Open Research Online

The Open University's repository of research publications
and other research outputs

Phylogenetics and Phylogeography in the Planktonic Diatom Genus *Chaetoceros*

Thesis

How to cite:

De Luca, Daniele (2019). Phylogenetics and Phylogeography in the Planktonic Diatom Genus Chaetoceros. PhD thesis. The Open University.

For guidance on citations see FAQs.

© 2019 The Author

Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

THE OPEN UNIVERSITY, MILTON KEYNES (UK)

STAZIONE ZOOLOGICA ANTON DOHRN, NAPLES (IT)

School of Life, Health and Chemical Sciences

Doctor of Philosophy (Ph.D.)

Phylogenetics and Phylogeography in the Planktonic Diatom Genus *Chaetoceros*

Daniele De Luca (M.Sc.)

Personal ID: F3576948

Director of Studies

Dr. Wiebe H.C.F. Kooistra

Stazione Zoologica Anton Dohrn, IT

External Supervisor

Prof. Dr. Christine A. Maggs

Bournemouth University, UK

Internal Supervisor

Dr. Diana Sarno

Stazione Zoologica Anton Dohrn, IT

September 2019

Abstract

The initial aims of this thesis were to assess the systematics of the planktonic diatom genus *Chaetoceros* and the phylogeographic patterns of selected species in this genus across spatial and temporal scales. As expected in every experiment, some initial ideas have been pursued as they were; others have taken a different route and led to different questions. Consequently, the systematics of *Chaetoceros* has become a multigene phylogeny and a revision of the classical taxonomic scheme for the family Chaetocerotaceae (Chapter II). Then, the phylogeographic approach, initially meant as a Sanger sequencing of a few genes from specimens collected around the world, turned into the analysis of the *C. curvisetus* cryptic species complex by using an approach which combines haplotype networks and metabarcoding data (Chapter IV). The mapping of this complex against a temporal metabarcoding dataset (MareChiara, Gulf of Naples, IT) has become a story of concerted evolution and has been extended to different *Chaetoceros* species and supported by a single strain 18S-V4 high throughput sequencing (Chapter V). Amid these experiments, the potential of metabarcoding data for biological recording was explored and tested in the whole genus *Chaetoceros* to assess diversity and distribution (Chapter III). Such data were integrated with classical ones from public repositories and literature and used to produce, among the other results, distribution maps of *Chaetoceros* species.

To my Mom,

For having taken on so many burdens,

allowing me to fully dedicate myself to this thesis

For having endured me during my rainy days

For having surrendered to the constant entry of books in the house:

one day all these books will drive us away.

Acknowledgements

It is almost impossible to thank all the people who made my time at the SZN an unforgettable and great experience. Most of them contributed with a simple “buongiorno”, a smile, or a chat on the terrace, and to them goes my gratitude. Others occupy a more special place in my heart. Therefore, my “special” thanks go…

…To my Director of Studies, Dr. Wiebe Kooistra, who gave me the opportunity to carry out my Ph.D. research and the freedom to develop my ideas, showing me the way, provide guidance and alternative routes, and not dictating the rules. This is the best gift I could have ever received.

… To my internal supervisor, Dr. Diana Sarno, for the taxonomic knowledge she shared with me, and her passion for diatoms, which she transmitted to me and which inspired me to immerse myself fully into the diversity of diatoms and, in particular, *Chaetoceros*.

… To my external supervisor, Prof. Dr. Christine A. Maggs, for the time spent together during her visits at the SZN, talking about my project, science in general and ordinary life. Her suggestions and enthusiasm were pure positive energy that prompted me to do better and better.

…To Dr. Roberta Piredda, for the invaluable time she invested in me teaching, discussing, listening and for her friendship. Thanks to her, I have domesticated the untamed beasts of “Linux”, “batching”, “server” and “big data”, I was introduced to the –omics world, and looked with new eyes to the ecology of communities. I own her everything I know about metabarcoding and much more about evolutionary methods for the study of biodiversity. Someone said: “together you are war machines”. Nothing more true.

… To Carmen Minucci, for the training in the lab and the constant help in carrying it on.

… To Dr. Grazia Quero, for being my friend and the best office partner I could ever have desired. Thank you for introducing me to the metagenomics world and for the constant help every time the machine took control over the man.

… To Dr. Isabella D'Ambra, for the morning chats and the courage with which she faces life every day. You are an example of success.

… To the old and new people of the Procaccini lab, my first house at the SZN: Gabriele, Miriam, Chiara, Emanuela, Domenico, Gaetano, Lazaro, Marlene, Lucia, Roberto, Jessica... Among you are head of departments, researchers, post-docs, but there is no "Dr" because you will be always my second family first, and then the brilliant scientists and lovely persons I have ever met. I am proud to have been a Procaccinidae.

… To Dr. Graziano Fiorito, my first "official" mentor at SZN for the esteem and affection he has always shown to me. I was proud to be "his compulsive student".

… To Pamela, Elena, Giovanna, Ruth, Caitlin, Paola and Andrea of the "OctoLab" for being my friends and all the lunches and "mozzarella days" we had together. Pam, we started this adventure almost together several years ago and I am proud of how far you went. The spirit of my "Pam!" along the corridors will accompany your for many years to come.

… To the people of the SBM: the big boss, Dr. Elio Biffali, who made me a better molecular biologist with his original way of teaching; Dr. Pasquale De Luca, for adopting me as "cousin" and witnessing all my feats; Elvira Mauriello and Raimondo Pannone, for all the support in molecular biology techniques and the chats and laughs together. I own you thousands of DNA sequences!

… To Gabriella Grossi and Daniela Consiglio for the support in the life as Ph.D. student and student's representative the first, and as reference student for the invited speakers the latter. Thanks for the passion you have for your work and the kindness showed.

Table of Contents

Chapter I - Introduction

Chapter II - Inferring the evolutionary history of Chaetocerotaceae

List of Figures

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

List of Tables

Chapter II

Chapter IV

Chapter V

List of acronyms and abbreviations

bp	base pair
BS	Bootstrap Support
COI	Cytochrome Oxydase subunit I
GBIF	Global Biodiversity Information Facility
LSU	Large Subunit ribosomal DNA
OBIS	Ocean Biogeographic Information System
OSD	Ocean Sampling Day
PP	Posterior Probability
*psb*A	photosystem II protein D1 gene
*rbc*L	ribulose bisphosphate carboxylase Large subunit gene
sp.	species (singular)
spp.	species (plural)
SSU	Small Subunit ribosomal DNA

Chapter I

Introduction

1.1. Diversity, the hallmark of living organisms

The most peculiar characteristic of life on Earth is its sheer and unfathomable diversity. It is so huge and widespread that, to date, we are still remarkably uncertain about how many species exist on Earth (May, 1992; Stork, 1993). Guesstimates vary by several orders of magnitude and show remarkably different levels of uncertainty, from 3-100 million species (May, 2010) to around 8.7 million (Mora et al., 2011). In contrast to our persistent uncertainty about the extent of biological diversity, our comprehension of the mechanisms giving rise to it is becoming well understood. The generation of biological diversity (biodiversity; Wilson, 1988), at least at the gene- and species levels, is an intrinsic property of evolution (Wilson, 1992). As outlined by Darwin in the *Origin* (1859), evolution, in its essence, is nothing but "descent with modification", a dualistic process ruled by chance and anti-chance factors (Mayr, 1997). The former is the modification, i.e. whatever heritable change affecting both the genotype and phenotype, produced "by chance" in the form of mutation, recombination or any other mechanism. One of the latter is natural selection, a mechanism that favours certain individuals over others with particular genetic attributes to fit in a specific environment, i.e., "survival of the fittest." Such combination of chance and "anti-chance" factors gives evolution flexibility and "goal-directedness" and makes it so powerful (Mayr, 1963).

Darwin (1859) clearly recognised that evolution follows two kinds of trajectories, one across time and another one across space (see also Wilson, 1992). The former, called *phyletic evolution* is a process of gradual change within a single population or metapopulation of a species, resulting in the gradual transition of an ancestral species into a new one (anagenesis). As consequence, phyletic evolution does not imply speciation and can be seen as a line from ancestral to descent taxa (Fig. 1.1A). This is the mode of evolution Darwin (1859) had in mind when explaining the action of natural selection. On the contrary, when changes occur over time and are spread over space, (e.g. if populations of a species become different and occupy different ecological niches or geographic areas),

then an ancestral species splits gradually into two or more daughter species (Fig. 1.1B). This process, called *divergent evolution* (Gulick, 1888) is nothing but the change within a lineage accompanied by speciation (cladogenesis), and can be drawn as a branching tree. Divergent evolution implies that several species may all exist at the same time and is the source of biodiversity that, using the words of E. O. Wilson (1992) is "a collateral effect of evolution".

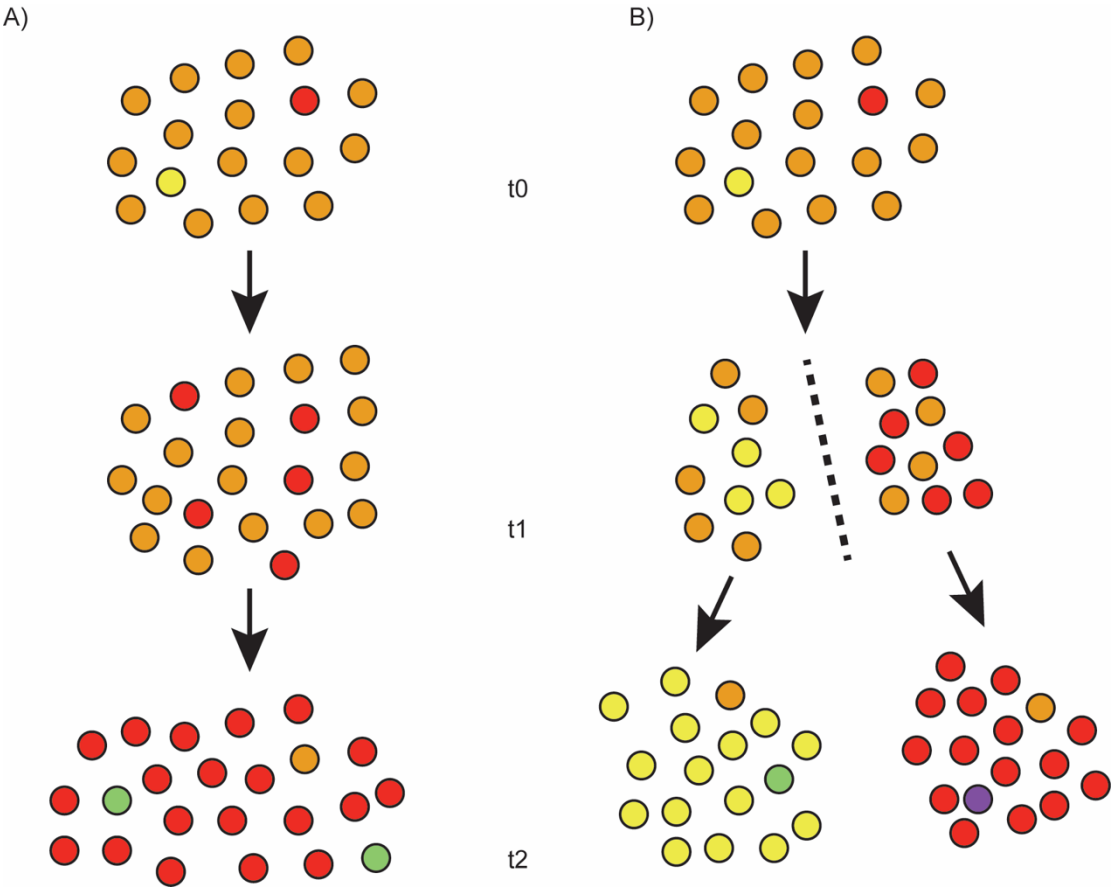


Fig. 1.1. Modes of evolution across space and time. (A) phyletic evolution (anagenesis); (B) divergent evolution (cladogenesis). Dots represent individuals of a population (species). Colours refer to variation among individuals.

As stated above, modification is an integral a part of evolution as common descent. Whatever change not transmitted to the offspring has no consequence for evolution. Furthermore, the rate at which mutations arise (mutation rate) is highly variable across organisms (Drake, 1999; Baer et al., 2007) and the spread of such mutations within a

population through gene flow, drift and selection following the rules of population genetics, eventually determines the evolution of organisms. All these processes affect the way we perceive "species". The adaptation to different ecological niches due to divergent natural selection and sexual reproduction have been indicated as the main factors responsible of genetic and phenotypic discontinuities between populations (Maynard Smith and Szathmáry, 1995; Coyne and Orr, 1998). Such discontinuities can also be observed at the molecular level. As already pointed out in the nineteenth century by the English geneticist William Bateson in its work *Materials for the study of variation* (1894), the variation of biological characters can be both continuous and discontinuous, and that "variations of a discontinuous nature may play a preponderant part in the constitution of a new species". But what if characters change slowly over time? What if there are no discontinuities? How do we recognise species in that case?

1.1.1. Diatom diversity and evolution

Diatoms are one of the most successful contemporary groups of photosynthetic eukaryotic microorganisms. The estimated number of species ranges from guesstimates of over 200,000 species (Mann and Droop, 1996) to more conservative, morphology-based estimations of 12,000 species (Guiry, 2012) and metabarcoding-based estimations of 4,748 OTUs (Malviya et al., 2016). Molecular phylogenetic studies (e.g. Medlin and Kaczmarska, 2004; Theriot et al., 2010) group diatoms in three main categories: the ancestral and paraphyletic radial centrics, the likewise paraphyletic multipolar centrics and the most recent and monophyletic pennates. Radial centrics seem to consist of few remnant lineages, with *Leptocylindrus* constituting an important bloom former in coastal regions all over the world (Nanjappa et al., 2014). Multipolar centrics contain two highly diverse clades, the Thalassiosirales and the Chaetocerotales, whilst the pennates are the most diverse group (Not et al., 2012).

Diatoms have a diplontic life cycle (i.e. they spend the most of their life as diploid organisms and form haploid gametes through meiosis) consisting of a long period (up to several years) during which cells divide mitotically and a brief period (a few days) during which sexual reproduction takes place. Mitotic divisions are constrained by the siliceous cell wall (frustule). Indeed, as vegetative growth goes on, two sibling cells with different valve size are produced: one identical to the parent cell and the other one slightly smaller. MacDonald (1869) and Pfitzer (1869) described this size diminution process independently over a century ago. Although some taxa have been shown to possess both physiological and morphological modifications to overcome size diminution (e.g. von Stosch, 1965; Round, 1972; Drebes, 1977; Gallagher, 1983), in most species size restoration occurs only through sexual reproduction (Edlund and Stoermer, 1997). Lewis (1984) argued that size reduction cannot be a mere consequence of the cell division mechanism in presence of a siliceous frustule, but must have an adaptive significance. He suggested that size reduction might act as a chronometer for sex, allowing diatoms to spread the high costs of sexual reproduction over several or many years (Lewis, 1984; Mann, 1989; Mock and Medlin, 2012).

Centric diatoms reproduce sexually through oogamy (i.e. production of non-motile, large cells, the oogonia, and small, motile ones, the sperm cells), whilst pennates do so therough isogamy (i.e. gametes of similar morphology differing in allele expression in one or more mating-type regions). In radial centrics the sperm cells do not include chloroplasts whereas in multipolar centrics (such as *Bacteriastrum* and *Chaeroceros*) the sperm carries a plastid, but this plastid generally does not contribute to the zygote. Instead in pennates, both gametes usually add a plastid to the zygote (Round et al., 1990). Centrics (including *Bacteriastrum* and *Chaeroceros*) are monoecious meaning that single strain can produce male and female gametes, and fertilise itself. This is a setback for crossing experiments and affects strain identity over time. Instead, pennates are dioecious, meaning that strains from the opposite mating type are needed to produce the next generation (Round et al., 1990).

Several diatom lineages can form resting stages in the form of spores or resting cells (McQuoid and Hobson, 1996). Resting cells are cells with condensed cytoplasm, less pigments in shrunken chloroplasts and thicker frustule than vegetative cells, but with the same shape of the vegetative cell (Lund, 1954). Instead, spores have a markedly different morphology from vegetative cells (i.e. a thick frustule, often ornamented with spines and other protuberances; Round et al., 1990).

There is increasing evidence that the evolutionary diversification of diatoms has taken place predominantly within sexual lineages. Indeed, there are no evidences of families or genera, even the most species rich, in which all the species are asexual or parthenogenetic (Mann, 1999). Natural diatom populations often consist of many fewer genotypes than individuals (except perhaps after mass auxosporulation), as a result of mitotic division and colony fragmentation (Richardson, 1995). In this scenario, it is possible that mutations occurring in a single individual are perpetuated quickly and indefinitely, eventually establishing a new species, as hypothesised by Goldschmidt (1940) and Small (1950). However, this mechanism is unlikely to work in sexual species (and most diatoms probably fall into this group) since a new species, arising through a macromutation in a single individual, would initially contain only one sex (Mann, 1999). However, in some species (e.g. *Chaetoceros*) individual strains can form male and female gametes (Round et al., 1990), making this scenario more likely to happen.

Divergence and speciation can apparently take place rapidly in diatoms, over periods of 1,000 - 10,000 years or less (e.g. Theriot, 1992). The availability of several diatom genomes has made it possible to estimate diversification rates at molecular level (Mock and Medlin, 2012). The bipolar centric diatom *Thalassiosira pseudonana* and the pennate *Phaeodactylum tricornutum*, known to have been diverging for about 90 million years, diverged of about 45% in their genomes (differences based on the percentage of amino acid identity of 4267 orthologous gene pairs, Bowler et al., 2008). In multicellular eukaryotes, a similar divergence (about 40%) is found between *Homo sapiens* and the

pufferfish *Takifugu rubripesis*, which have been diverging for the last 550 million years (Bowler et al. 2008). This comparison demonstrates that unicellular eukaryotes diverge faster than multilcellular counterparts, which might be related to a higher mutation rate, larger effective population size, and shorter generation times (Mock and Medlin, 2012). In multicellular organisms with small population size, advantageous mutations are rare and disrupted by sexual reproduction (Bromham, 2011); furthermore, the life histories longer than unicellular eukaryotes further reduce the diversification rates (Mock and Medlin, 2012).

1.2. The species problem

Despite the fact that species are the fundamental units of biology, the dispute about how to define them is still ongoing. Mayr (1982) argued that most of the confusion about what constitutes a species is due to the application of the term "species" to two fundamentally different logical categories. The first of these includes the use of the word species as synonymous of "kind of", to describe natural phenomena or things, like the words 'planet' or 'moon' (Mayr, 1996). In case of living-things, we refer to it as *species as taxon*, i.e. individuals that exists in space and time and have a historical continuity (Hull, 1976; 1978). The other meaning of species is *as taxonomic category*, to which taxa can belong. In this sense, the problem of species refer to the species as category and to the way (attributes) such categories are defined (Mayr, 1982). The pre-Linnaean concept of (biological) species was similar to the one used for non-living things: a species was defined by a set of unchangeable or slightly variable characteristics that allow us to recognise it from other such species. Therefore, for each species there was a model or "type" organism to which all the others must conform to be considered as members of the same species. However, after Darwin's *Origin*, it became clear that biological species are not immutable entities: they constantly change, in space and time and at any detectable level. Consequently, a good concept of biological species must take into the account this

variability and consider not *the* difference but the *degree* of difference as a threshold for delimiting members of the same or different species (Mayr, 1982). But what are these differences (or characteristics), how to choose them and who does the choosing?

Over the centuries, philosophers, physicians, naturalists and many other (categories of) people have spotted and used different sets of properties useful to identify species and distinguish them from others. Wilkins (2011) stated that there currently are seven operational definitions of species (reviewed in de Queiroz et al., 2007) with 27 variations, three more in respect to the 24 counted in Mayden (1997). These seven definitions *identify* species according to different properties: for the purposes of this thesis, I will focus on two of them: the morphological (also called phenetic) species concept and the phylogenetic species concept. The morphological species concept is the one that all the people, familiar or not with the biological sciences, use in their daily life. It encompasses all the organisms (individuals) that share a similar morphology and assumes that a "type species", identifiable through a "type specimen", exists. Nowadays we do not use anymore the term "type species" but still use "type specimen" in taxonomy, although with a different meaning, to designate the specimen to which the name of a genus or subgenus is taxonomically associated. It expresses the way in which classical taxonomists work: they collect different specimens, if possible, from different sites or regions, look at similarities and dissimilarities in their various traits, and make hypotheses on their relatedness. Unless these phenotypic characteristics have a selective advantage (i.e. are determined by the environment), generally they are indicative of a common ancestry due to interbreeding among individuals of the same species (which in turn is the so called "biological species concept").

On the contrary, the phylogenetic species concept is far less obvious and out of reach of not-professionals. In its simplest version, derived from the original by Eldredge and Cracraft (1980), it indicates "the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent" (Cracraft, 1983). It clearly

shows Darwin's footprint and the core of its theory: the similarity by descent. However, less clear is how we detect such relationships of ancestry and descent. A few decades ago, this was achieved by means of phylogenetic trees. One school of tree builders applied cladistics using morphological characters and their states, considering only those states that were "derived from a common ancestor and shared across its descendants" (synapomorphies); the relationships among taxa were then determined looking at the clades in the phylogenetic tree. A competing school of tree builders applied distance-based or probabilistic methods to infer phylogenies, taking into the account all the available characters and their states. If variable characters are plentiful, both methods tend to gravitate onto similar tree topologies because phylogenetic signal is additive whereas invariable characters and noise do not add anything to that signal (Lemey et al., 2009).

Nowadays the characters and their states that we use are mostly at molecular level, in the form of nucleotide positions (characters) and their states (the four nucleobases: A, T, C and G) in case of DNA and amino acid positions and their 20 possible states (the 20 amino acids) in case of proteins. Therefore, the characters that we use to identify species are nucleotide / amino acid positions in coding and non-coding DNA regions / proteins. There is no competition between morphological (phenotypic) and molecular characters as there is no universal best marker; each one can be used for a particular aim.

It should be noted that phylogenies inferred from whatever information is at hand are mere incomplete hypotheses of the real, but unknown, evolutionary history. Not all the information at hand is equally useful for phylogeny reconstruction; as stated by Avise (2012), "because phylogeny is "the stream of heredity", only genetically transmitted characters are informative to phylogenetic estimation". Even the information on such characters is not, by definition, adding to the phylogenetic resolution because both morphological and molecular characters can be affected by convergence (independent changes in different lineages converging on a similar or identical outcome). Phylogenies

inferred using just a very few characters (no matter their nature) are also of restricted use because they may deviate markedly from the true evolutionary history (Mayr 1982).

1.2.1. Species concepts in diatoms

As for most of the taxa, the morphological species concept has dominated diatom taxonomy and systematics from its beginning (e.g. Van Heurck, 1896). Diatom species began to be described in the first half of the 1800s (Fig. 1.2) based on morphological characters observable in light microscopy (e.g. Agardh, 1830-32; Kützing, 1833; Ehrenberg, 1838; reviewed in Mann, 2010a). At that time, species were considered discrete and immutable entities, and so in those early descriptions there was almost no discussion of species concepts, nor intraspecific variation taken into the account (Mann, 1999). From 1859 onward, when the publication of Darwin's *Origin* brought to attention the importance of varieties in the formation of new species, diatomologists started describing a huge number of species and varieties (e.g. Grunow, 1879; reviewed in Mann, 1999).

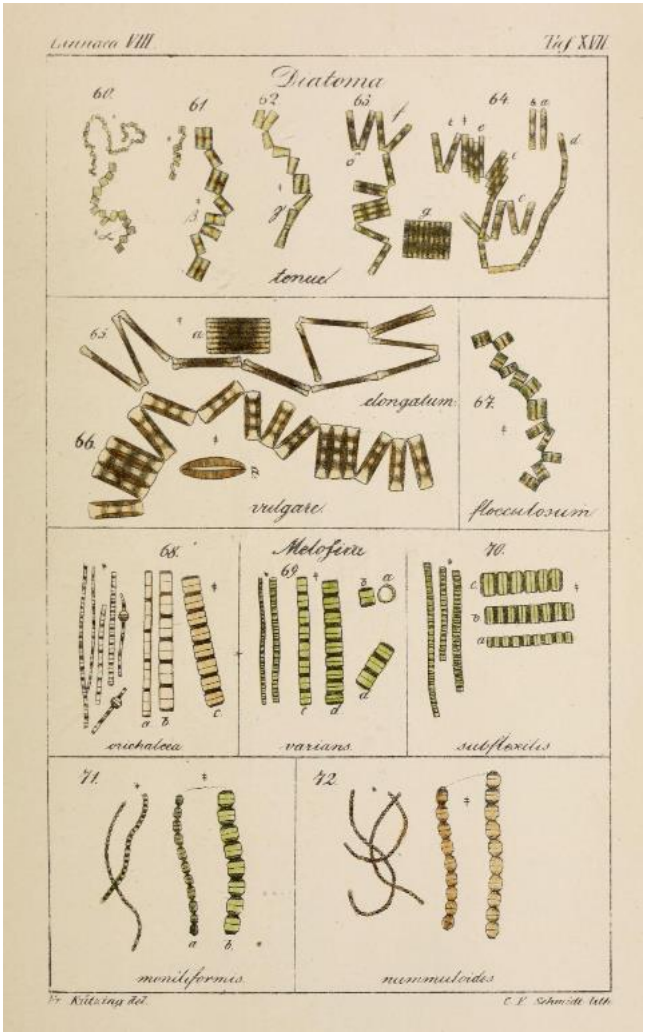


Fig. 1.2. Early example of diatom illustrations. Extracted from the "Synopsis Diatomearum, oder, Versuch einer systematischen Zusammenstellung der Diatomeen" by Kützing (1833).

Nowadays, the importance of intra-specific variation in diatoms is widely recognised and some authors have argued that no assertions should be made at the species level before considering it (Wood and Leatham, 1992). Intra-specific variation has been detected at different levels: i) at the individual (strain) level as clonal diversity (Rynearson and Armbrust, 2005; Ruggiero et al., 2018), heterozygosity (Rynearson and Armbrust, 2000), or phenotypic diversity (Gallagher et al., 1984; Gsell et al., 2012; Canesi and Rynearson, 2016); ii) at the population level as genetic differentiation among populations (Rynearson and Armbrust, 2004; Casteleyn et al., 2010) and phenotypic adaptation to different environments (Kremp et al., 2012). Empirical studies (reviewed in Godhe and Rynearson,

2017) have shown that intraspecific variation in diatoms is important in species' responses to environmental factors such as light, temperature, salinity and nutrient availability.

Despite the fact that identification of diatom species by means of morphological traits is hampered by such intra-specific diversity as well as by phenotypic plasticity and cryptic speciation, it is the easiest, quickest and cheapest way to identify taxa. The vast majority of diatom species descriptions are based on morphological features such as overall cell shape as well as the shape, size and ultrastructural detail of the siliceous cell wall elements comprising the frustule (Evans et al., 2007; Alverson, 2008).

To overcome the difficulties associated with morphological data, several attempts have been made to apply the biological species concept (BSC, Mayr 1942) to define and delimit diatom species (e.g. Amato et al., 2007; Kaczmarska et al., 2009; Quijano-Scheggia et al., 2009; De Decker et al., 2018). However, carrying out crossing experiments in diatoms is not without risk; for many species, the details of the sexual cycle have not been described, and even for those for which the phase is known, the triggers to commence sex are often not. Only in a few species does the experimenter have control over the process. Therefore, it remains impossible to test the validity of most diatom species under this concept. Furthermore, taxa to be tested are generally chosen based on the assumption that differences in their morphology are indicative of reproductive isolation, which often is not the case (Mann, 2010a).

The true revolution in diatom taxonomy and systematics arrived with the introduction of the phylogenetic species concept (Eldredge and Cracraft, 1980) and the use of molecular tools. The former provided a framework based on homology to analyse informative characters; the latter a quick, objective and cheap way to gather taxa. Homology can be estimated for both morphological and molecular characters: in the first case by a detailed knowledge of the morphological structure in question and its development; in the latter, aligning the bases (states) observed at the same positions (characters) in the nucleotide sequences obtained from different strains or specimens. The analysis of DNA or RNA

marker sequences has become particularly attractive for species discovery and classification in diatoms because homology is ascertained easily. For each of the many nucleotide positions in the sequence, the sequence markers exhibiting the appropriate level of variation can be chosen depending on the questions at hand, and the bases (states) at homologous positions (characters) can be scored easily, unambiguously and cost-effectively (Alverson, 2008; Mann, 2010a).

Nowadays diatom species are described and identified using a combination of morphological and genetic characters and, when available, information about their ecology and distribution. This approach, called *integrative taxonomy* (Dayrat, 2005) has the advantage of combining different properties of species and so providing a more robust framework for their inference.

1.3. Do species really exist?

Despite all the available species concepts, some authors have even questioned the existence of species. This is not to say that species *as taxa* are unreal; they are, but, *as category*, they are as artificial as all the other taxonomic ranks above the species level (Mishler, 1999). Species are the outcome of different evolutionary strategies and environmental factors; this is why an animal species cannot be compared with a plant or fungal species, not even to mention a prokaryotic "species." According to their motility, mating barriers, mutation rate, population size and many other factors, populations of a certain species are more or less dynamic and prone to changes over time and space. Generally, among botanists the attitude of denying biological species is prevalent (Bachmann, 1998; Mishler, 1999). Indeed, some plants form interspecific hybrids and, in some groups, phenotypic variation does not fit into discrete categories (Rieseberg and Willis, 2007). This was also the opinion of Darwin, who treated species as artificial constructs as genera, families and orders, asserting in the *Origin*: "I look at the term species as one arbitrarily given, for the sake of convenience, to a set of individuals closely

resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms" (Darwin, 1859). On the other hand, zoologists and especially the ones working on macrofauna, tend to recognise species because the reproductive barriers and morphological discontinuities are stronger or at least better defined. Among them, the most vehement defender of biological species was Ernst Mayr (1963; 1970; 1996; 1999), who happened to be an ornithologist.

Microbiologists have questioned the present species concepts and definitions used in microbiology (i.e. the morphological species concept for eukaryotic microorganisms and the DNA-based species definition for prokaryotic microorganisms) namely whether closely related isolates of bacteria or other microorganisms clustering into discrete groups have to be considered as different species (Spratt et al., 2006).

Within microbial eukaryotes, the sequencing of global samples of individuals of fungi and protists has shown that a vast diversity of genotypes exists and that this diversity is contained within relatively few morphologically recognised species that are globally distributed (Koufopanou et al., 2006; Spratt et al., 2006; Whitaker, 2006). In diatoms, it has been shown that these "phylogenetic species" that cannot be distinguished by morphology, are not simply the product of neutral genetic drift between geographically separate populations, because mating experiments have shown the presence of reproductive barriers (reviewed in Mann, 1999).

Two obvious differences underlying speciation between unicellular and multicellular organisms are that (i) population sizes tend to be much larger in the former and (ii) rates of homologous recombination can vary greatly, and lateral transfer can spread genes across large phylogenetic distances (Gogarten and Townsend, 2005; Spratt et al., 2006).

Since each species has its own evolutionary history, it is up to the specialist to ascertain if in its study system it is better to refer to species, or instead, to consider individuals, strains, populations, or meta-populations as the fundamental units of evolution. There is a common ancestry for all species, but not a common faith or definition.

My personal opinion on the matter embraces all the points discussed so far. I agree that every living organism is the product of unique and different historical, evolutionary and stochastic processes; therefore, in some cases it would be recommendable to refer to some taxa as species (e.g. when they form well distinct, homogeneous and recognisable reproductive units across time and space). In other cases (e.g. when reproductive barriers are labile), "species" are not arranged in discrete units and it would be better to consider lower taxonomic categories as the units of evolution below the species rank (e.g. metapopulations, populations, or even individuals or strains).

1.4. The need for classification

Classifications are arbitrary human constructs meant to group individual objects in categories based on a set of shared characters/properties. They are necessary when dealing with diversity, providing an effective tool for the storage and retrieval of information (Wheelis et al., 1992). However, the role of classifications is not limited to this. In his book "A System of Logic, Ratiocinative and Inductive" (1843), the English philosopher John Stuart Mill argued that a classification should serve to generate hypotheses. Similarly, Mayr (1969) supported this idea, but added that such hypotheses should have a strong likelihood of being true in order to produce reliable inferences.

The nature of a classification strictly depends on its intended function, and so there is no one "correct" classification (Wheelis et al., 1992). Biological classifications are just a kind, and their general meaning has changed profoundly over time. Early classifications of living things were utilitarian, attempting at explaining the plan of Creation (Agassiz, 1859), grouping organisms based on medical properties (Dioscoride, *De Materia Medica*) or physiological and reproductive traits (Aristotle), their "immutable essence" (Linnaeus) or simply analogies and differences (e.g. Cesalpino, *De plantis*, 1583) (Fig. 1.3). This has been their prevalent function until the formulation of the theory of common descent by Darwin and Wallace, when classification became phylogenetic (based on genealogy).

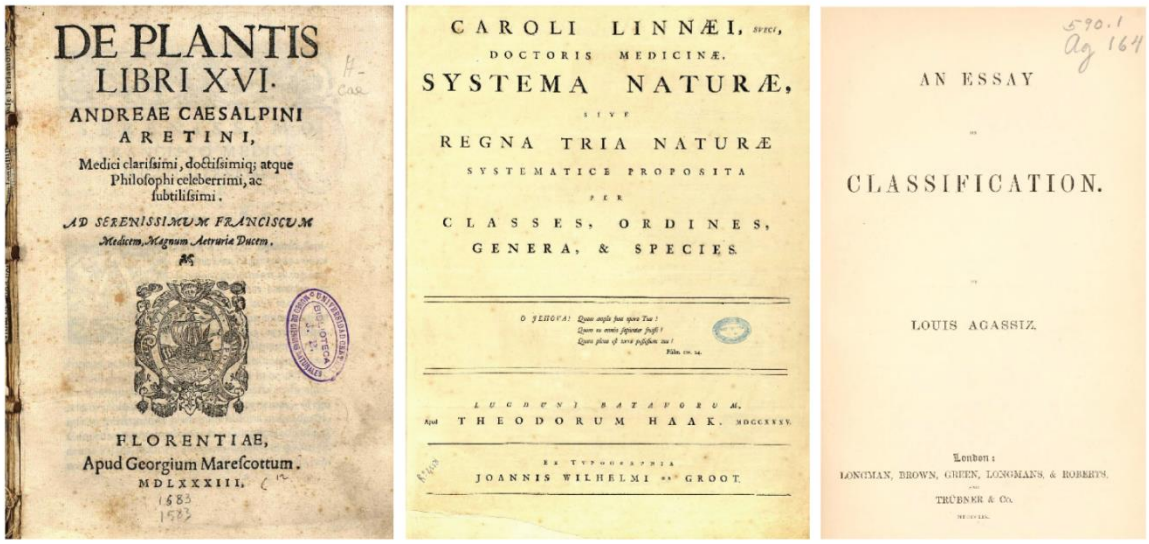


Fig. 1.3. Some classification essays from the sixteenth to the nineteenth century. From left to right: *De Plantis* (Cesalpino, 1583); *Systema Naturae* (Linnaeus, 1735); *An Essay on Classification* (Agassiz, 1859).

Despite some scientists as the French naturalist and mathematician Georges-Louis Leclerc, Comte de Buffon (1707-1788) and the English physician Erasmus Darwin (Charles Darwin's grandfather, 1731-1802) had considered the hypothesis that similar species could have derived from the same ancestral species, Charles Darwin was the first one to state it unequivocally (Mayr, 1982). An interesting outcome is that, despite being based on different perspectives, phenetic classifications often reflect phylogenetic ones. This is because similarity among organisms is fundamentally the result of common ancestry, as Darwin had understood. However, as outlined by Darwin himself in the *Origin* (1859) some organisms can be markedly different in morphology despite common descent because of radical modifications they underwent during evolution. A typical case is the one of birds and crocodiles (Arcosauromorpha), taxa that share a common ancestor but are extraordinarily different in their aspect due the different evolutionary trajectories they have followed. Nowadays, it is widely accepted that biological classifications should be both practical and phylogenetic, putting together organisms that have the greatest amount of shared characters due to common descent (Mayr, 1942; 1982).

At this point, it is important to highlight the distinction between classification and identification. As outlined by Simpson (1961) and Mayr (1969), classification and identification are two different things. Classification only involves groups, is based on the analysis of several different characters and searches for shared (synapomorphic) character states. On the contrary, identification is an individual-based process, requires the analysis of a few characters, and prefers to work with species-defining (autapomorphic) character states. Even if at the end of the identification process individuals are assigned to a particular group, this process cannot be called "classification" and so identification schemes are not classifications (Mayr, 1982). Both classification and identification are the object of study of taxonomy, whilst the study of the relationships among taxa is the field of systematics (Simpson, 1961; Mayr, 1969). Taxonomy and systematics have both benefited from the introduction of molecular approaches. In particular, in the last decade there has been a "renaissance" of taxonomic research due to introduction of DNA-based identification approaches (DNA barcoding).

1.5. DNA barcoding

The concept of DNA barcoding (i.e. the identification of taxa using short DNA sequences) is linked to one godfather, Paul Hebert, and one marker, the cytochrome oxidase subunit 1 (COI). The idea of identifying species with molecular markers can be traced back to the advent of molecular biology techniques in the early 1980's (Cristescu, 2014). Following the invention of PCR (Mullis and Faloona, 1987) and the development of universal primers (e.g. Kocher et al., 1989; Taberlet et al., 1991), Arnot et al. (1993) were the first to refer to "DNA barcodes" for species identification, amplifying the *Plasmodium falciparum* circumsporozoite (CS) gene to identify parasite stocks and lineages. However, the real revolution started when Hebert et al. (2003) proposed a system for the identification of animal taxa, called DNA barcoding, based on the use of a single gene marker, a 645 bp portion of the mitochondrial gene cytochrome c oxydase I (COI). A system based on DNA

barcodes provides both a way to *identify* taxa (e.g. the COI sequence for animals) and a way to *delimit* them from other such taxa (using a threshold of sequence divergence). In the same paper, Hebert et al. (2003) provided the example of 3% COI threshold dissimilarity value to delimit lepidopteran species and cite the > 2% cytochrome b threshold for vertebrates (Avise and Walker, 1999). The authors stressed multiple times that using a standard COI threshold for species delimitation, though appealing, should merely be considered as aid to the initial steps of the process. Unsurprisingly, the paper had its critics. Will et al. (2005) argued that "the real cutting-edge future for systematics and biodiversity research is integrative taxonomy, which uses a large number of characters, including DNA and many other types of data, to delimit, discover, and identify meaningful, natural species and taxa at all levels". They have not even spared the use of DNA barcoding for the identification of taxa, stating that "by emphasizing a single gene as a "universal barcode" (Powers, 2004), DNA barcoders are returning to an ancient, typological, single-character-system approach" (Will et al., 2005).

Rubinoff et al. (2006), instead, clarified that the opposition to DNA barcoding must not be intended as an opposition to the use of molecular tools in systematics and taxonomy in general. They argued that if DNA barcoding is intended as identification of species previously defined by other means, definition of new species by interpretation of DNA diversity as indicative of species diversity and operational units for ecological studies, there is no opposition to it. In spite of that, barcoding is actually functioning in a very different way from the original purpose for which it had been intended (i.e. identify known species and reveal those that are undescribed). Indeed, one of the criticisms raised by Rubinoff et al. (2006) is that barcoding papers have focused their attention on case studies where "cryptic species" were already suspected based on other sources of data (e.g. morphological or ecological data), thus violating the initial aim of identifying the unknown biodiversity.

From the practical point of view, an important limitation of DNA barcoding is that it relies on the assumption that speciation is generally accompanied by divergence in the sequence of the target gene. However, since sequence divergence is a stochastic process, some closely related species could not be resolved by barcoding, even if the chosen region of DNA evolves rapidly (Mann et al., 2010). Furthermore, some species might be impossible to barcode using a single gene simply because they are paraphyletic (Meyer and Paulay, 2005).

Whatever the pitfalls or drawbacks of a barcoding approach based on DNA sequences, it is unquestionable the impact that the idea of Paul Hebert and colleagues had on the study of biodiversity. Since its publication in 2003, their article has been cited more than 9000 times and it has opened the way to a new field of research. Even if the original idea of a "universal barcode" for all kingdoms of life has been abandoned, DNA barcodes are nowadays available for a huge number of taxa from all over the tree of life. For plants, two chloroplast genes, the large subunit of the rubisco enzyme (*rbc*L) and the maturase K (*mat*K) have been chosen (CBOL Plant Working Group, 2009), whilst for fungi the nuclear internal transcribed spacer region (ITS) (Schoch et al., 2012) and the V4 region of the gene coding for the small ribosomal subunit (V4-18S) for protists (Pawlowski et al., 2012) are the markers of choice.

1.6. From barcodes to metabarcodes

The recent technical advancements of massive parallel DNA sequencing technologies (e.g. next-generation sequencing platforms, NGS; Shendure and Ji, 2008; Glenn, 2011) have revolutionised many areas of scientific inquiry, taxonomy included. Providing millions of sequence-reads in a single experiment, NGS platforms have extended the classical, one-specimen-at-a-time Sanger sequencing identification of single specimens to the community level (Taberlet et al., 2012). This approach, called "metabarcoding", is a multispecies

identification method using massive parallel sequencing of a particular marker in environmental DNA or RNA samples (Cristescu, 2014). The significant decrease in the costs of massive sequencing and the ease of sampling and analysing multiple instead of individual specimens has led to an increase of metabarcoding studies for aquatic, microbial and soil communities (Schmidt et al., 2013; Valentini et al., 2016; Abdelfattah et al., 2018), as well as to its application to biodiversity surveillance and monitoring (Bohmann et al., 2014; Deiner et al., 2017). However, being "blind", metabarcoding approaches need a comprehensive taxonomic reference database, which is generated with the traditional barcoding approach on morphologically verified and curated specimens (Cristescu, 2014). Furthermore, its blindness is also extended to the unknown amount of species to identify in the community; this requires the primers used for the PCR to be highly versatile (amplify different target molecules with the same efficiency), in order not to miss species whose target sequences do not match well with the primers designed (Taberlet et al., 2012). Despite these and many other issue shared with the classical DNA-barcoding approach (use of a single target gene to identify taxa, PCR errors, etc.), DNA-metabarcoding has a potential that goes beyond biodiversity assessment and monitoring. It has proven to be an effective tool for diet assessment (Leray et al., 2013; De Barba et al., 2014; Kartzinel et al., 2015), species diversity and distribution (Nanjappa et al., 2014; Malviya et al., 2016; dos Santos et al., 2017; Tragin and Vaulot, 2019) and product authentication (Mishra et al., 2016; Raclariu et al., 2017; 2018). All the aforementioned studies show that we are still at the early stages of exploitation of DNA-metabarcoding potential, and it will be a powerful technique for many years to come.

1.6.1. DNA barcoding and metabarcoding in diatoms

The application of DNA barcoding to diatoms is no different, in principle, from that in other organisms i.e. to provide unambiguous identification of a specimen, using a short sequence of coding or noncoding DNA (Mann et al., 2010). Some characteristics found in

diatoms as cryptic speciation, different morphology across life cycle and culture conditions (Mann, 1999) make barcoding particularly advantageous in these organisms over classical morphological examinations (Mann et al., 2010).

To date, no universal barcode region for diatoms exists, but several markers have been considered and proposed within the nuclear, mitochondrial and chloroplast genomes (Moniz and Kaczmarska, 2009, Fig. 1.4).

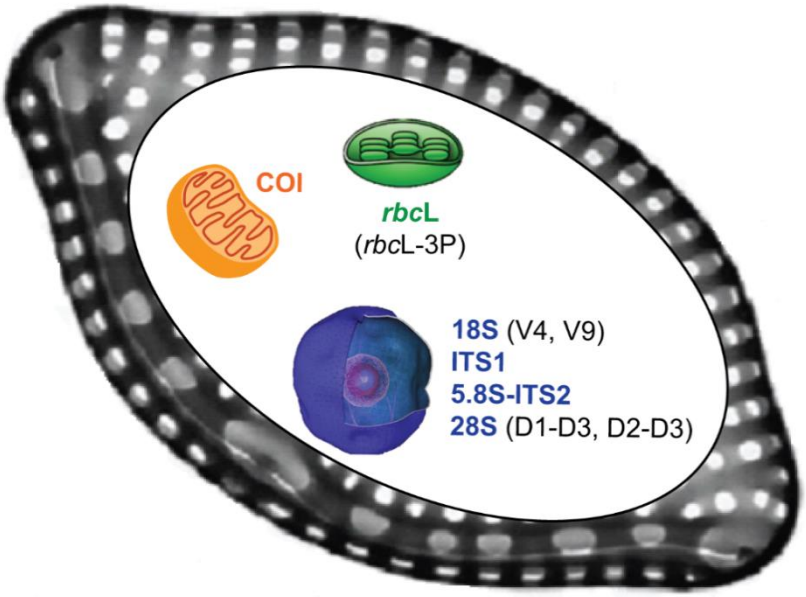


Fig. 1.4. Main target genes utilised for DNA barcoding in diatoms. Orange = mitochondrion; green = chloroplast; blue = nucleus.

The classical barcode genes used for animals (COI) and plants (*mat*K, *rbc*L) seem not to work well for diatoms and other protists. For COI, the main problem is lack of sufficiently conserved primer target regions across taxa (Evans et al., 2007; Moniz and Kaczmarska, 2009) and occurrence of introns (Ehara et al., 2000; Armbrust et al., 2004; Ravin et al., 2010). Plastid markers have been considered problematic for DNA barcoding due to both uniparental or biparental inheritance (Round et al., 1990; Jensen et al., 2003; Levialdi Ghiron et al., 2008). Nonetheless, the *rbc*L has been evaluated both in its entire length (~1400 bp) and as fragment at 3'-end (*rbc*L-3P, ~750 bp, Hamsher et al., 2011; ~540 bp, MacGillivary and Kaczmarska, 2011). Preliminary results suggested that the 3'-region is

more variable than the 5'-one and so discouraged the use of the whole gene (Hamsher et al., 2011). In spite of the fact that ease of amplification, sequencing, and alignment as well as lack of indels and introns make it a promising marker (MacGillivary and Kaczmarska, 2011), the low resolution at discerning closely related species in some groups and the aforementioned uncertain inheritance led to the conclusion of a better use of *rbc*L-3P region as complementary barcoding gene together with 5.8S-ITS2 rDNA region in a dual-locus DNA barcoding system (MacGillivary and Kaczmarska, 2011). This latter region was proposed by Moniz and Kaczmarska (2009, 2010) as candidate barcode based on its use at identifying protist, fungal and plant species (Wayne Litaker et al., 2007; Seifert, 2009; Chen et al., 2010). However, the ITS region is known to be difficult to align even in closely related species (Desdevises et al., 2000; Poisot et al., 2011) and to show infraspecific polymorphism due to non-concerted evolution (Harpke et al., 2006; Zheng et al., 2008), all factors that limit its applications in heterogeneous taxa.

Among nuclear DNA markers, and still within the rDNA cistron, most of the attention has been focused on the genes coding for the nuclear small and large subunit (SSU and LSU) RNAs of the ribosomes, (a.k.a. 18S and 28S rDNA, respectively). Due to its overall length, generally around 3,000 bp, barcoding has focused on the D1-D3 (~ 800 bp) and D2-D3 (~ 613 bp) regions in the LSU (Hamsher et al., 2011). These fragments are considered as variable as the *rbc*L-3P (Hamsher et al., 2011), and therefore, expected to resolve species- and sometimes population-level relationships (Alverson, 2008). However, these markers are unsuitable for current NGS platforms used in metabarcoding approaches because they are too long. Another drawback is that LSU reference sequences are available only for selected groups of organisms; not yet across the entire eukaryotic tree of life, not even across the diatom diversity. On the contrary, the SSU region has been used extensively in diatom phylogenies (Medlin et al., 1993; Kooistra and Medlin, 1996; Medlin et al., 1996; Medlin and Kaczmarska, 2004; Sarno et al., 2005; Sorhannus, 2007) and the huge number of reference sequences stored in public databases (e.g. PR^2, Guillou et al., 2012) essentially

covers the diversity of the diatoms. The validity of the various variable regions as barcoding target has been evaluated, in particular the V4 and V9 (Nelles et al., 1984). Recent results showed that the V4 region (~ 380-400 bp) can be considered the most promising candidate marker for DNA barcoding in diatoms given its ease of amplification, extensive reference library and variability, and universality of its primer target (Zimmermann et al., 2011; Luddington et al., 2012). It outperforms the V9 region in separating closely related species because of its greater length (~ 380 bp vs. 105 bp) and the fact that the V9 region is located at the very 3'-end of 18S gene, a region that is often sequenced incompletely or poorly (Gaonkar, 2017; Gaonkar et al., 2018). However, currently several V4 (BioMarKs, Massana et al., 2015; the Ocean Sampling Day, Kopf et al., 2015) and V9 (e.g. Tara Oceans, de Vargas et al., 2015) metabarcoding datasets are available to explore diversity and distribution of organisms (diatoms included) in world's oceans and to test the effectiveness of both regions in discriminating specific taxa. In this thesis, I will use the two global metabarcoding datasets, OSD (V4) and Tara Oceans (V9) to explore the diversity of *Chaetoceros* in the world's oceans.

1.7. Case study: the planktonic diatom family Chaetocerotaceae, with emphasis on the genus *Chaetoceros*

Diatoms (from the Greek word *diatomos*, "cut in half") are unicellular eukaryotes whose hallmark is the ornamented silica cell wall called frustule (Round et al. 1990). In the Tree of Life, diatoms are found in the superphylum Heterokonta (Stramenopiles, Adl et al., 2005), which includes unicellular eukaryotes that produce, at some point in their lifecycles, cells with two unequal flagella (Cavalier-Smith, 1986). Diatoms are one of the largest and ecologically most significant groups of organisms on Earth. They occur almost everywhere they can found adequately amount of light and water for photosynthesis: oceans, lakes, rivers, marshes, rock faces, and even on the feathers of some diving birds (Mann, 2010b).

Because of their abundance in marine plankton, diatoms are estimated to account for as much as 20% of global carbon fixation (Field et al., 1998).

From the ecological point of view, diatoms are generally divided in planktonic (suspended in open waters) and benthic (living on the floor of water basins). Planktonic diatoms dominate the phytoplankton of cold, nutrient-rich waters, such as upwelling areas of the oceans and recently circulated lake waters (Graham et al., 2016). Together with benthic ones, after death they are responsible for carbon sinking and accumulation of silica in sediments, contributing to the flux of nutrients (Smetacek, 1985; Willén, 1991).

The focus of my Ph.D. thesis is the planktonic diatom family Chaetocerotaceae Ralfs in Pritchard, with particular emphasis on the genus *Chaetoceros*. The genus *Chaetoceros* Ehrenberg, 1844 is common in the plankton worldwide and, together with the genus *Bacteriastrum* Shadbolt, 1854 constitutes the family Chaetocerotaceae Ralfs in Pritchard. The hallmark of the family is the presence of siliceous hollow spine-like extensions (setae), which protrude from the valve face or margin of the cell. Chaetocerotaceae belong to the bipolar centric diatoms, i.e., a clade or grade of diatoms with valves exhibiting a bi- or multipolar architecture, a circular pattern centre, a centrally located labiate process and apically located fields of poroids that are ultrastructurally distinct from the poroids in the remainder of the cell wall (frustule) elements. The setae are believed to have evolved from those apical pore fields. Among the differences between the two genera are the following: i) valvar symmetry, which is multipolar in *Bacteriastrum* (Fig. 1.5A) and bipolar in *Chaetoceros* (Fig. 1.5C); ii) seta number per valve, generally two in *Chaetoceros* (Fig. 1.5D) and more than two in *Bacteriastrum* (Fig. 1.5B); iii) valve outline, oval in the former (Fig. 1.5D) and circular in the latter (Fig. 1.5B); and iv) the number of species, hundreds in *Chaetoceros*, a few dozens in *Bacteriastrum*.

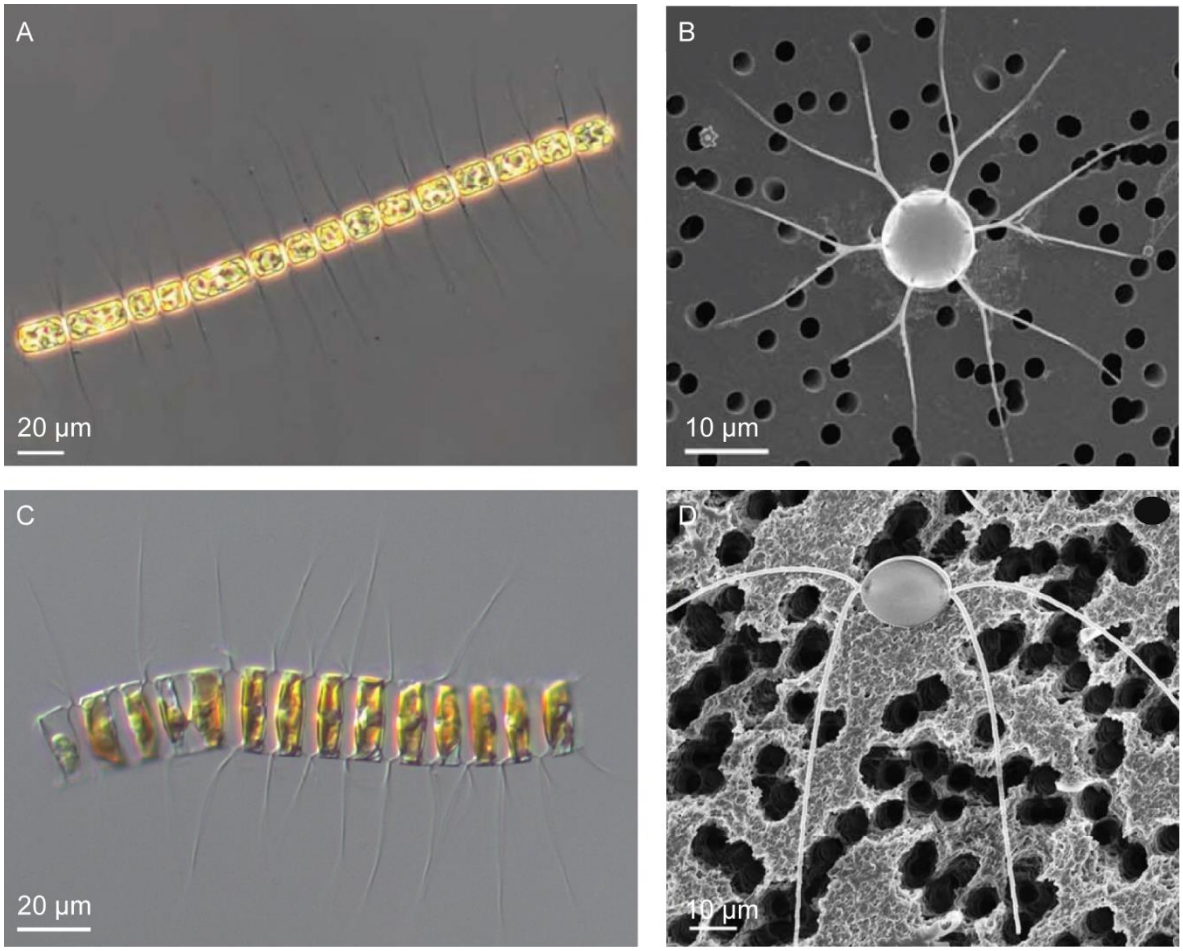


Fig. 1.5. Main morphological features of *Bacteriatrum* and *Chaetoceros*. (A) Girdle view of *B. furcatum* sp. 2 strain Na8A3 in LM; (B) Valval view of the same strain in SEM; (C) Girdle view of *C. debilis* sp. 3 strain Ch13A4 in LM; (D) Valval view of the same strain in SEM. Figures are from Gaonkar et al. (2018).

Gran (1897) divided the genus *Chaetoceros* in two subgenera, *Phaeoceros* Gran and *Hyalochaete*; the first includes species with multiple chloroplasts in the central body of the cell and in the setae, the second comprises species without plastids in the setae (Kooistra et al., 2010). Hendey (1964) changed the name of the subgenus *Phaeoceros* in *Chaetoceros* since the subgenus that includes the type species of the genus (*Chaetoceros dichaeta* Ehrenberg) has to keep the epithet of the genus, according to the rules of the botanical nomenclature. More recently, Hernández-Becerril (1993) created a third subgenus, *Bacteriastroidea* Hernández-Becerril to include a single species, *C. bacteriastroides*, exhibiting two different types of setae per valve.

Both genera are homothallic, i.e., micro and macrogametes (analogous to male and female gametes) are formed in one and the same clonal culture, but in different cells. Following gamete fusion, the resulting zygote develops through partial inflation into a specialised cell, the auxospore, which re-establishes the initial vegetative cell size. Furthermore, vegetative cells in many of the species can develop into resting spores anytime during the vegetative part of their life cycle (see Round et al. 1990). Resting spores are highly silicified, and often heavily armoured cells that go senescent and can survive under conditions adverse to growth. The spores sink to the sea floor and germinate whenever favourable conditions are restored. Simultaneous germination of massive numbers of spores can trigger sudden seasonal diatom blooms (McQuoid and Hobson, 1996). Alternatively, the spores can be sequestered in the sediment, where they provide a stratigraphic record (Suto, 2006). In the end, they can get fossilised, thus constituting an important carbon sink (Smetacek, 1985).

1.7.1. Fossil record of Chaetoceros

Vegetative cells of *Chaetoceros* leave no fossil record because these are weakly silicified and in most cases dissolve after the cell's death (Ishii et al., 2011). Instead, the heavily silicified resting spores are often preserved in near-shore sediments as fossils, frequently in association with other diatom fossils, providing useful information for reconstructing paleo-productivity and paleo-environmental changes (Akiba, 1986; Itakura, 2000). For these reasons, *Chaetoceros* fossils have been described as "spore genera" and they may represent extinct taxa (Ishii et al., 2011 and references therein). A large number of spore genera has been described, such as *Dicladia* Ehrenberg (1854), *Xanthiopyxis* Ehrenberg (1854), *Syndendrium* Ehrenberg (1854), *Liradiscus* Greville (1865) and *Monocladia* Suto (2003), all of which may be assignable to the genus *Chaetoceros* (Suto, 2005). These fossils are from the Paleogene (65-23 mya), in particular from the Eocene/Oligocene

boundary (~34 mya), the Oligocene/Miocene boundary (~23 mya) and the early/middle Miocene boundary (~15.9 mya, Suto et al., 2006).

However, age estimates of *Chaetoceros* from diatom phylogenies calibrated with molecular clocks and diatom fossils are far older than direct fossil evidence. A phylogeny of diatoms inferred using the small subunit of rDNA gene (18S) and calibrated with fossil records dated back the origin of *Chaetoceros* in the Cretaceous (around 120 mya, Sorhannus 2007). In another study, conducted using four molecular markers (SSU, LSU, *rbc*L and *psb*A) and performing molecular clock analysis the split between *Chaetoceros* and *Cymatosira* was found in the Jurassic, around 180 mya (Medlin, 2015).

1.7.2. The ecological and evolutionary importance of Chaetoceros

Chaetoceros possesses some characteristics that makes it the prime target for ecological and evolutionary studies in marine phytoplankton. Indeed, it; i) is one of the most species-rich genera among diatoms (Rines and Hargraves, 1988; Hasle and Syvertsen, 1996; Hernández-Becerril, 1996), with about 500 taxa attributable to species or "variants", and few more than 200 flagged as taxonomically accepted species (Guiry and Guiry, 2017); ii) is globally distributed, especially in upwelling regions (VanLandingham, 1968; Hasle and Syvertsen, 1996); iii) it accounts for 20–25% of the total marine primary production (Werner, 1977), especially in near-shore upwelling regions and coastal areas (Rines and Hargraves, 1988; Rines and Theriot, 2003). Furthermore, some species can be harmful during blooms, getting stuck in fish gills with their setae and causing mass mortality through limited oxygen uptake (Albright et al., 1993).

The success of this genus in terms of number of species, abundance, and global distribution is likely due to the combination of particular aspects of the life cycle (e.g. resting spore formation) and evolutionary novelty (the setae).

Many species of *Chaetoceros* form resting spores (Blasco, 1970; von Stosch et al., 1973; Hargraves and French, 1975), a strategy that allows them to escape situations in which

nutrient supplies are scarce, sinking to the sea floor and germinating when favourable conditions are restored. This characteristic is considered to be an evolutionary primitive trait (Simonsen, 1979) and typical of current neritic species (Ross and Sims, 1974).

The putative adaptive advantages in possessing setae have not been cleared; they might deter grazers, have a role in buoyancy or nutrient and CO_2 uptake (Smayda and Boleyn, 1966; Smetacek, 1985; Verity and Smetacek, 1996).

Chaetoceros is easy to identify at the generic level because of the setae, but it is difficult to identify at the species level since the morphological criteria used (e.g. colony formation, cell size and shape, intercellular spaces, number of chloroplasts, morphology and orientation of setae, etc.) are quite variable (Hargraves, 1979) and in many smaller species difficult to observe in LM. Factors such as the presence/absence of grazers, salinity changes, nutrient availability or prolonged culture conditions can alter the morphology of the species, thus creating uncertainties in the species identification.

In spite of that, integration of phylogenetic and morphological information on isolated strains has contributed to the characterisation of *Chaetoceros* species and to the discovery of cryptic and pseudo-cryptic species (Kooistra et al., 2010; Degerlund et al., 2012; Huseby et al., 2012; Chamnansinp et al., 2013; Li et al., 2017; Balzano et al., 2016; Gaonkar et al., 2017; 2018). More than 80 *Chaetoceros* strains and a dozen of *Bacteriastrum* have been characterised so far by morphological (light, scanning and transmission electron microscopy) and genetic (D1-D4 region of 28S rDNA) means (Gaonkar et al., 2018), thus providing a reference library of strains occurring in the Gulf of Naples and/or in other localities.

The PhD thesis of Gaonkar (2017) focused on: i) the molecular phylogeny of Chaetocerotaceae using 18S and partial 28S rDNA; ii) the diversity of Chaetocerotaceae in the Gulf of Naples (GoN) using a V4-18S metabarcoding approach; and iii) the analysis of the *C. socialis* species complex, with the description of two new species. The goals of i)

were: to understand how thoroughly the species diversity of the genera has been explored and what the relationships are between these genera; how many species are to be discovered yet; how common is cryptic diversity; and if morphological species delimitation has a genetic support. The V4-18S metabarcoding approach (point ii)) aimed at assessing: how many species occur in the GoN; how many species are found in the High-Throughput Sequencing (HTS) data but are still to be morphologically identified; how many of them occurring in the HTS data are known from elsewhere but have never been recorded in the cell counts at the GoN. Among the main results relevant to my thesis are the following: i) the 18S and 28S phylogenies do not resolve the position of *Bacteriastrum* with respect to *Chaetoceros*, and only terminal clades obtain significant support; ii) potential cryptic species exist within several morphologically defined species (e.g. *C. affinis*, *C. curvisetus*, *C. lorenzianus*, *C. socialis*); iii) the 18S-V4 region is generally better than V9 at discriminating terminal clades, with the same resolution of whole 18S gene, revealing to be a candidate target for metabarcoding studies.

1.7.3. Aim of Ph.D. thesis

Starting from the points discussed above, my PhD thesis has the following aims:

1. To produce a multi-gene phylogeny of the family Chaetocerotaceae integrating the pre-existing information of nuclear data with chloroplast and mitochondrial ones in order to assess if adding phylogenetic information helps towards resolving the phylogenetic history of the family (Chapter II);
2. To provides an assessment of the diversity and distribution of the genus *Chaetoceros* by integrating classical and novel primary biodiversity data (global metabarcoding dataset) (Chapter III);
3. To analyse the *C. curvisetus* species complex using the potential of spatial data contained in global metabarcoding datasets in the form of phylogenetic networks (Chapter IV);

4. To test the hypothesis of concerted evolution in *Chaetoceros* with an appropriate experimental design (single strain HTS and targeted analyses), starting from the data contained in a temporal metabarcoding dataset (MareChiara) (Chapter V).

References

Kociolek, J. P., Balasubramanian, K., Blanco, S., Coste, M., Ector, L., Liu, Y., … Witkowski, J. (2019). DiatomBase. Accessed at http://www.diatombase.org on 2019-06-02.

Abdelfattah, A., Malacrinò, A., Wisniewski, M., Cacciola, S. O., Schena, L. (2018). Metabarcoding: A powerful tool to investigate microbial communities and shape future plant protection strategies. *Biological Control, 120*, 1-10.

Adl, S. M., Simpson, A. G., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., ... Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology, 52*(5), 399-451.

Agardh, C. A. (1830–1832). *Conspectus criticus Diatomacearum, parts 1–4*. Berling, Lund.

Agassiz, L. (1859). Essay on classification, in *Contributions to the Natural History of United States*, vol. 1. Boston: Little, Brown & Co. Harvard University Press, Cambridge.

Akiba, F. (1986). Middle Miocene to Quaternary diatom biostratigraphy in the Nankai Trough and Japan Trench, and modified Lower Miocene through Quaternary diatom zones for middle-to-high latitudes of the North Pacific. In: *Initial Reports of the Deep Sea Drilfing Project* (H. Kagami, D.E. Karig et al., eds), *87*, 393-481. US. Government Printing Office, Washington, D.C., U.S.A.

Albright, L. J., Yang, C. Z., Johnson, S. (1993). Sub-lethal concentrations of the harmful diatoms, *Chaetoceros concavicornis* and *C. convolutus*, increase mortality rates of penned Pacific salmon. *Aquaculture, 117*(3-4), 215-225.

Alverson, A. J. (2008). Molecular systematics and the diatom species. *Protist, 159*(3), 339-353.

Amato, A., Kooistra, W. H. C. F., Ghiron, J. H. L., Mann, D. G., Pröschold, T., Montresor, M. (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist, 158*(2), 193-207.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., ... Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science, 306*(5693), 79-86.

Arnot, D. E., Roper, C., Bayoumi, R. A. (1993). Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Molecular and Biochemical Parasitology, 61*(1), 15-24.

Avise, J. C. (2012). *Molecular markers, natural history and evolution*. Springer Science & Business Media, Dordrecht.

Avise, J. C., Walker, D. (1999). Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. *Proceedings of the National Academy of Sciences, 96*(3), 992-995.

Bachmann, K. (1998). Species as units of diversity: an outdated concept. *Theory in Biosciences, 117*, 213-230.

Baer, C. F., Miyamoto, M. M., Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics, 8*(8), 619.

Balzano, S., Percopo, I., Siano, R., Gourvil, P., Chanoine, M., Marie, D., ... Sarno, D. (2016). Morphological and genetic diversity of Beaufort Sea diatoms with high

contributions from the *Chaetoceros neogracilis* species complex. *Journal of Phycology, 53*(1), 161-187.

Bateson, W. (1894). *Materials for the study of variation: treated with especial regard to discontinuity in the origin of species*. Macmillan, London.

Blasco, D. (1970). Estudio de la morfología de *Chaetoceros didymus* al microscopio electrónico. *Investigaciones Pesqueras, 34*(2), 149-155.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., ... De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution, 29*(6), 358-367.

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., ... Rayko, E. (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature, 456*(7219), 239-244.

Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1577), 2503-2513.

Canesi, K. L., Rynearson, T. A. (2016). Temporal variation of *Skeletonema* community composition from a long-term time series in Narragansett Bay identified using high-throughput DNA sequencing. *Marine Ecology Progress Series, 556*, 1-16.

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A. E., Kotaki, Y., Rhodes, L., ... Vyverman, W. (2010). Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences, 107*(29), 12952-12957.

Cavalier-Smith, T. (1986). The kingdom Chromista: origin and systematics. *Progress in Phycological Research, 4*, 309-347.

CBOL Plant Working Group., Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., ... Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, 106*(31), 12794-12797.

Cesalpino, A. (1583). *De plantis libri XVI*. Marescot, Florence.

Chamnansinp, A., Li, Y., Lundholm, N., Moestrup, Ø. (2013). Global diversity of two widespread, colony-forming diatoms of the marine plankton, *Chaetoceros socialis* (syn. *C. radians*) and *Chaetoceros gelidus* sp. nov. *Journal of Phycology, 49*(6), 1128-1141.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., ... Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS ONE, 5*(1), e8613.

Coyne, J. A., & Allen Orr, H. (1998). The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1366), 287-305.

Cracraft, J. (1983). Species concepts and speciation analysis. In *Current ornithology* (pp. 159-187). Springer, Boston, Massachusetts.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution, 29*(10), 566-571.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life*. John Murray, London.

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society, 85*(3), 407-417.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources, 14*(2), 306-323.

De Decker, S., Vanormelingen, P., Pinseel, E., Sefbom, J., Audoor, S., Sabbe, K., Vyverman, W. (2018). Incomplete reproductive isolation between genetically distinct sympatric clades of the pennate model diatom *Seminavis robusta*. *Protist, 169*(4), 569-583.

de Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology, 56*(6), 879-886.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science, 348*(6237), 1261605.

Degerlund, M., Huseby, S., Zingone, A., Sarno, D., Landfald, B. (2012). Functional diversity in cryptic species of *Chaetoceros socialis* Lauder (Bacillariophyceae). *Journal of Plankton Research, 34*(5), 416-431.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Desdevises, Y., Jovelin, R., Jousson, O., Morand, S. (2000). Comparison of ribosomal DNA sequences of *Lamellodiscus* spp. (Monogenea, Diplectanidae) parasitising *Pagellus* (Sparidae, Teleostei) in the North Mediterranean Sea: species divergence and coevolutionary interactions. *International Journal for Parasitology, 30*(6), 741-746.

dos Santos, A. L., Gourvil, P., Tragin, M., Noël, M. H., Decelle, J., Romac, S., Vaulot, D. (2017). Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *The ISME journal, 11*(2), 512.

Drake, J. W. (1999). The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences, 870*(1), 100-107.

Drebes, G. (1977). Sexuality. In: Werner, D. (Ed.). *The Biology of Diatom*. Blackwell Scientific, Oxford.

Edlund, M. B., Stoermer, E. F. (1997). Ecological, evolutionary, and systematic significance of diatom life histories. *Journal of Phycology, 33*(6), 897-918.

Ehara, M., Watanabe, K. I., Ohama, T. (2000). Distribution of cognates of group II introns detected in mitochondrial cox1 genes of a diatom and a haptophyte. *Gene, 256*(1-2), 157-167.

Ehrenberg, C. G. (1838). *Die Infusionsthierchen als vollkommene Organismen. Ein Blick in das tiefere organische Leben der Natur*. Leopold Voss, Leipzig.

Eldredge, N., Cracraft, J. (1980). *Phylogenetic patterns and the evolutionary process*. Columbia University Press, New York.

Evans, K. M., Wortley, A. H., Mann, D. G. (2007). An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist, 158*(3), 349-364.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science, 281*(5374), 237-240.

Gallagher, J. C. (1983). Cell enlargement in *Skeletonema costatum* (Bacillariophyceae). *Journal of Phycology, 19*(4), 539-542.

Gallagher, J. C., Wood, A. M., Alberte, R. S. (1984). Ecotypic differentiation in the marine diatom *Skeletonema costatum*: influence of light intensity on the photosynthetic apparatus. *Marine Biology, 82*(2), 121-134.

Gaonkar, C. C. (2017). *Diversity, Distribution and Evolution of the Planktonic Diatom Family Chaetocerotaceae*. Doctoral dissertation, The Open University.

Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Montresor, M., Sarno, D. (2017). Two new species in the *Chaetoceros socialis* complex (Bacillariophyta): *C. sporotruncatus* and *C. dichatoensis*, and characterization of its relatives, *C. radicans* and *C. cinctus*. *Journal of Phycology*, *53*(4), 889-907.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. *PloS ONE, 13*(12), e0208929.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759-769.

Godhe, A., Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1728), 20160399.

Gogarten, J. P., Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, *3*(9), 679-687.

Goldschmidt, R. (1940). *The Material Basis of Evolution*. Yale University Press, New Haven, Connecticut.

Graham, L. E., Graham, J. M., Wilcox, L. W., Cook, M. E. (2016). *Algae*. Pearson Prentice Hall, Upper Saddle River, New Jersey.

Gran, H. H. (1897). *Botanik. Prophyta: Diatomaceae, Silicoflagellata og Cilioflagellata. Den Norske Nordhavs Expedition 1876-1878*, 7, 1-36.

Grunow, A. (1879). New species and varieties of Diatomaceae from the Caspian Sea. Translated with additional notes by F Kitton. *Journal of the Royal Microscopical Society*, *2*, 677-691.

Gsell, A. S., de Senerpont Domis, L. N., Przytulska-Bartosiewicz, A., Mooij, W. M., Van Donk, E., Ibelings, B. W. (2012). Genotype-by-temperature interactions may help to maintain clonal diversity in *Asterionella formosa* (Bacillariophyceae). *Journal of Phycology*, *48*(5), 1197-1208.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... Christen, R. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597-D604.

Guiry, M. D. (2012). How many species of algae are there? *Journal of Phycology, 48*(5), 1057-1063.

Guiry, M. D., Guiry, G. M. (2017). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 09 June 2017.

Gulick, J. T. (1888). Divergent evolution through cumulative segregation. *Journal of the Linnean Society of London, Zoology, 20*(120), 189–274.

Hamsher, S. E., Evans, K. M., Mann, D. G., Poulíčková, A., Saunders, G. W. (2011). Barcoding diatoms: exploring alternatives to COI-5P. *Protist, 162*(3), 405-422.

Hargraves, P. E. (1979). Studies on marine plankton diatoms IV. Morphology of *Chaetoceros* resting spores. *Nova Hedwigia Beihefte, 64*, 99-120.

Hargraves, P. E., French, F. (1975). Observations on the survival of diatom resting spores. *Nova Hedwigia Beihefte*, *53*, 229-238.

Harpke, D., Peterson, A. (2006). Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Molecular Phylogenetics and Evolution, 41*(3), 579-593.

Hasle, G. R., Syvertsen, E. E. (1996). Marine diatoms. In: Tomas, C.R. (Ed.), *Identifying Marine Diatoms and Dinoflagellates*, pp. 5–385. Academic Press, San Diego.

Hebert, P. D., Cywinska, A., Ball, S. L., DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1512), 313-321.

Hendey, N. I. (1964). *An introductory account of the smaller algae of British coastal waters. Part. V. Bacillariophyceae. Fisheries Investigation, Series IV*. Otto Koeltz Science Publishers, Koenigstein.

Hernández-Becerril D. U. (1996). A morphological study of *Chaetoceros* species (Bacillariophyta) from the plankton of the Pacific Ocean of Mexico. *Bulletin of the Natural History Museum, Botany Series, 26*(1), 1–73.

Hernández-Becerril, D. U. (1993). Note on the morphology of two planktonic diatoms: *Chaetoceros bacteriastroides* and *C. seychellarus*, with comments on their

taxonomy and distribution. *Botanical journal of the Linnean Society, 111*(2), 117-128.

Hull, D. L. (1976). Are species really individuals? *Systematic Zoology*, *25*(2), 174-191.

Hull, D. L. (1978). A matter of individuality. *Philosophy of Science, 45*(3), 335-360.

Huseby, S., Degerlund, M., Zingone, A., Hansen, E. (2012). Metabolic fingerprinting reveals differences between northern and southern strains of the cryptic diatom *Chaetoceros socialis*. *European Journal of Phycology*, *47*(4), 480-489.

Itakura, S. (2000). Physiological ecology of the resting stage cells of coastal planktonic diatoms. *Bulletin of Fisheries and Environment of Inland Sea, 2*, 67–130.

Jensen, K. G., Moestrup, Ø., Schmid, A. M. M. (2003). Ultrastructure of the male gametes from two centric diatoms, *Chaetoceros laciniosus* and *Coscinodiscus wailesii* (Bacillariophyceae). *Phycologia, 42*(1), 98-105.

Kaczmarska, I., Ehrman, J. M., Moniz, M. B. J., Davidovich, N. (2009). Phenotypic and genetic structure of interbreeding populations of the diatom *Tabularia fasciculata* (Bacillariophyta). *Phycologia*, *48*(5), 391–403.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., ... Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences, 112*(26), 8019-8024.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences, 86*(16), 6196-6200.

Kooistra, W. H. C. F., Medlin, L. K. (1996). Evolution of the diatoms (Bacillariophyta): IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Molecular Phylogenetics and Evolution, 6*(3), 391-407.

Kooistra, W. H. C. F., Sarno, D., Hernández-Becerril, D. U., Assmy, P., Di Prisco, C., Montresor, M., (2010). Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia, 49,* 471-500.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., ... Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience, 4*(1), 27.

Koufopanou, V., Hughes, J., Bell, G., Burt, A. (2006). The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1475), 1941-1946.

Kremp, A., Godhe, A., Egardt, J., Dupont, S., Suikkanen, S., Casabianca, S., Penna, A. (2012). Intraspecific variability in the response of bloom-forming marine microalgae to changed climate conditions. *Ecology and Evolution, 2*(6), 1195-1207.

Kützing, F. T. (1833). Synopsis Diatomacearum oder Versuch einer systematischen Zusammenstellung der Diatomeen. *Linnaea, 8*: 529–620.

Lemey, P., Salemi, M., Vandamme, A. M. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, New York.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., ... Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology, 10*(1), 34.

Levialdi Ghiron, J. H., Amato, A., Montresor, M., Kooistra, W. H. C. F. (2008). Plastid inheritance in the planktonic raphid pennate diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae). *Protist, 159*(1), 91-98.

Lewis, W. M. (1984). The diatom sex clock and its evolutionary significance. *American Naturalist, 123*(1), 73–80.

Li, Y., Boonprakob, A., Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Hernández-Becerril, D., ... Lundholm, N. (2017). Diversity in the Globally Distributed Diatom Genus *Chaetoceros* (Bacillariophyceae): Three New Species from Warm-Temperate Waters. *PloS ONE, 12*(1), e0168887.

Luddington, I. A., Kaczmarska, I., Lovejoy, C. (2012). Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae). *PLoS ONE, 7*(9), e45664.

Lund, J. W. G. (1954). The seasonal cycle of the plankton diatom, *Melosira italica* (Ehr.) Kutz. subsp. *subarctica* O. Mull. *The Journal of Ecology*, 151-179.

MacDonald, J. D. (1869). On the structure of the diatomaceous frustule, and its genetic cycle. *Annals and Magazine of Natural History, 13*(4), 1-8.

MacGillivary, M. L., Kaczmarska, I. (2011). Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. *Journal of Eukaryotic Microbiology, 58*(6), 529-536.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., ... Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences, 113*(11), E1516-E1525.

Mann, D. G. (1999). The species concept in diatoms. *Phycologia, 38*(6), 437-495.

Mann, D. G. (2010a). Discovering diatom species: is a long history of disagreements about species-level taxonomy now at an end? *Plant Ecology and Evolution, 143*(3), 251-264.

Mann, D. G., Droop, S. J. M. (1996). Biodiversity, biogeography and conservation of diatoms. In: *Biogeography of freshwater algae*. Springer, Dordrecht.

Mann, D. G., Sato, S., Trobajo, R., Vanormelingen, P., Souffreau, C. (2010). DNA barcoding for species identification and discovery in diatoms. *Cryptogamie, Algologie, 31*(4), 557-577.

Mann, David G. (2010b). Diatoms. Version 07 February 2010 (under construction). http://tolweb.org/Diatoms/21810/2010.02.07 *in* The Tree of Life Web Project, http://tolweb.org/.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., ... de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology, 17*(10), 4035-4049.

May, R. M. (1992). How many species inhabit the earth? *Scientific American, 267*(4), 42-49.

May, R. M. (2010). Tropical arthropod species, more or less? *Science, 329*(5987), 41-42.

Mayden, R. L. (1997). A hierarchy of species concepts: The denouement in the saga of the species problem. Pages 381–424. In: *Species: The units of biodiversity* (M. F. Claridge, H. A. Dawah, and M. R. Wilson, eds.). Chapman and Hall, London.

Maynard Smith, J., Szathmary, E. (1995). *The Major Transitions in Evolution*. W.H. Freeman, Oxford.

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E. (1963). *Animal species and Evolution*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Mayr, E. (1969*). Principles of systematic zoology*. McGraw-Hill, New York.

Mayr, E. (1970). *Populations, species, and evolution: an abridgment of animal species and evolution (Vol. 19)*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Mayr, E. (1996). What is a species, and what is not? *Philosophy of Science, 63*(2), 262-277.

Mayr, E. (1997). *Evolution and the diversity of life: Selected essays*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Mayr, E. (1999). *Systematics and the origin of species, from the viewpoint of a zoologist*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

McQuoid, M. R., Hobson, L. A. (1996). Diatom resting stages. *Journal of Phycology, 32*(6), 889-902.

Medlin, L. K. (2015). A timescale for diatom evolution based on four molecular markers: reassessment of ghost lineages and major steps defining diatom evolution. *Vie et Milieu-Life and Environment, 65*(4), 219-238.

Medlin, L. K., Kaczmarska, I. (2004). Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia, 43*(3), 245-270.

Medlin, L. K., Kooistra, W. H., Gersonde, R., Wellbrock, U. (1996). Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. *Molecular Biology and Evolution, 13*(1), 67-75.

Medlin, L. K., Williams, D. M., Sims, P. A. (1993). The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. *European Journal of Phycology, 28*(4), 261-275.

Meyer, C. P., Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology, 3*(12), e422.

Mill, J. S. (1843). *A system of Logic Ratiocinative and Inductive—Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigations*. John W. Parker, London.

Mishler, B. D. (1999). Getting rid of species. In R. Wilson (ed.), *Species: New Interdisciplinary Essays*, pp.307-315. MIT Press, Cambridge, Massachusetts.

Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., Tiwari, R., Sundaresan, V. (2016). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology Journal, 14*(1), 8-21.

Mock, T., Medlin, L. K. (2012). Genomics and genetics of diatoms. In: *Advances in Botanical Research*. Academic Press.

Moniz, M. B., Kaczmarska, I. (2009). Barcoding diatoms: is there a good marker? *Molecular Ecology Resources, 9*, 65-74.

Moniz, M. B., Kaczmarska, I. (2010). Barcoding of diatoms: nuclear encoded ITS revisited. *Protist, 161*(1), 7-34.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., Worm, B. (2011). How many species are there on Earth and in the ocean? *PLoS Biology, 9*(8), e1001127.

Mullis, K. B., Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. In *Methods in Enzymology, 155* (pp. 189-204). Academic Press.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H. C. F., Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. *PLoS ONE, 9*(8), e103810.

Nelles, L., Fang, B. L., Volckaert, G., Vandenberghe, A., Wachter, R. D. (1984). Nucleotide sequence of a crustacean 18S ribosomal RNA gene and secondary structure of eukaryotic small subunit ribosoraal RNAs. *Nucleic Acids Research, 12*(23), 8749-8768.

Not, F., Siano, R., Kooistra, W. H., Simon, N., Vaulot, D., Probert, I. (2012). Diversity and ecology of eukaryotic marine phytoplankton. In: *Advances in Botanical Research*. Academic Press.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... de Vargas, C. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology, 10*(11), e1001419.

Pfitzer, E. (1869). Über den Bau und die Zellteilung der Diatomeen. *Botanische Zeitung*, *27*, 774-776.

Poisot, T., Verneau, O., Desdevises, Y. (2011). Morphological and molecular evolution are not linked in *Lamellodiscus* (Plathyhelminthes, Monogenea). *PLoS ONE, 6*(10), e26252.

Powers, T. (2004). Nematode molecular diagnostics: from bands to barcodes. *Annual Review of Phytopathology, 42*, 367-383.

Quijano-Scheggia, S. I., Garcés, E., Lundholm, N., Moestrup, Ø., Andree, K., Camp, J. (2009). Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic *Pseudo-nitzschia delicatissima* complex (Bacillariophyta), including the description of *P. arenysensis* sp. nov. *Phycologia, 48*(6), 492–509.

Raclariu, A. C., Heinrich, M., Ichim, M. C., de Boer, H. (2018). Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. *Phytochemical Analysis, 29*(2), 123-128.

Raclariu, A. C., Paltinean, R., Vlase, L., Labarre, A., Manzanilla, V., Ichim, M. C., ... de Boer, H. (2017). Comparative authentication of *Hypericum perforatum* herbal products using DNA metabarcoding, TLC and HPLC-MS. *Scientific Reports, 7*(1), 1291.

Ravin, N. V., Galachyants, Y. P., Mardanov, A. V., Beletsky, A. V., Petrova, D. P., Sherbakova, T. A., ... Grachev, M. A. (2010). Complete sequence of the

mitochondrial genome of a diatom alga *Synedracus* and comparative analysis of diatom mitochondrial genomes. *Current Genetics, 56*(3), 215-223.

Richardson, J. L. (1995). Dominance of asexuality in diatom life cycles: evolutionary, ecological and taxonomic implications. In: *Proceedings of the 13th International Diatom Symposium* (Ed. by D. Marino & M. Montresor). Biopress, Bristol.

Rieseberg, L. H., Willis, J. H. (2007). Plant speciation. *Science, 317*(5840), 910-914.

Rines, J. E., Theriot, E. C. (2003). Systematics of Chaetocerotaceae (Bacillariophyceae). I. A phylogenetic analysis of the family. *Phycological Research, 5*1(2), 83-98.

Rines, J.E.B., Hargraves, P.E., (1988). *The* Chaetoceros *Ehrenberg (Bacillariophyceae) flora of Narragansett Bay, Rhode Island, U.S.A.* Cramer J., Berlin.

Ross, R., Sims, P. A. (1974). Observations on family and generic limits in the centrales. *Nova Hedwigia Beihefte, 45,* 97-121.

Round, F. E. (1972). The problem of reduction of cell size during diatom cell division. *Nova Hedwigia, 23,* 291-303.

Round, F. E., Crawford, R. M. Mann, D. G. (1990). The Diatoms: Biology and Morphology of the Genera. *Cambridge University Press, Cambridge*.

Rubinoff, D., Cameron, S., Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity, 97*(6), 581-594.

Ruggiero, M. V., D'Alelio, D., Ferrante, M. I., Santoro, M., Vitale, L., Procaccini, G., Montresor, M. (2018). Clonal expansion behind a marine diatom bloom. *The ISME Journal, 12*(2), 463-472.

Rynearson, T. A., Armbrust, E. V. (2000). DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom *Ditylum brightwellii*. *Limnology and Oceanography, 45*(6), 1329-1340.

Rynearson, T. A., Armbrust, E. V. (2004). Genetic differentiation among populations of the planktonic marine diatom *Ditylum brightwellii* (Bacillariophyceae). *Journal of Phycology, 40*(1), 34-43.

Rynearson, T. A., Armbrust, E. V. (2005). Maintenance of clonal diversity during a spring bloom of the centric diatom *Ditylum brightwellii*. *Molecular Ecology, 14*(6), 1631-1640.

Sarno, D., Kooistra, W. H. C. F., Medlin, L. K., Percopo, I., Zingone, A. (2005). Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. *Journal of Phycology, 41*(1), 151-176.

Schmidt, P. A., Bálint, M., Greshake, B., Bandow, C., Römbke, J., Schmitt, I. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry, 65*, 128-132.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... Fungal Barcoding Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences, 109*(16), 6241-6246.

Seifert, K. A. (2009). Progress towards DNA barcoding of fungi. *Molecular Ecology Resources, 9*, 83-89.

Shendure, J., Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*(10), 1135-1145.

Simonsen, R. (1979). The diatom system: ideas on phylogeny. *Bacillaria, 2*, 9-71.

Simpson, G. G. (1961). *Principles of animal taxonomy*. Columbia University Press, New York.

Small, J. (1950). Quantitative evolution. XVI. Increase of species number in diatoms. *Annals of Botany, 14*, 91 – 113.

Smayda, T. J., Boleyn, B. J. (1966). Experimental observations on the flotation of marine diatoms. III. *Bacteriastrum hyalinum* and *Chaetoceros lauderi*. *Limnology and Oceanography, 11*(1), 35-43.

Smetacek, V. S. (1985). Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance. *Marine Biology, 84*(3), 239-251.

Sorhannus, U. (2004). Diatom phylogenetics inferred based on direct optimization of nuclear-encoded SSU rRNA sequences. *Cladistics, 20*(5), 487-497.

Sorhannus, U. (2007). A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology, 65*(1-2), 1-12.

Spratt, B. G., Staley, J. T., Fisher, M. C. (2006). Introduction: species and speciation in microorganisms. *Philophical Transactions of the Royal Society B: Biological Sciences, 361*, http://doi.org/10.1098/rstb.2006.1929.

Stork, N. E. (1993). How many species are there? *Biodiversity and Conservation, 2*(3), 215-232.

Suto, I. (2005). Taxonomy and biostratigraphy of the fossil marine diatom resting spore genera *Dicladia* Ehrenberg, *Monocladia* Suto and *Syndendrium* Ehrenberg in the North Pacific and Norwegian Sea. *Diatom Research, 20*(2), 351-374.

Suto, I. (2006). The explosive diversification of the diatom genus *Chaetoceros* across the Eocene/Oligocene and Oligocene/Miocene boundaries in the Norwegian Sea. *Marine Micropaleontology, 58*(4), 259-269.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology, 21*(8), 2045-2050.

Taberlet, P., Gielly, L., Pautou, G., Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology, 17*(5), 1105-1109.

Theriot, E. (1992). Clusters, species concepts, and morphological evolution of diatoms. *Systematic Biology*, *41*(2), 141-157.

Theriot, E. C., Ashworth, M., Ruck, E., Nakov, T., Jansen, R. K. (2010). A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution, 143*(3), 278-296.

Tragin, M., Vaulot, D. (2019). Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Scientific Reports, 9*(1), 5190.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929-942.

Van Heurck, H. F. (1896). *A Treatise on the Diatomaceae, containing introductory remarks on the structure, life history, collection, cultivation and preparation of diatoms, and a description and figure typical of every known genus, as well as a description and figure of every species found in the North Sea and countries bordering it, Including Great Britain, Belgium, & c.* Wheldon & Wesley, London.

VanLandingham, S.M., (1968). *Catalogue of the Fossil and Recent Genera and Species of Diatoms and their Synonyms: Part II.* Bacteriastrum *through* Coscinodiscus. Verlag von J. Cramer, Lehre, Germany.

Verity, P. G., Smetacek, V. (1996). Organism life cycles, predation, and the structure of marine pelagic ecosystems. *Marine Ecology Progress Series, 130*, 277-293. *Nova Hedwigia Beihefte, 53*, 1-35.

von Stosch, H. A. (1965). Manipulierung der Zellgrösse von Diatomeen im experiment. *Phycologia*, *5*(1), 21-44.

Wayne Litaker, R., Vandersea, M. W., Kibler, S. R., Reece, K. S., Stokes, N. A., Lutzoni, F. M., ... Tester, P. A. (2007). Recognizing dinoflagellate species using ITS rDNA sequences. *Journal of Phycology, 43*(2), 344-355.

Werner, D., (1977). Introduction with a note on taxonomy. In: Werner, D. (Ed.*), The Biology of Diatoms*. University of California Press, Berkeley, California.

Wheelis, M. L., Kandler, O., Woese, C. R. (1992). On the nature of global classification. *Proceedings of the National Academy of Sciences, 89*(7), 2930-2934.

Whitaker, R. J. (2006). Allopatric origins of microbial species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1475), 1975-1984.

Wilkins, J. (2011). Philosophically speaking, how many species concepts are there? *Zootaxa, 2765*, 58-60.

Will, K. W., Mishler, B. D., Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology, 54*(5), 844-851.

Willén, E. (1991). Planktonic diatoms-an ecological review. *Algological studies, 62*, 69-106.

Wilson, E. O. (1988). *Biodiversity*. National Academy Press.

Wilson, E. O. (1992). *The Diversity of Life*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Wood, A. M., Leatham, T. (1992). The species concept in phytoplankton ecology. *Journal of Phycology, 28*(6), 723-729.

Zheng, X., Cai, D., Yao, L., Teng, Y. (2008). Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Molecular Phylogenetics and Evolution, 48*(3), 892-903.

Zimmermann, J., Jahn, R., Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity and Evolution, 11*(3), 173.

Chapter II

Inferring the evolutionary history of Chaetocerotaceae

The material presented in this chapter has been published as Research Article:

"De Luca D., Sarno D., Piredda R., Kooistra W.H.C.F. (2019). A multigene phylogeny to infer the evolutionary history of Chaetocerotaceae (Bacillariophyta). *Molecular Phylogenetics and Evolution* 140, https://doi.org/10.1016/j.ympev.2019.106575".

As the author of this Elsevier article, I retain the right to include it in a thesis or dissertation (figures and tables included) according to the rules on copyright of Elsevier available at https://www.elsevier.com/about/our-business/policies/copyright#Author-rights.

2.1. Introduction

2.1.1. Systematics of Chaetocerotaceae

The planktonic diatom family Chaetocerotaceae Ralfs in Pritchard (1861) is one of the largest and most diverse marine diatom families (Cupp, 1943; Hernández-Becerril, 1996; Jensen and Moestrup, 1998; Rines and Hargraves, 1998; Shevchenko et al., 2006; Bosak and Sarno, 2017). It plays an ecologically important role, representing an important primary producer in coastal and offshore marine environments worldwide (Continuous Plankton Recorder Survey Team, 2004; Leblanc et al., 2012; Malviya et al., 2016). The family includes the extant genera *Bacteriastrum* Shadbolt and *Chaetoceros* Ehrenberg, which differ in the number of setae per valve. The former generally possesses many, regularly arranged along the valve margin whilst the latter exhibits usually just two, one at each end of the apical axis (Hasle and Syvertsen, 1996). Despite the ecological importance, little is known about the systematics of Chaetocerotaceae. *Bacteriastrum* is exclusively marine, with 11 taxonomically accepted species (Guiry and Guiry, 2018). Its cells are cylindrical in valve view and contain numerous plastids; intercalary setae usually fuse over a large part of their length and then bifurcate (i.e., appear to branch) whereas the terminal setae do not branch (Hasle and Syvertsen, 1996). Pavillard (1925) erected two sections, *Isomorpha* and *Sagittata*, based on the orientation of the terminal setae on the opposite terminal valves of a colony: in *Isomorpha* they are each other's mirror image whereas in *Sagittata* their orientation differs (Fig. 2.1). Within *Bacteriastrum*, *B. hyalinum* Lauder is the only species known to form resting spores (Drebes, 1972).

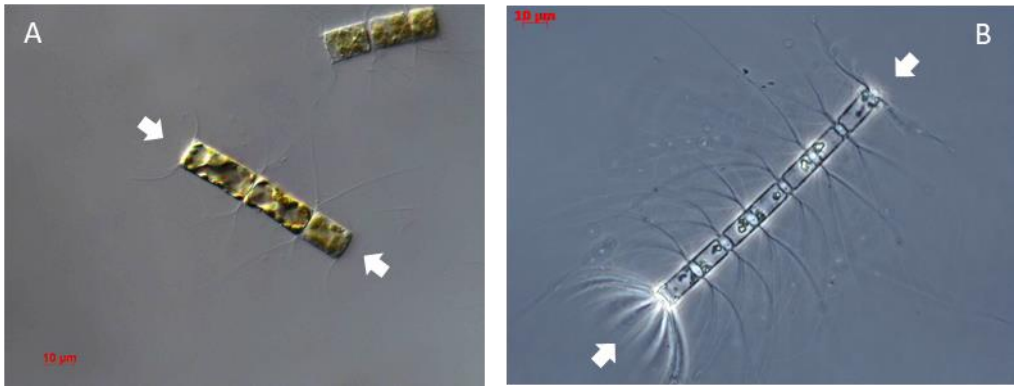


Fig. 2.1. Different orientation of terminal setae on the terminal valves of a colony of *Bacteriastrum* sections *Isomorpha* (A) and *Sagittata* (B). (A) *B. hyalinum* and (B) *B. elegans*.

Chaetoceros, with currently well over 200 taxonomically accepted species, is arguably the most diverse genus of planktonic diatoms in the marine realm (Guiry and Guiry, 2018). Most of current knowledge about its systematics dates back to the 19th century when, after the description of the material from the Antarctic expedition of Captain Rofs (1841-1843) by Ehrenberg (1844), several efforts have been made to fit this huge diversity into different taxonomic categories. The first attempt was made by Gran (1897), who divided *Chaetoceros* in two subgenera, *Phaeoceros* (now *Chaetoceros*) and *Hyalochaete*, basing on the distribution of chloroplasts. *Chaetoceros* has numerous small chloroplasts throughout the body of the cell and the setae, which are thick, very long, and armed with conspicuous spines (Hasle and Syvertsen, 1996). On the contrary, members of *Hyalochaete* have usually one or few chloroplasts only within the cell body, and setae are usually thin and more fragile (Fig. 2.2).

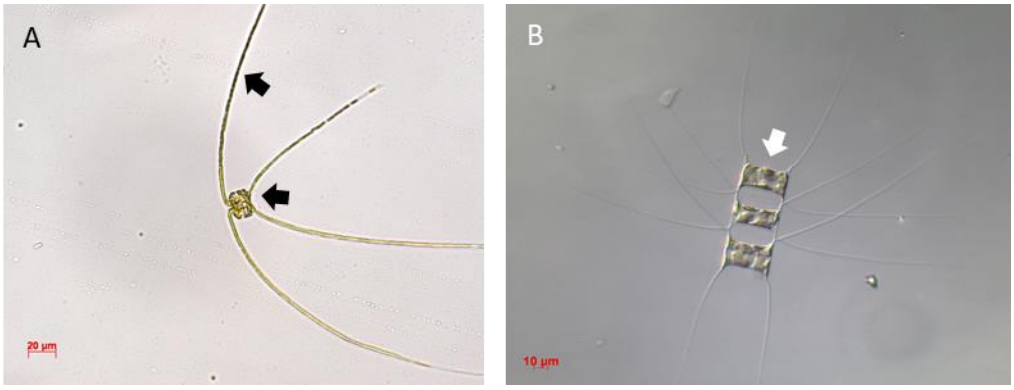


Fig. 2.2. Chloroplasts disposition in the subgenera *Chaetoceros* (A) and *Hyalochaete* (B) of *Chaetoceros*. (A) *C. peruvianus* 2 and (B) *C. decipiens*.

In addition, species belonging to the subgenus *Chaetoceros* exhibit rimoportulae (labiate processes) in both intercalary and terminal valves whereas in *Hyalochaete* these processes are observed only in terminal valves (Hasle and Syvertsen, 1996). *Chaetoceros* contains mostly oceanic species in which resting spores are lacking or unknown, except for *C. eibenii* (Grunow) Meunier (Jensen and Moestrup, 1998). After Gran, the two subgenera were further divided in sections by Ostenfeld (1903), Gran (1905) and, in recent times, by Hernández-Becerril (1991; 1993a; 1996), reaching the current number of 22 (Rines and Theriot, 2003). Furthermore, a third subgenus, *Bacteriastroidea*, was created for the only species *C. bacteriastroides* (Hernández-Becerril, 1993b).

Each of these infrageneric taxa are based on one or a few distinctive morphological features rather than on a formal cladistic analysis of all available characters and their states. Rines and Hargraves (1988) and Rines and Theriot (2003) pointed out some of these features are plastic, and so not reliable for a phylogenetic investigation. Cladograms inferred by Rines and Theriot (2003) from morphological information resolved *Bacteriastrum* inside paraphyletic *Hyalochaete*, which was resolved in its turn in paraphyletic subgenus *Chaetoceros*. Kooistra et al. (2010) reported similar topologies between phylogenies inferred from partial 28S rDNA sequences and from morphological information from the same strains. They resolved *Bacteriastrum* and monophyletic

subgenus *Chaetoceros* inside paraphyletic *Hyalochaete*. Yet, their study included fewer species than that of Rines and Theriot (2003).

Recent studies in Chaetocerotaceae have provided detailed morphological and ultrastructural illustrations as well as sequence data of numerous taxa, many of which are new to science (Kooistra et al., 2010; Li et al., 2013; 2017; Bosak et al., 2015; Gaonkar et al., 2017; 2018; Xu et al., 2019). However, most of these studies generally focused on the diversity within sections, and therefore the phylogenetic status of the investigated taxa remains to be resolved. Many studies used only the partial 28S rDNA as molecular marker, which poorly resolves the basal ramifications and therefore does not clarify relationships among the sections.

In this chapter, I infer a phylogeny of the family Chaetocerotaceae from a concatenated alignment of two nuclear (18S and 28S), two plastid (*rbc*L and *psb*A) and one mitochondrial (COI) gene gathered from 100 strains. Furthermore, I use the obtained tree to assess if the genera and the various infrageneric taxa are monophyletic as well as the validity of traditional classification scheme. This tree will also serve as a template to map characters and their states in future researches in order to reconstruct their evolutionary history. In this way, new insights will be gained on the evolution and diversification of one of the most species-rich and abundant marine planktonic diatom families.

2.2. Materials and methods

2.2.1. Taxon sampling, outgroups selection and DNA extraction

For this investigation, I used a total of 100 diatom strains (Table A2.1, Fig. A2.1, Appendix II), from all over the diversity of *Chaetoceros* (Rines and Hargraves, 1988; Guiry and Guiry, 2018) and *Bacteriastrum* (Van Landingham, 1968; Sarno et al., 1997; Godrijan et al., 2012; Guiry and Guiry, 2018). Most of the strains have been previously isolated from various localities (Table A2.1, Appendix II) and grown as monoclonal cultures in 74 ml polystyrene cell culture flasks (Corning Inc., NY, USA) filled with 30 ml

of f/2 medium at the following conditions: salinity of 36 ‰, 15 °C, 12:12 h light:dark cycle and a photon flux density of 50 μmol m^{-2} s^{-1} provided by cool white (40 W) fluorescent tubes (Gaonkar, 2017; Gaonkar et al., 2018). For the choice of outgroup sequences, I used the phylogenetic tree of diatoms by Theriot et al. (2015). I have chosen a nested set of taxa within the bipolar centric diatoms close to Chaetocerotaceae and for which there were GenBank sequences available for most, if not all, of the gene regions used in the present study, and from the same strain (Table A2.1, Appendix II). DNA was here extracted only for specimens not available in Gaonkar et al. (2018) and following the same protocol. DNA quantification was done by Nanodrop spectrophotometry.

2.2.2. Selection of genes, amplification and sequencing

To reconstruct the evolutionary history of Chaetocerotaceae, I used the information of five genes: two nuclear, encoding the small rDNA subunit (18S) and the D1 and D3 hypervariable domains of the large rDNA subunit (28S); two plastid, the rubisco large-subunit and the D1 protein-coding gene of the photosystem II (*rbc*L and *psb*A respectively); and a portion of the subunit I of cytochrome c oxidase gene (COI, mitochondrial). The sequences of 18S and 28S were mostly obtained from Gaonkar et al. (2018), except for the new strains here extracted and amplified (Table A2.1). All loci except 28S and COI were amplified for virtually their entire length using the primers listed in Table 2.1.

Table 2.1. List of the primers used for phylogenetic inference.

Gene	**Primer sequence (5'-3')**	**Reference**
18S		
SSU-F	TCYAAGGAAGGCAGCAGGCGC	Hamsher et al. (2011)
SSU-R	GTTTCAGCCTTGCGACCATACTCC	Ki et al. (2007)
28S		

D1R	ACCCGCTGAATTTAAGCATA	Scholin et al. (1994)
D3Ca	ACGAACGATTTGCACGTCAG	Scholin et al. (1994)
***rbc*L**		
F	GTGACCGTTACGAATCTGGTG	Fox and Sorhannus (2003)
R	CTGTTTCAGCGAAATCAGC	Fox and Sorhannus (2003)
***psb*A**		
*psb*A-F	AGTACCACATAATGGTTGTCGCC	Yoon et al. (2002)
*psb*A-R1	ACTTCATCAGCAGATTTTCGAC	Yoon et al. (2002)
COI		
GazF2	CAACCAYAAAGATATWGGTAC	Evans et al. (2007)
KEdtmR	CAAATAAAATTRATWGCWCCTAA	Evans et al. (2007)

PCR amplification protocols were adjusted according to the success or yield of amplification in different species. Regardless of the protocol, each reaction was conducted in a final volume of 20 µL consisting of: 5X Phusion HF Buffer, 0.2 mM dNTPs, 0.5 µM forward and reverse primers, 1 U Phusion® DNA Polymerase, approximately 50 ng of DNA and water to volume.

Nuclear genes were amplified at the conditions specified by Gaonkar et al. (2018). For *rbc*L and *psb*A genes, a first protocol including initial denaturation at 98 °C for 3 min and 34 cycles each with denaturation at 98 °C for 30 s, annealing at 62 to 45 °C (lowering the T of 0.5 °C/cycle) for 25 s, and extension at 72 °C for 1 min and 30 s was performed. In case of lack of amplification or poor yield, for *rbc*L the annealing temperature was lowered to 55-51.6 °C in steps of -0.1 °C/cycle, whilst for *psb*A to 52 °C. For the amplification of COI marker, the following protocol was applied: initial denaturation at 98 °C for 5 min, and 30 cycles each with a denaturation step at 98 °C for 1 min, annealing at 50 °C for 1 min, and extension at 72 °C for 45 s. The annealing temperature was lowered to 45 °C in samples providing poor yield. The amplification of 18S and 28S was carried out as specified in Gaonkar et al. (2017).

The success of PCRs was checked by electrophoresis in 1.5% agarose gel and 0.5 X TBE (Tris-Borate-EDTA). PCR products were purified either from agarose gel with the DNA Isolation Spin-Kit Agarose (AppliChem, Darmstadt, Germany) or directly from PCR tubes using the QIAquick® PCR Purification Kit (Qiagen, Hilden, Germany), whether multiple or single bands were observed following electrophoresis, respectively.

Purified DNA was sequenced using the BigDye Terminator v3.1 sequencing kit on a 48 capillaries-3730 DNA Analyzer (Life Technologies, ThermoFisher Scientific) at the Molecular Biology facility available at the SZN. PCR products were sequenced using both forward and reverse primers used for amplification. For 18S, two additional internal primers were used (Ch-528F and Ch-1055R, Gaonkar et al., 2018), whilst only one for *rbc*L, primer located at about 500 bp downstream the forward primer (*rbc*LinF, 5'-GTCGTGTAGTTTTCGAAG-3', present study).

2.2.3. Sequence editing and alignment

The electropherograms generated by Sanger sequencing were manually checked using Seq Scanner v2.0 (Applied Biosystems, ThermoFisher Scientific) and then, for 18S, 28S, *rbc*L and *psb*A, the resulting reads were assembled in contigs using ChromasPro v2.1.4 (Technelysium, Pty, Ltd) to generate the amplified fragment. For 18S and partial 28S data not generated in the present study, I used the sequences provided in Gaonkar et al. (2018) with introns removed. Sequences were aligned using ClustalX2 (Larkin et al., 2007) setting the parameters of pairwise and multiple alignment as specified in Hall (2004). Data were visualised and graphically edited in the R (R Core Team, 2018) working package apex (Jombart et al., 2017). Each gene matrix was then concatenated using Mesquite v3.51 (Maddison and Maddison, 2018) and visually checked.

2.2.4. Nucleotide composition and substitution saturation analyses

Base composition and substitution saturation are among the main factors known to affect phylogenetic reconstructions (Foster and Hickey, 1999; Moreira and Philippe, 2000; Theriot et al., 2015). Model-based phylogenetic methods usually assume that the aligned nucleotides evolve under homogeneous conditions (e.g. Jayaswal *et al.*, 2005), but the risk of phylogenetic errors increases if these conditions are violated (Ho and Jermiin, 2004; Jermiin *et al*., 2004). In order to detect putative base compositional heterogeneity in the dataset, I performed a χ^2 test of homogeneity of state frequencies across taxa on each gene partition (18S, 28S, *rbc*L, *psb*A, and COI) using PAUP* v4.0a (build 159) (Swofford, 2002). I also checked if substitution saturation was occurring at 3rd codon position of protein-coding alignments (*rbc*L, *psb*A, and COI) using the software DAMBE v6.4.107 (Xia, 2017). I calculated the proportion of invariant sites (P_{inv}) for the 3rd codon position of each gene using the NJ algorithm, and I used the obtained value to implement the saturation test by Xia et al. (2003). This test calculates the index of substitution saturation (I_{ss}) by sampling different subsets of sequences, and compares it to critical I_{ss} value ($I_{ss.c}$) at which the sequences will begin to fail to recover the true tree (Xia et al., 2003). Sequences are considered to have experienced little saturation when I_{ss} is significantly smaller than $I_{ss.c}$ (Xia et al., 2009).

2.2.5. Model selection and phylogenetic inference

I calculated the best-fitting model of nucleotide sequence evolution for each gene using the corrected Akaike information criterion (AICc) in PartitionFinder v.2.1.1 (Lanfear et al., 2016). The GTR+G+I model was favoured over the other models for all the genes considered. To ascertain if the evolutionary histories inferred from different cellular compartments were congruent, I inferred Maximum Likelihood (ML) trees using RAxML (Stamatakis, 2014) on the concatenated nuclear (18S and 28S) and plastid (*rbc*L and *psb*A) datasets and on the mitochondrial, single gene COI matrix. For ML inference, I conducted

100 ML tree searches under the GTR+G+I model of nucleotide substitution and then I calculated bootstrap support values by means of 1000 bootstrap replicates. The resulting nuclear, plastid and mitochondrial trees were checked for possible conflicts in their topology. Subsequently, I concatenated the five genes and inferred multigene phylogenetic trees using Maximum Parsimony (MP), ML and Bayesian Inference (BI). MP inference was conducted in PAUP* v4.0a (build 159) (Swofford, 2002). Heuristic tree searches comprised 10 random-addition replicates, TBR branch swapping, ACCTRAN character-state optimization, and gaps coded as missing data. Branch support was calculated by bootstrap analysis using 1000 bootstrap replicates. ML analysis was performed with IQ-TREE v1.6.8 (Nguyen et al., 2014) using the partition scheme suggested by PartitionFinder (GTR+G+I for each gene, -spp option), empirical base frequencies (+F option) and 1000 bootstrap replicates (-b option). A Bayesian tree was inferred using MrBayes v3.2.6 (Ronquist et al., 2012) using the GTR+G+I model (lset nst=6, rates=invgamma). All nucleotide substitution model parameters were unlinked across partitions and the different partitions were allowed to evolve at different rates (prset ratepr = variable). I ran four concurrent chains (one cold and three heated) for 10,000,000 generations and recorded samples every 1000 generations. Convergence and effective sample sizes (ESS) for all parameters were analysed in Tracer v.1.6 (Rambaut et al., 2014), the latter considered valid above the threshold of 200. Based on the results of Tracer analysis, I discarded the first 25% of the samples as burn-in.

2.2.6. Morphological sections and species assignment

In order to assign the strains here utilised to existing sections, I retrieved the descriptions of the morphological characteristics defining sections from the literature. For *Bacteriastrum*, I referred to Pavillard (1924; 1925) and Cupp (1943), whilst for *Chaetoceros* to Ostenfeld (1903), Gran (1897), Cupp (1943) and Hernández-Becerril (1996). I also integrated information from recent emendations or revisions of sections in

Chaetoceros (e.g. Li et al., 2016; Xu et al., 2019). Then, I assigned each taxon considered in the phylogenetic analysis to the relevant section using the morphological information provided in Gaonkar et al. (2018) and references therein. An illustration of a typical *Chaetoceros* species with the morphological terminology used here is provided in Fig. 2.3.

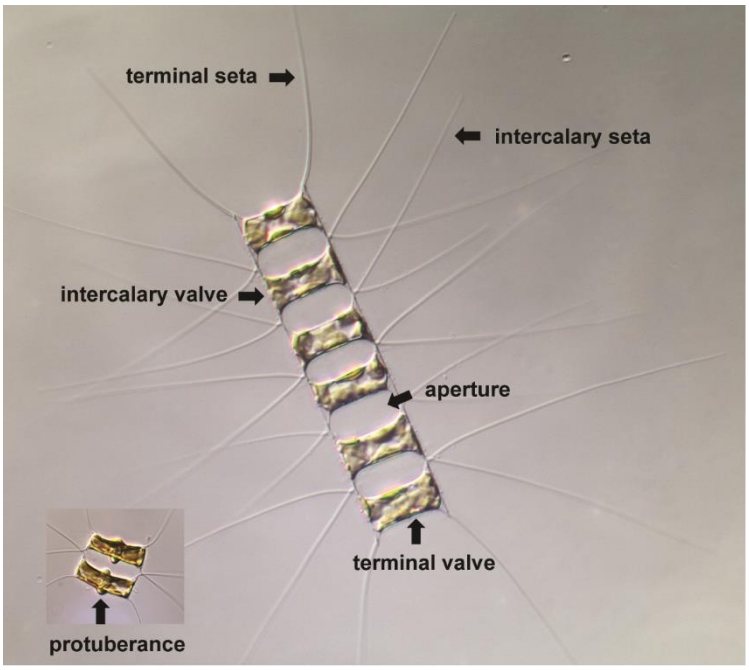


Fig. 2.3. Schematic representation of a typical *Chaetoceros* species, with the main morphological features relevant to this analysis.

2.3. Results

2.3.1. Dataset characteristics

Of all the genes here amplified, the highest amplification rate was obtained for *psb*A (83), followed by *rbc*L (74), and COI (52). For 18S and 28S, we used in total 92 and 88 sequences respectively from Gaonkar et al. (2018) plus six here amplified (three for 18S and three for 28S). A graphical overview of single gene and concatenated alignments is provided in Fig. 2.2. The low amplification success of COI is likely due to primer mismatches with their intended target regions. Indeed, the primers were developed against a conserved region within an exon of the pennate diatom *Sellaphora*. The known occurrence of introns in mitochondrial genomes of diatoms (Chaetocerotaceae included) as

well as the high substitution rate of the marker may have hampered primer-fit. The nucleotide sequences of *rbc*L, *psb*A and COI as well as newly generated 18S and 28S are available at the accession numbers listed in Table A2.1. The concatenated dataset (Table A2.2) included 100 strains (6 *Bacteriastrum* and 60 *Chaetoceros* species) and 5138 characters partitioned as follows: 18S (bp 1-1724), 28S (bp 1725-2495), *rbc*L (bp 2496-3806), *psb*A (bp 3807-4733), and COI (bp 4734-5138). The datasets organised per genomic compartment were as follows: 97 strains and 2495 characters for the nuclear data, 94 strains and 2238 characters for the plastid data, and 52 strains and 405 characters for the mitochondrial one.

I did not find any significant saturation at the 3rd codon positions of *rbc*L, *psb*A and COI genes ($I_{ss} < I_{ss.c}$, Table A2.3 in Appendix II) and, therefore, I assumed that the 1st and 2nd codon positions, known to evolve slower than the 3rd, are also not saturated. The results of this test indicated that the phylogenetic signal of such genes was not eroded by the substitution rates and that sequence similarity is largely due to homology. The χ^2 test of homogeneity of state frequencies across taxa detected no compositional heterogeneity ($p > 0.05$, Table A2.4 in Appendix II), so excluding its potential impact on phylogenetic inferences.

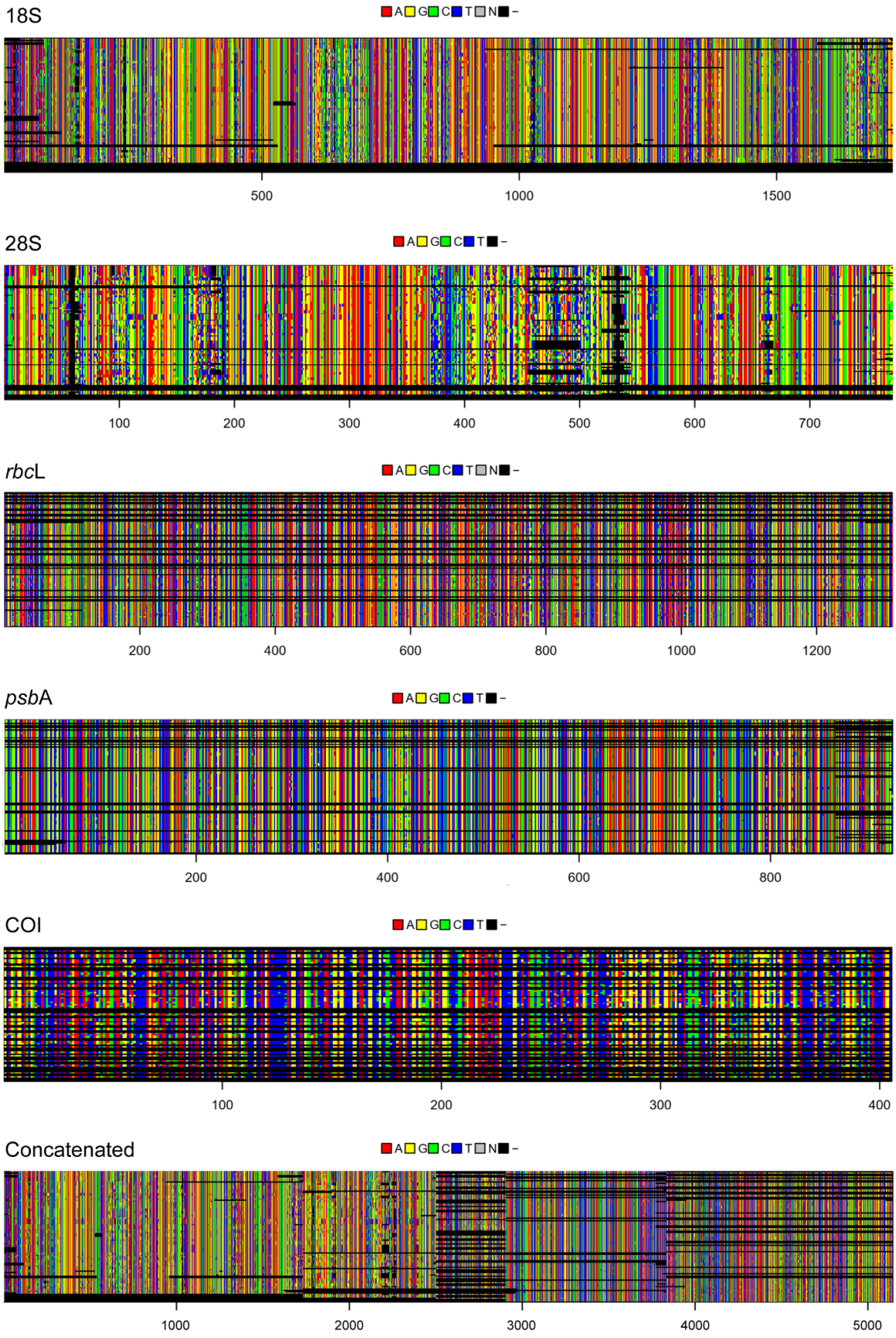


Fig. 2.4. Individual and concatenated sequence alignments of Chaetocerotaceae dataset. Each row represents an algal strain. N = undetermined bases, – = missing data.

2.3.2. Assignment of species to sections

The list and the description of the sections including the species included in the present study is provided in Table A2.5. I was able to assign most of the species to an extant section (Table A2.5, Appendix II) with few exceptions. Among the latter, was the group constituted by *C.* cf. *vixvisibilis* and *C.* sp. clades Na11C3, Na26B1, Na28A1 and Va7D2, encompassing heterogeneous taxa that did not show very distinctive morphological features. Other exceptions were *C.* cf. *pseudodichaeta*, *C. costatus* and *C. throndsenii*, which show distinctive morphological features not included in any extant section.

2.3.3. Nuclear, plastid and mitochondrial phylogenies

The concatenated nuclear ML tree (18S and partial 28S; Fig. A2.4) resolved *Bacteriastrum* as sister to the genus *Chaetoceros* with high bootstrap support (99%). The subgenus *Chaetoceros* was found to be monophyletic (73 BP), whilst *Hyalochaete* was paraphyletic. All terminal clades were fully resolved, whilst some internal nodes were poorly resolved (Fig. A2.2, Appendix II). The topology of the concatenated plastid tree (*rbc*L and *psb*A; Fig. A2.3 in Appendix II) was not in conflict with that of the nuclear tree, and where it was not in agreement, bootstrap support for those different relationships was not relevant. The only example for such a different relationship was the position of *Bacteriastrum*, which was recovered inside the genus *Chaetoceros*, though without bootstrap support (Fig. A2.3 in Appendix II). In general, the topology is as in the nuclear tree, but bootstrap support for many of the clades is low compared with the nuclear dataset. The COI tree (Fig. A2.4, Appendix II) was rooted using *Bacteriastrum* because no outgroup sequences were available in GenBank. The general topology of the mitochondrial tree resembles that of the trees inferred from the nuclear and plastid datasets, but the majority of the clades received insufficient bootstrap support. To summarise, there was no conflict in the tree topologies inferred from different genomic compartments.

2.3.4. Concatenated phylogenies

ML and BI phylogenies inferred from all the five gene regions concatenated showed the same topology (Fig. 2.5). *Bacteriastrum* formed a well-supported clade as sister to a clade comprising the genus *Chaetoceros* (72 BS, 0.99 PP). Within *Bacteriastrum*, phylogenetic relationships among taxa were well resolved but inconsistent with the sections (Fig. 2.5).

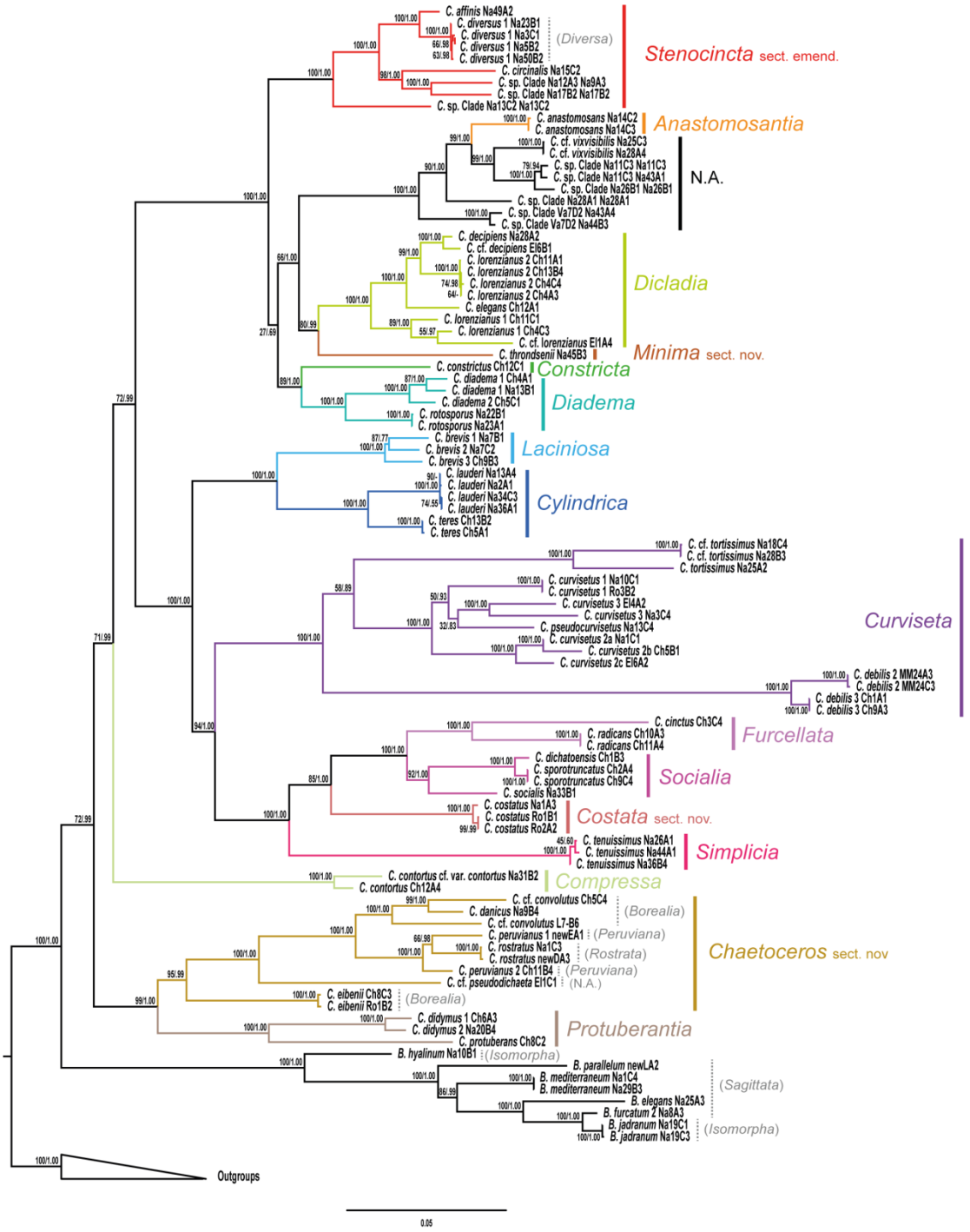


Fig. 2.5. Multigene Maximum Likelihood and Bayesian phylogenetic trees. Numbers at the basis of each node indicate the bootstrap support and the posterior probability respectively. Colours refer to the morphological section to which each taxon was assigned. In grey are indicated the rejected sections. N.A. = species not assigned to any existing section.

Within *Chaetoceros*, the first clade to branch off comprised in its turn a clade with taxa of section *Protuberantia* (*C. didymus* / *C. protuberans*) as sister to a clade with taxa of

subgenus *Chaetoceros.* Strong support for *Protuberantia* as sister to the subgenus *Chaetoceros* left *Hyalochaete* paraphyletic (Fig. 2.5). Within the subgenus *Chaetoceros*, section *Borealia* was not monophyletic. The remaining taxa in *Hyalochaete* were recovered in a clade (71 BS, 0.99 PP) in which a monophyletic section *Compressa* was resolved as sister to a clade with all remaining taxa (72 BS, 0.99 PP). This clade branched in its turn into two large and well supported clades. The lower one of these in Fig. 2.5 comprised the monophyletic sections *Laciniosa, Cylindrica,* C*urviseta, Furcellata, Socialia, Simplicia* and a clade with *C. costatus,* and the upper one included in essence three clades. One of these comprised section *Diversa* (only *C. diversus*) inside a paraphyletic section *Stenocincta.* A second one comprised a clade with *C. anastomosans* (Section *Anastomosantia*), *C.* cf. *vixvisibilis*, and strains belonging to a series of not yet formally described species for which morphological information is available in Gaonkar et al. (2018) as sister to a clade comprising the monophyletic section *Dicladia* and *C. throndsenii*. The third one contained the monophyletic section *Diadema* as sister to *C. constrictus* (section *Constricta*).

The MP tree was congruent with the ML and BI trees but exhibited a few poorly and unresolved relationships (Fig. A2.5). Nonetheless, the position of *Bacteriastrum* as sister genus to *Chaetoceros* was confirmed as well as monophyly of subgenus *Chaetoceros* within paraphyletic *Hyalochaete*. In summary, the three phylogenetic inference methods provided the same results, reinforcing the hypotheses of evolutionary relationships here inferred.

2.3.5. Comparison between morphological sections and molecular clades

Given the morphological assignment of taxa to sections and their phylogenetic positions in the concatenated ML and BI trees, I was able to name 16 clades in *Chaetoceros* and 2 in *Bacteriastrum* using the taxonomic division in sections (Table A2.4, Appendix II). A few taxa were not assigned to any section. These consisted of: i) a clade of species that have

not yet been formally described and for which an in-depth morphological and ultrastructural analysis is still needed (*C.* cf. *vixvisibilis*, *C.* spp. clades Na11C3, Na26B1, Na28A1 and Va7D2); ii) the minute species *C. throndsenii*; iii) *C. costatus*, and iv) *C.* cf. *pseudodichaeta*. As result, I emended one section, rejected seven and erected three new ones (Fig. 2.5, Table 2.1; see Discussion). The new classification system for the taxa here investigated is shown in Table 2.1. I also assigned to each section species for which both morphological and molecular information was available in literature (Table 2.2).

Table 2.2. Classification scheme of the family Chaetocerotaceae. Only sections including taxa utilised in the present study are shown. "Reference for description" refers to publications in which the section is described or amended. "Reference for assignation" refers to publications in which both morphological and molecular information of the species are available.

Genus *Bacteriastrum* Shadbolt	
No sectional division	
Genus *Chaetoceros* Ehrenberg	
Section	
Anastomosantia Ostenfeld	**Description:** setae united by a bridge. Chains mostly loose. **Reference for description:** Hernández-Becerril (1996). **Assigned species:** *C. anastomosans*. **Reference for assignation:** Gaonkar et al. (2018).
Chaetoceros sect. nov. Sarno, D. De Luca and Kooistra	**Description:** species with numerous chloroplasts in the cell body and in the setae. Robust, thick, and often very long setae armed with small, often elongated spines. Rimoportula on every valve with the exception of *C. pseudodichaeta*, which has rimoportula only on terminal valves. **Reference for description:** this study. **Assigned species:** *C. atlanticus*, *C. castracanei*, *C. convolutus*, *C. danicus*, *C. dichaeta*, *C. eibenii*, *C. peruvianus*, *C. pseudodichaeta*, *C. rostratus*. **Species assignation:** Gaonkar et al. (2018).
Compressa Ostenfeld; emended by Yang Li and Lundholm (in Xu et al., 2019)	**Description:** valves broadly elliptical to compress. Numerous small chloroplasts in each cell. Apertures usually moderately large. Terminal setae little different from others. Intercalary setae of two types: thin, common setae and heavy special setae. Heavy setae contorted with spiralling rows of spines and poroids, or heavy setae not visually contorted lacking rows of spines and poroids. Resting spores smooth or with a row of spicules. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Xu et al. (2019). **Assigned species:** *C. acadianus*, *C. bifurcatus*, *C. compressus*, *C. contortus*,

	C. hirtisetus, *C. millipedarius*. **Species assignation:** Chamnansinp et al. (2015); Gaonkar et al. (2018); Xu et al. (2019); Kaczmarska et al. (2019).
Constricta Ostenfeld	**Description:** cells with one or two chloroplasts and a marked constriction at the base of the valve mantle. Girdle at least one-third the length of the cell. Terminal setae mostly thicker than the others. Resting spores, when present, about the middle of the cell with numerous spines on both valves. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Assigned species:** *C. constrictus*. **Species assignation:** Gaonkar et al. (2018).
Costata sect. nov. Sarno, D. De Luca and Kooistra	**Description:** chains generally long, without differentiated terminal setae. One chloroplast. Each valve possesses four submarginal flattened protuberances, two on each pole of the valve, joining with those of the sibling valves. Girdle bands with a distinct thickened longitudinal rib at one edge also visible in LM. **Reference for description:** this study. **Assigned species.** *C. costatus* **Species assignation:** Gaonkar et al. (2018).
Curviseta Ostenfeld; emended by Gran	**Description:** chains usually curved, with setae all bent in one direction without special end cells. One chloroplast. **References for description:** Ostenfeld (1903); Gran (1905); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. curvisetus*, *C. debilis*, *C. pseudocurvisetus*, *C. tortissimus*. **Species assignation:** Gaonkar et al. (2018).
Cylindrica Ostenfeld	**Description:** cells with valves nearly circular (cylindrical). Apertures very narrow. Small, numerous chloroplasts. Terminal setae not thicker than others. Resting spores about middle of the cells, smooth or with spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. lauderi*, *C. teres.* **Species assignation:** Gaonkar et al. (2018).
Diadema (Ehrenberg) Ostenfeld; emended by Gran	**Description:** one chloroplast per cell. Chains long with conspicuous terminal setae. Primary valve of resting spores with branched processes or crown of spines, or sometimes smooth. **References for description:** Ostenfeld (1903); Gran (1905); Cupp (1943). **Assigned species:** *C. diadema*, *C. rotosporus*, *C. seiracanthus*, *Chaetoceros* sp. Clade Na13C1. **Species assignation:** Li et al. (2013); Gaonkar et al. (2018).
Dicladia (Ehrenberg) Gran; emended by Lebour	**Description:** multiple chloroplasts per cell and setae with large pores. Terminal and intercalary setae similar. Resting spores, when known, with two horns armed with small branches on primary valves. **References for description:** Gran (1905); Lebour (1930); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Assigned species:** *C. decipiens*, *C. elegans*, *C. laevisporus*, *C. lorenzianus*, *C. mannaii*, *C. mitra*, *C. pauciramosus.* **Species assignation:** Li et al. (2017); Chen et al. (2018).
Furcellata Ostenfeld	**Description:** chains generally loose, without differentiated terminal setae. One chloroplast. Resting cells eccentrically arranged in mother cell, lying close together two and two, with thick coalesced setae; with smooth valves or with short spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. cinctus*, *C. radicans*. **Species assignation:** Gaonkar et al. (2017).
Laciniosa Ostenfeld	**Description:** one or two chloroplasts per cell. Girdle rather long. Aperture large. Terminal setae usually thicker than the others, not diverging greatly.

	Resting spores smooth or with minute spines on primary valve, not in the middle of the cell. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. brevis*. **Species assignation:** Gaonkar et al. (2018).
Minima sect. nov. Sarno, D. De Luca and Kooistra	**Description:** very small, solitary species usually bearing one seta on a valve and one or two on the other. One chloroplast. Rimoportula very reduced in *C. throndsenii* and absent in *C. minimus*. **Reference for description:** this study. **Assigned species:** *C. minimus*, *C. throndsenii*. **Species assignation:** Gaonkar et al. (2018).
Protuberantia Ostenfeld; emended by Hernández-Becerril	**Description:** two chloroplasts per cell, each with a large pyrenoid situated in a protuberance in the middle of the valve surface. Valves with poroids. Resting spores paired with two long setae or free without setae. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Assigned species:** *C. didymus*, *C. protuberans*. **Species assignation:** Gaonkar et al. (2018).
Simplicia Ostenfeld	**Description:** cells small and fragile, generally single or two or three together. In case of chain formation, there is no differentiation of terminal setae. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. coloradensis*, *C. neogracilis*, *C. tenuissimus*. **Species assignation:** Li et al. (2016); Gaonkar et al. (2018).
Socialia Ostenfeld	**Description:** chains irregular and curved embedded in mucilage, forming irregularly spherical colonies. One chloroplast. Resting spores smooth or with small spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Assigned species:** *C. dichatoensis*, *C. gelidus*, *C. socialis*, *C. sporotruncatus*. **Species assignation:** Chamnansinp et al. (2013); Gaonkar et al. (2017).
Stenocincta Ostenfeld; emended by Sarno, D. De Luca and Kooistra	**Emended diagnosis:** one chloroplast per cell. Usually narrow aperture. Terminal setae generally thicker than the intercalary ones. Instead, *C. diversus* possesses thin terminal setae and generally two types of intercalary setae, differing in orientation and robustness. **References for description:** Ostenfeld (1903); Hernández-Becerril (1996); Gaonkar et al. (2018). **Morphologically assigned species:** *C. affinis*, *C. circinalis*, *C. diversus*, *Chaetoceros* sp. Clade Na12A3, *Chaetoceros* sp. Clade Na13C2, *Chaetoceros* sp. Clade Na17B2. **Species assignation:** Gaonkar et al. (2018).

2.4. Discussion

2.4.1. General comments to the dataset

The phylogenetic trees inferred from the three genomic compartments (nuclear, plastid and mitochondrial) are congruent, providing no indication of different evolutionary histories. According to this result, I conclude that during speciation of Chaetocerotaceae, the

corresponding gene copies in each species has been distributed in a pattern reflecting the parent species trees. This phenomenon is not universal, since gene trees and species trees do not always agree because of population-level lineage sorting (Pollard, et al., 2006), hybridization (McBreen and Lockhart, 2006), gene duplication and differential loss, and lateral gene transfer (LGT), where genes are exchanged between lineages (Dagan and Martin, 2006; Beiko et al., 2005; Leigh et al., 2008). However, most of phylogenetic studies dealing with the analysis of multiple genes often do not explicitly deal with the issue of congruence (Rokas et al., 2005; Qiu et al., 2006; James et al., 2006), making difficult to assess its extent across different taxa. In the few studies that analysed such issue in diatoms, no conflict among different gene trees was observed (e.g. Theriot et al., 2010; Souffreau et al., 2011; Kociolek et al., 2013).

In this study, as result of absence of conflicting topologies among trees, the concatenation of all the sequences increased the number of positively informative sites and so the phylogenetic signal (see e.g., Theriot et al., 2010; 2015). Indeed, the multigene tree shows better resolved relationships than the concatenated ones from each of the genomic compartments separately, as well as the trees based on single markers, e.g. in Gaonkar et al. (2018) and in Xu et al. (2019). Moreover, none of the markers included in our analysis shows saturation of the phylogenetic signal. Thus, I assume that the well supported clades in the concatenated tree can be used to make phylogenetically informed taxonomic decisions.

Most of the sections for which strains of multiple species have been included are monophyletic, and the synapomorphies of the clades are here used to validate, describe or emend the sections they belong to. For the purposes of this work, I aimed at a classification that is both supported phylogenetically and retains practical properties (Mayr, 1982; Benton, 2000). This approach is not mutually exclusive, considering that the objects of classifications should share similarities because of common descent (Mayr, 1942). I

retained only monophyletic sections, made emendations whenever possible and erected new sections only were complete and supported information was available.

2.4.2. Phylogenetic position of the genera Bacteriastrum *and* Chaetoceros

Results of the present study indicate that *Bacteriastrum* and *Chaetoceros* are each other's monophyletic sister genera. This finding contrasts with phylogenies inferred exclusively from partial 28S rDNA sequences, which resolve the former inside the latter, though with meagre support, if any (e.g., Bosak et al., 2015; Gaonkar et al., 2018). My results confirm the hypothesis that *Bacteriastrum* constitutes a genus different from *Chaetoceros* (e.g. Pritchard, 1861; Round et al., 1990; Hasle and Syvertsen, 1996). Within *Bacteriastrum*, the sections *Isomorpha* and *Sagittata* proposed by Pavillard (1925) are unsupported because the former is polyphyletic and the latter paraphyletic. This is because *B. jadranum*, placed in the section *Isomorpha* by Godrijan et al. (2012), is sister to members of section *Sagittata* and only distantly related to *B. hyalinum* (*Isomorpha*). The non-monophyly of these sections invalidates them and shows that their defining character states of terminal setae orientation are not synapomorphies. Thus, I reject the two sections since there is neither a phylogenetic reason nor a utilitarian one to maintain them.

2.4.3. Subgeneric division

The subgenus *Hyalochaete* was erected by Gran (1897) to include all the *Chaetoceros* species without chloroplasts in the setae. Therefore, this "catch-all" taxonomic category includes a highly diverse collection of species that basically share general features encountered in all *Chaetoceros* species; their only defining feature, absence of chloroplasts in the setae, is not a phylogenetically sound character state. Indeed, my results show that *Hyalochaete* is paraphyletic, and therefore, I reject it.

The subgenus *Chaetoceros* was formerly described as *Phaeoceros* by Gran (1897) to include species characterised by numerous plastids in both the cell body and the setae.

According to the concatenated phylogeny "plastids in the setae," (Gran, 1897) is a synapomorphy of subgenus *Chaetoceros*, whereas other features believed to define this subgenus, actually do not. Gran (1897) mentions "spores unknown" for subgenus *Chaetoceros.* None of the studies on species in this subgenus have reported spore formation, with one exception: *C. eibenii* (von Stosch et al., 1973; Jensen and Moestrup, 1998). Since this species is the first to branch off within the subgenus, ability to form resting spores, must have gone lost in the last common ancestor of the sister clade of *C. eibenii* because spore formation has been confirmed in most if not all of the other species in the genus *Chaetoceros* (Ishii et al., 2011). Thus, the absence of spores does not define the subgenus *Chaetoceros*. Another character state, "presence of rimoportulae in terminal as well as in intercalary valves" (see Hasle and Syvertsen, 1996) is a symplesiomorphy because strain El1C1 (Eilat, Israel) here identified as *C.* cf. *pseudodichaeta* resolves within the clade of the subgenus, but exhibits rimoportulae only in its terminal valves. Therefore, "presence of rimoportulae in terminal as well as in intercalary valves" is not a defining character state of the subgenus, either.

Although I cannot strictly reject the subgenus *Chaetoceros*, but having already rejected the subgenus *Hyalochaete* for its paraphyly, I argue that there is no utilitarian reason to keep it. It could be better treated as a new section *Chatoceros*, here proposed, to include all the species with chloroplasts not only in the central cell body, but also in the setae.

At this point, the only remaining subgenus is *Bacteriastroidea*. It was erected by Hernández-Becerril (1993b) to include the only species *C. bacteriastroides* for its peculiar morphology, intermediate between *Bacteriastrum* and *Chaetoceros* (cylindrical valves, intercalary ones with three pairs of setae, two of which very reduced). We agree that, considering the available data, it deserves a dedicated taxonomic category. However, no DNA is available for this species and, therefore, its molecular phylogenetic position within the Chaetocerotaceae is unknown. Genetic data may either confirm the validity of a dedicated taxonomic category or justify its inclusion into a pre-existing section. Therefore,

considering available information and following our way of action, I reject *Bacteriastroidea* as subgenus and consider it provisionally as a section of the genus *Chaetoceros*.

2.4.4. The sectional division

Ostenfeld (1903) was the first to subdivide the genus *Chaetoceros* into sections. Later authors (cit) added sections to accommodate species new to science that did not fit in the sections of Ostenfeld or to split pre-existing sections based on newly defined characters and their states. Newly described species are usually sorted without much ado into those existing sections (e.g. Li et al., 2016) or the sectional description needs to be emended only slightly to accommodate species new to science (Xu et al., 2019). However, some species such as *C. phuketensis* (Rines et al., 2000) are not. The results of my phylogenetic explorations show that the sections *Dicladia, Constricta, Diadema, Laciniosa, Cylindrica, Curviseta, Furcellata, Socialia, Simplicia, Compressa* and *Protuberantia* are monophyletic, and therefore, considered valid.

For the species that do not fit in any pre-existing section, a possible course of action would be to create a new section for every one of them showing a unique feature (e.g. section *Anastomosantia* for *C. anastomosans* with its silica bridge linking sibling setae; section *Rostrata* for *C. rostratus* with its fused rimoportulae of sibling valves). However, Rines et al. (2000) pointed out that this would lead *in extremis* to placing every morphologically distinct species in its own section, thereby defying the utilitarian purpose of sections. I recognise that some peculiar characters are important for species identification purposes, but I agree with Rines et al. (2000) to refrain from considering these as reasons to create new sections. Everytime the morphology of a new species does not quite fit the sectional description, I simply decided to emend the latter (see e.g., Xu et al., 2019).

For example, Ostenfeld (1903) placed *C. diversus* in a section called *Diversa* because this species' intercalary setae are far more robust than its terminal ones. This section was

maintained by Cupp (1943) and Hernández-Becerril (1996). However, the robust intercalary setae and delicate terminal setae of *C. diversus* resemble the robust terminal setae and delicate intercalary setae of the species in its paraphyletic "mother" section *Stenocincta*. Other characteristics such as a single plastid per cell and a narrow aperture are, in fact, shared between the two sections (Gaonkar et al., 2018). The phylogenetic position of *C. diversus* in our multigene phylogeny, inside the clade encompassing all the other taxa belonging to *Stenocincta*, provides further evidence of common ancestry. Therefore, I decided to emend the section *Stenocincta* to include taxa previously within *Diversa* and I hereby propose to reject the section *Diversa*.

Similarly, Ostenfeld (1903) and Hernández-Becerril (1996) considered the presence of spines on the spore valves the defining feature of the section *Diadema*. Instead, Cupp (1943), following Gran's (1905) emended description, provided a broader definition, which includes also features of the vegetative cells and which accommodates the spiny resting spores of *C. diadema* and *C. seiracanthus* as well as the smooth ones of *C. rotosporus* (Li et al., 2013). My phylogenetic tree supports the findings of Li et al. (2013) and Gaonkar et al. (2018) that the aforementioned taxa in section *Diadema* form a clade. Thus, all these species, including the recently described *C. rotosporus*, fit perfectly fine in the monophyletic section *Diadema*.

Most of the sections in my phylogeny are monophyletic and their defining character states are synapomorphies, but for some of them the taxonomic coverage is still low. Many more species need to be added to confirm their monophyly. For the sections *Compressa*, *Constricta*, *Rostrata* and *Simplicia* I was limited in testing their robustness because of the low numbers of species available, but I have no evidence that newly added will falsify the current classification into sections. For instance, in my phylogeny, only two species in section *Compressa* are available, but a phylogeny inferred from 28S rDNA sequences in Xu et al. (2019) shows this section to be monophyletic given far wider taxon coverage. Likewise, in my multigene phylogeny the section *Simplicia* is represented only by *C.*

tenuissimus. Li et al. (2016) included in their 18S rDNA phylogeny (fig. 25 of their paper) several other taxa in this section, most of which resolved together with *C. tenuissimus* in a weakly supported clade (clade IV). However, several strains that were poorly identified at morphological level, made some species polyphyletic and made the authors considering the section non monophyletic. In Gaonkar et al. (2018), *C. tenuissimus* forms a well-supported clade with taxonomically validated strains of *C. neogracilis* (section *Simplicia*; see Balzano et al., 2017). The 18S sequences of *C.* cf. *neogracile* in Li et al. (2016) are virtually identical to those in Balzano et al. (2017), and therefore they made me hypothesise that the section *Simplicia* is monophyletic.

A few species in my tree do not fit in any of the existing sections. For instance, Cupp (1943) placed *C. costatus* in the section *Stenocincta* whereas Lebour (1930) put it in *Curviseta* (under the name *C. adhaerens*). Instead, in my multigene phylogeny the strains of this species are recovered as nearest neighbour of a clade comprising sections *Socialia* and *Furcellata*. Neither morphological nor ultrastructural characters are shared with these neighbour sections. However, this species possesses several peculiar morphological features (e.g. four submarginal flattened protuberances joining with those of the sibling valves) that justify its inclusion into a dedicated section. Therefore, I propose to erect a new section for *C. costatus* (Section *Costata*). Such a section does not affect monophyly of the related sections *Furcellata, Socialia* and *Simplicia*.

The species description of *C. throndsenii* does not provide any assignment to a section (Marino et al., 1991). In the 18S and 28S trees by Gaonkar et al. (2018) it forms a clade with the morphologically similar species *C. minimus*, with which it shares a small cell size, a single cell habit, a reduction in the number of setae per cell (2 to 3) and a similarly shaped resting spore. These morphological similarities were already reported in Marino et al. (1991). I had no access to DNA of *C. minimus* and therefore this species was not included in my multigene tree. In my phylogeny, *C. throndsenii* is recovered on a long branch as sister to *Dicladia,* though it does not share any evident character state with this

section. Therefore, I propose placing *C. minimus* and *C. throndsenii* into a new section here called *Minima*.

Regarding *Anastomosantia*, Ostenfeld (1903) established this section for *C. anastomosans* based on the silica bridges linking sibling setae, which in our multigene tree constitutes an autapomorphy. Yet, this species is recovered in a well-supported clade with *C.* cf. *vixvisibilis* and a whole series of still undescribed taxa. *Chaetoceros dayaensis* (Li et al., 2015) also belongs to this clade (Gaonkar et al., 2018). Li et al. (2015) did not place *C. dayaensis* in any section nor did they establish a new section for it. In this case, I refrain from erecting a new section for the whole clade because the possible morphological synapomorphies defining this clade are still to be uncovered, and I keep the section *Anastomosantia* exclusively for *C. anastomosans*.

According to my multigene phylogeny, the sections *Borealia* (Ostenfeld 1903) and *Peruviana* (Hernández-Becerril, 1996) are not monophyletic and so I have to reject them. The presence of intercalary processes that link cells in chains in *C. rostratus*, defines the monotypic section *Rostrata* (Hernández-Becerril, 1998). Although these specialised processes are useful for taxonomic identification, they constitute an autapomorphy. Maintaining the section *Rostrata* requires the establishment of a whole series of additional sections in the clade, several of which will be monotypic, and without any clear defining character states, which is not particularly utilitarian. The same accounts for *C. atlanticus* and *C. dichaeta* in section *Atlantica* (Ostenfeld 1903), which are resolved in the phylogenies in Gaonkar et al. (2018) close to *C. peruvianus* (*Peruviana*) and *C. danicus* (*Borealia*). Therefore, I propose to reject not only *Borealia* and *Peruviana*, but also *Rostrata* and *Atlantica* and to erect a new Section *Chaetoceros* for all the taxa sharing the presence of chloroplasts in the central cell body as well as in the setae. The name *Chaetoceros* follows the rules of botanical nomenclature, according to which any

subdivision of a genus that includes the type specimen must adopt the name of the genus to which it is assigned.

Section *Chaetoceros* D. Sarno, D. De Luca and W.H.C.F. Kooistra, sect. nov.

Species with numerous chloroplasts in the cell body and in the setae. Robust, thick, and often very long setae armed with small, often elongated spines. Rimoportula on every valve with the exception of *C. pseudodichaeta*, which has rimoportula only on terminal valves (Fig. 2.6).

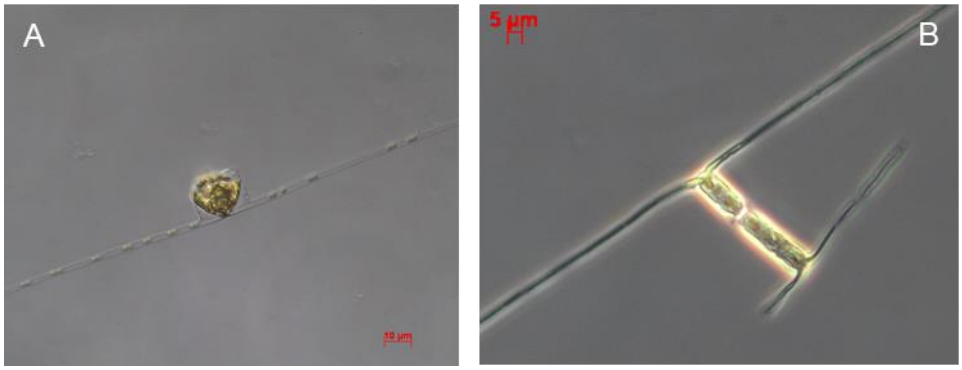


Fig. 2.6. *Chaetoceros danicus* (A) and *C. rostratus* (B), two members of the Section *Chaetoceros*.

Section *Costata* D. Sarno, D. De Luca and W.H.C.F. Kooistra, sect. nov.

Chains generally long, without differentiated terminal setae. One chloroplast. Each valve possesses four submarginal flattened protuberances, two on each pole of the valve, joining with those of the sibling valves. Girdle bands with a distinct thickened longitudinal rib at one edge also visible in LM (Fig. 2.7).

Fig. 2.7. *Chaetoceros costatus*, Section *Costata*. The arrow indicates the submarginal flattened protuberance typical of the Section.

Section *Minima* D. Sarno, D. De Luca and W.H.C.F. Kooistra, sect. nov.

Very small, solitary species usually bearing one seta on a valve and one or two on the other. One chloroplast. Rimoportula very reduced in *C. throndsenii* and absent in *C. minimus* (Fig. 2.8).

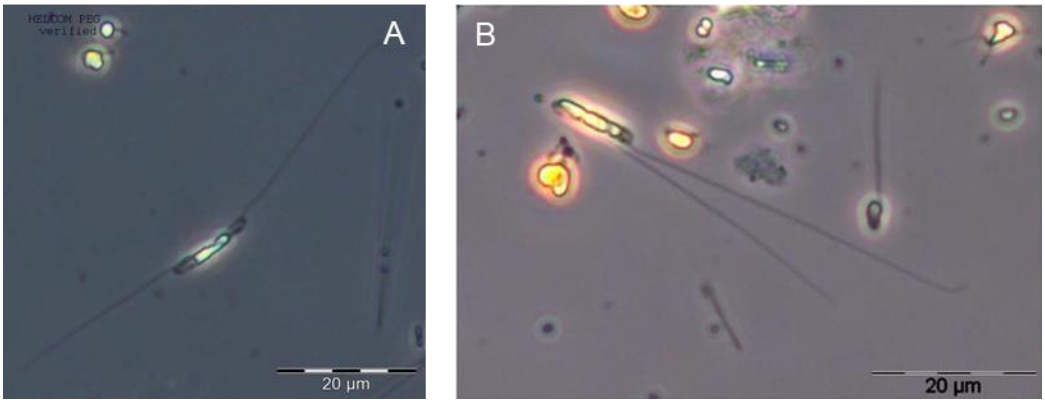


Fig. 2.8. *Chaetoceros minimus* (A) and *C. throndsenii* (B), two members of the Section *Minima*. Photo credit: Susanne Busch. From Nordic Microalgae (http://www.nordicmicroalgae.org).

Section *Stenocincta* Ostenfeld 1903 emend. D. Sarno, D. De Luca and W.H.C.F. Kooistra

Emended diagnosis: One chloroplast per cell. Usually narrow aperture. Terminal setae generally thicker than the intercalary ones. Instead, *C. diversus* possesses thin terminal

setae and generally two types of intercalary setae, differing in orientation and robustness (Fig. 2.9).

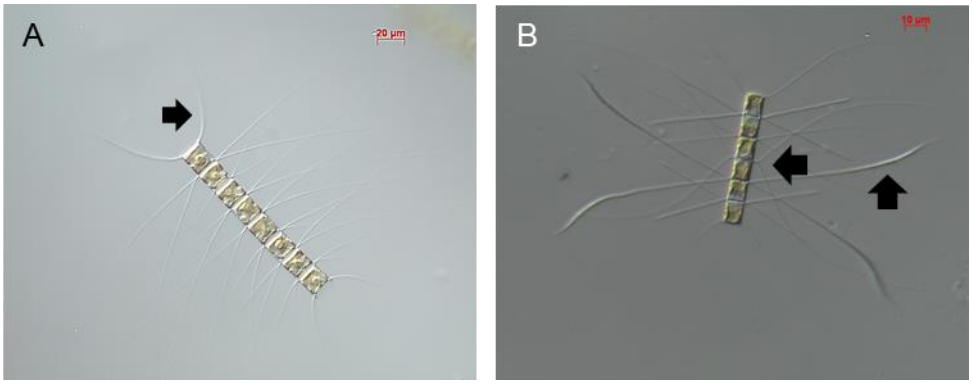


Fig. 2.9. *Chaetoceros affinis* (A) and *C. diversus* (B), two members of the Section *Stenocincta*. Arrows indicate some characteristics of the members of the Section, i.e. thick terminal setae (A) and two types of intercalary setae in *C. diversus* (B).

2.4.5. Future directions

The multigene analysis here inferred using the 18S, 28S, *rbc*L, *psb*A and COI genes has provided a robust phylogenetic hypothesis depicting the evolutionary history of Chaetocerotaceae. The comparison between infrageneric taxa based on morphology and the clades in the tree revealed congruence for most of them, falsified others, and highlighted that future work is needed on unresolved taxa. I rejected the three subgenera within *Chaetoceros* and seven sections (two in *Bacteriastrum* and five in *Chaetoceros*), emended one section and described three new ones. I refrained from elevating the sections into genera of their own. Splitting would be justified by the fact that the genetic distances among the chaetocerotacean sections are comparable with those observed among families or even orders in other diatom lineages. For instance, the Order Thalassiosirales has been split into a large series of narrowly defined genera. However, this has left the genus *Thalassiosira* paraphyletic. Although I have demonstrated that most of the Sections in *Chaetoceros* are monophyletic, I believe that the utilitarian principle has precedence. *Chaetoceros* is easily recognised because of its defining feature, the setae, which are

visible in LM. If the sections are elevated to genera, however, these new genera may not be recognised so easily. The sections can be further supported by ultrastructural features of valves and setae (Chamnansinp et al. 2015; Bosak and Sarno, 2017; Gaonkar et al., 2018; Xu et al., 2019) but this requires in depth comparison of the species in these sections. Future work may include the adding of new species to our phylogenetic tree as well as of new sections, in order to have a better and more complete view of the evolution of such important family of marine planktonic diatom. To date, this study represents a further step towards a better understanding of the evolution of Chaetocerotaceae.

References

Balzano, S., Percopo, I., Siano, R., Gourvil, P., Chanoine, M., Marie, D., Vaulot, D., Sarno, D. (2017). Morphological and genetic diversity of Beaufort Sea diatoms with high contributions from the *Chaetoceros neogracilis* species complex. *Journal of Phycology*, *53*(1), 161-187.

Beiko, R. G., Harlow, T. J., Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, *102*(40), 14332-14337.

Benton, M. J. (2000). Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead?. *Biological Reviews*, *75*(4), 633-648.

Bosak, S., Godrijan, J., Šilović, T. (2016). Dynamics of the marine planktonic diatom family Chaetocerotaceae in a Mediterranean coastal zone. *Estuarine, Coastal and Shelf Science*, *180*, 69-81.

Bosak, S., Sarno, D. (2017). The planktonic diatom genus *Chaetoceros* Ehrenberg (Bacillariophyta) from the Adriatic Sea. *Phytotaxa*, *314*(1), 1-44.

Chamnansinp, A., Li, Y., Lundholm, N., Moestrup, Ø. (2013). Global diversity of two widespread, colony-forming diatoms of the marine plankton, *Chaetoceros socialis*

(syn. *C. radians*) and *Chaetoceros gelidus* sp. nov. *Journal of Phycology*, *49*(6), 1128-1141.

Chamnansinp, A., Moestrup, Ø., Lundholm, N. (2015). Diversity of the marine diatom *Chaetoceros* (Bacillariophyceae) in Thai waters–revisiting *Chaetoceros compressus* and *Chaetoceros contortus*. *Phycologia*, *54*(2), 161-175.

Chen, Z. Y., Lundholm, N., Moestrup, Ø., Kownacka, J., Li, Y., 2018. *Chaetoceros pauciramosus* sp. nov. (Bacillariophyceae), a Widely Distributed Brackish Water Species in the *C. lorenzianus* Complex. *Protist, 169*(5), 615-631.

Continuous Plankton Recorder Survey Team. (2004). Continuous plankton records: plankton atlas of the North Atlantic Ocean (1958–1999). II. Biogeographical charts. *Marine Ecology Progress Series*, 11-75.

Cupp, E. E. (1943). *Marine plankton diatoms of the west coast of North America*. University of California Press, Berkeley.

Dagan, T., Martin, W. (2006). The tree of one percent. *Genome Biology, 7*(10), 118.

Drebes, G. (1972). The life history of the centric diatom *Bacteriastrum hyalinum* Lauder. *Nova Hedwigia, 39*, 95-110.

Ehrenberg, C. G. (1844). Einige vorläufige Resultate seiner Untersuchungen der ihm von der Südpolreise des Kapitän Ross. *Deutesche Akademie der Wissenschaften zu Berlin*, 182-207.

Evans, K. M., Wortley, A. H., Mann, D. G. (2007). An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist*, *158*(3), 349-364.

Foster, P. G., Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, *48*(3), 284-290.

Fox, M. G., Sorhannus, U. M. (2003). RpoA: A useful gene for phylogenetic analysis in diatoms. *Journal of Eukaryotic Microbiology*, *50*(6), 471-475.

Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Montresor, M., Sarno, D. (2017). Two new species in the *Chaetoceros socialis* complex (Bacillariophyta): *C. sporotruncatus* and *C. dichatoensis*, and characterization of its relatives, *C. radicans* and *C. cinctus*. *Journal of Phycology*, *53*(4), 889-907.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. *PLoS ONE*, *13*(12), e0208929.

Godrijan, J., Marić, D., Imešek, M., Janeković, I., Schweikert, M., Pfannkuchen, M. (2012). Diversity, occurrence, and habitats of the diatom genus *Bacteriastrum* (Bacillariophyta) in the northern Adriatic Sea, with the description of *B. jadranum* sp. nov. *Botanica marina*, *55*(4), 415-426.

Gran, H. H. (1897). Botanik. Prophyta: Diatomaceae, Silicoflagellata og Cilioflagellata. *Den Norske Nordhavs Expedition 1876-1878*, *7*, 1-36.

Gran, H. H. (1905). *Diatomeen in Brandt und Apstein, Nordisches Plankton, Botanischer Teil', vol. 19*. Kiel und Leipzig, Lipsius und Tischer.

Guiry, M. D., Guiry, G. M. (2018). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 14 March 2018.

Hall, B. G. (2004). *Phylogenetic trees made easy. A how-to manual, 2nd edition*. Sinauer Associates, Inc. Sunderland, Massachusetts.

Hamsher, S. E., Evans, K. M., Mann, D. G., Poulíčková, A., Saunders, G. W. (2011). Barcoding diatoms: exploring alternatives to COI-5P. *Protist*, *162*(3), 405-422.

Hasle, G. R., Syvertsen, E. E. (1996). Marine diatoms. In: *Identifying marine phytoplankton* (Ed. by C.R. Tomas), pp. 5-385. Academic Press, San Diego.

Hernández-Becerril, D. U. (1993b). Note on the morphology of two planktonic diatoms: *Chaetoceros bacteriastroides* and *C. seychellarus*, with comments on their

taxonomy and distribution. *Botanical Journal of the Linnean Society*, *111*(2), 117-128.

Hernández-Becerril, D. U. (1991). Note on the morphology of *Chaetoceros didymus* and *C. protuberans*, with some considerations on their taxonomy. *Diatom Research*, *6*(2), 289-297.

Hernández-Becerril, D. U. (1996). Morphological study of *Chaetoceros* species (Bacillariophyta) from the plankton of the Pacific Ocean of Mexico. *Bulletin of Natural History Museum of London (Botany)*, *26*(1), 1-73.

Hernández-Becerril, D. U., Granados, C. F. (1998). Species of the diatom genus *Chaetoceros* (Bacillariophyceae) in the plankton from the Southern Gulf of Mexico. *Botanica Marina*, *41*(1-6), 505-520.

Hernández-Becerril, D. U., Meave del Castillo, M. E., Lara Villa, M. A. (1993a). Observations on *Chaetoceros buceros* (Bacillariophyceae), a rare tropical planktonic species collected from the Mexican Pacific. *Journal of Phycology*, *29*(6), 811-818.

Ho, S. Y. W., Jermiin, L. S. (2004). Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, *53*(4), 623-637.

Ishii, K. I., Iwataki, M., Matsuoka, K., Imai, I. (2011). Proposal of identification criteria for resting spores of *Chaetoceros* species (Bacillariophyceae) from a temperate coastal sea. *Phycologia 50*, 351-362.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., ... Vilgalys, R. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, *443*(7113), 818-822.

Jayaswal, V., Jermiin, L. S., Robinson, J. (2005). Estimation of phylogeny using a general Markov model. *Evolutionary Bioinformatics*, *1*, 62-80.

Jensen, K. G., Moestrup, Ø. (1998). The genus *Chaetoceros* (Bacillariophyceae) in inner Danish coastal waters. *Nordic Journal of Botany*, *18*(1), 1-88.

Jermiin, L. S., Ho, S. W. H., Ababneh, F., Robinson, J., Larkum, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, *53*(4), 638-643.

Jombart, T., Archer, F., Schliep, K., Kamvar, Z., Harris, R., Paradis, E., Goudet, J., Lapp, H. (2017). apex: phylogenetics with multiple genes. *Molecular Ecology Resources*, *17*(1), 19-26.

Kaczmarska, I., Samanta, B., Ehrman, J. M., Porcher, E. M., 2019. Auxosporulation in *Chaetoceros acadianus* sp. nov. (Bacillariophyceae), a new member of the Section *Compressa*. Eur. J. Phycol. 54, 206-221.

Ki, J. S., Chang, K. B., Roh, H. J., Lee, B. Y., Yoon, J. Y., Jang, G. Y. (2007). Direct DNA isolation from solid biological sources without pretreatments with proteinase-K and/or homogenization through automated DNA extraction. *Journal of Bioscience and Bioengineering*, *103*(3), 242-246.

Kociolek, J. P., Stepanek, J. G., Lowe, R. L., Johansen, J. R., Sherwood, A. R. (2013). Molecular data show the enigmatic cave-dwelling diatom *Diprora* (Bacillariophyceae) to be a raphid diatom. *European Journal of Phycology*, *48*(4), 474-484.

Kooistra, W. H. C. F., Sarno, D., Hernández-Becerril, D. U., Assmy, P., Di Prisco, C., Montresor, M. (2010). Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia*, *49*(5), 471-500.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772-773.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T.

J., Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947-2948.

Leblanc, K., Arístegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.P., Kopczynska, E., Marshall, H., Peloquin, J., Piontkovski, S., … Yallop, M. (2012). A global diatom database – abundance, biovolume and biomass in the world ocean. *Earth System Science Data, 4*, 149-165.

Lebour, M. V. (1930). *The planktonic diatoms of northern seas*. The Ray Society, London.

Leigh, J. W., Susko, E., Baumgartner, M., Roger, A. J. (2008). Testing congruence in phylogenomic analysis. *Systematic Biology, 57*(1), 104-115.

Li, J., Kociolek, J. P., Gao, Y. (2016). *Chaetoceros coloradensis* sp. nov. (Bacillariophyta, Chaetocerotaceae), a new inland species from Little Gaynor Lake, Colorado, North America. *Phytotaxa*, *255*(3), 199-213.

Li, Y., Boonprakob, A., Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Hernández-Becerril, D., Chen, Z., Moestrup, Ø, Lundholm, N. (2017). Diversity in the globally distributed diatom genus *Chaetoceros* (Bacillariophyceae): Three new species from warm-temperate waters. *PLoS ONE, 12*, e0168887.

Li, Y., Lundholm, N., Moestrup, Ø. (2013). *Chaetoceros rotosporus* sp. nov. (Bacillariophyceae), a species with unusual resting spore formation. *Phycologia*, *52*(6), 600-608.

Li, Y., Zhu, S., Lundholm, N., Lü, S. (2015). Morphology and molecular phylogeny of *Chaetoceros dayaensis* sp. nov. (Bacillariophyceae), characterized by two 90 rotations of the resting spore during maturation. *Journal of Phycology, 51*(3), 469-479.

Maddison, W. P., Maddison, D. R. (2018). Mesquite: a modular system for evolutionary analysis. Version 3.51 http://www.mesquiteproject.org.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., … Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences, 113*(11), 1516-1525.

Marino, D., Giuffré, G., Montresor, M., Zingone, A. (1991). An electron microscope investigation on *Chaetoceros minimus* (Levander) comb. nov. and new observations on *Chaetoceros throndsenii* (Marino, Montresor and Zingone) comb. nov. *Diatom Research*, *6*(2), 317-326.

Mayr E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. Harvard University Press, Cambridge, Massachusetts.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution and inheritance*. Harvard University Press, Cambridge, Massachusetts.

McBreen, K., Lockhart, P. J. (2006). Reconstructing reticulate evolutionary histories of plants. *Trends in Plant Science, 11*(8), 398-404.

Moreira, D., Philippe, H. (2000). Molecular phylogeny: pitfalls and progress. *International Microbiology*, *3*(1), 9-16.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268-274.

Ostenfeld, C. H. (1903). Plankton from the sea around the Faeroes. *Botany of the Faeroes*, 558-611.

Pavillard J. (1924). Observations sur les Diatomees, 4eme serie. Le genre *Bacteriastrum*. *Bulletin de la Société Botanique de France, 71*(5), 1084-1090.

Pavillard, J. (1925). Bacillariales. Report on the Danish Oceanographical Expeditions 1908-1910 to the Mediterranean and adjacent seas, vol. 2. *Biology, 9*, 1-72.

Pollard, D. A., Iyer, V. N., Moses, A. M., Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics, 2*(10), e173.

Pritchard, A. (1861). *A history of Infusoria: including the Desmidiaceae and Diatomaceae, British and foreign*. Whittaker, London.

Qiu, Y. L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., ... davis, C. C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences*, *103*(42), 15511-15516.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rambaut, A., Suchard, M. A., Xie, D., Drummond, A. J. (2014). Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer.

Rines, J. E., Hargraves, P. E. (1988). The *Chaetoceros* Ehrenberg (Bacillariophyceae) flora of Narragansett Bay, Rhode Island, USA. *Bibliotheca Phycologica, 79*, 1-196.

Rines, J. E., Boonruang, P., Theriot, E. C. (2000). *Chaetoceros phuketensis* sp. nov. (Bacillariophyceae): a new species from the Andaman Sea. *Phycological Research*, *48*(3), 161-168.

Rines, J. E., Theriot, E. C. (2003). Systematics of Chaetocerotaceae (Bacillariophyceae). I. A phylogenetic analysis of the family. *Phycological Research*, *51*(2), 83-98.

Rokas, A., Krüger, D., Carroll, S. B. (2005). Animal evolution and the molecular signature of radiations compressed in time. *Science, 310*(5756), 1933-1938.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., 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.

Round, F. E., Crawford, R. M., Mann, D. G. (1990). *The diatoms. Biology & morphology of the genera*. Cambridge University Press, Cambridge.

Sarno, D., Zingone, A., Marino, D. (1997). *Bacteriastrum parallelum* sp. nov., a new diatom from the Gulf of Naples, and new observations on *B. furcatum* (Chaetocerotaceae, Bacillariophyta). *Phycologia*, *36*(4), 257-266.

Scholin, C. A., Herzog, M., Sogin, M., Anderson, D. M. (1994). Identification of group- and strain-specific genetic markers from globally distributed *Alexandrium* (Dinophyceae). II. Sequence analysis of fragments of the LSU rRNA gene. *Journal of Phycology* 30.6 (1994): 999-1011.

Shadbolt, G. (1854). A short description of some new forms of Diatomaceae from Port Natal. *Transactions of the Microscopical Society & Journal*, *2*(1), 13-18.

Shevchenko, O. G., Orlova, T. Y., Hernández-Becerril, D. U. (2006). The genus *Chaetoceros* (Bacillariophyta) from Peter the Great Bay, Sea of Japan. *Botanica Marina*, *49*(3), 236-258.

Souffreau, C., Verbruggen, H., Wolfe, A. P., Vanormelingen, P., Siver, P. A., Cox, E. J., ... Vyverman, W. (2011). A time-calibrated multi-gene phylogeny of the diatom genus Pinnularia. *Molecular Phylogenetics and Evolution*, *61*(3), 866-879.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Swofford, D. L. (2002). PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Massachusetts.

Theriot, E. C., Ashworth, M. P., Nakov, T., Ruck, E., Jansen, R. K. (2015). Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution*, *89*, 28-36.

Theriot, E. C., Ashworth, M., Ruck, E., Nakov, T., Jansen, R. K. (2010). A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution*, *143*(3), 278-296.

Van Landingham S. L. (1968). *Catalogue of the fossil and recent genera and species of diatoms and their synonyms. Part II.* Bacteriastrum *through* Coscinodiscus. Cramer, Vaduz.

von Stosch, H. A., Theil, G., Kowallik, K. (1973). Entwicklungsgeschichtliche Untersuchungen an zentrischen Diatomeen. V. Bau und Lebenszyklus von *Chaetoceros didymum*, mit beo-bachtungen über einige andere Arten der Gattung. *Helgoland Wiss Meer, 25*, 384–445.

Xia X. (2017). DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *Journal of Heredity*, *108*(4), 431-437.

Xia, X. Lemey, P. (2009). *Assessing substitution saturation with DAMBE*, in: Lemey, P., Salemi, M., Vandamme, A.M. (Eds.), The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny, 2nd edition. Cambridge University Press, New York.

Xia, X., Z. Xie, M. Salemi, L. Chen, Y. Wang. (2003). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*, *26*(1), 1-7.

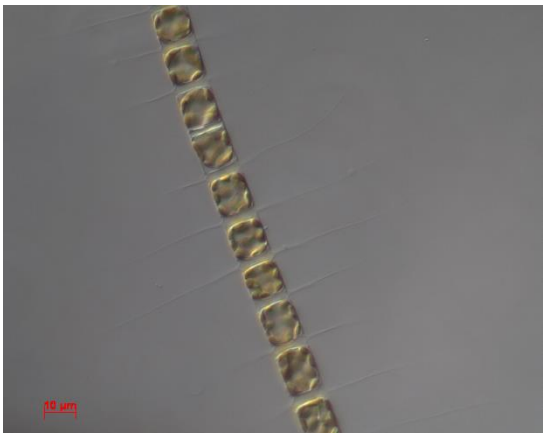
Xu, X. J., Chen, Z. Y., Lundholm, N., Li, Y. (2019). Diversity in the section *Compressa* of the genus *Chaetoceros* (Bacillariophyceae), with description of two new species from Chinese warm waters. *Journal of Phycology*, *55*(1), 104-117.

Yoon, H. S., Hackett, J. D., Bhattacharya, D. (2002). A single origin of the peridinin-and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. *Proceedings of the National Academy of Sciences*, *99*(18), 11724-11729.

Appendix II

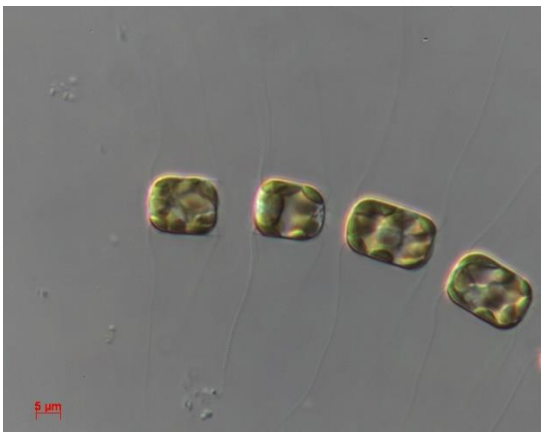
Fig. A2.1. Light microscopy photographies of *Bacteriastrum* and *Chaetoceros* species utilised in the present study. Pictures are from Gaonkar (2017), Gaonkar et al. (2018) and Dr. Wiebe Kooistra lab collection.

Bacteriastrum elegans



B. furcatum 2

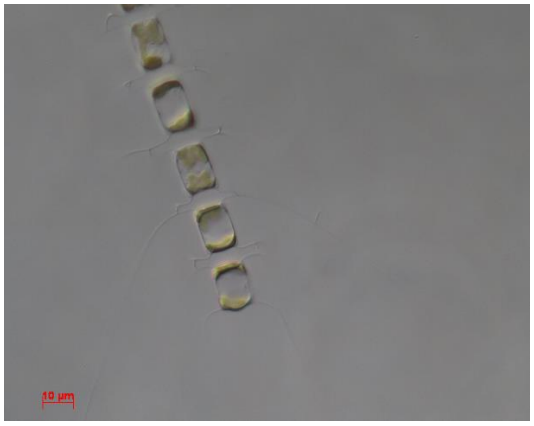
B. hyalinum



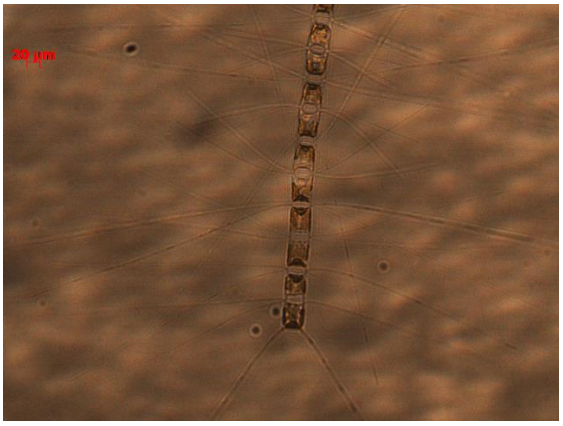
B. jadranum

B. mediterraneum

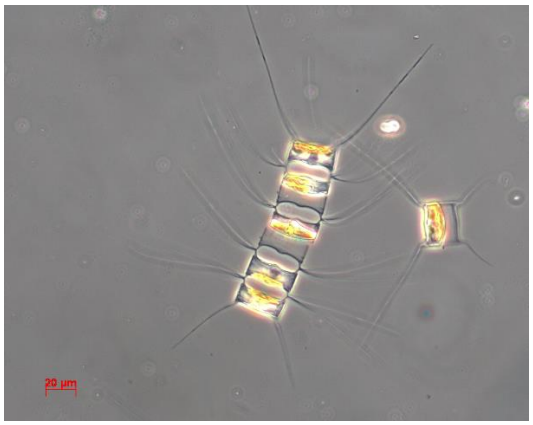
Chaetoceros affinis



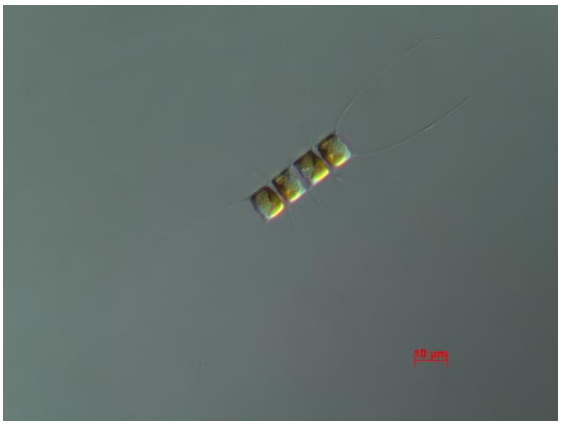
C. anastomosans



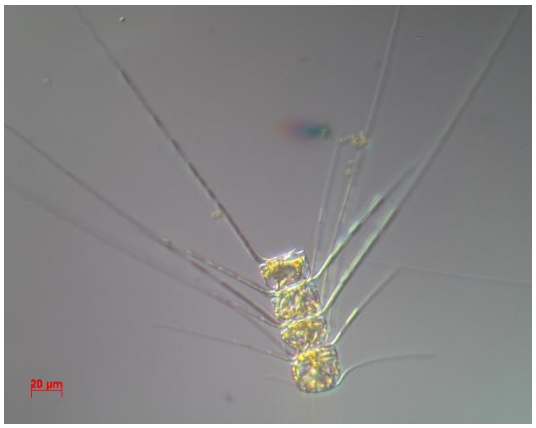
C. brevis 1



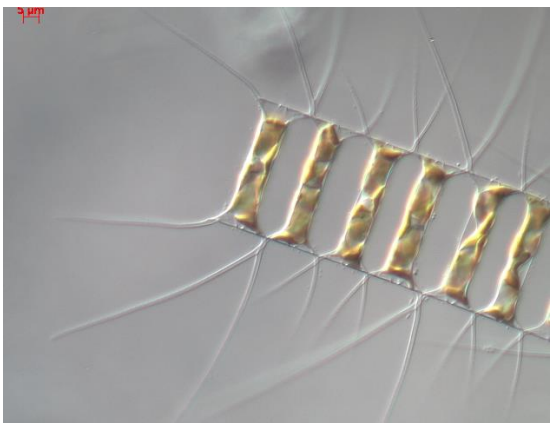
C. brevis 2



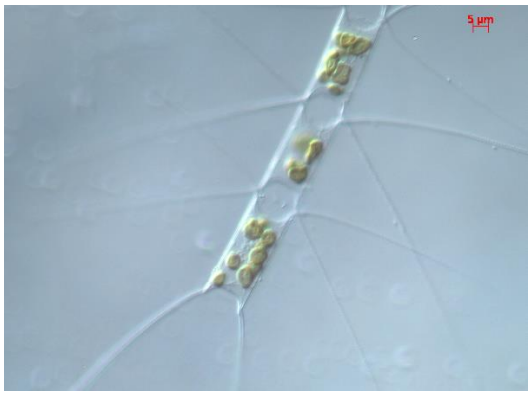
C. brevis 3



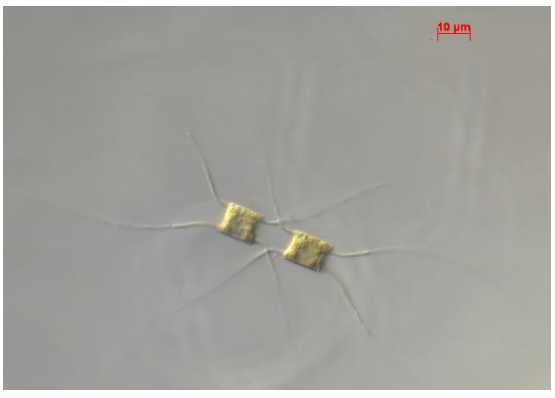
C. cf. *convolutus*



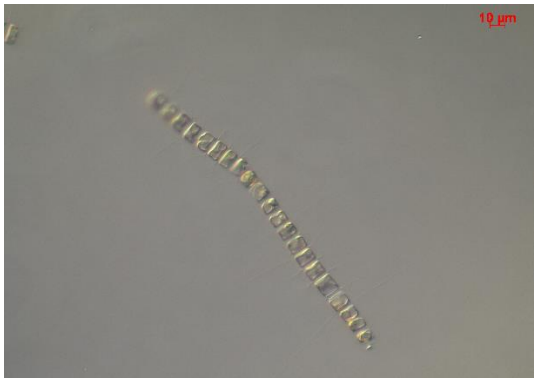
C. cf. *decipiens*



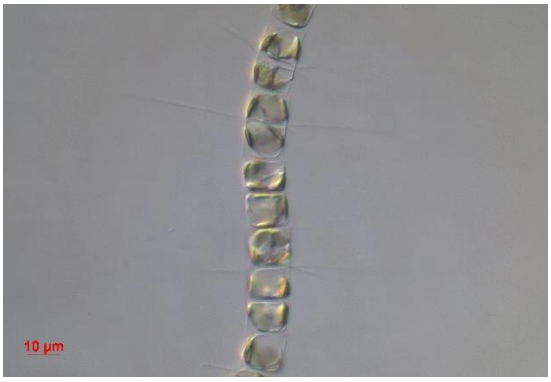
C. cf. *lorenzianus*



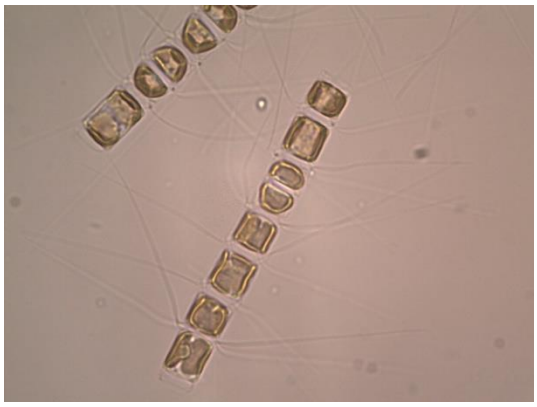
C. cf. *pseudodichaeta*



C. cf. *tortissimus*

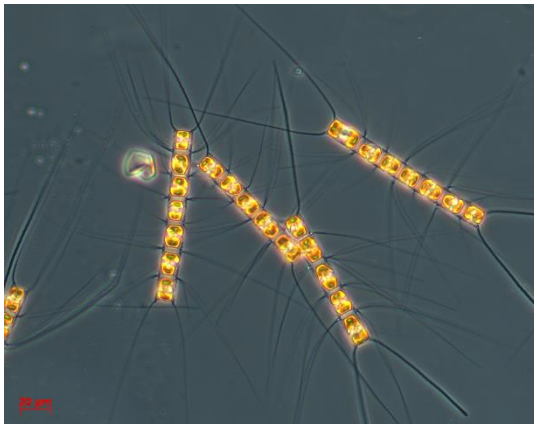


C. cf. *vixvisibilis*

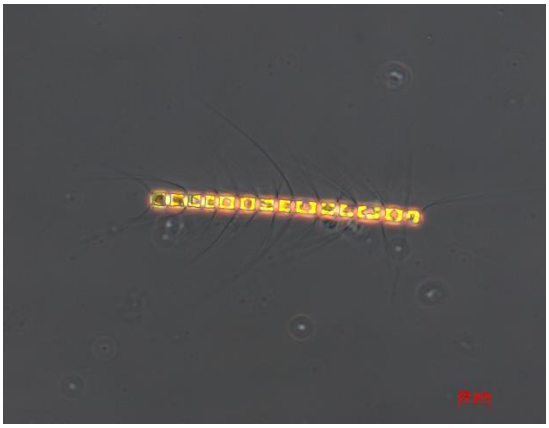


C. cinctus

C. circinalis

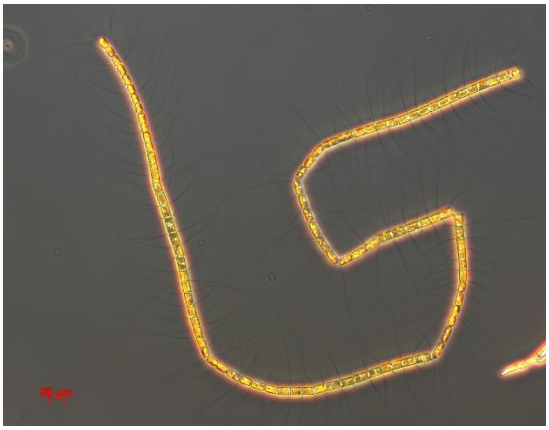


C. constrictus

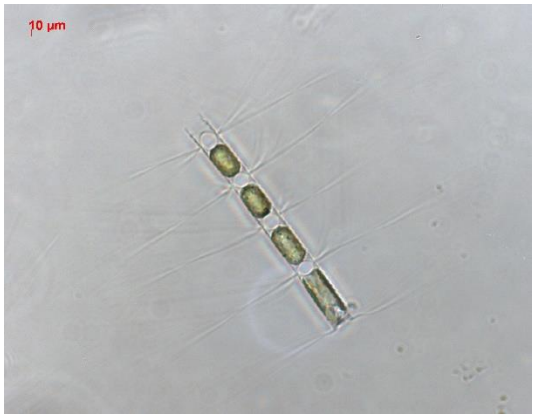


C. contortus

C. contortus cf. var. *contortus*



C. costatus



C. curvisetus 1

C. curvisetus 2

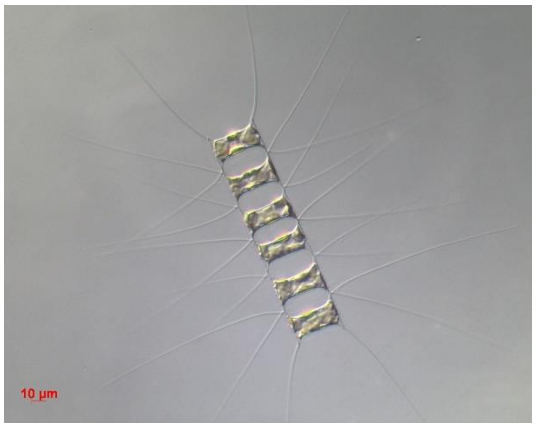
C. curvisetus 2c

C. curvisetus 3

C. curvisetus 3e

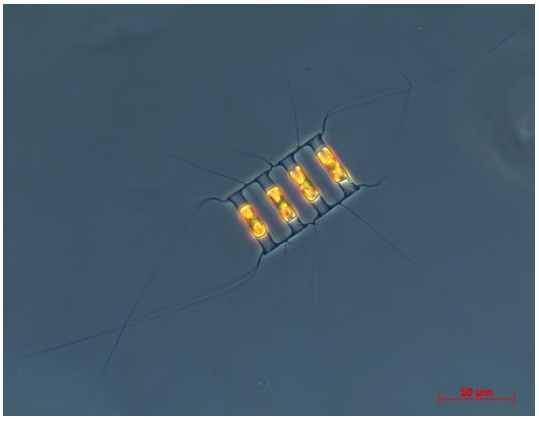
C. danicus

C. debilis 3



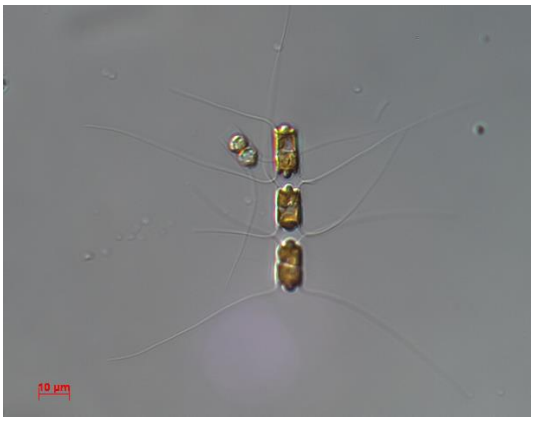
C. decipiens

C. diadema 1

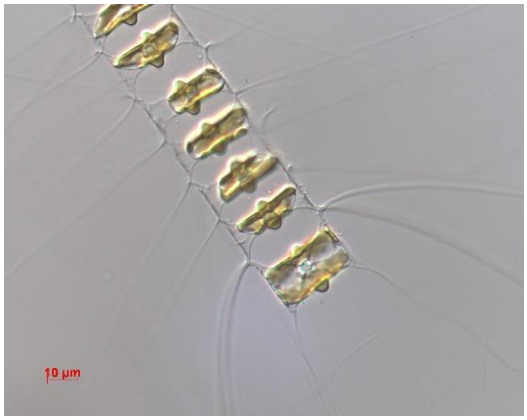


C. diadema 2

C. dichatoensis

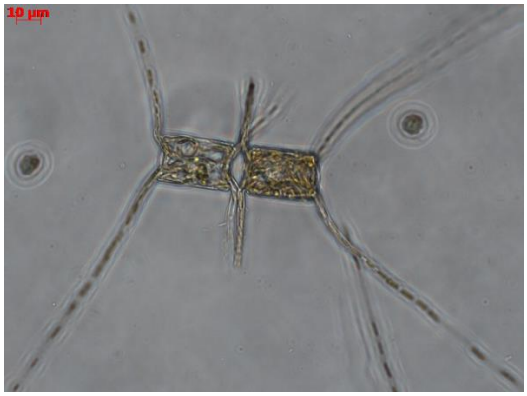


C. didymus 1

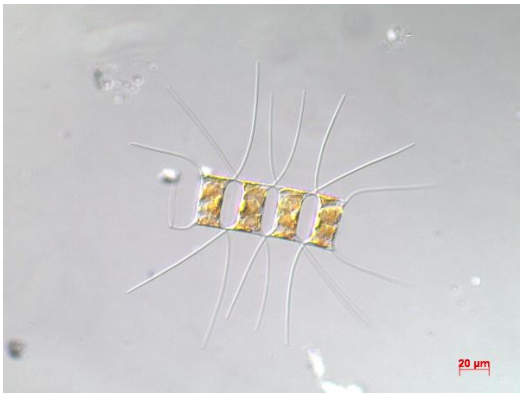


C. didymus 2

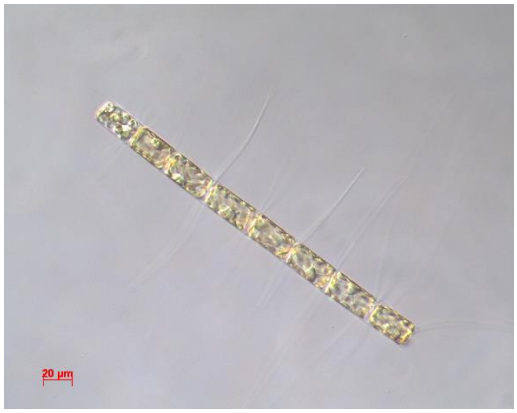
C. diversus 1



C. eibenii

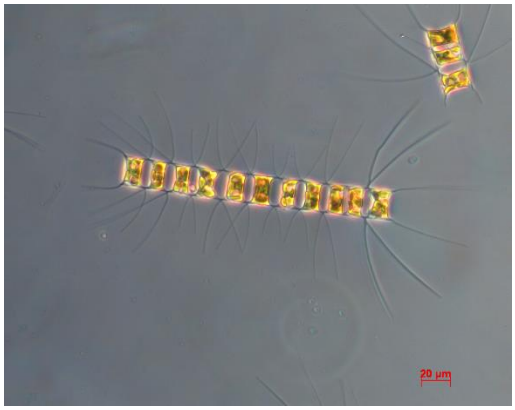


C. elegans



C. lauderi

C. lorenzianus 1

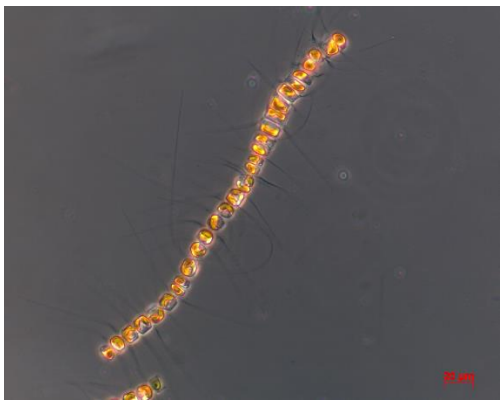


C. lorenzianus 2

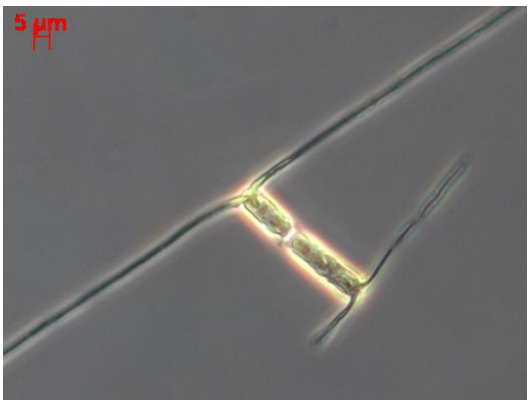
C. peruvianus 1

C. protuberans

C. pseudocurvisetus

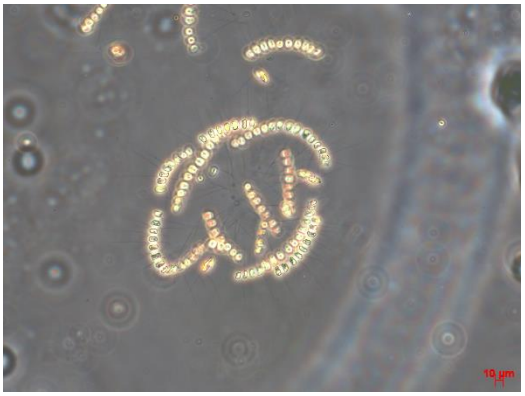


C. radicans



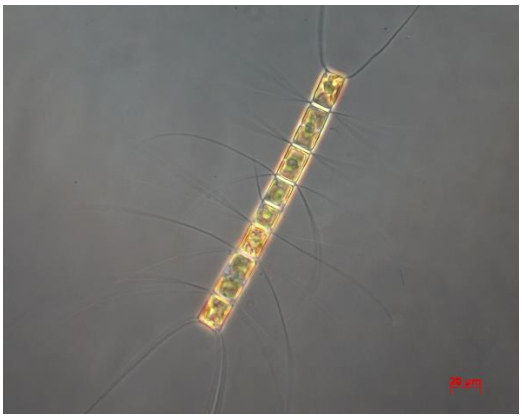
C. rostratus

C. rotosporus

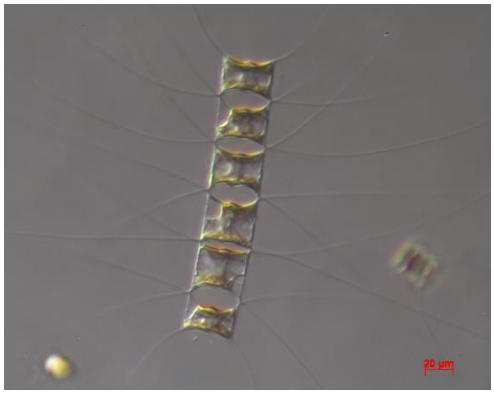


C. socialis

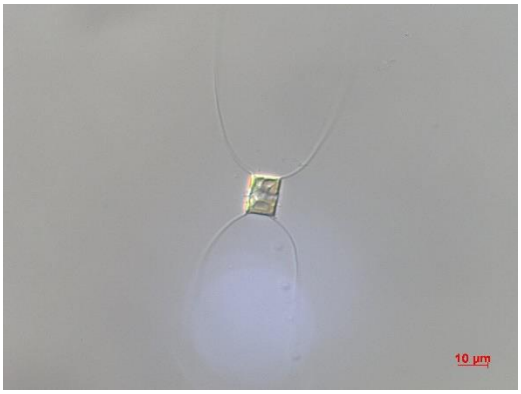
C. sp. Na11C3



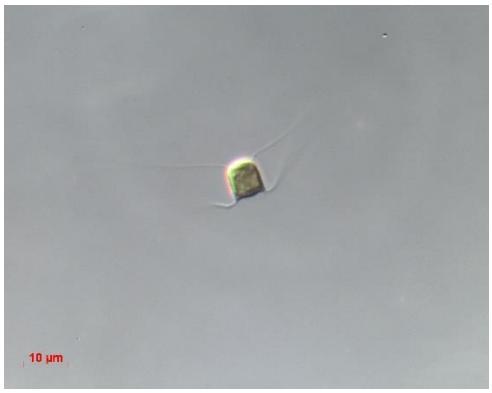
C. sp. Na12A3



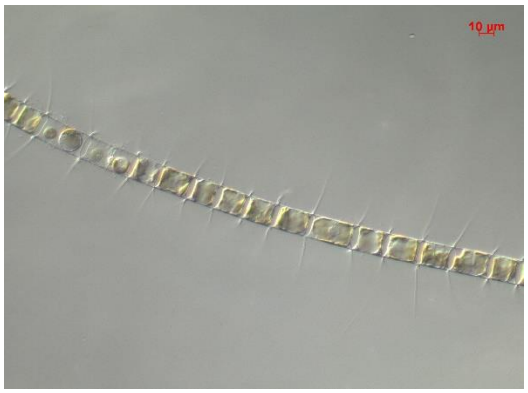
C. sp. Na13C2



C. sp. Na17B2

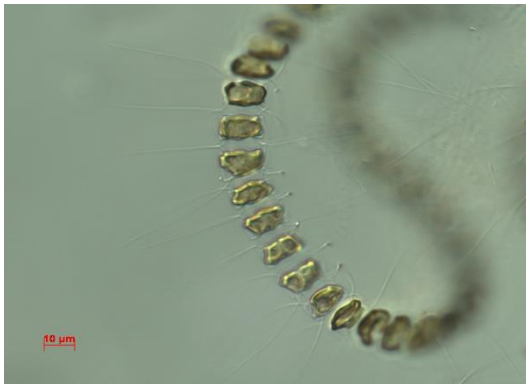


C. sp. Na26B1



C. sp. Na28A1

C. sp. Va7D2



C. sporotruncatus

C. teres

C. throndsenii

C. tortissimus

Fig. A2.2. Maximum Likelihood (ML) tree of concatenated nuclear genes (18S and 28S). Numbers at each node refer to bootstrap support after 1000 replicates.

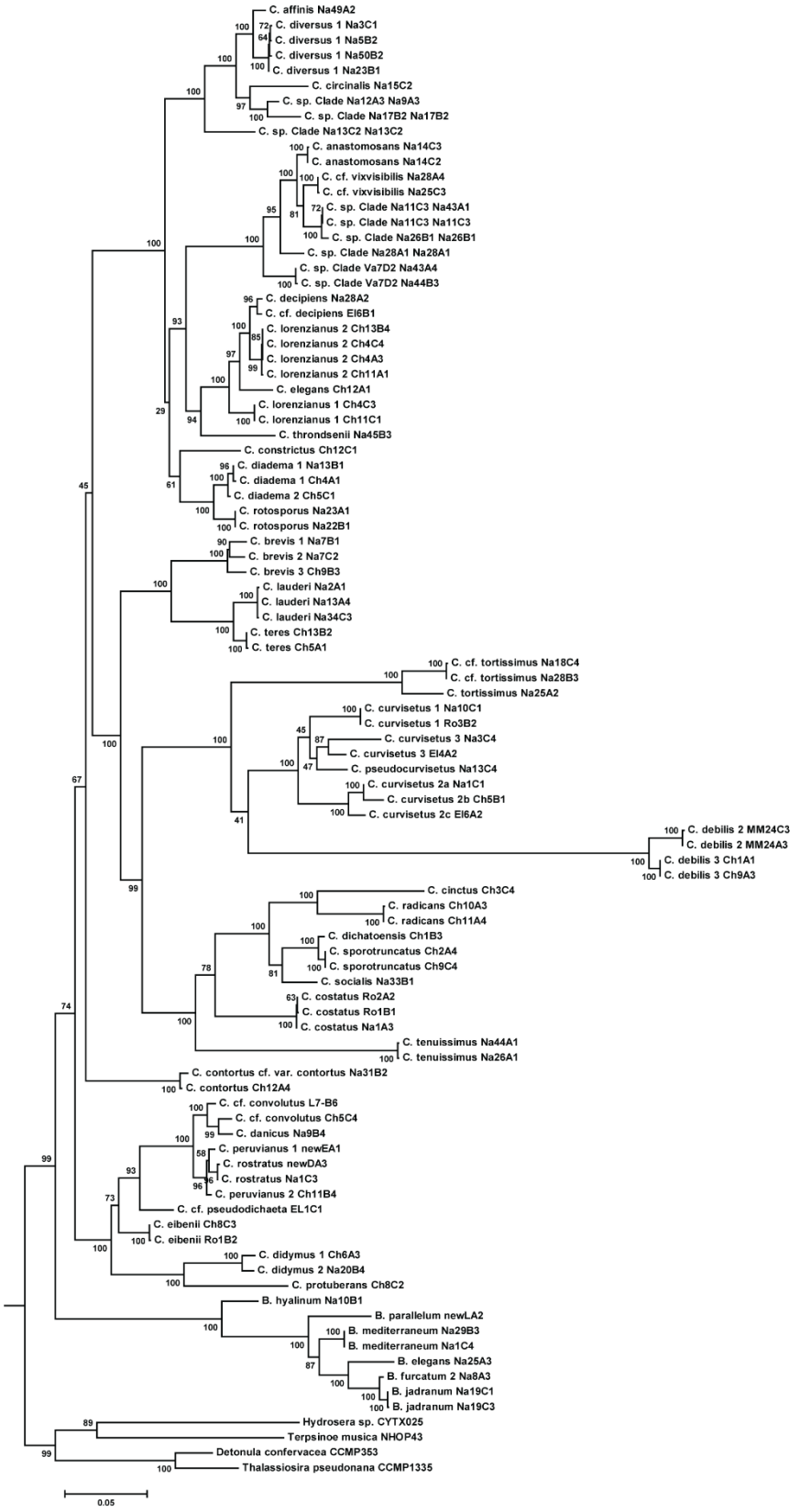


Fig. A2.3. Maximum Likelihood (ML) tree of concatenated plastid genes (*rbc*L and *psb*A). Numbers at each node refer to bootstrap support after 1000 replicates.

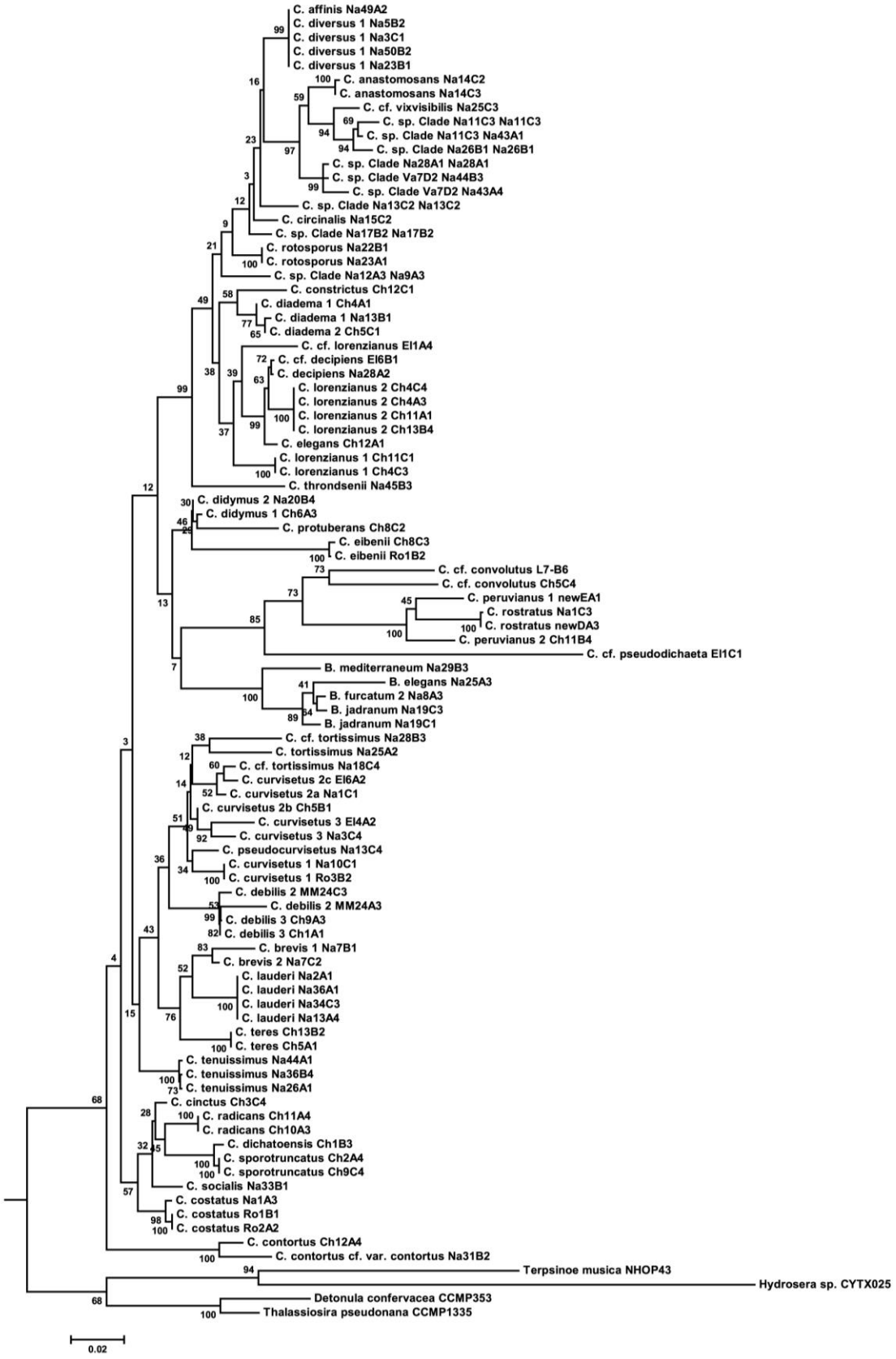


Fig. A2.4. Maximum Likelihood (ML) tree of mitochondrial COI gene. Numbers at each node refer to bootstrap support after 1000 replicates.

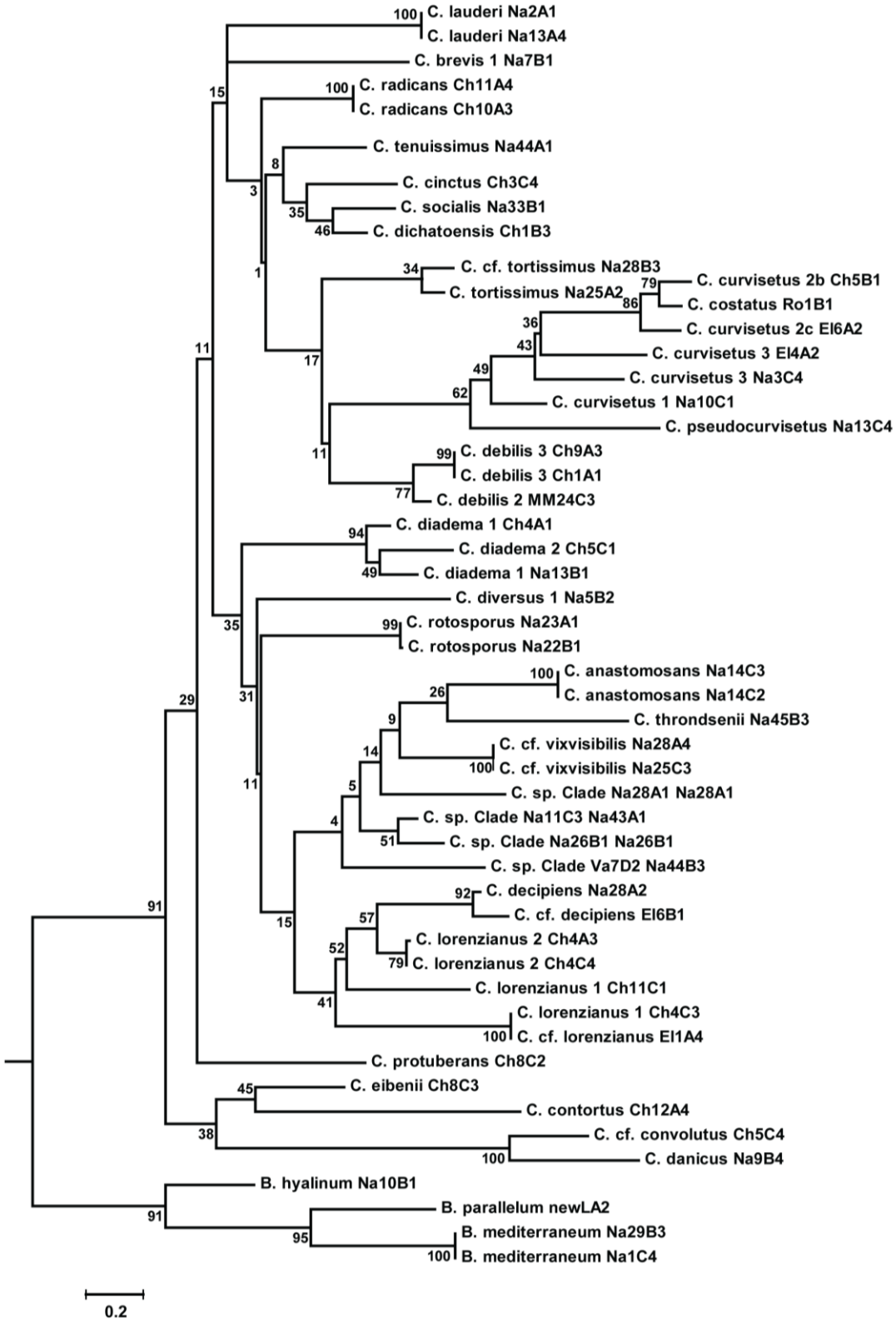


Fig. A2.5. Maximum Parsimony (MP) tree. Numbers at each node refer to bootstrap support after 1000 replicates.

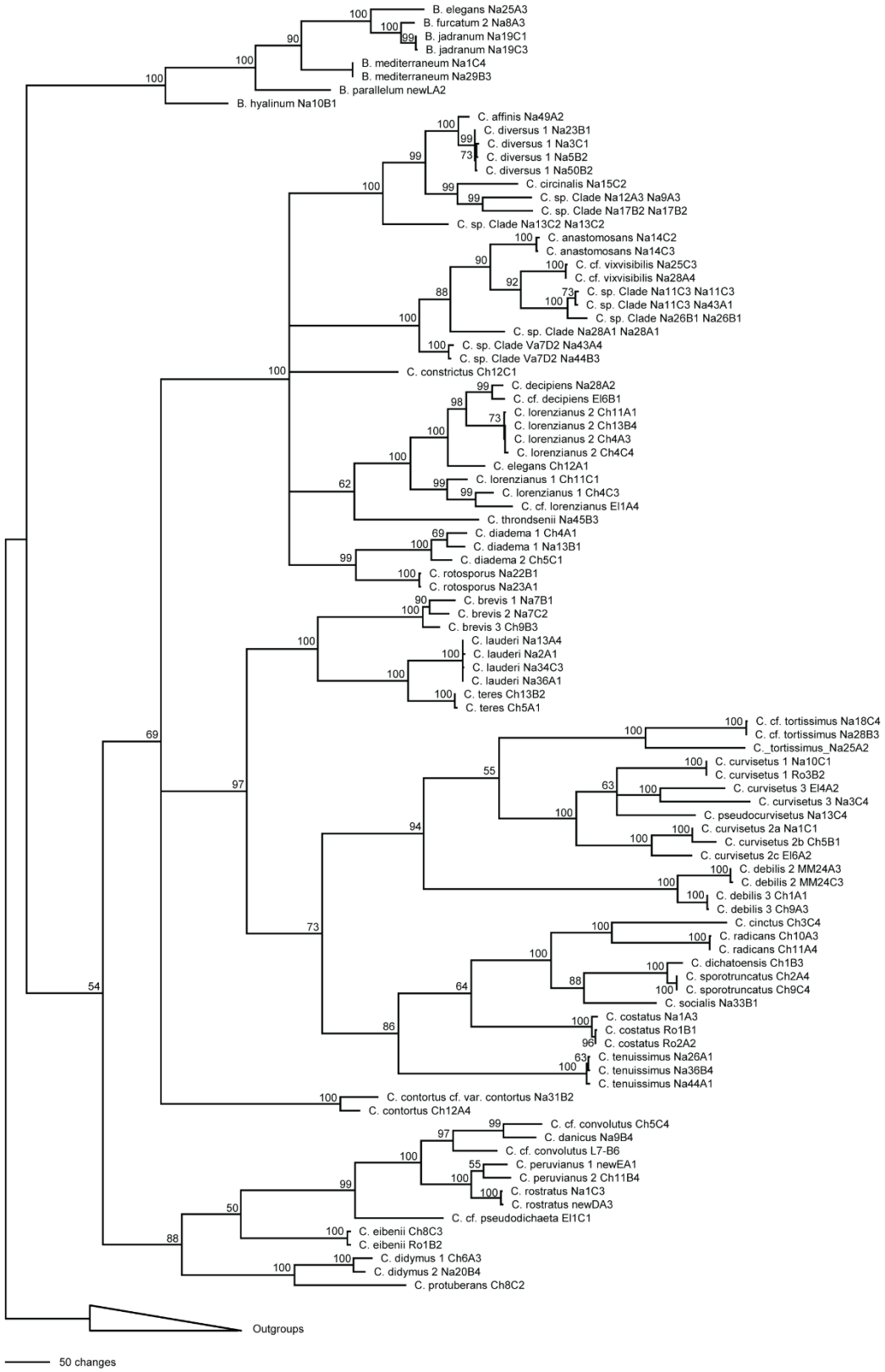


Table A2.1. List of taxa (species and strains) utilised in the present study, including sampling localities and dates and accession numbers for each gene amplified. NA = not available.

Species	Strain	Sampling locality	Sampling date	18S	28S	*rbc*L	*psb*A	COI	Reference
B. elegans	Na25A3	Gulf of Naples (Italy)	07/10/2014	MG972202	MG914436	NA	MK642358	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
B. furcatum 2	Na8A3	Gulf of Naples (Italy)	06/02/2014	MG972354	MG914439	MK642491	MK642359	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
B. hyalinum	Na10B1	Gulf of Naples (Italy)	19/03/2014	MG972205	MG914440	NA	NA	MK642438	Gaonkar et al. 2018 (18S/28S); this study (COI)
B. jadranum	Na19C1	Gulf of Naples (Italy)	30/07/2014	MG972356	MG914441	NA	MK642360	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
B. jadranum	Na19C3	Gulf of Naples (Italy)	30/07/2014	MG972357	MG914442	MK642492	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
B. mediterraneum	Na1C4	Gulf of Naples (Italy)	26/11/2013	MG972206	MG914444	NA	NA	MK642439	Gaonkar et al. 2018 (18S/28S); this study (COI)
B. mediterraneum	Na29B3	Gulf of Naples (Italy)	01/12/2014	NA	MG914446	MK642493	MK642361	MK642440	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
B. parallelum	newLA2	Gulf of Naples (Italy)	21/05/2013	MG972209	MG914447	NA	NA	MK642441	Gaonkar et al. 2018 (18S/28S); this study (COI)
C. affinis	Na49A2	Gulf of Naples (Italy)	26/10/2016	NA	MG914453	NA	MK642362	NA	Gaonkar et al. 2018 (28S); this study (psbA)
C. anastomosans	Na14C2	Gulf of Naples (Italy)	19/03/2014	MG972358	MG914456	MK642494	MK642363	MK642442	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. anastomosans	Na14C3	Gulf of Naples (Italy)	19/03/2014	MG972359	MG914457	MK642495	MK642364	MK642443	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. brevis 1	Na7B1	Gulf of Naples (Italy)	18/01/2014	MG972214	MG914464	MK642496	MK642365	MK642444	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. brevis 2	Na7C2	Gulf of Naples (Italy)	17/01/2014	MG972215	MG914467	NA	MK642366	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
C. brevis 3	Ch9B3	Concepción (Chile)	29/10/2013	MG972216	MG914468	NA	NA	NA	Gaonkar et al. 2018 (18S/28S)
C. cf. *convolutus*	Ch5C4	Las Cruces (Chile)	16/10/2013	MG972226	MG914482	MK642497	MK642367	MK642445	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. cf. *convolutus*	L7-B6	Lohafex experiment	NA	LC466960	NA	MK642498	NA	NA	This study
C. cf. *decipiens*	El6B1	Eilat, Red Sea (Israel)	31/01/2016	NA	LC466963	MK642499	MK642368	MK642446	This study
C. cf. *lorenzianus*	El1A4	Eilat, Red Sea (Israel)	31/01/2016	NA	NA	MK642500	MK642369	MK642447	This study
C. cf. *pseudodichaeta*	El1C1	Eilat, Red Sea (Israel)	31/01/2016	MG972306	MG914586	MK642501	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
C. cf. *tortissimus*	Na18C4	Gulf of Naples (Italy)	01/07/2014	MG972275	MG914640	NA	MK642370	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
C. cf. *tortissimus*	Na28B3	Gulf of Naples (Italy)	07/10/2014	MG972278	MG914643	MK642502	NA	MK642448	Gaonkar et al. 2018 (18S/28S); this study (rbcL/COI)

C. cf. *vixvisibilis*	Na25C3	Gulf of Naples (Italy)	07/10/2014	NA	MG914646	MK642503	MK642371	MK642449	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. cf. *vixvisibilis*	Na28A4	Gulf of Naples (Italy)	07/10/2014	MG972366	MG914648	NA	NA	MK642450	Gaonkar et al. 2018 (18S/28S); this study (COI)
C. cinctus	Ch3C4	Las Cruces (Chile)	16/10/2013	KY852266	KY852282	NA	MK642372	MK642451	Gaonkar et al. 2017 (18S/28S); this study (psbA/COI)
C. circinalis	Na15C2	Gulf of Naples (Italy)	24/04/2014	MG972362	MG914469	MK642504	MK642373	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. constrictus	Ch12C1	Las Cruces (Chile)	04/11/2013	MG972255	MG914471	MK642505	MK642374	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. contortus cf. var. *contortus*	Na31B2	Gulf of Naples (Italy)	07/04/2015	NA	MG914480	MK642506	MK642375	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. contortus	Ch12A4	Las Cruces (Chile)	04/11/2013	MG972222	MG914479	MK642507	MK642376	MK642452	Gaonkar et al. 2018 (18S/28S); this study (rbcL, psbA and COI)
C. costatus	Na1A3	Gulf of Naples (Italy)	26/11/2013	NA	MG914486	MK642509	MK642377	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. costatus	Ro1B1	Roscoff Estacade (France)	11/08/2014	MG972230	MG914490	MK642510	MK642378	MK642454	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. costatus	Ro2A2	Roscoff Estacade (France)	11/08/2014	NA	MG914492	MK642511	MK642379	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. curvisetus 1	Na10C1	Gulf of Naples (Italy)	19/03/2014	MG972232	MG914494	MK642512	MK642380	MK642455	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. curvisetus 1	Ro3B2	Roscoff Estacade (France)	11/08/2014	NA	MG914495	MK642513	MK642381	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. curvisetus 2a	Na1C1	Gulf of Naples (Italy)	26/11/2013	MG972235	MG914499	MK642514	MK642382	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. curvisetus 2b	Ch5B1	Las Cruces (Chile)	16/10/2013	MG972238	MG914506	NA	MK642383	MK642456	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. curvisetus 2c	El6A2	Eilat, Red Sea (Israel)	31/01/2016	LC466961	LC466964	MK642515	MK642384	MK642457	This study
C. curvisetus 3	El4A2	Eilat, Red Sea (Israel)	31/01/2016	LC466962	LC466965	MK642516	MK642385	MK642458	This study
C. curvisetus 3	Na3C4	Gulf of Naples (Italy)	26/11/2013	NA	MG914510	MK642517	MK642386	MK642459	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. danicus	Na9B4	Gulf of Naples (Italy)	19/03/2014	MG972243	MG914513	NA	NA	MK642460	Gaonkar et al. 2018 (18S/28S); this study (COI)
C. debilis 2	MM24-A3	Southern Ocean (Atlantic)	Oct. 2004	MG972247	EF423485	NA	MK642387	NA	Kooistra et al. 2010 (28S); Gaonkar et al. 2018 (18S); this study (psbA)
C. debilis 2	MM24-C3	Southern Ocean (Atlantic)	Oct. 2004	NA	EF423486	MK642518	NA	MK642461	Kooistra et al. 2010 (28S); this study (rbcL/COI)
C. debilis 3	Ch1A1	Las Cruces (Chile)	16/10/2013	MG972248	MG914516	MK642519	MK642388	MK642462	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. debilis 3	Ch9A3	Concepción (Chile)	29/10/2013	NA	MG914519	MK642520	MK642389	MK642463	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. decipiens	Na28A2	Gulf of Naples (Italy)	07/10/2014	NA	KY129900	MK642521	MK642390	MK642464	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. diadema 1	Ch4A1	Las Cruces (Chile)	16/10/2013	MG972254	MG914527	NA	MK642391	MK642465	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. diadema 1	Na13B1	Gulf of Naples (Italy)	19/03/2014	MG972218	MG914529	NA	MK642392	MK642466	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. diadema 2	Ch5C1	Las Cruces (Chile)	16/10/2013	MG97226	MG914534	MK64252	MK64239	MK64246	Gaonkar et al. 2018 (18S/28S); this study

				2		2	3	7	(rbcL/psbA/COI)
C. dichatoensis	Ch1B3	Las Cruces (Chile)	16/10/2013	KY852272	KY852299	MK642523	MK642394	MK642468	Gaonkar et al. 2017 (18S/28S); this study (rbcL/psbA/COI)
C. didymus 1	Ch6A3	Las Cruces (Chile)	16/10/2013	MG972270	MG914537	NA	MK642395	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
C. didymus 2	Na20B4	Gulf of Naples (Italy)	29/07/2014	MG972271	MG914538	NA	MK642396	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
C. diversus 1	Na23B1	Gulf of Naples (Italy)	10/09/2014	NA	MG914545	MK642524	MK642397	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. diversus 1	Na3C1	Gulf of Naples (Italy)	26/11/2013	MG972335	MG914542	MK642525	MK642398	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. diversus 1	Na50B2	Gulf of Naples (Italy)	07/11/2016	NA	MG914546	MK642526	MK642399	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. diversus 1	Na5B2	Gulf of Naples (Italy)	26/11/2013	MG972336	MG914543	MK642527	MK642400	MK642469	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. eibenii	Ch8C3	Concepción (Chile)	29/10/2013	MG972279	MG914547	MK642528	MK642401	MK642470	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. eibenii	Ro1B2	Roscoff Estacade (France)	11/08/2014	MG972280	MG914548	MK642529	MK642402	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. elegans	Ch12A1	Concepción (Chile)	29/10/2013	KX611421	KY129903	MK642530	MK642403	NA	Li et al. 2017 (18S/28S); this study (rbcL/psbA)
C. lauderi	Na13A4	Gulf of Naples (Italy)	19/03/2014	MG972284	MG914553	NA	MK642404	MK642471	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. lauderi	Na2A1	Gulf of Naples (Italy)	26/11/2013	MG972283	MG914552	NA	MK642405	MK642472	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. lauderi	Na34C3	Gulf of Naples (Italy)	28/07/2015	MG972285	MG914554	MK642531	MK642406	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. lauderi	Na36A1	Gulf of Naples (Italy)	26/08/2015	NA	NA	MK642532	MK642407	NA	This study
C. lorenzianus 1	Ch11C1	San Antonio (Chile)	01/11/2013	MG972290	NA	NA	MK642408	MK642473	Gaonkar et al. 2018 (18S); this study (psbA/COI)
C. lorenzianus 1	Ch4C3	Las Cruces (Chile)	16/10/2013	MG972287	MG914557	MK642533	MK642409	MK642474	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. lorenzianus 2	Ch11A1	Las Cruces (Chile)	31/10/2013	NA	MG914567	MK642534	MK642410	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. lorenzianus 2	Ch13B4	Las Cruces (Chile)	05/11/2013	NA	MG914569	MK642535	MK642411	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. lorenzianus 2	Ch4A3	Las Cruces (Chile)	16/10/2013	NA	MG914564	MK642536	MK642412	MK642475	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. lorenzianus 2	Ch4C4	Las Cruces (Chile)	16/10/2013	MG972292	MG914565	NA	MK642413	MK642476	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. peruvianus 1	newEA1	Gulf of Naples (Italy)	28/03/2013	MG972298	MG914573	MK642537	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
C. peruvianus 2	Ch11B4	Las Cruces (Chile)	01/11/2013	MG972296	MG914572	MK642538	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
C. protuberans	Ch8C2	Concepción (Chile)	29/10/2013	MG972299	MG914576	MK642539	MK642414	MK642477	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. pseudocurvisetus	Na13C4	Gulf of Naples (Italy)	19/03/2014	MG972304	MG914584	MK642540	MK642415	MK642478	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. radicans	Ch10A3	Las Cruces (Chile)	29/10/2013	KY852263	KY852291	MK642541	MK642416	MK642479	Gaonkar et al. 2017 (18S/28S); this study (rbcL/psbA/COI)

C. radicans	Ch11A4	Las Cruces (Chile)	01/11/2013	KY852262	KY852292	MK642542	MK642417	MK642480	Gaonkar et al. 2017 (18S/28S); this study (rbcL/psbA/COI)
C. rostratus	Na1C3	Gulf of Naples (Italy)	26/11/2013	MG972307	MG914588	MK642543	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
C. rostratus	newDA3	Gulf of Naples (Italy)	28/03/2013	MG972310	MG914591	MK642544	NA	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL)
C. rotosporus	Na22B1	Gulf of Naples (Italy)	10/09/2014	MG972350	MG914595	MK642545	MK642418	MK642481	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. rotosporus	Na23A1	Gulf of Naples (Italy)	10/09/2014	NA	MG914597	MK642546	MK642419	MK642482	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. socialis	Na33B1	Gulf of Naples (Italy)	14/07/2015	NA	KY852295	MK642547	MK642420	MK642483	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. sp. Clade Na11C3	Na11C3	Gulf of Naples (Italy)	19/03/2014	MG972328	MG914605	NA	MK642421	NA	Gaonkar et al. 2018 (18S/28S); this study (psbA)
C. sp. Clade Na11C3	Na43A1	Gulf of Naples (Italy)	15/03/2016	NA	MG914609	MK642548	MK642422	MK642484	Gaonkar et al. 2018 (28S); this study (rbcL/psbA/COI)
C. sp. Clade Na12A3	Na9A3	Gulf of Naples (Italy)	19/03/2014	NA	MG921671	MK642549	MK642423	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. sp. Clade Na13C2	Na13C2	Gulf of Naples (Italy)	19/03/2014	MG972344	MG921675	MK642550	MK642424	NA	This study
C. sp. Clade Na17B2	Na17B2	Gulf of Naples (Italy)	01/07/2014	MG972334	MG921677	MK642551	MK642425	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. sp. Clade Na26B1	Na26B1	Gulf of Naples (Italy)	07/10/2014	MG972329	MG914606	NA	MK642426	MK642485	Gaonkar et al. 2018 (18S/28S); this study (psbA/COI)
C. sp. Clade Na28A1	Na28A1	Gulf of Naples (Italy)	07/10/2014	MG972364	MG921679	MK642552	MK642427	MK642486	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. sp. Clade Va7D2	Na43A4	Gulf of Naples (Italy)	15/03/2016	NA	MG921681	NA	MK642428	NA	Gaonkar et al. 2018 (28S); this study (psbA)
C. sp. Clade Va7D2	Na44B3	Gulf of Naples (Italy)	15/03/2016	NA	MG921684	NA	MK642429	MK642487	Gaonkar et al. 2018 (28S); this study (psbA/COI)
C. sporotruncatus	Ch2A4	Las Cruces (Chile)	16/10/2013	KY852270	KY852297	MK642553	MK642430	NA	Gaonkar et al. 2017 (18S/28S); this study (rbcL/psbA)
C. sporotruncatus	Ch9C4	Concepción (Chile)	29/10/2013	NA	KY852298	MK642554	NA	NA	Gaonkar et al. 2017 (28S); this study (rbcL)
C. tenuissimus	Na26A1	Gulf of Naples (Italy)	07/10/2014	MG972314	MG914614	MK642555	MK642431	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. tenuissimus	Na36B4	Gulf of Naples (Italy)	26/08/2015	NA	NA	MK642556	MK642432	NA	This study
C. tenuissimus	Na44A1	Gulf of Naples (Italy)	31/05/2016	MG972315	MG914615	MK642557	MK642433	MK642488	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. teres	Ch13B2	Las Cruces (Chile)	05/11/2013	NA	MG914630	MK642558	MK642434	NA	Gaonkar et al. 2018 (28S); this study (rbcL/psbA)
C. teres	Ch5A1	Las Cruces (Chile)	16/10/2013	MG972317	MG914626	MK642559	MK642435	NA	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA)
C. throndsenii	Na45B3	Gulf of Naples (Italy)	31/05/2016	MG972323	MG914633	MK642560	MK642436	MK642489	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
C. tortissimus	Na25A2	Gulf of Naples (Italy)	07/10/2014	MG972325	MG914635	MK642561	MK642437	MK642490	Gaonkar et al. 2018 (18S/28S); this study (rbcL/psbA/COI)
Detonula confervacea	CCMP353	Culture collection	NA	HQ912617	NA	HQ912481	KM009482	NA	Theriot et al. 2010 (18S, rbcL); 2015 (psbA)
Hydrosera sp.	CYTX025	NA	NA	HQ912683	NA	HQ912547	NA	NA	Theriot et al. 2010

Terpsinoe musica	NHOP43	NA	NA	HQ912682	NA	HQ912546	KM009571	NA	Theriot et al. 2010 (18S, rbcL); 2015 (psbA)
Thalassiosira pseudonana	CCMP1335	Culture collection	NA	HQ912555	NA	HQ912419	KM009420	NA	Theriot et al. 2010 (18S, rbcL); 2015 (psbA)

Table A2.2. Tests of substitution saturation for *rbc*L (a), *psb*A (b) and COI (c) genes.

Analyses were performed on all sites.

(a)

N OTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.266	0.791	23.896	397	0.000	0.758	22.400	397	0.0000
8	0.273	0.746	20.824	397	0.000	0.634	15.912	397	0.0000
16	0.271	0.709	19.567	397	0.000	0.500	10.218	397	0.0000
32	0.279	0.695	18.871	397	0.000	0.368	4.007	397	0.0001

Note: two-tailed tests are used.

(b)

N OTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.171	0.781	24.153	246	0.0000	0.756	23.139	246	0.0000
8	0.186	0.734	18.902	246	0.0000	0.626	15.164	246	0.0000
16	0.187	0.680	16.850	246	0.0000	0.474	9.808	246	0.0000
32	0.203	0.683	15.443	246	0.0000	0.354	4.851	246	0.0000

Note: two-tailed tests are used.

(c)

N OTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.598	0.785	4.129	130	0.0001	0.801	4.479	130	0.0000
8	0.594	0.757	3.751	130	0.0003	0.681	1.987	130	0.0491
16	0.602	0.610	0.197	130	0.8438	0.459	3.447	130	0.0008
32	0.597	0.735	3.504	130	0.0006	0.460	3.462	130	0.0007

Note: two-tailed tests are used.

Table A2.3. Chi-squared test of homogeneity of state frequencies across taxa.

Taxon		A	C	G	T
B. elegans Na25A3	O	894.00	635.00	781.00	982.00
	E	883.39	625.82	777.39	1005.40
B. furcatum 2 Na8A3	O	1239.00	863.00	1059.00	1374.00
	E	1216.94	862.11	1070.92	1385.03
B. hyalinum Na10B1	O	769.00	507.00	716.00	845.00
	E	761.29	539.32	669.95	866.44
B. jadranum Na19C1	O	852.00	604.00	754.00	938.00
	E	844.74	598.44	743.39	961.42
B. jadranum Na19C3	O	1005.00	633.00	851.00	1043.00
	E	947.79	671.44	834.07	1078.70
B. mediterraneum Na1C4	O	763.00	512.00	699.00	843.00
	E	755.92	535.52	665.22	860.33
B. mediterraneum Na29B3	O	1366.00	933.00	1153.00	1560.00
	E	1344.94	952.79	1183.57	1530.71
B. parallelum newLA2	O	772.00	512.00	700.00	838.00
	E	757.26	536.47	666.40	861.86
C. affinis Na49A2	O	668.00	510.00	636.00	761.00
	E	690.98	489.51	608.08	786.43
C. anastomosans Na14C2	O	1331.00	942.00	1147.00	1528.00
	E	1327.76	940.63	1168.45	1511.16
C. anastomosans Na14C3	O	1348.00	954.00	1160.00	1546.00
	E	1343.86	952.03	1182.62	1529.48
C. brevis 1 Na7B1	O	1344.00	974.00	1210.00	1570.00
	E	1368.01	969.14	1203.87	1556.97
C. brevis 2 Na7C2	O	848.00	674.00	845.00	957.00
	E	891.97	631.90	784.95	1015.18
C. brevis 3 Ch9B3	O	644.00	476.00	666.00	667.00
	E	658.25	466.32	579.27	749.17
C. cf. *convolutus* Ch5C4	O	1382.00	931.00	1199.00	1573.00
	E	1364.53	966.67	1200.80	1553.00
C. cf. *convolutus* L7-B6	O	836.00	532.00	719.00	923.00
	E	807.71	572.21	710.80	919.28
C. cf. *decipiens* El6B1	O	1013.00	629.00	887.00	1072.00

	E	966.30	684.56	850.36	1099.77
C. cf. *lorenzianus* El1A4	O	910.00	656.00	809.00	995.00
	E	904.32	640.64	795.81	1029.23
C. cf. *pseudodichaeta* El1C1	O	1153.00	744.00	989.00	1273.00
	E	1116.04	790.64	982.13	1270.19
C. cf. *tortissimus* Na18C4	O	1349.00	940.00	1163.00	1555.00
	E	1343.59	951.84	1182.38	1529.18
C. cf. *tortissimus* Na28B3	O	748.00	507.00	699.00	816.00
	E	743.31	526.58	654.13	845.98
C. cf. *vixvisibilis* Na25C3	O	1023.00	718.00	878.00	1167.00
	E	1015.95	719.73	894.05	1156.27
C. cf. *vixvisibilis* Na28A4	O	1172.00	840.00	1024.00	1286.00
	E	1159.78	821.62	1020.62	1319.97
C. cinctus Ch3C4	O	1252.00	885.00	1096.00	1397.00
	E	1242.43	880.17	1093.36	1414.04
C. circinalis Na15C2	O	1259.00	902.00	1120.00	1411.00
	E	1259.07	891.96	1108.00	1432.97
C. constrictus Ch12C1	O	1357.00	975.00	1202.00	1562.00
	E	1367.48	968.76	1203.40	1556.36
C. contortus cf. var. *contortus* Na31B2	O	1249.00	928.00	1109.00	1402.00
	E	1257.99	891.20	1107.05	1431.75
C. contortus Ch12A4	O	1353.00	986.00	1189.00	1565.00
	E	1366.67	968.19	1202.69	1555.44
C. costatus Na1A3	O	1347.00	969.00	1184.00	1579.00
	E	1362.92	965.53	1199.39	1551.17
C. costatus Ro1B1	O	1248.00	912.00	1104.00	1410.00
	E	1254.24	888.54	1103.75	1427.48
C. costatus Ro2A2	O	1247.00	900.00	1114.00	1415.00
	E	1254.77	888.92	1104.22	1428.09
C. curvisetus 1 Na10C1	O	968.00	710.00	894.00	1137.00
	E	995.29	705.09	875.87	1132.76
C. curvisetus 1 Ro3B2	O	1351.00	963.00	1186.00	1565.00
	E	1359.16	962.87	1196.08	1546.89
C. curvisetus 2a Na1C1	O	1317.00	945.00	1166.00	1560.00
	E	1338.50	948.23	1177.90	1523.38
C. curvisetus 2b Ch5B1	O	1351.00	966.00	1181.00	1581.00

	E	1362.92	965.53	1199.39	1551.17
C. curvisetus 2c El6A2	O	775.00	506.00	729.00	845.00
	E	766.12	542.74	674.20	871.94
C. curvisetus 3 El4A2	O	901.00	644.00	791.00	1019.00
	E	900.29	637.79	792.27	1024.64
C. curvisetus 3 Na3C4	O	1153.00	744.00	959.00	1287.00
	E	1111.75	787.59	978.35	1265.31
C. danicus Na9B4	O	1376.00	952.00	1147.00	1593.00
	E	1359.96	963.44	1196.79	1547.81
C. debilis 2 MM24-A3	O	1377.00	952.00	1146.00	1593.00
	E	1359.96	963.44	1196.79	1547.81
C. debilis 2 MM24-C3	O	1334.00	943.00	1162.00	1562.00
	E	1341.98	950.70	1180.97	1527.35
C. debilis 3 Ch1A1	O	969.00	732.00	912.00	1112.00
	E	999.58	708.13	879.65	1137.64
C. debilis 3 Ch9A3	O	965.00	728.00	913.00	1119.00
	E	999.58	708.13	879.65	1137.64
C. decipiens Na28A2	O	1349.00	983.00	1204.00	1560.00
	E	1367.48	968.76	1203.40	1556.36
C. diadema 1 Ch4A1	O	1372.00	968.00	1170.00	1587.00
	E	1367.75	968.95	1203.64	1556.66
C. diadema 1 Na13B1	O	897.00	662.00	819.00	1005.00
	E	907.81	643.12	798.88	1033.20
C. diadema 2 Ch5C1	O	898.00	657.00	821.00	1009.00
	E	908.34	643.50	799.36	1033.81
C. dichatoensis Ch1B3	O	1239.00	897.00	1113.00	1383.00
	E	1242.97	880.55	1093.83	1414.65
C. didymus 1 Ch6A3	O	1240.00	895.00	1111.00	1378.00
	E	1240.82	879.03	1091.94	1412.21
C. didymus 2 Na20B4	O	1337.00	957.00	1188.00	1547.00
	E	1349.50	956.02	1187.58	1535.90
C. diversus 1 Na23B1	O	1356.00	951.00	1207.00	1585.00
	E	1368.28	969.33	1204.11	1557.28
C. diversus 1 Na3C1	O	1258.00	881.00	1127.00	1425.00
	E	1258.80	891.77	1107.76	1432.67
C. diversus 1 Na50B2	O	1262.00	884.00	1081.00	1411.00

	E	1244.58	881.69	1095.25	1416.48
C. diversus 1 Na5B2	O	943.00	738.00	951.00	1151.00
	E	1015.14	719.16	893.34	1155.36
C. eibenii Ch8C3	O	944.00	738.00	949.00	1152.00
	E	1015.14	719.16	893.34	1155.36
C. eibenii Ro1B2	O	1229.00	915.00	1143.00	1399.00
	E	1257.46	890.82	1106.58	1431.14
C. elegans Ch12A1	O	967.00	723.00	905.00	1140.00
	E	1002.26	710.03	882.01	1140.70
C. lauderi Na13A4	O	1348.00	950.00	1182.00	1566.00
	E	1354.06	959.26	1191.59	1541.09
C. lauderi Na2A1	O	1231.00	869.00	1080.00	1398.00
	E	1228.48	870.29	1081.08	1398.16
C. lauderi Na34C3	O	1229.00	865.00	1078.00	1396.00
	E	1225.79	868.39	1078.72	1395.10
C. lauderi Na36A1	O	1328.00	933.00	1155.00	1567.00
	E	1337.15	947.28	1176.72	1521.85
C. lorenzianus 1 Ch11C1	O	948.00	706.00	884.00	1134.00
	E	985.36	698.06	867.13	1121.46
C. lorenzianus 1 Ch4C3	O	846.00	545.00	715.00	898.00
	E	806.10	571.07	709.38	917.45
C. lorenzianus 2 Ch11A1	O	1050.00	680.00	935.00	1104.00
	E	1011.39	716.50	890.04	1151.08
C. lorenzianus 2 Ch13B4	O	1372.00	957.00	1181.00	1584.00
	E	1366.94	968.38	1202.93	1555.75
C. lorenzianus 2 Ch4A3	O	1343.00	974.00	1192.00	1568.00
	E	1362.38	965.15	1198.91	1550.56
C. lorenzianus 2 Ch4C4	O	1368.00	972.00	1157.00	1592.00
	E	1365.60	967.43	1201.75	1554.22
C. peruvianus 1 newEA1	O	1368.00	974.00	1161.00	1592.00
	E	1367.21	968.57	1203.17	1556.05
C. peruvianus 2 Ch11B4	O	1036.00	669.00	933.00	1115.00
	E	1007.09	713.45	886.26	1146.20
C. protuberans Ch8C2	O	1039.00	667.00	933.00	1116.00
	E	1007.63	713.83	886.73	1146.81
C. pseudocurvisetus Na13C4	O	1308.00	939.00	1161.00	1503.00

	E	1317.83	933.59	1159.71	1499.86
C. radicans Ch10A3	O	1313.00	940.00	1165.00	1505.00
	E	1321.05	935.87	1162.55	1503.52
C. radicans Ch11A4	O	1354.00	974.00	1188.00	1580.00
	E	1367.48	968.76	1203.40	1556.36
C. rostratus Na1C3	O	862.00	638.00	789.00	943.00
	E	867.29	614.41	763.22	987.08
C. rostratus newDA3	O	1168.00	819.00	962.00	1379.00
	E	1161.39	822.76	1022.04	1321.81
C. rotosporus Na22B1	O	1247.00	903.00	1132.00	1405.00
	E	1257.72	891.01	1106.82	1431.45
C. rotosporus Na23A1	O	1207.00	882.00	1087.00	1364.00
	E	1218.28	863.06	1072.10	1386.55
C. socialis Na33B1	O	1228.00	899.00	1115.00	1383.00
	E	1241.09	879.22	1092.18	1412.51
C. sp. Clade Na11C3 Na11C3	O	977.00	712.00	866.00	1121.00
	E	986.43	698.82	868.07	1122.68
C. sp. Clade Na11C3 Na43A1	O	1352.00	952.00	1134.00	1520.00
	E	1330.45	942.53	1170.81	1514.21
C. sp. Clade Na12A3 Na9A3	O	534.00	395.00	477.00	609.00
	E	540.71	383.06	475.83	615.40
C. sp. Clade Na13C2 Na13C2	O	636.00	459.00	549.00	773.00
	E	648.59	459.48	570.77	738.17
C. sp. Clade Na17B2 Na17B2	O	1271.00	904.00	1094.00	1423.00
	E	1259.07	891.96	1108.00	1432.97
C. sp. Clade Na26B1 Na26B1	O	1048.00	704.00	911.00	1102.00
	E	1010.31	715.74	889.09	1149.86
C. sp. Clade Na28A1 Na28A1	O	1262.00	903.00	1095.00	1410.00
	E	1253.16	887.78	1102.80	1426.26
C. sp. Clade Va7D2 Na43A4	O	1369.00	980.00	1175.00	1551.00
	E	1361.84	964.77	1198.44	1549.95
C. sp. Clade Va7D2 Na44B3	O	1224.00	904.00	1122.00	1384.00
	E	1243.50	880.93	1094.30	1415.26
C. sporotruncatus Ch2A4	O	1238.00	917.00	1132.00	1401.00
	E	1257.99	891.20	1107.05	1431.75
C. sporotruncatus Ch9C4	O	1308.00	896.00	1136.00	1494.00

	E	1297.17	918.95	1141.53	1476.34
C. tenuissimus Na26A1	O	1389.00	959.00	1157.00	1577.00
	E	1363.72	966.10	1200.10	1552.08
C. tenuissimus Na36B4	O	1043.00	740.00	871.00	1226.00
	E	1041.17	737.60	916.25	1184.98
C. tenuissimus Na44A1	O	833.00	527.00	697.00	846.00
	E	779.00	551.87	685.53	886.60
C. teres Ch13B2	O	1048.00	707.00	839.00	1173.00
	E	1010.85	716.12	889.56	1150.47
C. teres Ch5A1	O	1031.00	740.00	870.00	1187.00
	E	1027.22	727.71	903.97	1169.10
C. throndsenii Na45B3	O	889.00	630.00	743.00	1093.00
	E	900.29	637.79	792.27	1024.64
C. tortissimus Na25A2	O	798.00	593.00	674.00	930.00
	E	803.69	569.36	707.26	914.70
Detonula confervacea CCMP353	O	800.00	580.00	682.00	914.00
	E	798.59	565.74	702.77	908.89
Hydrosera sp. CYTX025	O	707.00	490.00	542.00	904.00
	E	709.23	502.44	624.13	807.19
Terpsinoe musica NHOP43	O	602.00	443.00	455.00	738.00
	E	600.55	425.45	528.50	683.50
Thalassiosira pseudonana CCMP1335	O	602.00	445.00	459.00	732.00
	E	600.55	425.45	528.50	683.50

Chi-square = 316.520842 (df=297), P = 0.20860911

Warning: This test ignores correlation due to phylogenetic structure.

Table A2.4. Traditional classification scheme for the family Chaetocerotaceae. Only sections including taxa utilised in the present study are shown. "References for description" refers to publications in which the section is described or amended. "Morphologically assigned species" refers to taxa assigned to the sections using information in Gaonkar et al. (2018) and references therein).

Genus *Bacteriastrum* Shadbolt	
Section	
Isomorpha Pavillard	**Description:** terminal setae of like construction and form on both ends of chain (isomorphic). Setae on both ends directed either outward from chain axis or toward the center. The two outer valves are therefore mirror images. **References for description:** Pavillard (1924); (1925); Cupp (1943). **Morphologically assigned species:** *B. hyalinum*, *B. jadranum*.
Sagittata Pavillard	**Description:** terminal setae on either end of chain different in form and direction (dimorphic). Setae of posterior valve directed outward from chain and running nearly parallel to chain axis, forming a bell-shaped space. Setae of other or anterior valve curved toward inner part of chain, or on their ends turned back toward the outside or in general deviating little from the valvar plane. **References for description:** Pavillard (1924); (1925); Cupp (1943). **Morphologically assigned species:** *B. elegans*, *B. furcatum* 2, *B. mediterraneum*, *B. parallelum.*
Genus *Chaetoceros* Ehrenberg	
Subgenus *Chaetoceros* (*Phaeoceros*) Gran	
Section	
Borealia Ostenfeld	**Description:** setae diverging in all directions; the directions of the setae of the one valve are often different from those of the other valve; the external process of the rimoportula in the centre of the valve absent. Apertures narrow. **References for description:** Ostenfeld (1903); Cupp (1943). **Morphologically assigned species:** *C.* cf. *convolutus*, *C. danicus*, *C. eibenii*.
Peruviana Hernández-Becerril	**Description:** cells solitary or in chains, heterovalvar. All setae robust, pointed towards the same end. Rimoportula present in every valve, excentrically placed. **Reference for description:** Hernández-Becerril (1996). **Morphologically assigned species:** *C. peruvianus* 1-2.
Rostrata Hernández-Becerril	**Description:** cells in chains, united by a linking central process. No apertures. Setae robust, with no fusion between sibling setae. Rimoportula on every valve, excentrically located. **Reference for description:** Hernández-Becerril (1998). **Morphologically assigned species:** *C. rostratus*.
Subgenus *Hyalochaete* Gran	
Section	

Anastomosantia Ostenfeld	**Description:** setae united by a bridge. Chains mostly loose. **References for description:** Ostenfeld (1903); Hernández-Becerril (1996). **Morphologically assigned species:** *C. anastomosans*.
Compressa Ostenfeld; emended by Yang Li and Lundholm (in Xu et al., 2019)	**Description:** valves broadly elliptical to compress. Numerous small chloroplasts in each cell. Apertures usually moderately large. Terminal setae little different from others. Intercalary setae of two types: thin, common setae and heavy special setae. Heavy setae contorted with spiralling rows of spines and poroids, or heavy setae not visually contorted lacking rows of spines and poroids. Resting spores smooth or with a row of spicules. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Xu et al. (2019). **Morphologically assigned species:** *C.* cf. var. *contortus*, *C. contortus*.
Constricta Ostenfeld	**Description:** cells with one or two chloroplasts and a marked constriction at the base of the valve mantle. Girdle at least one-third the length of the cell. Terminal setae mostly thicker than the others. Resting spores, when present, about the middle of the cell with numerous spines on both valves. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Morphologically assigned species:** *C. constrictus*.
Curviseta Ostenfeld; emended by Gran	**Description:** chains usually curved, with setae all bent in one direction without special end cells. One chloroplast. **References for description:** Ostenfeld (1903); Gran (1905); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C.* cf. *tortissimus*, C. *curvisetus* 1-2-3, *C. debilis* 2-3, *C. pseudocurvisetus*, *C. tortissimus*.
Cylindrica Ostenfeld	**Description:** cells with valves nearly circular (cylindrical). Apertures very narrow. Small, numerous chloroplasts. Terminal setae not thicker than others. Resting spores about middle of the cells, smooth or with spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. lauderi*, *C. teres*.
Diadema (Ehrenberg) Ostenfeld; emended by Gran	**Description:** one chloroplast per cell. Chains long with conspicuous terminal setae. Primary valve of resting spores with branched processes or crown of spines, or sometimes smooth. **References for description:** Ostenfeld (1903); Gran (1905); Cupp (1943). **Morphologically assigned species:** *C. diadema* 1-2, *C. rotosporus*.
Dicladia (Ehrenberg) Gran; emended by Lebour	**Description:** multiple chloroplasts per cell and setae with large pores. Terminal and intercalary setae similar. Resting spores, when known, with two horns armed with small branches on primary valves. **References for description:** Gran (1905); Lebour (1930); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Morphologically assigned species:** *C.* cf. *decipiens*, *C.* cf. *lorenzianus*, *C. decipiens*, *C. elegans*, *C. lorenzianus* 1-2.
Diversa Ostenfeld	**Description:** one chloroplast per cell. Short rigid chains. Inner setae of two kinds. Terminal setae less spread out than a special pair of setae in middle of cell. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. diversus*.
Furcellata Ostenfeld	**Description:** chains generally loose, without differentiated terminal setae. One chloroplast. Resting cells excentrically arranged in mother cell, lying close together two and two, with thick coalesced setae; with smooth valves or with short spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. cinctus, C. radicans*.

Laciniosa Ostenfeld	**Description:** one or two chloroplasts per cell. Girdle rather long. Aperture large. Terminal setae usually thicker than the others, not diverging greatly. Resting spores smooth or with minute spines on primary valve, not in the middle of the cell. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. brevis* 1-2-3.
Protuberantia Ostenfeld; emended by Hernández-Becerril	**Description:** two chloroplasts per cell, each with a large pyrenoid situated in a protuberance in the middle of the valve surface. Valves with poroids. Resting spores paired with two long setae or free without setae. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996); Gaonkar et al. (2018). **Morphologically assigned species:** *C. didymus* 1-2, *C. protuberans*.
Simplicia Ostenfeld	**Description:** cells small and fragile, generally single or two or three together. In case of chain formation, there is no differentiation of terminal setae. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. tenuissimus*.
Socialia Ostenfeld	**Description:** chains irregular and curved embedded in mucilage, forming irregularly spherical colonies. One chloroplast. Resting spores smooth or with small spines. **References for description:** Ostenfeld (1903); Cupp (1943); Hernández-Becerril (1996). **Morphologically assigned species:** *C. dichatoensis*, *C. socialis*, *C. sporotruncatus*.
Stenocincta Ostenfeld	**Description:** a single chloroplast per cell. Usually narrow aperture. Terminal setae curved, thicker than other setae. **References for description:** Ostenfeld (1903); Hernández-Becerril (1996); Gaonkar et al. (2018). **Morphologically assigned species:** *C. affinis*, *C. circinalis*, *Chaetoceros.* sp. Clade Na12A3, *Chaetoceros.* sp. Clade Na13C2, *Chaetoceros.* sp. Clade Na17B2.

Chapter III

Assessing diversity and distribution in Chaetoceros: integration of classical and novel strategies

The material presented in this chapter has been published as Research Article:

“De Luca D., Kooistra W.H.C.F, Sarno D., Gaonkar C.C., Piredda R. (2019). Global distribution and diversity of *Chaetoceros* (Bacillariophyta, Mediophyceae): integration of classical and novel strategies. PeerJ 7:e7410 https://doi.org/10.7717/peerj.7410”.

The paper has been published under the CC BY 4.0 licence. Therefore, I have all the rights to re-use the material contained in it (text, figures and tables).

3.1. Introduction

3.1.1. Primary Biodiversity Data: recording the occurrence of species

Primary Biodiversity Data can be defined as the basic attributes of observations or records of the occurrence of species (Anderson et al., 2016). For centuries, primary species-occurrence data were mostly obtained from taxonomic descriptions of specimens stored in museums, herbaria and private collections (Chapman, 2005). In the last few years, biological recording has evolved, particularly due to the involvement of citizens and the application of molecular tools (Isaac and Pocock, 2015; Pocock et al., 2015). Indeed, nowadays data are also gathered through satellite tracking and direct or remote observation (Croxall et al., 1993), frozen tissue collections and seed banks (Chapman, 2005), environmental DNA (August et al., 2015), and citizen science initiatives (Devictor et al., 2010; Hochachka et al., 2012).

Regardless their sources, data for biological recording are generally presence-only records (opportunistic incidence records, Peterson et al., 2011) since they do not report any info about the absence of the species into a particular area at the time of the survey. Furthermore, they are subject to bias in space and time, such as uneven sampling due to information gathered in urbanized or easily accessible areas and in suitable weather conditions for citizen science projects (Kéry et al., 2010; Isaac and Pocock, 2015) or a time series data for a small area in the case of checklists.

The uses of primary species-occurrence data in natural sciences are numerous and different, from the monitoring of biodiversity (Soberón and Peterson, 2009) and invasive species (Zanetos et al., 2005) to the identification and management of marine protected areas (Araújo and Williams, 2000) and development of conservation plans (Myers et al., 2000; Rondinini et al., 2006).

Field notes and checklists, such observations from early naturalists, scientific expeditions, and museum records, are among the most traditional data used by biologists to document

past patterns of species' distribution and abundance (Droege et al., 1998) and due to their source, they are generally highly reliable. These data are generally more biased in space than in time because related to an area chosen for being especially diverse for the taxon of interest and so intensively sampled over time (Prendergast et al., 1993).

The growth of biological records in recent decades led to the establishment of recording schemes and the organization and storing of such data in freely accessible online portals, such as the National Biodiversity Network Gateway (NBN Gateway; http://www.nbn.org.uk/) and GBIF (http://www.gbif.org/) (Isaac and Pocock, 2015; Powney and Isaac, 2015).

In recent years, environmental DNA (eDNA) data, defined as any DNA-containing trace left behind by organisms in the environment, were added to the list of PBDs (August et al. 2015; Lawson Handley, 2015). Despite eDNA metabarcoding can complement and overcome the limitations of conventional methods by targeting different species simultaneously and catching greater diversity, research is still needed to understand the complex spatial and temporal dynamics of the various eDNA types in the environment (Deiner et al., 2017). Among the different approaches for the characterisation of eDNA, DNA metabarcoding revealed useful at assessing species distribution of marine diatoms of the genus *Leptocylindrus* (Nanjappa et al., 2014), whilst metagenomics for the estimation of species abundance, distributions and richness in fungi (Unterseher et al., 2011).

3.1.2. Primary Biodiversity Data for planktonic species

Biodiversity data of planktonic species are traditionally gathered through samples collected over time, either once though opportunity, or many times through long term ecological research (LTER) projects at single sites (e.g. Helgoland Roads, MareChiara; Blanes Bay Microbial Observatory, Hawaii Ocean Time series), or once at each of many sites through expeditions (e.g., Challenger, Plankton-Expedition). A shortcoming of all these sampling

schemes is that they provide incomplete distribution maps of species with many "blank" regions and seasons for which information is lacking. Furthermore, species distribution patterns often reflect the distribution of the scientists studying the species, or tracks of expeditions (Droege et al., 1998). Sampling intensity is often skewed towards areas known to be diverse for taxa of interest because those areas attract the collectors (Prendergast et al., 1993). However, despite largely time-biased (i.e. sampling events occurring in single dates), these data have the advantage of providing information from areas difficult to access that would otherwise be unexplored (Ji et al., 2013).

Some of the initiatives in the plankton world tried to overcome these issues, such as the Sir Alister Hardy Foundation for Ocean Science's (SAHFOS) program of putting plankton recorders behind ships to sample tracks recurrently (Southward et al., 2005), and the involvement of the public in citizen science initiatives (Castilla et al., 2015; Busch et al., 2016). The results are usually available in form of taxonomic monographs, checklists, or species descriptions.

Among the freely accessible online portals where it is possible to check occurrence data for planktonic species are the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) (Isaac and Pocock, 2015; Powney and Isaac, 2015) and the Ocean Biogeographic Information System (OBIS; http://iobis.org/). GBIF contains occurrence data for both aquatic and terrestrial species gathered from different sources as natural history collections, environmental monitoring programmes, recording initiatives and citizen scientist projects. On the contrary, OBIS only focuses on world's ocean biodiversity and biogeographic data but uses the same sources of data as GBIF except for museum specimens and herbaria collections. Both facilities contain records that are processed according to the Darwin Core Standard (DwC, Wieczorek et al., 2012). Specific for algae is AlgaeBase (Guiry and Guiry, 2018), a repository of information with updated taxonomic info, images, bibliographic items and distributional records of algae curated by

phycologists. It focuses mainly on taxonomy, but provides also taxonomically reliable literature sources on distribution.

Molecular approaches revealed to be particular useful for the study of planktonic species, especially algae. Taxonomic assignment of specimens based on morphology alone can be inaccurate due to cryptic diversity or variation in morphological characters. This is why species identification is often done nowadays using DNA-based methods (Vanormelingen and Souffreau, 2010; Zimmermann et al., 2015). In addition, high throughput sequencing of taxonomically discriminative barcode regions (HTS metabarcoding), has revolutionised our capacity to gather biodiversity data from environmental samples allowing to identify of a plethora of species present in complex sample matrices or from mass collections of specimens.

HTS metabarcoding is particularly common in the study of marine microbial communities, as shown by several recent projects aimed at characterising the diversity and distribution of sea life. Examples are BioMarKs (http://www.biomarks.eu), the Cariaco Microbial Observatory (Edgcomb et al., 2011), Tara Oceans (https://oceans.taraexpeditions.org/en/m/about-tara/), Ocean Sampling Day, OSD (https://www.microb3.eu/osd.html), and time-series at aforementioned LTER stations. These initiatives are in many ways complementary and additive. For instance, Tara Oceans samples have been taken along a global oceanic trajectory on different dates, and the 18S rDNA-V9 region was used as metabarcode (e.g., Malviya et al., 2016), whereas OSD sampled globally as well, but at coastal sites, on a single day (June 21st summer solstice) and used the 18S rDNA-V4 region (e.g., Kopf et al., 2015). Their standardised procedures, including a centralised hub for laboratory work and data processing guaranteed consistency and data interoperability, and the resulting sequences and contextual data are now publicly available. Previous examples of the use of OSD or Tara Oceans datasets to map phytoplankton distribution were performed using only one of two datasets, without

integration of classical sources and at high taxonomic levels (e.g. Malviya et al., 2016; Lopes dos Santos et al., 2017; Penna et al., 2017; Tragin and Vaulot, 2018).

As result of all these activities, a wealth of different kinds of plankton biodiversity data is now available from various sources and in different formats, waiting to be applied to fields such as biogeography, biodiversity estimations, conservation and climate change biology. The integration of all these classical data sources and results from HTS metabarcoding may help to improve environmental monitoring, -management and -policy decisions (Kelly et al., 2014; Thomsen and Willerslev, 2015).

3.1.3. Aim of this work

In this work, I highlight the importance of the integration of classical and novel primary biodiversity data and the challenges related to them through the assessment of the global distribution of *Chaetoceros*. *Chaetoceros* is a highly diverse genus of marine planktonic diatoms (VanLandingham, 1968; Rines and Hargraves, 1988), and an abundant one globally (Guiry and Guiry, 2018). Molecular studies (e.g., Gaonkar et al., 2018) make it comparable to higher taxonomic categories (e.g. family or orders) in other diatom lineages. Cryptic diversity seems to be extensive in this group (Kooistra et al., 2010; Balzano et al., 2017; Gaonkar et al., 2017; Li et al., 2017) affecting the mapping of species distribution patterns based on morphological data.

I first explore the potential of different sources of occurrence data at assessing distribution and abundance of a highly diverse phytoplankton genus as well as its species richness in various regions all over the world. Then, I assess distribution patterns of *Chaetoceros* species using metabarcoding data and compare them with literature data in selected species in order to evaluate their potential and limits in biodiversity assessments.

3.2. Materials and Methods

3.2.1. Data collected from available public repositories, literature and checklists

In order to collect comprehensive info about the distribution of *Chaetoceros* species, I have developed a pipeline that is summed up in Fig. 3.1. I started my search consulting AlgaeBase (Guiry and Guiry, 2018). Upon typing "*Chaetoceros*" in the field 'search genus', I performed a preliminary filtering, taking into account only the taxonomically accepted species. For these, I retrieved the listed key literature to take note of the occurrence in the given area. In parallel, I searched Google Scholar for main checklists and distributional records in the literature using as keywords "*Chaetoceros*/phytoplankton distribution" and "*Chaetoceros*/phytoplankton checklist as well as "occurrence" and "biogeography". Papers resulting from cited literature were also considered. This approach allowed retrieving literature data compiled from experts of the field and so limiting misidentification of species.

I used all the papers focusing mainly on taxonomy containing info at the species-level and considered only names of taxonomically accepted species in Algaebase.

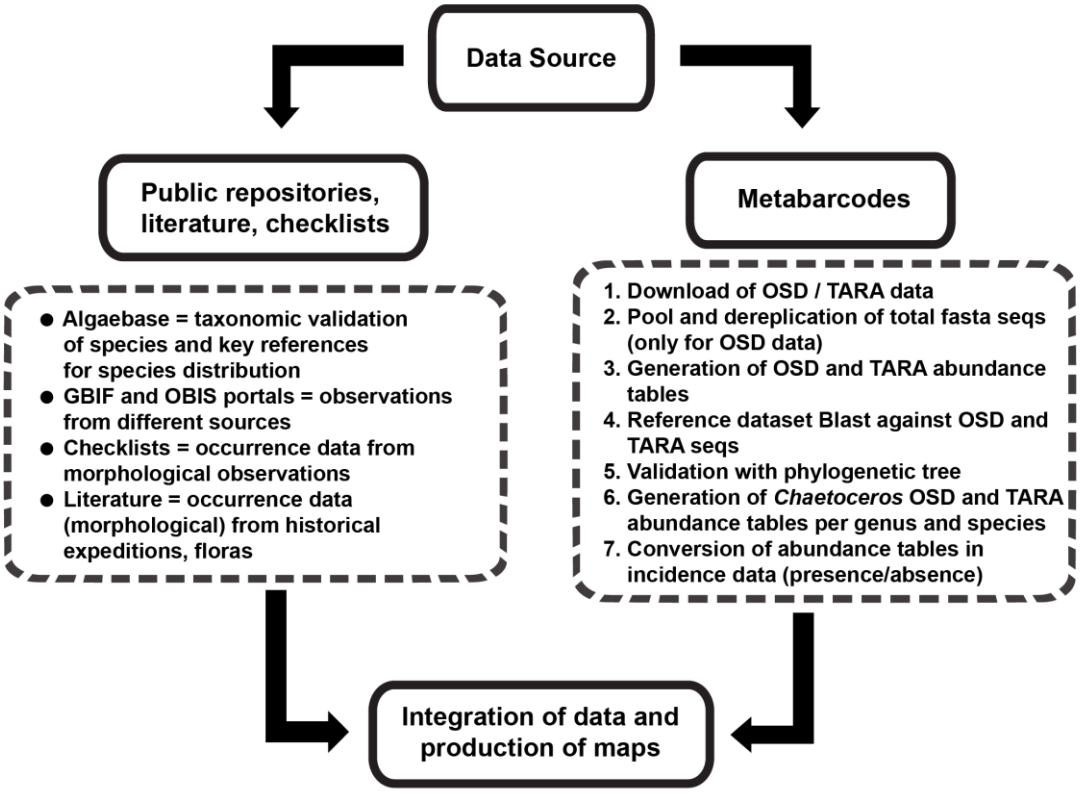


Fig. 3.1. Graphical representation of the main workflow utilised.

To include other sources of occurrence data at the genus level, I checked the Global Biodiversity Information Facility website (GBIF, https://www.gbif.org/) and the Ocean Biogeographic Information System (OBIS, http://iobis.org/). The former is an online tool including occurrence records of both terrestrial and aquatic species gathered from many sources, from museum specimens to geo-tagged smartphone photos. On the contrary, OBIS contains only records of marine species. Although many datasets are published in both, some are only in one (e.g. herbarium or museum collections containing marine species are only available in GBIF). Furthermore, despite the OBIS network is also included in GBIF, differences in updating procedures can cause temporary differences in results.

In both cases, I used the query "Chaetoceros" and downloaded the resulting occurrence data. Occurrence data generated from both databases were plotted using the R (R Core team, 2018) working packages "rgbif" (Chamberlain, 2017) and "robis" (Provoost and

Bosch, 2018) for GBIF and OBIS respectively. Data were plotted using the packages "maps" (Becker et al., 2018) and "ggplot2" (Wickham, 2016).

3.2.2. Data generated from molecular sources

I used the V4-18S metabarcoding data from the Ocean Sampling Day (OSD) initiative and the V9-18S metabarcoding data from the Tara Oceans expedition to obtain new insights on the global distribution of *Chaetoceros*. For the OSD dataset, I downloaded the V4 lgc workable data (e.g. data already pre-processed in order to derive common data sets on which to base follow-up analysis) available at the website https://mb3is.megx.net/osd-files?path=/2014/datasets/workable. Details of sampling protocols and the different kind of molecular data generated are available at https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data, whilst details of pre-processing can be found at https://github.com/MicroB3-IS/osd-analysis/wiki/Sequence-Data-Pre-Processing. The workable fasta files, downloaded for each of 144 geographical sampling sites, were pooled and I generated a total fasta file containing the non-redundant (unique) sequences and a table containing their distribution along the sites (Total OSD abundance table) using mothur v1.41.1 (Schloss et al., 2009).

For Tara Oceans dataset, I downloaded the V9-metabarcode dataset (De Vargas et al., 2017; Ibarbalz et al., 2019) available at https://doi.pangaea.de/10.1594/PANGAEA.873277 and at ENA website with acc. numb. PRJEB6610. Then, following the same pipeline described above, from the total 210 sampling sites, I generated a total unique fasta file and a Total Tara Oceans abundance table.

To generate distribution data, I used a high-quality data reference containing a selection of taxonomic validated *Chaetoceros* sequences of the 18S gene (Goankar et al., 2018). In particular, the reference barcode dataset included 202 *Chaetoceros*, 15 *Bacteriastrum* and 29 outgroup taxa. The fragments V4 and V9 were extracted from the full-length 18S genes

and aligned using MAFFT online (Katoh et al. 2017). In order to avoid mis-assignations at species level, for the two fragments (V4 and V9) I simulated several thresholds of clustering based on genetic distances (commands “dist.seqs” and “cluster” in mothur) (Schloss et al., 2009).

The V4 and the V9 reference sequences were used as queries for a local BLAST against the two global metabarcode datasets OSD and Tara Oceans. For the mapping at genus level, I set the threshold to 90 % of identity and from the outputs of BLAST I retained only the metabarcode hits having a query coverage with the reference > 370 bp in the analysis of V4 OSD dataset, and >105 bp for V9 Tara Oceans dataset. The metabarcodes extracted were aligned with the references, including outgroup taxa, using MAFFT online (Katoh et al., 2017) and two phylogenetic trees were then built in FastTree v2.1.8 (Price et al., 2010) using the GTR model and visualised in Archaeopteryx v0.9901 (Han and Zmasek, 2009). Metabarcode hits clustering within the outgroup clades were excluded from further analyses, whereas the others were considered as validated *Chaetoceros*. Their abundances and distributions were extracted from the Total OSD and Tara Oceans abundance tables to generate the *Chaetoceros*-genus OSD abundance table and *Chaetoceros*-genus Tara Oceans abundance table. For the mapping at species level, I first evaluated the information generated from the analyses described above for the V4 and V9 fragments (calculation of the genetic distances and simulation of several thresholds of clustering). Based on them, I extracted only the BLAST hits assigned in the range 100-99% of similarity. This range was identified as the best compromise between the precision required to an assignation at species level and the intra-species variation that could occur especially at global level. After the BLAST, I applied the same procedure described above for the genus level (alignment and generation of tree) to validate the assignations and we generated the *Chaetoceros*-species abundance table for the OSD and for Tara Oceans datasets.

The *Chaetoceros*-genus abundance tables were used both in term of occurrence and abundance of V4 and V9 reads in each sampling site. Abundance values were log10-transformed and plotted using *ggplot2* (Wickham, 2016).

Finally, to explore in detail the performances of classical and molecular data, I selected three species as case study: i) *C. tenuissimus* as test of cosmopolitan species; ii) *C. gelidus* as species with restricted distribution; iii) *C. neogracilis* as example of putative cryptic species complex.

3.3. Results

3.3.1. Data collected from available public repositories, literature and checklists

According to AlgaeBase, the genus *Chaetoceros* contained 370 species names and 172 intraspecific ones, 220 of which have been flagged as taxonomically accepted species based on the available literature (searched on 15/10/2018). This discrepancy is due to the occurrence of many homotypic or heterotypic synonyms in the literature as well as species of uncertain taxonomic status, which need taxonomic revision. I further filtered the 220 taxa flagged as taxonomically accepted (e.g. removing entries occurring twice) obtaining a final table (Table A3.1, Appendix III) with 175 entries at the date of the search. I considered the latter taxa in the count for species richness from literature data (see below).

The distribution map of *Chaetoceros* obtained using GBIF data (Fig. 3.2A) was based on 201,047 occurrence records from 1863 to 2018 (https://www.gbif.org/occurrence/charts?q=chaetoceros). Data were mostly from human observations (75.7 %) and preserved specimens (20.2 %) (GBIF.org, 14 September 2018, GBIF Occurrence Download https://doi.org/10.15468/dl.nofa8w). The definition of records is available at https://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/BasisOfRecord.html. Filtered occurrence data from GBIF are also available as supplementary info (Table A3.2, Appendix III). No information

from literature was available for *Chaetoceros* in GBIF data. Most of the observations were from the North Atlantic Ocean between 35° - 60° N and -80° W – 10°E (Continuous Plankton Recorder Dataset, SAHFOS, 83,513 counts; Réseau d'Observation et de Surveillance du Phytoplancton et des Phycotoxines, REPHY, 17,742 counts; QUADRIGE, 12,458 counts), followed by the Pacific coasts of North and Central America and Australia (Fig. 3.2A).

The distribution map obtained searching *Chaetoceros* in the OBIS database (Fig. 3.2B) contained 389,206 records from 1863 to 2016 (Table A3.2, Appendix III). Most of observations were from the World Ocean Database 2009 (119,592), followed by the Continuous Plankton Recorder (86,309) and the Japan Oceanographic Data Center Dataset (JODC, 31,388).

Chaetoceros occurrence data were found in 435 GBIF datasets and 179 OBIS datasets, of which 20 were shared (Table A3.2, Appendix III).

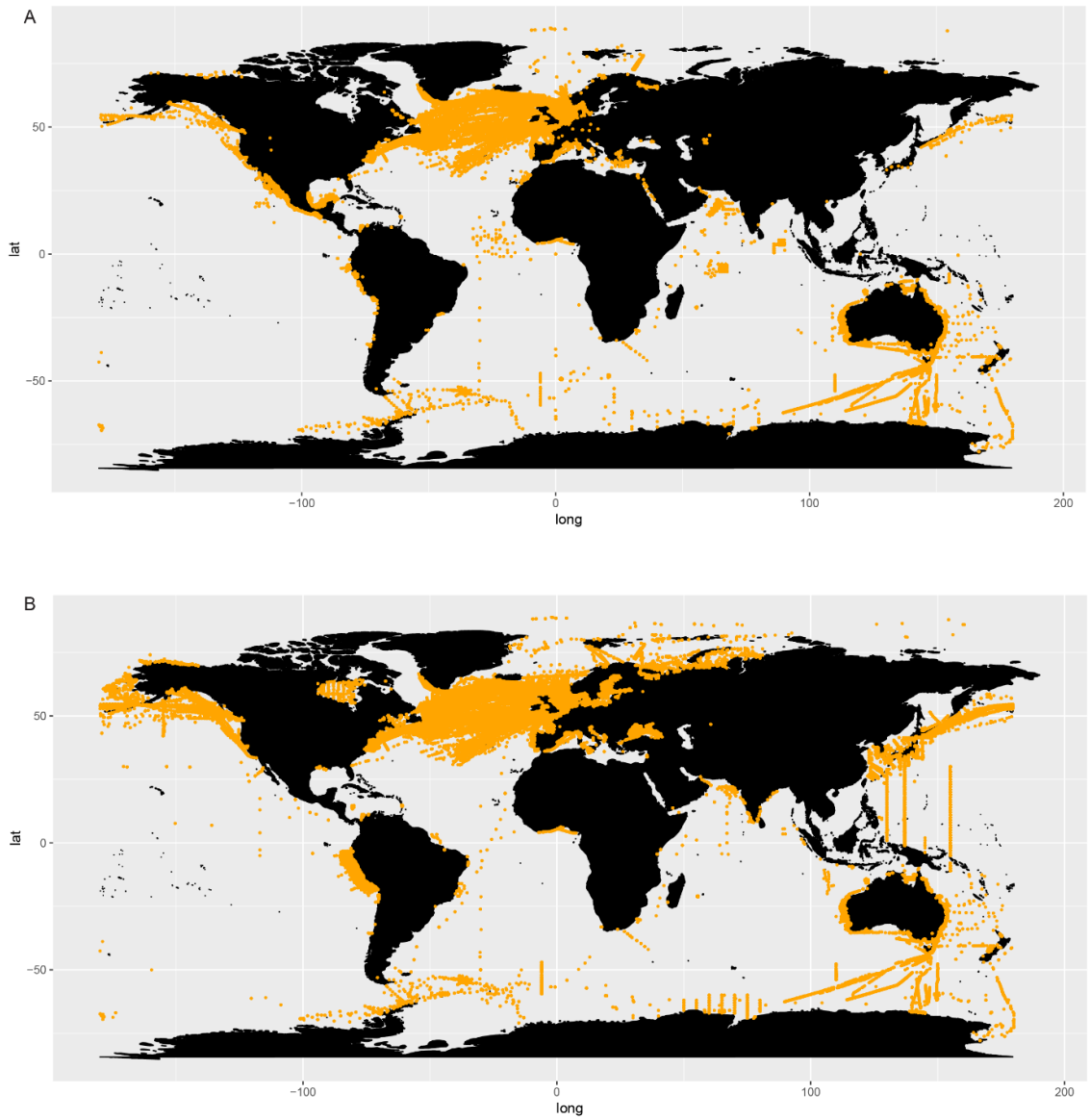


Fig. 3.2. Occurrence of *Chaetoceros* using (A) GBIF and (B) OBIS data.

The literature search conducted in Google Scholar and the other sources (see Material and Methods) resulted in 84 main bibliographic references reporting data of *Chaetoceros* occurrences (Table A3.3, Appendix III). These data encompassed both single observations and time series across the world, covering a period from 1873 to 2017 (Table A3.3, Appendix III). These data surely represent only a fraction of the whole existing literature (and the literature indexed in Google Scholar) but are representative of the principal checklists/floras compiled by expert taxonomists and of the spatial coverage were *Chaetoceros* is known to occur. None of these bibliographic references (checklists and

papers) was contained in GBIF or OBIS datasets (Table A3.2, Appendix III). According to these data, *Chaetoceros* species mostly occurred in the temperate to equatorial coastal waters of northern hemisphere and in the subtropical to tropical coastal waters of the southern one (Fig. 3.3).

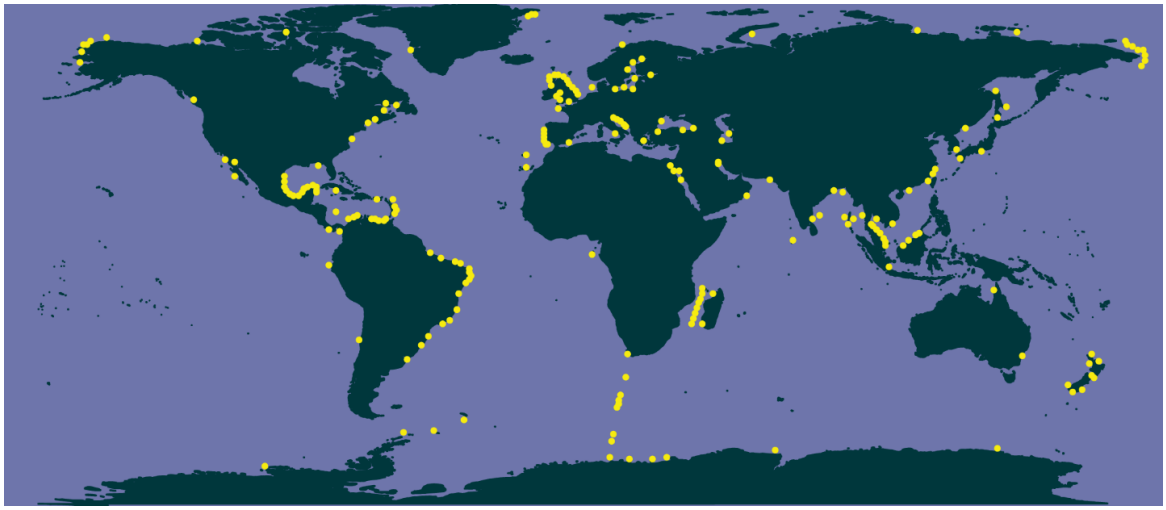


Fig. 3.3. Occurrence of *Chaetoceros* using literature data.

In terms of species richness, here defined as the number of valid species recorded in each locality's checklist, I found the highest values in the temperate waters of European coasts (North Sea, Baltic Sea, and middle Adriatic Sea, Fig. 3.4), followed by the tropical and subtropical waters of Brazil, Mozambique Channel and Indonesia (Fig. 3.4). The lowest number of species was found in the subpolar waters alongside the coasts of northern countries (Canada, Greenland, Norway and Russia) as well as in the equatorial ones of southern oceans (Fig. 3.4).

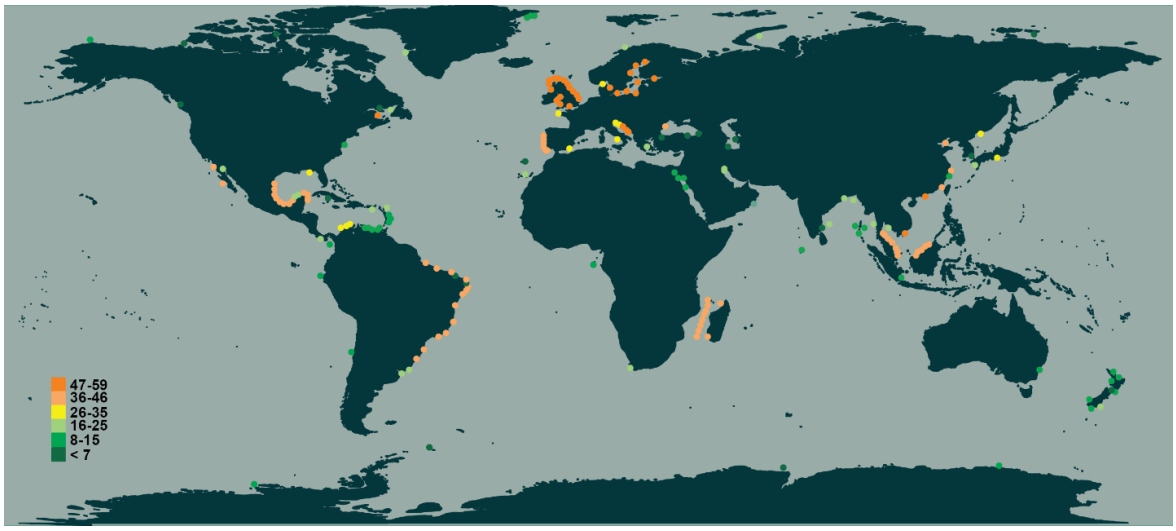


Fig. 3.4. Species richness of *Chaetoceros* estimated from literature data. Colours refer to the different classes of abundance (number of species recorded).

3.3.2. Data generated from molecular sources

Based on the generation of distances and simulation of clustering thresholds, the clustering at 100% similarity of the V4 *Chaetoceros* reference dataset (unique or non-redundant sequences) resulted in the collapse of only multiple strains from the same species, whereas the clustering at 99 % similarity threshold resulted in the collapse of several species (Table A3.4, Appendix III). On the contrary, in the V9 *Chaetoceros* reference dataset the clustering at 100% of similarity produced the collapse of different taxa generating more limitations in the mapping at species level (Table A3.4, Appendix III).

At genus level, I found occurrences of *Chaetoceros* taxa in 138 out of 144 OSD sampling sites (96%) and 146 out of 210 Tara Oceans stations (70%), highlighting very wide distribution of the genus (Fig. 3.5, Table A3.5, Appendix III).

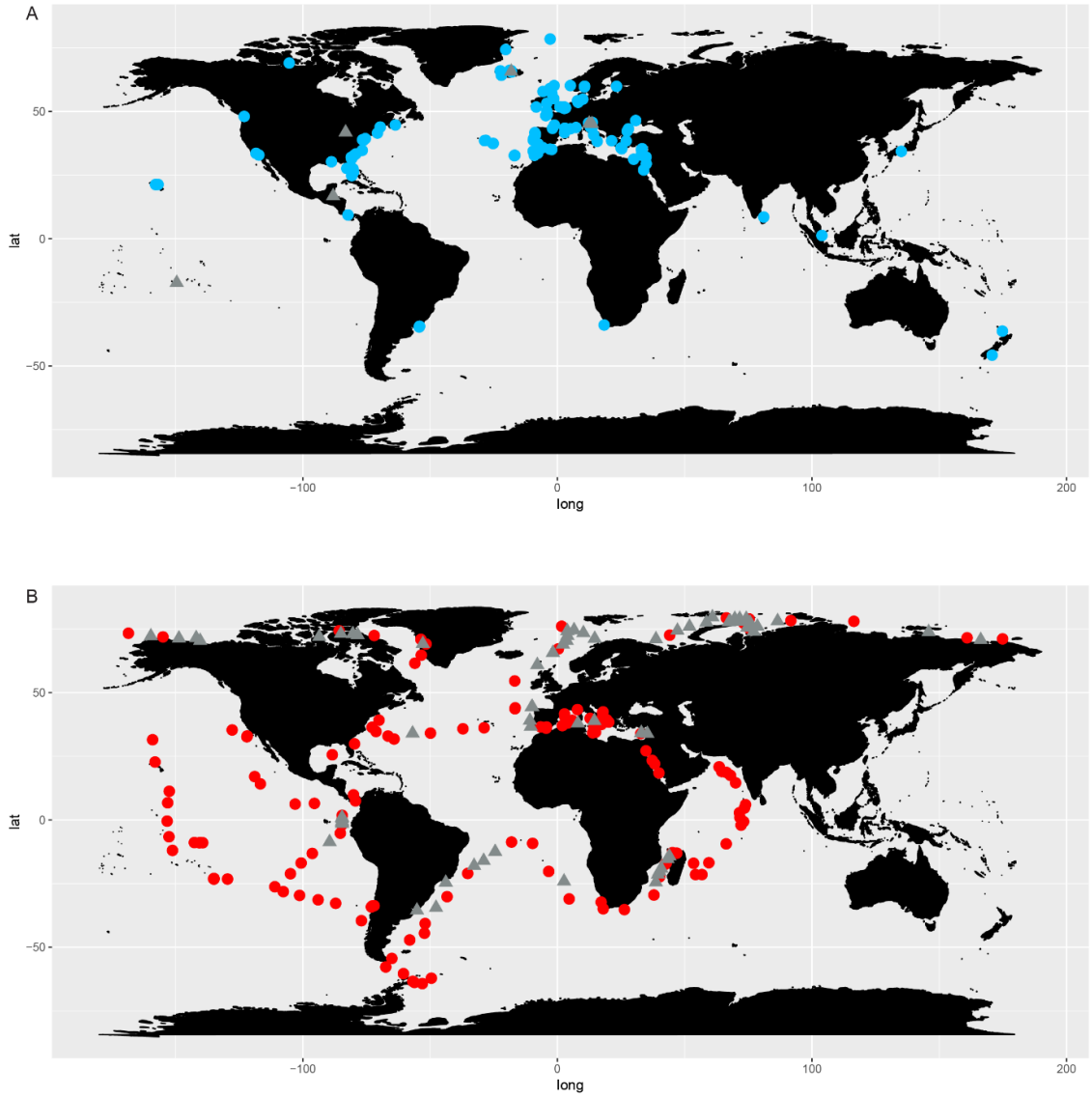


Fig. 3.5. *Chaetoceros* distribution according to OSD (A) and Tara Oceans (B) data. Dots indicate presence of *Chaetoceros* taxa in the sampling stations, whilst triangles their absence.

The plot of abundances, both in OSD and in Tara Oceans datasets, showed that *Chaetoceros* was equally abundant in the northern as in the southern hemisphere (Fig. 3.6). The highest abundances (in terms of reads) were mostly found in the polar to temperate regions of the two hemispheres, with some exceptions in the equatorial coastal waters of India and Indonesia (Fig. 3.6A). Lowest abundances were found in the subtropical to equatorial zones, especially in open ocean stations in the case of Tara Oceans dataset (Figure 6B), in the Red Sea for both datasets, and other few sites in the OSD dataset (Fig. 3.6A).

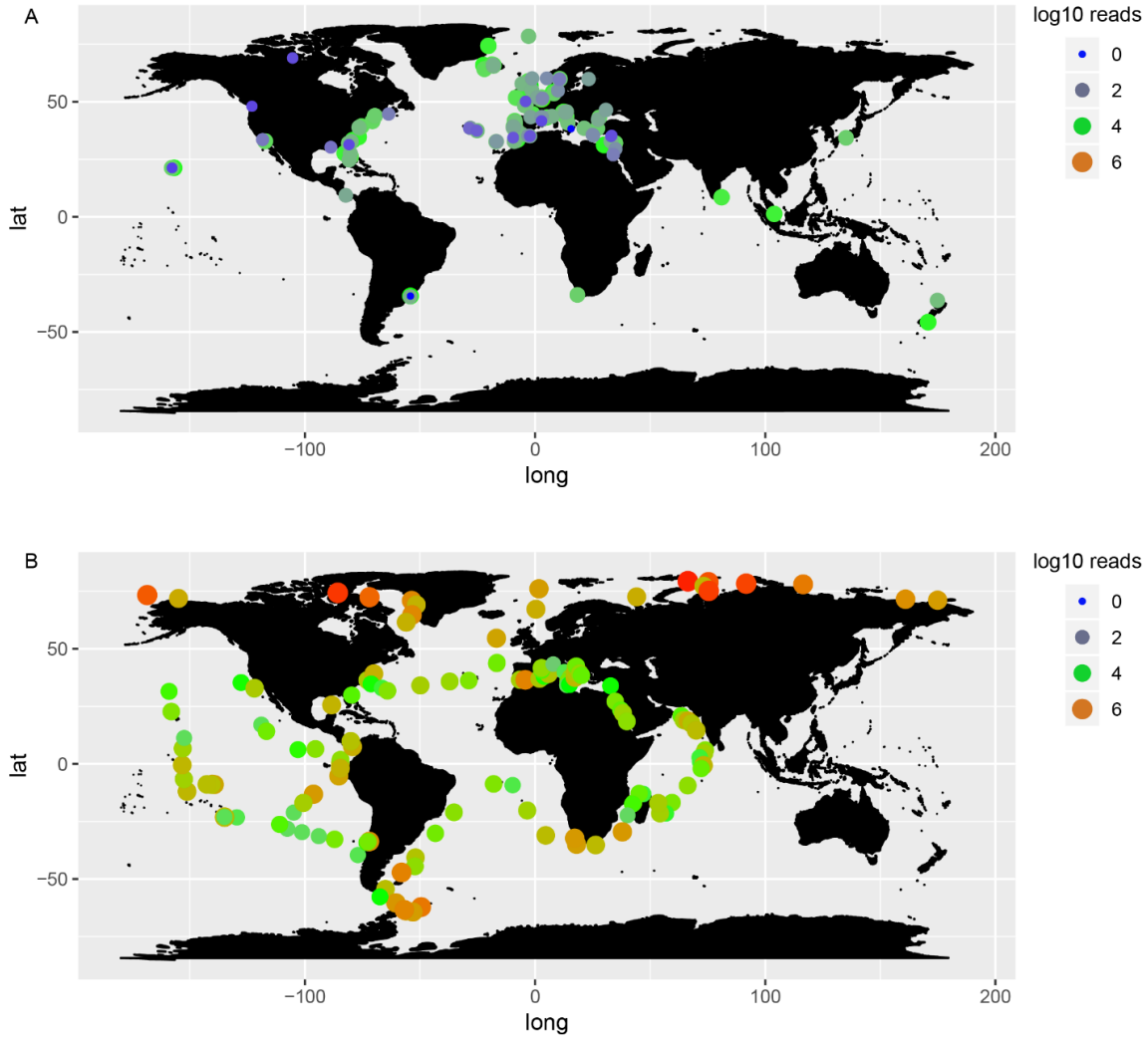


Fig. 3.6. Log10 abundance of *Chaetoceros* reads according to OSD (A) and Tara Oceans (B) datasets. Size and colours of the circles refer to the abundance.

At species level I generated, at 99% similarity threshold, a map of occurrence in the OSD and Tara Oceans datasets for each of the 69 *Chaetoceros* species (Figure A3.1, Appendix III). The only exceptions were *C.* cf. *vixvisibilis* Na16A3 and *C.* sp. Clade Na28A1 strain Na26C1, in which the collapse of barcodes prevented the plot of occurrences in Tara Oceans stations at species level.

The comparison of literature and genetic (metabarcoding) data in selected species of *Chaetoceros* (Fig. 3.7) showed consistency in the signal for *C. tenuissimus* and *C. gelidus*, and highlighted the occurrence of putative cryptic species in *C. neogracilis*.

In *C. tenuissimus*, literature (Fig. 3.7A) and metabarcoding data (Fig. 3.7B) confirmed a cosmopolitan distribution, with metabarcoding data providing new records for African, Asian and New Zealand coasts (Fig. 3.7B).

For *C. gelidus*, genetic data from OSD and Tara Oceans (Fig. 3.7D) confirmed the distribution area of literature data (field observations, Fig. 3.7C) but also included new records for Canada, North Scotland and Iceland (Fig. 3.7D). The species was also found in one OSD station in the Caribbean side of Panama coasts, but very low abundance (2 reads at 100% similarity).

C. neogracilis revealed to be an example of cryptic species complex. According to literature, the species was found both in the northern and southern hemisphere (Fig. 3.7E). On the contrary, occurrence data from metabarcoding revealed instead a distribution limited to the northern hemisphere, so covering just a small part of the distribution range known from literature data (Fig. 3.7F).

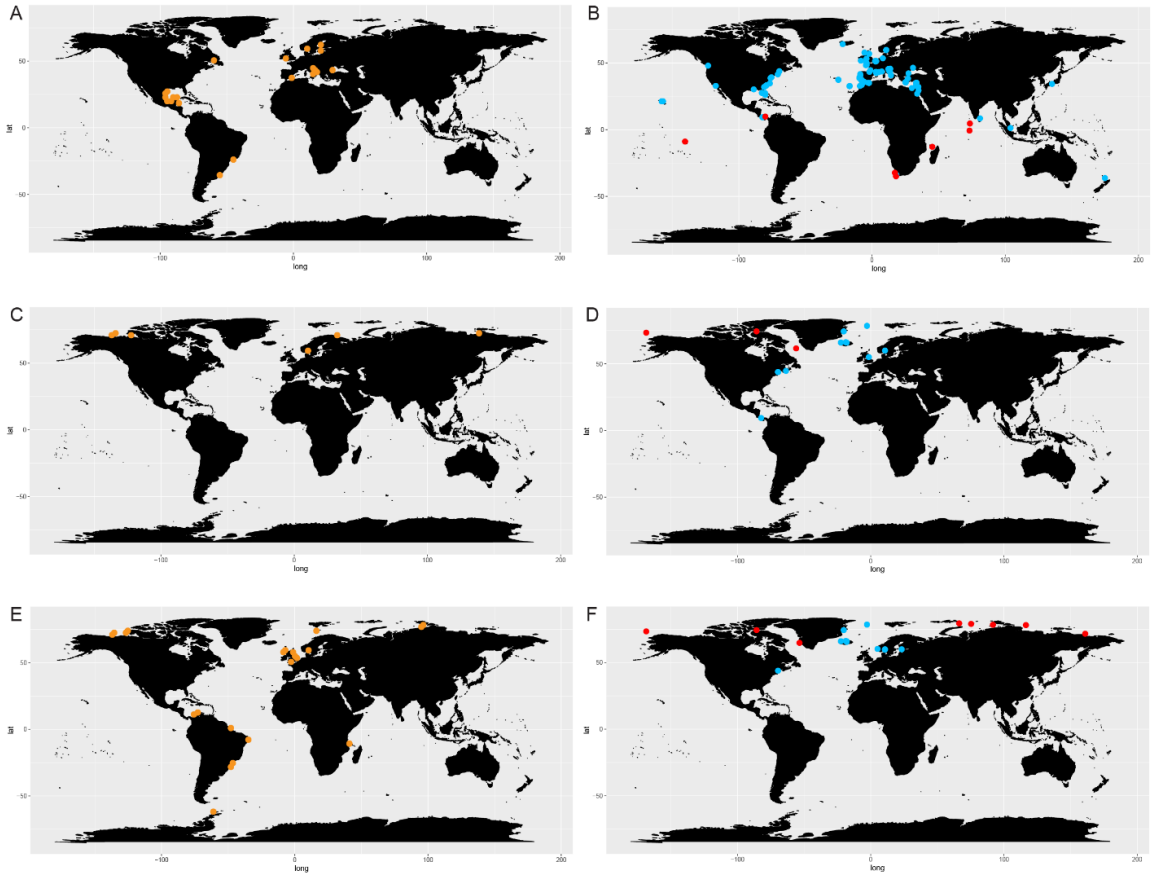


Fig. 3.7. Distribution of *C. tenuissimus* (A, B), *C. gelidus* (C, D) and *C. neogracilis* (E, F) according to literature (orange dots) and metabarcoding data (blue dots for OSD and red dots for Tara Oceans). Maps containing the sites considered for literature and metabarcoding data are found in Figure 3 and Figure A3.1 respectively.

3.4. Discussion

3.4.1. Global distribution of Chaetoceros

The more complete picture of *Chaetoceros* distribution was provided by the GBIF and OBIS platforms, which contain a huge amount of data from different sources (fossils, literature, machine and human observations, museum and herbarium specimens) and cover a wide time scale (in this case more than 150 years). Despite OBIS is a resource dedicated to marine organisms already included in GBIF database, I did not recover the same number of records and datasets from the two sources. Differences in updating data procedures are partially responsible for such temporary differences in results. Furthermore, some kinds of

information as museum collections are only available in GBIF, generating the necessity to interrogate both databases also in the case of marine species to ensure a complete mapping. The overview provided by the Google Scholar search of the main phytoplankton checklists is, despite the obvious limitations, able to provide the main distributional areas of the genus. Google Scholar can be considered as a convenient starting place to start a literature search, not a comprehensive endpoint. It has among its advantages the fact that is easily accessible to retrieve data that are otherwise stored in libraries' catalogues and databases and goes back in time in the scale of hundreds of years or more. Since this approach is highly sensitive to the kind and order of keywords used for the search, I cannot exclude the possibility of having missed some information, even if multiple searches were performed. However, I have retrieved datasets that are not included in GBIF or in OBIS. This aspect underlines that despite the big effort to generate and update these global databases, a minor part of the information could be lost. Furthermore, it highlights the difficulty for the researches to produce an exhaustive assessment of all the available data of a particular taxon. Probably, more effort is needed by the institutions from around the world to provide and share biodiversity datasets generated.

The two global metabarcoding datasets OSD and Tara Oceans, despite biased in time and space, provided an overall distribution map of the genus that is comparable to the one obtained from the sources discussed above. This clearly highlights that, despite some weaknesses (e.g. Coissac et al., 2012; Ficetola et al., 2015), the metabarcoding approach, in less than a decade from its diffusion, was able to compete with classical morphological records gathered over hundreds of years. At the moment, metabarcoding data cannot replace the classical ones, and should be seen as a powerful complement rather than a substitute of other data sources (Bush et al., 2017). For instance, the Tara Oceans dataset added new occurrence information for equatorial regions and other open ocean sites in the southern hemisphere, contributing to our knowledge in these still poorly investigated areas.

Despite both OSD and Tara Oceans datasets are open access, the extraction of information from these sources is not straightforward and requires some bioinformatic skills that are not common, especially among taxonomists.

My results showed that all data sources (GBIF, OBIS, Google Scholar search, OSD and Tara Oceans) support a cosmopolitan distribution of this genus as suggested by Rines and Hargraves (1988) using only classical sources, and Malviya et al. (2016) using only metabarcoding data. In terms of occurrence, *Chaetoceros* taxa showed a global distribution ranging from coastal areas to open ocean and from polar to tropical regions. However, the different data sources point out a prevalence of taxa in the temperate coastal waters between the temperate waters 60°N and 30°N and in the subtropical and equatorial ones between 30°N and 30°S. This can be due to the presence in such regions of various habitats (upwelling zones, lagoons, oligotrophic as well as eutrophic regions) and the marked seasonality in the water, which offer opportunities of co-existence of species by spatial or seasonal niche partitioning. Boreal regions are poorer probably because there is only the single summer season for phytoplankton growth.

With some exceptions (e.g. Hernández-Becerril and Granados, 1998 for the Gulf of Mexico and Hernández-Becerril, 1996 for the Mexican Pacific), the tropics are generally under-investigated for species diversity, though this is now ameliorated by recent studies in those regions (Li et al., 2013; 2017; Chamnansinp et al., 2015).

3.4.2. Abundance of Chaetoceros *at global scale*

In general, patterns of abundance in both molecular datasets suggest that *Chaetoceros* is equally abundant in the temperate to equatorial waters of northern and southern hemispheres, with the highest abundance in the Arctic region. A paucity of reads was generally observed from many sites located in the open ocean. This observation could reflect the well known hypothesis made on terrestrial ecosystems, according to which cold

to temperate regions contain less species but more abundant. However, multiple variables involved could alter such results. First and obvious is that the picture provided by metabarcoding data is very limited in space and time, and could not represent the real situation. Second, some species may have been collected during a bloom period, which could explain the high values of abundance. *Chaetoceros socialis*, for example, is known to be an important component of diatom blooms in the Barents Sea (Von Quillfeldt, 2000). Third, data here used (V4 and V9 regions) are from a multi-copy gene and since the copy number in *Chaetoceros* is unknown, this aspect, combined with a hypothetical bloom, could hamper our conclusions.

A previous mapping of *Chaetoceros* in Tara Oceans dataset was performed in Malviya et al. (2016) using only 46 stations. In the latter study, *Chaetoceros* was found to be highly abundant in the Southern Ocean and absent in the polar regions of the northern hemisphere. My analysis, using the complete Tara Oceans dataset (210 stations), showed that *Chaetoceros* is present also in the polar regions of the northern hemisphere, highlighting the fact that the wider the coverage of sampling and/or the integration from other source the better the resolution of distribution.

3.4.3. Integration of literature and metabarcoding data: three study cases in Chaetoceros

The direct comparison of literature and metabarcoding data in three selected species of *Chaetoceros* shows the power of novel molecular data coupled with classical occurrence data. In the case of *C. tenuissimus*, the molecular data allowed to increase the geographic range of distribution of this cosmopolitan species with new records in African, Asian and New Zealand coasts. Yet, in *C. gelidus* molecular data confirmed the previous knowledge on its restricted distribution in cold water, also adding new records for Canada, North Scotland and Iceland. For this reason, at the moment I interpret the occurrence of two reads found in one OSD station in the Caribbean coasts as a spot occurrence rather than a stable

geographic point. However, global changes could alter limits both in cosmopolitan or restricted species with consequent range expansion or contraction, highlighting the importance to generate baseline studies of the geographic distribution range of taxa to use as bases for future comparisons.

More complex is, instead, the case of *C. neogracilis*. The epithet *C. neogracilis* (*C. gracile* Schütt) has been attributed in the past to many small, unicellular *Chaetoceros* taxa collected worldwide (Rines and Hargraves, 1988). This led to considering the species cosmopolitan. A recent study by Balzano et al. (2017) from the Beaufort Sea (Canadian Arctic) revealed the occurrence of morphologically similar strains sharing identical 18S rDNA sequences, but belonging to four distinct genetic clades based on 28S rDNA, ITS-1 and ITS-2 markers. Since OSD and Tara oceans datasets are based on the 18S gene, I regarded these entities as one single species. In Balzano et al. (2017) they are also reported to co-occur at the stations they visited. The reference barcode from Balzano et al. (2017) blasted against the two datasets found identical sequences only in the cold waters of the northern hemisphere, so covering just a small part of the distribution range known from literature data. My results strongly suggest that under the name *C. neogracilis* there is a species restricted to polar regions of the northern hemisphere (as highlighted also by Balzano et al., 2017). Furthermore, as pointed out by these authors, a closely related species occurs in the cold waters of Antarctica, whilst the status of the *neogracilis* taxon reported in literature from South America and Africa is still to be determined. It could be a complex of taxa with similar morphology and, further samplings in these regions accompanied by genotyping of strains, will help clarifying the taxonomic status. However, the example highlights how the integration of several sources is required to a correct interpretation at species level of the patterns obtained from a metabarcoding sampling.

3.4.4. Assessing species distribution in Chaetoceros

The maps of occurrences generated using the OSD and Tara Oceans datasets for each of the 69 *Chaetoceros* species, provide new insights on biogeography in marine diatoms.

According to available literature, few endemic diatom species are known, and they are mostly freshwater (e.g. *Eunophora* in Tasmania and New Zealand, Vanormelingen et al., 2008 and *Cyclotella minuta* in Lake Baikal, Mackay et al., 2006) or from saline inland lakes (e.g. *Aulacoseira baicalensis*). Claims of putative endemic marine diatoms exist and are discussed in Mann and Vanormelingen (2013). In the specific case of *Chaetoceros*, Hernández-Becerril (1996) recognised that little efforts have been made to assess its world distribution but, starting from literature data available and personal observations, he grouped taxa according to major regions as inhabitants of cold waters, temperate to subtropical waters, world-wide warm waters and tropical and subtropical waters. Metabarcoding data here analysed suggested that cases of endemism or restricted geographical distributions can be also found in the marine environment. I detected species whose occurrence is limited to single basins as the Mediterranean Sea (*C. diversus* 1) or part of it (*C. throndsenii* in the Adriatic Sea) as well as distribution restricted to climatic zones (e.g. the polar to temperate zones for *C. constrictus*, *C. danicus* strain RCC2565, *C. debilis* 1 and *C. neogracilis*).

3.4.5. Future directions

In this chapter, I have highlighted both the importance of the integration of data and the challenges related to it, generating a comprehensive primary baseline of the geographic distribution range and diversity for *Chaetoceros*, one of the most diverse and abundant genera of marine planktonic diatoms. I have also stressed out that, at the moment, molecular and classical sources tend to be organised and maintained in separated repositories or infrastructures, forcing the users interested in integration of such sources to

do a not trivial work across the several sources of data and analyses. Certainly, molecular approaches can improve our knowledge both reducing mis-assignments (taxonomic lumping) in cryptic species complexes and helping for rare and small species not easily detected with traditional methods, especially for microbial (protist) species. However, this is not always true and, as I have shown, the short fragments used in metabarcoding can be identical in closely related taxa, not allowing a discrimination at species level (Cowart et al., 2015; Mordret et al., 2018; Piredda et., 2018). Nonetheless, metabarcoding data are a valuable source of primary biodiversity data.

The knowledge of the geographic range of species is a key issue in ecology, conservation and evolutionary biology allowing investigating causes and consequences of the limits. Climate change can alter these limits with consequent range expansion or contraction, and several examples have been reported (Walther et al., 2002; Parmesan and Yohe, 2003; McLachlan et al., 2005). This process is supposed to be underway, stressing the need to collect, integrate and summarise data available to create a Primary Biodiversity Data baseline. These collections provide bases for future comparisons or model predictions to support biodiversity change assessments.

References

Anderson, R. P., Araújo, M. B., Guisan, A., Lobo, J. M., Martinez-Meyer E., Townsend, A., Soberón, J. (2016). Are species occurrence data in global online repositories fit for modeling species distributions? The Case of the Global Biodiversity Information Facility (GBIF). Available at https://serval.unil.ch/resource/serval:BIB_768D188CEA5B.P001/REF.pdf (accessed 14 October 2018).

Araújo, M. B., Williams, P. H. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation*, *96*(3), 331-345.

August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., Jepson, P. (2015). Emerging technologies for biological recording. *Biological Journal of the Linnean Society, 115*(3), 731-749.

Balzano, S., Percopo, I., Siano, R., Gourvil, P., Chanoine, M., Marie, D., ... Sarno, D. (2017). Morphological and genetic diversity of Beaufort Sea diatoms with high contributions from the Chaetoceros neogracilis species complex. *Journal of Phycology, 53*(1), 161-187.

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., Deckmyn, A. (2018). maps: Draw geographical maps. R package version 3.3.0.

Busch, J. A., Price, I., Jeansou, E., Zielinski, O., van der Woerd, H. J. (2016). Citizens and satellites: Assessment of phytoplankton dynamics in a NW Mediterranean aquaculture zone. *International Journal of Applied Earth Observation and Geoinformation, 47*, 40-49.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., ... Dawson, T. P. (2017). Connecting Earth observation to high-throughput biodiversity data. *Nature Ecology & Evolution, 1*(7), 0176.

Castilla, E. P., Cunha, D. G. F., Lee, F. W. F., Loiselle, S., Ho, K. C., Hall, C. (2015). Quantification of phytoplankton bloom dynamics by citizen scientists in urban and peri-urban environments. *Environmental Monitoring and Assessment, 187*(11), 690.

Chamberlain, S. (2017). rgbif: Interface to the Global 'Biodiversity' Information Facility 'API'. R package version 0.9.8. Available at https://CRAN.R-project.org/package=rgbif (accessed 10 January 2019).

Chamnansinp, A., Moestrup, Ø., Lundholm, N. (2015). Diversity of the marine diatom *Chaetoceros* (Bacillariophyceae) in Thai waters–revisiting *Chaetoceros compressus* and *Chaetoceros contortus*. *Phycologia, 54*(2), 161-175.

Chapman, A. D. (2005). Uses of primary species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available at http://www.niobioinformatics.in/books/Uses%20of%20Primary%20Data.pdf (accessed 14 October 2018).

Coissac, E., Riaz, T., Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology, 21*(8), 1834-1847.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PloS ONE, 10*(2), e0117562.

Croxall, J. P., Briggs, D. R., Prince, P. A. (1993). Movements and interactions of the Wandering Albatrosses: the roles of satellite tracking and direct observations. *Sea Swallow*, *42*, 41-44.

de Vargas, C., Audic, S., Tara Oceans Consortium, Coordinators, Tara Oceans Expedition, Participants (2017). Total V9 rDNA information organized at the metabarcode level for the Tara Oceans Expedition (2009-2012). PANGAEA, https://doi.org/10.1594/PANGAEA.873277

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... Pfrender, M. E. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Devictor, V., Whittaker, R. J., Beltrame, C. (2010). Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions, 16*(3), 354-362.

Droege, S., Cyr, A., Larivée, J. (1998). Checklists: An under-used tool for the inventory and monitoring of plants and animals. *Conservation Biology, 12*(5), 1134-1138.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., ... Epstein, S. (2011). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *The ISME Journal, 5*(8), 1344-1356.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., ... Rayé, G. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources, 15*(3), 543-556.

Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Montresor, M., Sarno, D. (2017). Two new species in the *Chaetoceros socialis* complex (Bacillariophyta): *C. sporotruncatus* and *C. dichatoensis*, and characterization of its relatives, *C. radicans* and *C. cinctus*. *Journal of Phycology*, *53*(4), 889-907.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. *PLoS ONE*, *13*(12), e0208929.

Guiry, M. D., Guiry, G. M. (2018). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. Available at http://www.algaebase.org (accessed 02 June 2018).

Han, M. V., Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics, 10,* 356. https://doi.org/10.1186/1471-2105-10-356.

Hernández-Becerril, D. U. (1996). Morphological study of *Chaetoceros* species (Bacillariophyta) from the plankton of the Pacific Ocean of Mexico. *Bulletin of Natural History Museum of London (Botany), 26*(1), 1-73.

Hernández-Becerril, D. U., Granados, C. F. (1998). Species of the diatom genus *Chaetoceros* (Bacillariophyceae) in the plankton from the Southern Gulf of Mexico. *Botanica Marina*, *41*(1-6), 505-520.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W. K., Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution, 27*(2), 130-137.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ... Picheral, M. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell, 179*(5), 1084-1097.

Isaac, N. J., Pocock, M. J. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society, 115*(3), 522-531.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., ... Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters, 16*(10), 1245-1257.

Katoh, K., Rozewicki, J., Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 1-7. doi: 10.1093/bib/bbx108.

Kelly, R. P., Port, J. A., Yamahara, K. M., Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PloS ONE, 9*(1), e86175.

Kéry, M., Gardner, B., Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography, 37*(10), 1851-1862.

Kooistra, W. H. C. F., Sarno, D., Hernández-Becerril, D. U., Assmy, P., Di Prisco, C., Montresor, M. (2010). Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia, 49*(5), 471-500.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., … Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience, 4*(1), 27 DOI 10.1186/s13742-015-0066-5.

Lawson Handley, L. (2015). How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society, 115*(3), 750-766.

Li, Y., Lundholm, N., Moestrup, Ø. (2013). *Chaetoceros rotosporus* sp. nov. (Bacillariophyceae), a species with unusual resting spore formation. *Phycologia, 52*(6), 600-608.

Li, Y., Boonprakob, A., Gaonkar, C. C., Kooistra, W. H. C. F., Lange, C. B., Hernández-Becerril, D., ... Lundholm, N. (2017). Diversity in the globally distributed diatom genus *Chaetoceros* (Bacillariophyceae): Three new species from warm-temperate waters. *PloS ONE, 12*(1), e0168887.

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noël, M. H., Decelle, J., Romac, S., Vaulot, D. (2017). Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *The ISME Journal*, 11(2), 512-528.

Mackay, A. W., Ryves, D. B., Morley, D. W., Jewson, D. H., Rioual, P. (2006). Assessing the vulnerability of endemic diatom species in Lake Baikal to predicted future climate change: a multivariate approach. *Global Change Biology, 12*(12), 2297-2315.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., … Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences, 113*(11), 1516-1525.

Mann, D. G., Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology, 60*(4), 414-420.

McLachlan, J. S., Clark, J. S., Manos, P. S. (2005). Molecular indicators of tree migration capacity under rapid climate change. *Ecology, 86*(8), 2088-2098.

Mordret, S., Piredda, R., Vaulot, D., Montresor, M., Kooistra, W. H. C. F., Sarno, D. (2018). DINOREF: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. *Molecular Ecology Resources, 18*(5), 974-987.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature, 403*(6772), 853-858.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H. C. F., Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. *PLoS ONE, 9*(8), e103810.

Parmesan, C., Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature, 421*(6918), 37-42.

Penna, A., Casabianca, S., Guerra, A. F., Vernesi, C., Scardi, M. (2017). Analysis of phytoplankton assemblage structure in the Mediterranean Sea based on high-throughput sequencing of partial 18S rRNA sequences. *Marine Genomics, 36*, 49-55.

Peterson, A. T., Soberón, J., Pearson R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., Araujó, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press, Princeton.

Piredda, R., Claverie, J. M., Decelle, J., De Vargas, C., Dunthorn, M., Edvardsen, B., ... Zingone, A. (2018). Diatom diversity through HTS-metabarcoding in coastal European seas. *Scientific Reports, 8*(1), 18059.

Pocock, M. J., Roy, H. E., Preston, C. D., Roy, D. B. (2015). The Biological Records Centre: a pioneer of citizen science. *Biological Journal of the Linnean Society, 115*(3), 475-493.

Powney, G. D., Isaac, N. J. (2015). Beyond maps: a review of the applications of biological records. *Biological Journal of the Linnean Society, 115*(3), 532-542.

Prendergast, J. R., Wood, S. N., Lawton, J. H., Eversham, B. C. (1993). Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters, 1*(2), 39-53.

Price, M. N., Dehal, P. S., Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS ONE, 5*(3), e9490.

Provoost, P., Bosch, S. (2018). “robis: R Client to access data from the OBIS API.” Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO. R package version 1.0.1. Available at https://cran.r-project.org/package=robis (accessed 10 January 2019).

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/ (accessed 7 January 2019).

Rines, J. E., Hargraves, P. E. (1988). *The* Chaetoceros *Ehrenberg (Bacillariophyceae) flora of Narragansett Bay, Rhode Island, USA*. Lubrecht and Cramer, Berlin.

Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H., Possingham, H. P. (2006). Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, *9*(10), 1136-1145.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology, 75*(23), 7537-7541.

Soberón, J., Peterson, A. T. (2009). Monitoring biodiversity loss with primary species-occurrence data: toward national-level indicators for the 2010 target of the convention on biological diversity. *AMBIO: A Journal of the Human Environment, 38*(1), 29-35.

Southward, A. J., Langmead, O., Hardman-Mountford, N. J., Aiken, J., Boalch, G. T., Dando, P. R., ... & Hawkins, S. J. (2005). Long-term oceanographic and ecological research in the western English Channel. *Advances in Marine Biology, 47*, 1-105.

Thomsen, P. F., Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation, 183*, 4-18.

Tragin, M., Vaulot, D. (2018). Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. *Scientific Reports, 8*:14020.

VanLandingham, S. L. (1968). *Catalogue of the fossil and recent genera and species of diatoms and their synonyms. Part II.* Bacteriastrum *through* Coscinodiscus. Cramer, Lehre.

Vanormelingen, P., Verleyen, E., Vyverman, W. (2008). The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. *Biodiversity and Conservation, 17*, 393-405.

Vanormelingen, P., Souffreau, C. (2010). DNA barcoding for species identification and discovery in diatoms. *Cryptogamie, Algologie, 31*(4), 557-577.

Von Quillfeldt, C. H. (2000). Common diatom species in Arctic spring blooms: their distribution and abundance. *Botanica Marina, 43*(6), 499-516.

Walther, G. R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J., ... Bairlein, F. (2002). Ecological responses to recent climate change. *Nature, 416*(6879), 389-395.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

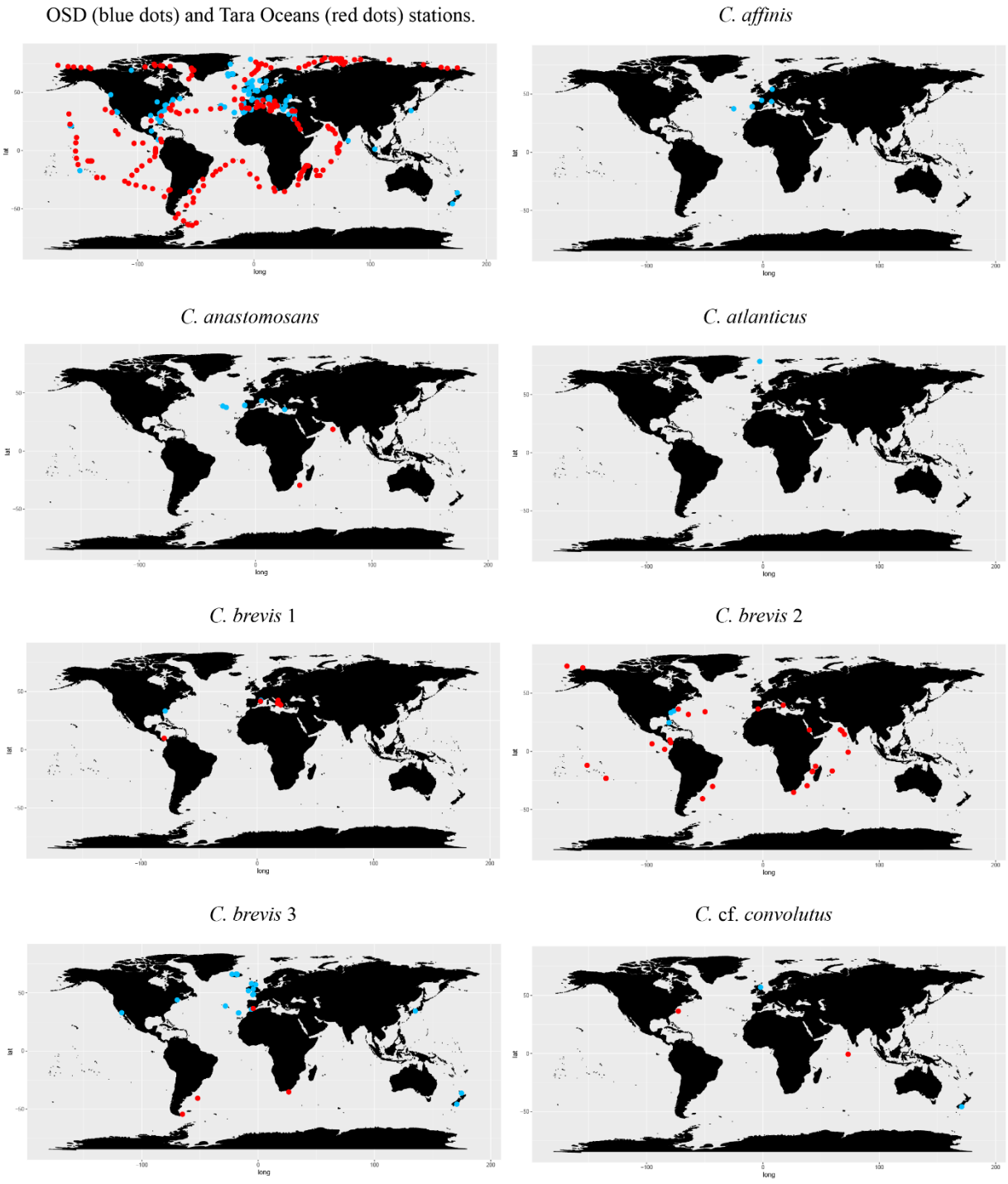
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ... Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS ONE, 7*(1), e29715.

Zenetos, A., Çinar, M. E., Pancucci-Papadopoulou, M. A., Harmelin, J. G., Furnari, G., Andaloro, F., ... Zibrowius, H. (2005). Annotated list of marine alien species in the Mediterranean with records of the worst invasive species. *Mediterranean Marine Science, 6*(2), 63-118.

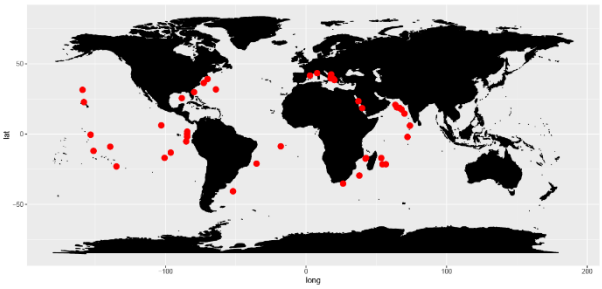
Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, *15*(3), 526-542.

Appendix III

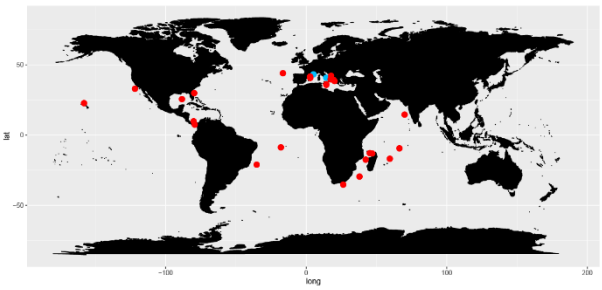
Fig. A3.1. Distribution maps of *Chaetoceros* species using OSD and Tara Oceans datasets.



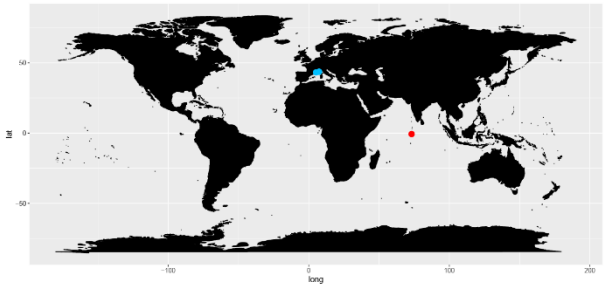
C. cf. *pseudodichaeta*



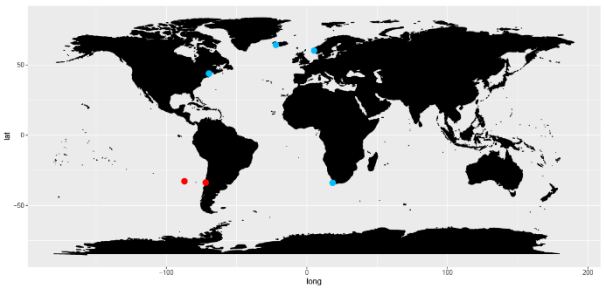
C. cf. *tortissimus*



C. cf. *vixvisibilis*

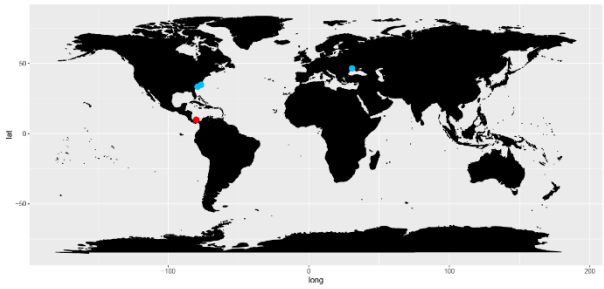


C. cinctus

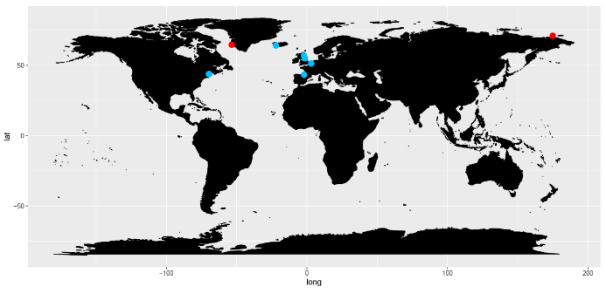


Note: Tara Oceans distribution (red dot) might refer to *C.* sp. Na28A1 due to identical reference sequences in Tara Oceans dataset.

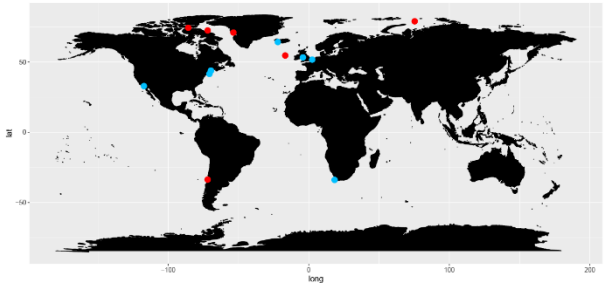
C. circinalis



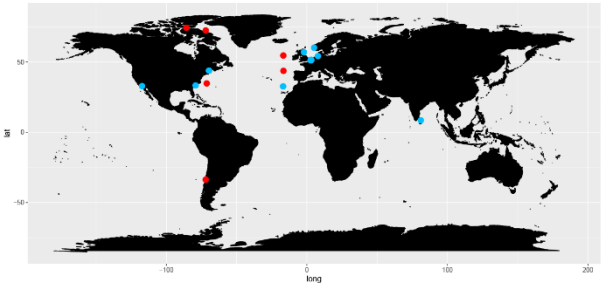
C. constrictus



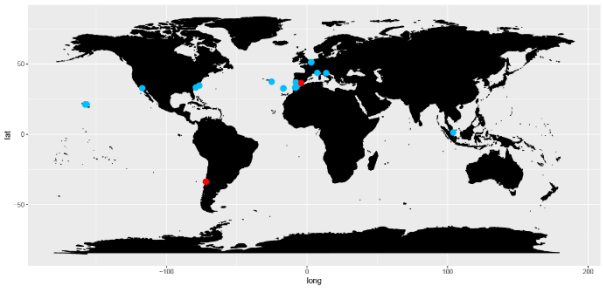
C. contortus



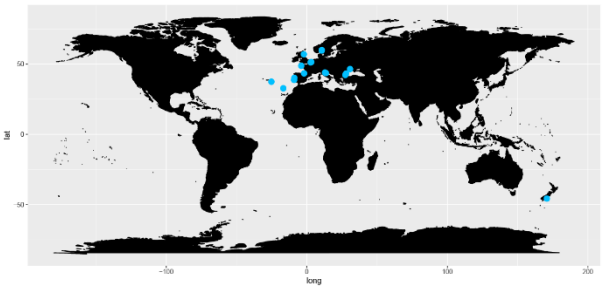
C. contortus cf. var. *contortus*



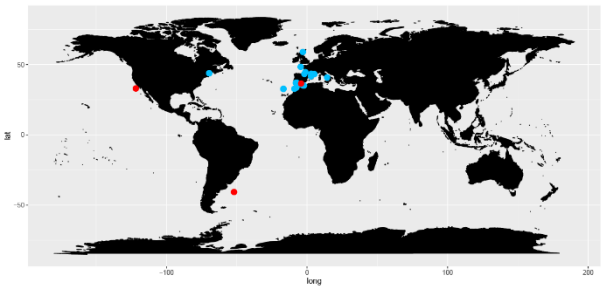
C. costatus



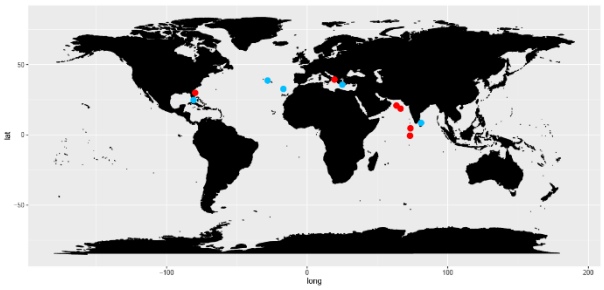
C. curvisetus 1



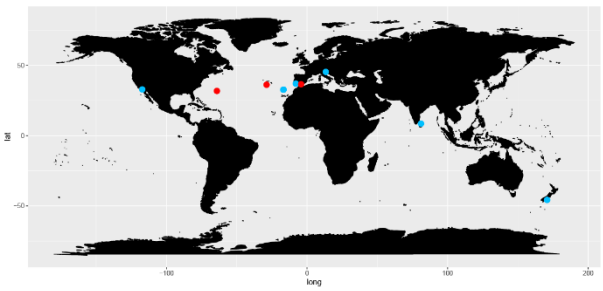
C. curvisetus 2



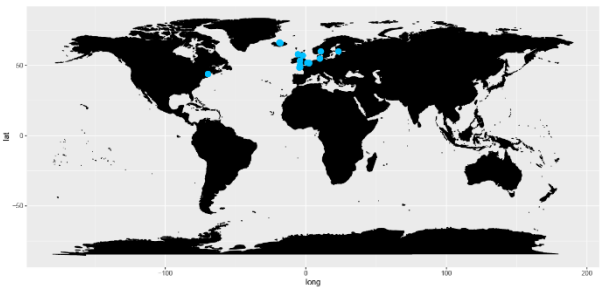
C. curvisetus 3



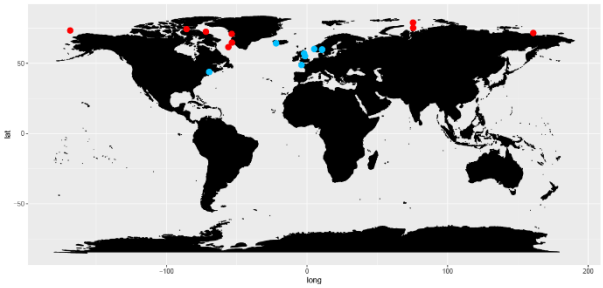
C. danicus (strain newCB1)



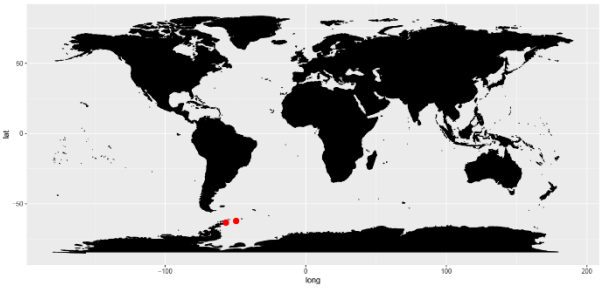
C. danicus (strain RCC2565)



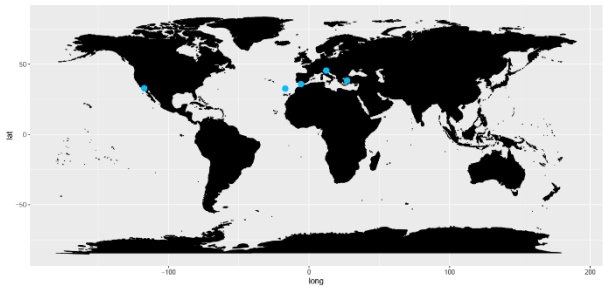
C. debilis 1



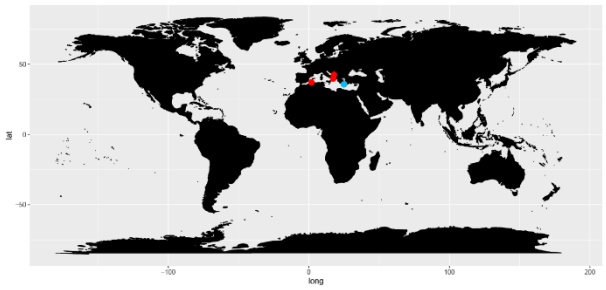
C. debilis 2 (strain L38-A2)



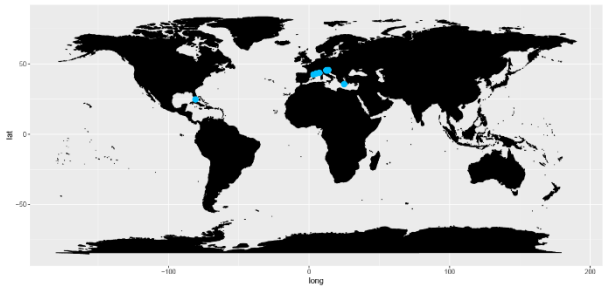
C. didymus 2



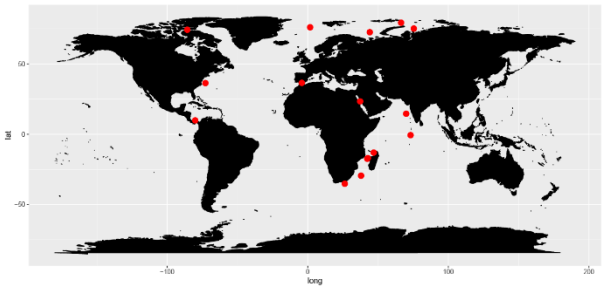
C. diversus 1



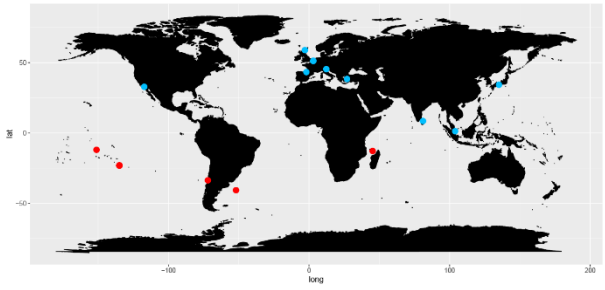
C. diversus 2



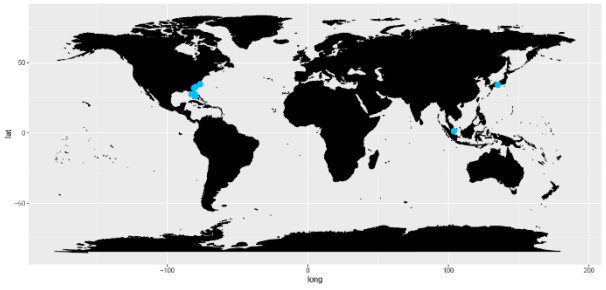
C. eibenii



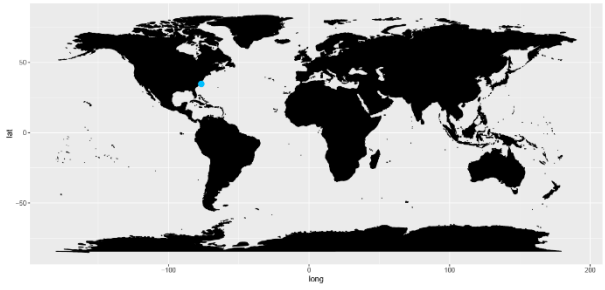
C. elegans (strain Ch12A1)



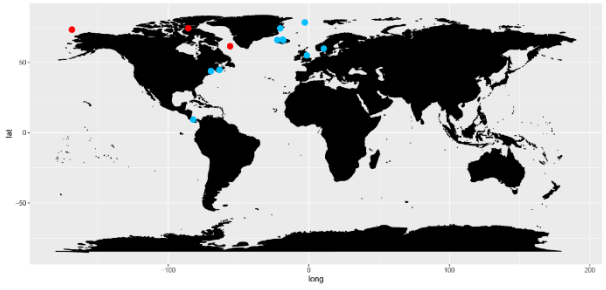
C. elegans (strain MC1001)



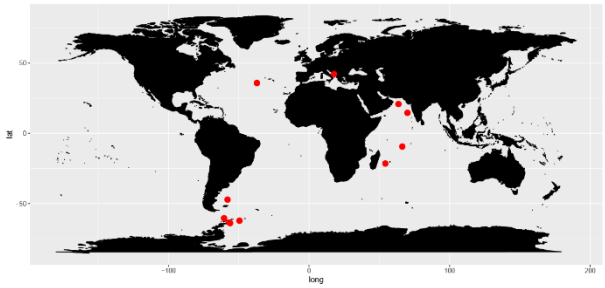
C. elegans (strain MC785)



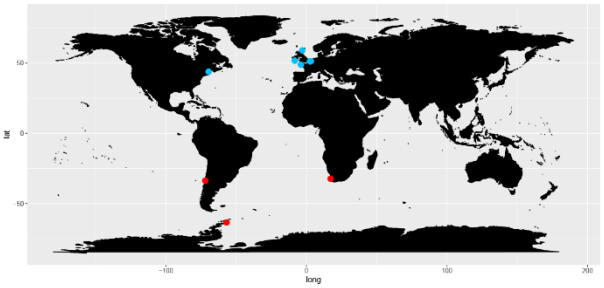
C. gelidus



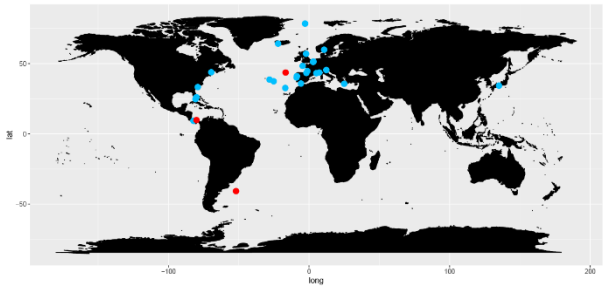
C. debilis 2 (strain MM24-A3)



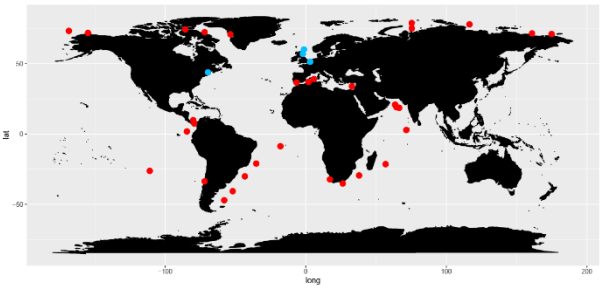
C. debilis 3



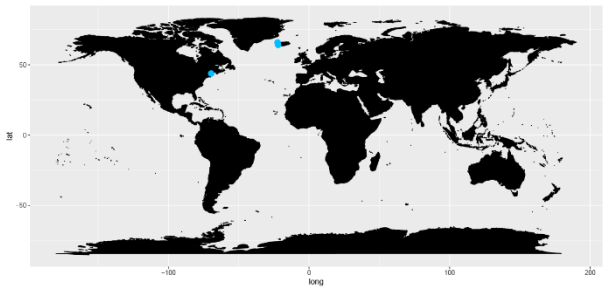
C. decipiens



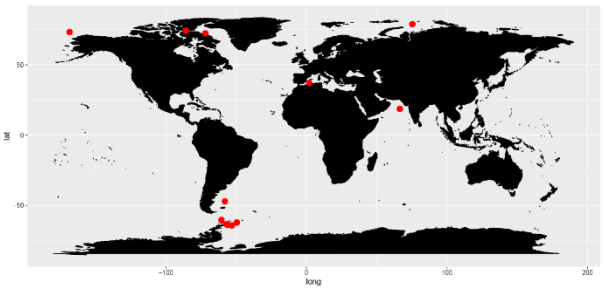
C. diadema 1



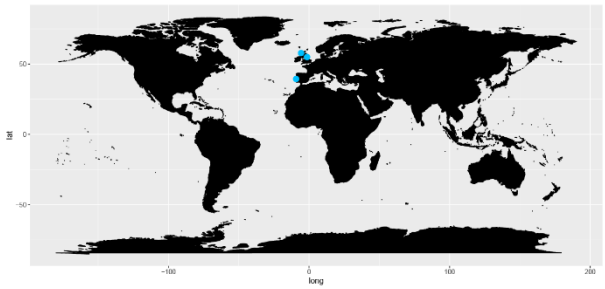
C. diadema 2



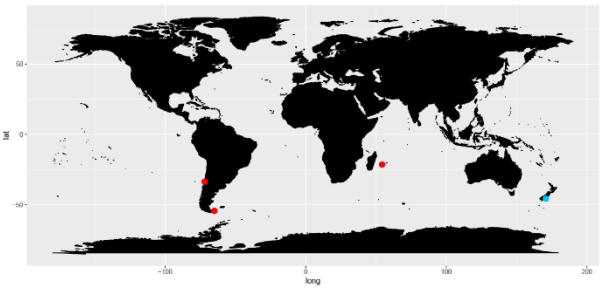
C. dichaeta



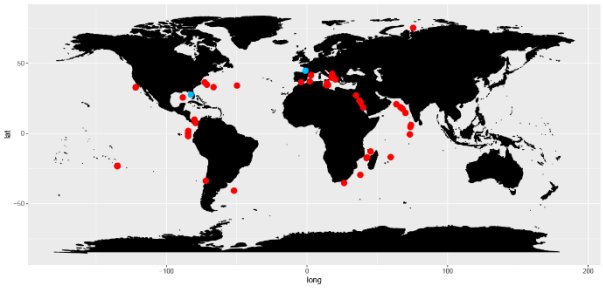
C. dichatoensis



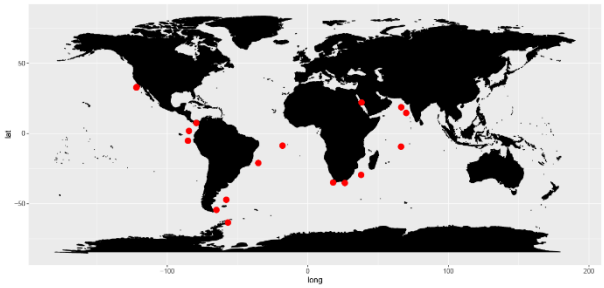
C. didymus 1



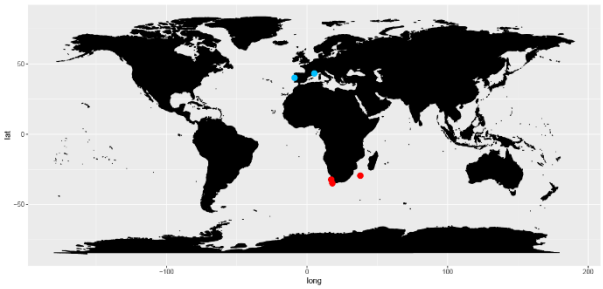
C. lauderi



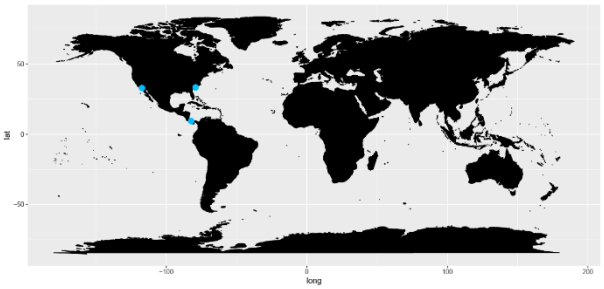
C. lorenzianus 1



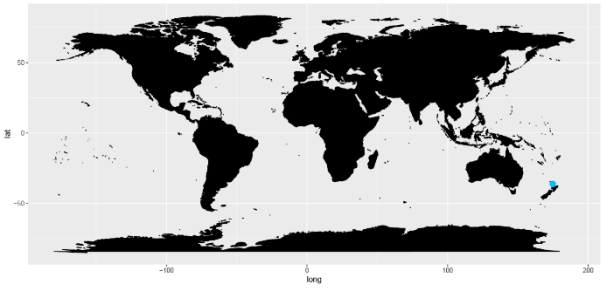
C. lorenzianus 2



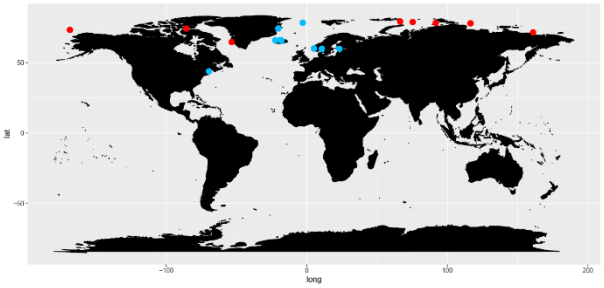
C. mannaii



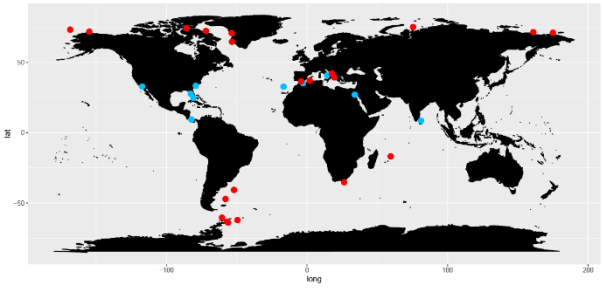
C. minimus



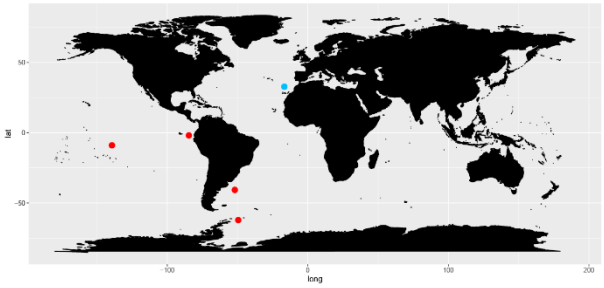
C. neogracilis



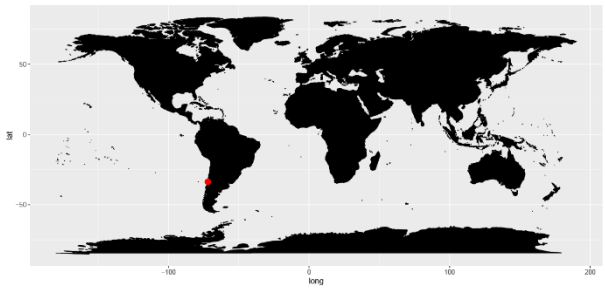
C. peruvianus 1



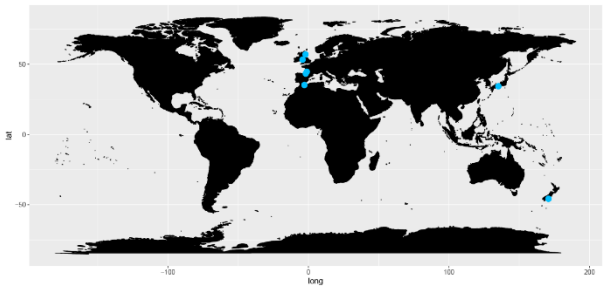
C. peruvianus 2



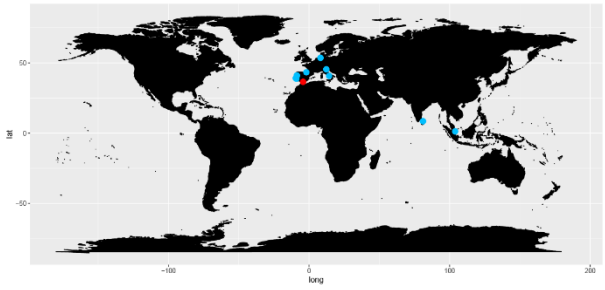
C. protuberans (strain Bristol)



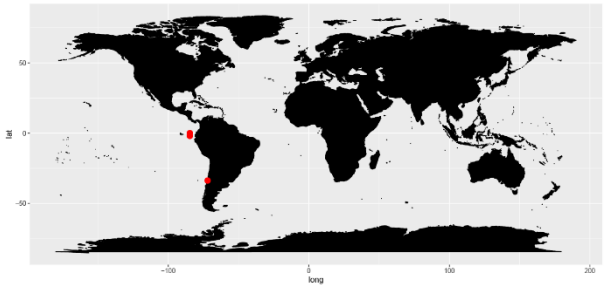
C. protuberans (strain newJC4)



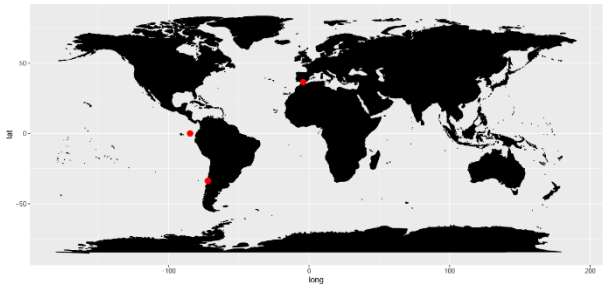
C. pseudocurvisetus



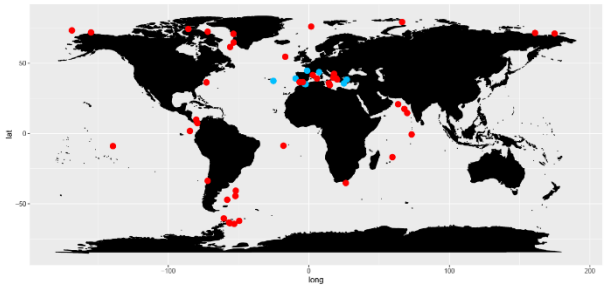
C. radicans (strain CCMP197)



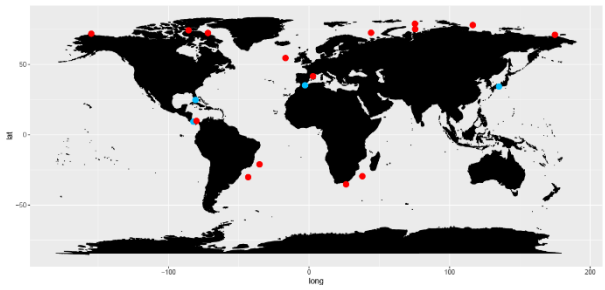
C. radicans (strain Ch11A4)



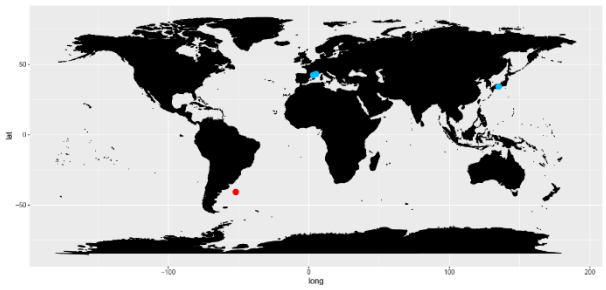
C. rostratus



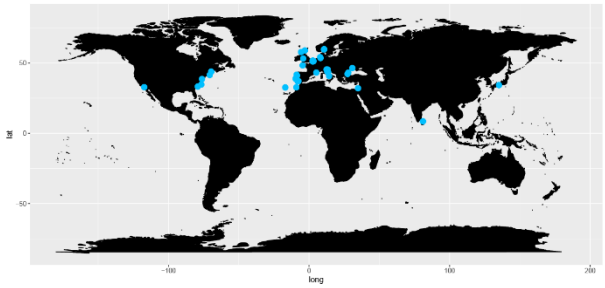
C. rotosporus



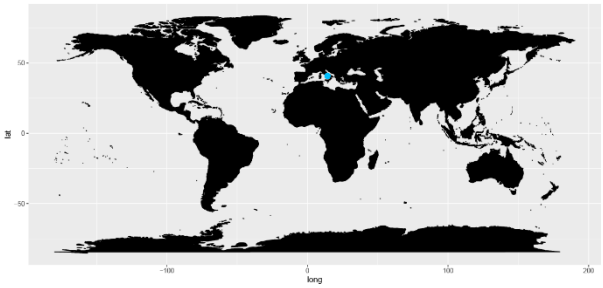
C. seiracanthus



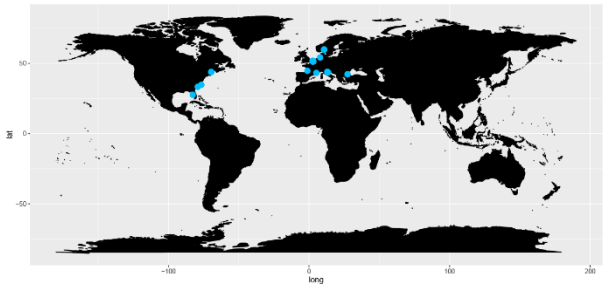
C. socialis



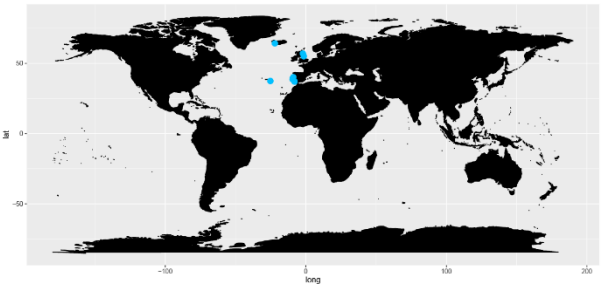
C. sp. Clade CDP22



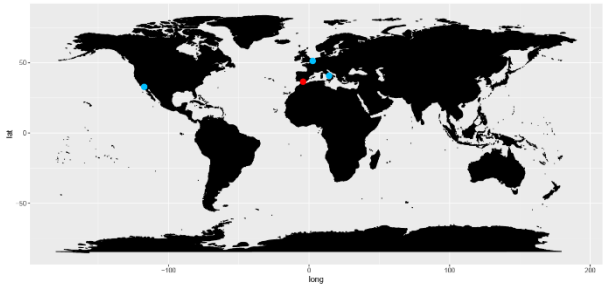
C. sp. Clade Na11C3



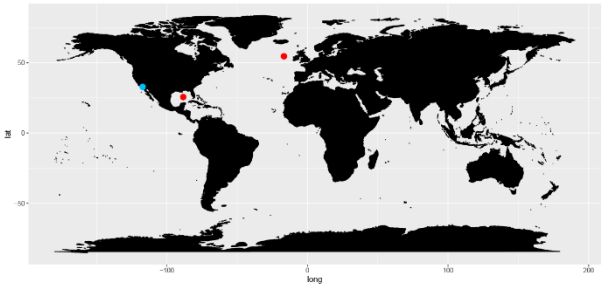
C. sp. Clade Na12A3



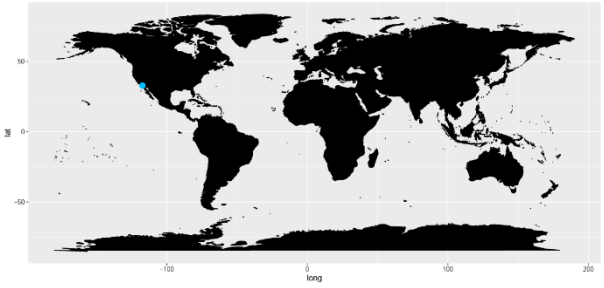
C. sp. Clade Na13C1



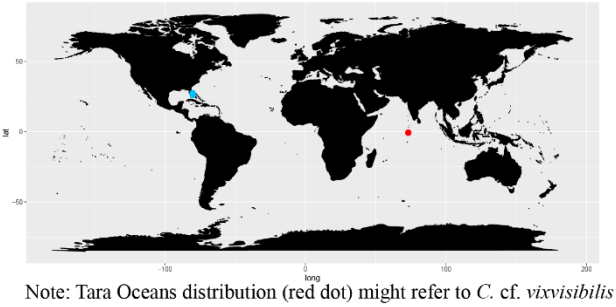
C. sp. Clade Na17B2



C. sp. Clade Na26B1

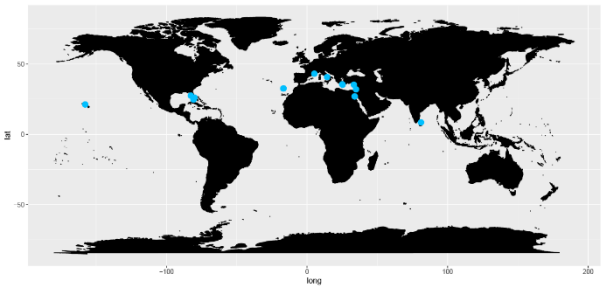


C. sp. Clade Na28A1

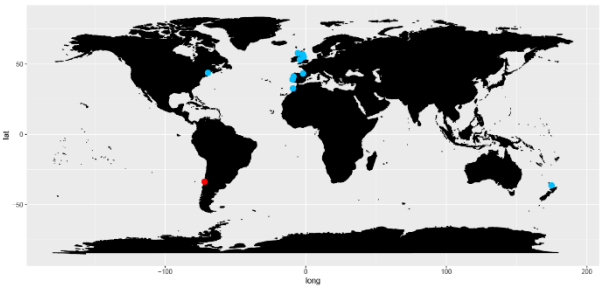


Note: Tara Oceans distribution (red dot) might refer to *C.* cf. *vixvisibilis* due to identical reference sequences in Tara Oceans dataset.

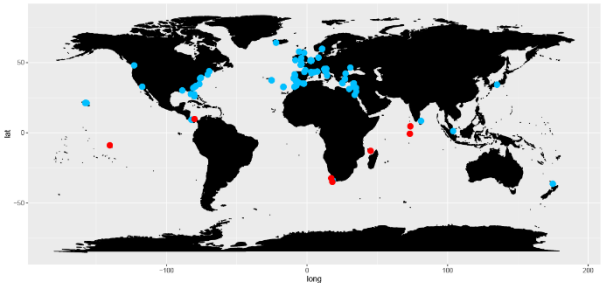
C. sp. Clade VA7D2



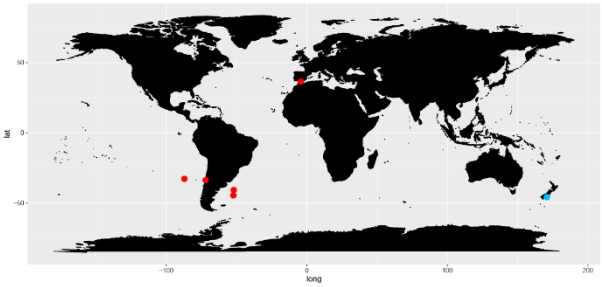
C. sporotruncatus



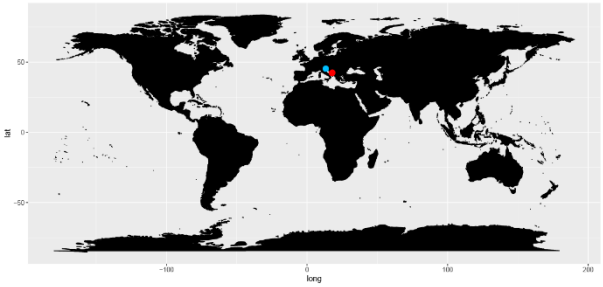
C. tenuissimus



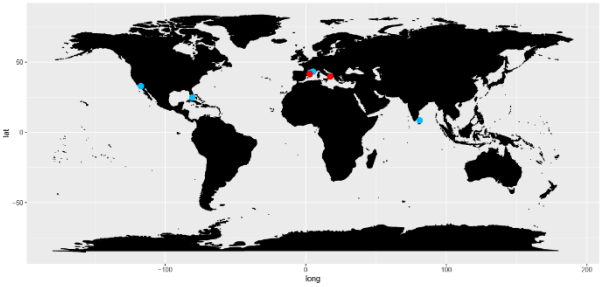
C. teres



C. throndsenii



C. tortissimus



For the Supplementary Tables cited in this Chapter, I remind to the online material published alongside the paper (https://peerj.com/articles/7410/#supplemental-information) as follows:

Reference in the Chapter	**Corresponding reference in the paper**
Table A3.1 - List of taxonomically accepted *Chaetoceros* taxa according to AlgaeBase	DOI: 10.7717/peerj.7410/supp-5
Table A3.2 - List of datasets included in GBIF and OBIS platforms	DOI: 10.7717/peerj.7410/supp-6
Table A3.3 - List of bibliographic references reporting *Chaetoceros* at given localities	DOI: 10.7717/peerj.7410/supp-7
Table A3.4 - Clustering test at 100 and 99% of similarity for V4 and V9 fragments	DOI: 10.7717/peerj.7410/supp-8
Table A3.5 - List of OSD and Tara Oceans stations	DOI: 10.7717/peerj.7410/supp-9

Chapter IV

Resolving the Chaetoceros curvisetus *cryptic species complex*

4.1. Introduction

4.1.1. Cryptic species complexes: origin, distribution and methodology of study

Cryptic species is a collective term generally used to indicate taxa that are morphologically indistinguishable to the observer but for which there is evidence (genetic, ecological, behavioural, etc.) of belonging to different evolutionary lineages (Mayr, 1970; Bickford et al., 2007). When many virtually identical species are involved, these groups of organisms are commonly referred to as "cryptic species complexes". Cryptic species may originate from recent divergence during speciation process (Fišer et al., 2018), which results in the lack of substantial morphological differences, or may be phylogenetically old and reproductively isolated from each other by strong biological barriers (Trontelj et al., 2009). In some cases, cryptic species can be phylogenetically unrelated and resulting from mimicry and convergence (Struck et al., 2018). In any case, they are real biological entities that have been inaccurately identified by taxonomists.

Increasing knowledge showed that cryptic species occur on all the branches of the tree of life and biogeographic regions (Pfenninger and Schwenk, 2007; Trontelj and Fišer, 2009). Furthermore, the frequency with which they are discovered using DNA sequence data calls for the integration of such methods in the process of species discovery and description (Bickford et al., 2007). The study of cryptic species has been approached in different ways, e.g. inferring phylogenies (e.g. Andrews et al., 2016), using species delimitation methods (e.g. Jörger et al., 2012; Crawford et al., 2013; Mills et al., 2017) and integrative taxonomy approaches (e.g. Gomes et al., 2015; Papakostas et al., 2016; Steiner et al., 2018). In general, all these approaches rely on the information gathered from Sanger sequencing of selected genes from a few sampled specimens (Lukhtanov et al., 2015; Saitoh et al., 2015; Iftikhar et al., 2016) and following inferred trees, to which morphological examinations can be added (e.g. integrative taxonomy approach). If genetic distances are large enough to justify an independent evolutionary lineage (e.g. a separate branch in a phylogenetic tree),

the occurrence of cryptic species is hypothesised. However, none of these approaches is free of pitfalls. First, phylogenetic trees may not be the best tool to visualise putative cryptic species, since they are well suited for representing evolutionary histories resulting from bifurcating speciation events and vertical changes within an ancestor-descent lineage (Huson et al., 2010). In the case of cryptic species, especially if they are the product of recent divergence, there could still be ongoing gene flow, which is better modelled by networks rather than phylogenetic trees. Second, it is often difficult to have a picture of the geographic variability of a species across its distributional area and what is indicated as putative cryptic species could also be a geographically isolated population that is undertaking a different evolutionary history.

Metabarcoding data from environmental samples have proven to be a powerful tool of non-invasive biodiversity assessment from species to community level (Cristescu, 2014; Deiner et al., 2017) and, in more recent times, a new source of biological records (Lawson Handley, 2015). They provide not only a bulk of sequence data for a gene region of interest of all the community sampled, but also information about their relative abundance, genetic variability and distribution. However, their application in ecology and evolution is still largely unexplored, especially as tool for inferring phylogenetic and phylogeographic relationships among taxa.

4.1.2. The Chaetoceros curvisetus *species complex*

The *Chaetoceros curvisetus* species complex currently includes several morphologically similar species that all share the straightforward characteristic of having the setae directed toward the outside of the chain spiral (Hasle and Syvertsen, 1996). Under light microscopy, some morphological features can distinguish among them, as the size of aperture between sibling cells (Kooistra et al., 2010), large and elliptical or nearly circular

in *curvisetus* (Fig. 4.1A) and narrow and oval in *pseudocurvisetus* (Fig. 4.1B). This is the visible effect of a very different valve morphology and a different type of cell junction.

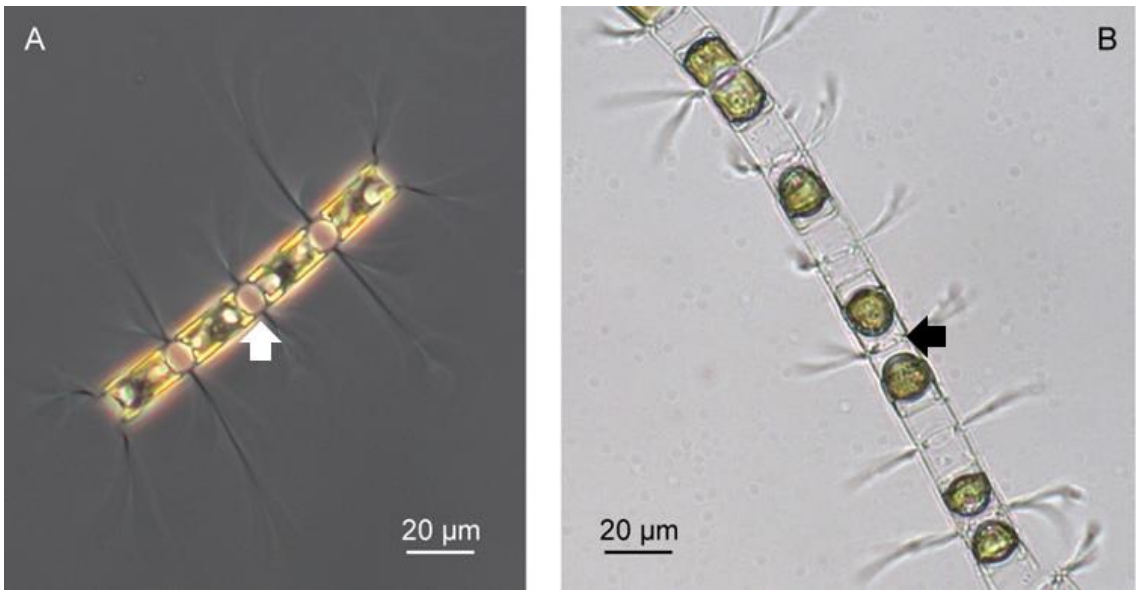


Fig. 4.1. *Chaetoceros curvisetus* (A) and *C. pseudocurvisetus* (B). The size and shape of aperture between sibling cells (see arrows) are useful characters for distinguishing these taxa.

All the species included in *Chaetoceros curvisetus* species complex form a monophyletic group, the section *Curviseta* (see Chapter II). To date, the only species that have been formally described are *Chaetoceros curvisetus* Cleve and *C. pseudocurvisetus* Mangin. A first molecular analysis using the hypervariable region (D1-D4) of the LSU rDNA gene revealed the occurrence of two distinct genetic clusters within *C. curvisetus* (Kooistra et al., 2010). A second screening, including more strains and sequences of LSU and SSU rDNA genes (Gaonkar et al., 2018) raised the number of genetic clusters in "curvisetus" to three. Furthermore, both studies highlighted the seemingly paraphyletic status of *C. curvisetus* due to a closer phylogenetic relationship among some "curvisetus" species to *pseudocurvisetus* strains than to other conspecifics. According to Gran (1897), *C. curvisetus* can be found throughout the year in the Atlantic Ocean and the Baltic Sea, but is especially abundant in summer and autumn. Hasle and Syvertsen (1996) indicated *C.*

curvisetus as a cosmopolitan mainly found in temperate and warm waters and *C. pseudocurvisetus* as an inhabitant of warm waters. My results of Chapter III (see Fig. A3.1) have shown that most of the *C. curvisetus* spp. have apparently no specific distribution restricted to particular habitats, with the exception of *C. curvisetus* 1, which was mostly found in cold-temperate waters. *C. pseudocurvisetus* was found in the warm waters of Indian coasts and Indonesia as well as in the Mediterranean Sea and nearby Atlantic Ocean (Fig. A3.1, Chapter III).

4.1.3. Objectives of the study

In this chapter, I use an 18S reference library of *C. curvisetus* species and close out-group taxa and map it against two global metabarcoding datasets: Tara Oceans and The Ocean Sampling Day 2014. The resulting data are used to generate a phylogenetic network in order to: 1) infer the number of the species within the complex; 2) explore the evolutionary relationships and the presence of gene flow among the members of the complex. Furthermore, I assess the distribution of the complex according to OSD and Tara Oceans data as well as the occurrence and abundance of each species delimited from the networks in Longhurst's biogeographic provinces (Longhurst, 2007).

I also explore the relative impact of sequence variability introduced by PCR and sequencing artefacts on the one hand and inter- and intraspecific variability on the other hand in metabarcoding data as well as the utility of using genetic distances to set boundaries across taxa.

4.2. Materials and Methods

4.2.1. Download and processing of metabarcoding data

To assess the phylogenetic relationships among members of the *C. curvisetus* species complex on a global scale, I used the V4-18S metabarcoding data from OSD and the V9-

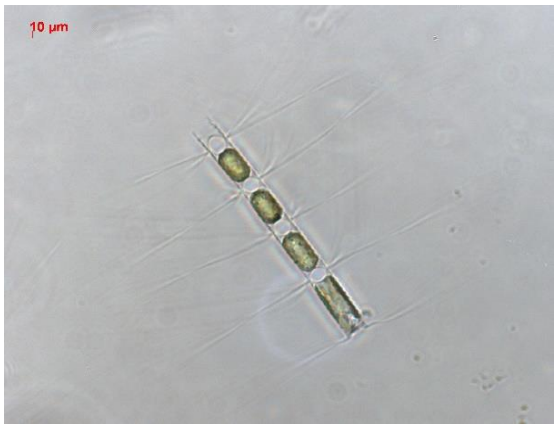
18S ones from Tara Oceans. OSD data were downloaded from https://mb3is.megx.net/osd-files?path=/2014/datasets/workable, whilst Tara Oceans data (De Vargas et al., 2017; Ibarbalz et al., 2019) from https://doi.pangaea.de/10.1594/PANGAEA.873277 and ENA website at acc. numb. PRJEB6610. For the OSD dataset, I pooled together the 144 workable fasta files from each sampling site and generated a total fasta file with the unique sequences and a table containing their abundance across the sites (Total OSD abundance table) using mothur v1.41.1 (Schloss et al., 2009). For the Tara Oceans dataset, I directly extracted a total unique fasta file and a Total Tara Oceans abundance table from the downloaded file containing sequences from 210 sampling sites.

To retrieve sequences of *C. curvisetus*-like taxa from these metabarcoding data, I started from the full length 18S rDNA sequences of *C. curvisetus* and *C. pseudocurvisetus* species and close outgroups (*C. tortissimus* and *C.* cf. *tortissimus*, Table 4.1) provided in Gaonkar et al. (2018) and used in Chapter II for phylogenetic inference. For V4 region, further barcodes were retrieved from NCBI, in particular for *C. curvisetus* (strain SKLMP_YG033, acc. numb. MG821562) and *C. pseudocurvisetus* (strain IRB, acc. numb. MG385841). In this chapter, numbers after *C. curvisetus* species' names (1, 2, 2c, 3 and 3e) refer to genetically defined species for which a formal description is not available yet, but that are discussed in Gaonkar et al. (2018) or in this thesis (e.g. Chapter II). Light miscroscopy photographies of these species are provided in Fig. 4.2. The wording "sp." followed by number (1, 2, 3 and 4) refers to hypothetical new species here identified.

Table 4.1. List of reference sequences utilised for gathering *C. curvisetus*-like taxa.

Taxon	**Strain**	**Accession Number**	**Reference for V4**	**Reference for V9**
C. curvisetus	SKLMP YG033	MG821562	yes	no
C. curvisetus 1	Na10C1	MG972232	yes	yes
C. curvisetus 2	Na1C1	MG972235	yes	yes

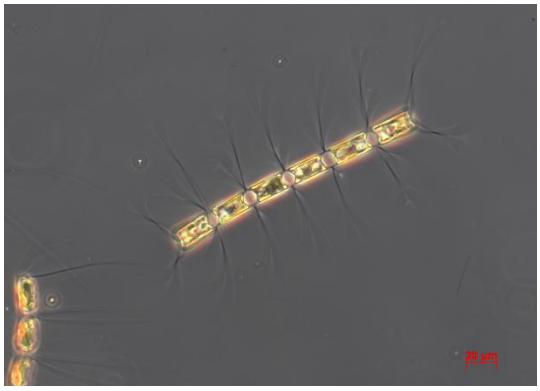
C. curvisetus 2c	El6A2	LC466961	yes	yes
C. curvisetus 3	newBB2	MG972241	yes	yes
C. curvisetus 3e	El4A2	LC466962	yes	yes
C. pseudocurvisetus	IRB	MG385841	yes	no
C. pseudocurvisetus	Na13C4	MG972304	yes	yes
C. cf. *tortissimus*	Na18C4	MG972275	yes	yes
C. tortissimus	Na25A2	MG972325	yes	yes



C. curvisetus 1

C. curvisetus 2

C. curvisetus 2c



C. curvisetus 3

C. curvisetus 3e

Fig. 4.2. Light microscopy photographies of the known members of the *C. curvisetus* species complex.

I extracted from the 18S region the V4 and V9 regions corresponding with the fragment amplified by the primers used in OSD and Tara Oceans. These fragments were clustered at several thresholds (100-99%) to ensure that different *C. curvisetus* species were not collapsed together (see Chapter III). The reference fragments were used as queries for a local BLAST to recover entries at 95% of similarity against the OSD and Tara Oceans datasets. The combined strategy of using both reference barcodes of close outgroups and a relaxed threshold (95%) allowed gathering in the metabarcoding datasets sequences of *C. curvisetus* like taxa for which reference barcodes could be unavailable.

The metabarcodes extracted were aligned with the references, including outgroup taxa, using MAFFT online (Katoh et al., 2017) and a phylogenetic tree was built in FastTree v2.1.8 (Price et al., 2010), using the GTR model. The resulting tree was visualised and modified in Archaeopteryx v0.9901 (Han and Zmasek, 2009), in order to remove false positive sequences clustering within outgroup clades and gather only *curvisetus*-like metabarcodes. This procedure was followed separately for V4 and V9 fragments. The sequences filtered through the previous procedure, were considered validated as *C. curvisetus*-like. The abundance and distribution of V4 and V9 *curvisetus* metabarcodes were extracted from the Total OSD and Tara Oceans abundance tables. At the end of the validation procedures, I generated four files: 1) the V4_OSD_curvi_validated.fasta file,

containing the sequences validated as *C. curvisetus* from OSD; 2) the V4_OSD_curvi_validated.count_table file, containing the distribution of each haplotype across the OSD sites; 3) the V9_TARA_curvi_validated.fasta file, containing the sequences validated as *C. curvisetus* from Tara Oceans; 4) the V9_TARA_curvi_validated.count_table file, containing the distribution of each haplotype across the Tara Oceans sites.

4.2.2. Phylogenetic haplotype network inference

Phylogenetic haplotype networks were used to circumscribe species within a species-complex in a non-dichotomous approach. For such inference, I used the statistical parsimony algorithm by Templeton et al. (1992) implemented in TCS network (Clement et al., 2002). This agglomerative algorithm collapses sequences in haplotypes and estimates the number of differences among them due to single substitutions and with a 95% statistical confidence (parsimony limit). Then, haplotypes (nodes in the network) are progressively connected among them by edges starting from the ones that differ by one change, then by two, three and so on until all the haplotypes have been connected into a single network or the parsimony limit has been reached. This kind of phylogenetic network was preferred over others (e.g. median joining networks, MJ) because it shows reticulations and includes unobserved haplotypes in the network as the MJ network but is computationally quicker.

Despite displaying the final output as networks, TCS approach differs from Swarm (Mahé et al., 2015). Swarm is a de novo clustering method that uses a clustering threshold (d) of nucleotide difference (a substitution, insertion, or deletion), whilst TCS works on a multi-alignment. Moreover, edges in Swarm networks carry no phylogenetic information, but are only a representation of the parameter d used, so it is not possible to infer relationships among OTUs as in TCS networks. Since I was interested in assessing the internal structure

(phylogenetic relationships) of my species complex and not only in the assessment of OTUs, I have chosen the TCS method over Swarm.

TCS algorithm was inferred and visualised as implemented within PopART v1.7 (Leigh and Bryant, 2015). Abundance of sequences was included in the inference. Each network was exported as table and nexus file.

Using the information contained in the table of haplotype of each TCS network, I delineated species following these criteria: 1) the sequences found within a node including the reference barcode were attributed to that species; 2) the sequences having mutations ≤ 2 in respect to the node with the reference and with abundance ≤ 3 were attributed to the that node; 3) nodes without reference and with mutations > 2 and abundance >3 respect to the ones with reference were considered as hypothetical new taxa. The latter were indicated as *C. curvisetus* sp. 1, sp. 2, etc.

After species inference, I took the representative sequence of each delimited species and inferred a phylogenetic tree (for V4 and V9 regions) for a rapid and supported visualisation of phylogenetic relationships among taxa. Maximum Likelihood (ML) trees were inferred using IQ-TREE v1.6.8 (Nguyen et al., 2014) under the TN+F+G4 model for V4 and the K2P+G4 model for V9 (suggested by ModelFinder, Kalyaanamoorthy et al., 2017) and 1000 bootstrap replicates for both datasets. The sequences of *C. tortissimus* and *C.* cf. *tortissimus* were used as outgroup.

4.2.3. Genetic divergence among species and variability within species

To quantify the relatedness of each species in terms of distances rather than number of mutations, I calculated the net genetic distances between pairs of species as implemented in MEGA6 (Tamura et al., 2013):

$dA = dXY – (dX + dY)/2,$

where dXY is the average distance between groups X and Y, and dX and dY are the mean within-group distances.

I used the Jukes-Cantor (JK) model of sequence evolution (Jukes and Cantor, 1969) to calculate the genetic distances across all metabarcodes of each species, which best fitted our data. I also calculated, using the same model, the minimum, maximum and average evolutionary divergence of sequences within nodes (the number of base substitutions per site from averaging over all sequence pairs within each group) using MEGA6 (Tamura et al., 2013). The presence of barcoding gap in the inferred species was explored. The barcoding gap was considered to occur if the maximum distance within species was lower than the minimum distance between species (Meyer and Paulay, 2005).

4.2.4. Global distribution of taxa belonging to the C. curvisetus *species complex*

I mapped the distribution of the members of the *C. curvisetus* species complex in world's oceans using the previously inferred species. First, from the abundance tables previously generated (V4_OSD_curvi_validated.count_table and V9_TARA_curvi_validated.count_table files), I summed the abundances of the haplotypes belonging to the same inferred species. Then, I plotted the occurrence of all *C. curvisetus* inferred species together on a world map divided in Longhurst's provinces. I also plotted the abundance of each species (in terms of reads) in Longhurst's provinces in the form of heatmaps.

To plot the occurrences, I downloaded the shapefiles containing the coordinates of Longhurst provinces (Longhurst, 2007) from the Marine Regions portal (http://www.marineregions.org/downloads.php#longhurst) and plotted them using the R package *rgdal* (Bivand et al., 2018) and the function ssplot in the daughter process "sp" (Pebesma and Bivand, 2005; Bivand et al., 2013). For abundances, I used the R (R Core

Team, 2019) working packages *phyloseq* (McMurdie and Holmes, 2013) and *ggplot2* (Wickham, 2016).

4.3. Results

After the extraction of the V4 and V9 regions, corresponding to the fragments amplified by the primers utilised in OSD and Tara Oceans datasets, from partial or full-length 18S sequences, I obtained 10 reference sequences for V4 and 8 for V9 (Table 4.1). Such difference was because two 18S sequences (*C. curvisetus* strain SKLMP YG033 and *C. pseudocurvisetus* strain IRB) did not cover the V9 region too.

4.3.1. Validation of C. curvisetus *candidate sequences*

Regarding the OSD dataset, following BLAST analysis I retrieved 4,223 sequences corresponding to 1,428 unique haplotypes including outgroups. After the validation of metabarcodes belonging to the *C. curvisetus* species complex by means of the phylogenetic tree-approach, I gathered 1,232 haplotypes, for a total of 3,804 sequences. Regarding Tara Oceans data, BLAST analysis returned 856,967 sequences corresponding to 2,247 unique haplotypes including outgroups. After validation, I eventually retrieved 68,210 sequences for 772 haplotypes belonging to the complex.

Metabarcodes validated as *C. curvisetus* (1,232 for OSD and 772 for Tara Oceans) were found in 60 out of 144 OSD sampling sites (41.7%) and 117 out of 210 Tara Oceans stations (55.7%) (Fig. 4.3, Table A4.1 in Appendix IV).

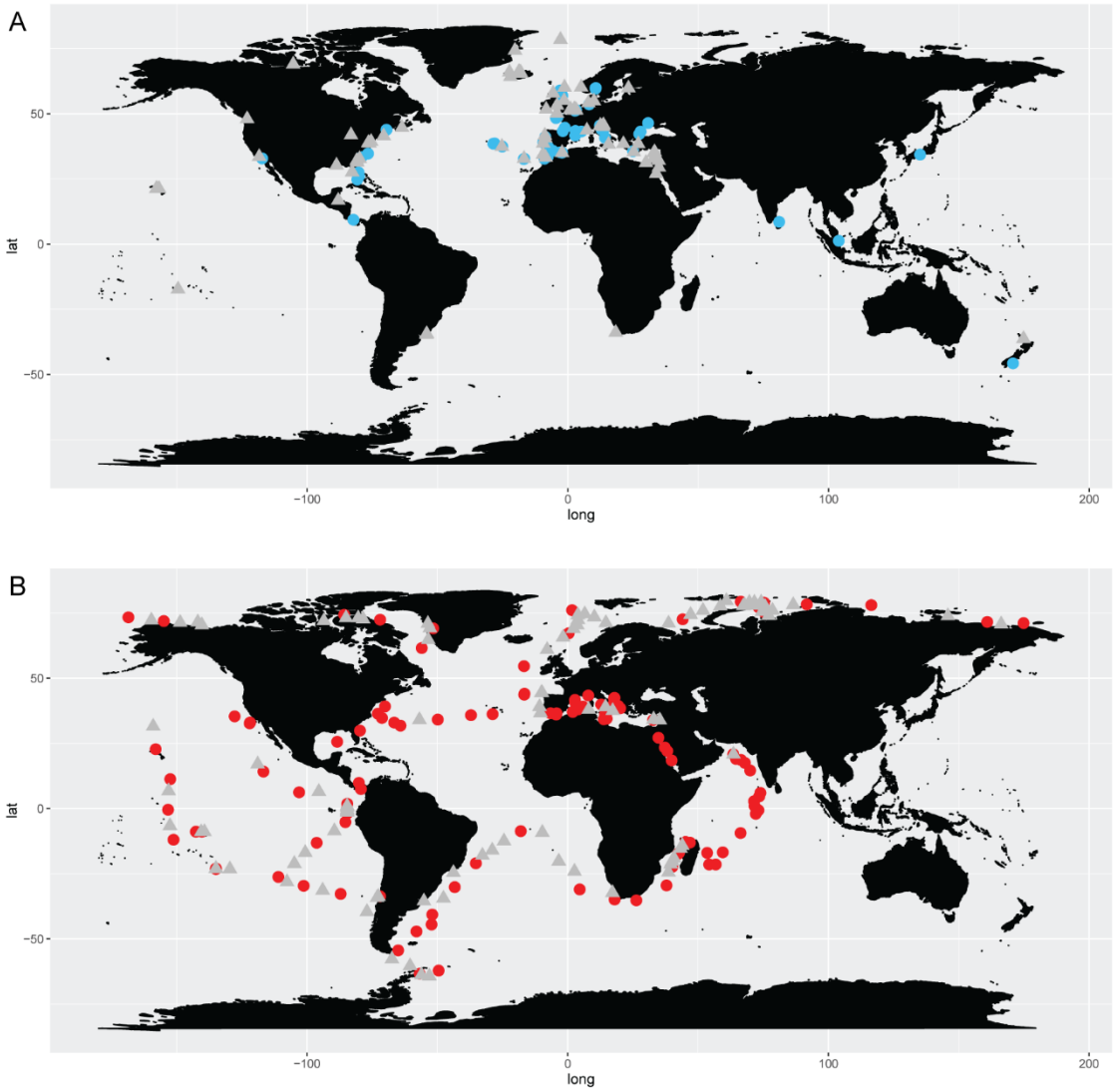


Fig. 4.3. Occurrence of taxa belonging to the *C. curvisetus* species complex in OSD (A) and Tara Oceans (B) datasets. Blue dots refer to occurrence in OSD data, whilst red dots in Tara Oceans data. Grey triangles indicate absence in the respective sampling site.

4.3.2. Phylogenetic haplotype networks

The haplotype network based on the OSD dataset (V4 region) contained seven nodes assigned to known species in the *C. curvisetus* complex plus two without a reference (Fig. 4.4). Most of the metabarcodes were assigned to *C. curvisetus* 1, 2 and 3, *C. curvisetus* strain SKLMP_YG033 and *C. pseudocurvisetus* (Fig. 4.4). No sequences were found for the barcode *C. curvisetus* 3e (El4A2) and only one for *C. curvisetus* 2c, both from the Red Sea. Many sequences clustered into two closely related nodes lacking barcodes (Fig. 4.4).

Moreover, the species *C. curvisetus* 3 is more closely related to *C. pseudocurvisetus* than other "curvisetus" species; *C. curvisetus* 3e (from the Red Sea) is closely related to *C. pseudocurvisetus* (two base changes) and distantly to *C. curvisetus* 1 (at least 12 mutations from the main edge). This latter node is separated by eight base changes from the other one referring to *C. curvisetus* strain SKLMP_YG033 from Hong Kong.

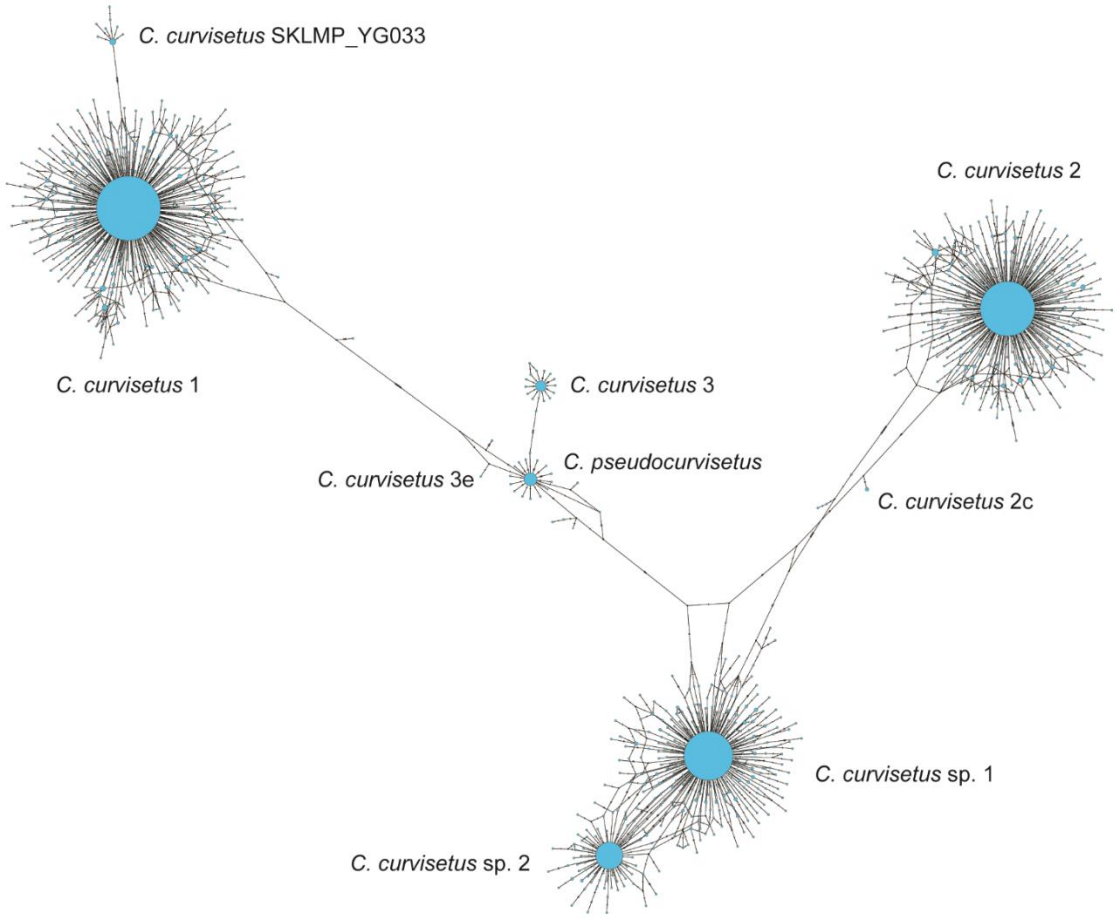


Fig. 4.4. TCS haplotype network for the *C. curvisetus* species complex according to OSD data. The size of the nodes refers to the abundance of the reads. Numbers or codes after *C. curvisetus* species' names refer to genetically and morphologically defined species within the *C. curvisetus* complex for which references are available (see Gaonkar et al., 2018). *C. curvisetus* sp. 1 and 2 refer to species in the *C. curvisetus* complex for which no reference sequences are available yet.

The haplotype network based on the Tara Oceans dataset (V9 region) contained six nodes assigned to a known *curvisetus* species plus two without a reference (Fig. 4.5). Most of the metabarcodes were assigned to *C. curvisetus* 2, followed by *C. curvisetus* 3, *C. curvisetus*

3e and all the others with comparable abundances. For both species *C. curvisetus* 3 and *C. curvisetus* 3e, a peripheral node with considerable abundance separating from the main one was observed and treated separately for further analyses (named *C. curvisetus* sp. 3 and sp. 4 respectively). Strains isolated from the Red Sea (*C. curvisetus* 2c and *C. curvisetus* 3e) were highly represented in terms of sequences in the Tara dataset when compared with the one of OSD.

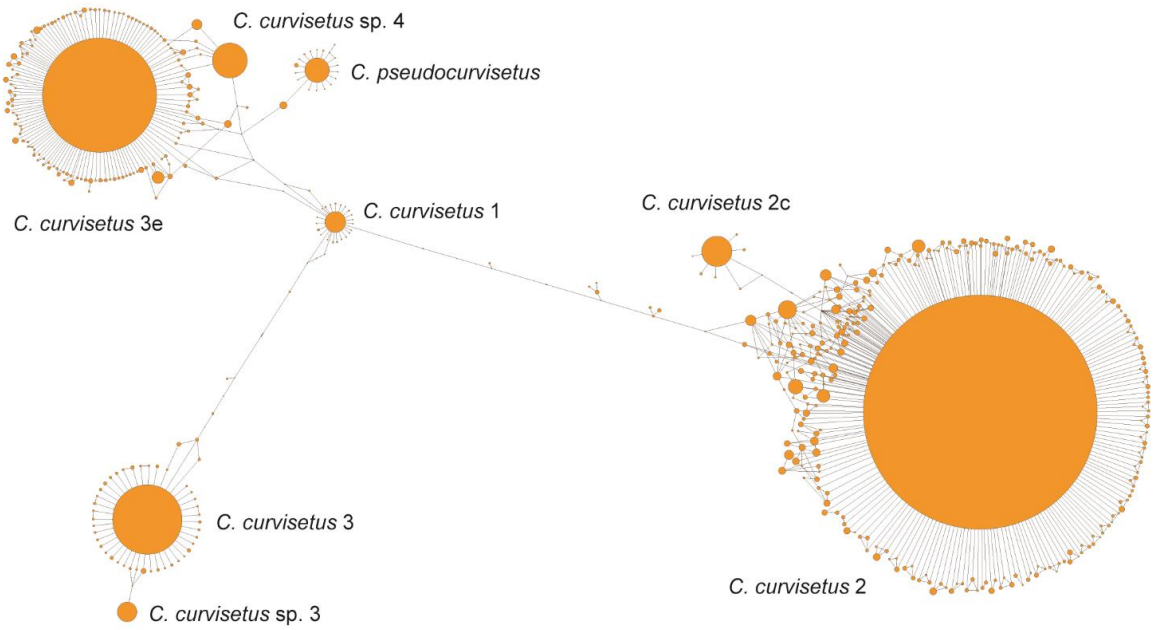


Fig. 4.5. TCS haplotype network for the *C. curvisetus* species complex according to Tara Oceans data. The size of the nodes refers to the abundance of the reads. Numbers or codes after *C. curvisetus* species' names refer to genetically and morphologically defined species within the *C. curvisetus* complex for which references are available (see Gaonkar et al., 2018). *C. curvisetus* sp. 3 and 4 refer to species in the *C. curvisetus* complex for which no reference sequences are available yet.

The comparison of V4 and V9 networks showed minor differences. In the former, the group encompassing the species *C. curvisetus* 3, *C. curvisetus* 3e and *C. pseudocurvisetus* acted as a bridge between *C. curvisetus* 1 and *C. curvisetus* 2 and the unassigned nodes, whilst *C. curvisetus* 2c did the same for *C. curvisetus* 2 and the two unassigned nodes. In the latter, the node attributed to *C. curvisetus* 1 was the pivot around which *C. curvisetus* 3,

the groups *C. curvisetus* 2 - *C. curvisetus* 2c and *C. curvisetus* 3e – *C. pseudocurvisetus* were collocated. In both networks, the relationships among main nodes were generally simple (without complex reticulations), indicating substantial lack of gene flow.

The Maximum likelihood tree inferred using the V4 representative sequences of each newly identified putative species plus the references confirmed that the taxa without reference barcodes, here indicated as *C. curvisetus* sp. 1 and sp. 2, are likely to constitute at least one new species (Fig. 4.6). These are closely related and share a common ancestor with *C. curvisetus* group 2 (Fig. 4.6).

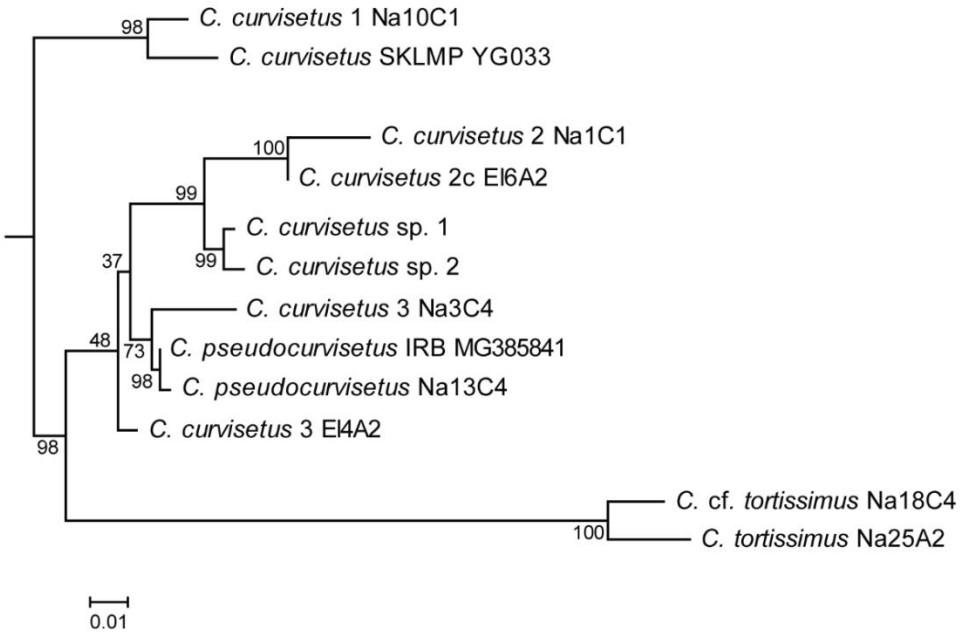


Fig. 4.6. Maximum Likelihood tree of the *C. curvisetus* species complex based on representative sequences of V4 data. Numbers at the basis of nodes indicate the support to branches after 1000 bootstrap replicates.

For the V9 tree, due to the shortness of the fragment, it was difficult to make hypotheses on the nature of the newly discovered taxa as well as about their phylogenetic relationships with other *curvisetus* species (Fig. 4.7). However, *C. curvisetus* sp. 3 seemed to be more differentiated to its sister taxon (*C. curvisetus* 3, 100 BS) than *C. curvisetus* sp. 4 to its own (*C. curvisetus* 3e, 61 BS, Fig. 4.7).

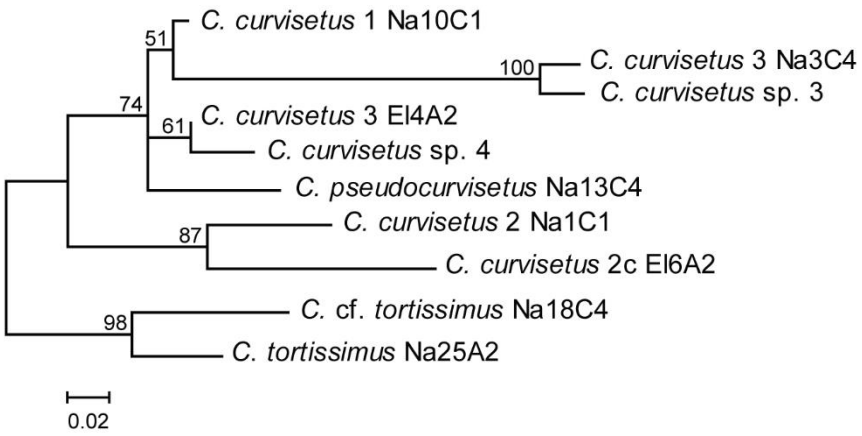


Fig. 4.7. Maximum Likelihood tree of the *C. curvisetus* species complex based on representative sequences of V9 data. Numbers at the basis of nodes indicate the support to branches after 1000 bootstrap replicates.

4.3.3. Genetic differentiation and variability

Genetic distances across inferred species for V4 and V9 regions were different in terms of absolute values, but the proportions were comparable (Table 4.2). For V4, the lowest interspecies genetic distance values were between *C. curvisetus* 3e and *C. pseudocurvisetus* (0.007) and *C. curvisetus* sp. 1 and sp. 2 (0.008, Table 4.2A), whilst the highest between *C. curvisetus* 1 and 2 (0.107) and *C. curvisetus* 2 and *C. curvisetus* strain SKLMP_YG033 (0.105) (Table 4.2A). For V9, values ranged from 0.368 (genetic distance between *C. curvisetus* 2c and 3) to 0.022 (*C. curvisetus* 3e and sp. 4) (Table 4.2B). For both V4 and V9 regions, the highest value of intraspecific divergence (0.105 and 0.049 respectively) was not lower than the minimum value of interspecific divergence (0.007 and 0.022 respectively). The lowest interspecies distances were lower than maxima intraspecies ones. Therefore, no threshold value was found within the complex to distinguish between inter- and intra-specific variability (barcoding gap).

Table 4.2. Pair-wise genetic differentiation between *C. curvisetus* species in OSD (A) and Tara Oceans (B) datasets. Genetic distances were calculated using the Jukes-Cantor model.

(A)

	***C. curvisetus* 1**	***C. curvisetus* 2**	***C. curvisetus* 2c**	***C. curvisetus* 3**	***C. curvisetus* 3e**	***C. curvisetus* SKLMP_YG033**	***C. pseudocurvisetus***	***C. curvisetus* sp. 1**	***C. curvisetus* sp. 2**
***C. curvisetus* 1**	-								
***C. curvisetus* 2**	0.107	-							
***C. curvisetus* 2c**	0.098	0.024	-						
***C. curvisetus* 3**	0.083	0.072	0.069	-					
***C. curvisetus* 3e**	0.054	0.034	0.022	0.025	-				
***C. curvisetus* SKLMP_YG033**	0.018	0.105	0.094	0.086	0.054	-			
C. pseudocurvisetus	0.062	0.063	0.049	0.023	0.007	0.063	-		
***C. curvisetus* sp. 1**	0.083	0.043	0.030	0.054	0.014	0.084	0.034	-	
***C. curvisetus* sp. 2**	0.085	0.046	0.032	0.057	0.018	0.086	0.037	0.008	-

(B)

	***C. curvisetus* 1**	***C. curvisetus* 2**	***C. curvisetus* 2c**	***C. curvisetus* 3**	***C. curvisetus* sp. 3**	***C. curvisetus* 3e**	***C. curvisetus* sp. 4**	***C. pseudocurvisetus***
***C. curvisetus* 1**	-							
***C. curvisetus* 2**	0.155	-						
***C. curvisetus* 2c**	0.189	0.134	-					
***C. curvisetus* 3**	0.181	0.241	0.368	-				
***C. curvisetus* sp. 3**	0.181	0.229	0.354	0.038	-			
***C. curvisetus* 3e**	0.038	0.179	0.154	0.204	0.204	-		
***C. curvisetus* sp. 4**	0.062	0.204	0.165	0.210	0.237	0.022	-	
C. pseudocurvisetus	0.079	0.203	0.191	0.204	0.230	0.079	0.073	-

Within each species, the mean evolutionary divergence over sequence pairs ranged from 0.000 (*C. curvisetus* 2c) to 0.055 (*C. curvisetus* 3e) for V4 region and from 0.000 (*C. curvisetus* sp. 3) to 0.017 (*C. curvisetus* 2) for V9 (Table 4.3).

Table 4.3. Average evolutionary divergence over sequence pairs within species. The number of base substitutions per site from averaging over all sequence pairs within each group are shown. Analyses were conducted using the Jukes-Cantor model. Dashes refer to species absent in that dataset.

	Divergence V4 region			**Divergence V9 region**		
Species	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
C. curvisetus 1	0.009	0.000	0.054	0.013	0.000	0.029
C. curvisetus 2	0.008	0.000	0.035	0.017	0.000	0.049
C. curvisetus 2c	0.000	0.000	0.000	0.012	0.000	0.029
C. curvisetus 3	0.007	0.000	0.021	0.013	0.000	0.029
C. curvisetus 3e	0.010	0.003	0.016	0.016	0.000	0.039
C. curvisetus SKLMP_YG033	0.013	0.003	0.105	-	-	-
C. curvisetus sp. 1	0.008	0.000	0.027	-	-	-
C. curvisetus sp. 2	0.007	0.000	0.098	-	-	-
C. curvisetus sp. 3	-	-	-	0.000	0.000	0.000
C. curvisetus sp. 4	-	-	-	0.016	0.000	0.039
C. pseudocurvisetus	0.010	0.003	0.035	0.014	0.000	0.029

4.3.4. Global distribution of taxa belonging to the C. curvisetus *species complex*

The plotting of occurrence data gathered from OSD and Tara Oceans metabarcoding data revealed that the species complex is cosmopolitan, and occurs in both coastal and open ocean waters at all latitudes from northern to southern hemisphere (Fig. 4.8A and B).

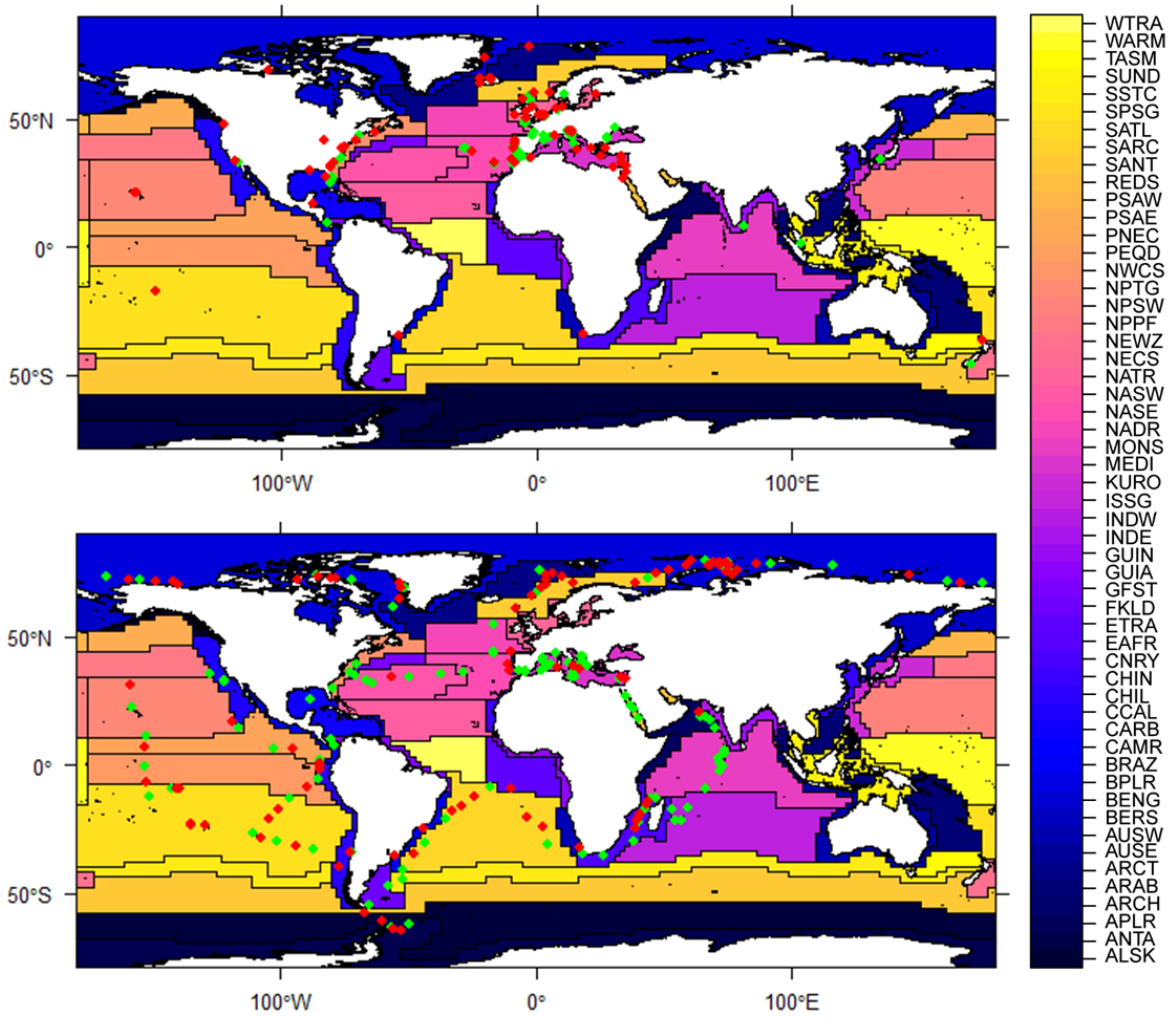


Fig. 4.8. Distribution of the *C. curvisetus* species complex in Longhurst provinces. (A) OSD data; (B) Tara Oceans data. Green dots indicate presence of the taxa in that station, whilst red dots indicate absence.

However, some species showed a specific pattern of occurrence and abundance across the different datasets. For instance, *C. curvisetus* 1 was found and revealed to be mostly abundant in polar (ARCT, BPLR) and temperate provinces (NADR, NECS, SSTC), whilst *C. curvisetus* 2 a typical generalist species (Fig. 4.9 and Fig. 4.10). Some species were rare in some datasets (e.g. *C. curvisetus* 2c in OSD) and completely absent in others (e.g. *C. curvisetus* sp. 1 and sp. 2 in Tara Oceans).

In the specific case of closely related taxa (e.g. *C. curvisetus* 1 - *C. curvisetus* strain SKLMP YG033 and *C. curvisetus* sp. 1 and sp. 2 in OSD; *C. curvisetus* 3 – *C. curvisetus* sp. 3 and *C. curvisetus* 3e – *C. curvisetus* sp. 4 in Tara Oceans), a peculiar occurrence

pattern was observed. In each couple, if the two taxa were sharply separated in the network (e.g. *C. curvisetus* 1 - *C. curvisetus* strain SKLMP YG033), in the heatmap I observed one of these occupying different provinces with opposite and generally of different environmental characteristics (e.g. tropical vs. temperate).

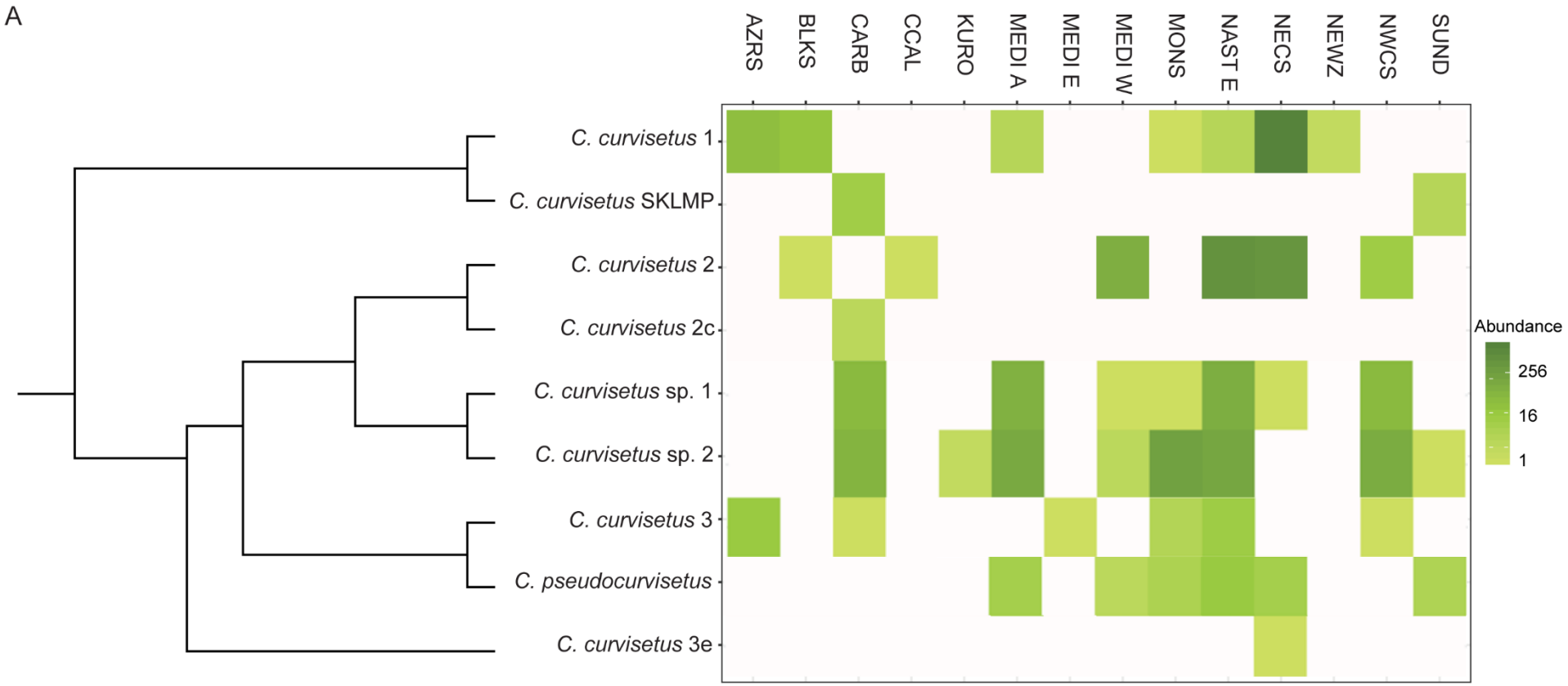


Fig. 4.9. Heatmap showing the abundance of *C. curvisetus* spp. in each Longhurst province according to OSD data. Abundance refers to the number of reads. Species are ordered according to phylogenetic position.

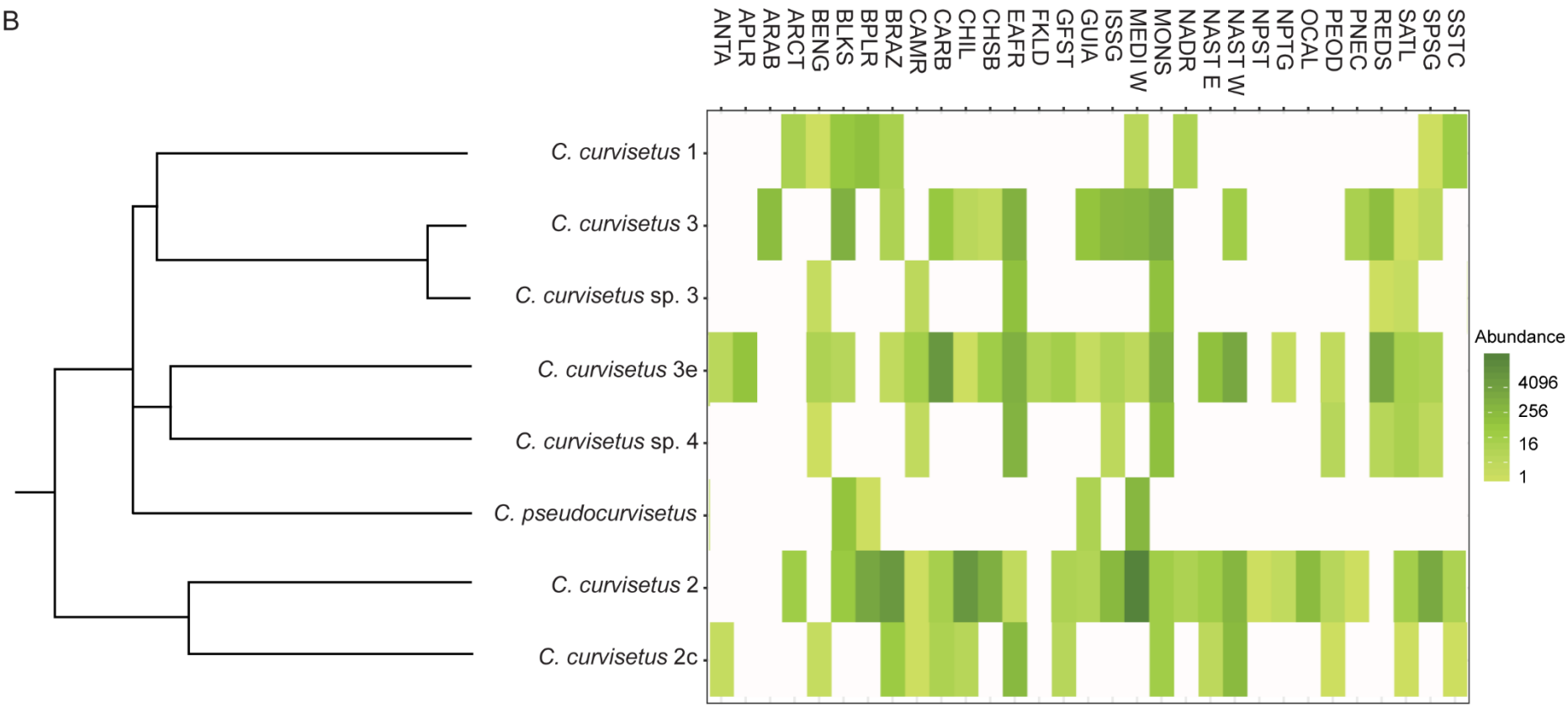


Fig. 4.10. Heatmap showing the abundance of *C. curvisetus* spp. in each Longhurst province according to Tara Oceans data. Abundance refers to the number of reads. Species are ordered according to phylogenetic position.

4.4. Discussion

In this chapter, I showed how the study of a cryptic species complex can be enhanced by the combining classical evolutionary approaches and huge amounts of diversity information contained in global metabarcoding datasets. In particular, I used the *C. curvisetus* species complex as case study and combined evolutionary approaches (haplotype networks, phylogenetic relationships, and genetic distances) with the most comprehensive and complementary global metabarcoding datasets available, the predominantly coastal OSD and the mainly oceanic Tara Oceans. The results obtained show the enormous potential of the integration of such methodologies for phylo- and biogeographic studies.

4.4.1. Phylogenetic relationships among taxa belonging to the C. curvisetus *species complex*

The use of haplotype networks allows a clear assessment and visualisation of the relationships among the taxa within the complex, which are not straightforward in the V4 and V9 trees in Gaonkar et al. (2018). The latter were based on references of a few specimens per species whereas the haplotype networks and their relative abundances provide insights in the populations of those species. In addition, phylogenetic trees are constrained in visualising speciation events as bifurcating processes, whereas haplotype networks can model evolution in a reticulated manner, best fitting cases of recent divergence as may occur in species complexes. The V4 and V9 TCS networks were presented slight differences related to different length of the regions (~384 and ~105 bp respectively). These differences are also found in the V4 and V9 phylogenetic trees in Gaonkar et al. (2018). Overall, the signal was consistent between the two datasets, allowing the inference of at least eight different species within the *C. curvisetus* species complex.

The network approach revealed also to be useful at detecting putative new species or isolated populations. These findings are supported by the high bootstrap values recovered in the tree obtained using the reference barcodes of *curvisetus* species and the representatives of unassigned nodes. The same is found in Tara Oceans V9 data for two taxa (*C. curvisetus* sp. 3 and 4), despite the fact that the separation from the other nodes is not as straightforward as in the OSD V4 network. Using the same OSD and Tara Oceans datasets but different taxa and a non-evolutionary approach (swarm OTU clustering), Pargana (2017) found a new clade close to *Leptocylindrus danicus* and several clades within *L. minimus* of uncertain taxonomic identity. Such inference of taxa from signature sequences (metabarcoding data) is just the first step of the process; the next step is to link such anonymous sequences to a reference of a specific taxon in order to be validated. This approach is called "reverse taxonomy" (Markmann and Tautz, 2005). In the case of metabarcoding data, the validation of anonymous sequences in the field is favoured by the use of abundance tables, which contain the information of occurrence and abundance in each sampled locality.

In general, the shape and size of nodes in the network, together with the number and structure of edges connecting them, can be considered as a primary hypothesis for species/population delimitation based on gene flow. In my networks, the signal of active gene flow between inferred species is weak but present. The absence of barcoding gap confirmed that signal, suggesting that the genetic barriers in part of the complex are not complete.

4.4.2. Distribution of taxa belonging to the C. curvisetus *species complex*

Chaetoceros curvisetus was reported by Gran (1897) as a common inhabitant of the Atlantic Ocean and the Baltic Sea, with peaks of abundance in summer and autumn. Hasle and Syvertsen (1996) indicated it as a cosmopolitan species mainly found in temperate and

warm waters. This was confirmed by my results in chapter III. In Chinese waters, the only references about the distribution of such species are the ones related to harmful algal blooms (Wang and Wu, 2009; Zhen et al., 2009), during which the species is particulary abundant. However, no production of toxins is known to date in any "curvisetus" species.

Instead, Hasle and Syvertsen (1996) considered *C. pseudocurvisetus* as an inhabitant of warm waters. This finding was partially confirmed by results of my analysis in this chapter and chapter III, in which the species was found not only in the Mediterranean Sea, the nearby Atlantic Ocean and the Indian Ocean, but also in the North Sea.

In general, results of my analysis using OSD and Tara Oceans dataset indicates that the *C. curvisetus* complex is cosmopolitan. Nonetheless, some species showed preference for particular environmental conditions. For example, *C. curvisetus* 1 occurs in cold to temperate waters, with the exception of the Mediterranean Sea. In the Gulf of Naples (Mediterranean Sea), Gaonkar (2017) found this species only during winter, supporting its preferences for cold environments. Similarly, but with an opposite trend, *C. curvisetus* strain SKLMP is only found in tropical seas. This is also interesting from the phylogenetic point of view, since these two taxa are sister species. This marked difference in climate preference between closely related species was also observed for other members of the complex, e.g. *C. curvisetus* 1 - *C. curvisetus* SKLMP, *C. curvisetus* sp. 1 – sp. 2, *C. curvisetus* 3 - *C. curvisetus* sp. 3 and *C. curvisetus* 3e – *C. curvisetus* sp. 4. The aforementioned pattern was more evident for sister taxa that were clearly separated in the network (*C. curvisetus* 1 - *C. curvisetus* SKLMP) than in others where gene flow was still on-going or the separation was recent (*C. curvisetus* sp. 1 – sp. 2, *C. curvisetus* 3 - *C. curvisetus* sp. 3 and *C. curvisetus* 3e – *C. curvisetus* sp. 4).

Other studies involving cryptic species have shown similar results. In the genus *Skeletonema* for example, the widely distributed species *Skeletonema costatum* sensu lato revealed to be a complex of several species (Sarno et al., 2005; 2007; Zingone et al., 2005).

Several of these appeared to be widely distributed as well, but within some broad climatological boundaries (cool-temperate *S. japonicum*; temperate to tropical *S. tropicum*; Kooistra et al., 2008). However, a few others such as *S. grethae* appeared to be more regional and apparently absent in climatologically comparable regions (Kooistra et al., 2008). More in general, Hasle (1976) already noticed that morphologically closely related diatom species were often found in different biogeographic regions. In the genera *Nitzschia* and *Thalassiosira*, she observed species only from the cold-water species of the Northern and Southern Hemispheres as well as from warm-water species, and cosmopolitan ones (Hasle, 1976). In *Leptocylindrus*, most species were found to be widespread across coastal waters (e.g. *L. convexus*, *L. danicus* and *L. hargravesii*) with the only exception of *L. minimus*, which was restricted to cold waters of the Northern Hemisphere (Pargana, 2017).

According to the "everything is everywhere" hypothesis (Baas Becking, 1934), most microbes form populations large enough to migrate efficiently and accumulate mutations that could be beneficial in particular environments (Shapiro et al., 2016). Speciation in the microbial world is therefore expected to involve little drift and geographical separation and more selection (Shapiro et al., 2016). Diatoms, for example, are believed to exhibit high intraspecific variability, which would be key for their adaptation to different environments (Godhe and Rynearson, 2017). It is possible that different strains of a species already possess beneficial mutations allowing them to adapt to different environments due to high intraspecific variability (see Godhe and Rynearson, 2017). Once a different environment is reached, some strains would be favoured by natural selection and, over time, accumulate other mutations that will finally differentiate them from the parental population, leading to speciation. In this context, the adaptation to different environments would be the factor triggering speciation in diatoms. In agreement with Hasle (1976), which surveyed the biogeographic trends of 26 diatom species, in this study I have observed that sister *C. curvisetus* species (e.g. *C. curvisetus* 1 - *C. curvisetus* SKLMP; *C. curvisetus* sp. 1 – sp. 2)

tend to be found in different biogeographic provinces with generally opposite environmental conditions (e.g. cold vs. warm environments). Data are far from conclusive to assert that adaptation to different environmental conditions triggers speciation in diatoms, but I have added other elements to support this hypothesis. Furthermore, all these studies emphasise once more the importance of correct identification of taxa at the species level to make adequate inferences on their distribution and ecology. In this context, metabarcoding data accompanied by a well-represented reference barcode library are a useful tool for primary hypotheses of species distribution.

4.4.3. Considerations on sequence variation in metabarcoding data

In this work, I have used the accepted barcode for protists (V4 region, Pawlowski et al., 2012) and the V9 region to study a cryptic species complex. Instead of a classical, Sanger-based approach of a multitude of geographic strains, I have used metabarcoding datasets (OSD and Tara Oceans), to take advantage of the data available for many sampling localities across the globe, which would have been difficult to sample with a classical sampling approach of establishing strains. As consequence of this choice, I had to work with thousands of sequences. Indeed, differently from a Sanger sequencing, which provides a single sequence as output (a consensus of all the amplified products), high-throughput techniques sequence every single molecule. Furthermore, since the 18S gene occurs in hundreds to thousands of copies within the genome, and sometime on multiple chromosomes (Alvarez and Wendel, 2003), the number of sequences to handle was even bigger. Such rDNA copies are expected to be homogenised by concerted evolution over time, but empirical studies suggest that this process is not perfect and multiple, polymorphic copies can persist within the genome (Alverson and Kolnick, 2005). When using environmental samples, 18S copies from different cistrons, chromosomes and

individuals are mixed together, rendering it difficult to discern between intra- and interspecific variability.

Using the network approach and simple criteria to infer sequences to a species (see M&M section), I have demonstrated that this is not an issue. Indeed, all these sequences resulting from the apparent failure of concerted evolution to achieve complete homogenisation, from geographic variability, from PCR and sequencing errors are arranged around the main node in which the "dominant haplotype" is located. All these dominant and peripheral haplotypes contribute to the definition of the species' overall genetic variation for this marker region. The dominant haplotype is here defined as the most abundant haplotype for a specific taxon, which also corresponds to the Sanger sequence in the species for which reference barcodes are available.

Furthermore, it is possible that the 18S copies escaping concerted evolution retain ancestral polymorphisms that can help assessing phylogenetic relationships among species.

In this context, I showed that the use of a multi-copy gene is not a disadvantage, but all these copies contribute to the evaluation of inter- and intra-species variation.

References

Álvarez, I., Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution, 29*(3), 417-434.

Alverson, A. J., Kolnick, L. (2005). Intragenomic nucleotide polymorphism among small subunit (18s) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *Journal of Phycology, 41*(6), 1248-1257.

Andrews, K. R., Williams, A. J., Fernandez-Silva, I., Newman, S. J., Copus, J. M., Wakefield, C. B., ... Bowen, B. W. (2016). Phylogeny of deepwater snappers (Genus Etelis) reveals a cryptic species pair in the Indo-Pacific and Pleistocene invasion of the Atlantic. *Molecular Phylogenetics and Evolution, 100*, 361-371.

Arbogast, B. S., Kenagy, G. J. (2001). Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography, 28*(7), 819-825.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography, 36*(1), 3-15.

Avise, J. C., (2000). *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., ... Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics, 18*(1), 489-522.

Baas Becking, L. G. M. (1934). *Geobiologie of Inleiding tot de Milieukunde*. The Hague: Van Stockum & Zoon.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., ... Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution, 22*(3), 148-155.

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer, New York.

Bivand, R., Keitt, T., Rowlingson, B. (2018). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.3-6. https://CRAN.R-project.org/package=rgdal.

Chan, L. M., Brown, J. L., Yoder, A. D. (2011). Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Molecular Phylogenetics and Evolution, 59*(2), 523-537.

Clement, M., Posada, D. C. K. A., Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology, 9*(10), 1657-1659.

Crawford, A. J., Cruz, C., Griffith, E., Ross, H., Ibanez, R., Lips, K. R., ... Crump, P. (2013). DNA barcoding applied to ex situ tropical amphibian conservation programme reveals cryptic diversity in captive populations. *Molecular Ecology Resources*, *13*(6), 1005-1018.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*(10), 566-571.

de Vargas C, Audic S, Tara Oceans Consortium, Coordinators, Tara Oceans Expedition, Participants. (2017). Total V9 rDNA information organized at the metabarcode level for the Tara Oceans Expedition (2009–2012). PANGAEA. DOI 10.1594/PANGAEA.873277.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872-5895.

Feuda, R., Bannikova, A. A., Zemlemerova, E. D., Di Febbraro, M., Loy, A., Hutterer, R., ... Colangelo, P. (2015). Tracing the evolutionary history of the mole, *Talpa europaea*, through mitochondrial DNA phylogeography and species distribution modelling. *Biological Journal of the Linnean Society*, *114*(3), 495-512.

Fišer, C., Robinson, C. T., Malard, F. (2018). Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology*, *27*(3), 613-635.

Gaonkar, C. C. (2017). *Diversity, Distribution and Evolution of the Planktonic Diatom Family Chaetocerotaceae*. PhD thesis, The Open University.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. *PloS ONE*, *13*(12), e0208929.

Godhe, A., Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1728), 20160399.

Gomes, L. C., Pessali, T. C., Sales, N. G., Pompeu, P. S., Carvalho, D. C. (2015). Integrative taxonomy detects cryptic and overlooked fish species in a neotropical river basin. *Genetica, 143*(5), 581-588.

Gran, H. H., (1897). Botanik. Prophyta: Diatomaceae, Silicoflagellata og Cilioflagellata. *Den Norske Nordhavs Expedition 1876–1878, 7*, 1–36.

Gutiérrez-Tapia, P., Palma, R. E. (2016). Integrating phylogeography and species distribution models: cryptic distributional responses to past climate change in an endemic rodent from the central Chile hotspot. *Diversity and Distributions, 22*(6), 638-650.

Han, M. V., Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics, 10*(1), 356.

Hasle G. R., Syvertsen E. E. (1996). Marine Diatoms. In: *Identifying marine phytoplankton* (Ed. By CR Tomas), pp. 5-585. Academic Press, San Diego.

Hasle, G. R. (1976). The biogeography of some marine planktonic diatoms. *Deep Sea Research, 23*, 319-338.

Hebert, P. D., Cywinska, A., Ball, S. L., deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1512), 313-321.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H., Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, 101*(41), 14812-14817.

Huson, D. H., Rupp, R., Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, New York.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ... Picheral, M. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell, 179*(5), 1084-1097.

Iftikhar, R., Ashfaq, M., Rasool, A., Hebert, P. D. (2016). DNA barcode analysis of thrips (Thysanoptera) diversity in Pakistan reveals cryptic species complexes. *PLoS ONE, 11*(1), e0146014.

Jörger, K. M., Norenburg, J. L., Wilson, N. G., Schrödl, M. (2012). Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. *BMC Evolutionary Biology, 12*(1), 245.

Jukes, T. H., Cantor, C. R. (1969). Evolution of protein molecules. In: Munro, H. N., (Ed.), *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods, 14*(6), 587-589.

Katoh, K., Rozewicki, J., Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics, 108,* 1–7.

Knowles, L. L. (2009). Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics, 40*, 593-612.

Kooistra, W. H. C. F., Sarno, D., Hernández-Becerril, D. U., Assmy, P., Di Prisco, C., Montresor, M. (2010). Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia, 49*(5), 471-500.

Kooistra, W. H., Sarno, D., Balzano, S., Gu, H., Andersen, R. A., Zingone, A. (2008). Global diversity and biogeography of *Skeletonema* species (Bacillariophyta). *Protist, 159*(2), 177-193.

Lawson Handley, L. (2015). How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society, 115*(3), 750-766.

Leigh, J. W., Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6*(9), 1110-1116.

Longhurst, A. (2007). *Ecological geography of the sea*. Academic Press, London.

Lukhtanov, V. A., Dantchenko, A. V., Vishnevskaya, M. S., Saifitdinova, A. F. (2015). Detecting cryptic species in sympatry and allopatry: analysis of hidden diversity in *Polyommatus* (Agrodiaetus) butterflies (Lepidoptera: Lycaenidae). *Biological Journal of the Linnean Society, 116*(2), 468-485.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ, 3*, e1420.

Markmann, M., Tautz, D. (2005). Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1462), 1917-1924.

Mayr, E. (1970). *Populations, species, and evolution: an abridgment of animal species and evolution (Vol. 19)*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

McMurdie, P. J., Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS ONE, 8*(4), e61217.

Meyer, C. P., Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology, 3*(12), e422.

Mills, S., Alcántara-Rodríguez, J. A., Ciros-Pérez, J., Gómez, A., Hagiwara, A., Galindo, K. H., ... Welch, D. B. M. (2017). Fifteen species in one: deciphering the *Brachionus plicatilis* species complex (Rotifera, Monogononta) through DNA taxonomy. *Hydrobiologia, 796*(1), 39-58.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H. C. F., Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. *PLoS ONE, 9*(8), e103810.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268-274.

Nielsen, R., Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Molecular Ecology, 18*(6), 1034-1047.

Papakostas, S., Michaloudi, E., Proios, K., Brehm, M., Verhage, L., Rota, J., ... Declerck, S. A. J. (2016). Integrative taxonomy recognizes evolutionary units despite widespread mitonuclear discordance: evidence from a rotifer cryptic species complex. *Systematic Biology, 65*(3), 508-524.

Pargana, A. (2017). *Functional and Molecular Diversity of the Diatom Family Leptocylindraceae*. PhD thesis, The Open University.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... Fiore-Donno, A. M. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology, 10*(11), e1001419.

Pebesma, E. J., Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News, 5*(2), https://cran.r-project.org/doc/Rnews/.

Pfenninger, M., Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology, 7*(1), 121.

Price, M. N., Dehal, P. S., Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saitoh, T., Sugita, N., Someya, S., Iwami, Y., Kobayashi, S., Kamigaichi, H., ... Nishiumi, I. (2015). DNA barcoding reveals 24 distinct lineages as cryptic bird species candidates in and around the Japanese Archipelago. *Molecular Ecology Resources*, *15*(1), 177-186.

Sarno, D., Kooistra, W. H. C. F., Medlin, L. K., Percopo, I., Zingone, A. (2005). Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. *Journal of Phycology*, *41*(1), 151-176.

Sarno, D., Kooistra, W. H., Balzano, S., Hargraves, P. E., & Zingone, A. (2007). Diversity in the genus *Skeletonema* (Bacillariophyceae): III. Phylogenetic position and morphological variability of *Skeletonema costatum* and *Skeletonema grevillei*, with the description of *Skeletonema ardens* sp. nov. *Journal of Phycology, 43*(1), 156-170.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537-7541.

Shapiro, B. J., Leducq, J. B., Mallet, J. (2016). What is speciation? *PLoS Genetics, 12*(3), e1005860.

Steiner, F. M., Csősz, S., Markó, B., Gamisch, A., Rinnhofer, L., Folterbauer, C., ... Schlick-Steiner, B. C. (2018). Turning one into five: Integrative taxonomy

uncovers complex evolution of cryptic species in the harvester ant Messor "structor". *Molecular Phylogenetics and Evolution, 127,* 387-404.

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., ... Dimitrov, D. (2018). Finding evolutionary processes hidden in cryptic species. *Trends in Ecology & Evolution, 33*(3), 153-163.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.

Templeton, A. R., Crandall, K. A., Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics, 132*(2), 619-633.

Trontelj, P., Douady, C. J., Fišer, C., Gibert, J., Gorički, Š., Lefébure, T., ... Zakšek, V. (2009). A molecular test for cryptic diversity in ground water: how large are the ranges of macro-stygobionts? *Freshwater Biology, 54*(4), 727-744.

Trontelj, P., Fišer, C. (2009). Perspectives: cryptic species diversity should not be trivialised. *Systematics and Biodiversity, 7*(1), 1-3.

Wang, J., Wu, J. (2009). Occurrence and potential risks of harmful algal blooms in the East China Sea. *Science of the Total Environment, 407*(13), 4012-4021.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Zhen, Y., Mi, T., Yu, Z. (2009). Detection of several harmful algal species by sandwich hybridization integrated with a nuclease protection assay. *Harmful Algae, 8*(5), 651-657.

Zingone, A., Percopo, I., Sims, P. A., Sarno, D. (2005). Diversity in the genus *Skeletonema* (Bacillariophyceae). I. A re-examination of the type material of *S. costatum* with the description of *S. grevillei* sp. nov. *Journal of Phycology, 41*(1), 140-150.

Appendix IV

Table A4.1. List of OSD and Tara Oceans sites in which were found metabarcodes validated as *C. curvisetus* spp.

OSD			Tara Oceans		
Station	**Longitude**	**Latitude**	**Station**	**Longitude**	**Latitude**
OSD2	-3.938	48.778	TARA_004	-6.553	36.563
OSD4	14.25	40.808	TARA_005	-4.406	36.030
OSD5	24.99	35.661	TARA_006	-4.251	36.529
OSD6	2.8	41.667	TARA_007	1.948	37.031
OSD13	27.909	43.176	TARA_008	3.966	38.011
OSD14	3.15	42.49	TARA_009	5.820	39.112
OSD22	5.175	43.226	TARA_010	2.865	40.668
OSD24	-2.88	35.193	TARA_011	2.798	41.666
OSD26	-5.75	35.82	TARA_012	7.899	43.348
OSD29	-80.283	27.469	TARA_014	12.858	39.902
OSD37	-80.093	26.103	TARA_016	15.454	37.398
OSD38	-80.784	24.745	TARA_017	14.306	36.258
OSD43	-117.257	32.867	TARA_018	14.288	35.756
OSD50	-1.925	43.333	TARA_019	13.865	34.216
OSD51	-82.266	9.348	TARA_020	14.973	34.451
OSD54	-69.641	43.844	TARA_022	17.400	39.729
OSD55	-69.578	43.86	TARA_023	17.729	42.176
OSD58	-76.671	34.718	TARA_024	17.956	42.457
OSD101	-16.711	32.742	TARA_025	19.421	39.333
OSD102	-16.91	32.646	TARA_026	20.188	38.431
OSD107	-9.38	39.14	TARA_030	32.789	33.929
OSD108	-8.966	38.757	TARA_031	34.819	27.151
OSD109	-9.012	38.677	TARA_032	37.254	23.391
OSD110	-8.869	40.145	TARA_033	38.218	22.057
OSD115	-9.385	39.134	TARA_034	39.884	18.445
OSD116	-9.219	39.415	TARA_036	63.524	20.824
OSD117	-7.504	37.167	TARA_038	64.576	19.017
OSD124	135.121	34.324	TARA_039	66.463	18.647
OSD131	27.401	42.245	TARA_040	67.984	17.500
OSD145	3.119	51.361	TARA_041	70.011	14.582
OSD147	81.052	8.522	TARA_042	73.919	5.992

OSD148	8.149	53.581	TARA_043	73.489	4.660
OSD153	-7.973	36.998	TARA_044	71.520	2.806
OSD154	-1.167	44.667	TARA_045	71.710	0.941
OSD155	10.599	59.816	TARA_046	73.162	-0.659
OSD156	10.72	59.9	TARA_047	72.164	-2.042
OSD157	10.628	59.622	TARA_048	66.320	-9.408
OSD158	-25.19	37.433	TARA_049	59.504	-16.808
OSD159	-4.552	48.359	TARA_050	56.795	-21.476
OSD162	-2.103	56.963	TARA_051	54.283	-21.476
OSD163	-2.973	58.957	TARA_052	53.508	-17.023
OSD166	2.9	43.433	TARA_053	46.923	-13.070
OSD173	3.14	51.441	TARA_054	45.226	-12.813
OSD177	2.702	51.186	TARA_057	42.742	-17.026
OSD60	-79.168	33.323	TARA_058	42.320	-17.455
OSD64	30.776	46.442	TARA_062	40.182	-22.339
OSD69	12.26	45.457	TARA_064	37.929	-29.508
OSD70	12.438	45.414	TARA_065	26.334	-35.226
OSD71	170.771	-45.744	TARA_066	18.016	-34.905
OSD74	-8.667	41.142	TARA_068	4.620	-31.039
OSD76	12.935	43.948	TARA_072	-18.006	-8.691
OSD77	13.073	43.851	TARA_076	-35.231	-21.029
OSD78	13.595	43.57	TARA_078	-43.323	-30.158
OSD81	-7.973	37.005	TARA_080	-51.952	-40.698
OSD91	-9.037	32.747	TARA_081	-52.214	-44.497
OSD92	-7.701	33.584	TARA_082	-58.012	-47.165
OSD94	-2.215	35.086	TARA_083	-65.023	-54.418
OSD95	103.917	1.268	TARA_085	-49.503	-62.176
OSD97	-28.602	38.53	TARA_088	-56.806	-63.386
OSD98	-28.13	38.64	TARA_092	-71.977	-33.697
			TARA_094	-87.093	-32.765
			TARA_096	-101.268	-29.655
			TARA_098	-110.992	-26.261
			TARA_100	-96.283	-13.162
			TARA_102	-85.270	-5.218
			TARA_106	-84.620	0.037
			TARA_109	-84.545	1.800

			TARA_110	-84.616	-1.913
			TARA_113	-134.920	-23.114
			TARA_114	-134.912	-23.130
			TARA_115	-134.931	-23.216
			TARA_116	-134.931	-23.217
			TARA_118	-135.009	-23.129
			TARA_120	-134.912	-23.012
			TARA_123	-140.304	-8.878
			TARA_125	-142.610	-8.890
			TARA_126	-151.208	-11.975
			TARA_128	-153.305	-0.469
			TARA_130	-152.462	11.265
			TARA_131	-158.052	22.746
			TARA_133	-127.750	35.343
			TARA_134	-121.986	32.667
			TARA_135	-121.832	32.983
			TARA_137	-116.699	14.161
			TARA_138	-103.017	6.216
			TARA_140	-79.312	7.471
			TARA_141	-80.086	9.834
			TARA_142	-88.417	25.602
			TARA_143	-79.682	29.885
			TARA_144	-72.815	36.369
			TARA_145	-70.076	39.163
			TARA_146	-71.248	34.731
			TARA_147	-66.533	32.954
			TARA_148	-64.145	31.782
			TARA_149	-49.840	34.098
			TARA_150	-37.102	35.800
			TARA_151	-28.801	36.194
			TARA_152	-16.662	43.668
			TARA_153	-16.564	44.034
			TARA_155	-16.755	54.597
			TARA_158	0.374	67.193
			TARA_163	1.689	76.078
			TARA_168	44.126	72.582

			TARA_173	75.345	78.939
			TARA_175	66.384	79.343
			TARA_178	73.235	77.234
			TARA_180	75.459	75.172
			TARA_188	91.725	78.304
			TARA_189	116.482	78.022
			TARA_191	160.961	71.549
			TARA_193	174.901	71.115
			TARA_194	-168.518	73.336
			TARA_196	-154.934	71.895
			TARA_201	-85.729	74.329
			TARA_205	-71.952	72.423
			TARA_208	-51.578	69.107
			TARA_210	-55.985	61.544

Chapter V

Concerted evolution

in Chaetoceros

5.1. Introduction

The first DNA reannealing and hybridisation studies conducted in the mid-1960-70s to unveil the structure and organisation of eukaryotic genomes showed that a large fraction of them was composed of repetitive regions (Britten and Waring, 1965; Britten and Kohne, 1968). The subsequent study of such regions revealed that, when comparing repetitive DNA families, there was greater sequence similarity within species than between species (Brown et al., 1972; Elder and Turner, 1995). Such observation was incompatible with the then common model of divergent evolution, according to which the differences in nucleotide sequence between different repeats of the same species were expected to be as large as those between repeats of different species (Nei and Rooney, 2005). Therefore, there had to be a mechanism responsible for the homogenisation of such sequences within an individual organism. The expression "concerted evolution" (Zimmer et al., 1980) was coined to indicate this phenomenon, by which an individual member of a gene family evolves in the same (concerted) way as all the other members of the family (Graur and Li, 1999).

The best-known example of concerted evolution is the rDNA cistron (Ganley and Kobayashi, 2007), but also other genes and non-coding regions (e.g. globins, immunoglobulins, heat-shock genes, histones) are known to evolve in this way (Long and Dawid, 1980; Liebhaber et al., 1981; Coen et al., 1982; Gojobori and Nei, 1984).

The exact mechanisms determining concerted evolution are still unclear. However, two processes, gene conversion and unequal crossing-over, are considered responsible for sequence homogeneity, the latter also causing fluctuations in number over evolutionary time (Lindegren, 1953; Holliday, 1964; Charlesworth et al., 1986). Despite this, the mechanism is not perfect and cases of deviations from such homogenisation have been detected in animals (Nikolaidis and Nei, 2004; Andrea et al., 2006), fungi (Li et al., 2013),

and especially in plants (Harpke and Peterson, 2006; Zheng et al., 2008; Xiao et al., 2010; Vilnet et al., 2012; Xu et al., 2017).

The extent of such non-homogenisation is particularly important in the case of the rDNA cistron, since it is the classical target for DNA barcoding studies in some taxa as fungi and protists (Pawlowski et al., 2012; Schoch et al., 2012; Stoeck et al., 2014). Therefore, understanding the inheritance of ribosomal genes and spacers is vital for taxonomic and systematic studies involving them.

So far, exceptions to concerted evolution have been spotted detecting noise in electropherograms and then cloning and sequencing subsamples of amplified products (e.g. Pillet et al., 2012; Naidoo et al., 2013). The resulting sequences were then put on a phylogenetic tree together with the ones from closely related species to ascertain the degree of similarity within and among species.

This approach has two main limitations: first, the number of detectable variants is constrained by the number of clones that are sequenced; second, there is no information about the abundance of each variant. Nowadays, metabarcoding techniques allow sequencing thousands of copies of a target region from environmental samples, bulk communities and even single specimens. The latter approach can be particularly useful to study concerted evolution.

A temporal metabarcoding analysis conducted in the LTER MareChiara (Gulf of Naples, Italy) across three years (48 dates) to unveil species diversity within the diatom family Chaetocerotaceae (Gaonkar, 2017) showed the following pattern. When a phylogenetic tree based on V4-18S metabarcodes was inferred, many terminal clades contained from few to tens of haplotypes, one of which was far more abundant than the others. Such a sequence, called “dominant haplotype”, was identical or nearly identical to the reference sequence (Sanger), when available, for that clade (Gaonkar, 2017). Furthermore, the relationship among “dominant” and “minor” haplotypes across species was consistent: when plotted on

a logarithmic scale, the dominant haplotype was of two orders more abundant than the others were. The number of detectable minor haplotypes in the environmental sample was function of the abundance of the dominant one: the more abundant the latter, the bigger the number of minor haplotypes.

However, the author did not discuss if such "minor" haplotypes were PCR or sequencing errors as well as intra- or inter-individual (strain) variation, but argued that such pattern can be considered as "the result of an equilibrium between the appearance of novel haplotypes, random drift, and the homogenizing effect of concerted evolution" (Gaonkar, 2017).

Based on the theory of concerted evolution, I formulated the hypothesis that the patterns observed at temporal scale in the 48 samples of MareChiara dataset were related to this phenomenon. To confirm or reject that hypothesis, I designed an experiment based on HTS of V4 region of 18S gene from single strains of different *Chaetoceros* species to test:

i) If the proportion between dominant and minor haplotypes in the environmental samples is also observed within individual strains;

ii) The identity between the sequence of the dominant haplotype in the HTS single strain both with the Sanger reference and with the sequence of the dominant haplotype in environmental metabarcoding for each species;

iii) The identity between the sequences with low abundance (minor haplotypes) found in the HTS single strain with the sequences found in the environmental samples;

iv) The pattern of phylogenetic networks for each *Chaetoceros* species using the metabarcoding dataset generated from the temporal distribution (48 dates).

5.2. Materials and Methods

5.2.1. Selection of taxa to study concerted evolution

In order to answer the aforementioned questions, I used part of the data from the thesis of Chetan Gaonkar (Gaonkar, 2017) and the metabarcoding data of Chaetocerotaceae from the LTER MareChiara (Gulf of Naples) deposited in GenBank at the accession numbers MK938374-MK940235 (414,041 reads). I started analysing the species *C. curvisetus* 2, from which the pattern of concerted evolution was first hypothesised (see Preface). Then, I used the HTS phylogenetic tree in Gaonkar (2017) inferred from the 48 dates of MareChiara to select other *Chaetoceros* species. In particular, I have chosen: i) a species occurring at high abundance all over the year and so displaying many minor haplotypes (*C. tenuissimus*); ii) a species with a marked seasonality displaying also a few minor haplotypes at high abundances (*C. costatus*); iii) a species displaying a single, lowly abundant, dominant haplotype (*C. anastomosans*); iv) two species without a clear delimitation that occurred in the same clade despite having different reference barcodes, and so with mixed minor haplotypes (*Chaetoceros* sp. Na11C3 and Na26B1). For each species, I selected outgroup taxa (Table 5.1) for subsequent validation of sequences gathered from BLAST analysis. The undescribed species *C.* sp. Na11C3 and *C.* sp. Na26B1 were analysed together because they were in the same clade in the NGS tree of Gaonkar (2017) despite having different barcodes.

Table 5.1. List of outgroup taxa for the validation of *Chaetoceros*-species sequences.

Species	**Outgroups**	**Accession number**
C. anastomosans	*C.* cf. *vixvisibilis* Na16A3 *Chaetoceros* sp. Na11C3	MG972367 MG972328
C. costatus	*C. cinctus* Ch6A2 *C. radicans* Ch2A2	KY852264 KY852259

C. curvisetus 2	*C.* cf. *tortissimus* Na18C4	MG972275
	C. tortissimus	MG972325
Chaetoceros sp. Na11C3 / Na26B1	*C. anastomosans* Na14C2	MG972358
	C. cf. *vixvisibilis* Na16A3	MG972367
	Chaetoceros clone HM347543	HM347543
C. tenuissimus	*C. neogracilis* 1 RCC2507	KT860998
	C. neogracilis 2 RCC2318	JN934684
	C. neogracilis 4 RCC2016	JF794049
	Chaetoceros sp.	AF145226
	Chaetoceros sp.	X85390

5.2.2. Analysis of environmental sequences

I used the metabarcoding data corresponding to 48 environmental samples collected in the LTER-MareChiara (Gulf of Naples, Italy) produced and processed by Gaonkar (2017). These data were sequenced in paired end (2 × 250 bp) on an Illumina MiSeq platform (see Gaonkar 2017 for further details) and are available in GenBank at the accession numbers MK938374-MK940235.

The procedure followed to retrieve sequences of selected species of *Chaetoceros* in the MareChiara dataset is similar to the one adopted in the previous chapter. In brief, I used the V4 region of my target species and close outgroups as queries for a local BLAST at 95%. The metabarcodes extracted were then aligned with the references and the outgroup taxa using MAFFT online (Katoh et al., 2017) and a phylogenetic tree was built in FastTree v2.1.8 (Price et al., 2010), using the GTR model. The resulting tree was visualised and modified in Archaeopteryx v0.9901 (Han and Zmasek, 2009) in order to remove sequences clustering within outgroup clades and gather only metabarcodes of the species of interest. The sequences retrieved were considered validated and used to retrieve the info of abundance using mothur v1.41.1 (Schloss et al., 2009).

5.2.3. Single strain HTS

Single strain metabarcoding was performed on: two strains of *C. anastomosans*, four strains of *C. costatus*, four strains of *C. curvisetus* sp. 2, one of *Chaetoceros* sp. Na26B1, two of *Chaetoceros* sp. Na11C3 and three strains of *C. tenuissimus* (Table 5.2).

Table 5.2. List of strains utilised for single-strain HTS.

Species	**Strain**
C. anastomosans	Na14C2 Na14C3
C. costatus	Na1A3 Na32B1 Ro1B1 Ro2A2
C. curvisetus 2	Ch5B2 Na1C1 Na19A2 Na20A4
Chaetoceros sp. Na11C3	Na11C3 Na43A1
Chaetoceros sp. Na26B1	Na26B1
C. tenuissimus	GB2a Na26A1 Na44A1

Abbreviations are as follows: Ch = Chile; Na = Naples; Ro = Roscoff. GB2a is a strain from the Gulf of Naples.

For each sample, I performed individual PCR in two steps: a first reaction for the amplification of the target sequence, and a second reaction (using the PCR product of the former one as template) to ligate proprietary adaptor sequence (P1) and unique 10–12 bp long identifier nucleotide key tags (barcodes) compatible with the GeneStudio S5 Ion

Torrent (Life Technologies). The obtained fragment contained all the information required for sequencing and differentiation of samples. The first amplification was conducted using the primers targeting the 18S-V4 region by Stoeck et al. (2010) modified by Piredda et al. (2016). PCRs were conducted in a final volume of 25 µL each containing: 3 ng of DNA, 1x Buffer HF, 0.2 mM dNTPs, 0.5 µM of each primer, 1U of Phusion High-Fidelity DNA polymerase (New England Biolabs Inc, Ipswich, MA) and water to volume. The thermal cycling profiles started with 98 °C for 30 s, followed by 10 cycles of denaturation at 98 °C for 10 s, annealing at 44 °C for 30 s, extension at 72 °C for 15 s, and then additional 15 cycles of denaturation at 98 °C for 10 s, annealing at 62 °C for 30 s and extension at 72 °C for 15 s, with a final extension at 72 °C for 7 min. PCR products (~470 bp) were visualised on 1.2% agarose gel and purified using the AMPure XP Beads kit (Agencourt Bioscience Corp., Beverly, MA, USA), at a concentration of 1.2× vol/vol, according to manufacturer's instructions. The second PCR was conducted in the same volume and using the same concentrations of reagents (DNA, dNTPs, Buffer and Taq). Adapter P1 was added at a concentration of 50 µM, whilst each barcode of 20 µM. The amplification profile was as follows: initial denaturation at 98 °C for 30 s; 5 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 30 s, extension at 72 °C for 15 s, and then a final extension at 72 °C for 7 min. The success of insertion of adapter and barcode in PCR products was checked by electrophoresis on 1.2% agarose gel (increase of size). Amplified products were purified as above and quantity and quality were determined with the Agilent DNA High Sensitivity Kit on the Bioanalyzer (Agilent) following the manufacturer's recommendations. Since not all PCRs amplified only the fragment of interest, prior to emulsion PCR an equal amount of all COI products was pooled and processed for fragment size selection (around 500 bp). This was done by running the pooled samples on 1.2% agarose gel together with a size standard and cutting the band of interest, which was then purified using the GenElute™ Gel Extraction Kit (Sigma-Aldrich). Emulsion PCR was

conducted in the Ion Chef System (Life Technologies) using 0.1 fmol/µL of the pool into a reaction volume of 50 µL. Massive-parallel sequencing was carried out using the Ion GeneStudio™ S5 System (Life Technologies).

5.2.4. Data pre-processing and analysis of single-strain HTS

From raw fastq data, adapters and primers were removed with cutadapt (Martin, 2011), allowing a maximum of three mismatches. All reads with a length < 350 bp and quality score < 20 were discarded.

Data obtained with Ion Torrent technology are known to have indel errors of an order of magnitude more frequent than substitution errors (Laehnemann et al., 2016), with most of indel errors caused by homopolymers. Furthermore, the Ion Torrent platform is known to have a higher indel error rate associated with the homopolymer region than the Illumina platform (Loman et al., 2012). In order to overcome this issue, I corrected indel errors using ICC v2.0.1 (Deng et al., 2013). This software starts filtering sequences based on length and quality and then blasts them against a reference. Successively, it retrieves the sequences in windows and proceeds with the correction, which is performed in clusters differing by homopolymer indels. As reference for BLAST, I used, for each species, the V4 region generated by Sanger sequencing of one of the strains listed in Table 5.2 since they are all identical.

5.2.5. Testing the concerted evolution hypothesis

Patterns of concerted evolution were detected by means of abundance plots, BLAST analysis and haplotype networks.

I plotted the abundance of the first most abundant 50 haplotypes for both environmental and single-strain samples in order to render the plots clearly readable. If the hypothesis of concerted evolution in action was correct, I expected to see a steep decrease in abundance

of minor haplotypes with respect to the dominant one. Plots were made in R (R Core Team, 2019) using the packages *ggplot2* (Wickham, 2016), *gridExtra* (Auguie, 2017) and *scales* (Wickham, 2018).

As second strategy, I blasted the validated environmental metabarcodes of each *Chaetoceros* species and the reference barcodes against those of the merged single-strains of each species. This was done in order to ascertain if: i) the most abundant haplotype in each single strain matched the reference barcode of that strain obtained with Sanger sequencing and with the dominant environmental haplotype; ii) the minor haplotypes in the strain were also found in the environmental samples.

Finally, as further check, I inferred haplotype networks for each species from environmental data (MareChiara) using the TCS method (Clement et al., 2000) implemented in PopART v1.7 (Leigh and Bryant, 2015). I only used metabarcodes with abundance ≥ 2 in order to reduce the number of sequences to be processed for network inference. Furthermore, for *C. costatus* and *C. tenuissimus*, I further reduced the number of haplotypes analysed considering only the ones with abundance ≥ 10 and ≥ 50 respectively, in order to obtain a clearer graphical visualisation of networks. Metabarcodes, spanning from 2011 to 2013, were pooled together in months, and a different colour was assigned to each of them. This was done to test the following hypothesis: if concerted evolution was in action, I would have observed not only a major node surrounded by smaller ones, but also a congruence in the temporal pattern (colour pattern in the nodes). If not, I could have observed multiple dominant haplotypes (multiple major nodes), without a correspondence between the temporal pattern in peripheral and major nodes.

5.3. Results

5.3.1. General characteristics of the datasets

The number of haplotypes retrieved for each species from the environmental dataset of MareChiara after the validation procedure described in section 5.3.2 is provided in Table 5.3. In this thesis, the term "haplotype" indicates the non-redundant (unique) sequences. The number of haplotypes utilised ranged from 15 (*C. anastomosans*) to 527 (*C.* sp. Na11C3). In *C. tenuissimus*, considering only the first 50 most abundant haplotypes, I recovered 121,321 sequences (Table 5.3).

Table 5.3. Number of environmental sequences and haplotypes utilised in this study.

Species	**N sequences utilised**	**N haplotypes utilised**
C. anastomosans	287 (abundance ≥ 2)	14 (abundance ≥ 2)
C. costatus	8,220 (abundance ≥ 10)	38 (abundance ≥ 10)
C. curvisetus 2	9,763 (abundance ≥ 2)	369 (abundance ≥ 2)
Chaetoceros sp. Na11C3	12,924 (abundance ≥ 2)	527 (abundance ≥ 2)
Chaetoceros sp. Na26B1	1,154 (abundance ≥ 2)	59 (abundance ≥ 2)
C. tenuissimus	121,321 (abundance ≥ 50)	102 (abundance ≥ 50)

For single strain HTS, the number of raw sequences ranged from 32,112 (*C. curvisetus* 2 Na1C1) to 516,766 (*Chaetoceros* sp. Na11C3 and, after pre-processing, from 19,185 (*C. curvisetus* 2 Na1C1) to 94,449 (*Chaetoceros* sp. Na11C3). The number of haplotypes used for following analyses ranged from a minimum of 2,002 (*C. curvisetus* 2 strain Na1C1) to a maximum of 4,696 (*C. costatus* strain Na32B1) (Table 5.4).

Table 5.4. Number of sequences before and after pre-processing and total number of haplotypes utilised in each strain. Pre-processing refers to removal of adapters, primers and correction with ICC.

Species/strains	**N raw sequences**	**N sequences after pre-processing**	**N haplotypes after pre-processing**
C. anastomosans			
Na14C2	427,364	62,284	4,310
Na14C3	431,665	62,183	3,970
C. costatus			
Na1A3	238,922	34,226	3,634
Na32B1	421,807	50,407	4,696
Ro1B1	274,436	37,489	4,170
Ro2A2	230,989	32,394	3,622
***C. curvisetus* 2**			
Ch5B2	161,145	39,735	2,985
Na1C1	32,112	19,185	2,002
Na19A2	120,545	34,287	2,794
Na20A4	117,234	34,149	2,738
***Chaetoceros* sp. Na11C3**			
Na11C3	516,766	94,449	5,055
Na43A1	259,525	54,973	4,444
***Chaetoceros* sp. Na26B1**			
Na26B1	273,039	56,985	3,360
C. tenuissimus			
GB2a	211,777	39,516	3,986
Na26A1	147,806	34,726	3,024
Na44A1	202,198	32,467	3,024

5.3.2. Abundance plots from environmental metabarcoding and single strain HTS

The plotting of the 50 most abundant haplotypes (Table A5.1 in Appendix V) from environmental metabarcoding data versus their abundance (log10 transformed) in each species (Fig. 5.1) showed a characteristic pattern. Indeed, in each species analysed, of all the haplotypes attributed to a particular species (environmental samples) there was one (the "dominant haplotype") that was far more abundant of all the others, of at least one order of magnitude (Fig. 5.1). All the other copies occurred in the environment at lower abundance.

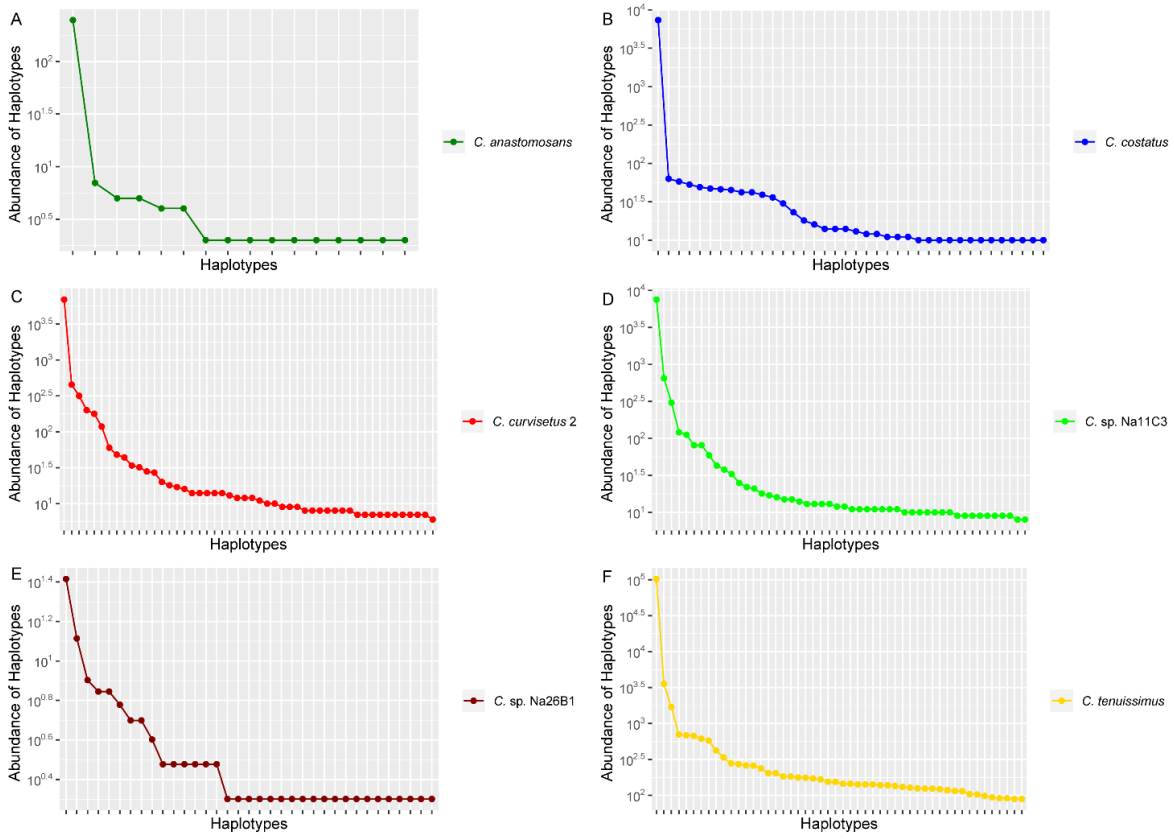


Fig. 5.1. Abundance plots for each *Chaetoceros* species from validated environmental sequences. (A) *C. anastomosans*; (B) *C. costatus*; (C) *C. curvisetus* 2; (D) *Chaetoceros* sp. Na11C3; (E) *Chaetoceros* sp. Na26B1; (F) *C. tenuissimus*. Only the first 50 most abundant haplotypes were plotted. Data were from the temporal metabarcoding dataset "MareChiara" (January 2011 to December 2013).

Patterns of abundance distribution in the HTS of single strains showed the same trend observed in the matabarcoding data of environmental samples (Fig. 5.2). Indeed, in each strain there was the same steep decrease in abundance of minor haplotypes in respect to the

dominant one. Furthermore, within the same species, the distribution of abundance of haplotypes was congruent (Fig. 5.2). The list of the 50 most abundant haplotypes from single strain HTS in each species, used for the plots, is available in the Appendix V as Table A5.2.

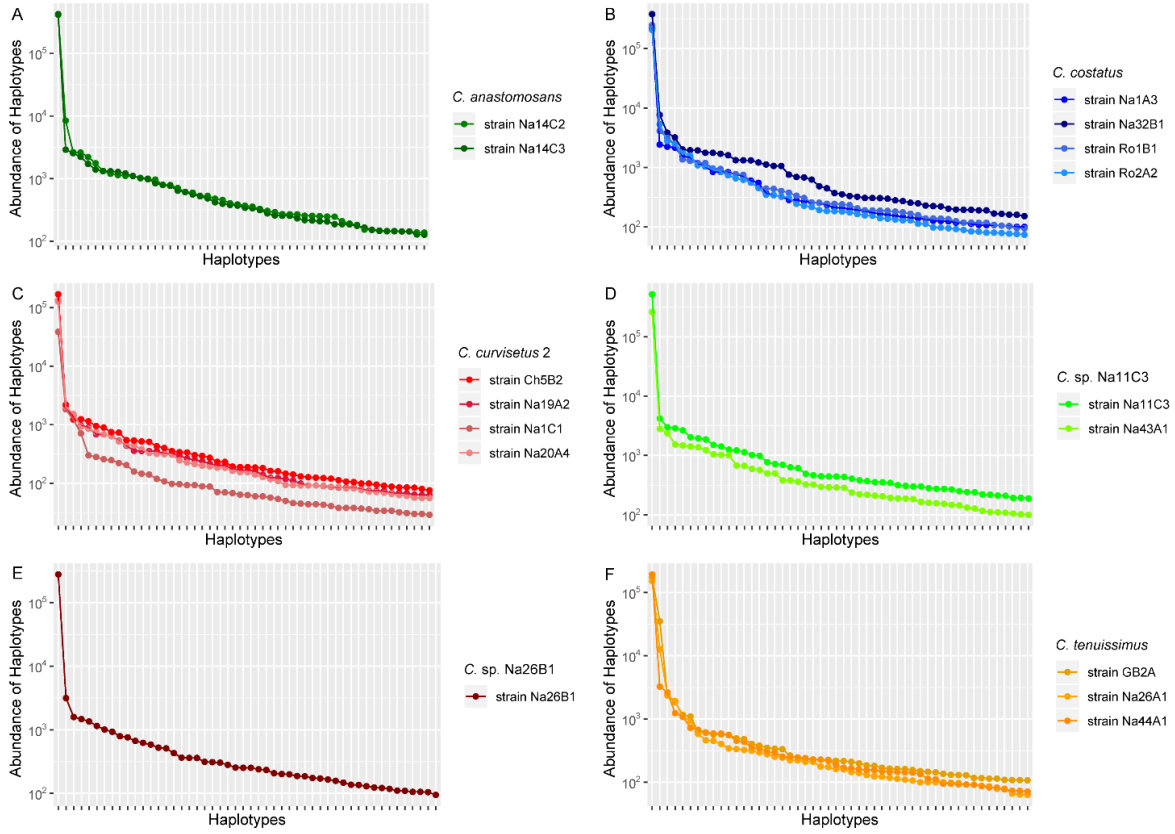


Fig. 5.2. Abundance plots for each strain analysed in different *Chaetoceros* species. (A) *C. anastomosans*; (B) *C. costatus*; (C) *C. curvisetus* 2; (D) *Chaetoceros* sp. Na11C3; (E) *Chaetoceros* sp. Na26B1; (F) *C. tenuissimus*. Data are from single strain high throughput sequencing. Only the first 50 most abundant haplotypes were plotted.

5.3.3. Blast of environmental haplotypes vs. single strain

Within each species, the dominant haplotypes of each strain were identical to each other. Therefore, for showing the results of BLAST analyses of single strains, I referred to just one haplotype (Table 5.5).

The result of BLAST analysis showed that the most abundant haplotype from environmental data as well as single strain HTS matched at 100% identity with the reference barcode (obtained with Sanger sequencing) of the species/strain it belonged (Table 5.5).

Table 5.5. Correspondence between the reference barcode (Sanger sequence) of each species and the dominant haplotypes of the environmental dataset (MareChiara) and single strain HTS. Since the reference sequences of the strains are identical to each other within the same species, only one has been chosen.

Species	**Reference sequence Accession number**	**Matching haplotype in MareChiara**	**% identity**	**Matching haplotype in single strain**	**% identity**
C. anastomosans	MG972358	M00390_81_000000000-AA7DR_1_2109_10899_14476	100	97KSI_03703_04635	100
C. costatus	KY852258	M00390_81_000000000-AA7DR_1_1112_20701_25092	100	97KSI_03062_04287	100
C. curvisetus 2	MG972239	M00390_81_000000000-AA7DR_1_1101_24335_7294	100	97KSI_04187_04119	100
Chaetoceros sp. Na11C3	MG972328	M00390_81_000000000-AA7DR_1_1101_6410_5509	100	97KSI_03663_01512	100
Chaetoceros sp. Na26B1	MG972329	M00390_81_000000000-AA7DR_1_1101_16198_12414	100	97KSI_01986_05212	100
C. tenuissimus	MG972311	M00390_81_000000000-AA7DR_1_1101_19390_3055	100	97KSI_00416_02071	100

Table 5.6. Summary of percentage of identity found between environmental haplotypes and single strain in each *Chaetoceros* species. Single strains have been merged together before BLAST analysis.

		% identity between MareChiara haplotypes and single HTS			
Species	**N hap**	**100**	**99.74-99.73**	**99.48-99.47**	**99.21-99.20**
C. anastomosans	14	42.9 %	50.0 %	7.1 %	-
C. costatus	38	73.7 %	26.3 %	-	-
C. curvisetus 2	369	53.6 %	41.5 %	4.9 %	-
C. sp. Na11C3	527	56.9 %	38.3 %	4.2 %	0.6 %
C. sp. Na26B1	59	45.8 %	49.2 %	5.0 %	-
C. tenuissimus	102	60.8 %	39.2 %	-	-

In most of the species, more than half of environmental haplotypes attributed were also found in single strains HTS at 100 % of identity. Overall, a match between environmental and single strain haplotypes was found for each species within the threshold of 99.20 % of identity (Table 5.6). This result support the hypothesis that the sequence variability observed in the environmental metabarcoding samples is part of infraspecific variation.

5.3.4. Phylogenetic networks from environmental samples

The inference of haplotype networks from the MareChiara dataset provided a graphical evidence to the occurrence of concerted evolution in the *Chaetoceros* species here analysed. Of all the species here analysed (Fig. 5.3 to Fig. 5.7), the temporal pattern observed in the node containing the dominant haplotype corresponded the temporal pattern of the other nodes containing haplotypes with lower abundance. This was particularly straightforward for *C. curvisetus* 2 (Fig. 5.5), *C.* sp. Na11C3 (Fig. 5.6, left network) and *C. tenuissimus* (Fig. 5.7). These were also the species with the highest number of haplotypes

utilised (369, 527 and 102 respectively). In *C. anastomosans* (Fig. 5.3) the pattern is almost absent due to the low number of sequences validated from the MareChiara dataset. However, in the HTS analysis of single strains (Fig. 5.2A), I have observed the expected pattern for concerted evolution in action.

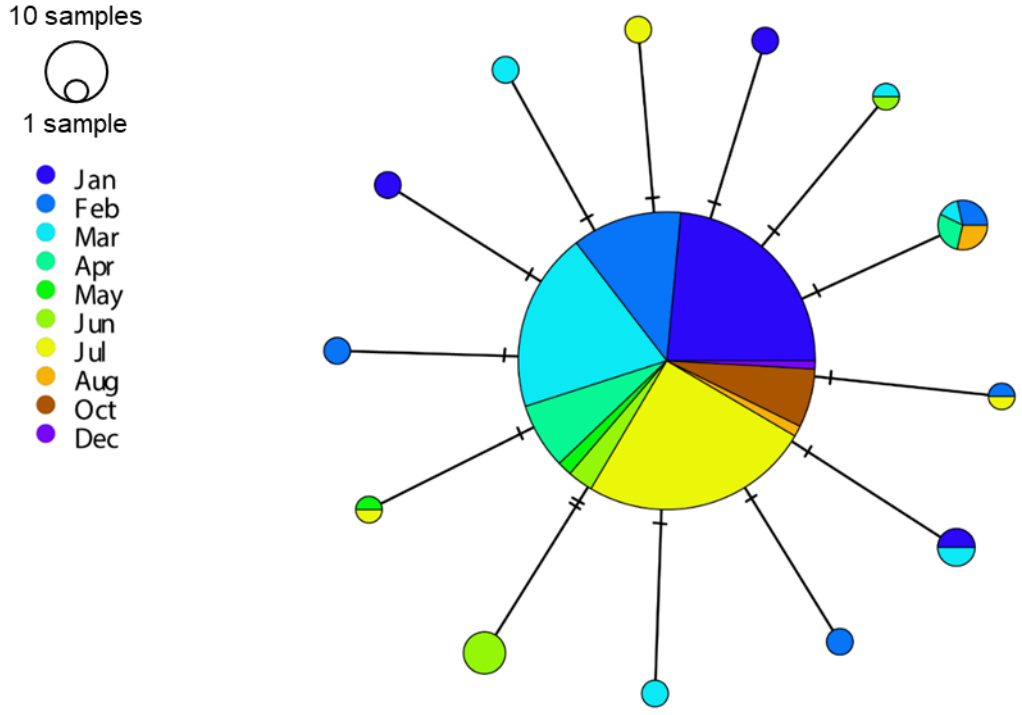


Fig. 5.3. TCS haplotype network for *C. anastomosans* inferred from the MareChiara temporal dataset. A total of 14 haplotypes with abundance ≥ 2 across 2011 and 2013 was used. Sample in the legend refers to the number of reads.

The removal of all the haplotypes with abundance ≤ 9 in *C. costatus* allowed a better visualisation of minor haplotypes (Fig. 5.4).

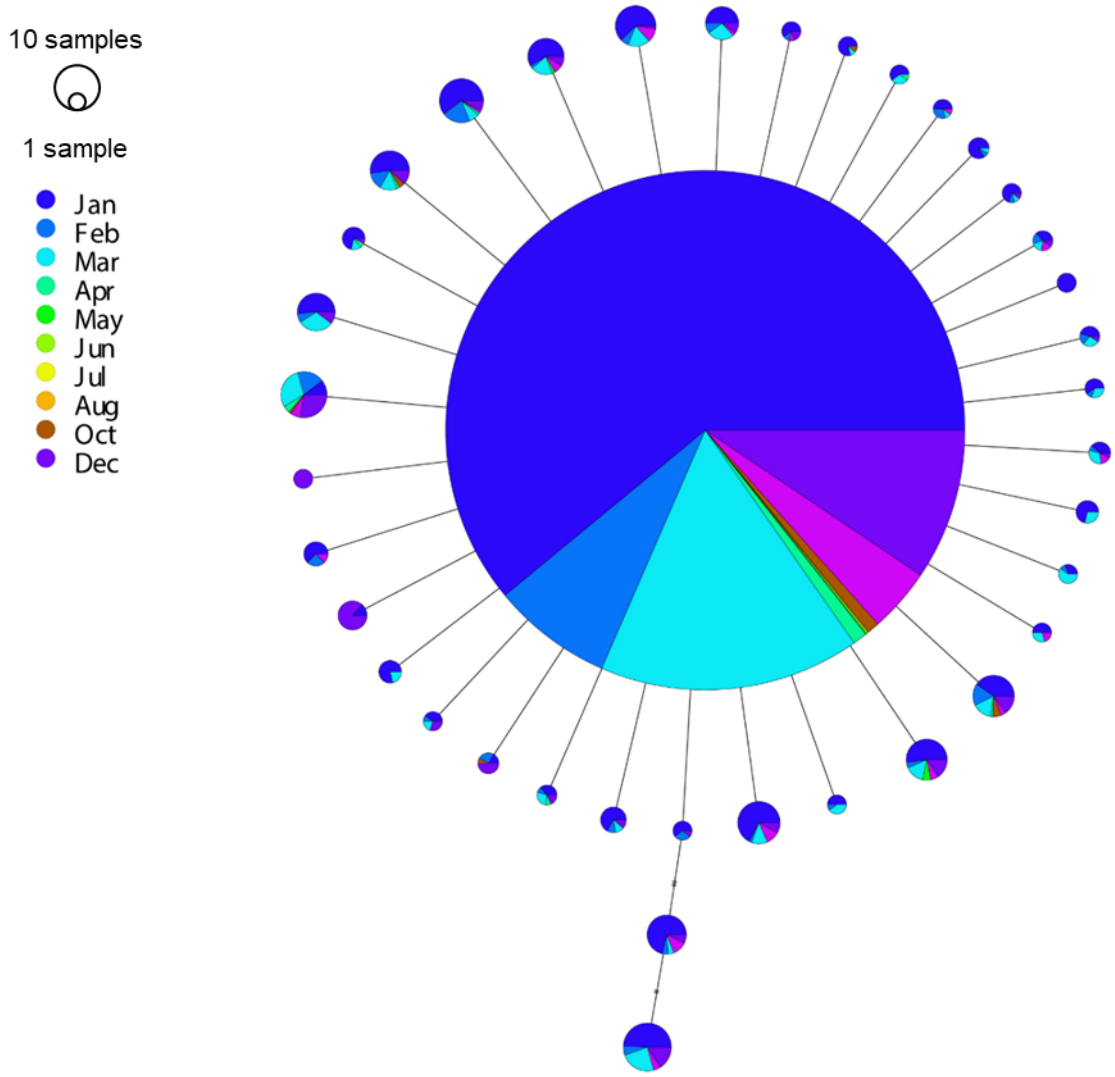


Fig. 5.4. TCS haplotype network for *C. costatus* inferred from the MareChiara temporal dataset. A total of 38 haplotypes with abundance ≥ 10 across 2011 and 2013 was used. Sample in the legend refers to the number of reads.

In *C. curvisetus* 2 (Fig. 5.5), I observed at least ten nodes (minor haplotypes) with many reads whose temporal distribution patterns mimic that of the node comprising the dominant haplotype (the big one).

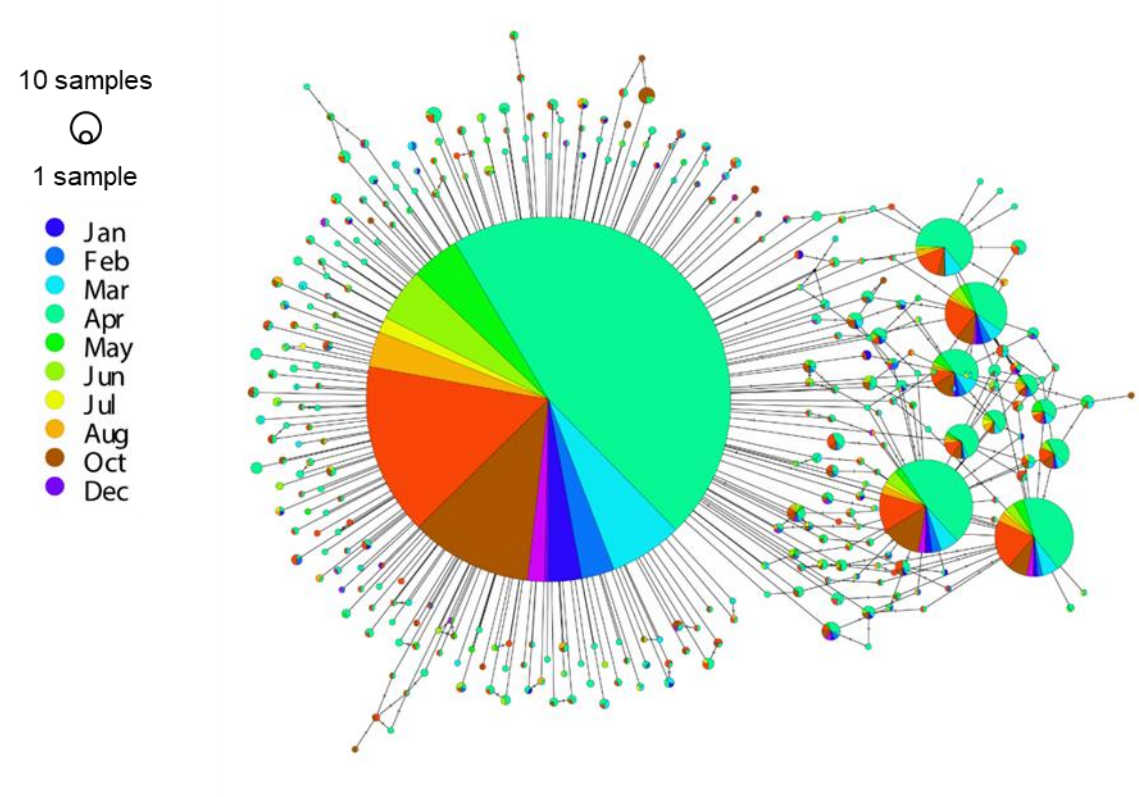


Fig. 5.5. TCS haplotype network for *C. curvisetus* 2 inferred from the MareChiara temporal dataset. A total of 369 haplotypes with abundance ≥ 2 across 2011 and 2013 was used. Sample in the legend refers to the number of reads.

For the closely related species *C.* sp. Na11C3 and *C.* sp. Na26B1, I have inferred a common TCS network since in the HTS phylogenetic tree in Gaonkar (2017) they were in the same clade with mixing minor haplotypes. In the network here inferred, they are on separated nodes, each with their minor haplotypes. The pattern of nodes expected in the case of concerted evolution is more evident for *C.* sp. Na11C3, where more sequences (527) were used; however, it is also observable, in reduced manner, in *C.* sp. Na26B1 (59 sequences utilised). An interest characteristic of such network is the fact that the two closely related species are differentiating each other in the occurrence across the year. The species *C.* sp. Na26B1was exclusively found, during the years 2011-2013, in the months from August to October, whilst *C.* sp. Na11C3 is particularly abundant in June and July and less in the other months (Fig. 5.6).

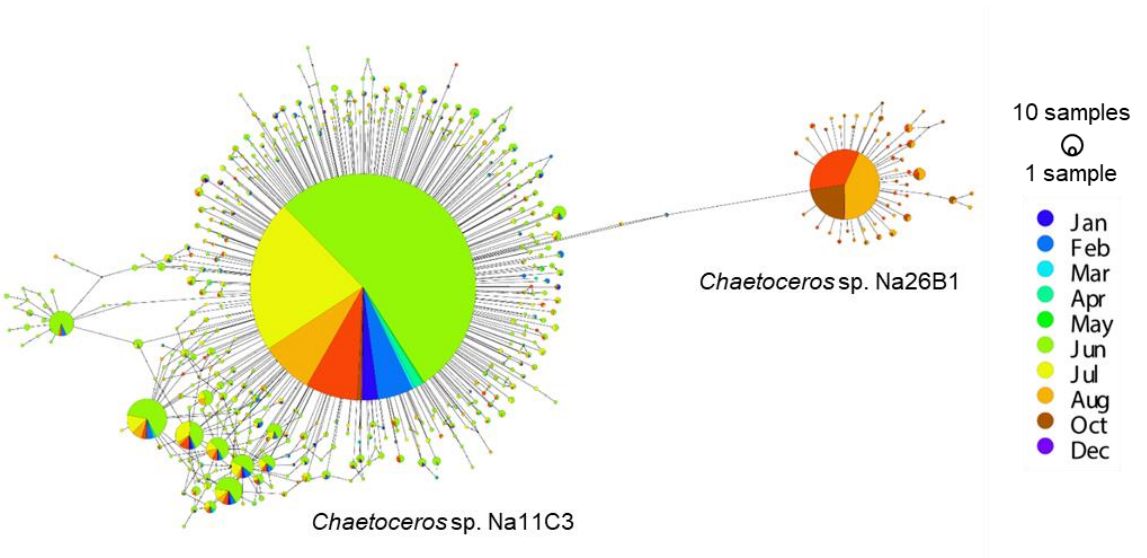


Fig. 5.6. TCS haplotype network for *Chaetoceros* sp. Na11C3 (left) and Na26B1 (right) inferred from the MareChiara temporal dataset. A total of 586 haplotypes (527 for *C.* sp. Na11C3 and 59 for *C.* sp. Na26B1) with abundance ≥ 2 across 2011 and 2013 was used. Sample in the legend refers to the number of reads.

Chaetoceros tenuissimus is perhaps the species in which the pattern of concerted evolution is more evident (Fig. 5.7). Indeed, almost all the nodes around the central one containing the dominant haplotype have a temporal pattern mimicking it. The visualisation of the first 50 most abundant haplotypes has reduced the noise due to haplotypes at low abundances (e.g. less than 10) that is observable in the networks of other species (e.g. Fig. 5.3, Fig. 5.5 and Fig. 5.6).

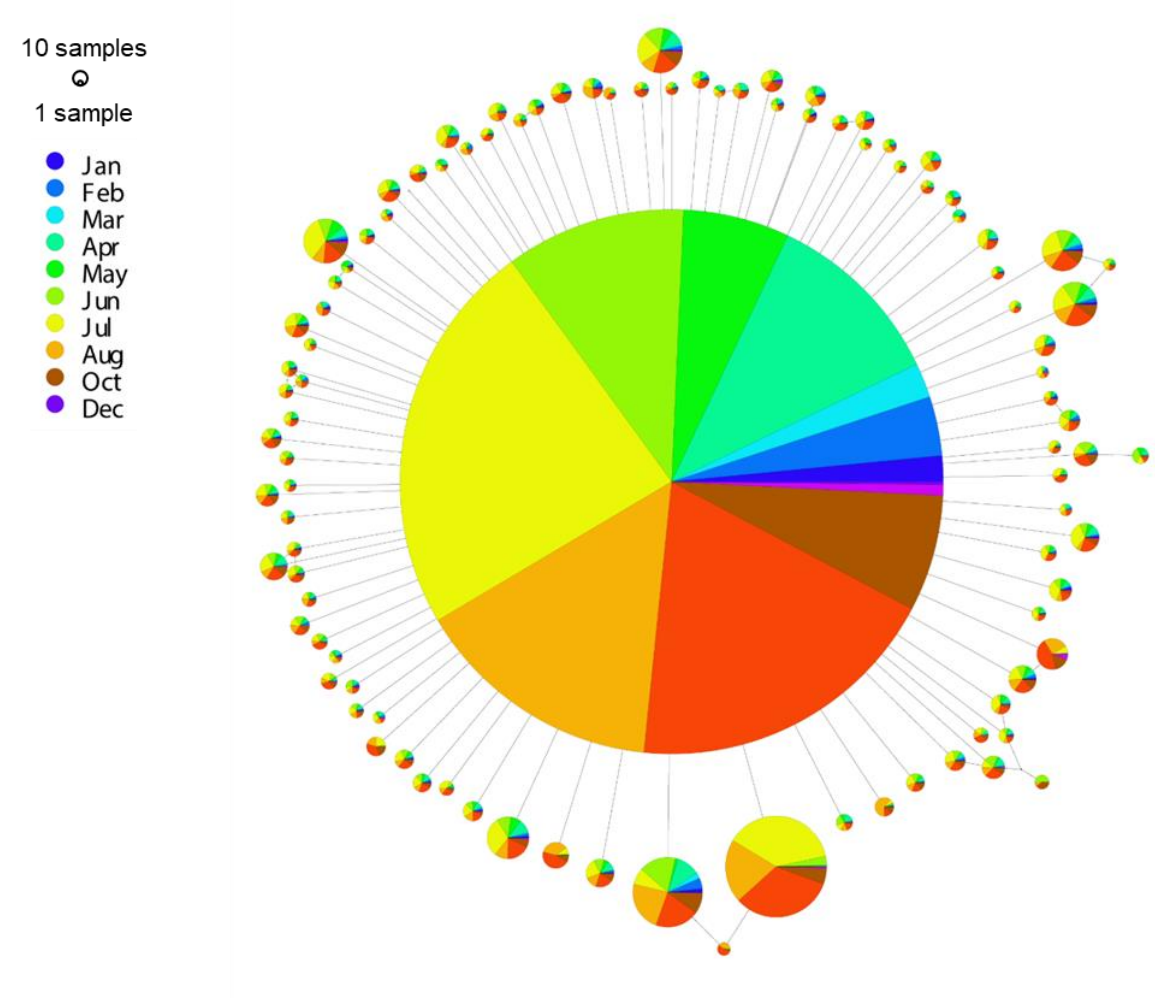


Fig. 5.7. TCS haplotype network for *C. tenuissimus* inferred from the MareChiara temporal dataset. A total of 102 haplotypes with abundance ≥ 50 across 2011 and 2013 was used. Sample in the legend refers to the number of reads.

5.4. Discussion

5.4.1. Concerted evolution in Chaetoceros

Since the first explanation of the process of concerted evolution in the rDNA cistron of *Xenopus* by Brown et al. (1972) using DNA-RNA hybridisation, this phenomenon has been observed and studied over years in different organisms using different techniques. Among the latter, Sanger sequencing of rDNA copies followed by phylogenetic analysis has been the most common approach (e.g. Vogler and DeSalle, 1994; Buckler et al., 1997; Li and Zhang, 2002; Xiao et al., 2010). In recent times, concerted evolution has also been

revealed by whole-genome shotgun sequence data (e.g. Ganley and Kobayashi, 2007) and chromosomal and array approaches (e.g. Kuhn et al., 2011; Bueno et al., 2016). However, to date there are no examples of studies that have dealt with concerted evolution using metabarcoding data or single-strain high throughput sequencing.

Thanks to the experimental design presented in this chapter, I have confirmed my hypothesis that the 18S gene is under concerted evolution in the *Chaetoceros* species here analysed. Furthermore, I have shown that it is possible to use a temporal metabarcoding dataset (with an adequate number of samples) to seek a first signal of this evolutionary phenomenon. Phylogenetic haplotype networks and the plots showing the distribution of the abundance of each haplotype were in accordance with the expectations of homogenisation. In particular, the occurrence in each strain and, more general, in each species of a haplotype (the "dominant" haplotype) far more abundant than all the others, confirmed my hypothesis of concerted evolution in action. In addition, the generation of single strain high throughput sequencing allowed me to prove at molecular level the patterns previously observed at level of ecological community (Gaonkar, 2017). This validation allowed distinguishing the presence of a real biological phenomenon due to infraspecific variation, instead of an artefact due to PCR errors or by-product of massive parallel sequencing. Based on the results obtained, I excluded that the variation found in the environment is an artefact of the methodology used. All the analyses here performed confirmed that the variation occurring in the temporal metabarcoding dataset is due to real variation present in the population and in representative individuals from that population. I did not perform any single-cell analysis, but instead, used a monoclonal culture of each *Chaetoceros* strain to perform high throughput sequencing, I am confident in asserting that the observed variation is intraindividual. This is because I have analysed the pattern of a multicopy gene that occurs in thousands of copies in the genome, and the probability that

any mutation possibly occurrying during culturing condition could have hampered the experiment is insignificant.

Minor variation among haplotypes is no sequencing artefact but results from concerted evolution not entirely succeeding in eliminating the emerging microvariation resulting from mutations and recombination. Indeed, BLAST analysis has shown that the haplotypes found in the environment also occur in the strains (are infraspecific variation). The abundance plots demonstrated also that both haplotypes from environmental metabarcoding and single strain HTS exhibit the same distribution pattern, with a dominant haplotype surrounded by several minor haplotypes. Furthermore, the dominant haplotypes of all the strains analysed were identical within the same species, as well as to these strains' Sanger sequences, and to the dominant metabarcode of that species in the environmental metabarcodes. This observed identity is in accordance with the way HTS and Sanger technologies work. A Sanger sequence can be considered as a consensus of all the targeted copies of a gene amplified. In this "consensus sequence", most of the weight will be carried by the most abundant sequence and therefore the Sanger sequence will read as the dominant haplotype. On the contrary, in massive parallel sequencing, every single copy present in the reaction tube will be sequenced, the only limit being constituted by reagents and sequencer characteristics. The dominant haplotype in the massive parallel sequencing is therefore the sequence that is "dominating" the aspect of the electropherogram in Sanger sequencing.

Based on my results, I hypothesise that in species in which besides concerted evolution other events have occurred, such as recent merging of two distinct populations, there might be multiple co-dominant haplotypes and their recombinants, a situation likely to result in messy, unreadable electropherograms. However, double peaks in electropherograms can also be due to different alleles occurring at similar frequencies in nuclear markers or to

heteroplasmy in the case of uniparental markers (e.g. mitochondrial and plastid genes). In this context, massive parallel sequencing can be of help at discriminating such situations.

5.4.2. Implications for DNA barcoding

Different regions of the rDNA cistron are targeted for DNA barcoding in several taxa. For example, the V4 region in the 18S gene is the currently recommended barcoding region for protists (Pawlowski et al., 2012), whilst the ITS region serves as such for fungi (Schoch et al., 2012). Some authors (e.g. Chase et al., 2007; Sonnenberg et al., 2007; Spooner, 2009) have argued that the concerted evolution process, known to affect ribosomal genes, may not be sufficiently effective to ensure complete sequence homogeneity. Therefore, knowing the extent of infraspecific variation and modality of evolution of such regions is vital to barcoding studies (Kane et al., 2012). The classical approach to the study of variants in rDNA genes is based on the cloning and Sanger sequencing of amplified products that produces noisy electropherograms. Studies targeting this region in different organisms revealed the occurrence of several different copies within each organism analysed and highlighted the potential risk for barcoding studies (e.g. Naidoo et al., 2013; Dakal et al., 2016). Indeed, one of the characteristics of a good DNA barcode is to have high interspecific divergence and low intraspecific variability (Kress and Erickson, 2008). Dakal et al. (2016) argued that the presence of several ribotypes within an individual shortens the barcoding gap and should be taken into consideration in barcoding studies of yeasts. However, what is lacking in these studies is information about the abundance of these "alternative" rDNA copies. Pillet et al. (2012) tried to predict the number of ribotypes in each specimen of *Elphidium macellum* (Foraminifera) correlating the number of clones screened with the number of ribotypes found. The authors argued that although some of less abundant ribotypes could be due to PCR artefacts, the high Spearman

correlation coefficient suggested that the real number of ribotypes in each individual could be underestimated (Pillet et al., 2012).

In this study, I have demonstrated that within each strain of several *Chaetoceros* species occur thousands of 18S ribotypes, one of which is far more abundant that all the others (the "dominant" haplotype). Because of such huge differences in abundance, the probability that a "minor" haplotype is sequenced with Sanger chemistry is almost null. In turn, this means that there is no risk associated to the use of the rDNA cistron as target gene in classical DNA barcoding studies. However, in metabarcoding studies these minor haplotypes can create a false rare diversity and therefore produce artefacts in diversity assessments.

My study also demonstrated that, when conducting metabarcoding experiments (from both environmental samples and bulk communities) or single strain HTS, the most abundant haplotype that is recovered for each species corresponds to the sequence that would be obtained by Sanger sequencing. Therefore, in case of a taxon for which a reference sequence is not available yet, the dominant haplotype retrieved from a metabarcoding dataset can be considered as such, and subsequently validated using Sanger sequencing when the specimen has been sampled.

5.4.3. Copy number across the Tree of Life and possible role of rDNA heterogeneity

The copy number of rDNA cistron has been estimated in different taxa along the Tree of Life. These studies have demonstrated that this number is highly variable: from 60 to 220 copies in fungi (Simon et al., 2005), 39-19,300 in animals (Prokopowich et al., 2003) and 150-26,048 in plants (Prokopowich et al., 2003). Among protists, ciliates harbour the highest number of rDNA copies, between 3,000 and 400,000 (Gong et al., 2013), followed by diatoms (1,057 to 12,812, Godhe et al., 2008) and dinoflagellates (200 to 1,200, Galluzzi et al., 2004). High variation among copies has been detected using a cloning and

sequencing approach in fungi (Simon and Weiß, 2008), dinoflagellates (Gribble and Anderson, 2007; Miranda et al., 2012), and Foraminifera (Pillet et al., 2012), as well as with genome sequencing in the plant genus *Asclepias* (Weitemier et al., 2015). However, the biological relevance of having many rDNA haplotypes is largely unknown. Part of such variation could be due imperfection of the mechanism that should homogenise all the copies among them. Another explanation, complementary to the former, is that there could be a selective advantage in possessing all these different copies. Indeed, in bacteria it has been shown that the number of copies of small rDNA gene correlates with the rate at which phylogenetically diverse bacteria respond to resource availability, with a high copy number leading to rapid colony formation (Klappenbach et al., 2000). In eukaryotes, the copy number of rDNA genes is unstable (Ganley and Kobayashi, 2014) and its stabilisation extends lifespan in yeast (Howitz et al., 2003). Always in yeast, it has been recently demonstrated that DNA replication stress induces a reduction in rDNA copy number in yeast (Salim et al., 2017). The possible role of rDNA heterogeneity in protists is to be unveiled yet.

5.4.4. Conclusions

In this chapter, I have shown how the analysis of ecological data by evolutionary approach can open unexpected scenarios. In this case, the analysis of temporal metabarcoding data analysed by phylogenetic networks, showed a pattern compatible with the theory of concerted evolution. The next experiment designed (the HTS of the strains) and the sets of analyses performed (plot of haplotype distribution, analysis of sequence similarity, evolutionary networks) confirmed the hypothesis in all the *Chaetoceros* species tested here, providing the first robust proof of concerted evolution in diatoms. Moreover, the simple approach to produce HTS of the strains can also be applied to other genera of diatoms or protists in order to understand the evolution of such gene region in different

marine taxa. In this sense, the use of metabarcoding or HTS data in general here shown is novel and powerful. However, the repercussions of this finding on metabarcoding studies are conflicting. On the one hand, I have demonstrated that the dominant haplotype perfectly matches with the Sanger reference sequence, validating the use of the metabarcoding technique for ecological studies. On the other hand, the high number of sequences occurring at low abundances (minor haplotypes) inflate the diversity assessments. In this study, I showed that at 99% of identity, all infraspecific variability is collapsed together. This is true for *Chaetoceros*, but the validity across other genera is to be tested yet. For studies using metabarcoding data at genus level, the clustering of sequences could be easily guided by evolutionary networks or trees, but for studies at community level the solution is more complicated. However, a possible course of action for future research could be to compare the results obtained in this study in *Chaetoceros* with other diatom and protist species, in order to understand the evolution of such gene region as well as the applicability of metabarcoding and high throughput sequencing in ecological and evolutionary studies in other marine organisms.

References

Andrea, L., Marini, M., Mantovani, B. (2006). Non-concerted evolution of the RET76 satellite DNA family in Reticulitermes taxa (Insecta, Isoptera). *Genetica, 128*(1-3), 123-132.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Britten, R. J., Waring, M. (1965). Renaturation of the DNA of higher organisms. *Carnegie Institution of Washington Yearbook, 64*, 316-333.

Britten, R. J., Kohne, D. E. (1968). Repeated sequences in DNA. *Science, 161*(3841), 529-540.

Brown, D. D., Wensink, P. C., Jordan, E. (1972). A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *Journal of Molecular Biology*, *63*(1), 57-73.

Buckler, E. S., Ippolito, A., Holtsford, T. P. (1997). The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics*, *145*(3), 821-832.

Bueno, D., Palacios-Gimenez, O. M., Martí, D. A., Mariguela, T. C., Cabral-de-Mello, D. C. (2016). The 5S rDNA in two *Abracris* grasshoppers (Ommatolampidinae: Acrididae): molecular and chromosomal organization. *Molecular Genetics and Genomics*, *291*(4), 1607-1613.

Charlesworth, B., Langley, C. H., Stephan, W. (1986). The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics*, *112*(4), 947-962.

Chase, M. W., Cowan, R. S., Hollingsworth, P. M., van den Berg, C., Madriñán, S., Petersen, G., ... Wilkinson, M. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon*, *56*(2), 295-299.

Clement, M., Posada, D., Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, *9*(10), 1657-1659.

Coen, E., Strachan, T., Dover, G. (1982). Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. *Journal of Molecular Biology*, *158*(1), 17-35.

Dakal, T. C., Giudici, P., Solieri, L. (2016). Contrasting patterns of rDNA homogenization within the *Zygosaccharomyces rouxii* species complex. *PLoS ONE*, *11*(8), e0160744.

Deng, W., Maust, B. S., Westfall, D. H., Chen, L., Zhao, H., Larsen, B. B., ... Mullins, J. I. (2013). Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. *Bioinformatics*, *29*(19), 2402-2409.

Elder Jr, J. F., Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology, 70*(3), 297-320.

Galluzzi, L., Penna, A., Bertozzini, E., Vila, M., Garcés, E., Magnani, M. (2004). Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a dinoflagellate). *Applied and Environmental Microbiology*, 70(2), 1199-1206.

Ganley, A. R. D., Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research, 17*(2), 184-191.

Ganley, A. R. D., Kobayashi, T. (2014). Ribosomal DNA and cellular senescence: new evidence supporting the connection between rDNA and aging. *FEMS Yeast Research, 14*(1), 49-59.

Gaonkar, C. C. (2017). *Diversity, Distribution and Evolution of the Planktonic Diatom Family Chaetocerotaceae*. PhD thesis, The Open University.

Godhe, A., Asplund, M. E., Härnström, K., Saravanan, V., Tyagi, A., Karunasagar, I. (2008). Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and Environmental Microbiology, 74*(23), 7174-7182.

Gojobori, T., Nei, M. (1984). Concerted evolution of the immunoglobulin VH gene family. *Molecular Biology and Evolution, 1*(2), 195-212.

Gong, J., Dong, J., Liu, X., Massana, R. (2013). Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist, 164*(3), 369-379.

Graur, D., Li, W.H. (1999). *Fundamentals of Molecular Evolution. Second edition*. Sinauer Associates, Sunderland, Massachusetts.

Gribble, K. E., Anderson, D. M. (2007). High intraindividual, intraspecific, and interspecific variability in large-subunit ribosomal DNA in the heterotrophic dinoflagellates *Protoperidinium*, *Diplopsalis*, and *Preperidinium* (Dinophyceae). *Phycologia*, *46*(3), 315-324.

Han, M. V., Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, *10*(1), 356.

Harpke, D., Peterson, A. (2006). Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Molecular Phylogenetics and Evolution*, *41*(3), 579-593.

Holliday, R. (1964). A mechanism for gene conversion in fungi. *Genetics Research*, *5*(2), 282-304.

Howitz, K. T., Bitterman, K. J., Cohen, H. Y., Lamming, D. W., Lavu, S., Wood, J. G., ... Sinclair, D. A. (2003). Small molecule activators of sirtuins extend *Saccharomyces cerevisiae* lifespan. *Nature*, *425*(6954), 191-196.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., Engels, J. M., Cronk, Q. (2012). Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany*, *99*(2), 320-329.

Katoh, K., Rozewicki, J., Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *108*, 1–7.

Klappenbach, J. A., Dunbar, J. M., Schmidt, T. M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, *66*(4), 1328-1333.

Kress, W. J., Erickson, D. L. (2008). DNA barcodes: genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences*, *105*(8), 2761-2762.

Kuhn, G. C., Küttler, H., Moreira-Filho, O., Heslop-Harrison, J. S. (2011). The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Molecular Biology and Evolution, 29*(1), 7-11.

Laehnemann, D., Borkhardt, A., McHardy, A. C. (2016). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics, 17*(1), 154-179.

Leigh, J. W., Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6*(9), 1110-1116.

Li, Y., Jiao, L., Yao, Y. J. (2013). Non-concerted ITS evolution in fungi, as revealed from the important medicinal fungus *Ophiocordyceps sinensis*. *Molecular Phylogenetics and Evolution, 68*(2), 373-379.

Li, D., Zhang, X. (2002). Physical Localization of the 18S-5· 8S-26S rDNA and Sequence Analysis of ITS Regions in *Thinopyrum ponticum* (Poaceae: Triticeae): Implications for Concerted Evolution. *Annals of Botany, 90*(4), 445-452.

Liebhaber, S. A., Goossens, M., Kan, Y. W. (1981). Homology and concerted evolution at the α1 and α2 loci of human α-globin. *Nature, 290*(5801), 26-29.

Lindegren, C. C. (1953). Gene conversion in Saccharomyces. *Journal of Genetics, 51*(3), 625-637.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology, 30*(5), 434-439.

Long, E. O., Dawid, I. B. (1980). Repeated genes in eukaryotes. *Annual Review of Biochemistry, 49*(1), 727-764.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), 10-12.

Miranda, L. N., Zhuang, Y., Zhang, H., Lin, S. (2012). Phylogenetic analysis guided by intragenomic SSU rDNA polymorphism refines classification of "*Alexandrium tamarense*" species complex. *Harmful Algae, 16*, 35-48.

Naidoo, K., Steenkamp, E. T., Coetzee, M. P., Wingfield, M. J., Wingfield, B. D. (2013). Concerted evolution in the ribosomal RNA cistron. *PLoS ONE, 8*(3), e59355.

Nei, M., Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics, 39*, 121-152.

Nikolaidis, N., Nei, M. (2004). Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. *Molecular Biology and Evolution, 21*(3), 498-505.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... Fiore-Donno, A. M. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology, 10*(11), e1001419.

Pillet, L., Fontaine, D., Pawlowski, J. (2012). Intra-genomic ribosomal RNA polymorphism and morphological variation in *Elphidium macellum* suggests inter-specific hybridization in Foraminifera. *PLoS ONE, 7*(2), e32373.

Piredda, R., Tomasino, M. P., D'erchia, A. M., Manzari, C., Pesole, G., Montresor, M., ... Zingone, A. (2016). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology, 93*(1), fiw200.

Price, M. N., Dehal, P. S., Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE, 5*(3), e9490.

Prokopowich, C. D., Gregory, T. R., Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome, 46*(1), 48-50.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Salim, D., Bradford, W. D., Freeland, A., Cady, G., Wang, J., Pruitt, S. C., Gerton, J. L. (2017). DNA replication stress restricts ribosomal DNA copy number. *PLoS Genetics, 13*(9), e1007006.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology, 75*(23), 7537-7541.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... Fungal Barcoding Consortium. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences, 109*(16), 6241-6246.

Simon, D., Moline, J., Helms, G., Friedl, T., Bhattacharya, D. (2005). Divergent histories of rDNA group I introns in the lichen family Physciaceae. *Journal of Molecular Evolution, 60*(4), 434-446.

Simon, U. K., Weiß, M. (2008). Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular Biology and Evolution, 25*(11), 2251-2254.

Sonnenberg, R., Nolte, A. W., Tautz, D. (2007). An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology, 4*(1), 6. doi:10.1186/1742-9994-4-6.

Spooner, D. M. (2009). DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. *American Journal of Botany, 96*(6), 1177-1189.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D., Breiner, H. W., Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a

highly complex eukaryotic community in marine anoxic water. *Molecular Ecology, 19,* 21-31.

Stoeck, T., Przybos, E., Dunthorn, M. (2014). The D 1-D 2 region of the large subunit ribosomal DNA as barcode for ciliates. *Molecular Ecology Resources, 14*(3), 458-468.

Vilnet, A., Konstantinova, N., Troitsky, A. (2012). Molecular phylogenetic data on reticulate evolution in the genus *Barbilophozia* Löske (Anastrophyllaceae, Marchantiophyta) and evidence of non-concerted evolution of rDNA in *Barbilophozia rubescens* allopolyploid. *Phytotaxa, 49*(1), 6-22.

Vogler, A. P., DeSalle, R. (1994). Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. *Molecular Biology and Evolution, 11*(3), 393-405.

Weitemier, K., Straub, S. C., Fishbein, M., Liston, A. (2015). Intragenomic polymorphisms among high-copy loci: a genus-wide study of nuclear ribosomal DNA in *Asclepias* (Apocynaceae). *PeerJ, 3,* e718.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wickham, H. (2018). scales: Scale Functions for Visualization. R package version 1.0.0. https://CRAN.R-project.org/package=scales.

Xiao, L. Q., Möller, M., Zhu, H. (2010). High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: incomplete concerted evolution and the origin of pseudogenes. *Molecular Phylogenetics and Evolution, 55*(1), 168-177.

Xu, B., Zeng, X. M., Gao, X. F., Jin, D. P., Zhang, L. B. (2017). ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Scientific Reports, 7,* 40057.

Zheng, X., Cai, D., Yao, L., Teng, Y. (2008). Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Molecular Phylogenetics and Evolution, 48*(3), 892-903.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., Wilson, A. C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences*, *77*(4), 2158-2162.

Appendix V

Table A5.1. List of the 50 most abundant haplotypes of MareChiara dataset and relative abundance. (A) *C. anastomosans*; (B) *C. costatus*; (C) *C. curvisetus* 2; (D) *Chaetoceros* sp. Na11C3; (E) *Chaetoceros* sp. Na26B1; (F) *C. tenuissimus*.

(A) *C. anastomosans*

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_2109_10899_14476	251
M00390_81_000000000-AA7DR_1_1111_15106_24806	7
M00390_80_000000000-AA759_1_1102_14009_18673	5
M00390_80_000000000-AA759_1_1104_18421_26995	4
M00390_81_000000000-AA7DR_1_1102_14691_18024	2
M00390_81_000000000-AA7DR_1_1103_24067_19555	2
M00390_80_000000000-AA759_1_1103_20661_12123	2
M00390_80_000000000-AA759_1_2108_19872_21486	2
M00390_80_000000000-AA759_1_1107_8089_22888	2
M00390_80_000000000-AA759_1_1101_25301_21125	2
M00390_80_000000000-AA759_1_2102_5911_13557	2
M00390_80_000000000-AA759_1_1109_4878_20074	2
M00390_80_000000000-AA759_1_1107_11931_10743	2
M00390_80_000000000-AA759_1_2103_22630_12031	2

(B) *C. costatus*

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_1112_20701_25092	7371
M00390_81_000000000-AA7DR_1_1101_15660_16312	63
M00390_81_000000000-AA7DR_1_1111_17729_14855	58
M00390_80_000000000-AA759_1_1103_9214_24775	53
M00390_81_000000000-AA7DR_1_1106_5105_16625	49
M00390_81_000000000-AA7DR_1_1105_3711_16331	47
M00390_80_000000000-AA759_1_2103_10847_8395	46
M00390_81_000000000-AA7DR_1_2101_22469_20255	45
M00390_81_000000000-AA7DR_1_1104_6258_7154	42
M00390_81_000000000-AA7DR_1_1106_23082_19031	42

M00390_81_000000000-AA7DR_1_1114_21762_16051	39
M00390_80_000000000-AA759_1_2109_20646_3307	36
M00390_81_000000000-AA7DR_1_2101_4247_16994	30
M00390_81_000000000-AA7DR_1_1103_13921_27562	23
M00390_81_000000000-AA7DR_1_1110_25882_24131	18
M00390_81_000000000-AA7DR_1_1107_10018_21762	16
M00390_80_000000000-AA759_1_1102_11565_2638	14
M00390_80_000000000-AA759_1_1102_7974_21450	14
M00390_80_000000000-AA759_1_2112_2026_13880	14
M00390_81_000000000-AA7DR_1_2103_5502_10960	13
M00390_81_000000000-AA7DR_1_1105_12696_21050	12
M00390_81_000000000-AA7DR_1_2107_8590_3849	12
M00390_81_000000000-AA7DR_1_1108_20055_11742	11
M00390_81_000000000-AA7DR_1_2105_17281_9878	11
M00390_81_000000000-AA7DR_1_2114_8675_4881	11
M00390_80_000000000-AA759_1_1102_24303_8392	10
M00390_80_000000000-AA759_1_1105_28775_12917	10
M00390_80_000000000-AA759_1_2108_12559_5387	10
M00390_81_000000000-AA7DR_1_1101_19338_9858	10
M00390_81_000000000-AA7DR_1_1102_17662_11871	10
M00390_81_000000000-AA7DR_1_1106_15323_19464	10
M00390_81_000000000-AA7DR_1_1109_16053_8409	10
M00390_81_000000000-AA7DR_1_1110_24824_25236	10
M00390_81_000000000-AA7DR_1_1114_26666_21508	10
M00390_81_000000000-AA7DR_1_2101_22791_7610	10
M00390_81_000000000-AA7DR_1_2102_15055_3213	10
M00390_81_000000000-AA7DR_1_2106_27602_17891	10
M00390_81_000000000-AA7DR_1_2114_2289_19059	10

(C) *C. curvisetus* 2

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_1101_24335_7294	6944
M00390_81_000000000-AA7DR_1_1109_20896_15345	453
M00390_81_000000000-AA7DR_1_1103_16288_26989	316
M00390_81_000000000-AA7DR_1_2103_28131_14780	200

M00390_81_000000000-AA7DR_1_1111_11450_2553	178
M00390_81_000000000-AA7DR_1_1101_4123_20747	118
M00390_81_000000000-AA7DR_1_1103_9271_16586	60
M00390_81_000000000-AA7DR_1_2104_16946_23798	48
M00390_81_000000000-AA7DR_1_1113_21922_15727	44
M00390_81_000000000-AA7DR_1_1112_21269_25157	34
M00390_81_000000000-AA7DR_1_1103_3248_11772	32
M00390_81_000000000-AA7DR_1_1108_10322_9155	28
M00390_81_000000000-AA7DR_1_1103_11972_18123	27
M00390_80_000000000-AA759_1_1101_6719_20035	20
M00390_80_000000000-AA759_1_1114_21412_18288	18
M00390_81_000000000-AA7DR_1_1106_11976_23254	17
M00390_80_000000000-AA759_1_2107_23256_10864	16
M00390_80_000000000-AA759_1_1104_25381_15393	14
M00390_80_000000000-AA759_1_1105_18022_3841	14
M00390_81_000000000-AA7DR_1_1107_28263_20903	14
M00390_81_000000000-AA7DR_1_1108_2122_12699	14
M00390_81_000000000-AA7DR_1_2105_8525_7938	14
M00390_80_000000000-AA759_1_2104_20058_6516	13
M00390_80_000000000-AA759_1_1114_19386_5392	12
M00390_80_000000000-AA759_1_2107_25970_19729	12
M00390_81_000000000-AA7DR_1_1111_10957_18509	12
M00390_81_000000000-AA7DR_1_1102_2708_17479	11
M00390_80_000000000-AA759_1_1101_22451_8109	10
M00390_80_000000000-AA759_1_1108_17827_14144	10
M00390_80_000000000-AA759_1_1108_21804_18503	9
M00390_80_000000000-AA759_1_1109_4803_21661	9
M00390_80_000000000-AA759_1_2108_15746_20584	9
M00390_80_000000000-AA759_1_1104_26914_13137	8
M00390_80_000000000-AA759_1_1113_27091_11483	8
M00390_80_000000000-AA759_1_2112_12719_20529	8
M00390_81_000000000-AA7DR_1_1102_4688_18121	8
M00390_81_000000000-AA7DR_1_1111_21225_11210	8
M00390_81_000000000-AA7DR_1_1111_23982_25410	8
M00390_81_000000000-AA7DR_1_2107_9396_11171	8
M00390_80_000000000-AA759_1_1107_6136_20046	7

M00390_80_000000000-AA759_1_1108_14909_22266	7
M00390_80_000000000-AA759_1_2104_18131_13043	7
M00390_81_000000000-AA7DR_1_1101_16629_13344	7
M00390_81_000000000-AA7DR_1_1101_2017_12577	7
M00390_81_000000000-AA7DR_1_1104_13759_24528	7
M00390_81_000000000-AA7DR_1_1104_24260_11901	7
M00390_81_000000000-AA7DR_1_1105_17846_11250	7
M00390_81_000000000-AA7DR_1_1114_11979_11975	7
M00390_81_000000000-AA7DR_1_2107_11725_16578	7
M00390_80_000000000-AA759_1_1105_12941_22574	6

(D) *Chaetoceros* sp. Na11C3

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_1101_6410_5509	9723
M00390_81_000000000-AA7DR_1_1111_20571_7220	304
M00390_81_000000000-AA7DR_1_1102_24763_6366	143
M00390_81_000000000-AA7DR_1_1101_11196_27974	139
M00390_81_000000000-AA7DR_1_1106_2424_15795	112
M00390_81_000000000-AA7DR_1_1103_25739_19440	105
M00390_81_000000000-AA7DR_1_1103_22194_6232	95
M00390_81_000000000-AA7DR_1_1110_26826_22971	53
M00390_81_000000000-AA7DR_1_1102_27566_10283	48
M00390_81_000000000-AA7DR_1_1102_17728_10049	46
M00390_81_000000000-AA7DR_1_1102_24856_10898	44
M00390_81_000000000-AA7DR_1_1109_4716_11773	34
M00390_81_000000000-AA7DR_1_1106_9998_3612	24
M00390_81_000000000-AA7DR_1_1105_25605_20745	24
M00390_81_000000000-AA7DR_1_1101_25478_16104	18
M00390_81_000000000-AA7DR_1_2103_19587_7983	17
M00390_81_000000000-AA7DR_1_2104_9541_20449	17
M00390_81_000000000-AA7DR_1_2111_16225_9778	17
M00390_81_000000000-AA7DR_1_2106_25799_8358	16
M00390_81_000000000-AA7DR_1_1109_12261_15569	16
M00390_81_000000000-AA7DR_1_1108_24721_11822	16
M00390_81_000000000-AA7DR_1_2103_24105_11792	15

M00390_81_000000000-AA7DR_1_1104_17994_3083	15
M00390_81_000000000-AA7DR_1_1102_9058_25666	15
M00390_81_000000000-AA7DR_1_1102_23857_26378	14
M00390_81_000000000-AA7DR_1_2103_16137_11494	14
M00390_81_000000000-AA7DR_1_2113_12929_19411	14
M00390_81_000000000-AA7DR_1_1113_27443_14488	13
M00390_81_000000000-AA7DR_1_1113_7128_23668	13
M00390_81_000000000-AA7DR_1_2104_13368_5663	13
M00390_81_000000000-AA7DR_1_1107_19739_2477	13
M00390_81_000000000-AA7DR_1_1102_15583_2449	13
M00390_81_000000000-AA7DR_1_1113_29235_17394	13
M00390_81_000000000-AA7DR_1_2112_11002_6236	13
M00390_81_000000000-AA7DR_1_1105_15319_18036	13
M00390_81_000000000-AA7DR_1_2111_24270_18806	12
M00390_81_000000000-AA7DR_1_1113_17825_8140	12
M00390_81_000000000-AA7DR_1_1104_19218_3762	12
M00390_81_000000000-AA7DR_1_1112_13352_5601	12
M00390_81_000000000-AA7DR_1_2105_20284_5220	12
M00390_81_000000000-AA7DR_1_2107_8323_22034	12
M00390_81_000000000-AA7DR_1_2102_20462_22578	11
M00390_81_000000000-AA7DR_1_1114_11222_26655	11
M00390_81_000000000-AA7DR_1_1104_22190_7531	11
M00390_81_000000000-AA7DR_1_1101_17105_13417	10
M00390_81_000000000-AA7DR_1_1112_9523_10659	10
M00390_81_000000000-AA7DR_1_2113_10927_8701	10
M00390_81_000000000-AA7DR_1_2109_27964_16114	10
M00390_81_000000000-AA7DR_1_1107_9732_26522	10
M00390_81_000000000-AA7DR_1_1113_16616_9226	10

(E) *Chaetoceros* sp. Na26B1

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_1101_16198_12414	941
M00390_81_000000000-AA7DR_1_1111_20391_3859	28
M00390_81_000000000-AA7DR_1_1101_9470_25874	13
M00390_81_000000000-AA7DR_1_1104_27372_19474	12

M00390_81_000000000-AA7DR_1_2105_25848_16533	11
M00390_81_000000000-AA7DR_1_2104_6697_8778	11
M00390_81_000000000-AA7DR_1_1109_19232_2903	9
M00390_81_000000000-AA7DR_1_1113_10186_25118	5
M00390_81_000000000-AA7DR_1_1104_22570_9473	5
M00390_81_000000000-AA7DR_1_1101_21520_5635	5
M00390_81_000000000-AA7DR_1_2108_15634_22907	4
M00390_81_000000000-AA7DR_1_1104_5102_14191	3
M00390_80_000000000-AA759_1_2108_18456_25450	3
M00390_80_000000000-AA759_1_2112_10861_13680	3
M00390_81_000000000-AA7DR_1_2101_5809_13550	3
M00390_81_000000000-AA7DR_1_2113_23019_13152	3
M00390_81_000000000-AA7DR_1_1110_13821_18158	3
M00390_81_000000000-AA7DR_1_2104_2589_16352	3
M00390_81_000000000-AA7DR_1_1102_17020_5965	3
M00390_81_000000000-AA7DR_1_2105_2682_19266	3
M00390_81_000000000-AA7DR_1_2107_6490_11438	3
M00390_80_000000000-AA759_1_1113_17535_10310	3
M00390_81_000000000-AA7DR_1_2111_19997_13380	3
M00390_81_000000000-AA7DR_1_1113_16061_26576	3
M00390_81_000000000-AA7DR_1_2104_11527_27123	3
M00390_81_000000000-AA7DR_1_1109_21257_16345	2
M00390_81_000000000-AA7DR_1_1114_20155_18983	2
M00390_81_000000000-AA7DR_1_2110_24767_14132	2
M00390_80_000000000-AA759_1_1111_7314_18049	2
M00390_81_000000000-AA7DR_1_1102_6690_11966	2
M00390_80_000000000-AA759_1_1114_14905_11144	2
M00390_81_000000000-AA7DR_1_2113_12146_19815	2
M00390_81_000000000-AA7DR_1_2105_24095_5901	2
M00390_81_000000000-AA7DR_1_1103_22901_8165	2
M00390_81_000000000-AA7DR_1_1103_17202_26051	2
M00390_81_000000000-AA7DR_1_1106_19721_26938	2
M00390_81_000000000-AA7DR_1_1108_8011_10045	2
M00390_80_000000000-AA759_1_1104_19320_19832	2
M00390_80_000000000-AA759_1_1107_13645_27564	2
M00390_80_000000000-AA759_1_1107_15354_12167	2

M00390_81_000000000-AA7DR_1_2107_20122_6543	2
M00390_80_000000000-AA759_1_2111_9841_2910	2
M00390_81_000000000-AA7DR_1_1105_28183_17335	2
M00390_81_000000000-AA7DR_1_2101_6397_18582	2
M00390_81_000000000-AA7DR_1_2110_7577_5967	2
M00390_81_000000000-AA7DR_1_1110_19023_18555	2
M00390_81_000000000-AA7DR_1_1103_7731_18975	2
M00390_81_000000000-AA7DR_1_2109_22304_14960	2
M00390_81_000000000-AA7DR_1_2112_8563_10293	2
M00390_81_000000000-AA7DR_1_2103_22817_8892	2

(F) *C. tenuissimus*

MareChiara haplotype	**abundance**
M00390_81_000000000-AA7DR_1_1101_19390_3055	102608
M00390_81_000000000-AA7DR_1_1101_17418_16093	3567
M00390_81_000000000-AA7DR_1_1111_22470_4248	1696
M00390_81_000000000-AA7DR_1_1101_19124_2835	704
M00390_81_000000000-AA7DR_1_1103_9398_12278	686
M00390_81_000000000-AA7DR_1_1101_25018_11322	673
M00390_81_000000000-AA7DR_1_1101_10404_7536	614
M00390_81_000000000-AA7DR_1_1101_19019_24975	578
M00390_81_000000000-AA7DR_1_1101_13180_26564	422
M00390_81_000000000-AA7DR_1_1108_24620_23287	337
M00390_81_000000000-AA7DR_1_1102_4023_9561	279
M00390_81_000000000-AA7DR_1_1101_21482_26412	272
M00390_81_000000000-AA7DR_1_1110_14154_2636	261
M00390_81_000000000-AA7DR_1_1101_7389_18078	259
M00390_81_000000000-AA7DR_1_1101_22681_24682	237
M00390_81_000000000-AA7DR_1_1104_3967_7974	204
M00390_81_000000000-AA7DR_1_2101_12520_13351	203
M00390_81_000000000-AA7DR_1_1102_7026_20095	183
M00390_81_000000000-AA7DR_1_1102_20566_15022	183
M00390_81_000000000-AA7DR_1_1102_21445_15640	177
M00390_81_000000000-AA7DR_1_2108_13462_27393	176
M00390_81_000000000-AA7DR_1_2105_5213_17961	172

M00390_81_000000000-AA7DR_1_1105_14401_23759	167
M00390_81_000000000-AA7DR_1_1105_20508_3587	155
M00390_81_000000000-AA7DR_1_1103_20846_6606	154
M00390_81_000000000-AA7DR_1_1110_20392_15399	145
M00390_81_000000000-AA7DR_1_1106_24238_20243	145
M00390_81_000000000-AA7DR_1_1102_18446_3596	142
M00390_81_000000000-AA7DR_1_1103_21707_22085	142
M00390_81_000000000-AA7DR_1_1103_16869_20837	142
M00390_81_000000000-AA7DR_1_1101_19579_23505	139
M00390_81_000000000-AA7DR_1_1112_4560_17960	138
M00390_81_000000000-AA7DR_1_1101_6592_8549	135
M00390_81_000000000-AA7DR_1_2102_25696_22060	131
M00390_81_000000000-AA7DR_1_1103_6137_19542	128
M00390_81_000000000-AA7DR_1_1101_25410_22880	125
M00390_81_000000000-AA7DR_1_2101_22809_10634	124
M00390_81_000000000-AA7DR_1_1108_16719_27399	124
M00390_81_000000000-AA7DR_1_1103_9573_27386	122
M00390_81_000000000-AA7DR_1_1105_12334_10159	118
M00390_81_000000000-AA7DR_1_1101_24859_6431	115
M00390_81_000000000-AA7DR_1_1107_22838_6403	114
M00390_81_000000000-AA7DR_1_1110_13450_22152	104
M00390_81_000000000-AA7DR_1_1102_21169_3932	103
M00390_81_000000000-AA7DR_1_1103_26982_11585	98
M00390_81_000000000-AA7DR_1_1107_22105_11345	94
M00390_81_000000000-AA7DR_1_2114_25314_9203	91
M00390_81_000000000-AA7DR_1_1112_1825_13639	91
M00390_81_000000000-AA7DR_1_1114_19091_18825	89
M00390_81_000000000-AA7DR_1_2102_24515_23008	89

Table A5.2. List of the 50 most abundant haplotypes in each strain and relative abundance. (A) *C. anastomosans*; (B) *C. costatus*; (C) *C. curvisetus* 2; (D) *Chaetoceros* sp. Na11C3; (E) *Chaetoceros* sp. Na26B1; (F) *C. tenuissimus*.

(A) *C. anastomosans*

Strain Na14C2		**Strain Na14C3**	
haplotype	**abundance**	**haplotype**	**abundance**
97KSI_03703_04635	415688	97KSI_00154_01105	425386
97KSI_01752_02923	8450	97KSI_00577_02517	2896
97KSI_04955_04530	2642	97KSI_03131_02564	2552
97KSI_04310_02070	2614	97KSI_03319_07327	2229
97KSI_01980_02933	2211	97KSI_01243_00573	1720
97KSI_04654_06561	1758	97KSI_02186_03827	1394
97KSI_01552_02038	1326	97KSI_03667_05941	1313
97KSI_03326_03053	1197	97KSI_02019_04661	1310
97KSI_05279_05020	1137	97KSI_01145_00779	1279
97KSI_03965_02929	1108	97KSI_04919_03953	1213
97KSI_00248_04839	1103	97KSI_05265_04107	1097
97KSI_01964_01784	1013	97KSI_03046_04251	1026
97KSI_03816_05769	976	97KSI_03775_05001	995
97KSI_03105_03043	965	97KSI_05249_04750	843
97KSI_04175_01929	796	97KSI_03376_00849	793
97KSI_00165_05792	787	97KSI_03553_02332	766
97KSI_00568_01197	706	97KSI_04764_06003	641
97KSI_00336_03283	613	97KSI_03031_02892	608
97KSI_03717_06068	596	97KSI_00478_01384	558
97KSI_02447_07481	531	97KSI_01772_02651	521
97KSI_03196_02045	529	97KSI_03655_04959	484
97KSI_04154_05725	477	97KSI_00857_03833	418
97KSI_02035_05239	455	97KSI_00955_01738	396
97KSI_01560_03047	404	97KSI_03270_04978	379
97KSI_03238_05718	385	97KSI_01084_01758	371
97KSI_02548_06343	369	97KSI_03644_05435	342
97KSI_04665_00700	355	97KSI_01172_01434	336
97KSI_02269_07091	328	97KSI_04142_04404	314

97KSI_03547_02119	304	97KSI_03298_04966	278
97KSI_00034_01952	291	97KSI_04874_02843	258
97KSI_01618_05102	270	97KSI_03771_03069	255
97KSI_00573_06877	269	97KSI_02225_02566	254
97KSI_02180_01133	268	97KSI_01576_06545	232
97KSI_03557_00823	255	97KSI_04634_01763	217
97KSI_00030_02400	254	97KSI_01486_06651	214
97KSI_00900_07078	251	97KSI_03032_01486	208
97KSI_04088_04995	247	97KSI_03334_07530	206
97KSI_02365_04076	246	97KSI_01143_04120	187
97KSI_05066_06224	208	97KSI_00278_03584	187
97KSI_00201_05488	191	97KSI_03940_00795	185
97KSI_00763_07243	182	97KSI_03995_05062	176
97KSI_03301_02434	153	97KSI_00634_01312	165
97KSI_02958_04282	152	97KSI_04734_03685	151
97KSI_03810_06648	148	97KSI_04481_03227	146
97KSI_05277_04300	145	97KSI_03217_03187	144
97KSI_01834_04060	145	97KSI_00127_04098	144
97KSI_03005_01905	143	97KSI_02835_03944	143
97KSI_03155_00406	143	97KSI_01840_03058	143
97KSI_03735_06349	138	97KSI_03710_00617	128
97KSI_02478_07476	137	97KSI_03269_07071	127

(B) *C. costatus*

Strain Na1A3		**Strain Na32B1**		**Strain Ro1B1**		**Strain Ro2A2**	
haplotype	**abundance**	**haplotype**	**abundance**	**haplotype**	**abundance**	**haplotype**	**abundance**
97KSI_03062_04287	218582	97KSI_00628_05777	379948	97KSI_02328_06963	247678	97KSI_02899_06467	206885
97KSI_04461_01525	2428	97KSI_01839_05734	7684	97KSI_04297_07118	4127	97KSI_04653_00957	5327
97KSI_00359_04771	2236	97KSI_03980_04497	3834	97KSI_03415_03185	3165	97KSI_00776_06235	2792
97KSI_00934_04772	2161	97KSI_00245_02587	3200	97KSI_01458_01176	2473	97KSI_03379_05484	2534
97KSI_03738_05059	1526	97KSI_03252_07546	2012	97KSI_01451_05587	1369	97KSI_03108_03694	1815
97KSI_03733	1488	97KSI_02626	1948	97KSI_00276	1292	97KSI_00461	1586

_06383		_04481		_02828		_03900	
97KSI_03566_01307	1182	97KSI_03425_05623	1936	97KSI_03331_00908	1197	97KSI_04453_04437	1089
97KSI_04193_00326	1035	97KSI_02826_02800	1766	97KSI_04180_00531	1187	97KSI_02194_06509	1051
97KSI_02146_04060	844	97KSI_01859_04368	1765	97KSI_03798_02577	963	97KSI_01997_03661	930
97KSI_00428_03998	841	97KSI_04321_01307	1698	97KSI_04398_02969	942	97KSI_03254_07267	849
97KSI_05083_06211	834	97KSI_00893_01512	1608	97KSI_03338_05011	799	97KSI_01028_00634	739
97KSI_00757_05950	769	97KSI_00684_05536	1324	97KSI_03392_04676	773	97KSI_02869_01248	657
97KSI_00328_04210	718	97KSI_00766_04377	1320	97KSI_03736_02517	751	97KSI_03886_00819	621
97KSI_00447_01854	602	97KSI_02083_06484	1316	97KSI_03878_02830	557	97KSI_03271_04439	565
97KSI_02676_01917	550	97KSI_05103_04700	1210	97KSI_02115_05093	452	97KSI_00762_01569	451
97KSI_00900_04043	370	97KSI_03802_00772	1121	97KSI_04812_06195	443	97KSI_01579_04283	347
97KSI_00084_03254	342	97KSI_00047_05046	1060	97KSI_05055_06665	437	97KSI_01025_02044	345
97KSI_04736_01030	325	97KSI_02666_06340	1058	97KSI_02974_07581	402	97KSI_02470_06461	322
97KSI_03255_05195	284	97KSI_03416_02170	756	97KSI_01900_07558	378	97KSI_02501_01075	310
97KSI_04923_06676	281	97KSI_01297_02043	692	97KSI_01611_00833	334	97KSI_03876_07253	245
97KSI_01914_00721	270	97KSI_00094_01653	679	97KSI_03526_00315	307	97KSI_00281_04404	227
97KSI_02612_02580	256	97KSI_02422_04520	625	97KSI_02492_03943	258	97KSI_03458_01614	216
97KSI_04675_03912	254	97KSI_00127_04231	484	97KSI_04823_02442	255	97KSI_04278_05885	192
97KSI_02446_01580	233	97KSI_00220_04465	446	97KSI_04666_02323	245	97KSI_02381_05846	185
97KSI_03666_01120	211	97KSI_03975_06608	371	97KSI_01536_02034	244	97KSI_00370_05371	184
97KSI_04563_03985	209	97KSI_00505_04671	354	97KSI_04748_03740	240	97KSI_02388_04250	181
97KSI_04272_05060	198	97KSI_02368_00940	331	97KSI_00426_05372	233	97KSI_03325_01069	175
97KSI_01217_04720	195	97KSI_00849_06887	322	97KSI_03620_06878	205	97KSI_04103_01395	165

97KSI_04115_01060	185	97KSI_04485_00803	307	97KSI_00552_01728	190	97KSI_00849_03412	156
97KSI_04778_06180	173	97KSI_00919_05841	306	97KSI_00993_01917	189	97KSI_04941_06156	154
97KSI_02430_06720	166	97KSI_01110_05077	305	97KSI_01041_06416	187	97KSI_02884_01016	140
97KSI_02255_01180	161	97KSI_00217_03921	298	97KSI_02571_02676	186	97KSI_03039_02937	137
97KSI_02255_01267	153	97KSI_01101_04050	280	97KSI_03825_06001	181	97KSI_04481_00662	133
97KSI_00425_02383	147	97KSI_03229_03447	272	97KSI_01370_05967	176	97KSI_04320_07295	129
97KSI_02036_04306	142	97KSI_04137_01183	255	97KSI_01177_05276	167	97KSI_00084_03847	129
97KSI_03261_03215	140	97KSI_03622_06607	249	97KSI_02743_03417	158	97KSI_01818_04512	113
97KSI_04172_02622	135	97KSI_04444_05213	229	97KSI_01933_05741	139	97KSI_01135_00710	112
97KSI_01125_01348	127	97KSI_03587_00348	224	97KSI_03468_00604	139	97KSI_00635_02930	98
97KSI_04637_00929	126	97KSI_00183_01315	221	97KSI_00970_06792	138	97KSI_01793_03053	97
97KSI_03306_02496	124	97KSI_03610_04478	202	97KSI_01985_06185	135	97KSI_01371_04503	95
97KSI_04767_05381	118	97KSI_02433_02259	196	97KSI_02829_01690	125	97KSI_04968_03484	92
97KSI_02697_01205	118	97KSI_01688_05159	194	97KSI_02441_00804	118	97KSI_01140_01861	89
97KSI_01280_03473	112	97KSI_00874_07345	194	97KSI_04833_03704	118	97KSI_04402_05294	84
97KSI_01124_06743	108	97KSI_04196_05067	190	97KSI_04444_04547	117	97KSI_01872_03686	83
97KSI_04242_05794	107	97KSI_01336_05679	190	97KSI_01514_01423	117	97KSI_01569_06297	80
97KSI_02500_06487	107	97KSI_04771_01799	168	97KSI_00284_04176	108	97KSI_01559_01938	80
97KSI_04999_03479	105	97KSI_04712_03820	166	97KSI_03280_00715	105	97KSI_03373_06479	79
97KSI_01644_01248	103	97KSI_02553_03747	161	97KSI_05218_04741	102	97KSI_03779_07288	77
97KSI_00512_02828	100	97KSI_00163_02311	160	97KSI_01260_05796	96	97KSI_02121_06291	76
97KSI_02519_03431	100	97KSI_05224_02816	152	97KSI_01469_07366	95	97KSI_02521_02232	74

(C) *C. curvisetus* 2

Strain Ch5B2		**Strain Na1C1**		**Strain Na19A2**		**Strain Na20A4**	
haplotype	**abundance**	**haplotype**	**abundance**	**haplotype**	**abundance**	**haplotype**	**abundance**
97KSI_04187_04119	169486	97KSI_02988_02860	38657	97KSI_01232_02148	130036	97KSI_01261_05089	127340
97KSI_01106_03045	2193	97KSI_02471_00580	1828	97KSI_03985_03982	2062	97KSI_01722_01232	2045
97KSI_00309_01364	1250	97KSI_04759_01311	1220	97KSI_01407_02088	1429	97KSI_01647_01659	1524
97KSI_01166_02524	1244	97KSI_00792_03430	712	97KSI_00902_00336	976	97KSI_04983_01608	936
97KSI_04760_02997	1145	97KSI_03343_07160	300	97KSI_04096_00422	901	97KSI_05130_06551	854
97KSI_02598_04820	951	97KSI_01161_05983	282	97KSI_05226_02343	690	97KSI_01615_00840	751
97KSI_01200_03715	898	97KSI_00302_02313	258	97KSI_04140_04400	682	97KSI_01784_01551	685
97KSI_04592_04855	749	97KSI_04935_03621	248	97KSI_01430_00514	644	97KSI_01783_06836	637
97KSI_01198_07171	734	97KSI_00304_02419	222	97KSI_03663_01767	534	97KSI_02103_02111	522
97KSI_02817_06374	547	97KSI_03970_02605	205	97KSI_00167_01641	441	97KSI_01656_03986	455
97KSI_01153_01578	539	97KSI_03163_06734	159	97KSI_03421_02183	359	97KSI_00310_01923	445
97KSI_04523_06290	518	97KSI_01093_01329	146	97KSI_01228_01919	355	97KSI_04272_03655	383
97KSI_00800_05590	511	97KSI_04278_06681	141	97KSI_00440_01212	355	97KSI_04561_00703	330
97KSI_03983_02065	431	97KSI_04914_06693	119	97KSI_01997_06893	355	97KSI_01915_04040	317
97KSI_03898_07206	401	97KSI_02864_04008	108	97KSI_01722_05814	322	97KSI_04127_00729	315
97KSI_01092_03272	355	97KSI_02828_07250	98	97KSI_04152_05193	320	97KSI_00846_05987	310
97KSI_00312_01166	340	97KSI_03777_00812	97	97KSI_00147_04517	297	97KSI_01232_01249	247
97KSI_04002_06564	337	97KSI_04824_04867	93	97KSI_02957_01043	265	97KSI_02158_07491	226
97KSI_02219_02561	303	97KSI_00384_04781	93	97KSI_04224_03867	251	97KSI_01896_03565	213
97KSI_02526_05461	294	97KSI_01749_07525	89	97KSI_05257_04824	238	97KSI_02472_02752	200
97KSI_01095	273	97KSI_05239	88	97KSI_03777	216	97KSI_05240	194

_03786		_01905		_04242		_04057	
97KSI_04945_03799	232	97KSI_00390_05805	71	97KSI_02284_05741	203	97KSI_02902_01662	185
97KSI_04797_02731	232	97KSI_00179_06381	70	97KSI_02522_01005	201	97KSI_01313_04662	184
97KSI_04369_00832	193	97KSI_00841_03065	68	97KSI_00627_03203	179	97KSI_04159_05252	163
97KSI_03520_01532	189	97KSI_04658_06207	64	97KSI_02241_07073	168	97KSI_03857_04134	155
97KSI_01235_00479	189	97KSI_00164_03428	63	97KSI_04204_05991	168	97KSI_02892_05702	154
97KSI_01745_01180	186	97KSI_01617_01725	60	97KSI_00788_05695	161	97KSI_04199_06533	140
97KSI_01129_02122	182	97KSI_00655_02063	60	97KSI_03233_02899	140	97KSI_05261_05641	128
97KSI_00233_01226	163	97KSI_01839_03444	57	97KSI_01150_06389	124	97KSI_04066_02721	118
97KSI_05082_02758	160	97KSI_03993_05173	54	97KSI_01472_03020	123	97KSI_02366_02585	109
97KSI_04969_01211	146	97KSI_00475_02161	50	97KSI_00670_03093	118	97KSI_03596_00164	94
97KSI_01567_06490	142	97KSI_04283_04904	46	97KSI_03840_05921	111	97KSI_02399_01540	93
97KSI_04918_01993	131	97KSI_03187_06770	45	97KSI_00436_03352	99	97KSI_01192_01527	92
97KSI_00291_01293	127	97KSI_00484_06656	44	97KSI_04117_05593	92	97KSI_05062_06515	92
97KSI_04516_06891	125	97KSI_05169_02592	44	97KSI_04751_05373	92	97KSI_05027_04129	92
97KSI_00354_04116	123	97KSI_02301_02704	43	97KSI_00821_01684	90	97KSI_03248_02447	91
97KSI_05059_01647	120	97KSI_00412_01918	41	97KSI_03880_03291	87	97KSI_00965_06607	88
97KSI_01749_03291	113	97KSI_01389_05988	38	97KSI_04942_05066	85	97KSI_04268_06469	87
97KSI_03466_02986	107	97KSI_00880_05224	38	97KSI_05268_02917	84	97KSI_03288_02240	82
97KSI_02187_05081	105	97KSI_01339_05825	38	97KSI_04321_05569	84	97KSI_02123_02627	82
97KSI_03037_03176	99	97KSI_03473_03134	37	97KSI_03837_00354	79	97KSI_01259_01169	77
97KSI_00500_04676	99	97KSI_02343_02905	36	97KSI_04046_05963	75	97KSI_01327_01761	72
97KSI_01635_03769	95	97KSI_04641_01976	34	97KSI_04046_05735	74	97KSI_00512_06372	71

97KSI_02223_03527	91	97KSI_00183_03877	34	97KSI_05022_02817	74	97KSI_04237_05733	69
97KSI_04408_04138	88	97KSI_02548_03583	34	97KSI_02974_01918	68	97KSI_00428_04657	66
97KSI_01432_01091	85	97KSI_00519_03708	32	97KSI_02191_05610	68	97KSI_04225_05852	63
97KSI_03267_01697	85	97KSI_00390_06887	31	97KSI_03698_00345	67	97KSI_03042_06349	61
97KSI_04292_00898	84	97KSI_02070_06063	30	97KSI_02031_07198	64	97KSI_01755_07378	57
97KSI_01710_03470	80	97KSI_01820_05872	30	97KSI_00604_02344	63	97KSI_01719_00881	56
97KSI_04277_05380	76	97KSI_04584_06203	29	97KSI_03424_02279	63	97KSI_03106_07585	56

(D) *Chaetoceros* sp. Na11C3

Strain Na11C3		**Strain Na43A1**	
haplotype	**abundance**	**haplotype**	**abundance**
97KSI_03663_01512	520646	97KSI_05086_04284	260766
97KSI_01342_06051	4195	97KSI_02675_04149	2804
97KSI_00143_02567	2986	97KSI_05077_02274	2363
97KSI_02348_07517	2877	97KSI_02364_06877	1538
97KSI_04012_04611	2642	97KSI_03896_06799	1477
97KSI_05274_03658	2035	97KSI_04969_03877	1415
97KSI_03415_05238	1943	97KSI_00933_04623	1374
97KSI_02801_00679	1861	97KSI_01108_02216	1231
97KSI_00313_02395	1500	97KSI_03678_01523	1046
97KSI_01574_04337	1415	97KSI_01402_03563	1024
97KSI_04502_04721	1253	97KSI_02210_02507	1008
97KSI_04085_03270	1214	97KSI_01126_05573	678
97KSI_04244_03016	1135	97KSI_04891_04888	670
97KSI_01133_03016	1015	97KSI_04085_02981	592
97KSI_03019_02005	987	97KSI_04710_02408	569
97KSI_00369_06436	769	97KSI_02632_05026	499
97KSI_01782_01635	716	97KSI_02256_07312	499
97KSI_00695_03999	697	97KSI_03830_00852	381
97KSI_00338_04739	631	97KSI_03157_05429	379
97KSI_05130_06212	610	97KSI_01936_02618	361

97KSI_04067_05041	490	97KSI_00470_01961	324
97KSI_00699_01850	463	97KSI_00321_03450	323
97KSI_04670_06771	446	97KSI_02720_02413	295
97KSI_00290_03562	442	97KSI_04880_02258	291
97KSI_01313_07257	441	97KSI_01053_01155	289
97KSI_02105_00631	434	97KSI_00706_01364	285
97KSI_01663_04289	406	97KSI_03581_07341	237
97KSI_03960_06202	381	97KSI_02780_04771	222
97KSI_04860_01981	368	97KSI_01439_02089	218
97KSI_02001_07437	355	97KSI_00474_03281	210
97KSI_00379_01738	354	97KSI_00930_04263	205
97KSI_02044_03097	345	97KSI_02248_07093	193
97KSI_05095_02527	315	97KSI_00605_05979	189
97KSI_04158_05183	307	97KSI_02477_01188	188
97KSI_02442_05201	301	97KSI_01208_01935	184
97KSI_01502_02279	301	97KSI_00578_03681	165
97KSI_03676_03501	282	97KSI_04031_04385	160
97KSI_05160_04842	274	97KSI_03910_06636	157
97KSI_01245_01743	273	97KSI_03780_01562	153
97KSI_02878_04016	267	97KSI_05148_03204	147
97KSI_03488_06956	250	97KSI_00415_01199	144
97KSI_00241_05213	242	97KSI_03355_04371	132
97KSI_01819_01022	240	97KSI_04900_03452	127
97KSI_02893_03642	221	97KSI_01187_03219	116
97KSI_01207_04749	218	97KSI_03857_01144	111
97KSI_03622_02935	211	97KSI_00834_06616	110
97KSI_01917_04547	209	97KSI_02957_07379	108
97KSI_01116_06337	193	97KSI_03619_02630	105
97KSI_00388_02382	193	97KSI_03042_00082	102
97KSI_04276_05558	189	97KSI_03335_06255	100

(E) *Chaetoceros* sp. Na26B1

Strain Na26B1	
haplotype	**abundance**
97KSI_01986_05212	277568
97KSI_03665_06329	3153

97KSI_04190_02727	1589
97KSI_00244_01964	1475
97KSI_03873_06785	1355
97KSI_03445_07568	1151
97KSI_02227_03525	1011
97KSI_03148_07131	932
97KSI_03876_04249	793
97KSI_04041_07453	763
97KSI_04283_00952	670
97KSI_00030_04326	624
97KSI_02890_03453	586
97KSI_04274_01473	525
97KSI_04440_02001	512
97KSI_03286_02649	430
97KSI_02922_01051	366
97KSI_04530_03070	362
97KSI_00963_05797	361
97KSI_04837_05019	314
97KSI_04106_03770	311
97KSI_00956_04969	305
97KSI_03789_01597	282
97KSI_03156_00523	255
97KSI_01224_05079	253
97KSI_02360_01917	253
97KSI_01739_00466	240
97KSI_04093_05675	234
97KSI_01022_00717	208
97KSI_02078_01146	202
97KSI_02771_00919	200
97KSI_02581_00543	188
97KSI_01203_00783	186
97KSI_00396_06508	175
97KSI_03062_03130	172
97KSI_04257_04121	166
97KSI_03025_01007	158
97KSI_03486_05518	148

97KSI_00497_06768	137
97KSI_03099_02556	136
97KSI_00186_01558	130
97KSI_01200_02891	124
97KSI_02695_01278	122
97KSI_03449_06178	118
97KSI_04225_02879	111
97KSI_03216_01548	110
97KSI_03040_07506	106
97KSI_00382_06202	106
97KSI_01606_05328	105
97KSI_01213_02474	95

(F) *C. tenuissimus*

Strain GB2a		**Strain Na26A1**		**Strain Na44A1**	
haplotype	**abundance**	**haplotype**	**abundance**	**haplotype**	**abundance**
97KSI_00416_02071	173565	97KSI_03171_05542	152018	97KSI_03436_06826	194285
97KSI_04642_06542	35170	97KSI_04894_03826	12696	97KSI_04590_06245	3252
97KSI_02170_01835	2362	97KSI_01619_03486	2362	97KSI_00987_03631	2640
97KSI_00830_02955	1849	97KSI_03111_06431	1931	97KSI_01520_02095	1240
97KSI_01788_02370	1163	97KSI_01659_06282	1077	97KSI_03196_06342	1089
97KSI_02477_05763	1097	97KSI_04937_05270	939	97KSI_01644_07473	726
97KSI_01971_07513	628	97KSI_03701_07245	583	97KSI_02090_05953	672
97KSI_00424_05431	605	97KSI_00722_03534	461	97KSI_03952_04583	615
97KSI_00326_00927	573	97KSI_04639_01300	452	97KSI_04492_01277	589
97KSI_04044_01178	571	97KSI_00576_03603	404	97KSI_04250_03866	587
97KSI_01112_02654	562	97KSI_01878_01695	343	97KSI_04749_02314	556
97KSI_01372_06128	494	97KSI_03866_06205	334	97KSI_04316_02987	454

97KSI_02195_05216	480	97KSI_05158_03582	324	97KSI_04034_05350	436
97KSI_05189_02224	403	97KSI_00651_04989	314	97KSI_03501_06489	354
97KSI_04045_06681	379	97KSI_01196_06645	295	97KSI_01303_03497	347
97KSI_04829_01720	351	97KSI_02081_03281	278	97KSI_04917_04159	298
97KSI_00301_04344	339	97KSI_03506_06519	254	97KSI_01488_03168	297
97KSI_01056_00315	333	97KSI_01741_07170	246	97KSI_02770_03320	254
97KSI_02347_01990	264	97KSI_01789_02080	222	97KSI_03487_03614	250
97KSI_02988_06637	242	97KSI_01767_04459	221	97KSI_04433_05038	247
97KSI_03746_05752	229	97KSI_03939_06957	211	97KSI_03404_02308	241
97KSI_00295_04180	229	97KSI_01102_00348	211	97KSI_04607_02502	227
97KSI_01187_02484	228	97KSI_04115_04425	177	97KSI_02123_04822	227
97KSI_02624_05741	226	97KSI_03788_05681	174	97KSI_02978_06414	219
97KSI_03283_05750	218	97KSI_01082_04725	161	97KSI_01510_01464	188
97KSI_02572_01604	216	97KSI_03767_05679	160	97KSI_02847_05417	169
97KSI_04520_03370	210	97KSI_05211_01456	145	97KSI_04978_05894	166
97KSI_03192_05461	199	97KSI_02376_04767	140	97KSI_02805_05175	160
97KSI_04613_06707	185	97KSI_05165_04141	130	97KSI_04346_06095	156
97KSI_01283_00600	181	97KSI_01005_07120	124	97KSI_00260_06263	155
97KSI_00716_06321	169	97KSI_00852_07059	122	97KSI_02069_03301	150
97KSI_03836_03871	164	97KSI_00826_05849	118	97KSI_03547_03341	146
97KSI_04152_02234	162	97KSI_01687_05543	113	97KSI_01793_01862	146
97KSI_02857_03435	161	97KSI_04583_02736	110	97KSI_05247_05539	143
97KSI_02055_02315	156	97KSI_01300_06380	106	97KSI_00065_02605	142

97KSI_04590_04548	150	97KSI_00362_02892	100	97KSI_01470_05884	135
97KSI_03686_01461	146	97KSI_04243_01143	100	97KSI_02523_06617	113
97KSI_04968_03228	143	97KSI_01489_06155	100	97KSI_00662_04965	111
97KSI_00793_04628	136	97KSI_02287_04277	100	97KSI_03533_06713	95
97KSI_00181_01728	131	97KSI_01058_03849	98	97KSI_03799_05793	95
97KSI_01418_01150	130	97KSI_02747_03509	97	97KSI_04125_01427	93
97KSI_00647_01272	129	97KSI_01051_04105	92	97KSI_00098_03753	92
97KSI_03612_06186	117	97KSI_03769_05802	91	97KSI_02907_02930	91
97KSI_00585_04337	116	97KSI_01363_07006	88	97KSI_01146_03689	87
97KSI_01798_07171	114	97KSI_00876_06594	86	97KSI_03794_02175	83
97KSI_00711_01124	114	97KSI_05231_01969	79	97KSI_04582_06501	82
97KSI_01741_07266	109	97KSI_01808_06608	79	97KSI_03682_02708	77
97KSI_00779_07182	108	97KSI_01199_05862	66	97KSI_01707_06393	73
97KSI_00236_03105	108	97KSI_01177_05951	64	97KSI_01801_04441	73
97KSI_03215_06439	107	97KSI_01867_03639	63	97KSI_01465_05425	71

Table A5.3. Percentage of identity between MareChiara haplotypes (query) and single strain ones (subject) after BLAST analysis.

Species	MareChiara haplotype	Single strain haplotype	% identity
C. anastomosans	M00390_80_000000000-AA759_1_1101_25301_21125	97KSI_02594_00890	100.00
	M00390_80_000000000-AA759_1_1102_14009_18673	97KSI_04102_07418	99.73
	M00390_80_000000000-AA759_1_1103_20661_12123	97KSI_00508_05944	100.00
	M00390_80_000000000-AA759_1_1104_18421_26995	97KSI_00908_01002	99.73
	M00390_80_000000000-AA759_1_1107_11931_10743	97KSI_02895_07021	100.00
	M00390_80_000000000-AA759_1_1107_8089_22888	97KSI_02466_02824	99.73
	M00390_80_000000000-AA759_1_1109_4878_20074	97KSI_00345_02439	99.73
	M00390_80_000000000-AA759_1_2102_5911_13557	97KSI_03016_04504	99.47
	M00390_80_000000000-AA759_1_2103_22630_12031	97KSI_03068_02457	99.73
	M00390_80_000000000-AA759_1_2108_19872_21486	97KSI_00443_03292	100.00
	M00390_81_000000000-AA7DR_1_1102_14691_18024	97KSI_01530_01155	99.73
	M00390_81_000000000-AA7DR_1_1103_24067_19555	97KSI_02722_02195	100.00
	M00390_81_000000000-AA7DR_1_1111_15106_24806	97KSI_03390_05249	99.73
	M00390_81_000000000-AA7DR_1_2109_10899_14476	97KSI_03703_04635	100.00
C. costatus	M00390_80_000000000-AA759_1_1102_11565_2638	97KSI_02291_06733	100.00
	M00390_80_000000000-AA759_1_1102_24303_8392	97KSI_00771_01206	100.00
	M00390_80_000000000-AA759_1_1102_7974_21450	97KSI_03318_04406	100.00
	M00390_80_000000000-AA759_1_1103_9214_24775	97KSI_03819_02360	99.74
	M00390_80_000000000-AA759_1_1105_28775_12917	97KSI_03425_05623	100.00
	M00390_80_000000000-AA759_1_2103_10847_8395	97KSI_04461_01525	100.00
	M00390_80_000000000-AA759_1_2108_12559_5387	97KSI_00595_04612	99.74
	M00390_80_000000000-AA759_1_2109_20646_3307	97KSI_03178_01877	100.00
	M00390_80_000000000-AA759_1_2112_2026_13880	97KSI_01201_05876	100.00
	M00390_81_000000000-AA7DR_1_1101_15660_16312	97KSI_03738_05059	100.00
	M00390_81_000000000-AA7DR_1_1101_19338_9858	97KSI_01923_04779	99.74
	M00390_81_000000000-AA7DR_1_1102_17662_11871	97KSI_00109_06212	100.00
	M00390_81_000000000-AA7DR_1_1103_13921_27562	97KSI_04517_03069	100.00
	M00390_81_000000000-AA7DR_1_1104_6258_7154	97KSI_03733_06383	100.00
	M00390_81_000000000-AA7DR_1_1105_12696_21050	97KSI_05089_03751	100.00
	M00390_81_000000000-AA7DR_1_1105_3711_16331	97KSI_00371_06560	100.00

	M00390_81_000000000-AA7DR_1_1106_15323_19464	97KSI_00757_05950	100.00
	M00390_81_000000000-AA7DR_1_1106_23082_19031	97KSI_05086_06023	99.74
	M00390_81_000000000-AA7DR_1_1106_5105_16625	97KSI_02188_05262	99.74
	M00390_81_000000000-AA7DR_1_1107_10018_21762	97KSI_05254_01755	100.00
	M00390_81_000000000-AA7DR_1_1108_20055_11742	97KSI_03905_04827	100.00
	M00390_81_000000000-AA7DR_1_1109_16053_8409	97KSI_03947_04173	100.00
	M00390_81_000000000-AA7DR_1_1110_24824_25236	97KSI_00054_04565	99.74
	M00390_81_000000000-AA7DR_1_1110_25882_24131	97KSI_00710_04425	100.00
	M00390_81_000000000-AA7DR_1_1111_17729_14855	97KSI_03089_07525	100.00
	M00390_81_000000000-AA7DR_1_1112_20701_25092	97KSI_03062_04287	100.00
	M00390_81_000000000-AA7DR_1_1114_21762_16051	97KSI_04115_01060	100.00
	M00390_81_000000000-AA7DR_1_1114_26666_21508	97KSI_01012_01800	100.00
	M00390_81_000000000-AA7DR_1_2101_22469_20255	97KSI_04678_04248	99.74
	M00390_81_000000000-AA7DR_1_2101_22791_7610	97KSI_01490_01634	100.00
	M00390_81_000000000-AA7DR_1_2101_4247_16994	97KSI_00611_03023	100.00
	M00390_81_000000000-AA7DR_1_2102_15055_3213	97KSI_00447_01854	100.00
	M00390_81_000000000-AA7DR_1_2103_5502_10960	97KSI_04599_02674	100.00
	M00390_81_000000000-AA7DR_1_2105_17281_9878	97KSI_00923_04621	99.74
	M00390_81_000000000-AA7DR_1_2106_27602_17891	97KSI_00530_01306	99.74
	M00390_81_000000000-AA7DR_1_2107_8590_3849	97KSI_04172_02622	100.00
	M00390_81_000000000-AA7DR_1_2114_2289_19059	97KSI_04330_04615	99.74
	M00390_81_000000000-AA7DR_1_2114_8675_4881	97KSI_00134_02557	100.00
***C. curvisetus* 2**	M00390_40_000000000-A6D16_1_1103_9508_5675	97KSI_00743_05902	100.00
	M00390_40_000000000-A6D16_1_1108_17804_11598	97KSI_04781_05500	99.74
	M00390_40_000000000-A6D16_1_1110_27567_9849	97KSI_03132_05290	99.74
	M00390_40_000000000-A6D16_1_1111_3542_11205	97KSI_03868_07408	100.00
	M00390_80_000000000-AA759_1_1101_11743_22610	97KSI_05189_05534	100.00
	M00390_80_000000000-AA759_1_1101_13895_25837	97KSI_04229_03933	99.74
	M00390_80_000000000-AA759_1_1101_22451_8109	97KSI_04700_05670	99.74
	M00390_80_000000000-AA759_1_1101_5468_24005	97KSI_03212_04472	100.00
	M00390_80_000000000-AA759_1_1101_6719_20035	97KSI_03028_04597	99.74
	M00390_80_000000000-AA759_1_1102_10472_8793	97KSI_02148_01247	99.74
	M00390_80_000000000-AA759_1_1102_24564_7650	97KSI_01344_03236	99.74
	M00390_80_000000000-AA759_1_1102_28325_10692	97KSI_01373_02440	99.74
	M00390_80_000000000-AA759_1_1103_10502_22016	97KSI_04567_02701	99.74

	M00390_80_000000000-AA759_1_1103_10733_19662	97KSI_00698_06928	99.74
	M00390_80_000000000-AA759_1_1103_20985_3495	97KSI_05102_02709	99.74
	M00390_80_000000000-AA759_1_1103_5353_19137	97KSI_03868_07408	99.74
	M00390_80_000000000-AA759_1_1103_7200_5218	97KSI_03612_05711	99.48
	M00390_80_000000000-AA759_1_1104_16447_20731	97KSI_02376_07020	100.00
	M00390_80_000000000-AA759_1_1104_17935_15234	97KSI_01122_03542	99.74
	M00390_80_000000000-AA759_1_1104_25381_15393	97KSI_03132_05290	100.00
	M00390_80_000000000-AA759_1_1104_26195_19039	97KSI_04516_04950	99.74
	M00390_80_000000000-AA759_1_1104_26914_13137	97KSI_02785_06860	100.00
	M00390_80_000000000-AA759_1_1105_10346_14685	97KSI_00461_03779	100.00
	M00390_80_000000000-AA759_1_1105_11673_22226	97KSI_03408_02291	100.00
	M00390_80_000000000-AA759_1_1105_12941_22574	97KSI_00781_04953	99.74
	M00390_80_000000000-AA759_1_1105_13984_10638	97KSI_03062_05179	100.00
	M00390_80_000000000-AA759_1_1105_18022_3841	97KSI_04187_04119	99.74
	M00390_80_000000000-AA759_1_1105_22867_22691	97KSI_04509_03201	100.00
	M00390_80_000000000-AA759_1_1106_18744_19999	97KSI_03155_05120	99.74
	M00390_80_000000000-AA759_1_1106_20541_6865	97KSI_04187_04119	99.74
	M00390_80_000000000-AA759_1_1106_4543_9571	97KSI_03612_05711	99.74
	M00390_80_000000000-AA759_1_1107_14518_27111	97KSI_01833_06973	100.00
	M00390_80_000000000-AA759_1_1107_19330_2927	97KSI_04567_02701	99.48
	M00390_80_000000000-AA759_1_1107_20145_4453	97KSI_04977_05303	100.00
	M00390_80_000000000-AA759_1_1107_23028_11214	97KSI_03304_05267	100.00
	M00390_80_000000000-AA759_1_1107_26025_8835	97KSI_01648_05844	99.74
	M00390_80_000000000-AA759_1_1107_5485_13612	97KSI_03052_03627	100.00
	M00390_80_000000000-AA759_1_1107_6136_20046	97KSI_03298_05173	99.74
	M00390_80_000000000-AA759_1_1108_14909_22266	97KSI_04567_02701	99.74
	M00390_80_000000000-AA759_1_1108_17827_14144	97KSI_01225_01807	99.74
	M00390_80_000000000-AA759_1_1108_18885_3870	97KSI_04167_01691	100.00
	M00390_80_000000000-AA759_1_1108_21041_24244	97KSI_00653_00804	100.00
	M00390_80_000000000-AA759_1_1108_21045_24223	97KSI_03873_04426	99.74
	M00390_80_000000000-AA759_1_1108_21804_18503	97KSI_01225_01807	99.48
	M00390_80_000000000-AA759_1_1109_10112_22292	97KSI_03449_02999	100.00
	M00390_80_000000000-AA759_1_1109_12487_7144	97KSI_03474_03329	99.74
	M00390_80_000000000-AA759_1_1109_20449_24122	97KSI_04964_06069	100.00
	M00390_80_000000000-AA759_1_1109_22338_15912	97KSI_04516_04950	99.74
	M00390_80_000000000-AA759_1_1109_22400_3193	97KSI_04516_04950	99.74

	M00390_80_000000000-AA759_1_1109_23473_17636	97KSI_01613_07083	100.00
	M00390_80_000000000-AA759_1_1109_4803_21661	97KSI_01376_05247	99.74
	M00390_80_000000000-AA759_1_1111_11786_2903	97KSI_01745_01180	100.00
	M00390_80_000000000-AA759_1_1111_19795_21223	97KSI_00831_05888	100.00
	M00390_80_000000000-AA759_1_1111_20150_4074	97KSI_04567_02701	99.74
	M00390_80_000000000-AA759_1_1111_24228_17282	97KSI_02911_07583	100.00
	M00390_80_000000000-AA759_1_1111_2812_18777	97KSI_04488_06894	99.74
	M00390_80_000000000-AA759_1_1111_2829_18786	97KSI_04488_06894	100.00
	M00390_80_000000000-AA759_1_1111_8624_21360	97KSI_04740_01204	99.74
	M00390_80_000000000-AA759_1_1112_11357_13565	97KSI_00867_07123	100.00
	M00390_80_000000000-AA759_1_1112_11663_13037	97KSI_00930_00962	100.00
	M00390_80_000000000-AA759_1_1112_22787_23939	97KSI_02977_04061	100.00
	M00390_80_000000000-AA759_1_1113_13540_17170	97KSI_04599_02055	99.48
	M00390_80_000000000-AA759_1_1113_19343_13398	97KSI_03764_04133	100.00
	M00390_80_000000000-AA759_1_1113_27091_11483	97KSI_03505_03492	100.00
	M00390_80_000000000-AA759_1_1113_27098_11499	97KSI_03505_03492	99.74
	M00390_80_000000000-AA759_1_1114_17684_18347	97KSI_04911_00819	100.00
	M00390_80_000000000-AA759_1_1114_19386_5392	97KSI_02927_04010	99.74
	M00390_80_000000000-AA759_1_1114_21412_18288	97KSI_04740_01204	100.00
	M00390_80_000000000-AA759_1_2102_14339_11186	97KSI_01137_04815	100.00
	M00390_80_000000000-AA759_1_2103_15394_7041	97KSI_04898_06275	99.74
	M00390_80_000000000-AA759_1_2104_10763_17096	97KSI_02194_02911	100.00
	M00390_80_000000000-AA759_1_2104_11244_3581	97KSI_00809_01739	100.00
	M00390_80_000000000-AA759_1_2104_18131_13043	97KSI_03612_05711	100.00
	M00390_80_000000000-AA759_1_2104_18695_10584	97KSI_04369_00832	100.00
	M00390_80_000000000-AA759_1_2104_19422_28482	97KSI_01336_03485	99.74
	M00390_80_000000000-AA759_1_2104_20058_6516	97KSI_00086_01422	99.74
	M00390_80_000000000-AA759_1_2105_12470_6531	97KSI_03569_02826	99.74
	M00390_80_000000000-AA759_1_2105_15923_6293	97KSI_04397_03614	100.00
	M00390_80_000000000-AA759_1_2106_11644_17478	97KSI_02828_02782	100.00
	M00390_80_000000000-AA759_1_2106_20114_18359	97KSI_01699_04103	100.00
	M00390_80_000000000-AA759_1_2106_23048_15729	97KSI_01958_06937	99.74
	M00390_80_000000000-AA759_1_2106_7808_13667	97KSI_04519_01603	100.00
	M00390_80_000000000-AA759_1_2107_13577_17164	97KSI_01635_03769	100.00
	M00390_80_000000000-AA759_1_2107_18121_26034	97KSI_04585_00974	100.00
	M00390_80_000000000-AA759_1_2107_23035_13452	97KSI_04514_04780	100.00

	M00390_80_000000000-AA759_1_2107_23256_10864	97KSI_03814_02058	99.74
	M00390_80_000000000-AA759_1_2107_25970_19729	97KSI_02136_06907	100.00
	M00390_80_000000000-AA759_1_2107_8480_12514	97KSI_03153_05335	99.74
	M00390_80_000000000-AA759_1_2108_15746_20584	97KSI_02216_02513	99.74
	M00390_80_000000000-AA759_1_2108_18539_23855	97KSI_04567_02701	99.74
	M00390_80_000000000-AA759_1_2108_21126_10885	97KSI_01202_06478	99.74
	M00390_80_000000000-AA759_1_2109_14137_19482	97KSI_01092_03272	100.00
	M00390_80_000000000-AA759_1_2109_21731_7346	97KSI_04700_05670	99.74
	M00390_80_000000000-AA759_1_2111_14210_25717	97KSI_04187_04119	99.74
	M00390_80_000000000-AA759_1_2111_16199_3148	97KSI_04294_00609	100.00
	M00390_80_000000000-AA759_1_2111_21201_19038	97KSI_04187_04119	100.00
	M00390_80_000000000-AA759_1_2111_5293_24441	97KSI_01745_01180	99.74
	M00390_80_000000000-AA759_1_2111_9689_14487	97KSI_01392_06664	100.00
	M00390_80_000000000-AA759_1_2112_12719_20529	97KSI_01478_00819	99.74
	M00390_80_000000000-AA759_1_2112_22183_15635	97KSI_03759_04687	100.00
	M00390_80_000000000-AA759_1_2112_26137_9178	97KSI_04599_02055	100.00
	M00390_80_000000000-AA759_1_2112_8585_7259	97KSI_05182_05960	99.74
	M00390_80_000000000-AA759_1_2113_15141_8792	97KSI_04567_02701	99.74
	M00390_80_000000000-AA759_1_2113_18012_8213	97KSI_00781_04953	99.74
	M00390_80_000000000-AA759_1_2113_19783_6098	97KSI_02148_01247	100.00
	M00390_80_000000000-AA759_1_2113_21317_28124	97KSI_02457_04915	99.74
	M00390_80_000000000-AA759_1_2113_8415_18494	97KSI_03303_06741	99.48
	M00390_80_000000000-AA759_1_2114_24627_13417	97KSI_00681_01119	99.74
	M00390_80_000000000-AA759_1_2114_25670_10168	97KSI_01478_00819	99.48
	M00390_80_000000000-AA759_1_2114_6455_19260	97KSI_04567_02701	99.74
	M00390_80_000000000-AA759_1_2114_7640_22822	97KSI_01613_07083	99.48
	M00390_80_000000000-AA759_1_2114_9719_14121	97KSI_03517_00836	100.00
	M00390_81_000000000-AA7DR_1_1101_13965_14741	97KSI_01235_00479	100.00
	M00390_81_000000000-AA7DR_1_1101_15023_10826	97KSI_03294_06151	100.00
	M00390_81_000000000-AA7DR_1_1101_15538_15864	97KSI_01046_01756	100.00
	M00390_81_000000000-AA7DR_1_1101_16237_17867	97KSI_02916_02536	100.00
	M00390_81_000000000-AA7DR_1_1101_16304_22244	97KSI_01322_06142	100.00
	M00390_81_000000000-AA7DR_1_1101_16629_13344	97KSI_04700_05670	99.74
	M00390_81_000000000-AA7DR_1_1101_17013_14347	97KSI_01479_06698	100.00
	M00390_81_000000000-AA7DR_1_1101_17523_22387	97KSI_04363_01439	99.74
	M00390_81_000000000-AA7DR_1_1101_19571_16176	97KSI_02797_04630	100.00

	M00390_81_000000000-AA7DR_1_1101_20071_12642	97KSI_04323_03374	99.74
	M00390_81_000000000-AA7DR_1_1101_2017_12577	97KSI_03931_03021	100.00
	M00390_81_000000000-AA7DR_1_1101_21181_9211	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_1101_23913_9308	97KSI_00630_02928	100.00
	M00390_81_000000000-AA7DR_1_1101_24335_7294	97KSI_04187_04119	100.00
	M00390_81_000000000-AA7DR_1_1101_24585_16205	97KSI_03895_06718	100.00
	M00390_81_000000000-AA7DR_1_1101_4123_20747	97KSI_04516_04950	100.00
	M00390_81_000000000-AA7DR_1_1101_5912_13393	97KSI_03762_04340	99.74
	M00390_81_000000000-AA7DR_1_1102_10301_25144	97KSI_01503_06678	99.74
	M00390_81_000000000-AA7DR_1_1102_12202_15435	97KSI_00389_02330	100.00
	M00390_81_000000000-AA7DR_1_1102_13087_23674	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1102_15618_9839	97KSI_00416_04850	100.00
	M00390_81_000000000-AA7DR_1_1102_18328_16560	97KSI_00692_06393	100.00
	M00390_81_000000000-AA7DR_1_1102_20565_21004	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1102_21274_21190	97KSI_02187_05081	100.00
	M00390_81_000000000-AA7DR_1_1102_23425_16102	97KSI_01228_06101	99.74
	M00390_81_000000000-AA7DR_1_1102_2708_17479	97KSI_04777_05330	100.00
	M00390_81_000000000-AA7DR_1_1102_4688_18121	97KSI_01708_06316	100.00
	M00390_81_000000000-AA7DR_1_1103_10087_12788	97KSI_03078_02645	99.74
	M00390_81_000000000-AA7DR_1_1103_11972_18123	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1103_13048_24800	97KSI_04943_01339	100.00
	M00390_81_000000000-AA7DR_1_1103_13712_25055	97KSI_03917_00716	99.74
	M00390_81_000000000-AA7DR_1_1103_16288_26989	97KSI_04599_02055	99.74
	M00390_81_000000000-AA7DR_1_1103_18238_11044	97KSI_03242_02279	100.00
	M00390_81_000000000-AA7DR_1_1103_23644_24204	97KSI_02484_02014	100.00
	M00390_81_000000000-AA7DR_1_1103_24932_21898	97KSI_04884_01290	100.00
	M00390_81_000000000-AA7DR_1_1103_3248_11772	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1103_7687_21888	97KSI_02337_03380	100.00
	M00390_81_000000000-AA7DR_1_1103_8660_16022	97KSI_03006_06832	100.00
	M00390_81_000000000-AA7DR_1_1103_9271_16586	97KSI_04700_05670	100.00
	M00390_81_000000000-AA7DR_1_1104_10534_24738	97KSI_00645_04722	100.00
	M00390_81_000000000-AA7DR_1_1104_10856_6132	97KSI_03958_00212	99.74
	M00390_81_000000000-AA7DR_1_1104_11311_3647	97KSI_00402_04614	99.74
	M00390_81_000000000-AA7DR_1_1104_11912_4321	97KSI_00510_03356	100.00
	M00390_81_000000000-AA7DR_1_1104_13759_24528	97KSI_04187_04119	99.74
	M00390_81_000000000-AA7DR_1_1104_13933_13441	97KSI_04053_03472	100.00

	M00390_81_000000000-AA7DR_1_1104_15906_14284	97KSI_01553_01460	100.00
	M00390_81_000000000-AA7DR_1_1104_17413_12794	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1104_17661_20107	97KSI_01309_03321	100.00
	M00390_81_000000000-AA7DR_1_1104_19050_21812	97KSI_04738_06689	100.00
	M00390_81_000000000-AA7DR_1_1104_19881_12816	97KSI_04537_05455	100.00
	M00390_81_000000000-AA7DR_1_1104_24260_11901	97KSI_00362_02486	100.00
	M00390_81_000000000-AA7DR_1_1105_11776_23421	97KSI_01033_04984	99.74
	M00390_81_000000000-AA7DR_1_1105_12813_20116	97KSI_01261_03143	100.00
	M00390_81_000000000-AA7DR_1_1105_17846_11250	97KSI_04452_06333	100.00
	M00390_81_000000000-AA7DR_1_1105_20843_14258	97KSI_04700_05670	99.74
	M00390_81_000000000-AA7DR_1_1105_2109_17189	97KSI_04400_03556	99.74
	M00390_81_000000000-AA7DR_1_1105_21846_21709	97KSI_03281_06196	100.00
	M00390_81_000000000-AA7DR_1_1105_21981_8612	97KSI_04187_04119	100.00
	M00390_81_000000000-AA7DR_1_1105_23437_11795	97KSI_04256_05357	100.00
	M00390_81_000000000-AA7DR_1_1105_24588_13039	97KSI_01975_02265	99.74
	M00390_81_000000000-AA7DR_1_1105_5005_12698	97KSI_02492_01578	99.74
	M00390_81_000000000-AA7DR_1_1105_8241_13607	97KSI_04704_05753	99.74
	M00390_81_000000000-AA7DR_1_1106_10037_24355	97KSI_01376_05247	99.48
	M00390_81_000000000-AA7DR_1_1106_11382_10477	97KSI_01131_01904	100.00
	M00390_81_000000000-AA7DR_1_1106_11976_23254	97KSI_04213_06207	99.74
	M00390_81_000000000-AA7DR_1_1106_12010_23897	97KSI_01381_04784	100.00
	M00390_81_000000000-AA7DR_1_1106_13620_7801	97KSI_04740_01204	99.74
	M00390_81_000000000-AA7DR_1_1106_14632_23952	97KSI_04187_04119	100.00
	M00390_81_000000000-AA7DR_1_1106_19291_13940	97KSI_04700_05670	99.74
	M00390_81_000000000-AA7DR_1_1106_19693_18158	97KSI_00500_04676	100.00
	M00390_81_000000000-AA7DR_1_1106_20593_21372	97KSI_03423_03772	100.00
	M00390_81_000000000-AA7DR_1_1106_20985_7779	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_1106_2628_14595	97KSI_02881_04736	100.00
	M00390_81_000000000-AA7DR_1_1106_4722_10235	97KSI_04694_05800	100.00
	M00390_81_000000000-AA7DR_1_1106_6581_15713	97KSI_02077_05387	100.00
	M00390_81_000000000-AA7DR_1_1106_7821_12016	97KSI_01051_05261	100.00
	M00390_81_000000000-AA7DR_1_1107_12787_9145	97KSI_00233_01226	100.00
	M00390_81_000000000-AA7DR_1_1107_16021_15870	97KSI_05130_02585	99.74
	M00390_81_000000000-AA7DR_1_1107_20388_17913	97KSI_04599_02055	99.48
	M00390_81_000000000-AA7DR_1_1107_24621_22614	97KSI_00382_04346	100.00
	M00390_81_000000000-AA7DR_1_1107_25117_20516	97KSI_01537_06394	100.00

	M00390_81_000000000-AA7DR_1_1107_28263_20903	97KSI_04709_01874	100.00
	M00390_81_000000000-AA7DR_1_1107_3250_20669	97KSI_01101_04979	100.00
	M00390_81_000000000-AA7DR_1_1108_10322_9155	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_1108_12794_24631	97KSI_02027_05875	99.74
	M00390_81_000000000-AA7DR_1_1108_13168_23363	97KSI_04459_05934	100.00
	M00390_81_000000000-AA7DR_1_1108_13693_10959	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_1108_16364_8323	97KSI_03887_04659	100.00
	M00390_81_000000000-AA7DR_1_1108_16650_22482	97KSI_01183_05812	100.00
	M00390_81_000000000-AA7DR_1_1108_2122_12699	97KSI_04187_04119	100.00
	M00390_81_000000000-AA7DR_1_1108_23772_25200	97KSI_04709_01874	99.74
	M00390_81_000000000-AA7DR_1_1108_6254_16529	97KSI_04621_00925	100.00
	M00390_81_000000000-AA7DR_1_1108_8069_19260	97KSI_00797_03564	100.00
	M00390_81_000000000-AA7DR_1_1108_8120_22089	97KSI_05117_03368	100.00
	M00390_81_000000000-AA7DR_1_1108_8675_7801	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_1108_9560_13088	97KSI_03716_06622	100.00
	M00390_81_000000000-AA7DR_1_1109_12067_10106	97KSI_01033_04984	100.00
	M00390_81_000000000-AA7DR_1_1109_14468_20736	97KSI_00536_04951	100.00
	M00390_81_000000000-AA7DR_1_1109_16920_17915	97KSI_00966_04610	100.00
	M00390_81_000000000-AA7DR_1_1109_18275_20293	97KSI_00738_03516	100.00
	M00390_81_000000000-AA7DR_1_1109_18303_27624	97KSI_04599_02055	99.48
	M00390_81_000000000-AA7DR_1_1109_18809_20901	97KSI_00809_01739	99.74
	M00390_81_000000000-AA7DR_1_1109_20896_15345	97KSI_04567_02701	100.00
	M00390_81_000000000-AA7DR_1_1109_25683_7709	97KSI_04599_02055	99.48
	M00390_81_000000000-AA7DR_1_1109_27444_12760	97KSI_01225_01807	99.48
	M00390_81_000000000-AA7DR_1_1109_7114_11295	97KSI_04232_05452	99.74
	M00390_81_000000000-AA7DR_1_1110_10678_22797	97KSI_04740_06629	100.00
	M00390_81_000000000-AA7DR_1_1110_12644_15465	97KSI_04304_01090	100.00
	M00390_81_000000000-AA7DR_1_1110_13579_11643	97KSI_03885_03542	99.74
	M00390_81_000000000-AA7DR_1_1110_15284_13262	97KSI_01392_06664	99.74
	M00390_81_000000000-AA7DR_1_1110_20140_11098	97KSI_02279_02709	100.00
	M00390_81_000000000-AA7DR_1_1110_22128_7754	97KSI_01033_04984	99.74
	M00390_81_000000000-AA7DR_1_1110_22410_16801	97KSI_00614_05119	99.74
	M00390_81_000000000-AA7DR_1_1110_23222_14938	97KSI_00474_06618	100.00
	M00390_81_000000000-AA7DR_1_1110_24983_14727	97KSI_03608_04036	100.00
	M00390_81_000000000-AA7DR_1_1110_5200_17928	97KSI_04523_06290	100.00
	M00390_81_000000000-AA7DR_1_1110_6826_9413	97KSI_02940_05822	100.00

	M00390_81_000000000-AA7DR_1_1110_7895_25543	97KSI_04187_04119	99.74
	M00390_81_000000000-AA7DR_1_1111_10820_10057	97KSI_03419_06418	99.74
	M00390_81_000000000-AA7DR_1_1111_10957_18509	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_1111_11450_2553	97KSI_03419_06418	100.00
	M00390_81_000000000-AA7DR_1_1111_11655_22760	97KSI_05153_02135	99.74
	M00390_81_000000000-AA7DR_1_1111_12745_18011	97KSI_01936_07377	99.74
	M00390_81_000000000-AA7DR_1_1111_14758_9950	97KSI_01795_00693	100.00
	M00390_81_000000000-AA7DR_1_1111_17102_24469	97KSI_03419_06418	99.74
	M00390_81_000000000-AA7DR_1_1111_21225_11210	97KSI_03419_06418	99.74
	M00390_81_000000000-AA7DR_1_1111_23759_6534	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_1111_23982_25410	97KSI_03419_06418	99.74
	M00390_81_000000000-AA7DR_1_1111_4473_13567	97KSI_03419_06418	99.48
	M00390_81_000000000-AA7DR_1_1111_4991_12258	97KSI_03419_06418	99.74
	M00390_81_000000000-AA7DR_1_1112_10694_20471	97KSI_01973_03126	100.00
	M00390_81_000000000-AA7DR_1_1112_21269_25157	97KSI_04905_06288	99.74
	M00390_81_000000000-AA7DR_1_1112_23142_16163	97KSI_02136_06907	99.74
	M00390_81_000000000-AA7DR_1_1112_25078_11343	97KSI_00681_01119	100.00
	M00390_81_000000000-AA7DR_1_1113_10609_21251	97KSI_01843_06084	100.00
	M00390_81_000000000-AA7DR_1_1113_10765_18473	97KSI_05112_05261	100.00
	M00390_81_000000000-AA7DR_1_1113_12685_22420	97KSI_04771_06275	100.00
	M00390_81_000000000-AA7DR_1_1113_14513_14532	97KSI_00996_02030	100.00
	M00390_81_000000000-AA7DR_1_1113_20888_17331	97KSI_03875_07024	99.74
	M00390_81_000000000-AA7DR_1_1113_21488_11049	97KSI_04958_02941	99.74
	M00390_81_000000000-AA7DR_1_1113_21922_15727	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_1113_22720_18263	97KSI_00916_05827	100.00
	M00390_81_000000000-AA7DR_1_1113_6128_17592	97KSI_02957_04702	100.00
	M00390_81_000000000-AA7DR_1_1113_6426_8240	97KSI_04187_04119	99.74
	M00390_81_000000000-AA7DR_1_1114_11979_11975	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_1114_13133_7846	97KSI_03666_03024	100.00
	M00390_81_000000000-AA7DR_1_1114_18358_7289	97KSI_01219_04037	100.00
	M00390_81_000000000-AA7DR_1_1114_22337_17115	97KSI_01225_01807	99.48
	M00390_81_000000000-AA7DR_1_1114_24438_25877	97KSI_04599_02055	99.74
	M00390_81_000000000-AA7DR_1_1114_25565_18230	97KSI_03684_06839	100.00
	M00390_81_000000000-AA7DR_1_1114_25609_5817	97KSI_04363_05641	99.74
	M00390_81_000000000-AA7DR_1_1114_7843_16895	97KSI_00300_06550	100.00
	M00390_81_000000000-AA7DR_1_2101_10014_19409	97KSI_00942_03578	100.00

	M00390_81_000000000-AA7DR_1_2101_10115_7341	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_2101_10868_23841	97KSI_03637_07116	99.74
	M00390_81_000000000-AA7DR_1_2101_10940_15484	97KSI_04704_05753	99.74
	M00390_81_000000000-AA7DR_1_2101_14832_23380	97KSI_00220_06028	100.00
	M00390_81_000000000-AA7DR_1_2101_15480_21794	97KSI_01527_02635	100.00
	M00390_81_000000000-AA7DR_1_2101_20624_13445	97KSI_01035_01669	99.74
	M00390_81_000000000-AA7DR_1_2101_4020_19252	97KSI_01503_06678	100.00
	M00390_81_000000000-AA7DR_1_2101_8847_9320	97KSI_04187_04119	100.00
	M00390_81_000000000-AA7DR_1_2102_10529_27652	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_2102_11141_21062	97KSI_04600_02136	100.00
	M00390_81_000000000-AA7DR_1_2102_15062_25843	97KSI_04187_04119	99.74
	M00390_81_000000000-AA7DR_1_2102_15898_7673	97KSI_00312_01166	100.00
	M00390_81_000000000-AA7DR_1_2102_19853_11337	97KSI_03424_06077	100.00
	M00390_81_000000000-AA7DR_1_2102_20662_22225	97KSI_01198_07171	100.00
	M00390_81_000000000-AA7DR_1_2102_21895_22191	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_2102_22737_9163	97KSI_00703_00467	100.00
	M00390_81_000000000-AA7DR_1_2102_23080_14611	97KSI_04187_04119	99.74
	M00390_81_000000000-AA7DR_1_2102_2718_17085	97KSI_04701_06228	100.00
	M00390_81_000000000-AA7DR_1_2102_6652_14554	97KSI_05059_01647	100.00
	M00390_81_000000000-AA7DR_1_2102_8866_19993	97KSI_00339_03063	100.00
	M00390_81_000000000-AA7DR_1_2102_9792_14823	97KSI_01035_01669	100.00
	M00390_81_000000000-AA7DR_1_2103_13742_19266	97KSI_02888_05734	99.74
	M00390_81_000000000-AA7DR_1_2103_14572_23446	97KSI_01055_01616	100.00
	M00390_81_000000000-AA7DR_1_2103_14675_24426	97KSI_03268_06624	100.00
	M00390_81_000000000-AA7DR_1_2103_18249_17554	97KSI_02856_07618	99.74
	M00390_81_000000000-AA7DR_1_2103_21243_13448	97KSI_03596_05515	100.00
	M00390_81_000000000-AA7DR_1_2103_21256_18739	97KSI_02216_02513	100.00
	M00390_81_000000000-AA7DR_1_2103_23271_13389	97KSI_00572_05103	100.00
	M00390_81_000000000-AA7DR_1_2103_24127_12230	97KSI_03570_03883	100.00
	M00390_81_000000000-AA7DR_1_2103_24691_16368	97KSI_01225_01807	99.48
	M00390_81_000000000-AA7DR_1_2103_28131_14780	97KSI_01648_05844	100.00
	M00390_81_000000000-AA7DR_1_2103_3964_21181	97KSI_03875_07024	99.74
	M00390_81_000000000-AA7DR_1_2103_6459_9580	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_2103_8016_22078	97KSI_05182_05960	100.00
	M00390_81_000000000-AA7DR_1_2104_10109_8976	97KSI_00619_03563	100.00
	M00390_81_000000000-AA7DR_1_2104_11978_13811	97KSI_02927_04010	99.74

	M00390_81_000000000-AA7DR_1_2104_12564_15378	97KSI_03958_00212	99.74
	M00390_81_000000000-AA7DR_1_2104_16946_23798	97KSI_00567_05066	99.74
	M00390_81_000000000-AA7DR_1_2104_18077_27447	97KSI_01225_01807	99.74
	M00390_81_000000000-AA7DR_1_2104_21699_15198	97KSI_02356_06598	99.74
	M00390_81_000000000-AA7DR_1_2104_4704_10529	97KSI_03083_01515	100.00
	M00390_81_000000000-AA7DR_1_2104_5389_13932	97KSI_01743_04579	100.00
	M00390_81_000000000-AA7DR_1_2105_11652_17368	97KSI_00189_01245	100.00
	M00390_81_000000000-AA7DR_1_2105_12150_4581	97KSI_02406_00831	99.74
	M00390_81_000000000-AA7DR_1_2105_14502_17429	97KSI_02111_02597	100.00
	M00390_81_000000000-AA7DR_1_2105_14886_21243	97KSI_00538_04716	99.74
	M00390_81_000000000-AA7DR_1_2105_15248_23860	97KSI_03585_02346	100.00
	M00390_81_000000000-AA7DR_1_2105_15413_12831	97KSI_03569_02826	100.00
	M00390_81_000000000-AA7DR_1_2105_19110_16140	97KSI_04715_02096	100.00
	M00390_81_000000000-AA7DR_1_2105_3066_19474	97KSI_01422_06975	100.00
	M00390_81_000000000-AA7DR_1_2105_6558_12191	97KSI_02301_04052	100.00
	M00390_81_000000000-AA7DR_1_2105_8525_7938	97KSI_04781_05500	100.00
	M00390_81_000000000-AA7DR_1_2106_14426_18041	97KSI_01797_07274	99.74
	M00390_81_000000000-AA7DR_1_2106_16661_24663	97KSI_00936_03491	100.00
	M00390_81_000000000-AA7DR_1_2106_17213_17299	97KSI_00475_06811	99.74
	M00390_81_000000000-AA7DR_1_2106_19924_20621	97KSI_03294_06151	99.74
	M00390_81_000000000-AA7DR_1_2106_8835_23021	97KSI_04042_02077	99.74
	M00390_81_000000000-AA7DR_1_2107_10446_7832	97KSI_03958_00212	99.74
	M00390_81_000000000-AA7DR_1_2107_11725_16578	97KSI_03958_00212	100.00
	M00390_81_000000000-AA7DR_1_2107_13036_14378	97KSI_01244_07404	100.00
	M00390_81_000000000-AA7DR_1_2107_14537_22178	97KSI_02742_07076	100.00
	M00390_81_000000000-AA7DR_1_2107_14756_21371	97KSI_01851_04108	100.00
	M00390_81_000000000-AA7DR_1_2107_17842_13515	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_2107_20255_15389	97KSI_05184_04990	100.00
	M00390_81_000000000-AA7DR_1_2107_9396_11171	97KSI_03303_06741	99.74
	M00390_81_000000000-AA7DR_1_2108_12821_27054	97KSI_04599_02055	99.48
	M00390_81_000000000-AA7DR_1_2108_15229_13188	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_2108_16729_15068	97KSI_04516_04950	99.74
	M00390_81_000000000-AA7DR_1_2108_17265_11023	97KSI_04975_03734	100.00
	M00390_81_000000000-AA7DR_1_2108_3833_11029	97KSI_03885_03542	100.00
	M00390_81_000000000-AA7DR_1_2108_4505_22110	97KSI_01081_00420	100.00
	M00390_81_000000000-AA7DR_1_2108_6822_9219	97KSI_02082_05813	100.00

	M00390_81_000000000-AA7DR_1_2108_9341_20249	97KSI_03919_06615	100.00
	M00390_81_000000000-AA7DR_1_2109_13907_11340	97KSI_02136_06907	99.74
	M00390_81_000000000-AA7DR_1_2109_14206_25147	97KSI_01117_02699	100.00
	M00390_81_000000000-AA7DR_1_2109_14423_9938	97KSI_04296_03028	99.74
	M00390_81_000000000-AA7DR_1_2109_16921_13436	97KSI_00446_02613	100.00
	M00390_81_000000000-AA7DR_1_2109_19339_19157	97KSI_00407_00871	99.74
	M00390_81_000000000-AA7DR_1_2109_4746_14889	97KSI_02023_02420	99.74
	M00390_81_000000000-AA7DR_1_2110_10351_12783	97KSI_03803_06221	100.00
	M00390_81_000000000-AA7DR_1_2110_17018_9908	97KSI_00632_03290	99.74
	M00390_81_000000000-AA7DR_1_2110_22214_8296	97KSI_00291_01293	100.00
	M00390_81_000000000-AA7DR_1_2110_6835_21029	97KSI_04599_02055	99.48
	M00390_81_000000000-AA7DR_1_2111_10281_13264	97KSI_05182_04204	100.00
	M00390_81_000000000-AA7DR_1_2111_10486_7906	97KSI_04516_04950	99.48
	M00390_81_000000000-AA7DR_1_2111_16150_12284	97KSI_03840_03784	100.00
	M00390_81_000000000-AA7DR_1_2111_20084_7863	97KSI_04388_06583	100.00
	M00390_81_000000000-AA7DR_1_2111_22522_16753	97KSI_03620_01563	100.00
	M00390_81_000000000-AA7DR_1_2111_24941_11256	97KSI_00473_05399	100.00
	M00390_81_000000000-AA7DR_1_2111_3921_17933	97KSI_01033_04984	99.74
	M00390_81_000000000-AA7DR_1_2111_6272_21662	97KSI_02082_05813	99.74
	M00390_81_000000000-AA7DR_1_2111_9310_9000	97KSI_04567_02701	99.74
	M00390_81_000000000-AA7DR_1_2111_9782_11142	97KSI_00440_06051	100.00
	M00390_81_000000000-AA7DR_1_2112_14241_20548	97KSI_01735_04967	100.00
	M00390_81_000000000-AA7DR_1_2112_14745_11643	97KSI_02492_01578	100.00
	M00390_81_000000000-AA7DR_1_2112_16329_13134	97KSI_00717_00658	100.00
	M00390_81_000000000-AA7DR_1_2112_16913_20967	97KSI_04410_06275	99.74
	M00390_81_000000000-AA7DR_1_2113_16196_18346	97KSI_01648_05844	99.74
	M00390_81_000000000-AA7DR_1_2113_4950_10754	97KSI_04190_05506	99.74
	M00390_81_000000000-AA7DR_1_2113_7363_14186	97KSI_00293_03871	100.00
	M00390_81_000000000-AA7DR_1_2114_15051_13851	97KSI_00293_03871	99.74
	M00390_81_000000000-AA7DR_1_2114_15116_23862	97KSI_01888_03164	100.00
	M00390_81_000000000-AA7DR_1_2114_17389_18710	97KSI_04707_04348	99.74
	M00390_81_000000000-AA7DR_1_2114_20908_10400	97KSI_01340_02432	100.00
***C.* sp. Na11C3**	M00390_40_000000000-A6D16_1_1108_17304_25880	97KSI_03021_03598	100.00
	M00390_80_000000000-AA759_1_1101_24925_5895	97KSI_03096_05080	99.73
	M00390_80_000000000-AA759_1_1102_11036_5510	97KSI_01656_03252	99.73

	M00390_80_000000000-AA759_1_1102_11862_19658	97KSI_00844_05835	99.73
	M00390_80_000000000-AA759_1_1102_22265_13042	97KSI_02846_03329	99.73
	M00390_80_000000000-AA759_1_1103_18744_24933	97KSI_00993_01026	99.47
	M00390_80_000000000-AA759_1_1103_24919_24925	97KSI_04387_04031	99.73
	M00390_80_000000000-AA759_1_1103_4831_10564	97KSI_00321_05775	99.73
	M00390_80_000000000-AA759_1_1103_6746_6120	97KSI_00460_06255	99.73
	M00390_80_000000000-AA759_1_1104_18700_4291	97KSI_01067_00604	100.00
	M00390_80_000000000-AA759_1_1104_2328_18588	97KSI_03817_04140	100.00
	M00390_80_000000000-AA759_1_1105_12350_25470	97KSI_01943_03397	99.73
	M00390_80_000000000-AA759_1_1105_12930_17652	97KSI_01165_00618	99.73
	M00390_80_000000000-AA759_1_1105_19903_27035	97KSI_03369_02217	99.73
	M00390_80_000000000-AA759_1_1105_24209_15762	97KSI_00321_05775	99.73
	M00390_80_000000000-AA759_1_1105_7297_13164	97KSI_04017_02014	100.00
	M00390_80_000000000-AA759_1_1106_6160_18009	97KSI_04258_02565	100.00
	M00390_80_000000000-AA759_1_1107_13696_15050	97KSI_03473_05700	99.73
	M00390_80_000000000-AA759_1_1107_25024_7083	97KSI_04469_05210	100.00
	M00390_80_000000000-AA759_1_1107_9830_24499	97KSI_02750_07324	100.00
	M00390_80_000000000-AA759_1_1108_18612_20000	97KSI_03920_02034	100.00
	M00390_80_000000000-AA759_1_1108_22314_4157	97KSI_03781_03810	99.73
	M00390_80_000000000-AA759_1_1108_8881_16788	97KSI_00487_03761	100.00
	M00390_80_000000000-AA759_1_1109_10815_17518	97KSI_01106_03871	100.00
	M00390_80_000000000-AA759_1_1109_8552_5622	97KSI_02911_03818	100.00
	M00390_80_000000000-AA759_1_1111_11445_6058	97KSI_00893_05900	100.00
	M00390_80_000000000-AA759_1_1111_12930_26967	97KSI_01165_00618	99.73
	M00390_80_000000000-AA759_1_1111_13044_25175	97KSI_00595_04086	100.00
	M00390_80_000000000-AA759_1_1111_24180_24201	97KSI_04796_03244	100.00
	M00390_80_000000000-AA759_1_1112_17929_5348	97KSI_01837_05566	100.00
	M00390_80_000000000-AA759_1_1112_29011_17675	97KSI_00453_01744	100.00
	M00390_80_000000000-AA759_1_1112_6328_6889	97KSI_01122_03589	99.73
	M00390_80_000000000-AA759_1_1112_6677_23858	97KSI_00976_05951	99.73
	M00390_80_000000000-AA759_1_1113_27929_14998	97KSI_03191_01259	100.00
	M00390_80_000000000-AA759_1_1114_10356_13250	97KSI_04550_07000	100.00
	M00390_80_000000000-AA759_1_1114_20338_13729	97KSI_00289_02270	100.00
	M00390_80_000000000-AA759_1_1114_2317_12526	97KSI_02212_04011	100.00
	M00390_80_000000000-AA759_1_1114_25768_22209	97KSI_01002_05814	100.00
	M00390_80_000000000-AA759_1_1114_7144_11894	97KSI_04095_00209	100.00

	M00390_80_000000000-AA759_1_2102_20128_17451	97KSI_02684_01976	100.00
	M00390_80_000000000-AA759_1_2102_20595_5515	97KSI_02134_01725	99.73
	M00390_80_000000000-AA759_1_2102_20779_24907	97KSI_03765_01443	99.73
	M00390_80_000000000-AA759_1_2103_15597_11925	97KSI_00886_06722	100.00
	M00390_80_000000000-AA759_1_2103_21692_23377	97KSI_00790_03258	100.00
	M00390_80_000000000-AA759_1_2104_12973_8670	97KSI_00200_03141	100.00
	M00390_80_000000000-AA759_1_2104_14747_24209	97KSI_02067_02314	99.73
	M00390_80_000000000-AA759_1_2105_5356_7057	97KSI_00379_01738	100.00
	M00390_80_000000000-AA759_1_2105_8963_20511	97KSI_01849_03036	100.00
	M00390_80_000000000-AA759_1_2105_9651_14559	97KSI_03820_01798	99.73
	M00390_80_000000000-AA759_1_2106_10278_6592	97KSI_00775_03521	99.73
	M00390_80_000000000-AA759_1_2106_12686_25312	97KSI_04095_00209	99.73
	M00390_80_000000000-AA759_1_2106_14994_22331	97KSI_00252_04181	100.00
	M00390_80_000000000-AA759_1_2106_18502_26342	97KSI_01005_02504	99.73
	M00390_80_000000000-AA759_1_2106_24441_24093	97KSI_03894_07053	100.00
	M00390_80_000000000-AA759_1_2107_6172_18553	97KSI_04009_00303	99.73
	M00390_80_000000000-AA759_1_2108_12418_7174	97KSI_04174_03701	100.00
	M00390_80_000000000-AA759_1_2108_19837_10015	97KSI_01119_01619	100.00
	M00390_80_000000000-AA759_1_2109_13318_11905	97KSI_05137_01822	100.00
	M00390_80_000000000-AA759_1_2111_17490_13024	97KSI_01628_05466	100.00
	M00390_80_000000000-AA759_1_2111_8013_25115	97KSI_03342_02492	100.00
	M00390_80_000000000-AA759_1_2112_10650_15784	97KSI_02933_06047	100.00
	M00390_80_000000000-AA759_1_2112_15082_14392	97KSI_03474_05985	100.00
	M00390_80_000000000-AA759_1_2113_16044_22931	97KSI_00901_02972	100.00
	M00390_80_000000000-AA759_1_2113_20573_11702	97KSI_03894_07053	99.73
	M00390_80_000000000-AA759_1_2113_2975_15819	97KSI_02690_03895	100.00
	M00390_80_000000000-AA759_1_2113_4748_10009	97KSI_00960_02295	100.00
	M00390_80_000000000-AA759_1_2113_8235_13827	97KSI_00748_06170	99.73
	M00390_80_000000000-AA759_1_2113_9768_17940	97KSI_04376_00331	100.00
	M00390_80_000000000-AA759_1_2114_27206_21239	97KSI_04067_05041	100.00
	M00390_81_000000000-AA7DR_1_1101_11196_27974	97KSI_02154_04878	100.00
	M00390_81_000000000-AA7DR_1_1101_12251_3908	97KSI_00460_06255	99.73
	M00390_81_000000000-AA7DR_1_1101_13232_27438	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1101_16158_15998	97KSI_04341_05203	100.00
	M00390_81_000000000-AA7DR_1_1101_16159_26878	97KSI_01279_04505	100.00
	M00390_81_000000000-AA7DR_1_1101_17105_13417	97KSI_04670_06771	100.00

	M00390_81_000000000-AA7DR_1_1101_17107_13439	97KSI_00533_03380	100.00
	M00390_81_000000000-AA7DR_1_1101_18409_11139	97KSI_03488_06956	99.73
	M00390_81_000000000-AA7DR_1_1101_18743_18805	97KSI_03833_00784	100.00
	M00390_81_000000000-AA7DR_1_1101_19191_4495	97KSI_00748_06170	100.00
	M00390_81_000000000-AA7DR_1_1101_20003_21142	97KSI_00890_05779	100.00
	M00390_81_000000000-AA7DR_1_1101_21290_7677	97KSI_00774_00457	100.00
	M00390_81_000000000-AA7DR_1_1101_24432_18027	97KSI_05250_05331	99.73
	M00390_81_000000000-AA7DR_1_1101_25478_16104	97KSI_00460_06255	100.00
	M00390_81_000000000-AA7DR_1_1101_27613_11799	97KSI_01782_01635	100.00
	M00390_81_000000000-AA7DR_1_1101_27935_22163	97KSI_01784_04142	100.00
	M00390_81_000000000-AA7DR_1_1101_6410_5509	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1101_7376_5104	97KSI_02216_04247	99.73
	M00390_81_000000000-AA7DR_1_1101_9097_5324	97KSI_03093_06409	100.00
	M00390_81_000000000-AA7DR_1_1102_12058_20294	97KSI_04981_02568	100.00
	M00390_81_000000000-AA7DR_1_1102_12871_5484	97KSI_01214_00826	100.00
	M00390_81_000000000-AA7DR_1_1102_13228_10139	97KSI_01758_00919	99.73
	M00390_81_000000000-AA7DR_1_1102_13427_6631	97KSI_01427_02313	100.00
	M00390_81_000000000-AA7DR_1_1102_13729_26766	97KSI_02732_06679	100.00
	M00390_81_000000000-AA7DR_1_1102_15210_14459	97KSI_00961_03556	100.00
	M00390_81_000000000-AA7DR_1_1102_15583_2449	97KSI_02189_04009	100.00
	M00390_81_000000000-AA7DR_1_1102_16054_27848	97KSI_05057_04395	100.00
	M00390_81_000000000-AA7DR_1_1102_17728_10049	97KSI_01062_02647	99.73
	M00390_81_000000000-AA7DR_1_1102_17746_10054	97KSI_03757_00347	100.00
	M00390_81_000000000-AA7DR_1_1102_18329_3761	97KSI_02189_04009	99.73
	M00390_81_000000000-AA7DR_1_1102_19523_26650	97KSI_03496_05087	99.73
	M00390_81_000000000-AA7DR_1_1102_23857_26378	97KSI_04568_05809	100.00
	M00390_81_000000000-AA7DR_1_1102_24082_17124	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_1102_24563_10548	97KSI_04778_03296	100.00
	M00390_81_000000000-AA7DR_1_1102_24763_6366	97KSI_00844_05835	100.00
	M00390_81_000000000-AA7DR_1_1102_24856_10898	97KSI_00321_05775	100.00
	M00390_81_000000000-AA7DR_1_1102_25954_22369	97KSI_03894_07053	99.73
	M00390_81_000000000-AA7DR_1_1102_27534_22303	97KSI_03139_06556	99.73
	M00390_81_000000000-AA7DR_1_1102_27566_10283	97KSI_02911_03818	99.73
	M00390_81_000000000-AA7DR_1_1102_3969_15352	97KSI_03774_02908	100.00
	M00390_81_000000000-AA7DR_1_1102_6352_12255	97KSI_01744_01568	99.73
	M00390_81_000000000-AA7DR_1_1102_7851_19962	97KSI_00154_01935	99.73

	M00390_81_000000000-AA7DR_1_1102_9046_16993	97KSI_03139_06556	100.00
	M00390_81_000000000-AA7DR_1_1102_9058_25666	97KSI_01733_00992	99.73
	M00390_81_000000000-AA7DR_1_1103_10845_13680	97KSI_03269_00122	100.00
	M00390_81_000000000-AA7DR_1_1103_10959_18065	97KSI_03820_01798	99.47
	M00390_81_000000000-AA7DR_1_1103_11912_2490	97KSI_00378_06507	100.00
	M00390_81_000000000-AA7DR_1_1103_13104_26209	97KSI_01230_04601	100.00
	M00390_81_000000000-AA7DR_1_1103_13537_14886	97KSI_00533_02475	99.73
	M00390_81_000000000-AA7DR_1_1103_14016_18079	97KSI_03360_01646	99.73
	M00390_81_000000000-AA7DR_1_1103_15051_28199	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_1103_16291_4409	97KSI_00719_05982	100.00
	M00390_81_000000000-AA7DR_1_1103_16292_4430	97KSI_02552_02912	100.00
	M00390_81_000000000-AA7DR_1_1103_20280_15740	97KSI_02296_00693	99.73
	M00390_81_000000000-AA7DR_1_1103_22194_6232	97KSI_05069_01868	100.00
	M00390_81_000000000-AA7DR_1_1103_23853_17280	97KSI_00313_02395	100.00
	M00390_81_000000000-AA7DR_1_1103_24022_20791	97KSI_01713_00803	99.73
	M00390_81_000000000-AA7DR_1_1103_24081_22178	97KSI_01969_05288	99.73
	M00390_81_000000000-AA7DR_1_1103_24189_10432	97KSI_02060_05024	100.00
	M00390_81_000000000-AA7DR_1_1103_24556_22354	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_1103_25739_19440	97KSI_03817_04140	99.73
	M00390_81_000000000-AA7DR_1_1103_26170_13373	97KSI_00363_01769	99.73
	M00390_81_000000000-AA7DR_1_1103_4701_17747	97KSI_03114_00768	99.73
	M00390_81_000000000-AA7DR_1_1104_10127_18000	97KSI_03988_01657	100.00
	M00390_81_000000000-AA7DR_1_1104_11340_20596	97KSI_00221_01329	99.73
	M00390_81_000000000-AA7DR_1_1104_14676_18009	97KSI_02374_02761	99.73
	M00390_81_000000000-AA7DR_1_1104_15985_17789	97KSI_04809_01292	99.73
	M00390_81_000000000-AA7DR_1_1104_16767_17706	97KSI_04576_01800	100.00
	M00390_81_000000000-AA7DR_1_1104_17994_3083	97KSI_03516_04747	100.00
	M00390_81_000000000-AA7DR_1_1104_18910_10184	97KSI_00338_04739	100.00
	M00390_81_000000000-AA7DR_1_1104_19012_23767	97KSI_04475_04635	100.00
	M00390_81_000000000-AA7DR_1_1104_19218_3762	97KSI_04884_05044	99.73
	M00390_81_000000000-AA7DR_1_1104_19364_12375	97KSI_01039_05703	99.73
	M00390_81_000000000-AA7DR_1_1104_22190_7531	97KSI_03913_02426	100.00
	M00390_81_000000000-AA7DR_1_1104_25563_18750	97KSI_02203_06434	99.73
	M00390_81_000000000-AA7DR_1_1104_25731_13884	97KSI_05069_01868	100.00
	M00390_81_000000000-AA7DR_1_1104_27007_19260	97KSI_01166_04899	100.00
	M00390_81_000000000-AA7DR_1_1104_27341_15601	97KSI_00844_05835	99.73

	M00390_81_000000000-AA7DR_1_1104_6825_21363	97KSI_01814_03706	99.74
	M00390_81_000000000-AA7DR_1_1104_8517_8323	97KSI_03497_07247	100.00
	M00390_81_000000000-AA7DR_1_1105_11045_15726	97KSI_03862_01719	99.73
	M00390_81_000000000-AA7DR_1_1105_11940_23531	97KSI_03606_06975	100.00
	M00390_81_000000000-AA7DR_1_1105_14740_11197	97KSI_02035_06902	100.00
	M00390_81_000000000-AA7DR_1_1105_15131_25255	97KSI_02518_06093	100.00
	M00390_81_000000000-AA7DR_1_1105_15319_18036	97KSI_00784_05608	99.73
	M00390_81_000000000-AA7DR_1_1105_18468_5300	97KSI_00924_07082	100.00
	M00390_81_000000000-AA7DR_1_1105_20428_22822	97KSI_00525_06461	100.00
	M00390_81_000000000-AA7DR_1_1105_20439_16228	97KSI_01945_03846	100.00
	M00390_81_000000000-AA7DR_1_1105_23214_8852	97KSI_03817_04140	99.73
	M00390_81_000000000-AA7DR_1_1105_24774_19922	97KSI_03820_01798	99.47
	M00390_81_000000000-AA7DR_1_1105_25605_20745	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1105_26175_14727	97KSI_00556_05581	100.00
	M00390_81_000000000-AA7DR_1_1105_27585_16880	97KSI_00143_02567	100.00
	M00390_81_000000000-AA7DR_1_1105_27967_11774	97KSI_00844_05835	99.73
	M00390_81_000000000-AA7DR_1_1105_28780_20199	97KSI_04520_04321	99.47
	M00390_81_000000000-AA7DR_1_1105_3268_17783	97KSI_03114_00768	99.73
	M00390_81_000000000-AA7DR_1_1105_3321_19115	97KSI_02216_06531	100.00
	M00390_81_000000000-AA7DR_1_1105_7167_22593	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1105_8517_22664	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1106_10610_5650	97KSI_04167_06628	99.73
	M00390_81_000000000-AA7DR_1_1106_11577_10270	97KSI_02152_04729	99.73
	M00390_81_000000000-AA7DR_1_1106_16338_21693	97KSI_02337_07107	100.00
	M00390_81_000000000-AA7DR_1_1106_21055_5887	97KSI_00074_04973	99.73
	M00390_81_000000000-AA7DR_1_1106_22057_8800	97KSI_01307_00715	100.00
	M00390_81_000000000-AA7DR_1_1106_22530_20421	97KSI_03676_03501	100.00
	M00390_81_000000000-AA7DR_1_1106_23551_25375	97KSI_00369_06436	100.00
	M00390_81_000000000-AA7DR_1_1106_2424_15795	97KSI_03820_01798	100.00
	M00390_81_000000000-AA7DR_1_1106_25771_7692	97KSI_05095_02527	100.00
	M00390_81_000000000-AA7DR_1_1106_27392_21231	97KSI_02067_02314	99.73
	M00390_81_000000000-AA7DR_1_1106_6645_12895	97KSI_00532_02482	100.00
	M00390_81_000000000-AA7DR_1_1106_9051_6587	97KSI_02082_03344	100.00
	M00390_81_000000000-AA7DR_1_1106_9364_16076	97KSI_04830_04581	99.73
	M00390_81_000000000-AA7DR_1_1106_9998_3612	97KSI_01222_07450	100.00
	M00390_81_000000000-AA7DR_1_1107_14821_4905	97KSI_00266_01420	100.00

	M00390_81_000000000-AA7DR_1_1107_17406_12355	97KSI_02853_00993	100.00
	M00390_81_000000000-AA7DR_1_1107_1776_15506	97KSI_03188_00866	100.00
	M00390_81_000000000-AA7DR_1_1107_18348_21528	97KSI_02881_05835	99.73
	M00390_81_000000000-AA7DR_1_1107_18834_15725	97KSI_03765_01443	99.73
	M00390_81_000000000-AA7DR_1_1107_19739_2477	97KSI_04276_03724	100.00
	M00390_81_000000000-AA7DR_1_1107_19859_6310	97KSI_03371_01805	100.00
	M00390_81_000000000-AA7DR_1_1107_20279_27480	97KSI_02977_01371	100.00
	M00390_81_000000000-AA7DR_1_1107_21130_10078	97KSI_03335_06240	99.73
	M00390_81_000000000-AA7DR_1_1107_21253_22489	97KSI_01733_00992	100.00
	M00390_81_000000000-AA7DR_1_1107_22475_25010	97KSI_02332_06237	100.00
	M00390_81_000000000-AA7DR_1_1107_23588_16444	97KSI_01195_03946	100.00
	M00390_81_000000000-AA7DR_1_1107_25678_10393	97KSI_00191_01163	100.00
	M00390_81_000000000-AA7DR_1_1107_27207_11882	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_1107_3484_19190	97KSI_02867_04498	100.00
	M00390_81_000000000-AA7DR_1_1107_4779_15550	97KSI_00477_04066	100.00
	M00390_81_000000000-AA7DR_1_1107_6217_8406	97KSI_04958_03810	99.73
	M00390_81_000000000-AA7DR_1_1107_6332_24750	97KSI_02154_04878	99.47
	M00390_81_000000000-AA7DR_1_1107_6818_21960	97KSI_04158_02254	100.00
	M00390_81_000000000-AA7DR_1_1107_6975_8420	97KSI_01107_06748	99.73
	M00390_81_000000000-AA7DR_1_1107_8730_9843	97KSI_01262_05868	100.00
	M00390_81_000000000-AA7DR_1_1107_9732_26522	97KSI_01417_02170	99.73
	M00390_81_000000000-AA7DR_1_1108_11603_11868	97KSI_00863_01684	100.00
	M00390_81_000000000-AA7DR_1_1108_11619_11906	97KSI_05154_03279	100.00
	M00390_81_000000000-AA7DR_1_1108_13180_7077	97KSI_00434_06523	99.73
	M00390_81_000000000-AA7DR_1_1108_16881_27613	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1108_17545_2966	97KSI_02027_03223	100.00
	M00390_81_000000000-AA7DR_1_1108_20835_27333	97KSI_05007_01508	100.00
	M00390_81_000000000-AA7DR_1_1108_21114_11099	97KSI_01359_00277	100.00
	M00390_81_000000000-AA7DR_1_1108_24721_11822	97KSI_00355_00902	99.74
	M00390_81_000000000-AA7DR_1_1108_3171_18890	97KSI_04130_00630	99.73
	M00390_81_000000000-AA7DR_1_1108_3479_14144	97KSI_01817_02528	100.00
	M00390_81_000000000-AA7DR_1_1108_3955_11068	97KSI_03515_01655	100.00
	M00390_81_000000000-AA7DR_1_1108_4832_11089	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_1108_6656_18128	97KSI_01155_04044	100.00
	M00390_81_000000000-AA7DR_1_1108_9798_3001	97KSI_00687_06967	100.00
	M00390_81_000000000-AA7DR_1_1109_10311_3106	97KSI_05069_01868	99.73

	M00390_81_000000000-AA7DR_1_1109_11107_25766	97KSI_04233_07078	100.00
	M00390_81_000000000-AA7DR_1_1109_11125_25767	97KSI_00804_06204	100.00
	M00390_81_000000000-AA7DR_1_1109_12261_15569	97KSI_00290_02888	100.00
	M00390_81_000000000-AA7DR_1_1109_12394_10561	97KSI_02835_02126	100.00
	M00390_81_000000000-AA7DR_1_1109_15314_9521	97KSI_02722_02279	100.00
	M00390_81_000000000-AA7DR_1_1109_15429_18641	97KSI_03101_05043	100.00
	M00390_81_000000000-AA7DR_1_1109_18421_5919	97KSI_00678_05143	100.00
	M00390_81_000000000-AA7DR_1_1109_22201_12066	97KSI_00196_02555	99.73
	M00390_81_000000000-AA7DR_1_1109_22429_18036	97KSI_04451_01423	100.00
	M00390_81_000000000-AA7DR_1_1109_23147_6138	97KSI_02911_03818	99.73
	M00390_81_000000000-AA7DR_1_1109_26584_8352	97KSI_00460_06255	99.73
	M00390_81_000000000-AA7DR_1_1109_3355_19727	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_1109_3918_8950	97KSI_01714_04568	99.73
	M00390_81_000000000-AA7DR_1_1109_4716_11773	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_1109_5908_24290	97KSI_04201_03749	100.00
	M00390_81_000000000-AA7DR_1_1109_7800_4262	97KSI_04158_05183	100.00
	M00390_81_000000000-AA7DR_1_1109_8052_12092	97KSI_04040_04857	100.00
	M00390_81_000000000-AA7DR_1_1110_12634_14213	97KSI_00906_05965	99.73
	M00390_81_000000000-AA7DR_1_1110_13121_24818	97KSI_00078_05612	100.00
	M00390_81_000000000-AA7DR_1_1110_13803_24457	97KSI_04376_07129	100.00
	M00390_81_000000000-AA7DR_1_1110_14924_26720	97KSI_02034_05599	100.00
	M00390_81_000000000-AA7DR_1_1110_16850_27753	97KSI_01819_01022	100.00
	M00390_81_000000000-AA7DR_1_1110_19679_13356	97KSI_03345_04692	100.00
	M00390_81_000000000-AA7DR_1_1110_19763_2849	97KSI_01511_05514	100.00
	M00390_81_000000000-AA7DR_1_1110_20110_10982	97KSI_02615_01104	99.73
	M00390_81_000000000-AA7DR_1_1110_24225_5889	97KSI_04219_07235	100.00
	M00390_81_000000000-AA7DR_1_1110_24768_13379	97KSI_03684_05943	100.00
	M00390_81_000000000-AA7DR_1_1110_25292_12984	97KSI_03285_02895	100.00
	M00390_81_000000000-AA7DR_1_1110_25756_19987	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1110_26826_22971	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1110_28181_21380	97KSI_02246_03171	100.00
	M00390_81_000000000-AA7DR_1_1110_2836_12986	97KSI_04625_06294	100.00
	M00390_81_000000000-AA7DR_1_1110_3332_21204	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1110_5330_8573	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1110_5925_18140	97KSI_02848_01222	100.00
	M00390_81_000000000-AA7DR_1_1110_5995_9355	97KSI_00138_05029	100.00

	M00390_81_000000000-AA7DR_1_1110_6357_22219	97KSI_00701_06124	100.00
	M00390_81_000000000-AA7DR_1_1110_7052_18315	97KSI_00160_02389	99.21
	M00390_81_000000000-AA7DR_1_1110_7268_15251	97KSI_00645_05634	100.00
	M00390_81_000000000-AA7DR_1_1110_7314_13849	97KSI_03013_02536	99.73
	M00390_81_000000000-AA7DR_1_1110_9986_27720	97KSI_02077_06508	100.00
	M00390_81_000000000-AA7DR_1_1111_10416_18691	97KSI_03195_01710	99.73
	M00390_81_000000000-AA7DR_1_1111_12461_22533	97KSI_04818_02139	99.47
	M00390_81_000000000-AA7DR_1_1111_13086_24853	97KSI_03894_07053	99.73
	M00390_81_000000000-AA7DR_1_1111_13321_12720	97KSI_05130_06212	99.73
	M00390_81_000000000-AA7DR_1_1111_15203_3656	97KSI_03496_05087	99.73
	M00390_81_000000000-AA7DR_1_1111_16048_6783	97KSI_03414_07393	99.73
	M00390_81_000000000-AA7DR_1_1111_17794_9340	97KSI_02060_05024	99.73
	M00390_81_000000000-AA7DR_1_1111_18908_14831	97KSI_00343_06741	99.73
	M00390_81_000000000-AA7DR_1_1111_19580_15224	97KSI_04874_06788	99.73
	M00390_81_000000000-AA7DR_1_1111_20571_7220	97KSI_00704_01982	99.73
	M00390_81_000000000-AA7DR_1_1111_21053_3978	97KSI_01222_07450	99.73
	M00390_81_000000000-AA7DR_1_1111_21418_20226	97KSI_00321_05775	99.73
	M00390_81_000000000-AA7DR_1_1111_21616_8389	97KSI_02337_07107	99.73
	M00390_81_000000000-AA7DR_1_1111_22503_11817	97KSI_01067_00604	99.73
	M00390_81_000000000-AA7DR_1_1111_23705_18078	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1111_26410_18073	97KSI_02154_04878	99.47
	M00390_81_000000000-AA7DR_1_1111_2828_19038	97KSI_03817_04140	99.47
	M00390_81_000000000-AA7DR_1_1111_29568_13591	97KSI_00844_05835	99.73
	M00390_81_000000000-AA7DR_1_1111_3142_11361	97KSI_02911_03818	99.47
	M00390_81_000000000-AA7DR_1_1111_4345_8627	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1111_5255_13740	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_1111_5764_8356	97KSI_00609_02172	100.00
	M00390_81_000000000-AA7DR_1_1111_5829_6427	97KSI_03515_01655	99.73
	M00390_81_000000000-AA7DR_1_1111_6782_23881	97KSI_00321_05775	99.73
	M00390_81_000000000-AA7DR_1_1111_7551_7003	97KSI_00363_01769	99.73
	M00390_81_000000000-AA7DR_1_1111_9746_13407	97KSI_05028_01463	99.47
	M00390_81_000000000-AA7DR_1_1112_12407_14878	97KSI_04884_05044	99.73
	M00390_81_000000000-AA7DR_1_1112_13352_5601	97KSI_03357_06194	99.73
	M00390_81_000000000-AA7DR_1_1112_14682_17903	97KSI_00784_05608	100.00
	M00390_81_000000000-AA7DR_1_1112_15586_17537	97KSI_02348_07517	100.00
	M00390_81_000000000-AA7DR_1_1112_16796_22389	97KSI_00663_06675	100.00

	M00390_81_000000000-AA7DR_1_1112_17859_8164	97KSI_02057_02771	99.73
	M00390_81_000000000-AA7DR_1_1112_18973_7862	97KSI_01574_04337	100.00
	M00390_81_000000000-AA7DR_1_1112_20622_21848	97KSI_01172_05928	99.73
	M00390_81_000000000-AA7DR_1_1112_21063_5655	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1112_21557_14292	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1112_21936_7863	97KSI_03665_02501	99.73
	M00390_81_000000000-AA7DR_1_1112_22279_16278	97KSI_04170_06566	100.00
	M00390_81_000000000-AA7DR_1_1112_2632_13931	97KSI_00108_02998	99.74
	M00390_81_000000000-AA7DR_1_1112_26948_23469	97KSI_02399_03900	99.73
	M00390_81_000000000-AA7DR_1_1112_29292_16098	97KSI_00460_06255	99.73
	M00390_81_000000000-AA7DR_1_1112_4447_20796	97KSI_00363_01769	99.47
	M00390_81_000000000-AA7DR_1_1112_6379_19715	97KSI_00363_01769	99.73
	M00390_81_000000000-AA7DR_1_1112_7360_10892	97KSI_00844_05835	99.73
	M00390_81_000000000-AA7DR_1_1112_7621_14101	97KSI_00647_03319	99.73
	M00390_81_000000000-AA7DR_1_1112_8191_10479	97KSI_03998_01614	100.00
	M00390_81_000000000-AA7DR_1_1112_9503_18674	97KSI_02395_02753	100.00
	M00390_81_000000000-AA7DR_1_1112_9523_10659	97KSI_01594_05894	100.00
	M00390_81_000000000-AA7DR_1_1113_11527_9270	97KSI_03369_02217	100.00
	M00390_81_000000000-AA7DR_1_1113_14267_3433	97KSI_04670_06771	99.73
	M00390_81_000000000-AA7DR_1_1113_16579_4869	97KSI_03013_02536	100.00
	M00390_81_000000000-AA7DR_1_1113_16616_9226	97KSI_04839_06993	99.74
	M00390_81_000000000-AA7DR_1_1113_17568_13740	97KSI_02221_01862	99.47
	M00390_81_000000000-AA7DR_1_1113_17825_8140	97KSI_01744_01568	100.00
	M00390_81_000000000-AA7DR_1_1113_20112_16241	97KSI_00509_03810	100.00
	M00390_81_000000000-AA7DR_1_1113_21216_24600	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1113_2240_17861	97KSI_00122_02414	100.00
	M00390_81_000000000-AA7DR_1_1113_22428_24990	97KSI_02898_05450	99.73
	M00390_81_000000000-AA7DR_1_1113_22706_5313	97KSI_01107_06748	100.00
	M00390_81_000000000-AA7DR_1_1113_24059_6884	97KSI_04465_06527	99.73
	M00390_81_000000000-AA7DR_1_1113_27443_14488	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_1113_29235_17394	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_1113_5250_15472	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_1113_6961_16472	97KSI_04599_03647	99.73
	M00390_81_000000000-AA7DR_1_1113_7128_23668	97KSI_03360_01646	100.00
	M00390_81_000000000-AA7DR_1_1113_7303_19521	97KSI_03820_01798	99.47
	M00390_81_000000000-AA7DR_1_1114_11218_23596	97KSI_04078_01285	100.00

	M00390_81_000000000-AA7DR_1_1114_11222_26655	97KSI_00483_01396	100.00
	M00390_81_000000000-AA7DR_1_1114_11759_4336	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_1114_12633_10649	97KSI_01969_05288	100.00
	M00390_81_000000000-AA7DR_1_1114_12771_2177	97KSI_03525_00856	99.73
	M00390_81_000000000-AA7DR_1_1114_13097_17018	97KSI_01642_06778	100.00
	M00390_81_000000000-AA7DR_1_1114_13281_21120	97KSI_00532_04248	100.00
	M00390_81_000000000-AA7DR_1_1114_13862_7325	97KSI_04926_02920	100.00
	M00390_81_000000000-AA7DR_1_1114_15645_15671	97KSI_01035_03095	100.00
	M00390_81_000000000-AA7DR_1_1114_16470_27354	97KSI_01714_04568	100.00
	M00390_81_000000000-AA7DR_1_1114_16640_23341	97KSI_01924_01514	100.00
	M00390_81_000000000-AA7DR_1_1114_16932_3958	97KSI_02146_02453	99.73
	M00390_81_000000000-AA7DR_1_1114_17435_20606	97KSI_03488_06956	100.00
	M00390_81_000000000-AA7DR_1_1114_17987_9687	97KSI_01098_03584	100.00
	M00390_81_000000000-AA7DR_1_1114_20938_21424	97KSI_01222_07450	99.73
	M00390_81_000000000-AA7DR_1_1114_24743_8698	97KSI_00284_01155	100.00
	M00390_81_000000000-AA7DR_1_1114_25048_16944	97KSI_04945_00944	100.00
	M00390_81_000000000-AA7DR_1_1114_27901_17468	97KSI_03114_00768	100.00
	M00390_81_000000000-AA7DR_1_1114_3976_9516	97KSI_04946_01213	100.00
	M00390_81_000000000-AA7DR_1_1114_8755_21908	97KSI_00163_03224	100.00
	M00390_81_000000000-AA7DR_1_2101_10559_24670	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_2101_10960_18184	97KSI_03340_07170	100.00
	M00390_81_000000000-AA7DR_1_2101_11179_10541	97KSI_00074_04973	100.00
	M00390_81_000000000-AA7DR_1_2101_11461_26930	97KSI_03894_00766	100.00
	M00390_81_000000000-AA7DR_1_2101_11474_28568	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_2101_13016_5366	97KSI_00290_02888	99.73
	M00390_81_000000000-AA7DR_1_2101_15617_6895	97KSI_01117_04819	100.00
	M00390_81_000000000-AA7DR_1_2101_17640_3858	97KSI_04387_04031	99.73
	M00390_81_000000000-AA7DR_1_2101_19155_24278	97KSI_02874_07598	100.00
	M00390_81_000000000-AA7DR_1_2101_19403_26289	97KSI_02876_02931	100.00
	M00390_81_000000000-AA7DR_1_2101_21889_17023	97KSI_01222_07450	99.73
	M00390_81_000000000-AA7DR_1_2101_24497_5984	97KSI_05069_01868	99.47
	M00390_81_000000000-AA7DR_1_2101_5271_10562	97KSI_04737_01816	99.73
	M00390_81_000000000-AA7DR_1_2101_5652_21179	97KSI_04797_06930	100.00
	M00390_81_000000000-AA7DR_1_2101_5702_15459	97KSI_00647_03319	100.00
	M00390_81_000000000-AA7DR_1_2101_9413_13422	97KSI_00252_04181	99.73
	M00390_81_000000000-AA7DR_1_2101_9925_11464	97KSI_01207_04749	100.00

	M00390_81_000000000-AA7DR_1_2102_10796_16083	97KSI_01503_02896	100.00
	M00390_81_000000000-AA7DR_1_2102_14629_27271	97KSI_00844_05835	99.73
	M00390_81_000000000-AA7DR_1_2102_15928_24133	97KSI_03976_06338	100.00
	M00390_81_000000000-AA7DR_1_2102_16304_20120	97KSI_02922_04379	100.00
	M00390_81_000000000-AA7DR_1_2102_17962_19458	97KSI_03363_07025	100.00
	M00390_81_000000000-AA7DR_1_2102_19871_14399	97KSI_03292_07193	99.73
	M00390_81_000000000-AA7DR_1_2102_20221_17635	97KSI_01355_02233	99.74
	M00390_81_000000000-AA7DR_1_2102_20462_22578	97KSI_04141_05765	100.00
	M00390_81_000000000-AA7DR_1_2102_21439_19560	97KSI_01541_01528	100.00
	M00390_81_000000000-AA7DR_1_2102_21454_6492	97KSI_04722_01603	100.00
	M00390_81_000000000-AA7DR_1_2102_21772_5076	97KSI_02554_03871	100.00
	M00390_81_000000000-AA7DR_1_2102_4208_16235	97KSI_02221_01862	100.00
	M00390_81_000000000-AA7DR_1_2102_5549_6635	97KSI_04586_03967	100.00
	M00390_81_000000000-AA7DR_1_2102_7362_11629	97KSI_05246_02257	99.47
	M00390_81_000000000-AA7DR_1_2103_16137_11494	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2103_17071_11070	97KSI_01122_03589	100.00
	M00390_81_000000000-AA7DR_1_2103_19587_7983	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2103_20177_16769	97KSI_03525_00856	99.73
	M00390_81_000000000-AA7DR_1_2103_21453_17225	97KSI_03512_07397	100.00
	M00390_81_000000000-AA7DR_1_2103_23032_14402	97KSI_00907_04235	100.00
	M00390_81_000000000-AA7DR_1_2103_24105_11792	97KSI_02622_01430	100.00
	M00390_81_000000000-AA7DR_1_2103_24798_23009	97KSI_04839_06993	99.74
	M00390_81_000000000-AA7DR_1_2103_25106_12067	97KSI_04825_01640	100.00
	M00390_81_000000000-AA7DR_1_2103_6765_6807	97KSI_04852_06284	99.73
	M00390_81_000000000-AA7DR_1_2103_7329_8598	97KSI_04576_06137	100.00
	M00390_81_000000000-AA7DR_1_2104_10106_24750	97KSI_04712_01509	99.47
	M00390_81_000000000-AA7DR_1_2104_10279_18534	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2104_13368_5663	97KSI_03496_05087	100.00
	M00390_81_000000000-AA7DR_1_2104_13530_11913	97KSI_04822_06010	100.00
	M00390_81_000000000-AA7DR_1_2104_16410_14060	97KSI_02200_00828	100.00
	M00390_81_000000000-AA7DR_1_2104_17095_27177	97KSI_00534_04066	100.00
	M00390_81_000000000-AA7DR_1_2104_17292_20281	97KSI_04462_02230	100.00
	M00390_81_000000000-AA7DR_1_2104_18396_11569	97KSI_03249_05481	100.00
	M00390_81_000000000-AA7DR_1_2104_19307_2758	97KSI_04387_04031	100.00
	M00390_81_000000000-AA7DR_1_2104_21076_21928	97KSI_00781_05708	100.00
	M00390_81_000000000-AA7DR_1_2104_22220_11603	97KSI_04586_06236	99.73

	M00390_81_000000000-AA7DR_1_2104_22244_23398	97KSI_02911_03818	99.47
	M00390_81_000000000-AA7DR_1_2104_2423_14612	97KSI_03133_02303	100.00
	M00390_81_000000000-AA7DR_1_2104_25479_17374	97KSI_01117_04819	99.73
	M00390_81_000000000-AA7DR_1_2104_27065_20549	97KSI_02277_04861	99.73
	M00390_81_000000000-AA7DR_1_2104_3257_12020	97KSI_00949_04344	100.00
	M00390_81_000000000-AA7DR_1_2104_9541_20449	97KSI_03765_01443	100.00
	M00390_81_000000000-AA7DR_1_2105_10580_9052	97KSI_01098_03584	99.73
	M00390_81_000000000-AA7DR_1_2105_18439_8573	97KSI_03923_02963	99.73
	M00390_81_000000000-AA7DR_1_2105_18702_2725	97KSI_03774_02908	99.73
	M00390_81_000000000-AA7DR_1_2105_19315_22983	97KSI_02634_04753	100.00
	M00390_81_000000000-AA7DR_1_2105_19521_16980	97KSI_04809_01292	100.00
	M00390_81_000000000-AA7DR_1_2105_20284_5220	97KSI_00678_05143	99.73
	M00390_81_000000000-AA7DR_1_2105_23909_17506	97KSI_00774_00457	99.73
	M00390_81_000000000-AA7DR_1_2105_28041_20617	97KSI_04520_04321	99.20
	M00390_81_000000000-AA7DR_1_2105_4475_13710	97KSI_03489_04334	100.00
	M00390_81_000000000-AA7DR_1_2105_8673_27165	97KSI_04520_04321	99.20
	M00390_81_000000000-AA7DR_1_2106_10345_7008	97KSI_03960_06202	100.00
	M00390_81_000000000-AA7DR_1_2106_11559_3665	97KSI_01098_06747	99.73
	M00390_81_000000000-AA7DR_1_2106_14439_4393	97KSI_02146_02453	99.73
	M00390_81_000000000-AA7DR_1_2106_14857_5762	97KSI_01062_02647	100.00
	M00390_81_000000000-AA7DR_1_2106_15824_28437	97KSI_02449_07412	100.00
	M00390_81_000000000-AA7DR_1_2106_16960_9965	97KSI_02460_03842	99.73
	M00390_81_000000000-AA7DR_1_2106_17290_11056	97KSI_02562_02328	99.73
	M00390_81_000000000-AA7DR_1_2106_22729_7720	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_2106_23257_22061	97KSI_03648_01276	100.00
	M00390_81_000000000-AA7DR_1_2106_25799_8358	97KSI_03665_02501	100.00
	M00390_81_000000000-AA7DR_1_2106_7578_14962	97KSI_02897_06461	100.00
	M00390_81_000000000-AA7DR_1_2106_7709_25022	97KSI_00695_03999	100.00
	M00390_81_000000000-AA7DR_1_2106_9774_24846	97KSI_03637_05047	100.00
	M00390_81_000000000-AA7DR_1_2107_10851_13942	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_2107_11521_18891	97KSI_02789_03576	100.00
	M00390_81_000000000-AA7DR_1_2107_13072_6385	97KSI_01708_02371	100.00
	M00390_81_000000000-AA7DR_1_2107_16543_15439	97KSI_02902_02071	100.00
	M00390_81_000000000-AA7DR_1_2107_18074_27173	97KSI_00587_06465	100.00
	M00390_81_000000000-AA7DR_1_2107_18786_20459	97KSI_03195_01710	100.00
	M00390_81_000000000-AA7DR_1_2107_18971_4875	97KSI_03019_07346	99.47

	M00390_81_000000000-AA7DR_1_2107_25509_9716	97KSI_03765_01443	99.73
	M00390_81_000000000-AA7DR_1_2107_5893_23497	97KSI_02161_00967	100.00
	M00390_81_000000000-AA7DR_1_2107_7192_20096	97KSI_04414_02542	99.73
	M00390_81_000000000-AA7DR_1_2107_8323_22034	97KSI_03019_07346	99.74
	M00390_81_000000000-AA7DR_1_2107_8410_19743	97KSI_03663_01512	99.73
	M00390_81_000000000-AA7DR_1_2107_8697_14774	97KSI_02881_05835	99.73
	M00390_81_000000000-AA7DR_1_2107_9244_24116	97KSI_02789_03576	99.73
	M00390_81_000000000-AA7DR_1_2107_9651_25453	97KSI_00974_03851	100.00
	M00390_81_000000000-AA7DR_1_2108_12545_11221	97KSI_04072_04602	100.00
	M00390_81_000000000-AA7DR_1_2108_13761_25321	97KSI_04730_00564	100.00
	M00390_81_000000000-AA7DR_1_2108_13867_19238	97KSI_02144_02403	100.00
	M00390_81_000000000-AA7DR_1_2108_15404_9580	97KSI_01534_00636	100.00
	M00390_81_000000000-AA7DR_1_2108_22035_24001	97KSI_02450_02524	100.00
	M00390_81_000000000-AA7DR_1_2108_24423_17837	97KSI_02041_00872	100.00
	M00390_81_000000000-AA7DR_1_2108_2829_12406	97KSI_00349_04546	100.00
	M00390_81_000000000-AA7DR_1_2108_4621_18479	97KSI_03682_04245	100.00
	M00390_81_000000000-AA7DR_1_2108_4686_7578	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2108_5362_18382	97KSI_00212_02403	100.00
	M00390_81_000000000-AA7DR_1_2108_7220_16086	97KSI_04737_01816	100.00
	M00390_81_000000000-AA7DR_1_2108_8038_17551	97KSI_05267_04117	100.00
	M00390_81_000000000-AA7DR_1_2108_9332_3887	97KSI_01206_04859	99.73
	M00390_81_000000000-AA7DR_1_2109_12662_19654	97KSI_00877_07286	100.00
	M00390_81_000000000-AA7DR_1_2109_13978_16069	97KSI_01865_02366	99.73
	M00390_81_000000000-AA7DR_1_2109_14327_12220	97KSI_04946_01213	99.73
	M00390_81_000000000-AA7DR_1_2109_17536_14372	97KSI_04244_03016	100.00
	M00390_81_000000000-AA7DR_1_2109_22862_17949	97KSI_04095_06760	100.00
	M00390_81_000000000-AA7DR_1_2109_23326_14538	97KSI_02135_03690	100.00
	M00390_81_000000000-AA7DR_1_2109_23367_19635	97KSI_04734_01105	99.73
	M00390_81_000000000-AA7DR_1_2109_24966_14488	97KSI_03585_00511	99.73
	M00390_81_000000000-AA7DR_1_2109_26240_17823	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_2109_27384_10333	97KSI_02154_04878	99.47
	M00390_81_000000000-AA7DR_1_2109_27964_16114	97KSI_04520_04321	99.73
	M00390_81_000000000-AA7DR_1_2109_6324_15238	97KSI_01807_03311	99.73
	M00390_81_000000000-AA7DR_1_2109_8450_15790	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_2110_10348_21199	97KSI_04818_02139	99.47
	M00390_81_000000000-AA7DR_1_2110_11897_25381	97KSI_03308_05912	100.00

	M00390_81_000000000-AA7DR_1_2110_12101_15963	97KSI_00270_02577	100.00
	M00390_81_000000000-AA7DR_1_2110_12819_22314	97KSI_02872_03000	99.73
	M00390_81_000000000-AA7DR_1_2110_14165_25549	97KSI_01033_06090	99.73
	M00390_81_000000000-AA7DR_1_2110_14301_8821	97KSI_03622_02935	100.00
	M00390_81_000000000-AA7DR_1_2110_15881_27179	97KSI_04805_04634	100.00
	M00390_81_000000000-AA7DR_1_2110_22066_18385	97KSI_04834_01018	100.00
	M00390_81_000000000-AA7DR_1_2110_26557_22032	97KSI_00844_05835	100.00
	M00390_81_000000000-AA7DR_1_2110_7360_6899	97KSI_03991_07212	100.00
	M00390_81_000000000-AA7DR_1_2110_9273_13717	97KSI_01479_03480	99.47
	M00390_81_000000000-AA7DR_1_2111_15139_12520	97KSI_00468_02426	100.00
	M00390_81_000000000-AA7DR_1_2111_16225_9778	97KSI_04271_02812	100.00
	M00390_81_000000000-AA7DR_1_2111_21122_24228	97KSI_01594_05894	99.73
	M00390_81_000000000-AA7DR_1_2111_24270_18806	97KSI_01206_04859	100.00
	M00390_81_000000000-AA7DR_1_2111_26413_19083	97KSI_02154_04878	99.73
	M00390_81_000000000-AA7DR_1_2111_27232_15351	97KSI_00844_05835	100.00
	M00390_81_000000000-AA7DR_1_2111_5228_18238	97KSI_02152_02890	99.73
	M00390_81_000000000-AA7DR_1_2111_6691_25308	97KSI_00893_06529	99.73
	M00390_81_000000000-AA7DR_1_2111_7271_24960	97KSI_03256_02872	100.00
	M00390_81_000000000-AA7DR_1_2111_9175_9245	97KSI_01206_04859	99.73
	M00390_81_000000000-AA7DR_1_2112_11002_6236	97KSI_03975_01169	99.73
	M00390_81_000000000-AA7DR_1_2112_11893_4010	97KSI_02227_02888	99.73
	M00390_81_000000000-AA7DR_1_2112_13597_23479	97KSI_03220_07398	100.00
	M00390_81_000000000-AA7DR_1_2112_14035_13127	97KSI_04009_00303	100.00
	M00390_81_000000000-AA7DR_1_2112_16269_12869	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2112_19587_25970	97KSI_04147_05597	100.00
	M00390_81_000000000-AA7DR_1_2112_2070_14066	97KSI_02366_03370	100.00
	M00390_81_000000000-AA7DR_1_2112_20747_7120	97KSI_02781_04932	100.00
	M00390_81_000000000-AA7DR_1_2112_21936_11046	97KSI_04852_06284	99.73
	M00390_81_000000000-AA7DR_1_2112_23769_13792	97KSI_03041_04640	99.73
	M00390_81_000000000-AA7DR_1_2112_2708_10279	97KSI_01005_02504	100.00
	M00390_81_000000000-AA7DR_1_2112_28846_15117	97KSI_02337_07107	99.73
	M00390_81_000000000-AA7DR_1_2112_6003_19825	97KSI_03876_06624	100.00
	M00390_81_000000000-AA7DR_1_2112_6092_23830	97KSI_00775_03521	100.00
	M00390_81_000000000-AA7DR_1_2112_6103_23850	97KSI_05069_01868	99.73
	M00390_81_000000000-AA7DR_1_2113_10012_16875	97KSI_00742_03662	99.73
	M00390_81_000000000-AA7DR_1_2113_10927_8701	97KSI_02898_05450	100.00

	M00390_81_000000000-AA7DR_1_2113_12507_23692	97KSI_03561_05871	99.73
	M00390_81_000000000-AA7DR_1_2113_12929_19411	97KSI_00993_01026	99.73
	M00390_81_000000000-AA7DR_1_2113_15500_21922	97KSI_05130_06212	100.00
	M00390_81_000000000-AA7DR_1_2113_16191_16892	97KSI_02913_03689	99.73
	M00390_81_000000000-AA7DR_1_2113_19441_18399	97KSI_00775_01027	100.00
	M00390_81_000000000-AA7DR_1_2113_21239_18948	97KSI_00529_01262	100.00
	M00390_81_000000000-AA7DR_1_2113_23441_12993	97KSI_01185_04616	100.00
	M00390_81_000000000-AA7DR_1_2113_2379_18430	97KSI_02154_04878	99.47
	M00390_81_000000000-AA7DR_1_2113_5142_16083	97KSI_03820_01798	99.73
	M00390_81_000000000-AA7DR_1_2113_6080_24551	97KSI_04818_02139	99.73
	M00390_81_000000000-AA7DR_1_2113_6687_19954	97KSI_04905_04178	100.00
	M00390_81_000000000-AA7DR_1_2113_9152_17289	97KSI_04365_03064	100.00
	M00390_81_000000000-AA7DR_1_2114_13157_21784	97KSI_03663_01512	100.00
	M00390_81_000000000-AA7DR_1_2114_17524_25803	97KSI_00299_01386	100.00
	M00390_81_000000000-AA7DR_1_2114_18666_9774	97KSI_03830_02551	100.00
	M00390_81_000000000-AA7DR_1_2114_22473_15974	97KSI_00787_03593	100.00
	M00390_81_000000000-AA7DR_1_2114_22648_22014	97KSI_00500_04401	100.00
	M00390_81_000000000-AA7DR_1_2114_23836_19289	97KSI_04738_02066	100.00
	M00390_81_000000000-AA7DR_1_2114_25330_24954	97KSI_02897_07535	100.00
	M00390_81_000000000-AA7DR_1_2114_9691_5295	97KSI_02154_04878	99.73
***C*. sp. Na26B1**	M00390_80_000000000-AA759_1_1104_19320_19832	97KSI_01885_05851	99.47
	M00390_80_000000000-AA759_1_1107_13645_27564	97KSI_01885_05851	99.73
	M00390_80_000000000-AA759_1_1107_15354_12167	97KSI_01911_00721	99.73
	M00390_80_000000000-AA759_1_1111_7314_18049	97KSI_00361_06437	100.00
	M00390_80_000000000-AA759_1_1113_17535_10310	97KSI_04093_05675	100.00
	M00390_80_000000000-AA759_1_1114_14905_11144	97KSI_00794_03497	100.00
	M00390_80_000000000-AA759_1_2108_18456_25450	97KSI_01986_05212	99.73
	M00390_80_000000000-AA759_1_2111_9841_2910	97KSI_01986_05212	99.73
	M00390_80_000000000-AA759_1_2112_10861_13680	97KSI_01986_05212	99.73
	M00390_81_000000000-AA7DR_1_1101_16198_12414	97KSI_01986_05212	100.00
	M00390_81_000000000-AA7DR_1_1101_21520_5635	97KSI_04233_02052	99.73
	M00390_81_000000000-AA7DR_1_1101_28091_16249	97KSI_04826_01611	99.73
	M00390_81_000000000-AA7DR_1_1101_9470_25874	97KSI_01986_05212	99.73
	M00390_81_000000000-AA7DR_1_1102_17020_5965	97KSI_03486_05518	100.00
	M00390_81_000000000-AA7DR_1_1102_6690_11966	97KSI_00666_01179	100.00

	M00390_81_000000000-AA7DR_1_1103_17202_26051	97KSI_01364_03881	100.00
	M00390_81_000000000-AA7DR_1_1103_22901_8165	97KSI_01348_06387	100.00
	M00390_81_000000000-AA7DR_1_1103_7731_18975	97KSI_02811_03406	100.00
	M00390_81_000000000-AA7DR_1_1104_22570_9473	97KSI_01986_05212	100.00
	M00390_81_000000000-AA7DR_1_1104_27372_19474	97KSI_01885_05851	100.00
	M00390_81_000000000-AA7DR_1_1104_5102_14191	97KSI_00610_01398	100.00
	M00390_81_000000000-AA7DR_1_1105_28183_17335	97KSI_01986_05212	99.73
	M00390_81_000000000-AA7DR_1_1106_19721_26938	97KSI_01612_01703	99.47
	M00390_81_000000000-AA7DR_1_1106_22711_22894	97KSI_03374_05196	99.73
	M00390_81_000000000-AA7DR_1_1107_20021_20682	97KSI_04167_01063	100.00
	M00390_81_000000000-AA7DR_1_1107_25259_9417	97KSI_03931_01058	99.73
	M00390_81_000000000-AA7DR_1_1108_8011_10045	97KSI_01686_02713	100.00
	M00390_81_000000000-AA7DR_1_1109_15430_22502	97KSI_03277_06223	99.73
	M00390_81_000000000-AA7DR_1_1109_19232_2903	97KSI_01133_03320	99.73
	M00390_81_000000000-AA7DR_1_1109_21257_16345	97KSI_00130_03851	100.00
	M00390_81_000000000-AA7DR_1_1110_13821_18158	97KSI_03288_01515	100.00
	M00390_81_000000000-AA7DR_1_1110_19023_18555	97KSI_02078_01146	100.00
	M00390_81_000000000-AA7DR_1_1111_20391_3859	97KSI_00435_06774	99.73
	M00390_81_000000000-AA7DR_1_1113_10186_25118	97KSI_01984_02000	100.00
	M00390_81_000000000-AA7DR_1_1113_16061_26576	97KSI_04875_05703	99.73
	M00390_81_000000000-AA7DR_1_1114_20155_18983	97KSI_00136_06038	99.73
	M00390_81_000000000-AA7DR_1_2101_11669_5727	97KSI_03460_00492	100.00
	M00390_81_000000000-AA7DR_1_2101_20292_12467	97KSI_03862_07101	100.00
	M00390_81_000000000-AA7DR_1_2101_5809_13550	97KSI_01986_05212	99.73
	M00390_81_000000000-AA7DR_1_2101_6397_18582	97KSI_01986_05212	100.00
	M00390_81_000000000-AA7DR_1_2103_22817_8892	97KSI_03058_02877	99.47
	M00390_81_000000000-AA7DR_1_2104_11527_27123	97KSI_05262_02726	99.73
	M00390_81_000000000-AA7DR_1_2104_2589_16352	97KSI_03449_06178	99.73
	M00390_81_000000000-AA7DR_1_2104_6697_8778	97KSI_03758_05617	100.00
	M00390_81_000000000-AA7DR_1_2105_20409_3391	97KSI_03921_02093	99.73
	M00390_81_000000000-AA7DR_1_2105_24095_5901	97KSI_01251_06803	100.00
	M00390_81_000000000-AA7DR_1_2105_25848_16533	97KSI_02800_01959	99.73
	M00390_81_000000000-AA7DR_1_2105_2682_19266	97KSI_03758_05617	99.73
	M00390_81_000000000-AA7DR_1_2107_20122_6543	97KSI_01941_01649	99.73
	M00390_81_000000000-AA7DR_1_2107_6490_11438	97KSI_03758_05617	99.73
	M00390_81_000000000-AA7DR_1_2108_10756_6054	97KSI_03923_05450	99.73

	M00390_81_000000000-AA7DR_1_2108_15634_22907	97KSI_01393_04532	100.00
	M00390_81_000000000-AA7DR_1_2109_22304_14960	97KSI_02828_01344	99.73
	M00390_81_000000000-AA7DR_1_2110_24767_14132	97KSI_00329_05507	99.73
	M00390_81_000000000-AA7DR_1_2110_7577_5967	97KSI_01986_05212	99.73
	M00390_81_000000000-AA7DR_1_2111_19997_13380	97KSI_04190_02727	100.00
	M00390_81_000000000-AA7DR_1_2112_8563_10293	97KSI_02913_07422	100.00
	M00390_81_000000000-AA7DR_1_2113_12146_19815	97KSI_00915_03044	100.00
	M00390_81_000000000-AA7DR_1_2113_23019_13152	97KSI_01986_05212	99.73
C. tenuissimus	M00390_81_000000000-AA7DR_1_1101_10404_7536	97KSI_04069_00269	99.74
	M00390_81_000000000-AA7DR_1_1101_12020_2561	97KSI_02748_05003	100.00
	M00390_81_000000000-AA7DR_1_1101_12750_12735	97KSI_02983_06904	100.00
	M00390_81_000000000-AA7DR_1_1101_12972_26204	97KSI_00416_02071	99.74
	M00390_81_000000000-AA7DR_1_1101_13180_26564	97KSI_02658_06876	99.73
	M00390_81_000000000-AA7DR_1_1101_17418_16093	97KSI_04894_03826	100.00
	M00390_81_000000000-AA7DR_1_1101_19019_24975	97KSI_02001_02819	100.00
	M00390_81_000000000-AA7DR_1_1101_19124_2835	97KSI_03794_07273	100.00
	M00390_81_000000000-AA7DR_1_1101_19390_3055	97KSI_00416_02071	100.00
	M00390_81_000000000-AA7DR_1_1101_19499_23881	97KSI_01509_04247	100.00
	M00390_81_000000000-AA7DR_1_1101_19579_23505	97KSI_00974_05580	99.74
	M00390_81_000000000-AA7DR_1_1101_19596_23537	97KSI_01970_01571	100.00
	M00390_81_000000000-AA7DR_1_1101_20308_10428	97KSI_00368_02076	100.00
	M00390_81_000000000-AA7DR_1_1101_21482_26412	97KSI_01763_06817	100.00
	M00390_81_000000000-AA7DR_1_1101_22681_24682	97KSI_00447_06375	100.00
	M00390_81_000000000-AA7DR_1_1101_22809_8498	97KSI_00596_07083	99.74
	M00390_81_000000000-AA7DR_1_1101_24859_6431	97KSI_02268_05990	99.74
	M00390_81_000000000-AA7DR_1_1101_25018_11322	97KSI_00093_05704	100.00
	M00390_81_000000000-AA7DR_1_1101_25410_22880	97KSI_00814_03179	100.00
	M00390_81_000000000-AA7DR_1_1101_27830_16975	97KSI_03912_02912	100.00
	M00390_81_000000000-AA7DR_1_1101_28042_10043	97KSI_00404_01043	100.00
	M00390_81_000000000-AA7DR_1_1101_6592_8549	97KSI_04782_03918	99.74
	M00390_81_000000000-AA7DR_1_1101_7389_18078	97KSI_00652_03222	100.00
	M00390_81_000000000-AA7DR_1_1102_13482_6478	97KSI_01911_03168	100.00
	M00390_81_000000000-AA7DR_1_1102_18446_3596	97KSI_01987_05213	100.00
	M00390_81_000000000-AA7DR_1_1102_20566_15022	97KSI_03923_05350	99.74
	M00390_81_000000000-AA7DR_1_1102_21169_3932	97KSI_01798_07171	99.74

	M00390_81_000000000-AA7DR_1_1102_21445_15640	97KSI_01290_05422	99.74
	M00390_81_000000000-AA7DR_1_1102_23013_11309	97KSI_03894_07381	100.00
	M00390_81_000000000-AA7DR_1_1102_25920_11969	97KSI_00250_04935	99.74
	M00390_81_000000000-AA7DR_1_1102_28179_19330	97KSI_02564_05554	99.74
	M00390_81_000000000-AA7DR_1_1102_4023_9561	97KSI_00416_02071	100.00
	M00390_81_000000000-AA7DR_1_1102_7026_20095	97KSI_02003_04853	100.00
	M00390_81_000000000-AA7DR_1_1102_8293_25185	97KSI_02113_03202	100.00
	M00390_81_000000000-AA7DR_1_1103_11989_21537	97KSI_01150_04900	100.00
	M00390_81_000000000-AA7DR_1_1103_16869_20837	97KSI_01982_02174	99.74
	M00390_81_000000000-AA7DR_1_1103_20219_18162	97KSI_04538_02642	100.00
	M00390_81_000000000-AA7DR_1_1103_20846_6606	97KSI_01561_07286	99.74
	M00390_81_000000000-AA7DR_1_1103_21707_22085	97KSI_03738_02676	100.00
	M00390_81_000000000-AA7DR_1_1103_26982_11585	97KSI_03391_02324	100.00
	M00390_81_000000000-AA7DR_1_1103_4069_22966	97KSI_02642_05819	99.74
	M00390_81_000000000-AA7DR_1_1103_6137_19542	97KSI_02347_01990	100.00
	M00390_81_000000000-AA7DR_1_1103_9398_12278	97KSI_00416_02071	100.00
	M00390_81_000000000-AA7DR_1_1103_9573_27386	97KSI_01137_07114	100.00
	M00390_81_000000000-AA7DR_1_1104_19600_23603	97KSI_00182_03020	100.00
	M00390_81_000000000-AA7DR_1_1104_25687_10430	97KSI_03871_06777	99.74
	M00390_81_000000000-AA7DR_1_1104_3853_19559	97KSI_00416_02071	100.00
	M00390_81_000000000-AA7DR_1_1104_3967_7974	97KSI_05087_02289	99.74
	M00390_81_000000000-AA7DR_1_1105_12334_10159	97KSI_00476_02252	99.74
	M00390_81_000000000-AA7DR_1_1105_14401_23759	97KSI_03045_03726	100.00
	M00390_81_000000000-AA7DR_1_1105_19354_7845	97KSI_01283_00600	100.00
	M00390_81_000000000-AA7DR_1_1105_20508_3587	97KSI_03925_06427	100.00
	M00390_81_000000000-AA7DR_1_1105_25599_18512	97KSI_00543_01739	99.74
	M00390_81_000000000-AA7DR_1_1106_19455_13394	97KSI_00373_04914	99.74
	M00390_81_000000000-AA7DR_1_1106_24238_20243	97KSI_03847_07014	99.74
	M00390_81_000000000-AA7DR_1_1107_22105_11345	97KSI_02624_05741	100.00
	M00390_81_000000000-AA7DR_1_1107_22838_6403	97KSI_04102_01861	99.74
	M00390_81_000000000-AA7DR_1_1108_11446_5613	97KSI_02124_07354	100.00
	M00390_81_000000000-AA7DR_1_1108_16719_27399	97KSI_01118_03038	99.74
	M00390_81_000000000-AA7DR_1_1108_24620_23287	97KSI_01788_02370	100.00
	M00390_81_000000000-AA7DR_1_1110_13450_22152	97KSI_00416_02071	99.74
	M00390_81_000000000-AA7DR_1_1110_14154_2636	97KSI_02805_02859	100.00
	M00390_81_000000000-AA7DR_1_1110_20392_15399	97KSI_04483_06080	99.74

	M00390_81_000000000-AA7DR_1_1111_17352_7130	97KSI_04894_03826	99.74
	M00390_81_000000000-AA7DR_1_1111_22470_4248	97KSI_04474_06763	100.00
	M00390_81_000000000-AA7DR_1_1111_4309_13169	97KSI_00650_03757	100.00
	M00390_81_000000000-AA7DR_1_1112_15167_7060	97KSI_03963_04914	99.74
	M00390_81_000000000-AA7DR_1_1112_1825_13639	97KSI_04945_03094	99.74
	M00390_81_000000000-AA7DR_1_1112_4560_17960	97KSI_01367_04277	100.00
	M00390_81_000000000-AA7DR_1_1113_11707_26787	97KSI_00447_06861	100.00
	M00390_81_000000000-AA7DR_1_1113_6922_6952	97KSI_01418_01150	100.00
	M00390_81_000000000-AA7DR_1_1113_7867_24551	97KSI_02642_05460	99.74
	M00390_81_000000000-AA7DR_1_1114_18087_7032	97KSI_03211_03001	100.00
	M00390_81_000000000-AA7DR_1_1114_19091_18825	97KSI_00729_00865	100.00
	M00390_81_000000000-AA7DR_1_2101_10811_4358	97KSI_05110_01509	99.74
	M00390_81_000000000-AA7DR_1_2101_12520_13351	97KSI_00982_02890	100.00
	M00390_81_000000000-AA7DR_1_2101_14615_2053	97KSI_05085_06319	99.74
	M00390_81_000000000-AA7DR_1_2101_22809_10634	97KSI_00416_02071	99.74
	M00390_81_000000000-AA7DR_1_2101_7082_9254	97KSI_00058_04755	100.00
	M00390_81_000000000-AA7DR_1_2102_15853_27724	97KSI_01562_07249	100.00
	M00390_81_000000000-AA7DR_1_2102_24515_23008	97KSI_01667_02618	99.74
	M00390_81_000000000-AA7DR_1_2102_25696_22060	97KSI_00416_02071	100.00
	M00390_81_000000000-AA7DR_1_2102_3405_15446	97KSI_04363_04801	99.74
	M00390_81_000000000-AA7DR_1_2103_25851_13587	97KSI_04752_05541	100.00
	M00390_81_000000000-AA7DR_1_2103_3110_14542	97KSI_02761_06988	100.00
	M00390_81_000000000-AA7DR_1_2104_10184_3810	97KSI_04183_05614	99.74
	M00390_81_000000000-AA7DR_1_2104_6780_20861	97KSI_00093_05704	99.74
	M00390_81_000000000-AA7DR_1_2105_22163_6261	97KSI_02635_05560	100.00
	M00390_81_000000000-AA7DR_1_2105_5213_17961	97KSI_02364_02710	100.00
	M00390_81_000000000-AA7DR_1_2105_8136_7588	97KSI_04453_01298	99.74
	M00390_81_000000000-AA7DR_1_2107_28029_16628	97KSI_00974_05580	100.00
	M00390_81_000000000-AA7DR_1_2108_11075_7121	97KSI_01270_06801	100.00
	M00390_81_000000000-AA7DR_1_2108_13462_27393	97KSI_00451_06500	99.74
	M00390_81_000000000-AA7DR_1_2108_15462_20574	97KSI_01170_03272	100.00
	M00390_81_000000000-AA7DR_1_2109_6008_17235	97KSI_04277_06569	100.00
	M00390_81_000000000-AA7DR_1_2110_12921_14598	97KSI_01873_04488	100.00
	M00390_81_000000000-AA7DR_1_2111_25786_13526	97KSI_04746_03424	99.74
	M00390_81_000000000-AA7DR_1_2113_21512_11176	97KSI_02884_04325	100.00
	M00390_81_000000000-AA7DR_1_2113_23993_5770	97KSI_03391_02324	99.74

	M00390_81_000000000-AA7DR_1_2114_12713_17929	97KSI_01150_06546	100.00
	M00390_81_000000000-AA7DR_1_2114_24850_7921	97KSI_02564_05554	100.00
	M00390_81_000000000-AA7DR_1_2114_25314_9203	97KSI_02622_03603	100.00

Chapter VI

Concluding remarks and future perspectives

6.1. Concluding remarks

This Ph.D. thesis embodies my contribution to the understanding of the evolution of the marine diatom family Chaetocerotaceae and, in particular, of the genus *Chaetoceros* by means of molecular data. In some cases, molecular data have been used in their canonical way and proved to be conclusive for the purposes they were intended to. This was, for example, the case of the multigene phylogeny inferred in Chapter II to assess the evolutionary history of Chaetocerotaceae. In other cases, molecular data (especially in the form of metabarcoding data), have been used in a new, different way and played the role of main actors in stories that went beyond *Chaetoceros* or diatoms in general. This is what happened in Chapters III, IV and V, in which I have designed a series of experiments that have shown the potential of metabarcoding data in so far unexplored contexts.

For **Chapter II**, I started my experiments with the initial idea of inferring a multigene phylogeny of the family Chaetocerotaceae to resolve terminal or internal relationships that were poorly supported in previous nuclear phylogenies (e.g. Kooistra et al., 2010; Gaonkar et al., 2018). Then, considering that in our lab we had reached a considerable number of strains of Chaetocerotaceae belonging to different species around the world and that there was a renewed interest in revision of sections triggered by the discovery of new species (e.g. Li et al., 2013; 2016; Xu et al., 2019), I decided to change my plans. I kept the initial idea of inferring a multigene phylogeny for the family Chaetocerotaceae, but I also decided to test the traditional classification scheme based in generic and infrageneric (subgenera and sections) divisions using the inferred phylogeny as backbone. Taxonomies are not neutral, but they reflect (or even create) the hypothesis on the structure of living world (Gould and Vrba, 1982). When one looks at how people classify things, one also understands how they think (Foucault, 1970). Therefore, I aimed at a classification scheme that was supported phylogenetically but also retained practical properties, following the thinking of Mayr (1982) and Benton (2000). My classification of Chaetocerotaceae had to

group together species similar because of common descent (phylogenetically informative) and in the meantime allow these groups to aid people in the identification of new or already known species (utilitarian principle, practical purpose). I dusted off the traditional classification scheme and, with some adjustments (emendation of one section, rejection of seven and erection of three new ones) and I made it fit to the clades of the inferred multigene phylogeny. This work made it possible to keep most of the traditional systematic terminology but in the light of a modern and updated interpretation. I tried to avoid leaving clades nameless wherever and whenever I could, because I believe that things without a name tend to be disregarded. Furthermore, giving priority to the utilitarian criterion, I refrained from classifying the major, well-supported clades within *Chaetoceros* into their own genera, since *Chaetoceros* species are easily recognised by their defining feature, the setae, whereas each of such more narrowly defined genera would not be recognised so easily. Splitting would have created a series of genera that are not always easy to distinguish.

In **Chapter III** I have shown how the integration of classical occurrence data and new ones (metabarcoding data) can be used to obtain a comprehensive assessment of the distribution of species, especially of microscopic ones such as protists. Classical occurrence data as reports of scientific expeditions, floras and faunas and checklists have formed the main sources of primary biodiversity data for inferring species distribution (Droege et al., 1998; Chapman, 2005). In recent years, occurrence data have also been gathered from a large variety of sources as satellite tracking and direct or remote observation (He et al., 2015), frozen tissue collections and seed banks (Chapman, 2005), environmental DNA (August et al., 2015), and citizen science initiatives (Devictor et al., 2010; Hochachka et al., 2012). However, a big step forward has been done with the adding of DNA information to classical approaches. This kind of data, have revolutionised the study of protistan diversity (Leray and Knowlton, 2016; Caron and Hu, 2018) that was

before exclusively based on morphological studies. For marine protists indeed, there are several challenges related to the assessment of diversity and distribution at different taxonomic levels, the species one being particularly difficult. Cryptic diversity is widespread (Smayda, 2011; Amato et al., 2019) and traditional analyses based on microscopy are time-consuming and require taxonomic expertise (Culverhouse, 2007). The availability of global metabarcoding datasets as Ocean Sampling Day (OSD, Kopf et al., 2015) and Tara Oceans (de Vargas et al., 2015) has offered a valuable source of sequence and occurrence data that fostered the assessment of diversity and distribution of several marine taxa (de Vargas et al., 2015; Malviya et al., 2016; Tragin and Vaulot, 2018; 2019). In contrast to Tara Oceans, which sampled different marine regions at different times of the year, OSD is a simultaneous sampling of coastal regions (mostly Northern Hemisphere), which allows analysis of spatial distribution patterns of species without the impact of seasonality (Tragin and Vaulot, 2019). For my thesis work, I decided to use the information available in these metaborcoding datasets together with other stored in public repositories (GBIF and OBIS) as well as phytoplankton checklists or floras to show how the integration of these data can contribute to insight in the biogeography and diversity at the genus- and species-level in *Chaetoceros*. I extracted *Chaetoceros* records from GBIF and OBIS, collected literature data by means of a Google Scholar search and mapped *Chaetoceros* references barcodes against OSD (144 sites) and Tara Oceans (210 sites). I compared the resolution of these different data sources in determining the global distribution of the genus and provided examples, at the species level, of detection of cryptic species, endemism and cosmopolitan or restricted distributions. Of all the non-molecular data, the most complete picture of *Chaetoceros* distribution was provided by the GBIF and OBIS platforms, which contain a huge amount of data from different sources and cover a wide time scale. The search on Google Scholar could be considered as a convenient starting place to commence a literature search but not an endpoint. The two

global metabarcoding datasets OSD and Tara Oceans provided an overall distribution of the genus that was comparable to the one obtained from GBIF and OBIS. This proved that, despite their bias in space and time, metabarcoding data can compete with classical occurrence data gathered over hundreds of years. I also produced maps for the genus containing info about occurrence, species richness and abundance, as well as *Chaetoceros* species distribution maps from OSD and Tara Oceans data. Finally yet importantly, in this chapter I have provided a pipeline to study occurrence and diversity of taxa for which reference barcodes and metabarcoding data are available.

As stated in the Abstract of this Ph.D. thesis, the initial aim of **Chapter IV** was to infer the phylogeographic pattern of selected *Chaetoceros* species by means of Sanger sequencing of a few genes from specimens collected around the world. Then, it turned into the analysis of the *C. curvisetus* species complex inferring haplotype networks from metabarcoding data. The choice of changing strategy was made to take advantage of the global metabarcoding datasets of OSD and Tara Oceans, which together covered about 350 sampling sites across coastal and open ocean waters of both hemispheres. Reaching even a small fraction of such sampling localities would have been hard considering the duration of a Ph.D. program, and the costs related to the selection and sequencing of target gene regions quite high. Then, the change of the subject, from the comparison of the phylogeographic patterns of different *Chaetoceros* species to the analysis of a species complex, was a consequence of the results I obtained from the multigene phylogeny inferred in Chapter II. Indeed, some phylogenetic relationships among *C. curvisetus* species were not fully resolved even including more loci, which made me suppose that the relationships among them were more complex than simple dichotomies. Therefore, I decided to explore the patterns of genetic variation of 18S gene (V4 and V9 regions) across space to reconstruct the evolutionary relationships of the aforementioned species.

Metabarcoding data have been used so far in the form of phylogenetic trees or OTU clustering (Nanjappa et al., 2014; Gaonkar, 2017; Pargana, 2017; Tragin and Vaulot, 2019) to delimit species in protists, but none has built haplotype networks with them.

For my experiment, I started from a set of reference barcodes of *C. curvisetus* spp. produced by Gaonkar et al. (2018) and myself (e.g. strains of the Red Sea, see Chapter II) and two global metabarcoding datasets (OSD and Tara Oceans). The latter datasets allowed me to explore the genetic diversity within my cryptic species complex in a way that would have been hard to reach with classical Sanger sequencing data. Since the object of my study was a species complex, I supposed that the best way to analyse it was by means of phylogenetic haplotype networks rather than phylogenetic trees. Therefore, I set up several criteria to delimit species from my networks. Then, I validated at molecular level the species inferred above by means of inference of Maximum Likelihood phylogenetic trees and calculation of genetic distances. After this, I moved to an ecological level, and I have mapped the inferred *C. curvisetus* species in the biogeographic provinces of Longhurst (2007) using the information contained in the two global metabarcoding datasets. This latter exercise allowed me to test from the ecological perspective the species I have inferred from genetic data.

In conclusion, I confirmed as species the initial taxa for which I had reference barcodes and that there are four more molecularly defined taxonomic units (MOTUs) that need further investigation, some of which are likely to constitute species new to science. Furthermore, within the *C. curvisetus* species complex it seems to still be gene flow.

The final experiment of this Ph.D. thesis, described in **Chapter V**, initially was not planned at all and resulted from some preliminary results of Chapter IV. It is a story of concerted evolution of 18S gene in several *Chaetoceros* species and inferred from metabarcoding data. Concerted evolution is the mode of evolution of some genes and non-coding regions across all the major branches of the Tree of Life and was first detected by

hybridisation studies and successively by phylogenetic approaches (Graur and Li, 1999). Here, for the first time, I showed how metabarcoding data and single strain high throughput sequencing (HTS) could be used to study this biological phenomenon. Using such data in the form of abundance plots, BLAST analysis and haplotype networks, I have demonstrated that concerted evolution is occurring in all of the investigated species, and all the methodologies here used for its detection are conclusive and easy to perform.

The work presented in this chapter also demonstrated that there are no consequences for DNA barcoding due to the occurrence, within each *Chaetoceros* strain, of thousands of 18S ribotypes. Indeed, one of the copies, identified as the dominant haplotype, is far more abundant that all the others that the probability that a "minor" haplotype is sequenced with Sanger chemistry is almost null. I have also demonstrated that, when conducting metabarcoding experiments (from both environmental samples and bulk communities) or single strain HTS, the most abundant haplotype that is recovered for each species corresponds to the sequence that would be obtained by Sanger sequencing. However, this study also highlighted that the high number of sequences occurring at low abundances (minor haplotypes) can inflate diversity assessments inferred from metabarcoding data.

In conclusion, my Ph.D. thesis:

- is a contribution to the systematics of the family Chaetocerotaceae (Chapter II);
- provides an assessment of the diversity and distribution of the genus *Chaetoceros* by integrating classical and novel primary biodiversity data (Chapter III);
- shows a new way to analyse a cryptic species complex using the potential of spatial data contained in global metabarcoding datasets in the form of phylogenetic networks (Chapter IV);
- illustrates how starting from the data contained in a temporal metabarcoding dataset (MareChiara), I have formulated the hypothesis of concerted evolution in

Chaetoceros that was successively tested with an appropriate experimental design (single strain HTS and targeted analyses).

6.2. Future perspectives

During the 3 years of my Ph.D. program, I performed several experiments that have contributed to my understanding of the diversity and evolution of the family Chaetocerotaceae and, in particular, of the genus *Chaetoceros*. However, no experiment is to be considered definitive and, in this sense, my Ph.D. opens several research perspectives. Strictly related to the work performed in this thesis, I see the following possibilities. The multigene phylogeny (Chapter II), although taxonomically comprehensive regarding the known diversity, leaves a few additional sections still to be investigated. Future work needs to include species not treated here to test the validity of my proposed classification system, and to provide a more comprehensive view of the evolutionary history of this family. In particular, the addition of molecular data for *C. bacteriastroides* will clarify its phylogenetic position in the family Chaetocerotaceae. This species exhibits features of *Bacteriastrum* and *Chaetoceros*. Hernández-Becerril (1993) placed this species into its own subgenus (*Bacteriastroidea*), and here I considered it as section until new data become available.

Another course of action is to infer a cladogram from morphological characters and their states, especially ultrastructural ones, to ascertain if and in how far its topology agrees with that of the molecular phylogeny. Ultrastructural details of valves and setae are increasingly becoming available for Chaetocerotaceae (e.g. Chamnansinp et al., 2015; Bosak and Sarno, 2017; Gaonkar et al., 2018; Xu et al., 2019). The sections here investigated could be further supported by these data, providing a more robust basis to the hypothesis of evolutionary relationships here inferred.

About the inference of genus and species distributions through integration of classical and novel strategies (Chapter III), future research could include the application of this pipeline to different taxa. Integration of metabarcoding occurrence data with classical ones can be used for conservation planning (Rondinini et al., 2006; Newbold, 2010), species distribution models (Elith et al., 2006; Lütolf et al., 2006), and many other ecological and evolutionary applications. Similarly, metabarcoding data could be used to infer phylogeography or analyse other marine species complexes, as I did in Chapter IV. In addition, the molecular data analysed in this thesis, especially in Chapter III and IV (analysis of the genus *Chaetoceros* and the *C. curvisetus* species complex respectively) could be integrated with the large amount of imaging data produced by the TARA Oceans initiative. These imaging data, collectively included in the T.A.O.M.I (TAra Oceans Marine biology Imaging) platform, refer to observations of plankton organisms (from a few micrometres to one centimetre) gathered from flow analysis, microscopy and macro-photography. Images from flow analysis were obtained using the FlowCam, an equipment consisting of a cytometer and a microscope enabling to swiftly follow organisms of very different sizes, whilst microscopy pictures were taken by means of stereomicroscopy, fluorescent microscopy and fluorescent microscopy with phase. Finally, T.A.O.M.I. platform also includes a video and macro-photographies of large planktonic organisms (e.g. larvae, jellyfish, etc.), as well as corals and macroscopic algae. Further information is available at https://oceans.taraexpeditions.org/en/m/science/news/imaging-during-tara-oceans/.

The work on concerted evolution in *Chaetoceros* illustrated in Chapter V could be integrated with an assessment of rDNA copy number in the species here investigated using the combination of single-cell approaches and Digital PCR. Indeed, starting with the extraction of RNA from a single cell, using Digital PCR will be possible to count all the copies of rDNA genes (or a specific gene of the cistron). These data will be then compared

with the ones here obtained from high throughput sequencing to determine the number of rDNA copies occurring within each *Chaetoceros* strain. Besides this, it would be interesting to analyse the distribution of 18S-V4 haplotypes here obtained over time in many species to assess how concerted evolution interplays with other processes (drift, geographic patterning, migration, gene flow, spore capital).

Apart from establishing HTS metabarcode time series in many coastal and oceanic regions and connecting the obtained data for meta-analysis, I believe there are several topics related to the ones treated in my thesis worthy of being investigated. For example, I believe that time has come to sequence the genomes of several *Chaetoceros* species, taking advantage of reduction of time and costs of sequencing technologies. The occurrence within the genus of species with different life-strategies and habits (e.g. spore formers vs. non spore formers; coastal vs. oceanic species), morphological features (chloroplasts migrating in the setae vs. only in the cell body) and different usage of silica for setae formation (thin vs. thick), just to cite a few examples, allows for comparative genomics studies. Indeed, comparing the genomes of closely related species with different ecological and/or morphological traits (e.g. the members of the sections *Chaetoceros* versus *Protuberantia*) may reveal genetic factors responsible for these characteristics. Furthermore, the comparison of the genomes of *Chaetoceros* species versus other diatoms (publicly available), will shed light on the structure and function of core genes responsible for silica production and translocation, putative new genes involved in the formation of setae as well as carbon metabolism (e.g. C3 vs. C4).

References

Amato, A., Kooistra, W. H. C. F., Montresor, M. (2019). Cryptic diversity: a long-lasting issue for diatomologists. *Protist, 170*(1), 1-7.

August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., Jepson, P. (2015). Emerging technologies for biological recording. *Biological Journal of the Linnean Society, 115*(3), 731-749.

Benton, M. J. (2000). Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? *Biological Reviews, 75*(4), 633-648.

Bosak, S., Sarno, D. (2017). The planktonic diatom genus *Chaetoceros* Ehrenberg (Bacillariophyta) from the Adriatic Sea. *Phytotaxa*, *314*(1), 1-44.

Caron, D. A., Hu, S. K. (2018). Are We Overestimating Protistan Diversity in Nature?. *Trends in Microbiology, 27*(3), 197-205.

Chamnansinp, A., Moestrup, Ø., Lundholm, N. (2015). Diversity of the marine diatom *Chaetoceros* (Bacillariophyceae) in Thai waters–revisiting *Chaetoceros compressus* and *Chaetoceros contortus*. *Phycologia*, *54*(2), 161-175.

Chapman, A. D. (2005). Uses of primary species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available at http://www.niobioinformatics.in/books/Uses%20of%20Primary%20Data.pdf (accessed 14 October 2018).

Culverhouse, P. F. (2007). Human and machine factors in algae monitoring performance. *Ecological Informatics, 2*(4), 361-366.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science, 348*(6237), 1261605.

Devictor, V., Whittaker, R. J., Beltrame, C. (2010). Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions, 16*(3), 354-362.

Droege S., Cyr, A., Larivée, J. (1998). Checklists: An Under-Used Tool for the Inventory and Monitoring of Plants and Animals. *Conservation Biology, 12*(5), 1134-1138.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., ... Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129-151.

Foucault, M. (1970). *The Order of Things*. Random House, New York.

Gaonkar, C. C. (2017). *Diversity, Distribution and Evolution of the Planktonic Diatom Family Chaetocerotaceae*. PhD thesis, The Open University.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. *PLoS ONE*, *13*(12), e0208929.

Gould, S. J., Vrba, E. S. (1982). Exaptation—a missing term in the science of form. *Paleobiology*, *8*(1), 4-15.

Graur, D., Li, W. H. (1999). *Fundamentals of Molecular Evolution. Second edition.* Sinauer Associates, Sunderland, Massachusetts.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M. N., Schmidtlein, S., ... Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models?. *Remote Sensing in Ecology and Conservation*, *1*(1), 4-18.

Hernández-Becerril, D. U. (1993). Note on the morphology of two planktonic diatoms: *Chaetoceros bacteriastroides* and *C. seychellarus*, with comments on their taxonomy and distribution *Botanical Journal of the Linnean Society*, *111*(2), 117-128.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W. K., Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), 130-137.

Kooistra, W. H. C. F., Sarno, D., Hernández-Becerril, D. U., Assmy, P., Di Prisco, C., Montresor, M. (2010). Comparative molecular and morphological phylogenetic

analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia, 49*(5), 471-500.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., … Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience, 4*(1), 27 DOI 10.1186/s13742-015-0066-5.

Leray, M., Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1702), 20150331.

Li, Y., Lundholm, N., Moestrup, Ø. (2013). *Chaetoceros rotosporus* sp. nov. (Bacillariophyceae), a species with unusual resting spore formation. *Phycologia, 52*(6), 600-608.

Li, J., Kociolek, J. P., Gao, Y. (2016). *Chaetoceros coloradensis* sp. nov. (Bacillariophyta, Chaetocerotaceae), a new inland species from Little Gaynor Lake, Colorado, North America. *Phytotaxa, 255*(3), 199-213.

Longhurst, A. (2007). *Ecological geography of the sea*. Academic Press, London.

Lütolf, M., Kienast, F., Guisan, A. (2006). The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology, 43*(4), 802-815.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., … Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences, 113*(11), 1516-1525.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution and inheritance*. Harvard University Press, Cambridge, Massachusetts.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H. C. F., Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. *PLoS One, 9*(8), e103810.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography, 34*(1), 3-22.

Pargana, A. (2017). *Functional and Molecular Diversity of the Diatom Family Leptocylindraceae*. PhD thesis, The Open University.

Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H., Possingham, H. P. (2006). Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters, 9*(10), 1136-1145.

Smayda, T. J. (2011). Cryptic planktonic diatom challenges phytoplankton ecologists. *Proceedings of the National Academy of Sciences*, *108*(11), 4269-4270.

Tragin, M., Vaulot, D. (2018). Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. *Scientific Reports, 8*(1), 14020.

Tragin, M., Vaulot, D. (2019). Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Scientific Reports, 9*(1), 5190.

Xu, X. J., Chen, Z. Y., Lundholm, N., Li, Y. (2019). Diversity in the section *Compressa* of the genus *Chaetoceros* (Bacillariophyceae), with description of two new species from Chinese warm waters. *Journal of Phycology*, *55*(1), 104-117.