement between GMYC clusters and species names for the datasets used in this study.

Fig. S3: Complete ML and Bayesian trees of *tuf*A alignment.

Fig. S4: Complete ML and Bayesian trees of ITS1 alignment.

Fig. S5: Worldwide species distribution of the eleven large distromatic foliose *Ulva* species.

Fig. S6: Complete ML COX1 and nrRNA-ITS trees of the *Fucus* dataset

Fig. S7: Complete ML COX1 tree of the *Porphyra*+*Pyropia* dataset

File S1: List of scripts and software used in this study.

File S2: R script for GMYC delineation and percentage species name agreement within GMYC cluster.

File S3: Python script to retrieve GPS coordinates from a list of NCBI accession numbers.

File S4: Python script to retrieve specimen names from a list of NCBI accession numbers.

Competing interests

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank Ricardo Bermejo (NUI Galway), Lars Brunner and Sarah Reed (Scottish Association for Marine Science), Dan Smale and Cat Wilding (Marine Biological Organisation), Tim van Berkel and Caroline Warwick-Evans (The Cornish Seaweed Company) and Wave Crookes (SeaGrown), Helena Abreu (Alga +), for providing some of the strains used in this study. This work was funded by the European Union Horizon 2020 programme (project ID 727892, GenialG - GENetic diversity exploitation for Innovative Macro-ALGal biorefinery, http://genialgproject.eu/), SFI Frontiers for the Future (Project Pristine Coasts, grant number 19/FFP/6841) and the European Union Northern Periphery and Arctic Programme (project number 366, SW-GROW - Innovations for Seaweed Producers in the Northern Periphery Area project; http://sw-grow.interreg-npa.eu/).

The authors would like to thank the reviewers and MER editors for their constructive criticism of the manuscript and helpful comments.

Data Availability Statement

The data that support the findings of this study are openly available in the NCBI at https://www.ncbi.nlm.nih.gov/, reference numbers MT894471- MT895108. Scripts and pipeline are available in GitHub: https://github.com/FortAnt/BarcodeAnalysis.

Authors Contributions.

AF and RS designed the experiments; all authors provided biological material; AF performed the experiments; AF, MM, MDG and RS analysed the results; KC and PP provided administrative and technical support; AF, MDG, MM and RS wrote the manuscript. All authors reviewed the manuscript.

References

Barbier, M., Charrier, B., Araujo, R., Holdt, S., Jacquemin, B., Rebours, C., . . . Charrier, B. (2019). PEGASUS-PHYCOMORPH European guidelines for a sustainable aquaculture of seaweeds. *COST action FA1406. Roscoff, France*.

Biancarosa, I., Espe, M., Bruckner, C., Heesch, S., Liland, N., Waagbø, R., . . . Lock, E. (2017). Amino acid composition, protein content, and nitrogen-to-protein conversion factors of 21 seaweed species from Norwegian waters. *Journal of Applied Phycology, 29*(2), 1001-1009.

Cao, M., Xu, K., Yu, X., Bi, G., Liu, Y., Kong, F., . . . Mao, Y. (2020). A chromosome-level genome assembly of Pyropia haitanensis (Bangiales, Rhodophyta). *Molecular Ecology Resources, 20*(1), 216-227. doi:https://doi.org/10.1111/1755-0998.13102

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973.

Chowdhary, A., Singh, A., Singh, P. K., Khurana, A., & Meis, J. F. (2019). Perspectives on misidentification of Trichophyton interdigitale/Trichophyton mentagrophytes using internal transcribed spacer region sequencing: Urgent need to update the sequence database. *Mycoses, 62*(1), 11-15. doi:https://doi.org/10.1111/myc.12865

Coat, G., Dion, P., Noailles, M.-C., De Reviers, B., Fontaine, J.-M., Berger-Perrot, Y., & Loiseaux-De Goër, S. (1998). *Ulva armoricana* (Ulvales, Chlorophyta) from the coasts of Brittany (France). II. Nuclear rDNA ITS sequence analysis. *European Journal of Phycology, 33*(1), 81-86.

Cui, J., Monotilla, A. P., Zhu, W., Takano, Y., Shimada, S., Ichihara, K., . . . Hiraoka, M. (2018). Taxonomic reassessment of Ulva prolifera (Ulvophyceae, Chlorophyta) based on specimens from the type locality and Yellow Sea green tides. *Phycologia, 57*(6), 692-704.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods, 9*(8), 772-772.

de Queiroz, K. (2006). The PhyloCode and the Distinction between Taxonomy and Nomenclature. *Systematic Biology, 55*(1), 160-162. doi:10.1080/10635150500431221

Dion, P., De Reviers, B., & Coat, G. (1998). Ulva armoricana sp. nov.(Ulvales, Chlorophyta) from the coasts of Brittany (France). I. Morphological identification. *European Journal of Phycology, 33*(1), 73-80.

Du, G., Wu, F., Mao, Y., Guo, S., Xue, H., & Bi, G. (2014). DNA barcoding assessment of green macroalgae in coastal zone around Qingdao, China. *Journal of Ocean University of China, 13*(1), 97-103. doi:10.1007/s11802-014-2197-1

Dugon, M. M., Black, A., & Arthur, W. (2012). Variation and specialisation of the forcipular apparatus of centipedes (Arthropoda: Chilopoda): A comparative morphometric and microscopic investigation of an evolutionary novelty. *Arthropod Structure & Development, 41*(3), 231-243. doi:https://doi.org/10.1016/j.asd.2012.02.001

Dunbar, J. P., Vitkauskaite, A., O'Keeffe, D. T., Fort, A., Sulpice, R., & Dugon, M. M. (2021). Bites by the noble false widow spider Steatoda nobilis can induce Latrodectus-like symptoms and vector-borne bacterial infections with implications for public health: a case series. *Clinical Toxicology*, 1-12.

Estoup, A., & Guillemaud, T. (2010). Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology, 19*(19), 4113-4130.

Fort, A., Guiry, M. D., & Sulpice, R. (2018). Magnetic beads, a particularly effective novel method for extraction of NGS-ready DNA from macroalgae. *Algal Research, 32*, 308-313. doi:https://doi.org/10.1016/j.algal.2018.04.015

Fort, A., Lebrault, M., Allaire, M., Esteves-Ferreira, A. A., McHale, M., Lopez, F., . . . Sulpice, R. (2019). Extensive variations in diurnal growth patterns and metabolism among *Ulva spp.* strains. *Plant physiology, 180*(1), 109-123.

Fort, A., Linderhof, C., Coca-Tagarro, I., Inaba, M., McHale, M., Cascella, K., . . . Sulpice, R. (2021). A sequencing-free assay for foliose Ulva species identification, hybrid detection and bulk biomass characterisation. *Algal Research-Biomass Biofuels and Bioproducts, 55*, 102280. doi:ARTN 102280

10.1016/j.algal.2021.102280

Fort, A., Mannion, C., Fariñas-Franco, J. M., & Sulpice, R. (2020). Green tides select for fast expanding *Ulva* strains. *Science of the Total Environment, 698*, 134337.

Fort, A., McHale, M., Cascella, K., Potin, P., Usadel, B., Guiry, M. D., & Sulpice, R. (2021). Foliose Ulva Species Show Considerable Inter-Specific Genetic Diversity, Low Intra-Specific Genetic Variation, and the Rare Occurrence of Inter-Specific Hybrids in the Wild. *Journal of Phycology, 57*(1), 219-233. doi:10.1111/jpy.13079

Fujisawa, T., & Barraclough, T. G. (2013). Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Systematic Biology, 62*(5), 707-724.

Garg, A., Leipe, D., & Uetz, P. (2019). The disconnect between DNA and species names: lessons from reptile species in the NCBI taxonomy database. *Zootaxa, 4706*(3), 401-407.

Guiry, M., & Guiry, G. (2020). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. 2020. *URL: http://www. algaebase. org*.

Hayden, H. S., & Waaland, J. R. (2002). Phylogenetic systematics of the Ulvaceae (Ulvales, Ulvophyceae) using chloroplast and nuclear DNA sequences. *Journal of Phycology, 38*(6), 1200-1212.

Heesch, S., Broom, J. E. S., Neill, K. F., Farr, T. J., Dalen, J. L., & Nelson, W. A. (2009). *Ulva*, *Umbraulva* and *Gemina*: genetic survey of New Zealand taxa reveals diversity and introduced species. *European Journal of Phycology, 44*(2), 143-154. doi:10.1080/09670260802422477

Hiraoka, M., Shimada, S., Uenosono, M., & Masuda, M. (2004). A new green-tide-forming alga, Ulva ohnoi Hiraoka et Shimada sp. nov.(Ulvales, Ulvophyceae) from Japan. *Phycological Research, 52*(1), 17-29.

Hofmann, L. C., Nettleton, J. C., Neefus, C. D., & Mathieson, A. C. (2010). Cryptic diversity of Ulva (Ulvales, Chlorophyta) in the Great Bay Estuarine System (Atlantic USA): introduced and indigenous distromatic species. *European Journal of Phycology, 45*(3), 230-239. doi:10.1080/09670261003746201

Horimoto, R., Masakiyo, Y., & Ichihara, K. (2011). Enteromorpha-like Ulva (Ulvophyceae, Chlorophyta) growing in the Todoroki River, Ishigaki Island, Japan, with special reference to Ulva meridionalis Horimoto et Shimada, sp. nov. *Bull. Natl. Mus. Nat. Sci. Ser. B Bot, 37*, 155-167.

Hughey, J. R., Gabrielson, P. W., Maggs, C. A., & Mineur, F. (2021a). Genomic analysis of the lectotype specimens of European Ulva rigida and Ulva lacinulata (Ulvaceae, Chlorophyta) reveals the ongoing misapplication of names. *European Journal of Phycology*, 1-11. doi:10.1080/09670262.2021.1914862

Hughey, J. R., Gabrielson, P. W., Maggs, C. A., Mineur, F., & Miller, K. A. (2021b). Taxonomic revisions based on genetic analysis of type specimens of Ulva conglobata, U. laetevirens, U. pertusa and U. spathulata (Ulvales, Chlorophyta). *Phycological Research, 69*(2), 148-153.

Hughey, J. R., Maggs, C. A., Mineur, F., Jarvis, C., Miller, K. A., Shabaka, S. H., & Gabrielson, P. W. (2019). Genetic analysis of the Linnaean *Ulva lactuca* (Ulvales, Chlorophyta) holotype and related type specimens reveals name misapplications, unexpected origins, and new synonymies. *Journal of Phycology, 55*(3), 503-508.

Johannesson, K., & Andre, C. (2006). Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Molecular Ecology, 15*(8), 2013-2029.

Kang, E. J., Kim, J.-H., Kim, K., Choi, H.-G., & Kim, K. Y. (2014). Re-evaluation of green tide-forming species in the Yellow Sea. *Algae, 29*(4), 267-277.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics, 20*(4), 1160-1166.

Kirkendale, L., Saunders, G. W., & Winberg, P. (2013). A molecular survey of *Ulva* (Chlorophyta) in temperate Australia reveals enhanced levels of cosmopolitanism. *Journal of Phycology, 49*(1), 69-81.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics, 35*(21), 4453-4455.

Kraft, L. G., Kraft, G. T., & Waller, R. F. (2010). Investigations into southern Australian *Ulva* (Ulvophyceae, Chlorophyta) taxonomy and molecular phylogeny indicate both cosmopolitanism and endemic cryptic species *Journal of Phycology, 46*(6), 1257-1277.

Krupnik, N., Paz, G., Douek, J., Lewinsohn, E., Israel, A., Carmel, N., . . . Maggs, C. A. (2018). Native, invasive and cryptogenic *Ulva* species from the Israeli Mediterranean Sea: risk and potential. *Mediterranean Marine Science, 19*(1), 132-146.

Leal, M. C., Hilário, A., Munro, M. H., Blunt, J. W., & Calado, R. (2016). Natural products discovery needs improved taxonomic and geographic information. *Natural Product Reports, 33*(6), 747-750.

Leigh, J. W., & Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6*(9), 1110-1116. doi:https://doi.org/10.1111/2041-210X.12410

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. *European Journal of Phycology, 49*(2), 179-196. doi:10.1080/09670262.2014.904524

Leliaert, F., Zhang, X., Ye, N., Malta, E. j., Engelen, A. H., Mineur, F., . . . De Clerck, O. (2009). Research note: identity of the Qingdao algal bloom. *Phycological Research, 57*(2), 147-151.

Lemoine, F., Domelevo Entfellner, J. B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature, 556*(7702), 452-456. doi:10.1038/s41586-018-0043-0

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Loughnane, C. J., McIvor, L. M., Rindi, F., Stengel, D. B., & Guiry, M. D. (2008). Morphology, *rbc*L phylogeny and distribution of distromatic *Ulva* (Ulvophyceae, Chlorophyta) in Ireland and southern Britain. *Phycologia, 47*(4), 416-429. doi:10.2216/PH07-61.1

Malta, E.-J., Draisma, S., & Kamermans, P. (1999). Free-floating Ulva in the southwest Netherlands: species or morphotypes? A morphological, molecular and ecological comparison. *European Journal of Phycology, 34*(5), 443-454.

Melton, J. T., Collado-Vides, L., & Lopez-Bautista, J. M. (2016). Molecular identification and nutrient analysis of the green tide species Ulva ohnoi M. Hiraoka & S. Shimada, 2004 (Ulvophyceae, Chlorophyta), a new report and likely nonnative species in the Gulf of Mexico and Atlantic Florida, USA. *Aquatic invasions, 11*(3), 225-237.

Mendez, M., Rosenbaum, H. C., Subramaniam, A., Yackulic, C., & Bordino, P. (2010). Isolation by environmental distance in mobile marine species: molecular ecology of franciscana dolphins at their southern range. *Molecular Ecology, 19*(11), 2212-2228.

Miladi, R., Manghisi, A., Minicante, S. A., Genovese, G., Abdelkafi, S., & Morabito, M. (2018). A DNA barcoding survey of *Ulva* (Chlorophyta) in Tunisia and Italy reveals the presence of the overlooked alien *U. ohnoi*. *Cryptogamie, Algologie*.

Monotilla, A. P., Nishimura, T., Adachi, M., Tanii, Y., Largo, D. B., & Hiraoka, M. (2018). Examination of prezygotic and postzygotic isolating barriers in tropical Ulva (Ulvophyceae, Chlorophyta): evidence for ongoing speciation. *Journal of Phycology, 54*(4), 539-549.

Nasehi, A., Al-Sadi, A. M., Esfahani, M. N., Ostovar, T., Rezaie, M., Atghia, O., . . . Javan-Nikkhah, M. (2019). Molecular re-identification of Stemphylium lycopersici and Stemphylium solani isolates deposited in NCBI GenBank and morphological characteristics of Malaysian isolates. *European Journal of Plant Pathology, 153*(3), 965-974.

O'Kelly, C. J., Kurihara, A., Shipley, T. C., & Sherwood, A. R. (2010). Molecular assessment of *Ulva spp.* (Ulvophyceae, Chlorophyta) in the Hawaiian islands. *Journal of Phycology, 46*(4), 728-735.

Packer, L., Gibbs, J., Sheffield, C., & Hanner, R. (2009). DNA barcoding and the mediocrity of morphology. *Mol Ecol Resour, 9 Suppl s1*(s1), 42-50. doi:10.1111/j.1755-0998.2009.02631.x

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E., & Stamatakis, A. (2010). How Many Bootstrap Replicates Are Necessary? *Journal of Computational Biology, 17*(3), 337-354. doi:10.1089/cmb.2009.0179

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., . . . Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology, 55*(4), 595-609.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793-808.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology, 21*(8), 1864-1877. doi:10.1111/j.1365-294X.2011.05239.x

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology, 67*(5), 901-904. doi:10.1093/sysbio/syy032

Rautenberger, R., Fernandez, P. A., Strittmatter, M., Heesch, S., Cornwall, C. E., Hurd, C. L., & Roleda, M. Y. (2015). Saturating light and not increased carbon dioxide under ocean acidification drives photosynthesis and growth in Ulva rigida (Chlorophyta). *Ecology and evolution, 5*(4), 874-888.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., . . . Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology, 61*(3), 539-542.

Saunders, G. W., & Kucera, H. (2010). An evaluation of *rbc*L, *tuf*A, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. *Cryptogamie, Algologie, 31*(4), 487-528.

Sauriau, P.-G., Dartois, M., Becquet, V., Aubert, F., Huet, V., Bréret, M., . . . Pante, E. (2021). Multiple genetic marker analysis challenges the introduction history of Ulva australis (Ulvales, Chlorophyta) on French coasts. *European Journal of Phycology*, 1-13.

South, A. (2011). rworldmap: a new R package for mapping global data. *R Journal, 3*(1).

Steinhagen, S., Karez, R., & Weinberger, F. (2019). Cryptic, alien and lost species: molecular diversity of *Ulva sensu lato* along the German coasts of the North and Baltic Seas. *European Journal of Phycology, 54*(3), 466-483. doi:10.1080/09670262.2019.1597925

Suchard, M. A., & Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics, 25*(11), 1370-1376. doi:10.1093/bioinformatics/btp244

Sutherland, J. E., Lindstrom, S. C., Nelson, W. A., Brodie, J., Lynch, M. D., Hwang, M. S., . . . Oliveira, M. C. (2011). A new look at an ancient order: generic revision of the Bangiales (Rhodophyta) 1. *Journal of Phycology, 47*(5), 1131-1151.

Tan, I. H., Blomster, J., Hansen, G., Leskinen, E., Maggs, C. A., Mann, D. G., . . . Stanhope, M. J. (1999). Molecular phylogenetic evidence for a reversible morphogenetic switch controlling the gross morphology of two common genera of green seaweeds, *Ulva* and *Enteromorpha*. *Molecular Biology and Evolution, 16*(8), 1011-1018.

Tang, C. Q., Humphreys, A. M., Fontaneto, D., & Barraclough, T. G. (2014). Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology and Evolution, 5*(10), 1086-1094. doi:10.1111/2041-210x.12246

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution, 24*(2), 110-117.

Vranken, S., Bosch, S., Peña, V., Leliaert, F., Mineur, F., & De Clerck, O. (2018). A risk assessment of aquarium trade introductions of seaweed in European waters. *Biological Invasions, 20*(5), 1171-1187.

Wichard, T., Charrier, B., Mineur, F., Bothwell, J. H., Clerck, O. D., & Coates, J. C. (2015). The green seaweed *Ulva*: a model system to study morphogenesis. *Frontiers in Plant Science, 6*(72). doi:10.3389/fpls.2015.00072

Yang, L.-E., Deng, Y.-Y., Xu, G.-P., Russell, S., Lu, Q.-Q., & Brodie, J. (2020). Redefining Pyropia (Bangiales, Rhodophyta): Four New Genera, Resurrection of Porphyrella and Description of Calidia pseudolobata sp. nov. From China. *Journal of Phycology, 56*(4), 862-879. doi:https://doi.org/10.1111/jpy.12992

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics, 29*(22), 2869-2876.

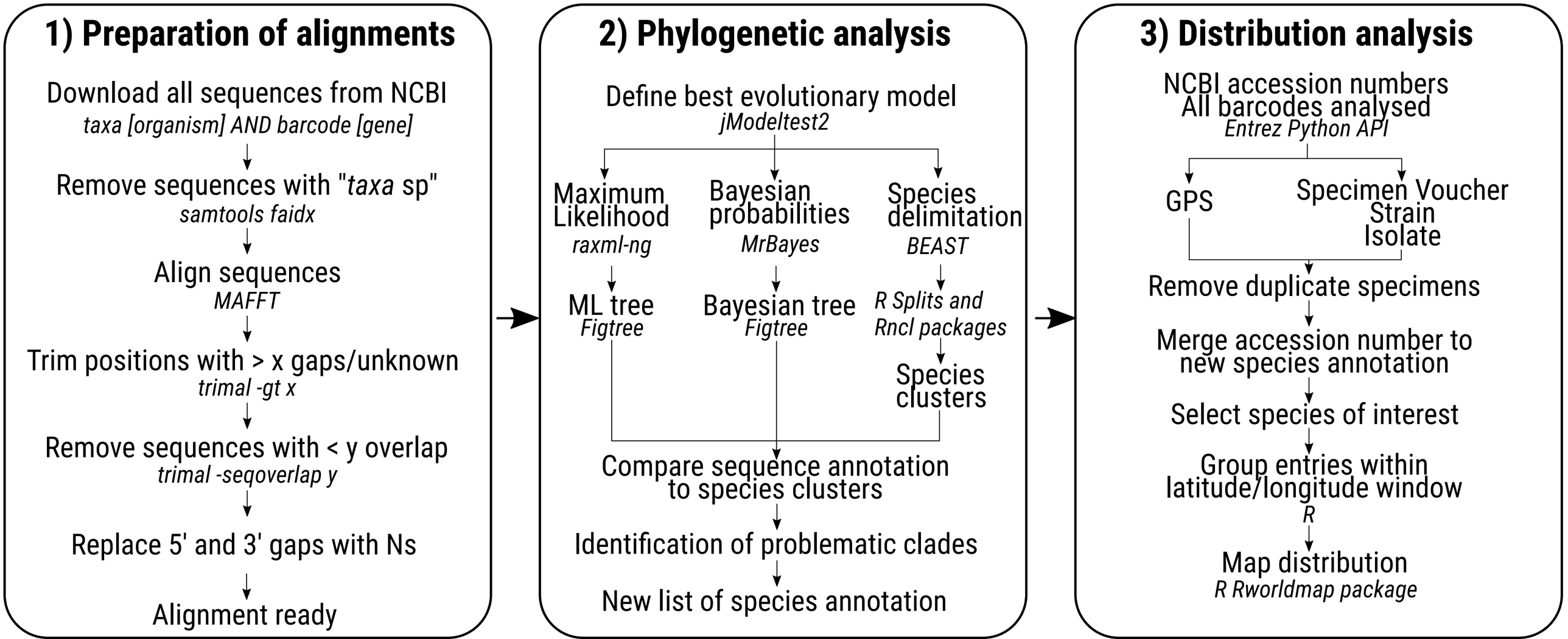
Zhao, W., Dong, L., Hong, D. D., Brodie, J., Chen, W. Z., Tien, D. D., . . . Yang, L. E. (2021). Haplotype networks of Phycocalidia tanegashimensis (Bangiales, Rhodophyta) indicate a probable invasion from the South China Sea to Brazil. *Marine Biodiversity, 51*(2), 1-11. doi:ARTN 33

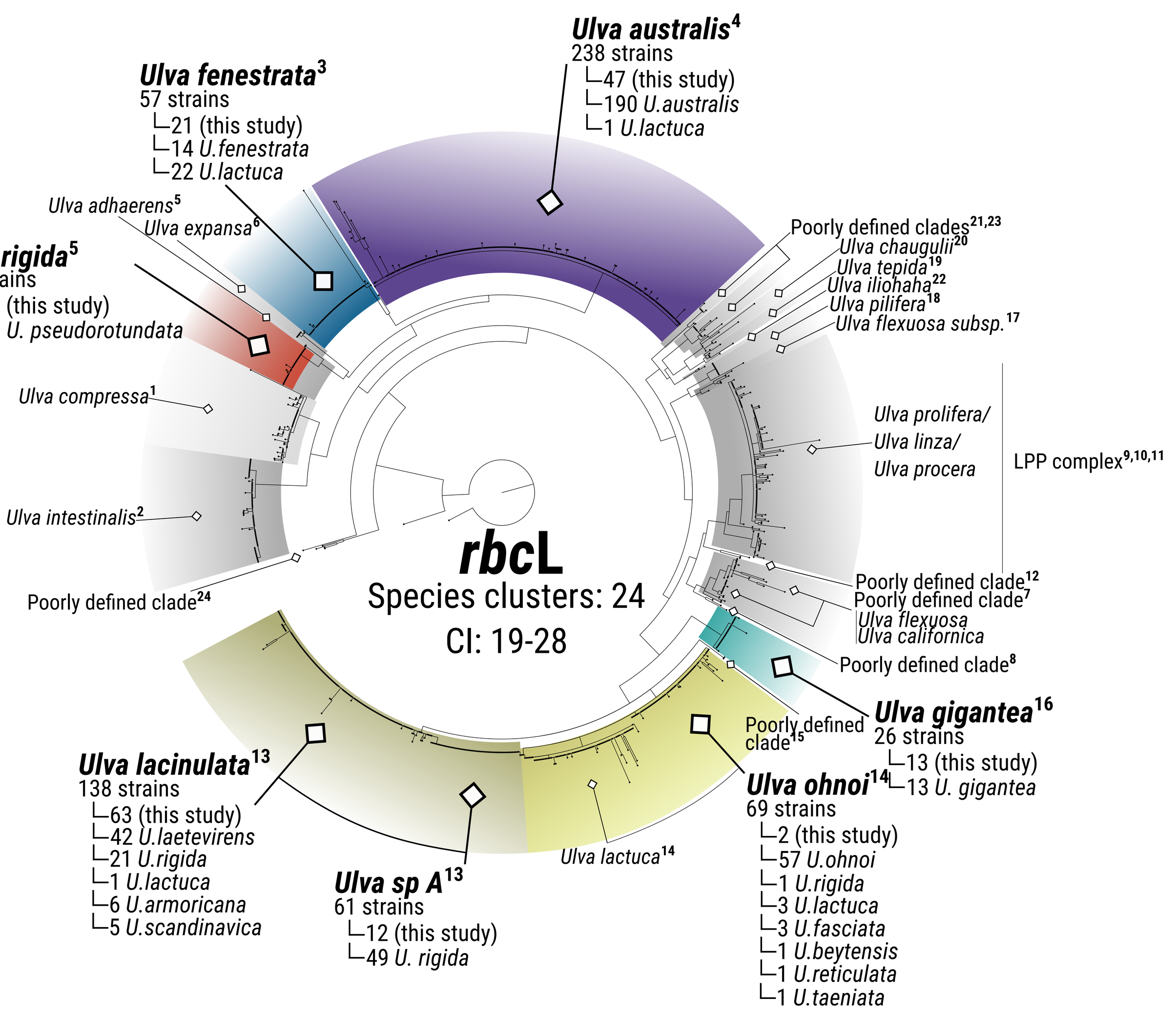
10.1007/s12526-021-01177-w

Species	Synonymous name	Reference
Ulva lactuca Linnaeus	*Ulva fasciata* Delile	Hughey et al, 2019
Ulva australis Areschoug	*Ulva pertusa* Kjellman, *Ulva laetevirens* Areschoug	Kraft et al, 2010, Hughey et al, 2021
Ulva compressa Linnaeus	*Ulva mutabilis* Föyn	Steinhagen et al. 2019
Ulva expansa (Setchell) Setchell & N.L.Gardner	*Ulva lobata* (Kützing) Harvey	Hughey et al, 2019
Ulva lacinulata (Kützing)	*Ulva scandinavica* Bliding, *Ulva armoricana* (Dion, Reviers & Coat)	Hughey et al, 2021, this study

Species clade	This study	Total	% misannotated	This study	Total	% misannotated	This study	Total	% misannotated
U. arasakii	0	11	0	0	1	0	0	10	0
U. lacinulata	62	134	49.3	63	138	46.4	59	140	57.9
U. australis	48	90	2.2	47	238	0.4	43	175	0
U. expansa	0	4	0	0	6	0	0	32	0
U. fenestrata	21	53	37.7	21	57	38.6	21	225	38.2
U. gigantea	15	25	0	13	26	0	14	32	0
U. lactuca	0	26	38.5	0	58	17.2	0	16	37.5
U. ohiohilulu	0	0	0	0	0	0	0	9	0
U. ohnoi	3	35	11.4	2	69	14.5	3	92	4.3
U. rigida	18	26	28	20	30	37	18	24	25
U. sp. A	15	50	70	12	61	80	11	40	73

Species	**NCBI ITS accession**	**NCBI *rbc*L accession**	**NCBI *tuf*A accession**	**Reference**
Ulva australis	MT894708	MT160564	MT160674	Fort et al, 2021
Ulva lacinulata	MW544060*	MW543061*	MT160697	Hughey et al, 2021, Fort et al, 2021
Ulva sp. A	MT894534	MT160573	MT160683	Fort et al, 2021
Ulva ohnoi	AB116031*	AB116037*	MT894753	Hiraoka et al, 2004; This study
Ulva rigida	MW544059*	MW543060*	MT160722	Hughey et al, 2021, Fort et al, 2021
Ulva gigantea	MT894480	MT160606	MT160716	Fort et al, 2021; this study
Ulva lactuca	AY260561†	MK456395*	MF172082†	Hayden et al, 2004, Hughey et al, 2019, Miladi et al, 2018
Ulva fenestrata	MT894725	MK456393*	MT160728	Fort et al, 2021, Hughey et al, 2019
Ulva arasakii	AB097650	AB097621	MK992126	Shimada et al, 2003, Kang et al, 2019
Ulva expansa	MH730161*	MH730975*	MH731007*	Hughey et al, 2018
Ulva ohiohilulu	KT881224*	KT932996*	KT932977*	Spalding et al, 2016
† Annotated as *U. fasciata*				
*** holotype/lectotype sequence**				


1) Preparation of alignments
Download all sequences from NCBI
taxa [organism] AND barcode [gene]
Remove sequences with "taxa sp"
samtools faidx
Align sequences
MAFFT
Trim positions with > x gaps/unknown
trimal -gt x
Remove sequences with < y overlap
trimal -seqoverlap y
Replace 5' and 3' gaps with Ns
Alignment ready
2) Phylogenetic analysis
Define best evolutionary model
jModeltest2
Maximum Likelihood
raxml-ng
Bayesian probabilities
MrBayes
Species delimitation
BEAST
ML tree
Figtree
Bayesian tree
Figtree
R Splits and Rncl packages
Species clusters
Compare sequence annotation to species clusters
Identification of problematic clades
New list of species annotation
3) Distribution analysis
NCBI accession numbers
All barcodes analysed
Entrez Python API
GPS
Specimen Voucher
Strain
Isolate
Remove duplicate specimens
Merge accession number to new species annotation
Select species of interest
Group entries within latitude/longitude window
R
Map distribution
R Rworldmap package


Ulva australis[4]
238 strains
47 (this study)
190 U.australis
1 U.lactuca
Ulva fenestrata[3]
57 strains
21 (this study)
14 U.fenestrata
22 U.lactuca
Ulva adhaerens[5]
Ulva expansa[6]
Ulva rigida[5]
30 strains
19 (this study)
11 U. pseudorotundata
Poorly defined clades[21,23]
Ulva chaugulii[20]
Ulva tepida[19]
Ulva iliohaha[22]
Ulva pilifera[18]
Ulva flexuosa subsp.[17]
Ulva compressa[1]
Ulva prolifera/
Ulva linza/
Ulva procera
LPP complex[9,10,11]
Ulva intestinalis[2]
Poorly defined clade[24]
rbcL
Species clusters: 24
CI: 19-28
Poorly defined clade[12]
Poorly defined clade[7]
Ulva flexuosa
Ulva californica
Poorly defined clade[8]
Ulva gigantea[16]
26 strains
13 (this study)
13 U. gigantea
Poorly defined clade[15]
Ulva ohnoi[14]
69 strains
2 (this study)
57 U.ohnoi
1 U.rigida
3 U.lactuca
3 U.fasciata
1 U.beytensis
1 U.reticulata
1 U.taeniata
Ulva lacinulata[13]
138 strains
63 (this study)
42 U.laetevirens
21 U.rigida
1 U.lactuca
6 U.armoricana
5 U.scandinavica
Ulva sp A[13]
61 strains
12 (this study)
49 U. rigida
Ulva lactuca[14]

Ulva gigantea[22]
32 strains
└14 (this study)
└18 *U.gigantea*

Ulva ohnoi[15,16]
92 strains
└3 (this study)
└85 *U.ohnoi*
└3 *U.lactuca*
└1 *U.prolifera*

Ulva lactuca[18]

Ulva sp. A[14]
40 strains
└11 (this study)
└29 *U.rigida*

Poorly defined clades[19,20,21]

Ulva lacinulata[13]
140 strains
└59 (this study)
└62 *U.laetevirens*
└19 *U.rigida*

Poorly defined clade[17]

Ulva shanxiensis[39]

LLP complex

Ulva linza[27,28,29]

Ulva prolifera[30,31,32]

Ulva stenophylla[33]

Ulva iliohaha[38]

Ulva pilifera[37]

Ulva chaugulii[35]

Poorly defined clade[34]

Ulva tepida[36]

Ulva flexuosa[24]

Ulva californica[23]

Ulva torta[25]

Ulva ohiohilulu[40]

Ulva compressa[9,10,11,12]

Ulva intestinalis[6,7,8]

Ulva expansa[5]

Ulva rigida[4]
24 strains
└18 (this study)
└6 *U.pseudorotundata*

tufA
Species clusters: 40
CI: 37-46

Ulva australis[3]
175 strains
└43 (this study)
└132 *U.australis*

Ulva arasakii[1]

Ulva fenestrata[2]
225 strains
└21 (this study)
└119 *U.fenestrata*
└86 *U.lactuca*

ITS1

Species clusters: 42

CI: 34-59

Ulva lacinulata[9]
134 strains
└62 (this study)
└44 *U.laetevirens*
└21 *U.rigida*
└3 *U.scandinavica*
└3 *U.armoricana*
└1 *U.fenestrata*

Ulva ohnoi[17]
35 strains
└3 (this study)
└21 *U.ohnoi*
└4 *U.lactuca*
└7 *U.reticulata*

Ulva sp. A[10]
52 strains
└15 (this study)
└36 *U.rigida*
└1 *U.lactuca*

Ulva gigantea[16]
25 strains
└15 (this study)
└10 *U.gigantea*

Ulva fenestrata[6]
53 strains
└21 (this study)
└12 *U.fenestrata*
└19 *U.lactuca*
└1 *U.californica*

Ulva australis[7]
90 strains
└48 (this study)
└40 *U.australis*
└2 *U.rotundata*

Ulva rigida[4]
26 strains
└18 (this study)
└6 *U.pseudorotundata*
└1 *U.adhaerens*

Ulva taeniata[12], *beytensis*[18] *reticulata*[19]

Ulva muscoides[11]

Ulva lactuca[13,14,15]

Ulva ovata[21]

Poorly defined clade[20]
Ulva intestinalis
Ulva tepida
Ulva shanxiensis
Ulva paschisma

Ulva chaugulii[42]

Poorly defined clade[8]

Ulva torta[33]

Ulva partita[34]

Poorly defined clade[22]

Ulva californica[28,29]

Ulva flexuosa[26,27]

Ulva tanneri[32]

Ulva simplex[30]

Poorly defined clade[31]

LLP complex[35,36,37,40,41]
Ulva linza/
Ulva prolifera/
Ulva procera

Poorly defined clade[24]

Ulva meridionalis[23]

Poorly defined clade[25]

Ulva flexuosa subsp[38,39]

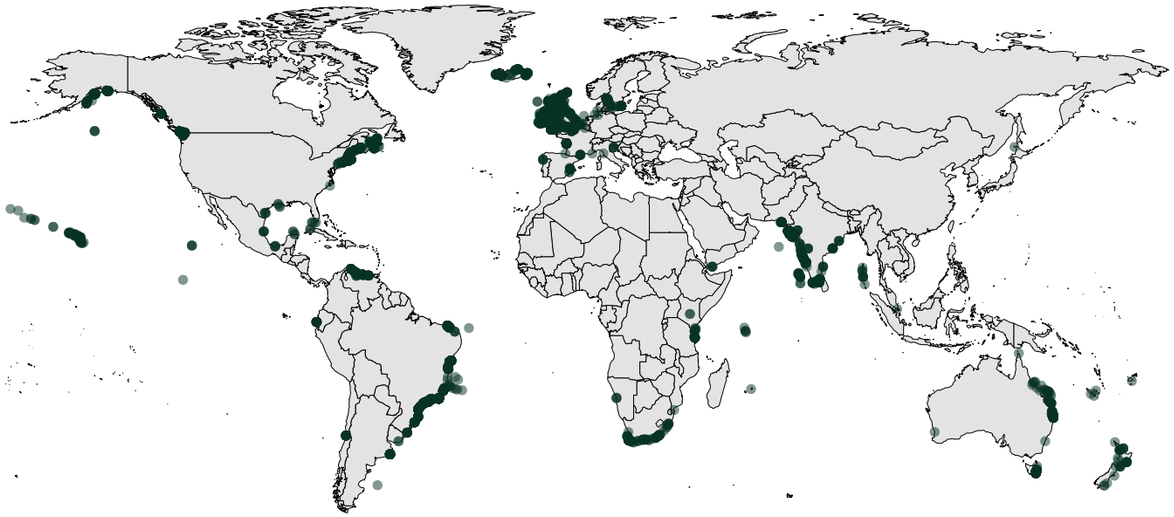
Ulva intestinalis[2]

Ulva arasakii[3]

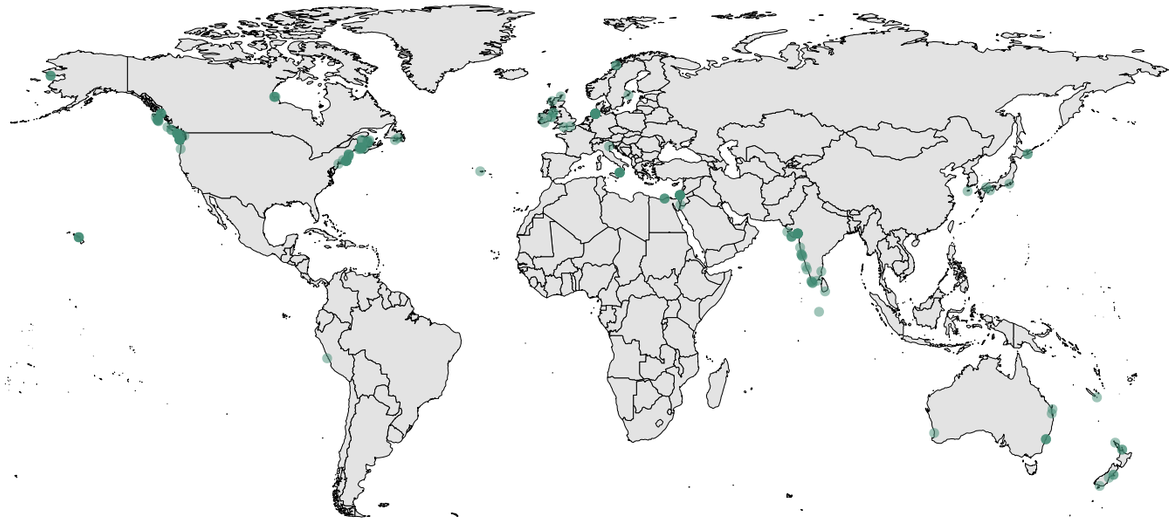
Ulva expansa[5]

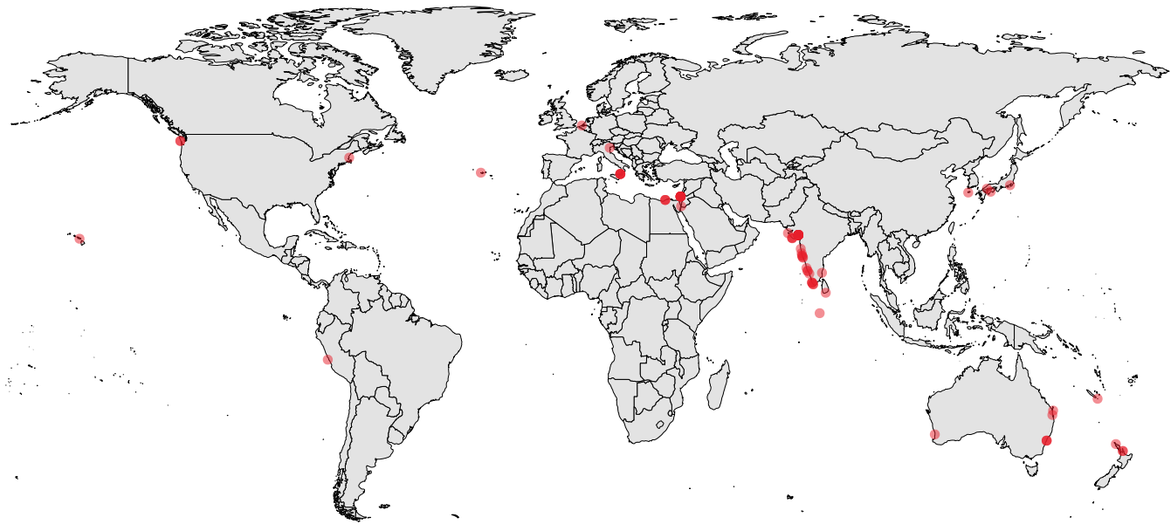
Ulva compressa[1]

OBIS records

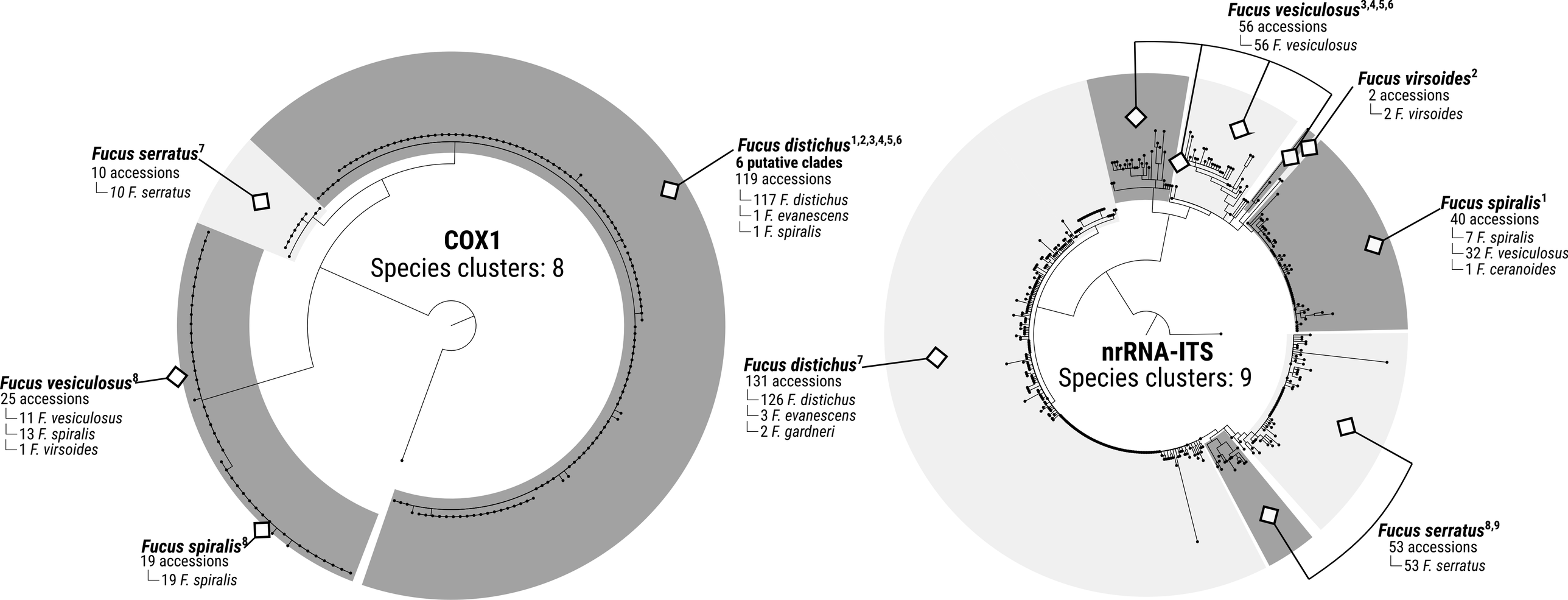


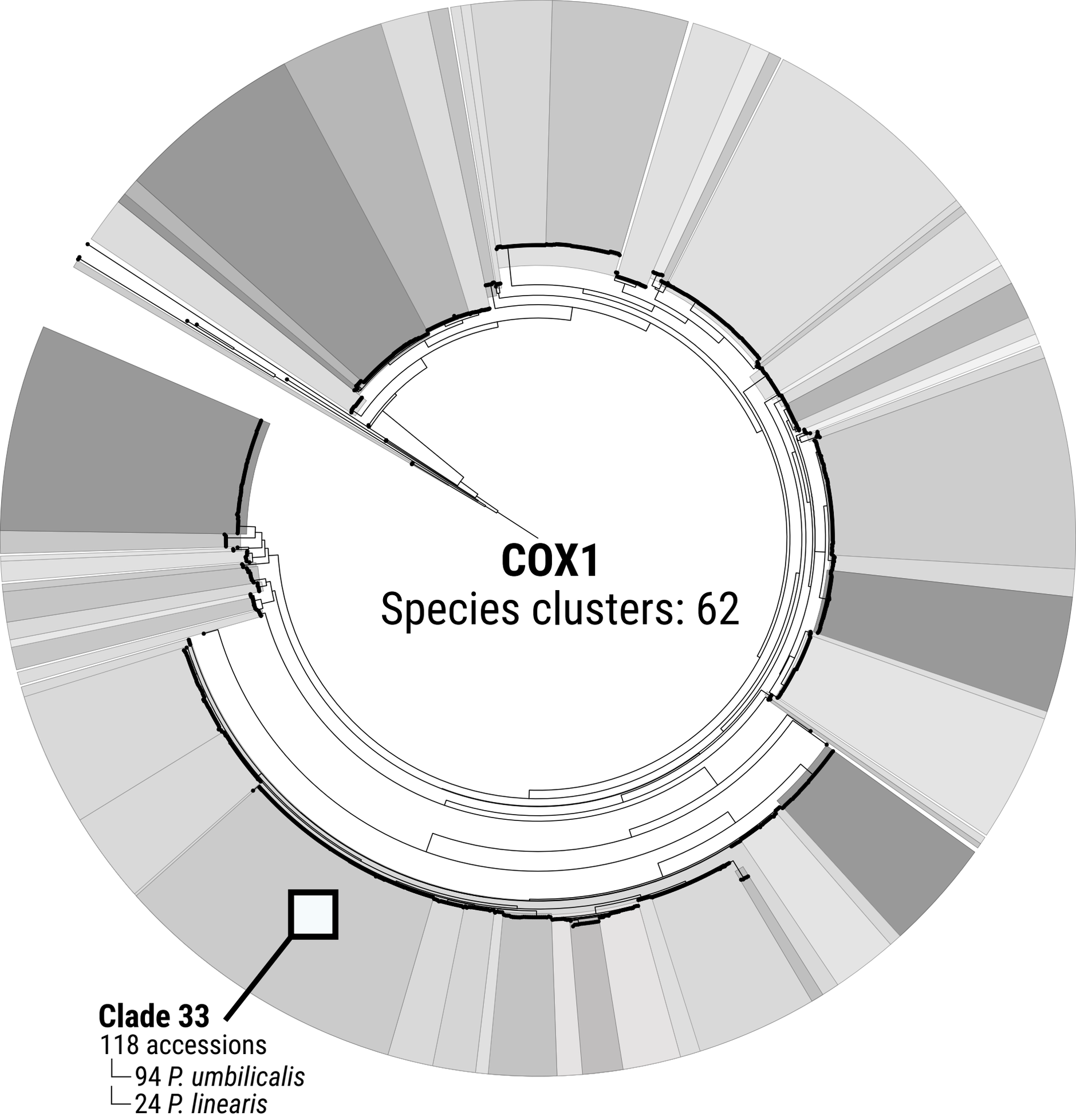
NCBI records

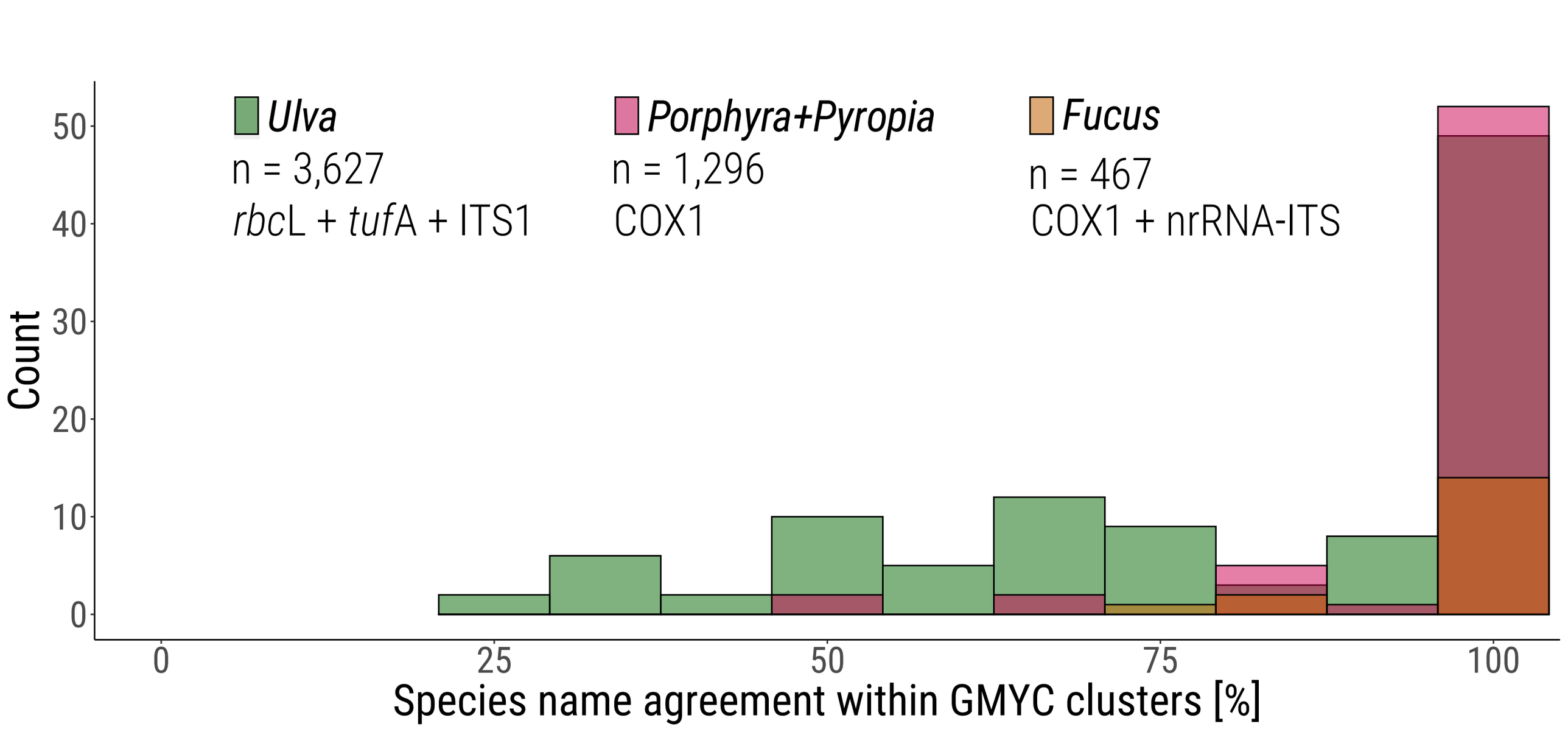




NCBI records - reanalysed


COX1
Species clusters: 8
Fucus distichus1,2,3,4,5,6
6 putative clades
119 accessions
117 F. distichus
1 F. evanescens
1 F. spiralis
Fucus serratus7
10 accessions
10 F. serratus
Fucus vesiculosus8
25 accessions
11 F. vesiculosus
13 F. spiralis
1 F. virsoides
Fucus spiralis8
19 accessions
19 F. spiralis
nrRNA-ITS
Species clusters: 9
Fucus vesiculosus3,4,5,6
56 accessions
56 F. vesiculosus
Fucus virsoides2
2 accessions
2 F. virsoides
Fucus spiralis1
40 accessions
7 F. spiralis
32 F. vesiculosus
1 F. ceranoides
Fucus distichus7
131 accessions
126 F. distichus
3 F. evanescens
2 F. gardneri
Fucus serratus8,9
53 accessions
53 F. serratus


COX1
Species clusters: 62
Clade 33
118 accessions
94 P. umbilicalis
24 P. linearis


Ulva
n = 3,627
rbcL + tufA + ITS1
Porphyra+Pyropia
n = 1,296
COX1
Fucus
n = 467
COX1 + nrRNA-ITS
Count
0
10
20
30
40
50
0
25
50
75
100
Species name agreement within GMYC clusters [%]