Exhaustive reanalysis of barcode sequences from public repositories highlights ongoing misidentifications and impacts taxa diversity and distribution: a case study of the Sea Lettuce.

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

[1]NUI Galway
[2]Station Biologique de Roscoff
[3]Scottish Association for Marine Science
[4]University of Aveiro
[5]Universidade de Aveiro
[6]University of Porto Interdisciplinary Centre of Marine and Environmental Research
[7]Wageningen University & Research

November 24, 2020

Abstract

Sea Lettuce (*Ulva spp.*; Ulvophyceae, Ulvales, Ulvaceae) is an important ecological and economical entity, with a worldwide distribution and is a well-known source of near-shore blooms blighting many coastlines. Species of *Ulva* are frequently misidentified in public repositories, including herbaria and gene banks, making species identification based on traditional barcoding hazardous. We investigated the species distribution of 295 individual distromatic foliose strains from the North East Atlantic by traditional barcoding or next generation sequencing. We found seven distinct species, and 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 in the NCBI database. We estimate that 21% of the entries pertaining to foliose species are misannotated. In the extreme case of *U. lactuca*, 65% of the entries are erroneously labelled specimens of another *Ulva* species, typically *U. fenestrata*. In addition, 30% of *U. rigida* entries are misannotated, *U. rigida* being relatively rare and often misannotated *U. laetevirens*. Furthermore, *U. armoricana* and *U. scandinavica* present as being synonymous to *U. laetevirens*. 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. Altogether, exhaustive taxonomic clarification by aggregation of a library of barcode sequences highlights misannotations, and delivers an improved representation of *Ulva* species diversity and distribution. This approach could be easily adapted to other taxa.

1. Introduction

Species of the genus *Ulva* the type and name-bringing genus of the Ulvophyceae, Ulvales and Ulvaceae, are a genetically diverse group of green macroalgal species ubiquitous in the worlds ocean, brackish and even in freshwater environments. Some 400 *Ulva* species have been described of which about 90 are currently recognised taxonomically. Many of these taxa are uncommon or rare and only about 25 have been frequently reported (Guiry & Guiry 2020; unpublished). The morphology of *Ulva* species can be grouped into two general types, one containing foliose "sheet-like" species (distromatic foliose blades commonly known as

the "Sea Lettuce"), and another with tubular or partially tubular thalli (monostromatic tubes formerly recognized as *Enteromorpha* genera). Phenotypic plasticity between tubular and foliose morphotypes can be based on both abiotic and biotic factors (Wichard et al., 2015). Due to such phenotypic plasticity in response to environmental factors, and relatively subtle morphological differences between species, particularly in the distromatic foliose taxa (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 of *Ulva* spp. relies on the amplification and sequencing of specific loci in the genome, most often using 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; Miladi et al., 2018; O'Kelly, Kurihara, Shipley, & Sherwood, 2010). The sequences obtained from those barcodes are then compared with sequences publicly available in repositories, such as the National Center for Biotechnology Information (NCBI). Typically, NCBI sequences with $> 99\%$ identity compared with the query sequence are considered as belonging to the same species and used for the classification of the species of the sequenced individual. The risk in such case is that the species attributed to the matching sequence present in the NCBI can be erroneous, leading to the misidentification of the investigated individual. A most recent example in distromatic foliose species was highlighted by (Hughey et al., 2019), where the authors sequenced the holotype of *Ulva lactuca* Linnaeus, as well as the holotype of *Ulva fenestrata* Postels & Ruprecht, and discovered a serious misapplication of names in subsequent published work. Given those findings, a significant number of *Ulva lactuca* individuals reported in the literature in the North East Atlantic (Biancarosa et al., 2017; Loughnane, McIvor, Rindi, Stengel, & Guiry, 2008; Steinhagen, Karez, & Weinberger, 2019), actually belong to *Ulva fenestrata* . Since the extent of misannotated NCBI entries has not to date been characterised, it can be difficult to assign a species to a sequence with confidence when the sequence of interest closely matches NCBI entries with several species names.

Here, we employed DNA barcoding (*rbc* L, *tuf* A, ITS1) on 185 strains of distromatic foliose *Ulva* from the North East Atlantic, and used the next-generation sequencing data and species delimitation from our previous study containing another 110 strains (Fort et al., 2020), 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 highlight the extent of misannotations in the sequences 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.

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, country of origin, species and GPS coordinates 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 chloroplastic *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 et al., 2020). 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 $4{,}000g$ 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 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 13$^{\text{th}}$ of July 2020), in addition to the sequences from our previous study [Fort et al, 2020]. 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 et al, 2020), 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 -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.

2.4 Phylogenetic analyses

We used both maximum likelihood and Bayesian MCMC phylogenetic analyses for the ITS1, *rbc* L and *tuf* A datasets. 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 (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; Fort et al., 2020), with a General Mixed Yule Coalescent model (Fujisawa & Barraclough, 2013; Pons et al., 2006) in BEAST, and 50 millions Markov Chain Monte Carlo (MCMC). Convergence was confirmed 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/*).

2.5 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 S2** and **S3**), restricted to vouchers assigned in our analysis as belonging to the eleven main distromatic foliose *Ulva* species [i.e *U. australis* Areschoug, *U. fenestrata* Postels & Ruprecht, *U. lactuca* Linnaeus, *U. gigantea* (Kützing) Bliding, *U. laetevirens* Areschoug, *U.ohnoi* M.Hiraoka & S.Shimada, *U. rigida* C.Agardh, *U. pseudorotundata* M.Cormaci, G.Furnari & G.Alongi [[?] *U. rotundata* Bliding], *U. expansa*(Setchell) Setchell & N.L.Gardner, *U. arasakii* Chihara and*U. ohiohilulu* H.L.Spalding & A.R.Sherwood]. Publications associated with NCBI entries missing GPS coordinates and/or location of origin were manually searched to retrieve GPS coordinates where available. Where the voucher did not contain coordinates or publication, we used the affiliated address of the authors. 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** . Latitudes and longitudes were grouped in multiple of two degrees to merge entries from similar geographical areas, creating windows of ˜ 12,100 km^2. 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 et al, 2020, as well as all available *rbc* L sequences from*Ulva* entries (see Materials and Methods). From the *rbc* L alignment, we generated a Maximum Likelihood phylogenetic tree containing 1,245 sequences. GMYC analysis revealed the presence of 24 clades containing more than two sequences (confidence interval 19-28) (**Fig. 2**). Since several species names have been found to be synonymous, we used the species names listed in **Table 1** as our reference**.** Of these, ten belong to obligatory distromatic foliose species, namely *Ulva arasakii* , *Ulva pseudorotundata* , *Ulva expansa* , *Ulva fenestrata* ,*Ulva australis* , *Ulva gigantea* , *Ulva ohnoi* ,*Ulva lactuca* , *Ulva rigida* and *Ulva laetevirens* . The GMYC species delimitation, however, failed to discriminate between five species. *Ulva laetevirens* and *U. rigida* are shown to be conspecific, as well as a single clade containing both *U. lactuca*and *U. ohnoi,* and another clade containing *U. pseudorotundata* 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 185 samples sequenced in this study originating from the North East Atlantic belong to seven distinct clades, with 19 samples identified as*U. pseudorotundata* , 21 samples as *U. fenestrata* , 47 as*U. australis* ,

13 as *U. gigantea* , 2 as *U. ohnoi* , 12 as *U. rigida* and 63 as *U. laetevirens* . Because most holotypes of those species have not been sequenced yet (apart from *Ulva fenestrata* ; (Hughey et al., 2019)), we based our species attribution with comparisons from sequences from (Fort et al., 2020) and from entries present in the NCBI database, with the caveat that indeed the species names could change once holotype sequences become available.

All *U. pseudorotundata* samples showed >99% similarity with those described by (Fort et al., 2020) and the five*rbc* L *U. pseudorotundata* vouchers in the NCBI database from (Biancarosa et al., 2017; Loughnane et al., 2008), all originating from Ireland and Brittany, except for one strain from Spain (Fort et al. 2020). *Ulva fenestrata* sequences were annotated based on their >99% identity with the holotype of *U. fenestrata*(Hughey et al., 2019). *Ulva australis* was identified based on >99% identity with those of (Kraft, Kraft, & Waller, 2010) and (Heesch et al., 2009). This clade shows no discrepancy, with all individuals of either *U. australis* or *U. pertusa*(which are synonymous; see **Table 1**) being present within the clade. For *U. gigantea* (13 individuals in this study, 10 in (Fort et al., 2020) and 3 from (Loughnane et al., 2008)), all entries appear well annotated. 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 specimen), 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* Forsskal and one *U. taeniata*(Setchell) Setchell & N.L.Gardner. Most entries originate from the same unpublished population set (number 452119310). Next, the *U. rigida* clade contains 12 strains from this study, 29 described in (Fort et al., 2020), as well as 20 *U. rigida* entries from the NCBI, described in (Heesch et al., 2009; Rautenberger et al., 2015) as well as NCBI entry EU484408 from (Loughnane et al., 2008). Finally, the *U. laetevirens* clade containing 138 strains appears more problematic, with several cases of likely species misidentification. This clade contains 63 individuals from this study, 38 individuals from (Fort et al., 2020), and four *U. laetevirens* entries [two from (Kraft et al., 2010), Port Phillip, South Australia, the type locality of *U.laetevirens* (Guiry & Guiry, 2020), and two from China (Du et al., 2014)]. However, 21 entries in the *U. laetevirens* clade were assigned as *U. rigida* , indicating a likely common confusion between *U. rigida* and *U. laetevirens* . Interestingly, all six *U. armoricana* entries and all five *U. scandinavica*entries also cluster within the *U. laetevirens* clade, with all six *U. armoricana* entries showing 100% identity with *U. laetevirens* individuals (e.g., NCBI voucher EU933943 and most entries from this study and (Fort et al., 2020)). Two out of five *U. scandinavica* entries are indistinguishable from *U. laetevirens*ones, and the other three possess a single polymorphic site. Altogether, *U. armoricana* and *U. scandinavica* are more likely to be synonymous with *U. laetevirens* .

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** . Finally, the *U. lactuca* clade contains 58 sequences, 48 of which are annotated as either *U. lactuca* or *U. fasciata* (which are synonymous, **Table 1**), two erroneous *U. ohnoi* , four erroneous *U. reticulata* , two *U. taeniata,* one *U. beytensis* one *U. laetevirens,*all from the same population set (# 452119310, same as for *U. ohnoi* misannotated sequences).

Concerning other species, *U. compressa* Linnaeus and *U. intestinalis* Linnaeus are well defined, with no misidentifcation for *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* .

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

We performed the same analysis using the *tuf* A barcode (**Fig. 3** , **Fig. S2** 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, as expected, the *U. fenestrata* clade shows the same name misapplication with *U. lactuca* , with 225 individuals, 21 in this study, 11 in (Fort et al., 2020), 107 *U. fenestrata* entries and 86 *U. lactuca*

entries. The *U. lactuca* clade contains 16 sequences, ten of which annotated as *U. fasciata* [[?] *U. lactuca*]. *Ulva australis* and *U.gigantea tuf* A clades appears well defined, with no name misapplication, similar to the *rbc* L results. *U. pseudorotundata tuf* A sequences only contain individuals described in this study and in (Fort et al., 2020). *Ulva ohnoi* is generally well circumscribed, with 92 *U. ohnoi* vouchers (J. H. Kang et al., 2019; Krupnik et al., 2018; Lee, Kang, & Kim, 2019; Melton et al., 2016; Miladi et al., 2018), but also the presence of three *U. fasciata* entries and one *U. prolifera* entry, all from unpublished studies (population set number 452119404, same as for *U. lactuca* misannotated sequences). Interestingly, while 18 *U. rigida tuf* A sequences are present in the NCBI dataset (Steinhagen et al., 2019; Wolf, Sciuto, Andreoli, & Moro, 2012), all belong to the *U. laetevirens* clade. Indeed, the *U. rigida*clade only contains sequences from this study and (Fort et al., 2020). The *U. laetevirens* clade contains 59 strains identified in this study, 38 identified previously in (Fort et al., 2020), and 25 from the NCBI database (J. H. Kang et al., 2019; Mao, Kim, Wilson, & Yarish, 2014; Miladi et al., 2018; Saunders & Kucera, 2010). 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* , 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.

3.3 Analysis of all ITS1 sequences from public repositories

Finally, the analysis was repeated on the ITS1 barcode dataset (**Fig. 4** , **Fig. S3** 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. fenestrata* clade (21 in this study, 11 in (Fort et al., 2020)) only contains a single *U. fenestrata* NCBI entry AY260562, annotated/submitted by (Hillary S Hayden & Waaland, 2002), and 19 erroneous *U. lactuca* . In addition, this clade contains a single *U. californica* entry, likely misannotated. We found two misannotated *U. pseudorotundata* sequences, which belong to the *U. australis* clade. *Ulva pseudorotundata* and *U. gigantea* ITS1 sequences are only described in the present study and (Fort et al., 2020). The *U. ohnoi* clade contains three sequences from this study, as well as 23 *U. ohnoi* vouchers [described in (Hiraoka et al., 2004; Lawton, Mata, de Nys, & Paul, 2013; Monotilla et al., 2018)]. Three erroneous *U. fasciata* sequences were found, all from unpublished sources. The *U. rigida* clade contains 50 sequences (14 from this study, 29 from (Fort et al., 2020), the rest from (Coat et al., 1998; Hillary S Hayden & Waaland, 2002; Hillary S. Hayden & Waaland, 2004; Tan et al., 1999). An extraneous *U. lactuca* voucher was found within the *U. rigida* clade. Finally, the *U. laetevirens* clade contains 134 sequences with 62 from this study, 38 from (Fort et al., 2020), and only six of NCBI entries annotated as *U. laetevirens* (described by (Du et al., 2014; Kraft et al., 2010; Mao et al., 2014)). Of the other sequences, 21 are annotated as *U. rigida* , and one as *U. fenestrata* . As for*rbc* L results, we found *U. armoricana* and *U. scandinavica* within the *U. laetevirens* clade, all of which show 100% identity with most other *U. laetevirens* sequences. Regarding *U. expansa* , the clade contains four vouchers, two annotated as *U. expansa* and two as *U. lobata* , with 100% identity within the clade.

With regard to narrow-tubular species, the Linza-Procera-Prolifera (LPP) complex is poorly delimited, with NCBI entries of all three species intertwined within a large clade. 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 *U. tepida* Y.Masakiyo & S.Shimada clade contains several entries annotated as *U. intestinalis.*

4) Discussion

4.1) Species delimitation using three common barcodes should be avoided.

In this study, we endeavoured to exhaustively 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 amount of sequences yields relatively large species clusters confidence intervals. For instance, using *rbc* L did not allow to separate certain taxa that were previously shown to be separate species (Fort et al., 2020; Hiraoka et al., 2004; Hughey et al., 2019), such as *U. rigida* and *U. laetevirens* or *U. ohnoi* and *U. lactuca* . Such a discrepancy is inherent to large-scale species delimitation analyses when using a 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 we were able to separate *U. laetevirens* and *U. rigida* , which is in agreement with our previous study (Fort et al., 2020). *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 clusters. 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 groups. Similarly, *U. linza* ,*U. compressa* , *U. intestinalis* and *U. prolifera*clades are separated into several sub-groups. Altogether, precise species delimitation analysis on single barcodes used here and elsewhere in the literature should be avoided, and should ideally be performed on a larger amount of genetic information, such as full organellar genomes (Fort et al., 2020).. Hence, outside of the six foliose species studied in (Fort et al., 2020), and those with sequenced organellar genomes (Cai et al., 2018; Cai, Wang, Zhou, He, & Jiao, 2017; Hughey et al., 2019; Hughey, Miller, & Gabrielson, 2018; Wang, Cai, Zhou, He, & Jiao, 2017; Zhou, Wang, Zhang, Cai, & He, 2016), precise species delimitation of all the available barcode data of the *Ulva* genera should be avoided.

Interestingly, the number of "species names" in the entries from the NCBI dataset is 56, with nine of which being classified as synonymous. Out of the 47 unique species names remaining, this analysis, despite its limitations, found ˜40 species clusters containing more than two sequences, thus broadly agreed the present number of species described in NCBI. These numbers are significantly lower than that of the number of currently accepted species taxonomically (90 according to (Guiry & Guiry, 2020)). 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 the site they originate from resampled, as the NCBI database likely only contains a subset of all *Ulva* species.

4.2) 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. 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 significant. Indeed, ˜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 classified as *U. lactuca* . We support the use of *U. fenestrata* holotype described by (Hughey et al., 2019) as the baseline for this species (**Table 2**). Furthermore, our study shows that *U. rigida* and *U. laetevirens* are also commonly misannotated in public repositories, which was recently 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 et al., 2020). Unfortunately, holotype specimens for both species are not available. Where necessary, lectotype, neotype and corresponding epitype specimens should be designated and sequenced to conclusively assign species names to each clade. In addition, we show here that *U. armoricana*(described by (Dion, De Reviers, & Coat, 1998)) and *U. scandinavica* (Battelli & Tan, 1998) are synonymous of *U. laetevirens* . *Ulva scandinavica* was previously thought to be synonymous with *U. rigida* (Loughnane et al., 2008) but this analysis suggests it is in fact synonymous to *U. laetevirens* . Hence, caution should be taken when assigning species names to those vouchers.

Overall, the analysis of large foliose *Ulva* species showed ˜21% of misannotated entries in the NCBI database, and 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** to**S3** the trees of all three barcodes in to allow researchers to use the search function of pdf viewers for searching specific vouchers and identifying to which clade they belong. Alternatively, **Table S3** contains all of the accession numbers of the foliose species highlighted here, as well as our proposed species attribution. Finally, we propose in **Table 2** a list of reference NCBI accessions for all three barcodes of the eleven large foliose *Ulva* species. 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* .

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; E. J. 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. Since hybrids between *U. linza* and *U. prolifera* species have been shown to be viable (Xie et al., 2020), the matter of species delimitation within that clade remains to be resolved. 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.

Altogether, the potential for misidentifications in public repositories should not be overlooked, and similar analyses could be performed on different taxa of interest to highlight misannotated sequences/species and provide the scientific community with a list of appropriate reference accessions or proposed re-annotations.

4.3) Global distribution of foliose *Ulva spp.*

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. 5**). Notably, there is a lack of genetic data from those *Ulva* species for the African coast, South America and South East Asia/North East Oceania, as is the case for many taxa. While this does not preclude the presence of those eleven *Ulva* species, it shows that more genetic data originating from those areas and released in public repositories is needed to precisely characterise the global distribution of foliose *Ulva*species. For instance, foliose *Ulva* is used commercially in South Africa for more than a decade but its species identity is currently unknown (Bolton, Robertson-Andersson, Shuuluka, & Kandjengo, 2009), with no NCBI sequence originating from this country. It is also likely that new/un-sequenced species are present within those areas.

From the locations containing genetic information, *U. laetevirens*and *U. australis* are the most widely distributed, with the highest number of unique specimens sequenced (**Fig. 5** , top panel, **Fig. 6**). They are present in the Atlantic coast, the Mediterranean sea, East Asia, the Americas, Australia and New Zealand. Interestingly however, *U. australis* is conspicuously absent from the Irish and British coasts, despite a large number of individuals in nearby regions such as Brittany and the Netherlands. Indeed, the density of sampling and sampling dates in Ireland in particular (**Table S1**) likely allowed for an exhaustive capture of the species diversity of foliose *Ulva* in Ireland, and no *U. australis* were recovered. Thus, the absence of *U. australis* in Britain and Ireland remains to be explained

Three species are present in narrow latitudes (**Fig. 5** , middle panel: *U. fenestrata* is present in the cold to temperate waters of the North and South hemispheres, while *U. lactuca* and *U. ohnoi* favour warmer waters. Finally, the other six large foliose*Ulva* species are more geographically localised (**Fig. 5** , bottom panel), with *U. pseudorotundata* restricted to the East Atlantic coast, *U. gigantea* to the North Atlantic, *U. rigida* to the East Atlantic coast and New Zealand, *U. expansa* to the North-East Pacific, and finally *U. arasakii* and *U. ohiohilulu* restricted to the sea of Japan and Hawai'i, respectively.

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 contradiction with our results (**Fig. S4**). This poses a significant challenge since 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, Hilario, Munro, Blunt, & Calado, 2016). This highlights the need to both improve the identification of *Ulva* species and change the European food regulation by inclusion of the *Ulva* species which are effectively consumed at present under the name of "*Ulva lactuca* ".

Taken together, the distribution results indicate that *Ulva*species might have varying degrees of specialization, from the most tolerant to a wide range of environments (e.g., *U. laeteviren* s and *U. australis*), to more localised species such as *U. pseudorotundata* . It also highlights the worldwide co-occurrence of foliose *Ulva* species, since most areas contain more than one species.

ConclusionsDue to the increasingly large number of sequences being deposited in public repositories, it is becoming important to revisit and reassess the genetic information of taxa of interest, to highlight ongoing species identification issues and potentially reassign names to previously uncharacterised synonymous species. Here, we investigated all *Ulva* sequences in the NCBI public repository for three common barcodes, as a contribution to clarify the species composition and annotation of the *Ulva* genus worldwide, with a focus on large foliose species. This dataset can be used for future species identification, accession validation and classification purposes, to ensure accurate representation of the species names within the databases. Interestingly, the worldwide distribution of foliose *Ulva* species presented here show that only a few species are present throughout the world's oceans, and that most foliose species belong to specific geographical locations. The analytical framework described here in detail could be transferred to any other taxa of interest, particularly those that 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 semitubular species and/or species with no representative in this study. Shaded and 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 vouchers 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 semitubular species and/or species with no representative in this study. Shaded and colored clades represent species clusters determined using GMYC. Full trees including bootstrap values and bayesian posterior probabilities are available in**Fig. S2** .

Fig. 4: Maximum Likelihood phylogenetic tree of 1,245 *Ulva spp.* ITS1 sequences, and description of the vouchers 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 semitubular species and/or species with no representative in this study. Shaded and colored clades represent species clusters determined using GMYC. Full trees including bootstrap values and bayesian posterior probabilities are available in**Fig. S3** .

Fig. 5: Worldwide species distribution of the eleven large distromatic foliose *Ulva* species. Top: widely distributed species, middle: species present in narrow latitudes, bottom: localised species. The size of the circles is proportional to the number of specimen entries within a ˜12,100 km^{2}radius. The map contains 1336 unique specimens.

Fig. 6: Number of unique NCBI specimens sequences available for each foliose species.

Table 1: Names and synonyms used in this study.

Table 2: Proposed reference sequences for foliose *Ulva*species

Species	NCBI ITS accession	NCBI *rbc*L accession	NCBI *tuf*A accession	Reference
Ulva australis	MT894708	MT160564	MT160674	Fort et al, 2020
Ulva laetevirens	MT894611	MT160587	MT160697	Fort et al, 2020
Ulva rigida	MT894503	MT160586	MT160696	Fort et al, 2020
Ulva ohnoi	AB116031	AB116037	MK992234	Shimada et al, 200
Ulva pseudorotundata	MT894650	MT160609	MT160719	Fort et al, 2020
Ulva gigantea	MT894472	MT160566	MT160676	Fort et al, 2020
Ulva lactuca	EU933990 +	MK456395	MF172082 +	Kraft et al, 2010, H
Ulva fenestrata	MT894736	MK456393	MK456404	Fort et al, 2020, Hu
Ulva arasakii	AB097650	AB097621	MK992126	Shimada et al, 200
Ulva expansa	MH730161	MH746437	MH731007	Hughey et al, 2018
Ulva ohiohilulu	NA	NA	KT932996	Spalding et al, 201

+Annotated as *U. fasciata*

Supplementary Data

Table S1: List of samples, species, GPS coordinates and NCBI vouchers 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: Complete ML and Bayesian trees of *tuf*A alignment.

Fig. S3: Complete ML and Bayesian trees of ITS1 alignment.

Fig. S4: Comparison of OBIS and NCBI records of *U. lactuca*.

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

File S2: Python script to retrieve GPS coordinates from a list of NCBI accession numbers.

File S3: 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/*)

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.

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* .

Battelli, C., & Tan, I. H. (1998). *Ulva scandinavica Bliding,(Chlorophyta): a new species for the Adriatic Sea.* Paper presented at the Annales: Annals for Istran and Mediterranean Studies.

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.

Bolton, J., Robertson-Andersson, D., Shuuluka, D., & Kandjengo, L. (2009). Growing *Ulva* (Chlorophyta) in integrated systems as a commercial crop for abalone feed in South Africa: a SWOT analysis.*Journal of Applied Phycology, 21* (5), 575-583.

Cai, C., Wang, L., Jiang, T., Zhou, L., He, P., & Jiao, B. (2018). The complete mitochondrial genomes of green tide algae *Ulva flexuosa*(Ulvophyceae, Chlorophyta). *Conservation Genet. Resour., 10* (3), 415-418.

Cai, C., Wang, L., Zhou, L., He, P., & Jiao, B. (2017). Complete chloroplast genome of green tide algae *Ulva flexuosa*(Ulvophyceae, Chlorophyta) with comparative analysis. *PloS one, 12* (9).

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.

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.

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

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., McHale, M., Cascella, K., Potin, P., Usadel, B., Guiry, M. D., & Sulpice, R. (2020). 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, n/a* (n/a). 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.

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.

Hayden, H. S., & Waaland, J. R. (2004). A molecular systematic study of*Ulva* (Ulvaceae, Ulvales) from the northeast Pacific.*Phycologia, 43* (4), 364-382. doi:10.2216/i0031-8884-43-4-364.1

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., 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.

Hughey, J. R., Miller, K. A., & Gabrielson, P. W. (2018). Mitogenome analysis of a green tide forming *Ulva* from California, USA confirms its identity as *Ulva expansa* (Ulvaceae, Chlorophyta).*Mitochondrial DNA Part B, 3* (2), 1302-1303.

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.

Kang, J. H., Jang, J. E., Kim, J. H., Byeon, S. Y., Kim, S., Choi, S. K., . . . Lee, H. J. (2019). Species composition, diversity, and distribution of the genus Ulva along the coast of Jeju Island, Korea based on molecular phylogenetic analysis. *PloS one, 14* (7), e0219958.

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.

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.

Lawton, R. J., Mata, L., de Nys, R., & Paul, N. A. (2013). Algal bioremediation of waste waters from land-based aquaculture using*Ulva* : selecting target species and strains. *PLoS One, 8* (10), e77344.

Leal, M. C., Hilario, 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.

Lee, H. W., Kang, J. C., & Kim, M. S. (2019). Taxonomy of *Ulva*causing blooms from Jeju Island, Korea with new species, *U. pseudo-ohnoi sp. nov.* (Ulvales, Chlorophyta). *Algae, 34* (4), 253-266. doi:10.4490/algae.2019.34.12.9

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., Lopez-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., Davila 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.

Mao, Y., Kim, J. K., Wilson, R., & Yarish, C. (2014). The appearance of Ulva laetevirens (Ulvophyceae, Chlorophyta) in the northeast coast of the United States of America. *Journal of Ocean University of China, 13* (5), 865-870.

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.

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.

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.

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.

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., Hohna, 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.

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

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

Vranken, S., Bosch, S., Pena, 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.

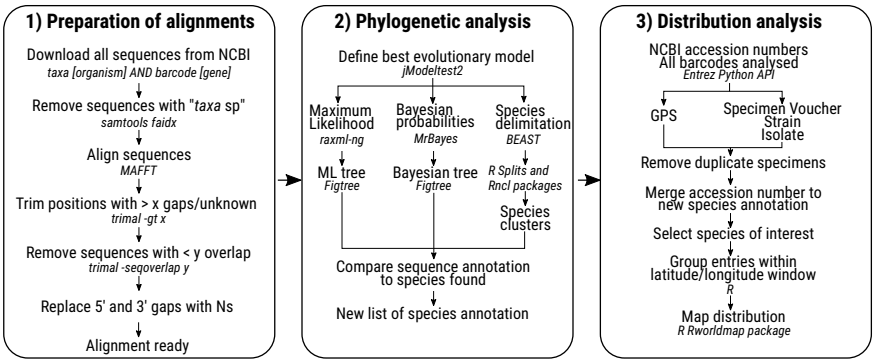
Wang, L., Cai, C., Zhou, L., He, P., & Jiao, B. (2017). The complete chloroplast genome sequence of *Ulva linza* . *Conservation genetics resources, 9* (3), 463-466.

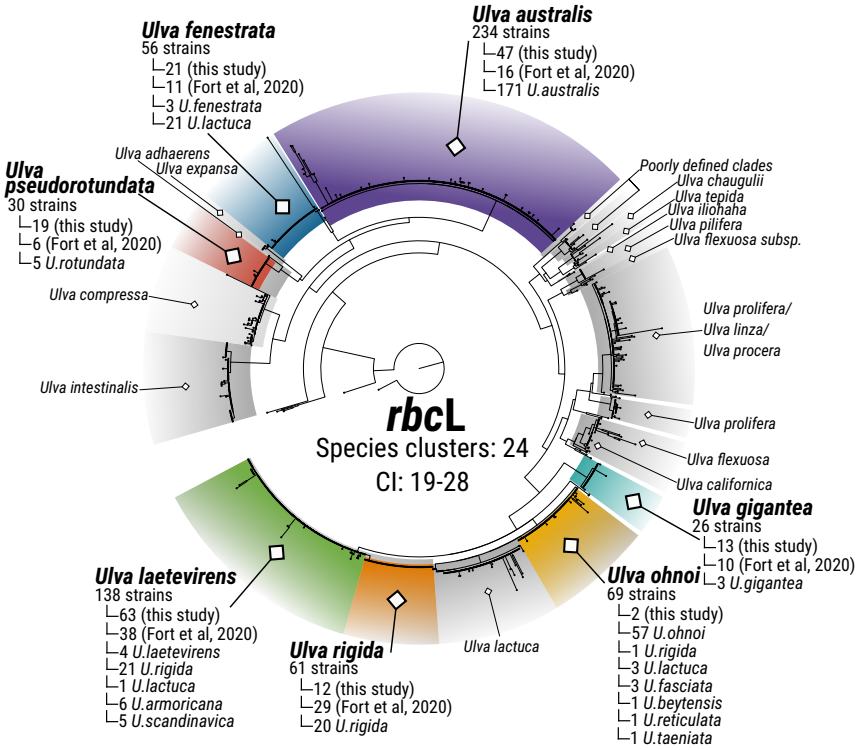
Wichard, T., Charrier, B., Mineur, F., Bothwell, J. H., Clerck, O. D., & Coates, J. C. (2015). The green seaweed *Ulva* : a model system to study morphogenesis. *Frontiers in Plant Science, 6* (72). doi:10.3389/fpls.2015.00072

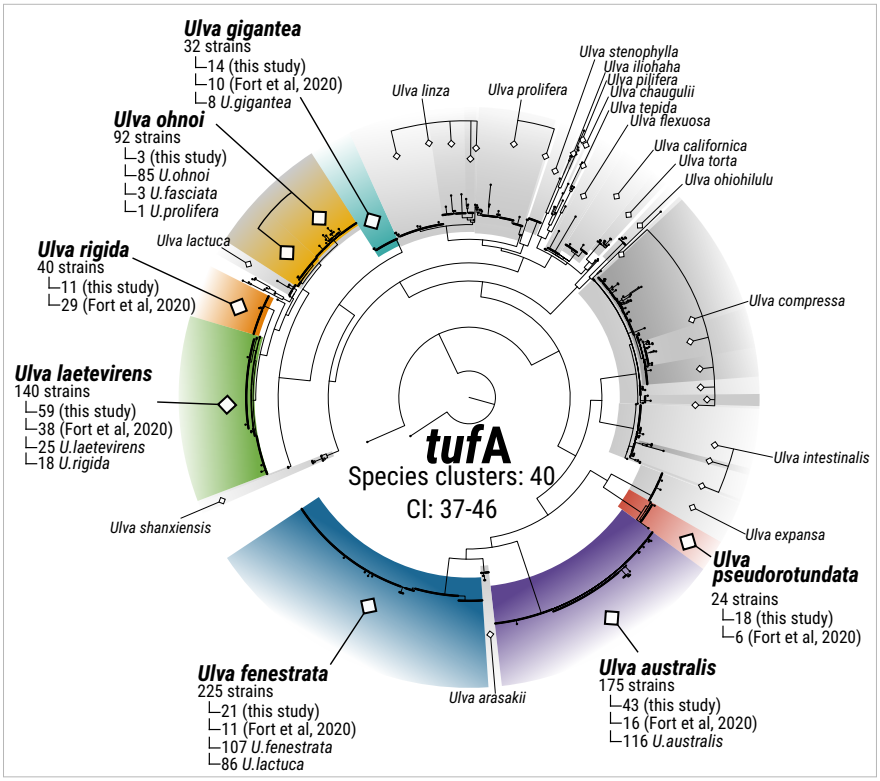
Wolf, M. A., Sciuto, K., Andreoli, C., & Moro, I. (2012). *Ulva*(Chlorophyta, Ulvales) biodiversity in the North Adriatic Sea (Mediterranean, Italy): cryptic species and new introductions.*Journal of Phycology, 48* (6), 1510-1521.

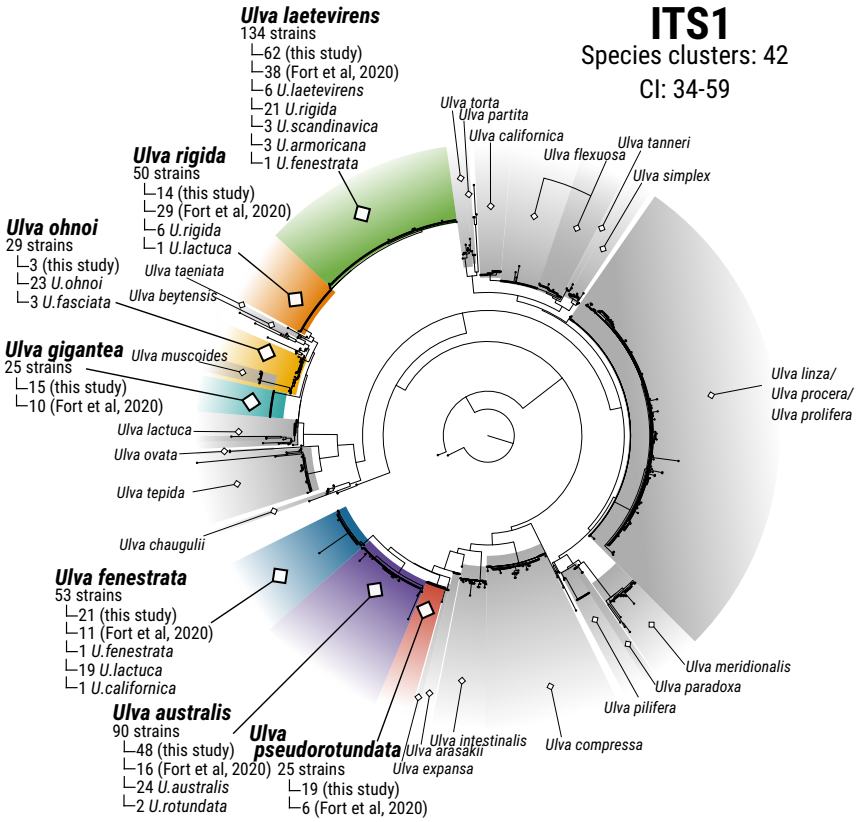
Xie, E., Xu, R., Zhang, J., Xu, C., Huang, B., Zhu, W., & Cui, J. (2020). Growth characteristics of hybrids produced by closely related *Ulva* species. *Aquaculture, 519* , 734902. doi:https://doi.org/10.1016/j.aquaculture.2019.734902

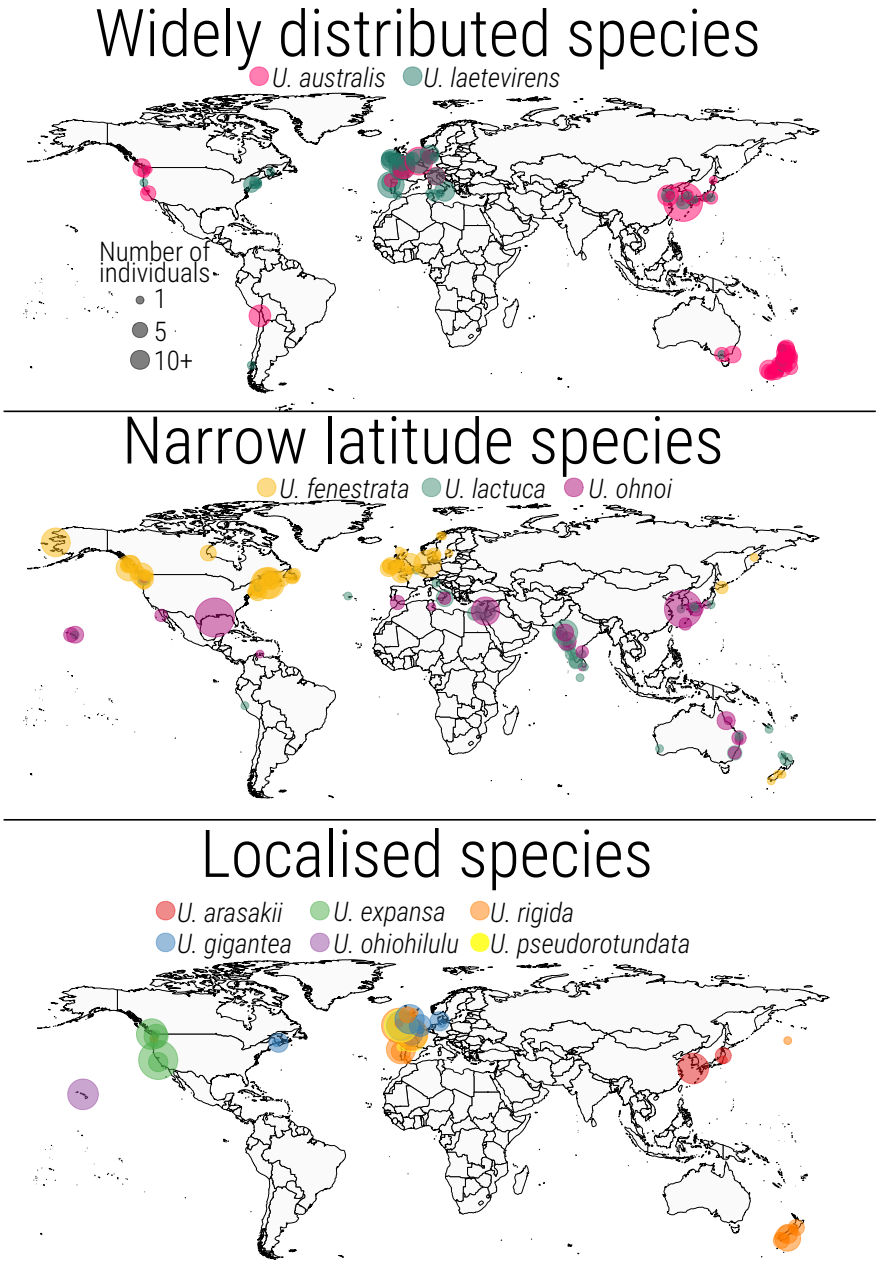
Zhou, L. J., Wang, L. K., Zhang, J. H., Cai, C., & He, P. M. (2016). Complete mitochondrial genome of *Ulva linza* , one of the causal species of green macroalgal blooms in Yellow Sea, China. *Mitochondrial DNA Part B-Resources, 1* (1), 31-33. doi:10.1080/23802359.2015.1137806

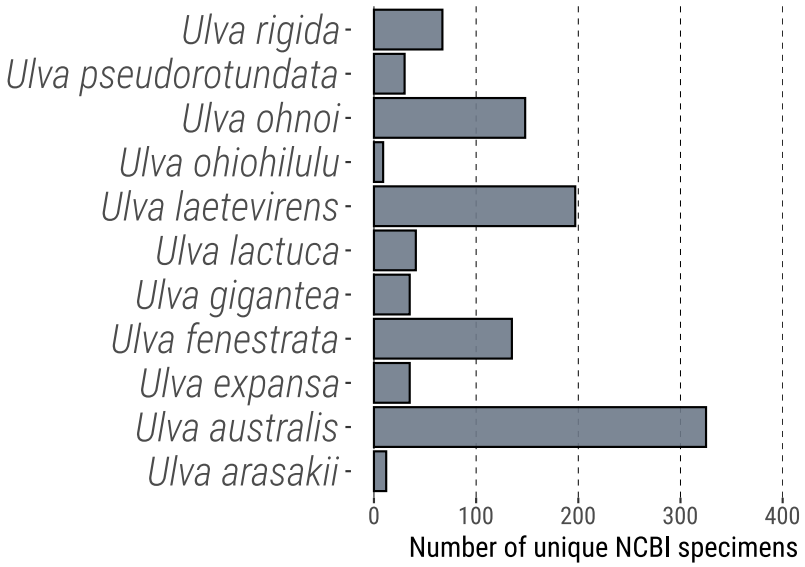












Hosted file

`Table 1.ods` available at https://authorea.com/users/378378/articles/494914-exhaustive-reanalysis-of-barcode-sequences-from-public-repositories-highlights-ongoing-misidentifications-and-impacts-taxa-diversity-and-distribution-a-case-study-of-the-sea-lettuce

Hosted file

`Table 2.xlsx` available at https://authorea.com/users/378378/articles/494914-exhaustive-reanalysis-of-barcode-sequences-from-public-repositories-highlights-ongoing-misidentifications-and-impacts-taxa-diversity-and-distribution-a-case-study-of-the-sea-lettuce